Overview of the Web Retrieval Task at the Third NTCIR Workshop

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

† National Institute of Informatics (NII)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{eguchi, oyama, emi, kando}@nii.ac.jp

‡ Shirayuri College
1-25 Midorigaoka, Chofu-shi, Tokyo 182-8525, Japan
kuriyama@shirayuri.ac.jp

Abstract

This paper gives an overview of the Web Retrieval Task that was conducted from 2001 to 2002 at the Third NTCIR Workshop. In the Web Retrieval Task, we attempted to assess the retrieval effectiveness of Web search engine systems using a common data set, and built a re-usable test collection suitable for evaluating Web search engine systems. With these objectives, we constructed 100-gigabyte and 10-gigabyte document data that were mainly gathered from the '.jp' domain. Participants were allowed to access those data only within the 'Open Laboratory' located at the National Institute of Informatics. Relevance judgments were performed on the retrieved documents, which were written in Japanese or English, by considering the relationshiop between the pages referenced by hyperlinks. Some evaluation measures were also applied to individual system results submitted by the participants.

Keywords: *Evaluation Method, Test Collection, Web Information Retrieval.*

1 Introduction

This paper gives an overview of the Web Retrieval Task at the Third NTCIR Workshop ('NTCIR Web Task') [5, 6, 15]. The essential objective of the NTCIR Web Task was 'to research the retrieval of large-scale Web document data that have a structure composed of tags and links, and that are written in Japanese or English'. Using this NTCIR Web Task, we built a re-usable test collection that was suitable for evaluating Web search engine systems, and evaluated the retrieval effectiveness of a certain number of Web search engine systems. In this work, the test collection was composed of: a document set, the topics —*i.e.*, statements of information needs—, and a list of relevance judgment results for each topic.

The overall task design was considered from the Web retrieval aspects, as described in Section **2**.

The participants created queries using selected topics, and performed searches using the topics from 100-gigabyte and/or 10-gigabyte document data that were mainly gathered from the '.jp' domain, and then submitted the run results to the organizers. At that time, the participants were allowed to process those data only within the 'Open Laboratory' located at the National Institute of Informatics. The organizers then assessed the relevance of the run results. The details of the document set, the topics, and the relevance assessment are described in Section **3**.

Using the run results and the relevance assessment result, the organizers evaluated the overall effectiveness of the search systems. The evaluation measures were also considered from the aspects of Web retrieval, as described in Section **4**.

Sixteen groups enrolled to participate in the NTCIR Web Task, and seven of these groups ('active participating groups') submitted the run results. A summary of the participation and the evaluation of the results can be found in Section **5**.

The NTCIR Web Task was carried out according to the following schedule:

Aug. 1, 2001 call for participation

Jan. 15, 2002 access permissions granted for the document set

Feb. 8, 2002 distribution of the dry-run topics

Feb. 18, 2002 submission of the dry-run results

Apr. 15, 2002 distribution of the relevance judgment results and evaluation results of the dry-run

Apr. 25, 2002 distribution of the formal-run topics

May 13, 2002 submission of the formal-run results

Aug. 6, 2002 distribution of the relevance judgment results and evaluation results of the formal run

Oct. 8-10, 2002 workshop meeting and round-table discussion

The dry-run was performed so that the participants and organizers —*i.e.*, the authors of this paper— could gain experience in the procedure for the NTCIR Web Task using a small number of topics, as this is our first attempt at the Web retrieval evaluation workshop.

2 Task Description

The NTCIR Web Task was composed of the following tasks for the two document data types composed of: (I) 100 gigabytes, and (II) 10 gigabytes, respectively.

(A) 'Survey Retrieval Tasks'
- **(A1)** 'Topic Retrieval Task'
- **(A2)** 'Similarity Retrieval Task'

(B) 'Target Retrieval Task'

(C) 'Optional Tasks'
- **(C1)** 'Search Results Classification Task'
- **(C2)** 'Speech-Driven Retrieval Task'

The objectives and procedures of the Survey Retrieval Tasks and Target Retrieval Task are described in Sections **2.1** and **2.2**, respectively. We describe in Section **2.3** an overview of the 'Search Results Classification Task' and 'Speech-Driven Retrieval Task,' which are parts of the Optional Tasks.

2.1 Survey Retrieval Tasks

The Survey Retrieval Tasks assumed the user model where the user attempted to comprehensively find documents relevant to his/her information needs. Three types of query were supposed: query term(s) and sentence(s) as discussed in Section **2.1.1**, and query document(s) as discussed in Section **2.1.2**.

2.1.1 Topic Retrieval Task

The Topic Retrieval Task is similar to a traditional adhoc retrieval against scientific documents or newspapers, etc. [17, 12, 10], 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 also optionally submit their run results using other topic fields. 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. 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.

The participating groups were also asked to submit 'evidential passages', *i.e.*, those parts of each retrieved document that provided evidence by which the search system computed the relevance, although the submission of evidential passages was not made mandatory. We considered that the evidential passages might be useful for a complementary evaluation. Unfortunately no evidential passages were submitted with the results.

2.1.2 Similarity Retrieval Task

The Similarity Retrieval Task was a new task, with the objectives of (i) evaluating the similarity search methods driven by one query document with using context given by query, and (ii) evaluating the relevance feedback methods driven by a few training documents suitable for the Web environment.

In the Similarity Retrieval Task, we specified mandatory and optional runs as follows:

mandatory The first term specified by ⟨RDOC⟩ tag in the topic had to be used. The ⟨TITLE⟩ tag could also be used. The second and third terms in the ⟨RDOC⟩ tag could not be used for the query.

optional The first term in the ⟨RDOC⟩ tag had to be used. The ⟨TITLE⟩ tag could also be used. The second and third terms in the ⟨RDOC⟩ tag could also be used.

The details of the topic formats, such as the ⟨RDOC⟩ tag, and the others used can be found in Section **3.2.1**. The mandatory conditions were used for Objective (i), and the optional conditions were used for Objective (ii).

The methodology for results submission was the same as that for the Topic Retrieval Task. The relevance judgments were performed using the criteria contained only in the entire statement of a topic, and not by the contents of certain specified relevant document(s) in the ⟨RDOC⟩ tag of the topic, as described in Section **3.4**.

Another interesting point was (iii) to evaluate similarity searching by assessing one query document without using the context given by the query. However, in this case, relevance judgments should be performed using only the criteria given by the contents

of the specified relevant document(s) in ⟨RDOC⟩. We did not adopt (iii) because the relevance judgments are more expensive.

2.2 Target Retrieval Task

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. The mandatory runs were the same as those of the Topic Retrieval Task. Several evaluation measures were applied, as is described in Section **4**. The methodology of results submission was almost the same as that used in the Topic Retrieval Task, except that the number of the retrieved documents submitted was 20, and not 1,000.

This task description was different from the TREC High Precision Tracks [3, 4], in which an assessor was asked to find the 10 (or 15) most relevant documents possible within five minutes for each topic. However, the final goal of our task was somewhat similar to the one for the TREC High Precision Tracks from the point of view that the precision of top-ranked documents was important.

While previous TREC Web Tracks [10, 8, 9] performed evaluations using precision at top-ranked document levels as well as precision-recall-related measures at cutoff levels of 100 or more, we performed the Survey Retrieval Task and the Target Retrieval Task separately. Therefore, the number of topics could be larger for the Target Retrieval Task, because relevance judgments for the Target Retrieval Task are not generally more expensive than those for the Survey Retrieval Tasks. Moreover, the search systems focused on the precision of the top-ranked documents, so those focused on the comprehensiveness of relevant documents could be evaluated separately.

2.3 Optional Tasks

The participants could freely submit proposals relating to their own research interests using the document sets contained in the above tasks. These proposals were adopted as one of the tasks, and were investigated in detail if they involved several participants. Consequently, two 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 textual documents on the Web.

2.3.1 Search Results Classification Task

The Search Results Classification Task tried to evaluate techniques for supporting user navigation by means of classification-based output presentation when the user submits very short queries, *e.g.*, only one term.

The participants were expected to perform searches using only the lead term specified in the ⟨TITLE⟩ of the topic, classify the search results into some labeled groups, and then submit the first 200 resulting documents. The classification processing could be performed on more than the top 200 documents retrieved.

For example, when using 'Hidetoshi Nakata', who is a famous Japanese soccer player, as the query, the results were supposed to be classified into 'sites', 'schedules', 'magazines or TV programs', 'photographs' and 'supporters' diaries'. We did not set a limit on the number of classes. Hierarchical classification was also acceptable. The labels of the classes could be topical terms that represented the classification, typical page titles, or machine-like identification codes, *e.g.*, 'cluster A' and 'cluster B'.

In evaluating the Search Results Classification Task, we considered the following aspects of the evaluation method:

- whether the classifications are easily understood or not
- the number of classes
- the number of documents included in each class
- the relevance of each class to the documents in it
- the number of classes that include the relevant documents and their distribution
- whether the required information can be found or not.

However, very unfortunately, no classification results were submitted.

2.3.2 Speech-Driven Retrieval Task

The systems to be evaluated were driven by spoken queries that were created by reading the topics aloud, and searching against the Web documents. This task was proposed by Fujii and Itou, and the details of the task description and the evaluation results can be found in reference [7].

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 had to be explicitly specified for the test collections. As our first challenge of the Web retrieval evaluation workshop, we adopted the following method to construct several possible collections.

- Extract a part of the crawled Web pages, and then define a set of URLs as document data that will be used for searching.
- Provide the document data for searching.

As this method is the same as that of conventional test collections, many well-known techniques can be utilized to identify relevant document sets and to evaluate systems, and this made the constructed test collection re-usable.

In the NTCIR Web Task, we prepared two types of document data gathered from the '.jp' domain: (i) document data over 100 gigabytes ('NW100G-01'), and (ii) 10-gigabyte document data ('NW10G-01'). Almost all the documents were written in Japanese or English, but some were written in other languages. We also provided two separate lists of documents that were connected from the individual documents included in the NW100G-01 and NW10G-01 data, respectively, but not limiting to the '.jp' domain. These four data sets were used for searching in the NTCIR Web Task.

The crawling strategies are described in Section **3.1.1**, and the definition of the document set is described in Section **3.1.2**. The participants were allowed to process the NW100G-01 and NW10G-01 data only inside the 'Open Laboratory' located at the National Institute of Informatics in Japan, as described in Section **3.1.3**.

3.1.1 Crawling

A crawler fetches Web pages to construct the document data, and then accumulates them, following the links in each page that has been fetched. It keeps track of the URLs it has yet to follow, and the URLs it has already tracked. The crawling was performed from August 29, 2001 to November 12, 2001 under the following conditions:

Web sites HTTP servers on the '.jp' domain[1]

ports Any

file formats HTML files or plain text files; The file formats were detected by 'Content-Type' information in individual HTTP headers and Web pages.

Firstly, the crawling program[2] discovered Web sites to be fetched, and then tried to fetch the Web pages from them. The following crawling strategies were applied under the previously mentioned conditions of Web sites, ports and file formats.

Web site discovery

(1) Specify the starting point. We specified it as the entrance page of the National Institute of Informatics ⟨http://www.nii.ac.jp/⟩.

(2) Extract links from the root page of a discovered site, and try to fetch 20 pages.

(3) Detect links out of the fetched pages.

(4) Extract newly discovered sites. A site was identified by the host name, not the IP address.

(5) Discard alias sites and non-working ones.

(6) Recursively and concurrently perform steps (2)–(5) until the discovery rates become relatively small.

Web page fetching

(1) For each site discovered in the process above, add the root page's URL to a URL list.

(2) Try to fetch the page at the top of the URL list.

(3) Discard and go to step (2) if it is a duplicated page or an inappropriate page according to the previously mentioned conditions.

(4) Extract links from the fetched page that are connected to pages on the same site.

(5) Add the newly discovered pages to the end of the URL list.

(6) Perform steps (2)–(5) until the number of fetched pages reaches 2,000 or the URL list ends[3].

Here, we identified a 'root page' by describing a host name and its port number suffixed by a slash, expecting them to be URL strings. While crawling, we discarded the following kinds of pages:

- pages that obstructed building a document collection. These include non-text pages such as images and archive files whose content types were indicated as plain text in error."
- pages that caused looped paths

[1] After crawling the Web pages, we extracted the links from the fetched pages —these were not limited to the '.jp' domain—, and expanded the document set as described in Section **3.1.2**. Consequently, the document data for searching mainly came from the '.jp' domain, but not exclusively so.

[2] We used 'Livelink Spider' provided by the Open Text Corporation, as the crawling program.

[3] This limitation was determined from our experience. After fetching Web pages, we adjusted the maximum number of pages within a site to 1,300 for the 100-gigabyte data set or 20 for the 10-gigabyte data set.

- dynamically generated pages. However, we did not discard the first 10 of these.

3.1.2 Definition of Document Set and Document Data

We extracted a subset of the Web pages gathered through crawling, and using this, we constructed the document data to be provided to the participants. Moreover, we extracted the links from the document data, not limiting them to the '.jp' domain, checked existence of the pages, and created a list of URLs to expand the document data. Consequently, we defined the 'document set' as having the two following components: (1) 'document data for providing,' that is, 100-gibabyte NW100G-01 and 10-gigabyte NW10G-01 data, and (2) 'document data for reference'. We built two sets of the reference document data (2) consisted of pages that were connected from any of the documents included in the NW100G-01 and NW10G-01 data, respectively. They could be used for link analysis and, consequently, included in the search results. The small-scale document data, *i.e.*, NW10G-01, was a subset of the large-scale document data, NW100G-01.

The statistical characteristics of NW100G-01 and NW10G-01 are shown in **Tables 1**, **2**, **3** and **4**.

For NW100G-01 and NW10G-01, the following files are provided:

- the page data and the metadata, *i.e.*, the fetched URL, the time spent crawling, the http headers, etc.; one file per site
- a list of the crawled sites
- a list of the alias sites
- a list of the crawled pages
- a list of the duplicated pages
- a list of the referenced pages, *i.e.*, pages in the 'document data for reference'
- a list of the links from the crawled pages to the crawled and referenced pages

We defined a 'Web document' as being an individual page datum and its metadata. Hereafter, in this paper, we will simply refer to a Web document in general terms as being a 'document', unless specifying a particular type. Each field of the Web document was flanked by a pair of tags having one of the meanings defined below. A sample of such Web documents can be seen in **Figure 1**.

