Information Retrieval, 5, 41–59, 2002

Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop

KAZUKO KURIYAMA kuriyama@nii.ac.jp
NORIKO KANDO kando@nii.ac.jp
National Institutes of Informatics (NII), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

TOSHIHIKO NOZUE tnozue@cl.aoyama.ac.jp
Aoyama Gakuin University, 4-4-25 Shibuya, Shibuya-ku, Tokyo 150-8366, Japan

KOJI EGUCHI eguchi@nii.ac.jp
National Institutes of Informatics (NII), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Received May 1, 2000; Revised April 9, 2001; Accepted May 1, 2001

Abstract. We have conducted a study to: (1) verify the exhaustiveness of pooling for the purpose of constructing a large-scale test collection, and (2) examine whether a difference in the number of pool documents can affect the relative evaluation of IR systems. We carried out the experiments using search topics, their relevance assessments, and the search results that were submitted for both the pre-test and test of the first NTCIR Workshop.

Our results verified the efficiency and the effectiveness of the pooling method, the exhaustiveness of the relevance assessments, and the reliability of the evaluation using the test collection based on the pooling method.

Keywords: test collection, IR system, evaluation methodology, relevance assessment, exhaustiveness of pooling

1. Introduction

1.1. The NTCIR project

We have been constructing a Japanese test collection, for the NACSIS-NII[1] Test Collection for Information Retrieval (NTCIR) Project[2] (NTCIR 2001). The first NTCIR Workshop was held from November 1998 to September 1999 using the NTCIR-1 data (NACSIS Test Collection 1) (preliminary version) (Kando et al. 1999a, Kando and Nozue 1999). This was the first evaluation workshop for Japanese text retrieval, and was similar to that used at the Text REtrieval Conference (TREC) (Voorhees and Harman 2000). It consisted of pre-test and test evaluations. The participating groups submitted their search results for the pre-test on December 2, 1998, and for the test on March 4, 1999.

1.2. The purpose of our experiments

For the construction of a large-scale test collection using the pooling method, the questions we must consider from the viewpoint of testing IR systems are as follows.

(1) *Exhaustiveness of the document pool*: The pooling method is known to be an efficient approach for collecting documents that are likely to be relevant (Gilbert and Sparck Jones 1979). Documents outside the pool are assumed to be not relevant, and so are not judged. Therefore, the question to be answered is, "How can we exhaustively pool candidates for the relevant documents"?

(2) *Reliability of the test collection as a tool for system testing*: A test collection is a tool used for the relative testing of different IR systems. For this purpose, the lists of the relevant documents in the test collection must necessarily be fair for all systems. Therefore, the point to be addressed is whether pooling affects inter-system comparisons. That is, rankings of the search results from the different systems.

The pooling method can be adopted for collecting as many relevant documents as possible. However, the goal of constructing a large test collection by pooling is to not to collect all the total relevant documents, but to collect as many relevant documents needed such that the test collections can enable an unbiased comparison to be carried out sufficiently among the various systems using different algorithms.

1.3. Test collections and pooling methods

A test collection for an IR system test consists of: (1) documents, (2) search topics, and (3) relevance assessments for each search topic. When constructing a test collection, ideally one would judge all the documents for each search topic and make an exhaustive list of the relevant documents. However, this is not feasible for a large-scale database that contains tens of thousands of documents.

The pooling method (Gilbert and Sparck Jones 1979) is a well-known method for effectively and efficiently collecting candidates for the relevant documents in a large-scale test collection. In this approach, the top X documents retrieved by various systems using different retrieval algorithms for each topic are pooled, and then each document in a pool is judged by human assessors. Since 1992, the TREC has constructed large-scale test collections by the pooling method.

Recently, the Move-To-Front (MTF) pooling method was proposed as an improved variation on the pooling method (Cormack et al. 1998). In contrast to the pooling method, the MTF pooling method prioritizes the search results, and pools many more documents from the results with top priority, which are then judged. It has been shown that the MTF pooling method effectively produces a collection with considerably fewer judgments than would be required for the pooling method (Cormack et al. 1998). However, there remains a question for IR systems testing: whether it is unfair to change the number of documents pooled from each search result according to its priority.

Therefore, we experimented with various pooling methods to verify: (1) the exhaustiveness of the document pool, and (2) the reliability of the test collection as a tool for system testing.

1.4. Organization of this paper

In Section 2 we report on the test collection NTCIR-1, and on the first NTCIR Workshop pre-test and test results.

To verify (1) the exhaustiveness of the document pools, we discuss the experimental pooling using the search results for the pre-test in Section 3, and similarly for the test in the first half of Section 4. As an investigation on (2) the reliability of the relevance assessments, we report on the system testing for the test in the second half of Section 4. Finally, we summarize our results in Section 5.