- ⟨NW:DOC⟩ specifies the boundary of a Web document.
- ⟨NW:META⟩ indicates the metadata that includes the followings:
  - ⟨NW:DOCID⟩ indicates the document identification number.

```
⟨NW:DOC⟩
⟨NW:META⟩
⟨NW:DOCID⟩NW010616091⟨/NW:DOCID⟩
⟨NW:DATE⟩Mon, 05 Nov 2001 09:46:11 GMT ⟨/NW:DATE⟩
⟨NW:CTYPE⟩text/html⟨/NW:CTYPE⟩
⟨NW:URL⟩http://www.nii.ac.jp/⟨/NW:URL⟩
⟨NW:HTTPH⟩HTTP/1.1 200 OK
Date: Tue, 06 Nov 2001 02:24:19 GMT
Server: Apache/1.3.14 (Unix)
Last-Modified: Mon, 05 Nov 2001 09:46:11 GMT
ETag: "ae663-4dce-3be65fe3"
Accept-Ranges: bytes
Content-Length: 19918
Connection: close
Content-Type: text/html
⟨/NW:HTTPH⟩
⟨/NW:META⟩
⟨NW:DATA⟩⟨NW:DSIZE⟩19852⟨/NW:DSIZE⟩  ⟨!DOCTYPE
HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"⟩
⟨HTML lang="ja"⟩
⟨HEAD⟩
⟨META HTTP-EQUIV="Content-Type" CONTENT= "text/html;
charset=ISO-2022-JP"⟩
⟨TITLE⟩NII -The National Institute of Informatics-⟨/TITLE⟩
  .
  .
  .
⟨/HTML⟩
⟨/NW:DATA⟩
⟨/NW:DOC⟩
```

Figure 1. A sample Web document from NW100G-01

  - ⟨NW:DATE⟩ indicates the crawled time.
  - ⟨NW:CTYPE⟩ indicates the 'Content-Type', *i.e.*, 'text/html' or 'text/plain'.
  - ⟨NW:URL⟩ indicates the URL strings.
  - ⟨NW:HTTPH⟩ indicates the HTTP headers.
- ⟨NW:DATA⟩ indicates the page data that started by ⟨NW:DSIZE⟩ and followed by original contents of the page.
  - ⟨NW:DSIZE⟩ indicates the size of the page data represented in bytes.

Web documents gathered from the same site were bundled in a file. The following three versions of document data were provided to the participants:

- one that consisted of the original page data without any processing
- one that had all the page data described in Japanese character codes converted to the 'EUC' coding system
- a cooked one in which HTML tags and the commented-out parts had been eliminated from all the page data. Keywords specified by META tags were retained but marked with an indicator at the head of the line.

No other data preprocessing was performed on the document data.

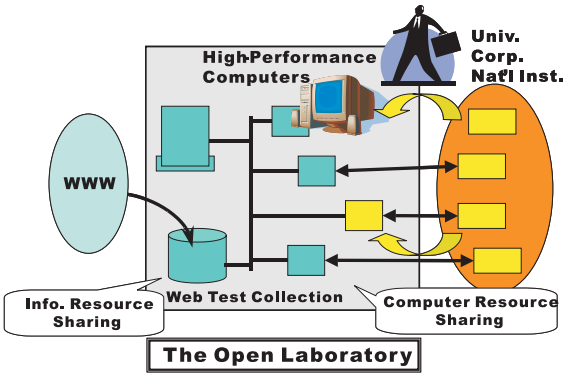


Figure 2. A Framework of the Open Laboratory

3.1.3 The Open Laboratory

Participants were allowed to use the document sets only within the National Institute of Informatics (NII), because the data sets were too large to handle easily and there were some restrictions on the delivery of the original Web contents. Participants used the computer resources in the 'Open Laboratory' located at NII to perform data processing[4], *e.g.*, indexing of the original document data. Participants could take the resulting data, *e.g.*, index files, and perform experiments on them in their own laboratories, as shown in **Figure 2**. Participants could thus join the NTCIR Web Task even though they did not have sufficient computer resources.

The following is a summary of the computer resources available at the Open Laboratory:

- a shared file server that provides the document data, etc.
- host computers: Sun Blade, Linux or Windows 2000 with 2-gigabyte memory; one for each participating group
- auxiliary storage: 500 gigabytes of storage for each participating group that uses the large-scale document data
- data backup facilities: DVD-R, magnetic tape equipment, etc.
- software: basic software according to the participants' requirement
- network environments; The individual host computers are connected to the Internet through an exclusive segment that is protected by a firewall.
- remote access; Remote access to the individual host computers is controlled by the firewall. Remote access from host computers to the outside is also controlled.
- take-in machines; We accept take-in machines as far as the space, the power supply, management conditions and other circumstances allow.

[4]To perform the data processing, remote access to the individual host computers in the Open Laboratory was allowed.

3.2 Topics

3.2.1 Topic Format

The organizers provide 'topics' that were statements of information needs rather than queries.

The topic format was basically inherited from previous NTCIR Workshops [12], except for the definitions of the ⟨TITLE⟩, ⟨RDOC⟩, and ⟨USER⟩ tags, and the format of the ⟨NARR⟩ tag. The usable fields and mandatory fields varied according to the tasks described in Section **2**. 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⟩ provided up to three terms that were specified by the topic creator, simulating the query terms in real Web search engines. The topic creator was instructed in advance not to be excessively conscious of individual features of search engine systems. The topic creator also selected one of the following three search strategies, deemed suitable for obtaining the needed information using search engines, and then, according to the selected strategy, the topic creator specified up to three terms for inputting into the search engine. The terms specified by the ⟨TITLE⟩ tag were listed in the order of importance for searching. The title has the attribute of 'CASE', which indicated the type of search strategy as follows:
  - **(a)** All of the terms had the same, or had strongly related meanings.
  - **(b)** All of the terms had different meanings that corresponded to different semantic categories.
  - **(c)** Only two of the three terms had the same meaning, or strongly related meanings, and the other term had a different meaning. They were specified by the attribute, 'RELAT' in ⟨TITLE⟩ tag.

  For example, participants using a Boolean search could use the OR operator for strategy (a), the AND operator for strategy (b), and a combination of the AND and the OR operators for strategy (c).
- ⟨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

⟨TOPIC⟩
⟨NUM⟩0004⟨/NUM⟩
⟨TITLE CASE="c" RELAT="2-3"⟩ コンピューターウィルス, 予防, 対策 ⟨/TITLE⟩
⟨DESC⟩ コンピューターウィルスの予防方法や対策法について説明している文章を探したい ⟨/DESC⟩
⟨NARR⟩⟨BACK⟩ インターネット利用が爆発的に普及する中でコンピュータウィルスは日常的な問題にまで近づいてきている。そこでどのような予防法をととり、もし感染してしまった際にはどのような対処をすればよいのか知っておきたい。⟨/BACK⟩⟨RELE⟩ 適合文書は、コンピュータウィルスへの予防・対策についての情報を提供するもの。被害届出やウィルスの種類についてのみ述べているものは適合としない。特定のウィルスについてのみ情報を提供するページは部分的適合とする。⟨/RELE⟩⟨/NARR⟩
⟨CONC⟩ コンピュータウィルス, ワーム, 情報セキュリティ, 不正アクセス, 予防, 対策, 感染 ⟨/CONC⟩
⟨RDOC⟩NW003214039, NW013338047, NW013315769
⟨/RDOC⟩
⟨USER⟩ 大学院修士 1 年, 男性, 検索歴 5 年 ⟨/USER⟩
⟨/TOPIC⟩

(a) An original sample topic

⟨TOPIC⟩
⟨NUM⟩0004⟨/NUM⟩
⟨TITLE CASE="c" RELAT="2-3"⟩computer virus, preventive, countermeasure⟨/TITLE⟩
⟨DESC⟩I want to find sentences that explain preventives or countermeasures against computer viruses.⟨/DESC⟩
⟨NARR⟩⟨BACK⟩Because the use of the Internet has spread explosively, computer viruses have become a serious problem in our daily lives. I want to know what kind of preventives are required, and what kind of countermeasures I should take when my computer becomes infected.⟨/BACK⟩ ⟨RELE⟩Relevant documents must provide some information on preventives or countermeasures against computer viruses. Documents that describe only the victim's reports or the types of computer viruses are regarded as not relevant. Pages that provide some information on a particular virus are regarded as partially relevant.⟨/RELE⟩⟨/NARR⟩
⟨CONC⟩computer virus, worm, information security, illegal access, preventive, countermeasure, infection⟨/CONC⟩
⟨RDOC⟩NW003214039, NW013338047, NW013315769
⟨/RDOC⟩
⟨USER⟩1st year of Master Course, Male, 5 years of search experience⟨/USER⟩
⟨/TOPIC⟩

(b) An English translation of a sample topic

Figure 3. A sample topic for NTCIR Web Task (dry-run) and its English translation

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.

- ⟨CONC⟩ ('concepts') provided the synonyms, related terms, or broader terms that were defined by the topic creator.
- ⟨RDOC⟩ ('given relevant documents') provided document identification numbers of up to three relevant documents that were used for the 'Similarity Retrieval' method described in Section **2.1**. To describe ⟨RDOC⟩, the topic creator first selected up to three relevant documents from the ranked documents retrieved by the organizer's search system.
- ⟨USER⟩ ('user attributes') provided the attributes of the topic creator, *i.e.*, job title, gender, and search experience.

All of the above topics were written in Japanese. A topic example and its English translation are shown in **Figure 3**.

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 discarded 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 '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⟩.
- To describe ⟨RDOC⟩, the topic creator selected the three most relevant pages out of the top 20 results retrieved by an organizer's search system. This process was performed before we delivered the topics to the participating groups.

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 organizers' search system to discard inappropriate topics before delivering topics. As a result, 140 topics were

selected for the formal run, and we delivered them to the participants[5].

Second, we performed 'shallow pooling', which is a sampling method that takes the 20 highest-ranked documents from each run result submitted by a participant [6], ranking them in order of a meta-search-engine strategy. We applied the 'Borda Count' voting algorithm [1] as our ranking strategy. By assessing the relevance of each document included in the 'shallow pool,' we discarded 35 topics and used the remaining 105 topics for the next step.

Third, we carefully assessed the relevance of the document set obtained through 'deep pooling', which will be described in Section **3.3.2**. As a result, we decided to discard nine topics for reasons such as having few relevant documents, and tried to use the remaining 96 topics for the evaluation of individual run results.

We planned to evaluate the Target Retrieval Task in the formal run using all 96 topics, and the Survey Retrieval Tasks using only 47 topics, about the half of the 96 topics. Unfortunately, however, for unexpected reasons, it was hard to perform the relevance assessment of the Target Retrieval Task for some of the 96 topics. Consequently, we used the 47 topics[6] for evaluating both the Survey Retrieval Tasks and the Target Retrieval Task.

3.3.2 Deep Pooling

Using the topics of the Survey Retrieval Task, we perform '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 conventional information retrieval evaluation workshops [17, 10, 12]. 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 Survey Retrieval Tasks.

In the pooling task, we took the top 100 ranked documents from each run results. Moreover, we performed ranking the pooled documents in order of the meta-search-engine strategy, using the same process as in the shallow pooling stage.

We did not perform any additional manual searches to improve the comprehensiveness of relevant documents [12, 14]. However, the organizers ran their own search system and added the run results to the participants' run results, attempting to improve the comprehensiveness of the pool.

[5]We delivered seven topics to the participants for the dry-run.

[6]The identification numbers of the 47 topics were: 0008, 0010, 0011, 0012, 0013, 0014, 0015, 0016, 0017, 0018, 0019, 0020, 0022, 0023, 0024, 0027, 0028, 0029, 0030, 0031, 0032, 0033, 0034, 0035, 0036, 0037, 0038, 0039, 0040, 0041, 0042, 0043, 0044, 0046, 0047, 0048, 0049, 0052, 0053, 0056, 0057, 0058, 0059, 0060, 0061, 0062 and 0063.

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. Human assessors judged the relevance of each document in the pool, using the multiple document models described in Section **3.4.1**, using an assessing system as described in **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**. In addition, they chose the best documents, as described in **3.4.4**, and made other assessments from other aspects, as described in Section **3.4.5**.

3.4.1 Document Models

Web pages are represented in various ways, so that in one example, an 'information unit' on the Web could be hyper-linked pages, while in another, it could be an individual page, or a passage included on a page.

Previous Web retrieval evaluation workshops assumed an information unit on the Web to be a page [10, 8, 9]. According to this assumption, a 'hub page' [13] 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.

The NTCIR Web Task attempted to incorporate two other assumptions into the relevance assessment. These assume that hyper-linked pages or a passage are an information unit, so we defined the following three 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.

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.

Passage-unit document model This was where the assessor specified the passages that provided

evidence of relevance, which he/she used to judge the passages relevant.

3.4.2 Assessment System

The assessment system that we used in the NTCIR Web Task ran on our HTTP server, and was available through CGIs. All the pooled documents to be assessed were ranked by a meta-search-engine strategy, as described in **3.3.2**, 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.

3.4.3 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—, so that we also refered to these kinds of relevance as 'absolute relevance'. 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.

3.4.4 Relative Relevance

Voorhees found little agreement between multiple assessors' judgments concerning the best document on a topic, and pointed out that the evaluation using the best documents was less stable [18]. However, the best documents were important, considering the ways in which real Web search engine systems are used. Trying to relax the aforementioned problems, we assessed multiple best documents for each topic rather than one for each topic, although we assigned one assessor for each topic.

The assessors chose, out of the pooled documents, a small number of documents that were most relevant to the statement of the topic with priority of relevance: *e.g.*, the best, the second- best and the third-best documents. These best documents should not be duplicated, strongly similar or linked to each other. The assessors also found those documents that duplicated, were strongly similar to or linked from/to any of the best documents as long as possible. We refer to these kinds of relevance as 'relative relevance' in contrast to absolute relevance.

The results of relative relevance judgments can be used for the weighted reciprocal rank measure as described in Section **4.3**.

Moreover, we administered a questionnaire to all the assessors asking why they chose each of the three best documents, and whether there were any reasons besides relevance to the statement of a topic, such as the following:

- the amount of relevant information
- a degree of detail on the relevant information
- reliability of the relevant information or the page
- freshness of the relevant information or the page
- readability of the relevant information or the page
- richness of the hyper-links to pages that give useful information related to relevant information
- others

3.4.5 Additional Assessment

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. The assessors judged the relevance of the pooled documents only on the basis of the information given in Japanese or English. Moreover, the assessors judged the duplication, coherence and reliability of the documents for further investigation, as follows:

Not Japanese or English The assessors found those documents with page data not described in Japanese or English.

Duplication The assessors found as many duplicated documents corresponding to the page data of relevant documents as possible.

Coherence The assessors judged the coherence of the relevant documents by classifying documents as less coherent if the topic-related content of the document was less than one third of the entire document.

Reliability The assessors judged the reliability of the relevant documents in terms of whether or not the relevant information in them was reliable according to their knowledge. At that time, they were allowed to consider whether or not the page data seemed to give reliable information on the basis of the entire page data or by the name of the organization, which could sometimes be determined from URL strings. The assessors also reported the reasons why they judged documents as less reliable.

4 Evaluation Measures

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 20 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, and (ii) those with discounted cumulative gain. For the Target Retrieval Task, we applied the three types of measures: the aforementioned measures in (i) and (ii), and weighted reciprocal rank measure ((iii)).

Although the one-click-distance document model was partly applied in the relevance assessment, as described in Section **3.4.1**, 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 document sets, according to which of the two document models was used.

4.1 Precision and Recall

As an evaluation measure for the run results of the Survey Retrieval Tasks, we used the 'average precision (non-interpolated)' measure taken over all the relevant documents, and the 'R-precision', *i.e.*, the precision after $|R|$ documents were retrieved, where $|R|$ indicates the number of relevant documents for each topic. We also computed the 'recall-level precision' for 11 points of recall, and the 'document-level precision' after 5, 10, 15, 20, 30, and 100 documents were retrieved, respectively [2]. These evaluation measures[7] have been used in conventional information retrieval evaluation workshops [17, 12, 10].

On the other hand, as a measures for the Target Retrieval Task, we used the document-level precision after 5, 10, 15, and 20 documents were retrieved.

The aforementioned measures based on precision and/or recall often required the multi-grade relevance to be mapped into binary relevance, so that we supposed the following two relevance levels in using the measures:

Relevance level 0 We considered the document to be relevant if it was highly relevant, and otherwise considered it to be irrelevant.

Relevance level 1 We considered the document to be relevant if it was highly relevant or fairly relevant, and otherwise considered it to be irrelevant.

Relevance level 2 We considered the document to be relevant if it was highly relevant, fairly relevant, or partially relevant. Otherwise, we considered it to be irrelevant.

[7]These 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⟩.

We computed the mean values of the abovementioned measures over all the topics for each run result according to Relevance levels 1 and 2, omitting level 0 because the number of highly relevant documents was small for some topics. Relevance level 1 can be regarded as a rigid criterion, and level 2 a relaxed one.

4.2 Discounted Cumulative Gain

We adopted 'Discounted Cumulative Gain' measure [11, 18] ('DCG') as one of the evaluation measures suitable for multi-grade relevance. The DCG is represented by the following equations:

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

$$\mathrm{g}(i) = \begin{cases} h & \text{if } \mathrm{d}(i) \in H \\ a & \text{if } \mathrm{d}(i) \in A \\ b & \text{if } \mathrm{d}(i) \in B \end{cases} \quad (2)$$

where $\mathrm{d}(i)$ indicates the i-th-ranked document, and H, A and B indicate the sets of highly relevant, fairly relevant, and partially relevant documents, respectively. We set the magnitude of the gain indicated in Equation (2) to the following two relevance levels[8] :

Rigid level $(h, a, b) = (3, 2, 0)$,

Relaxed level $(h, a, b) = (3, 2, 1)$.

We set the base of the logarithmic function as $b = 2$ in Equation (1). The DCG was derived from the 'cumulative gain' measure [11], as indicated in Equation (3), and modified in that the gain $\mathrm{g}(i)$ at rank i was discounted as being divided by a logarithmic rank i.

$$\mathrm{cg}(i) = \begin{cases} \mathrm{g}(1) & \text{if } i = 1 \\ \mathrm{cg}(i-1) + \mathrm{g}(i) & \text{otherwise}. \end{cases} \quad (3)$$

For each run result, we computed the DCG value by the 1,000th-ranked document for the Survey Retrieval Task, and by the 20th-ranked document for the Target Retrieval Task, and then calculated the mean values of the DCG at the respective rank over all the topics.

4.3 Weighted Reciprocal Rank

The 'Mean Reciprocal Rank' measure [16] ('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. In the NTCIR Web Task, we proposed a new measure, the 'Weighted Reciprocal Rank' ('WRR') as the mean value of the wrr, defined by the following equations over all the topics:

$$\mathrm{wrr}(m) = \max\left(\mathrm{r}(m)\right) , \quad (4)$$

[8]The rigid and relaxed levels respectively correspond to Relevance levels 1 and 2 as described in Section **4.1**.

$$\mathrm{r}(m) = \begin{cases} \delta_h / (i - 1/\beta_h) & \text{if } (\mathrm{d}(i) \in H \ \wedge \ 1 \leq i \leq m) \\ \delta_a / (i - 1/\beta_a) & \text{if } (\mathrm{d}(i) \in A \ \wedge \ 1 \leq i \leq m) \\ \delta_b / (i - 1/\beta_b) & \text{if } (\mathrm{d}(i) \in B \ \wedge \ 1 \leq i \leq m) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

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.

WRR is one of the generalized measures of MRR that is suitable for the multi-grade relevance. In other words, MRR is a special case of WRR with binary relevance. Therefore, in Equation (5), the term $(-1/\beta_x)$, $x \in \{h, a, b\}$ can be omitted when the value of β_x is sufficiently large.

We computed the WRR values under the conditions where m was set to 5, 10, 15, and 20, and the combinations of δ_x and β_x were set as below, supposing that two of the relevance levels indicated in Section **4.1** applied.

Relevance level 1 $(\delta_h, \delta_a, \delta_b) = (1, 1, 0)$, $(\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$

Relevance level 2 $(\delta_h, \delta_a, \delta_b) = (1, 1, 1)$, $(\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$

We also computed the number and percentage of topics for which no relevant documents were retrieved under the conditions of respective cut-off levels of 5, 10, 15, and 20, with the two relevance levels mentioned.

It should be noted that Eqs. (4) and (5) should be replaced by the following for evaluations using a small number of the best documents that were obtained by the assessment of relative relevance, as described in Section **3.4.4**.

$$\mathrm{wrr}(m) = \max_{1 \leq j \leq k} \left(\max \left(\mathrm{r}_j(m) \right) \right) , \quad (6)$$

or

$$\mathrm{wrr}(m) = (1/k) \sum_{1 \leq j \leq k} \alpha(j) \cdot \max \left(\mathrm{r}_j(m) \right) , \quad (7)$$

$$\mathrm{r}_j(m) = \begin{cases} 1 / (i - 1/\beta(j)) & \text{if } (\mathrm{d}(i) \in BEST_j \\ & \qquad \wedge \ 1 \leq i \leq m) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $BEST_j$ indicates the set of the j-th best document and its related documents, *i.e.*, those that were duplicated, were strongly similar to or out-linked from the j-th best document, and $\alpha(j)$ and $\beta(j)$ indicate the weight function that should satisfy $0 \leq \alpha(j+1) \leq \alpha(j) \leq 1$, and $\beta(j+1) \geq \beta(j) > 1$, respectively. In Eq. (8), the term $(-1/\beta(j))$ can be omitted when the value of $\beta(j)$ is sufficiently large.

4.4 An Evaluation Method Considering Duplication

When duplicate pages appear in the Web search engine results, they are often unwelcome for users. We proposed an evaluation method that considers duplication, as follows:

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

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

We designed this evaluation method by supposing it to be combined with the precision-recall-related measures described in Section **4.1**, or the DCG measure described in Section **4.2**.

The assessor judged not only the relevance but also the duplication on the documents in the pool. We also detected the completely duplicated documents as the complement of the human-judged documents. This method can be used in the same way for documents that are strongly similar to each other, or connected by hyperlinks. Moreover, this evaluation method using groups of related documents is expected to be used for the evaluation of search engines using topic distillation techniques, by combining it with the relevance judgments according to the one-click-distance document model.

5 Evaluation Results

5.1 Summary of Participation

Six groups, listed below in alphabetical order of affiliations, submitted their completed run results[9], with the organizers also submitting the results from their own search system along with those of the participants in an attempt to improve the comprehensiveness of the pool.

- Nara Institute of Science and Technology, and Communication Research Laboratory
- NEC Corporation
- Osaka Kyoiku University
- University of Aizu
- University of Library and Information Science, and National Institute of Advanced Industrial Science and Technology
- University of Tokyo, and RICOH Co. Ltd.

The individual participating groups pursued various objectives. We summarize them as follows (listed in alphabetical order of group IDs):

[9] Although seven of the participating groups submitted run results, one group had submitted their run results for only half of the topics by the due submission date.

GRACE Experimented with pseudo-relevance feedback based on a probabilistic model, and re-ranking methods using link analysis based on Kleinberg's HITS.

K3100 Experimented with a retrieval method using not only the textual contents of a page but also the anchor text that pointed to the page or its site.

NAICR Integrated multiple retrieval results with and without score normalization. The retrieval module was based on OKAPI —a probablistic model approach—only using textual content of the Web documents.

OASIS Experimented on a distributed search system, where the document set was divided into 10 independent subsets. The retrieval module was based on a vector space model using only the textual content of the Web documents.

OKSAT Experimented with long gram-based indices using textual contents of Web documents. The retrieval module was based on a probabilistic model.

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

UAIFI Experimented with a speech-driven retrieval system, where speech recognition and text retrieval modules were integrated. The text retrieval module was based on a probabilistic model using only the textual content of the Web documents. The run results of the text retrieval module were included in the pool, but the speech-driven retrieval results were not.

Summaries of the run result submissions of each participating group can be found in **Table 5**, and the details can be found in papers of the participating groups in this proceedings.

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 (as described in Section **3.4.1**)

(DM_1) One-click-distance document model

(DM_2) Page-unit document model

Relevance Levels (as described in Section **4**)

(RL_1) Rigid relevance level This also means Relevance level 1.

(RL_2) Relaxed relevance level This also means Relevance level 2.

5.3 Summary of Evaluation Results

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 6** and **7**, 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 DM_2 and RL_1 for the Survey Retrieval Task, and the precision at 10 document-level in DM_2 and RL_1 for the Target Retrieval Task. In the tables, each evaluation values were averaged over all the 47 topics.

Selected recall-precision and DCG curves against the 100-gigabyte data set are also shown in **Figures 4**, **5**, **6** and **7**. The curves for the 10-gigabyte data set are also shown in **Figures 8**, **9**, **10** and **11**. 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 **2.1.1**. In each graph, the explanatory notes report the run ID codes, which are ranked in order of the average precision in the case of the Survey Retrieval Task, and the precision at 10 document-level in the case of the Target Retrieval Task. 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 Web Retrieval Task at the Third NTCIR Workshop. To evaluate the task, we have built 100-gigabyte and 10-gigabyte document sets that were mainly gathered from the '.jp' domain. Participants used the computer resources in the 'Open Laboratory' located at NII to perform data processing using the original document data. 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.

The evaluation using the results of some additional assessment is currently in progress. The detailed analysis of the evaluation results will be performed. One of our future tasks is to develop an evaluation using groups of related documents, *i.e.*, strongly similar to each other or connected by hyperlinks. We expect such an evaluation will be used on search engines using topic distillation techniques [13], by combining the relevance judgments according to the one-click-distance document model.

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] J. A. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 276–284, New Orleans, Louisiana, USA, Sep. 2001.

[2] R. Baeza-Yates, editor. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] C. Buckley. TREC 6 High-Precision Track. In E. Voorhees and D. K. Harman, editors, *Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pages 69–71. NIST Special Publication 500-240, 1996.

[4] C. Buckley. The TREC 7 High Precision Track. In E. Voorhees and D. K. Harman, editors, *Proceedings of the 7th Text REtrieval Conference (TREC-7)*, pages 57–63. NIST Special Publication 500-242, 1997.

[5] K. Eguchi, K. Oyama, E. Ishida, K. Kuriyama, and N. Kando. Evaluation design of Web Retrieval Task in the Third NTCIR Workshop. In *The 11th International World Wide Web Conference (WWW2002)*, number poster-22, Honolulu, Hawaii, USA, May 2002.

[6] K. Eguchi, K. Oyama, E. Ishida, K. Kuriyama, and N. Kando. The Web Retrieval Task and its evaluation in the Third NTCIR Workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, pages 375–376, Tampere, Finland, Aug. 2002.

[7] A. Fujii and K. Itou. Evaluating speech-driven IR in the NTCIR-3 Web Retrieval Task. In *Proceedings of the 3rd NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, to appear.

[8] D. Hawking. Overview of the TREC-9 Web Track. In E. Voorhees and D. K. Harman, editors, *Proceedings of the 9th Text REtrieval Conference (TREC-9)*, pages 97–112. NIST Special Publication 500-249, 2000.

[9] D. Hawking and N. Craswell. Overview of the TREC-2001 Web Track. In E. Voorhees and D. K. Harman, editors, *Proceedings of the 10th Text REtrieval Conference (TREC-2001)*, pages 61–68. NIST Special Publication 500-250, 2001.

[10] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the TREC-8 Web Track. In E. Voorhees and D. K. Harman, editors, *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 131–149. NIST Special Publication 500-246, 1999.

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

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

[13] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM SIAM Symposium on Discrete Algorithms*, San Francisco, California, USA, Jan. 1998.

[14] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi. Pooling for a large-scale test collection : An analysis of the search results from the First NTCIR Workshop. *Information Retrieval*, 5(1):41–59, feb 2002.

[15] NTCIR Web Task Organizers. NTCIR Web Task. http://research.nii.ac.jp/ntcir/web/.

[16] E. Voorhees. The TREC-8 Question Answering Track report. In E. Voorhees and D. K. Harman, editors, *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 77–82. NIST Special Publication 500-246, 1999.

[17] E. Voorhees and D. K. Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In E. Voorhees and D. K. Harman, editors, *Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pages 1–24. NIST Special Publication 500-240, 1997.

[18] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 74–82, Sep. 2001.

Table 1. Fundamental statistics of NW100G-01

# of crawled sites	97,561 (# of aliased sites: 3,285 is not included)
maximum # of pages within a site	1,300
# of crawled pages (*i.e.,* # of pages included in the document data for providing)	11,038,720 (# of aliased sites: 419,709 is not included)
# of pages for searching (of which existence are comfirmed)	15,364,404
# of links (connected from the crawled pages to the pages for searching)	64,365,554

Table 2. Statistics on links of NW100G-01

		# of links	# of pages	# of sites
1.	# of links connected from the crawled pages (only text files)	78,175,556		
2.	# of links connected to the pages for searching (of which existence are confirmed)	64,365,554		
3.	# of links connected to the pages not for searching (of which existence could not be confirmed)	13,810,002 (3./1.=0.176)		
2.	# of links that connected to the pages for searching (of which existence are comfirmed), and # of their destination pages	64,365,554	15,182,651	
2-1-1.	# of links that are closed in the crawled pages, and # of their destination pages	53,928,019	10,857,715	
2-1-2.	# of links that are not closed in the crawled pages, and # of their destination pages	10,437,535	4,324,936	
2-2-1.	# of links that connected to the pages within the same sites, and # of their destination pages	56,673,429	14,218,861	
2-2-1-1.	In 2-2-1., # of links that connected to the crawled pages, and # of their destination pages	49,960,354	10,800,231	
2-2-1-2.	In 2-2-1., # of links that connected outside the crawled pages, and # of their destination pages	6,713,075	3,418,630	
2-2-2.	# of links that connected to the pages on another crawled site, and # of their destination pages	5,563,383	729,754	
2-2-2-1.	In 2-2-2., # of links that connected to the crawled pages, and # of their destination pages	3,967,665	344,487	
2-2-2-2.	In 2-2-2., # of links that connected outside the crawled pages, and # of their destination pages	1,595,718	385,267	
2-2-3.	# of links that connected outside the crawled sites, and # of their destination pages and sites	2,128,742	600,437	237,432
3.	# of links connected to the pages not for searching (of which existence could not be confirmed), and # of their destination pages	13,810,002 (3./1.=0.176)		
3-1-1.	# of links connected to the pages within the same sites, and # of their destination pages	8,525,716	5,863,863	
3-1-2.	# of links and pages that connected to the pages on another crawled site, and # of their destination pages	1,789,643	687,553	
3-1-3.	# of links, pages and sites that connected outside the crawled sites, and # of their destination pages and sites	3,494,643	1,047,306	217,554

Table 3. Fundamental statistics of NW10G-01

# of crawled sites	97,561 (# of aliased sites: 3,285 is not included)
maximum # of pages within a site	20
# of crawled pages (*i.e.,* # of pages included in the document data for providing)	1,445,466 (# of aliased sites: 141,574 is not included)
# of pages for searching (of which existence are comfirmed)	4,849,714
# of links (connected from the crawled pages to the pages for searching)	9,885,538