2. The first NTCIR workshop

2.1. Test collection NTCIR-1

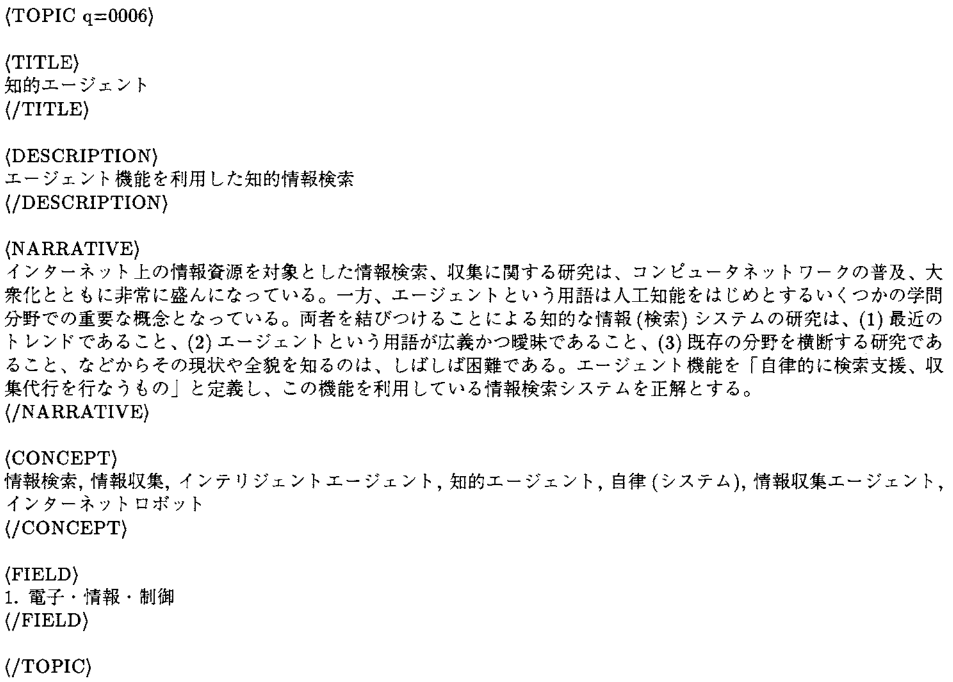
The test collection NTCIR-1 (NACSIS Test Collection 1) consisted of the following.

(1) *Document collection*. The NTCIR-1 contained three document collections: the JE Collection, the J Collection, and the E Collection. The JE Collection contained 339, 483 documents, more than half of which were present as English-Japanese paired. The J and E Collections were constructed by extracting the respective Japanese or English parts of the documents in the JE Collection. The documents were composed of the author abstracts of papers presented at conferences hosted by 65 Japanese academic societies, with a wide variation in their lengths and subject domains. A document contains the following fields: “title”, “author”, “name of conference”, “date of conference”, “abstract”, and “keywords” (which were assigned by the author(s) of the documents). These were extracted from the original database, “Academic Conference Papers”, provided by NACSIS.

(2) *Search topics*. A search topic contains SGML-like tags. A topic consists of a title for the topic, a description, a detailed narrative, a list of concepts, and field(s) in Japanese. A title can be used as a very short query that is often submitted to Internet search engines. NTCIR-1 contained 30 training topics, and 53 test topics. Figure 1 shows an example of a search topic, and figure 2 shows its English translation.

(3) *Relevance assessments for each search topic*. The relevance assessments for each topic were undertaken separately by two assessors, and then cross-checked. The final assessment was based on negotiations between the two assessors and determined by the primary assessor, and creator, of the topic (who was also one of the two assessors). The assessment assigned one of three possible grades; relevant (A), partially-relevant (B), and non-relevant (C).

(4) *Tagged corpus*. The tagged corpus contained detailed part-of-speech tags and was used in the term extraction tasks in the Workshop (Kageura et al. 1997).



⟨TOPIC q=0006⟩

⟨TITLE⟩
知的エージェント
⟨/TITLE⟩

⟨DESCRIPTION⟩
エージェント機能を利用した知的情報検索
⟨/DESCRIPTION⟩

⟨NARRATIVE⟩
インターネット上の情報資源を対象とした情報検索、収集に関する研究は、コンピュータネットワークの普及、大衆化とともに非常に盛んになっている。一方、エージェントという用語は人工知能をはじめとするいくつかの学問分野での重要な概念となっている。両者を結びつけることによる知的な情報(検索) システムの研究は、(1) 最近のトレンドであること、(2) エージェントという用語が広義かつ曖昧であること、(3) 既存の分野を横断する研究であること、などからその現状や全貌を知るのは、しばしば困難である。エージェント機能を「自律的に検索支援、収集代行を行なうもの」と定義し、この機能を利用している情報検索システムを正解とする。
⟨/NARRATIVE⟩

⟨CONCEPT⟩
情報検索, 情報収集, インテリジェントエージェント, 知的エージェント, 自律 (システム), 情報収集エージェント, インターネットロボット
⟨/CONCEPT⟩

⟨FIELD⟩
1. 電子・情報・制御
⟨/FIELD⟩

⟨/TOPIC⟩

Figure 1. An example of a search topic.