Table 4. Statistics on links of NW10G-01

		# of links	# of pages	# of sites
1.	# of links connected from the crawled pages (only text files)	11,642,167		
2.	# of links connected to the pages for searching (of which existence are confirmed)	9,885,538		
3.	# of links connected to the pages not for searching (of which existence could not be confirmed)	1,756,629 (3./1.=0.150)		
2.	# of links that connected to the pages for searching (of which existence are comfirmed), and # of their destination pages	9,885,538	4,810,115	
2-1-1.	# of links that are closed in the crawled pages, and # of their destination pages	4,978,298	1,405,928	
2-1-2.	# of links that are not closed in the crawled pages, and # of their destination pages	4,907,240	3,404,187	
2-2-1.	# of links that connected to the pages within the same sites, and # of their destination pages	8,427,690	4,461,635	
2-2-1-1.	In 2-2-1., # of links that connected to the crawled pages, and # of their destination pages	4,303,577	1,349,118	
2-2-1-2.	In 2-2-1., # of links that connected outside the crawled pages, and # of their destination pages	4,124,113	3,112,517	
2-2-2.	# of links that connected to the pages on another crawled site, and # of their destination pages	1,084,263	193,122	
2-2-2-1.	In 2-2-2., # of links that connected to the crawled pages, and # of their destination pages	674,721	122,070	
2-2-2-2.	In 2-2-2., # of links that connected outside the crawled pages, and # of their destination pages	409,542	71,052	
2-2-3.	# of links that connected outside the crawled sites, and # of their destination pages and sites	373,585	155,358	73,916
3.	# of links connected to the pages not for searching (of which existence could not be confirmed), and # of their destination pages	1,756,629 (3./1.=0.150)		
3-1-1.	# of links connected to the pages within the same sites, and # of their destination pages	822,318	597,855	
3-1-2.	# of links and pages that connected to the pages on another crawled site, and # of their destination pages	934,311	442,940	
3-1-3.	# of links, pages and sites that connected outside the crawled sites, and # of their destination pages and sites	620,663	262,735	80,169

Table 5. Summary of run result submission

Task	RunID	QMethod	TopicPart	LinkInfo	Task	RunID	QMethod	TopicPart	LinkInfo
I-A1	GRACE-LA1-1	automatic	T	cont	II-A1	GRACE-SA1-1	automatic	T	cont
I-A1	GRACE-LA1-2	automatic	T	link&cont	II-A1	GRACE-SA1-2	automatic	T	cont
I-A1	GRACE-LA1-3	automatic	D	cont	II-A1	GRACE-SA1-3	automatic	D	cont
I-A1	GRACE-LA1-4	automatic	D	link&cont	II-A1	GRACE-SA1-4	automatic	D	cont
I-A1	K3100-05	automatic	T	link&cont	II-A1	K3100-01	automatic	T	link&cont
I-A1	K3100-06	automatic	T	link&cont	II-A1	K3100-02	automatic	T	link&cont
I-A1	K3100-07	automatic	D	link&cont	II-A1	K3100-03	automatic	D	link&cont
I-A1	K3100-08	automatic	D	link&cont	II-A1	K3100-04	automatic	D	link&cont
I-A1	NAICR-I-A1-1	automatic	D	cont	II-A1	NAICR-II-A1-1	automatic	D	cont
I-A1	NAICR-I-A1-2	automatic	D	cont	II-A1	NAICR-II-A1-2	automatic	D	cont
I-A1	NAICR-I-A1-3	automatic	D	cont	II-A1	NAICR-II-A1-3	automatic	D	cont
I-A1	NAICR-I-A1-4	automatic	T	cont	II-A1	NAICR-II-A1-4	automatic	T	cont
I-A1	OKSAT-WEB-F-02	interactive	TD	cont	II-A1	OASIS11	automatic	D	cont
I-A1	OKSAT-WEB-F-04	automatic	T	cont	II-A1	OASIS12	automatic	D	cont
I-A1	OKSAT-WEB-F-06	automatic	D	cont	II-A1	OKSAT-WEB-F-01	interactive	TD	cont
I-A1	ORGREF-LA1-1	automatic	T	cont	II-A1	OKSAT-WEB-F-03	automatic	T	cont
I-A1	ORGREF-LA1-2	automatic	T	cont	II-A1	OKSAT-WEB-F-05	automatic	D	cont
I-A1	ORGREF-LA1-3	automatic	T	cont	II-A1	ORGREF-SA1-1	automatic	T	cont
I-A1	ORGREF-LA1-4	automatic	T	cont	II-A1	ORGREF-SA1-2	automatic	T	cont
I-A1	ORGREF-LA1-5	automatic	T	cont	II-A1	ORGREF-SA1-3	automatic	T	cont
I-A1	ORGREF-LA1-6	automatic	T	cont	II-A1	ORGREF-SA1-4	automatic	T	cont
I-A1	UAIFI1	automatic	D	cont	II-A1	ORGREF-SA1-5	automatic	T	cont
I-A1	UAIFI2	automatic	D	cont	II-A1	ORGREF-SA1-6	automatic	T	cont
I-A1	UAIFI3	automatic	T	cont	II-A1	UAIFI10	automatic	D	cont
I-A1	UAIFI4	automatic	T	cont	II-A1	UAIFI11	automatic	T	cont
					II-A1	UAIFI12	automatic	T	cont
					II-A1	UAIFI9	automatic	D	cont
I-A2	GRACE-LA2-1	automatic	TR[1]	cont	II-A2	GRACE-SA2-1	automatic	TR[1]	cont
I-A2	GRACE-LA2-2	automatic	TR[1]	cont	II-A2	GRACE-SA2-2	automatic	TR[1]	cont
I-A2	GRACE-LA2-3	automatic	TR[1]	cont	II-A2	GRACE-SA2-3	automatic	TR[1]	cont
I-A2	GRACE-LA2-4	automatic	TR[1]	cont	II-A2	GRACE-SA2-4	automatic	TR[1]	cont
I-A2	NAICR-I-A2-1	automatic	D	cont	II-A2	NAICR-II-A2-1	automatic	TR[1]	cont
I-A2	NAICR-I-A2-2	automatic	T	cont	II-A2	NAICR-II-A2-2	automatic	TR[1]	cont
I-A2	NAICR-I-A2-3	automatic	TR[1]	cont	II-A2	NAICR-II-A2-3	automatic	T	cont
I-A2	NAICR-I-A2-4	automatic	TR[1]	cont	II-A2	NAICR-II-A2-4	automatic	D	cont
I-B	GRACE-LB-1	automatic	T	cont	II-B	GRACE-SB-1	automatic	T	cont
I-B	GRACE-LB-2	automatic	T	link&cont	II-B	GRACE-SB-2	automatic	T	cont
I-B	GRACE-LB-3	automatic	D	cont	II-B	GRACE-SB-3	automatic	D	cont
I-B	GRACE-LB-4	automatic	D	link&cont	II-B	GRACE-SB-4	automatic	D	cont
I-B	K3100-13	automatic	T	link&cont	II-B	K3100-09	automatic	T	link&cont
I-B	K3100-14	automatic	T	link&cont	II-B	K3100-10	automatic	T	link&cont
I-B	K3100-15	automatic	D	link&cont	II-B	K3100-11	automatic	D	link&cont
I-B	K3100-16	automatic	D	link&cont	II-B	K3100-12	automatic	D	link&cont
I-B	NAICR-I-B-1	automatic	T	cont	II-B	NAICR-II-B-1	automatic	D	cont
I-B	NAICR-I-B-2	automatic	D	cont	II-B	NAICR-II-B-2	automatic	D	cont
I-B	NAICR-I-B-3	automatic	D	cont	II-B	NAICR-II-B-3	automatic	D	cont
I-B	NAICR-I-B-4	automatic	D	cont	II-B	NAICR-II-B-4	automatic	T	cont
I-B	ORGREF-LB-1	automatic	T	cont	II-B	ORGREF-SB-1	automatic	T	cont
I-B	ORGREF-LB-2	automatic	T	cont	II-B	ORGREF-SB-2	automatic	T	cont
I-B	ORGREF-LB-3	automatic	T	cont	II-B	ORGREF-SB-3	automatic	T	cont
I-B	ORGREF-LB-4	automatic	T	cont	II-B	ORGREF-SB-4	automatic	T	cont
I-B	ORGREF-LB-5	automatic	T	cont	II-B	ORGREF-SB-5	automatic	T	cont
I-B	ORGREF-LB-6	automatic	T	cont	II-B	ORGREF-SB-6	automatic	T	cont
I-B	UAIFI5	automatic	D	cont	II-B	UAIFI13	automatic	D	cont
I-B	UAIFI6	automatic	D	cont	II-B	UAIFI14	automatic	D	cont
I-B	UAIFI7	automatic	T	cont	II-B	UAIFI15	automatic	T	cont
I-B	UAIFI8	automatic	T	cont	II-B	UAIFI16	automatic	T	cont

Task: Indicates the types of the tasks. 'I' indicates a task using the 100-gigabyte data set and 'II' one using the 10-gigabyte data set. The detailed task descriptions are explained in Section **2**.
RunID: Indicates the identification codes of the system run results. Each one starts with the group ID.
QMethod: 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. The characters 'T', 'D' and 'R[n]' respectively indicate TITLE, DESC and the nth document specified in RDOC.
LinkInfo: Indicates whether or not the system used link information in Web documents. The notation 'link&cont' indicates that the links and contents were used; 'cont' indicates that only contents were used.

Table 6. Selected evaluation results of the Survey Retrieval Tasks

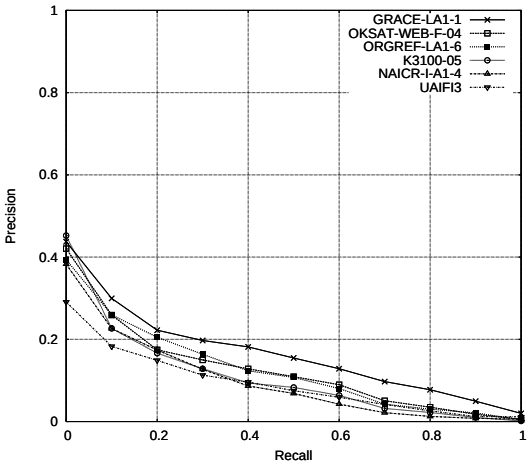
Task	QMethod	Topic Part	RunID	LinkInfo	$DM_2\&RL_1$				$DM_2\&RL_2$				$DM_1\&RL_1$			
					aprec	rprec	dcg(100)	dcg(1K)	aprec	rprec	dcg(100)	dcg(1K)	aprec	rprec	dcg(100)	dcg(1K)
I-A1	automatic	T	GRACE-LA1-1	cont	0.1506	0.1707	7.3478	13.0492	0.2109	0.2345	9.2695	16.4611	0.1551	0.1986	9.3798	17.3195
	automatic	T	GRACE-LA1-2	link&cont	0.1489	0.1739	7.2873	12.8568	0.2061	0.2361	9.1492	16.1874	0.1548	0.2019	9.3503	17.1396
	automatic	T	OKSAT-WEB-F-04	cont	0.1151	0.1469	6.3920	11.0055	0.1520	0.1972	7.8993	13.7063	0.1104	0.1536	7.3605	13.7549
	automatic	T	ORGREF-LA1-6	cont	0.1111	0.1571	6.0655	10.6693	0.1447	0.1939	7.5225	13.4621	0.1111	0.1603	7.0813	13.5672
	automatic	T	K3100-05	link&cont	0.0951	0.1449	5.8189	10.3197	0.1300	0.1821	7.2762	13.0700	0.1013	0.1560	7.3532	13.8044
	automatic	T	K3100-06	link&cont	0.0937	0.1369	5.7608	10.2257	0.1281	0.1747	7.1691	12.9111	0.1002	0.1486	7.2143	13.5953
	automatic	T	NAICR-I-A1-4	cont	0.0875	0.1298	5.3408	10.0245	0.1135	0.1751	6.8675	12.5391	0.0904	0.1443	6.7687	13.3463
	automatic	T	ORGREF-LA1-5	cont	0.0833	0.1176	4.7709	9.5219	0.1135	0.1623	6.2605	12.0449	0.0868	0.1280	5.7715	12.3219
	automatic	T	UAIFI3	cont	0.0815	0.1204	5.1242	10.1186	0.1346	0.1913	6.9208	13.1294	0.0981	0.1477	7.0491	14.0841
	automatic	T	UAIFI4	cont	0.0808	0.1128	4.9741	10.2023	0.1280	0.1837	6.6714	13.0869	0.0938	0.1432	6.7929	13.9697
	automatic	T	ORGREF-LA1-3	cont	0.0654	0.1039	4.5514	8.0706	0.1009	0.1560	6.0680	10.4609	0.0685	0.1164	5.8615	10.6385
	automatic	T	ORGREF-LA1-1	cont	0.0596	0.0998	4.2086	7.7335	0.0879	0.1339	5.2586	9.7006	0.0588	0.1030	4.7826	9.8060
	automatic	T	ORGREF-LA1-4	cont	0.0526	0.0846	4.1272	7.2806	0.0827	0.1350	5.5365	9.4269	0.0554	0.0983	5.2200	9.5099
	automatic	T	ORGREF-LA1-2	cont	0.0449	0.0728	3.4951	7.2076	0.0653	0.1096	4.3731	8.9193	0.0455	0.0792	4.0205	9.0462
	automatic	D	GRACE-LA1-4	link&cont	0.1555	0.1856	8.0395	13.7795	0.2088	0.2310	9.7968	16.7512	0.1488	0.1918	9.5314	16.9439
	automatic	D	GRACE-LA1-3	cont	0.1548	0.1799	8.0356	13.7948	0.2082	0.2302	9.7899	16.7633	0.1479	0.1906	9.5193	16.9506
	automatic	D	OKSAT-WEB-F-06	cont	0.1236	0.1473	6.4498	11.4801	0.1559	0.1902	7.7792	14.2475	0.1206	0.1607	7.4291	14.3643
	automatic	D	NAICR-I-A1-3	cont	0.0907	0.1363	5.7470	10.7955	0.1148	0.1798	7.1050	13.1694	0.0935	0.1529	7.0400	13.9398
	automatic	D	K3100-07	link&cont	0.0897	0.1164	5.2714	8.7714	0.1162	0.1524	6.4217	11.0198	0.0926	0.1325	6.6369	11.4676
	automatic	D	UAIFI1	cont	0.0855	0.1247	5.6399	10.5625	0.1257	0.1809	7.1328	13.1817	0.0982	0.1547	7.5055	14.1661
	automatic	D	UAIFI2	cont	0.0843	0.1256	5.5013	10.5842	0.1184	0.1772	6.9297	13.1267	0.0928	0.1461	7.2712	14.1109
	automatic	D	NAICR-I-A1-2	cont	0.0756	0.0990	3.9179	7.6978	0.0945	0.1257	4.7994	9.4511	0.0845	0.1102	5.2435	10.2021
	automatic	D	K3100-08	link&cont	0.0737	0.0991	4.5122	8.9811	0.0980	0.1368	5.5802	11.2223	0.0756	0.1150	5.6447	11.3948
	automatic	D	NAICR-I-A1-1	cont	0.0736	0.0929	3.4572	6.9776	0.0882	0.1222	4.2856	8.5957	0.0841	0.1083	4.9067	9.3417
	interactive	TD	OKSAT-WEB-F-02	cont	0.1238	0.1535	6.9065	12.6003	0.1674	0.2032	8.3054	15.3811	0.1191	0.1682	8.0375	15.5999
I-A2	automatic	TR[1]	GRACE-LA2-3	cont	0.1977	0.2229	10.0645	15.7602	0.2546	0.2770	11.9122	19.1600	0.1998	0.2464	12.3606	20.3841
	automatic	TR[1]	GRACE-LA2-2	cont	0.1966	0.2207	9.9764	15.4430	0.2553	0.2808	11.9402	18.8438	0.1990	0.2291	12.2912	20.1235
	automatic	TR[1]	GRACE-LA2-1	cont	0.1913	0.2092	9.7183	14.9258	0.2430	0.2762	11.5838	18.1920	0.1959	0.2317	11.9774	19.7194
	automatic	TR[1]	GRACE-LA2-4	cont	0.1769	0.2077	9.0702	15.0199	0.2357	0.2570	10.9365	18.4044	0.1811	0.2275	11.2150	19.3453
	automatic	TR[1]	NAICR-I-A2-3	cont	0.0873	0.1073	5.5473	9.6440	0.1000	0.1415	6.3849	11.3242	0.0886	0.1281	7.2916	12.5037
	automatic	TR[1]	NAICR-I-A2-4	cont	0.0860	0.1183	4.9069	8.3797	0.1059	0.1352	5.7773	10.0328	0.0943	0.1344	7.0367	11.4648
	automatic	T	NAICR-I-A2-2	cont	0.0823	0.1214	5.3067	9.9524	0.1102	0.1718	6.7786	12.4229	0.0898	0.1445	6.7268	13.3339
	automatic	D	NAICR-I-A2-1	cont	0.0908	0.1356	5.7532	10.7333	0.1168	0.1788	7.1618	13.1620	0.0935	0.1537	7.0990	13.8678
II-A1	automatic	T	GRACE-SA1-2	cont	0.2164	0.1966	4.1269	5.0773	0.2433	0.2429	5.1583	6.3925	0.1775	0.1935	5.8389	7.7713
	automatic	T	GRACE-SA1-1	cont	0.2056	0.1978	3.7720	4.8112	0.2266	0.2372	4.7265	6.0589	0.1768	0.1912	5.6464	7.6963
	automatic	T	OKSAT-WEB-F-03	cont	0.1749	0.1702	3.0193	3.8194	0.1936	0.2069	3.8146	4.9079	0.1271	0.1496	4.2058	5.8870
	automatic	T	ORGREF-SA1-6	cont	0.1747	0.1769	3.0717	3.8409	0.1873	0.1918	3.8668	4.8973	0.1478	0.1682	4.6581	6.1227
	automatic	T	UAIFI11	cont	0.1430	0.1396	2.5043	3.4155	0.1491	0.1663	3.2681	4.4459	0.1201	0.1468	4.2076	5.9526
	automatic	T	UAIFI12	cont	0.1342	0.1281	2.4010	3.3096	0.1407	0.1566	3.1988	4.3501	0.1125	0.1302	4.0756	5.8087
	automatic	T	K3100-01	link&cont	0.1333	0.1341	2.7853	3.5351	0.1641	0.1709	3.6344	4.6416	0.1145	0.1364	4.3749	5.7201
	automatic	T	K3100-02	link&cont	0.1297	0.1335	2.7752	3.5293	0.1613	0.1702	3.6193	4.6315	0.1137	0.1359	4.3485	5.7052
	automatic	T	ORGREF-SA1-1	cont	0.1243	0.1465	2.3034	3.1777	0.1298	0.1553	2.8820	4.0060	0.0951	0.1272	3.2905	4.8968
	automatic	T	ORGREF-SA1-5	cont	0.1214	0.1158	2.6300	3.2386	0.1304	0.1358	3.2642	4.1666	0.1217	0.1510	4.2670	5.5502
	automatic	T	ORGREF-SA1-3	cont	0.1101	0.1109	1.8230	2.4912	0.1041	0.1310	2.3043	3.2148	0.0885	0.1131	2.9778	4.3754
	automatic	T	ORGREF-SA1-2	cont	0.1101	0.1360	2.1614	2.9670	0.1102	0.1408	2.6588	3.7180	0.0863	0.1203	3.0921	4.6049
	automatic	T	NAICR-II-A1-4	cont	0.1028	0.0971	2.5783	3.3424	0.1268	0.1577	3.2079	4.1906	0.0965	0.1331	3.8590	5.2079
	automatic	T	ORGREF-SA1-4	cont	0.0803	0.0892	1.5944	2.2843	0.0882	0.1204	2.0438	2.9751	0.0731	0.0897	2.6996	4.0951
	automatic	D	GRACE-SA1-4	cont	0.2264	0.2368	3.8688	4.7892	0.2458	0.2535	4.7648	5.9638	0.1694	0.2031	5.3314	7.2810
	automatic	D	GRACE-SA1-3	cont	0.1970	0.2168	3.7560	4.7709	0.2241	0.2435	4.6328	5.9133	0.1604	0.1915	5.3131	7.4203
	automatic	D	OKSAT-WEB-F-05	cont	0.1859	0.1935	3.2372	4.0581	0.2228	0.2180	4.1085	5.2259	0.1397	0.1591	4.4914	6.2484
	automatic	D	UAIFI9	cont	0.1408	0.1368	2.7048	3.7329	0.1577	0.1605	3.4496	4.7394	0.1259	0.1503	4.4458	6.3179
	automatic	D	UAIFI10	cont	0.1330	0.1247	2.6003	3.6667	0.1529	0.1574	3.3869	4.6950	0.1161	0.1426	4.2757	6.1605
	automatic	D	NAICR-II-A1-2	cont	0.1286	0.1084	2.0882	3.2505	0.1460	0.1597	2.5680	4.0171	0.0971	0.1287	2.9105	4.8595
	automatic	D	K3100-03	link&cont	0.1244	0.1070	2.3855	2.8715	0.1588	0.1582	3.1519	3.7956	0.1120	0.1357	3.6493	4.4917
	automatic	D	NAICR-II-A1-1	cont	0.1234	0.1016	2.1688	3.3191	0.1404	0.1514	2.6938	4.1096	0.0871	0.1148	2.9350	4.9196
	automatic	D	K3100-04	link&cont	0.1213	0.1152	2.6217	3.2188	0.1530	0.1615	3.3607	4.2460	0.0993	0.1216	3.7586	4.9564
	automatic	D	NAICR-II-A1-3	cont	0.1183	0.1170	2.8585	4.0558	0.1356	0.1629	3.3936	4.9340	0.1044	0.1310	4.3235	6.3616
	automatic	D	OASIS12	cont	0.0406	0.0366	1.2292	2.0637	0.0481	0.0479	1.4979	2.5557	0.0296	0.0389	1.6223	3.0087
	automatic	D	OASIS11	cont	0.0385	0.0366	1.1471	1.9719	0.0453	0.0435	1.3921	2.4401	0.0275	0.0378	1.4870	2.8542
	interactive	TD	OKSAT-WEB-F-01	cont	0.2236	0.2305	3.9013	4.7711	0.2436	0.2513	4.8161	5.9357	0.1635	0.1896	5.2637	7.0661
II-A2	automatic	TR[1]	GRACE-SA2-2	cont	0.2810	0.2676	4.8953	5.8038	0.3310	0.3173	5.8673	7.0763	0.2264	0.2396	6.8790	8.8193
	automatic	TR[1]	GRACE-SA2-3	cont	0.2781	0.2497	5.0227	5.8906	0.3410	0.3201	6.0741	7.2131	0.2284	0.2502	6.9720	8.9581
	automatic	TR[1]	GRACE-SA2-1	cont	0.2637	0.2453	4.5736	5.5302	0.3153	0.3092	5.5584	6.7654	0.2133	0.2441	6.6967	8.5307
	automatic	TR[1]	GRACE-SA2-4	cont	0.2361	0.2395	4.4749	5.2703	0.2696	0.2825	5.4708	6.5560	0.2017	0.2301	6.2571	8.1387
	automatic	TR[1]	NAICR-II-A2-2	cont	0.1899	0.1862	3.8552	4.6794	0.2319	0.2458	4.5697	5.5567	0.1438	0.1747	5.0891	6.7373
	automatic	TR[1]	NAICR-II-A2-1	cont	0.1716	0.1667	3.2509	4.2335	0.2033	0.2275	3.6848	4.9340	0.1519	0.1785	4.9626	6.5970
	automatic	T	NAICR-II-A2-3	cont	0.1187	0.1204	2.8332	3.8501	0.1558	0.1649	3.5881	4.8696	0.1026	0.1290	4.1889	6.0097
	automatic	D	NAICR-II-A2-4	cont	0.1176	0.1247	2.8257	3.9193	0.1378	0.1643	3.3619	4.8037	0.1059	0.1364	4.3418	6.1920

RunID indicates the identification codes of the system run results, as shown in **Table 5**.
aprec indicates the average precision (non-interpolated).
rprec 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.

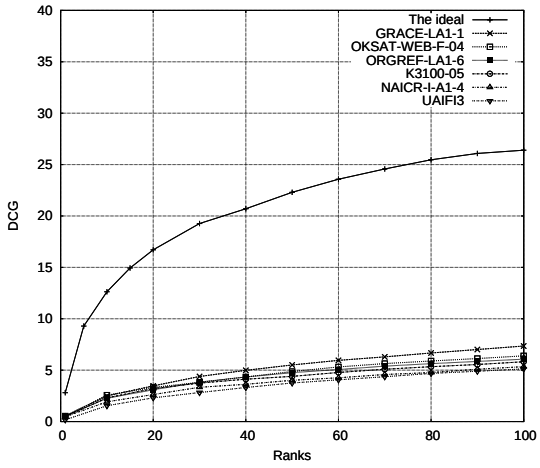
Table 7. Selected evaluation results of the Target Retrieval Task

Task	QMethod	Topic Part	RunID	LinkInfo	DM_2&RL_1				DM_2&RL_2				DM_1&RL_1			
					prec(10)	dcg(10)	wrr(10)	%nf(10)	prec(10)	dcg(10)	wrr(10)	%nf(10)	prec(10)	dcg(10)	wrr(10)	%nf(10)
I-B	automatic	T	GRACE-LB-1	cont	0.2213	2.4940	0.3618	0.2979	0.3511	3.2212	0.4767	0.2979	0.2532	2.9598	0.4060	0.2553
	automatic	T	GRACE-LB-2	link&cont	0.2106	2.3901	0.3581	0.2766	0.3404	3.0686	0.4627	0.2553	0.2447	2.8802	0.4030	0.2553
	automatic	T	K3100-14	link&cont	0.2106	2.3196	0.3544	0.2553	0.2830	2.7305	0.4553	0.1915	0.2447	2.7441	0.3691	0.2553
	automatic	T	K3100-13	link&cont	0.2085	2.3072	0.3796	0.1915	0.2936	2.7736	0.4822	0.1277	0.2426	2.7400	0.3983	0.1915
	automatic	T	ORGREF-LB-6	cont	0.1915	2.2709	0.3346	0.3191	0.2745	2.7288	0.4301	0.2766	0.2170	2.4918	0.3422	0.3191
	automatic	T	UAIFI8	cont	0.1468	1.6883	0.2751	0.3404	0.2702	2.3850	0.4141	0.3191	0.1979	2.3787	0.3766	0.2553
	automatic	T	UAIFI7	cont	0.1426	1.5409	0.2338	0.3404	0.2830	2.3470	0.4058	0.3191	0.1872	2.1654	0.3506	0.2766
	automatic	T	NAICR-I-B-1	cont	0.1383	1.7538	0.2992	0.3830	0.2085	2.0991	0.3370	0.2979	0.1787	2.2595	0.3446	0.3404
	automatic	T	ORGREF-LB-5	cont	0.1340	1.6262	0.2943	0.4255	0.2362	2.1592	0.4036	0.3191	0.1681	1.9934	0.3319	0.3830
	automatic	T	ORGREF-LB-3	cont	0.1213	1.4486	0.3016	0.4043	0.2319	2.1360	0.4534	0.3191	0.1638	1.9411	0.3563	0.3617
	automatic	T	ORGREF-LB-4	cont	0.1106	1.2752	0.2646	0.4681	0.2021	1.8616	0.4053	0.3617	0.1447	1.6978	0.3368	0.3830
	automatic	T	ORGREF-LB-1	cont	0.1064	1.2370	0.1986	0.6383	0.1596	1.4746	0.2311	0.5319	0.1128	1.2778	0.2008	0.6170
	automatic	T	ORGREF-LB-2	cont	0.0809	0.9163	0.1148	0.7021	0.1106	1.0343	0.1369	0.5957	0.0851	0.9510	0.1169	0.6809
	automatic	D	GRACE-LB-4	link&cont	0.2340	2.9319	0.4214	0.3191	0.3340	3.5002	0.5020	0.2979	0.2681	3.3213	0.4634	0.2553
	automatic	D	GRACE-LB-3	cont	0.2340	2.9707	0.4204	0.3617	0.3298	3.5222	0.5023	0.3191	0.2681	3.3602	0.4589	0.2979
	automatic	D	UAIFI5	cont	0.1851	1.9955	0.2543	0.3617	0.2894	2.5765	0.3832	0.2553	0.2319	2.6208	0.3446	0.2553
	automatic	D	UAIFI6	cont	0.1830	2.0094	0.3008	0.3830	0.2745	2.5031	0.3910	0.2979	0.2319	2.6879	0.3808	0.2340
	automatic	D	NAICR-I-B-2	cont	0.1596	1.8877	0.3538	0.3191	0.2489	2.3694	0.4682	0.2340	0.1872	2.2507	0.4134	0.2766
	automatic	D	K3100-15	link&cont	0.1574	1.9228	0.3542	0.3830	0.2298	2.3331	0.4892	0.2128	0.1787	2.2030	0.3742	0.3191
	automatic	D	NAICR-I-B-3	cont	0.1319	1.5588	0.2267	0.5745	0.1787	1.8170	0.2776	0.5532	0.1723	2.0386	0.2574	0.5532
	automatic	D	NAICR-I-B-4	cont	0.1319	1.5209	0.2213	0.5745	0.1809	1.7859	0.2848	0.5106	0.1681	2.0232	0.2678	0.5319
	automatic	D	K3100-16	link&cont	0.1191	1.4750	0.2881	0.4468	0.1809	1.8565	0.4211	0.3191	0.1426	1.7639	0.3086	0.3830
II-B	automatic	T	GRACE-SB-2	cont	0.1745	2.3978	0.4222	0.3617	0.2638	2.9225	0.5197	0.2553	0.2255	2.9720	0.4517	0.3191
	automatic	T	GRACE-SB-1	cont	0.1681	2.1770	0.4297	0.2553	0.2511	2.6727	0.5454	0.1702	0.2191	2.8935	0.5146	0.2128
	automatic	T	ORGREF-SB-6	cont	0.1277	1.8553	0.4331	0.2979	0.1936	2.2583	0.5110	0.2553	0.1617	2.3035	0.4998	0.2766
	automatic	T	K3100-09	link&cont	0.1170	1.6852	0.3987	0.2766	0.1809	2.0590	0.4764	0.1915	0.1681	2.2144	0.4223	0.2553
	automatic	T	K3100-10	link&cont	0.1170	1.6844	0.3880	0.2766	0.1787	2.0512	0.4658	0.1915	0.1681	2.2087	0.4117	0.2553
	automatic	T	ORGREF-SB-5	cont	0.1149	1.3497	0.2431	0.4043	0.1638	1.6300	0.3142	0.3191	0.1681	2.0671	0.3433	0.3191
	automatic	T	NAICR-II-B-4	cont	0.1021	1.4110	0.3314	0.3404	0.1596	1.7312	0.4254	0.2766	0.1383	1.7645	0.3463	0.3404
	automatic	T	ORGREF-SB-1	cont	0.0979	1.4558	0.3333	0.3617	0.1489	1.7820	0.4081	0.2766	0.1234	1.7580	0.3821	0.2979
	automatic	T	UAIFI15	cont	0.0915	1.1879	0.2889	0.4468	0.1532	1.5306	0.3538	0.3830	0.1319	1.7779	0.3931	0.3830
	automatic	T	UAIFI16	cont	0.0851	1.0937	0.2587	0.4681	0.1532	1.4818	0.3445	0.3617	0.1298	1.7609	0.3805	0.3830
	automatic	T	ORGREF-SB-2	cont	0.0851	1.3603	0.3063	0.4255	0.1319	1.6324	0.3855	0.3191	0.1128	1.6736	0.3606	0.3191
	automatic	T	ORGREF-SB-3	cont	0.0745	0.9305	0.2474	0.5106	0.1191	1.1893	0.3033	0.3830	0.1191	1.5963	0.3751	0.4468
	automatic	T	ORGREF-SB-4	cont	0.0617	0.8083	0.1975	0.5745	0.1106	1.0859	0.2770	0.4043	0.1106	1.4979	0.3206	0.4681
	automatic	D	GRACE-SB-4	cont	0.1702	2.1963	0.3970	0.2979	0.2596	2.7014	0.4878	0.2340	0.2213	2.7438	0.4339	0.2553
	automatic	D	GRACE-SB-3	cont	0.1574	2.1215	0.3714	0.2766	0.2340	2.5717	0.4553	0.2128	0.1979	2.6903	0.4606	0.2340
	automatic	D	K3100-11	link&cont	0.1191	1.6266	0.3197	0.3830	0.1809	2.0473	0.4690	0.2553	0.1766	2.2604	0.3662	0.3191
	automatic	D	K3100-12	link&cont	0.1170	1.5794	0.3327	0.4043	0.1702	1.9287	0.4495	0.3191	0.1617	2.1115	0.3757	0.3617
	automatic	D	UAIFI13	cont	0.1064	1.3941	0.3003	0.4255	0.1681	1.7433	0.3480	0.3617	0.1617	2.1268	0.4221	0.3617
	automatic	D	NAICR-II-B-1	cont	0.0957	1.2957	0.2628	0.5319	0.1447	1.5883	0.3227	0.4894	0.1213	1.5585	0.2828	0.5106
	automatic	D	UAIFI14	cont	0.0915	1.2614	0.2719	0.4894	0.1511	1.6203	0.3435	0.4043	0.1426	1.9744	0.4166	0.3830
	automatic	D	NAICR-II-B-3	cont	0.0787	1.0807	0.2045	0.5957	0.1340	1.3784	0.2604	0.5319	0.1128	1.4262	0.2543	0.5319
	automatic	D	NAICR-II-B-2	cont	0.0745	1.0700	0.2172	0.5532	0.1234	1.3504	0.2810	0.5106	0.1043	1.3520	0.2446	0.5319

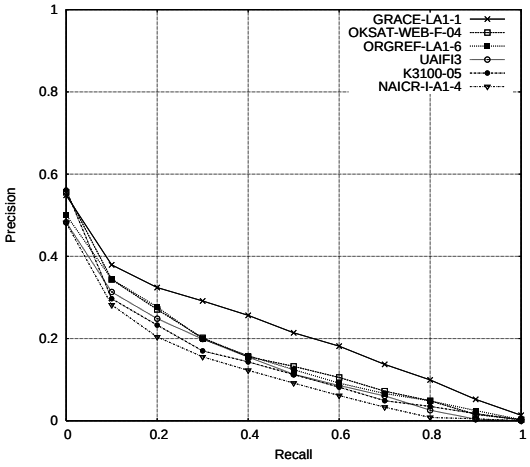
RunID indicates the identification codes of the system run results, as shown in **Table 5**.
prec(10) indicates the precision at the 10-document level.
dcg(10) indicates the DCG value at the 10-document level.
wrr(10) indicates the WRR value at the 10-document level.
%nf(10) indicates the percentage of topics for which no relevant documents were retrieved at the 10-document level.



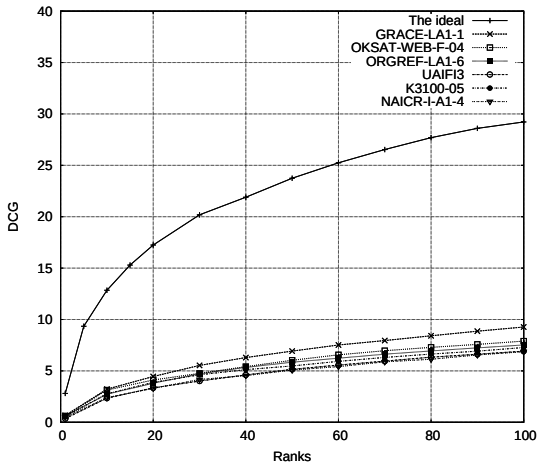
Recall-precision curves for the I-A1 'automatic' and 'TITLE-only' runs without considering links (rigid relevance level)



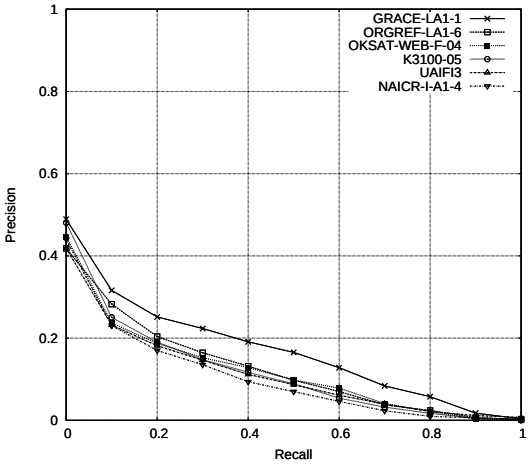
DCG curves for the I-A1 'automatic' and 'TITLE-only' runs without considering links (rigid relevance level)



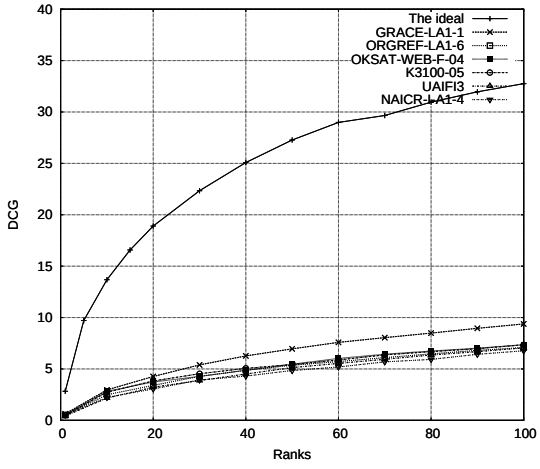
Recall-precision curves for the I-A1 'automatic' and 'TITLE-only' runs without considering links (relaxed relevance level)



DCG curves for the I-A1 'automatic' and 'TITLE-only' runs without considering links (relaxed relevance level)

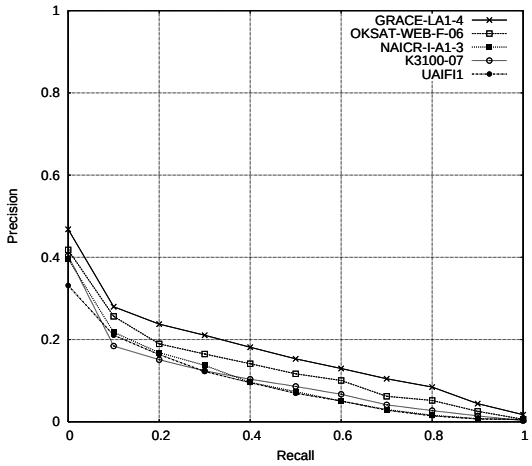


Recall-precision curves for the I-A1 'automatic' and 'TITLE-only' runs with considering links (rigid relevance level)

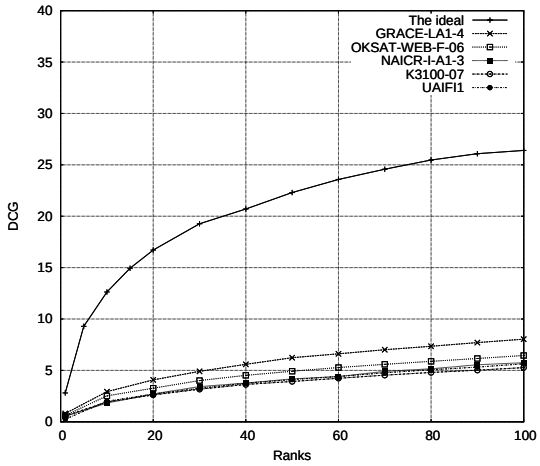


DCG curves for the I-A1 'automatic' and 'TITLE-only' runs with considering links (rigid relevance level)

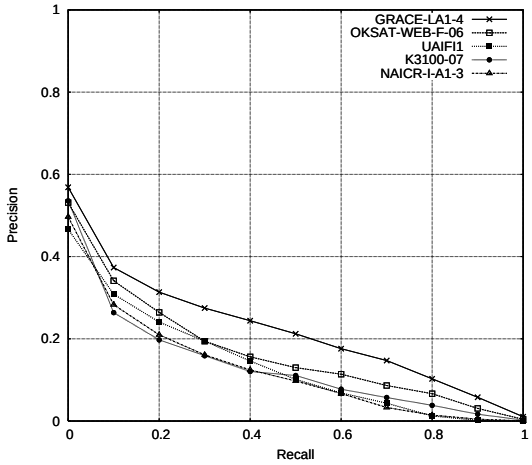
Figure 4. Recall-precision and DCG curves for the I-A1 'automatic' and 'TITLE-only' runs



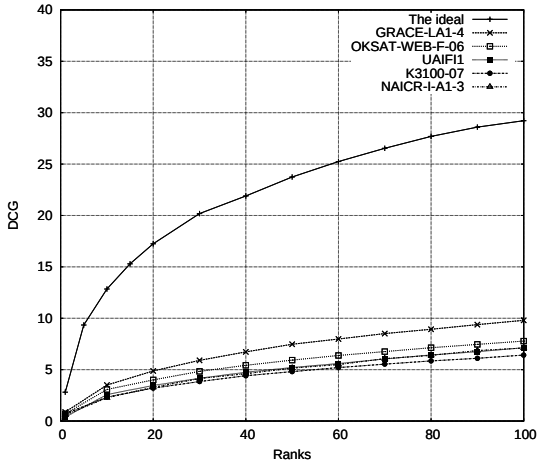
Recall-precision curves for the I-A1 'automatic' and 'DESC-only' runs without considering links (rigid relevance level)



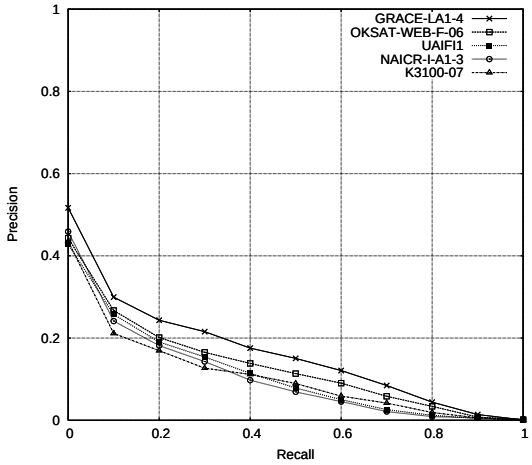
DCG curves for the I-A1 'automatic' and 'DESC-only' runs without considering links (rigid relevance level)



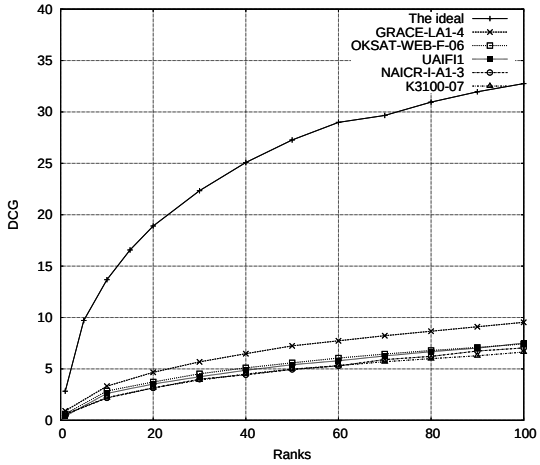
Recall-precision curves for the I-A1 'automatic' and 'DESC-only' runs without considering links (relaxed relevance level)



DCG curves for the I-A1 'automatic' and 'DESC-only' runs without considering links (relaxed relevance level)

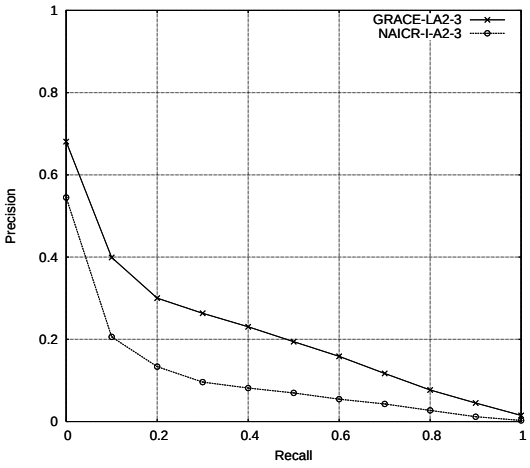


Recall-precision curves for the I-A1 'automatic' and 'DESC-only' runs with considering links (rigid relevance level)

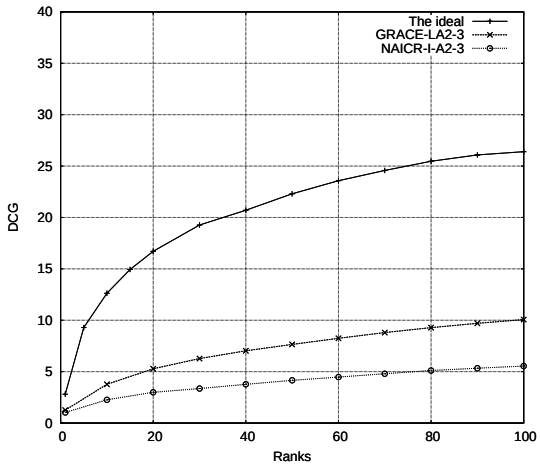


DCG curves for the I-A1 'automatic' and 'DESC-only' runs with considering links (rigid relevance level)

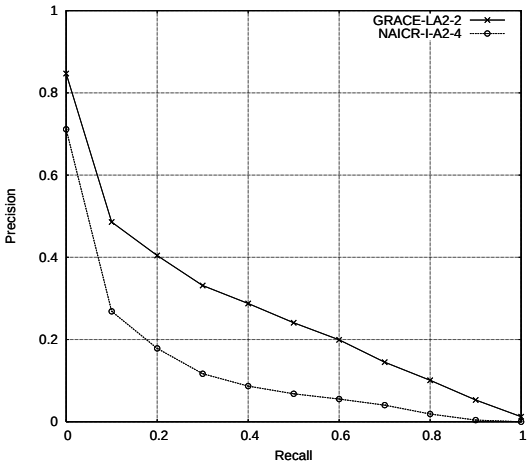
Figure 5. Recall-precision and DCG curves for the I-A1 'automatic' and 'DESC-only' runs



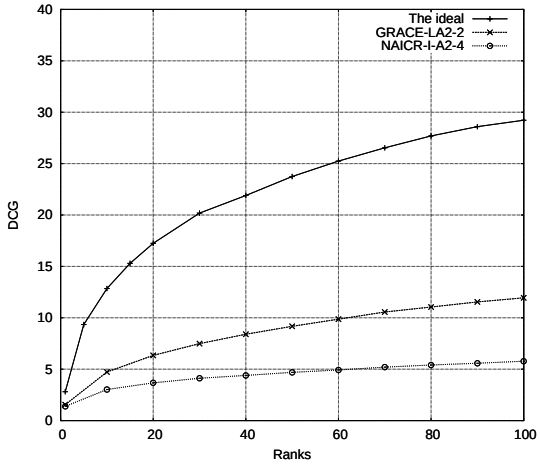
Recall-precision curves for the I-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (rigid relevance level)



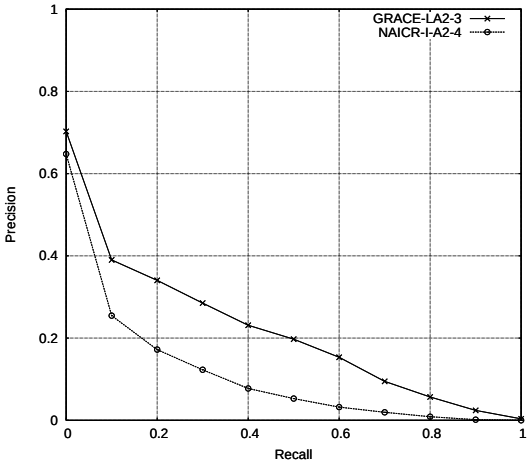
DCG curves for the I-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (rigid relevance level)



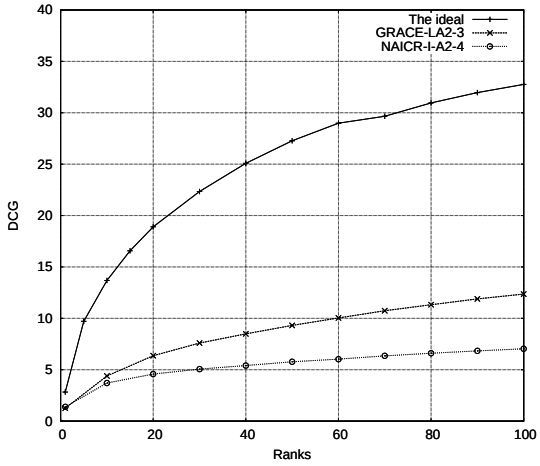
Recall-precision curves for the I-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (relaxed relevance level)



DCG curves for the I-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (relaxed relevance level)

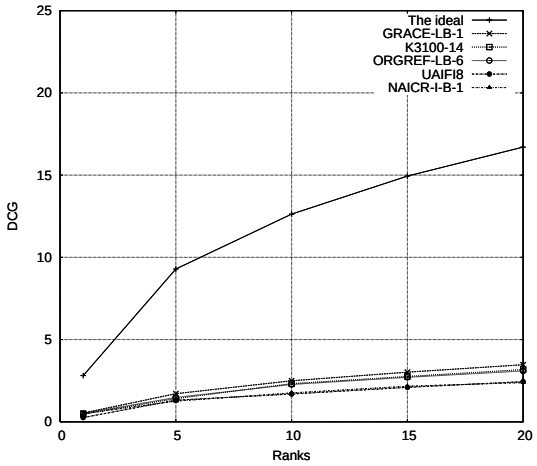


Recall-precision curves for the I-A2 'automatic' and 'TITLE-and-RDOC[1]' runs with considering links (rigid relevance level)

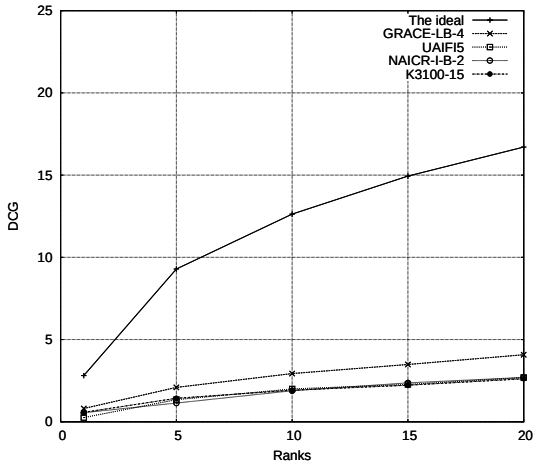


DCG curves for the I-A2 'automatic' and 'TITLE-and-RDOC[1]' runs with considering links (rigid relevance level)

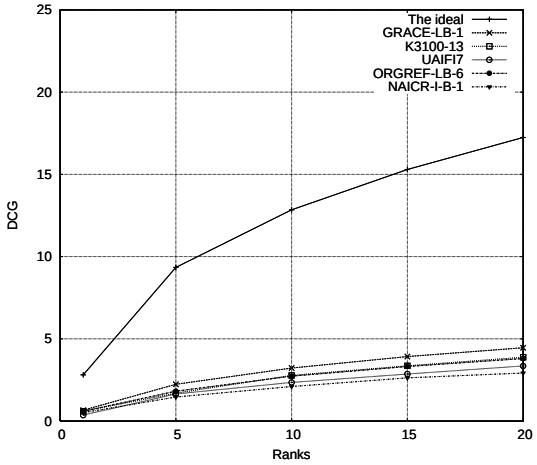
Figure 6. Recall-precision and DCG curves for the I-A2 'automatic' runs



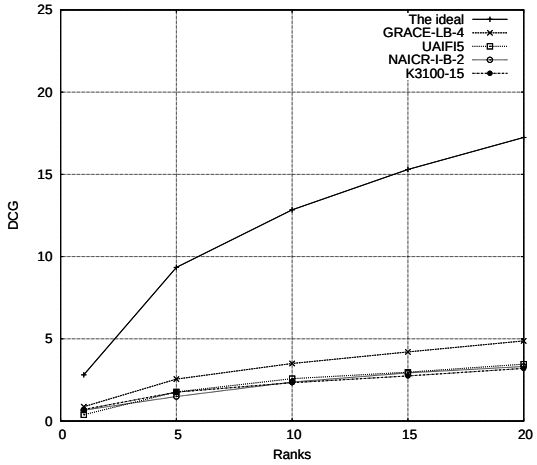
DCG curves for the I-B 'automatic' and 'TITLE-only' runs without considering links (rigid relevance level)



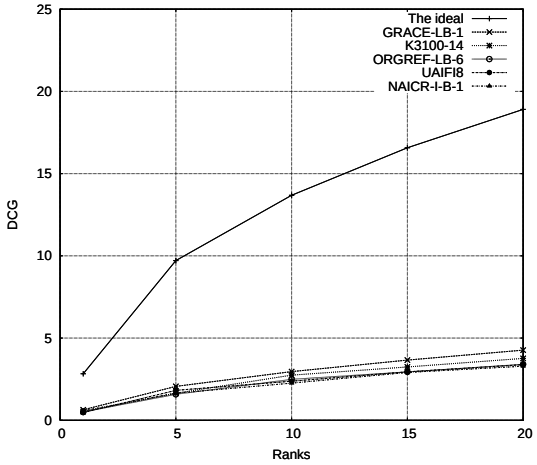
DCG curves for the I-B 'automatic' and 'DESC-only' runs without considering links (rigid relevance level)



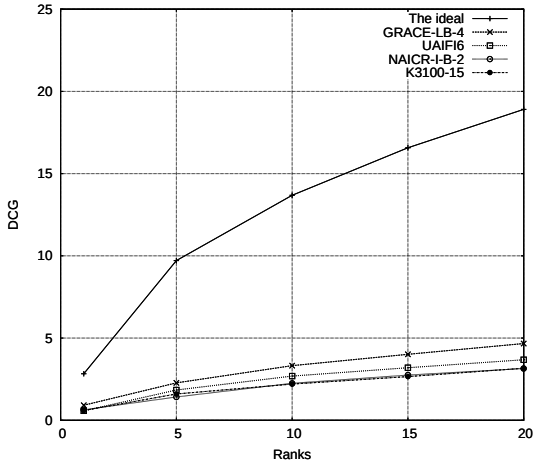
DCG curves for the I-B 'automatic' and 'TITLE-only' runs without considering links (relaxed relevance level)



DCG curves for the I-B 'automatic' and 'DESC-only' runs without considering links (relaxed relevance level)

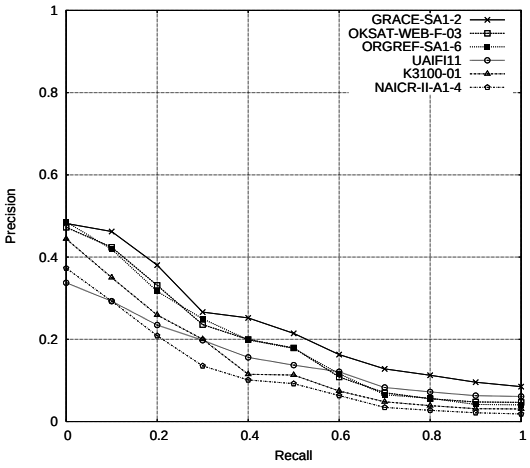


DCG curves for the I-B 'automatic' and 'TITLE-only' runs with considering links (rigid relevance level)

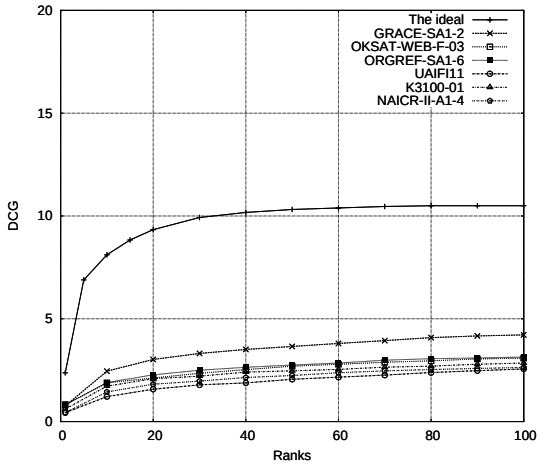


DCG curves for the I-B 'automatic' and 'DESC-only' runs with considering links (rigid relevance level)

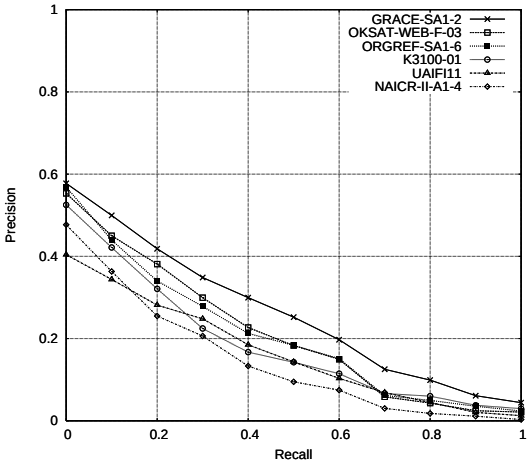
Figure 7. DCG curves for the I-B 'automatic' runs



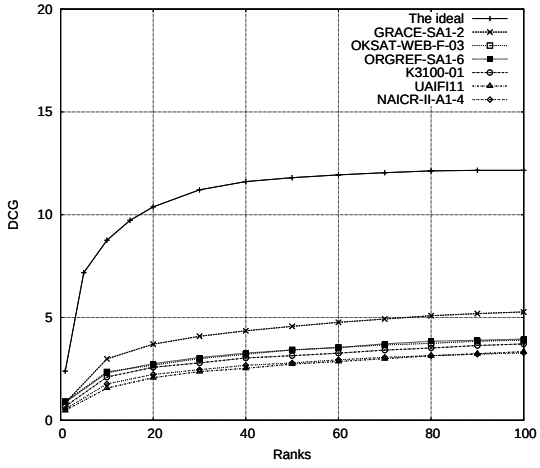
Recall-precision curves for the II-A1 'automatic' and 'TITLE-only' runs without considering links (rigid relevance level)



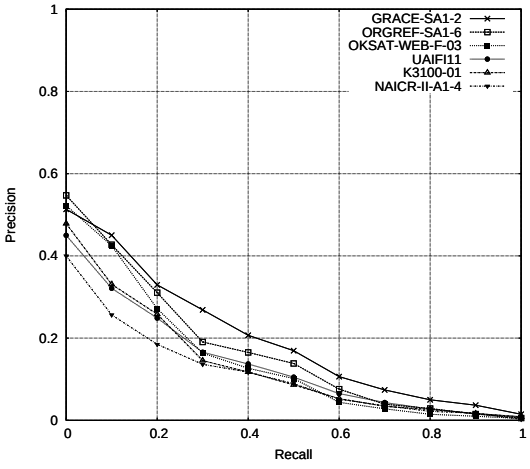
DCG curves for the II-A1 'automatic' and 'TITLE-only' runs without considering links (rigid relevance level)



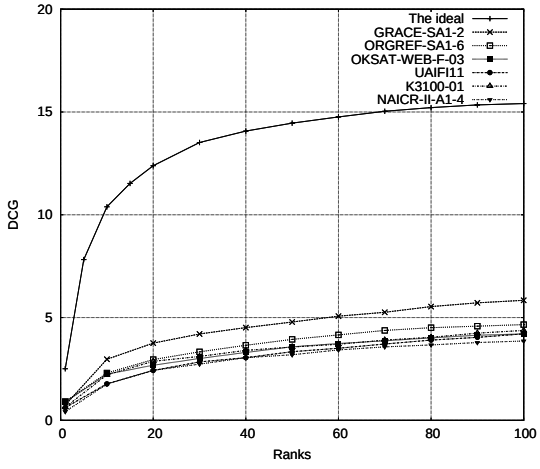
Recall-precision curves for the II-A1 'automatic' and 'TITLE-only' runs without considering links (relaxed relevance level)



DCG curves for the II-A1 'automatic' and 'TITLE-only' runs without considering links (relaxed relevance level)

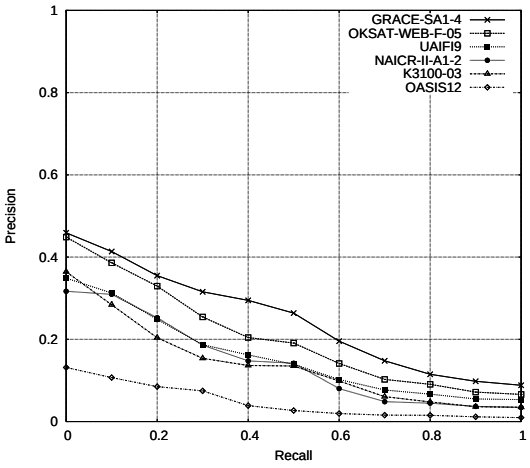


Recall-precision curves for the II-A1 'automatic' and 'TITLE-only' runs with considering links (rigid relevance level)

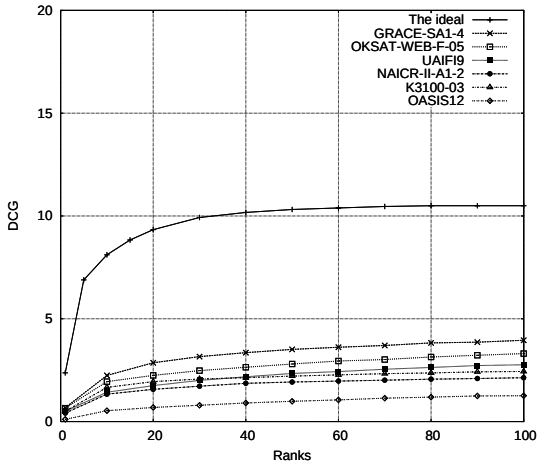


DCG curves for the II-A1 'automatic' and 'TITLE-only' runs with considering links (rigid relevance level)

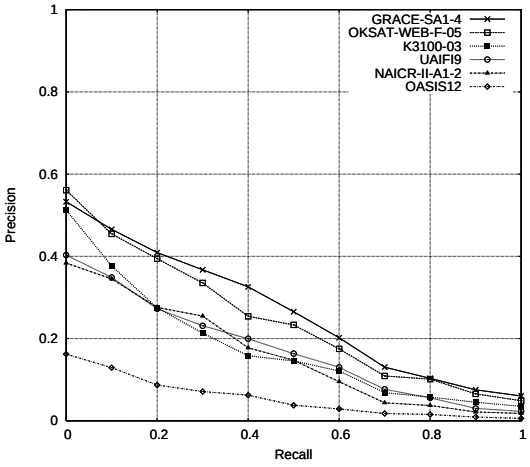
Figure 8. Recall-precision and DCG curves for the II-A1 'automatic' and 'TITLE-only' runs



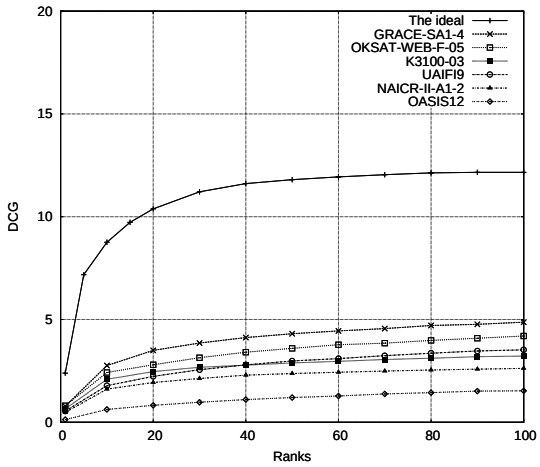
Recall-precision curves for the II-A1 'automatic' and 'DESC-only' runs without considering links (rigid relevance level)



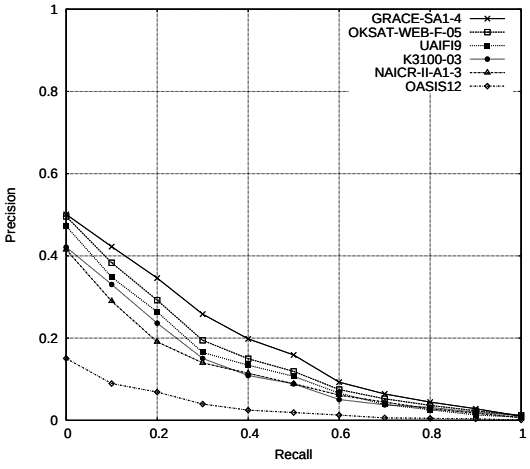
DCG curves for the II-A1 'automatic' and 'DESC-only' runs without considering links (rigid relevance level)



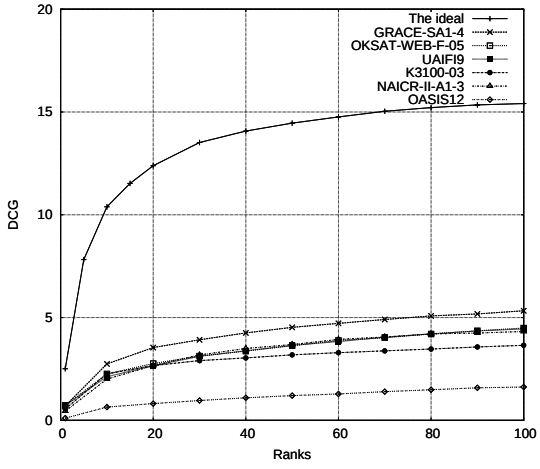
Recall-precision curves for the II-A1 'automatic' and 'DESC-only' runs without considering links (relaxed relevance level)



DCG curves for the II-A1 'automatic' and 'DESC-only' runs without considering links (relaxed relevance level)

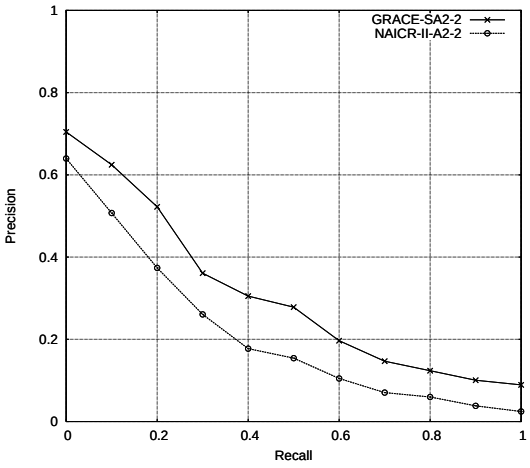


Recall-precision curves for the II-A1 'automatic' and 'DESC-only' runs with considering links (rigid relevance level)

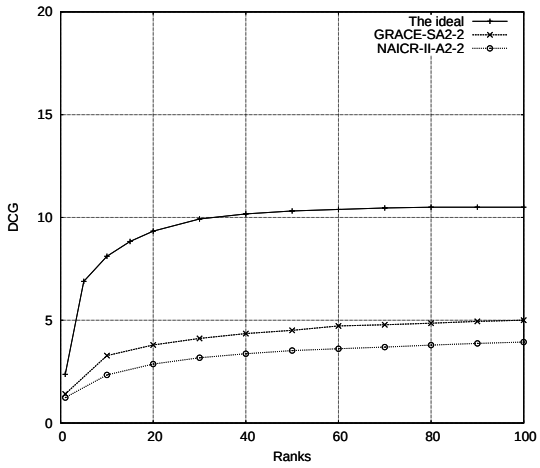


DCG curves for the II-A1 'automatic' and 'DESC-only' runs with considering links (rigid relevance level)

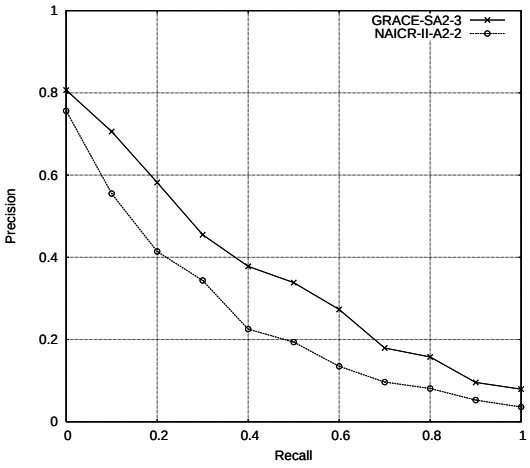
Figure 9. Recall-precision and DCG curves for the II-A1 'automatic' and 'DESC-only' runs



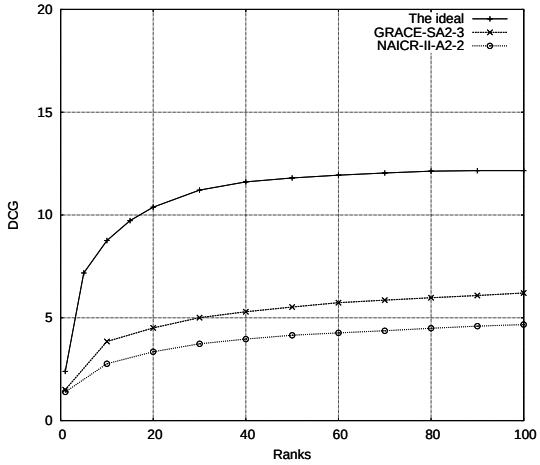
Recall-precision curves for the II-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (rigid relevance level)



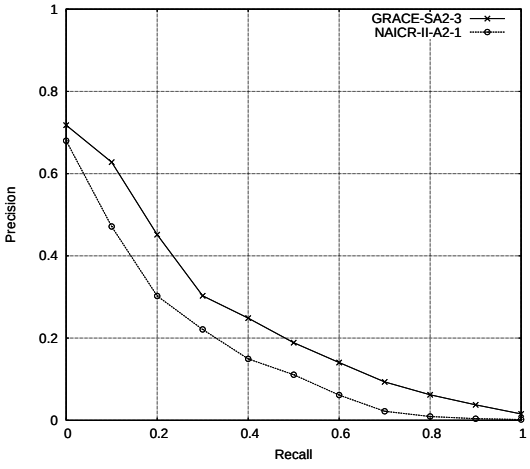
DCG curves for the II-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (rigid relevance level)



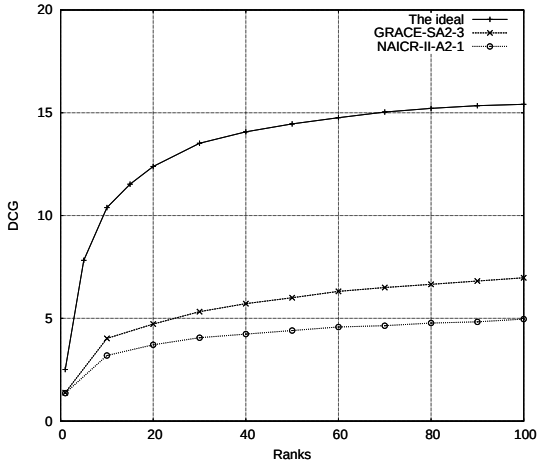
Recall-precision curves for the II-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (relaxed relevance level)



DCG curves for the II-A2 'automatic' and 'TITLE-and-RDOC[1]' runs without considering links (relaxed relevance level)

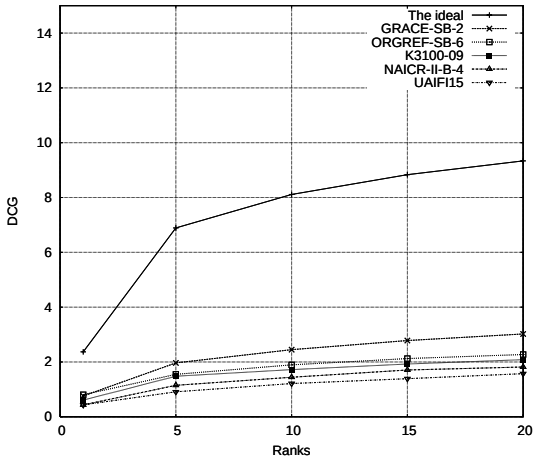


Recall-precision curves for the II-A2 'automatic' and 'TITLE-and-RDOC[1]' runs with considering links (rigid relevance level)

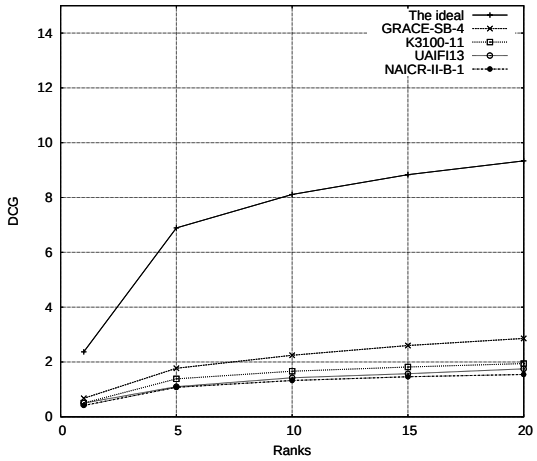


DCG curves for the II-A2 'automatic' and 'TITLE-and-RDOC[1]' runs with considering links (rigid relevance level)

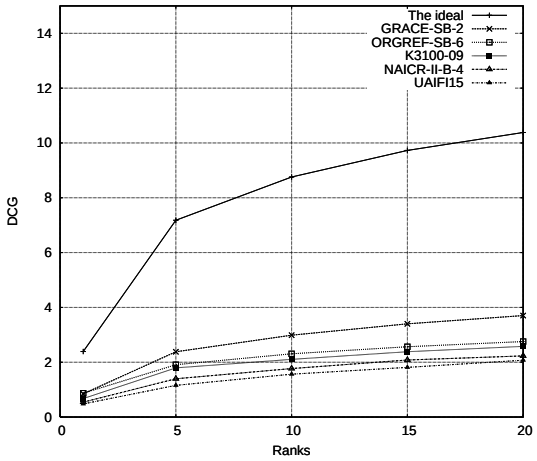
Figure 10. Recall-precision and DCG curves for the II-A2 'automatic' runs



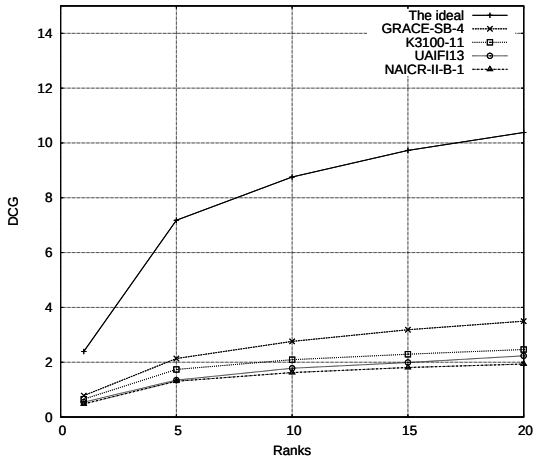
DCG curves for the II-B 'automatic' and 'TITLE-only' runs without considering links (rigid relevance level)



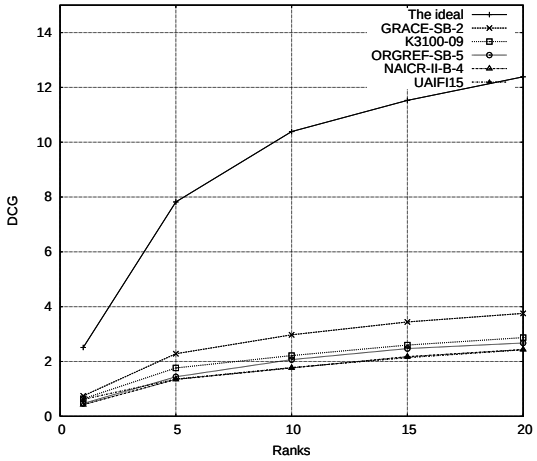
DCG curves for the II-B 'automatic' and 'DESC-only' runs without considering links (rigid relevance level)



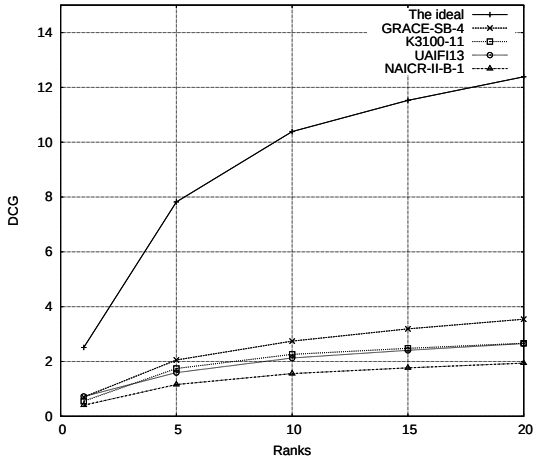
DCG curves for the II-B 'automatic' and 'TITLE-only' runs without considering links (relaxed relevance level)



DCG curves for the II-B 'automatic' and 'DESC-only' runs without considering links (relaxed relevance level)



DCG curves for the II-B 'automatic' and 'TITLE-only' runs with considering links (rigid relevance level)



DCG curves for the II-B 'automatic' and 'DESC-only' runs with considering links (rigid relevance level)

Figure 11. DCG curves for the II-B 'automatic' runs