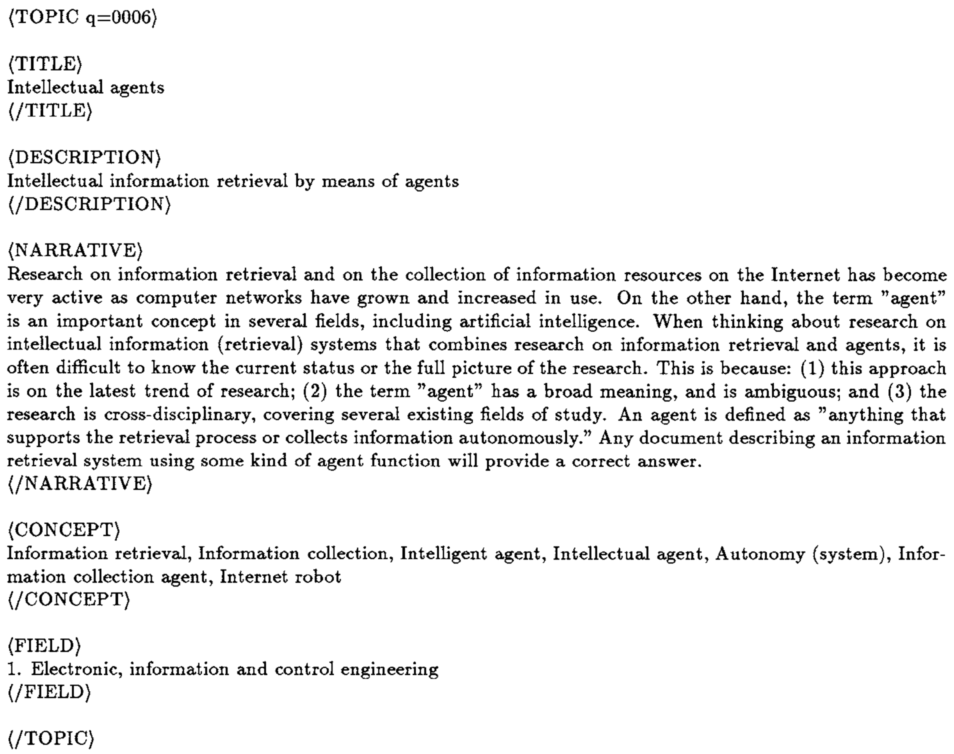
2.2. *Outlines of the pre-test and test*

2.2.1. The tasks of pre-test and test. Each of the participants in the first NTCIR Workshop had conducted one or more of the following tasks.

The Ad Hoc IR task—to investigate the retrieval performance of systems that search a static set of documents using new search topics. The documents were in Japanese and English (JE Collection), and the topics were in Japanese for this task.
The Cross-Lingual IR task—an ad hoc task in which the documents were in English (E Collection) and the topics were in Japanese.
The Mono-Lingual IR task—an ad hoc task in which the documents were in Japanese (J Collection) and the topics were in Japanese. This was an optional task.

2.2.2. The pre-test. For NTCIR-1, we had prepared beforehand preliminary relevance assessments for the training search topics through pooling using the IR systems of the NACSIS. In order to examine the exhaustiveness of these relevance assessments and the reliability of the test collection, we carried out a pre-test on December 2, 1998 for the first NTCIR Workshop (Kando et al. 1999a).

In the pre-test, the participating groups submitted search results for 30 training topics, and we then completed the relevance assessments by adding new relevant documents that were found in the search results.



⟨TOPIC q=0006⟩

⟨TITLE⟩
Intellectual agents
⟨/TITLE⟩

⟨DESCRIPTION⟩
Intellectual information retrieval by means of agents
⟨/DESCRIPTION⟩

⟨NARRATIVE⟩
Research on information retrieval and on the collection of information resources on the Internet has become very active as computer networks have grown and increased in use. On the other hand, the term "agent" is an important concept in several fields, including artificial intelligence. When thinking about research on intellectual information (retrieval) systems that combines research on information retrieval and agents, it is often difficult to know the current status or the full picture of the research. This is because: (1) this approach is on the latest trend of research; (2) the term "agent" has a broad meaning, and is ambiguous; and (3) the research is cross-disciplinary, covering several existing fields of study. An agent is defined as "anything that supports the retrieval process or collects information autonomously." Any document describing an information retrieval system using some kind of agent function will provide a correct answer.
⟨/NARRATIVE⟩

⟨CONCEPT⟩
Information retrieval, Information collection, Intelligent agent, Intellectual agent, Autonomy (system), Information collection agent, Internet robot
⟨/CONCEPT⟩

⟨FIELD⟩
1. Electronic, information and control engineering
⟨/FIELD⟩

⟨/TOPIC⟩

Figure 2. An example of a search topic (English translation).

For the pre-test, the ten participating groups submitted a total of 23 sets of search results for the Ad Hoc IR task, the Cross-Lingual IR task, and the optional Mono-Lingual IR task as a baseline for the CLIR search results. The 23 sets comprised of 16 sets from the ten groups for the Ad Hoc IR task, five sets from four groups for the Cross-Lingual IR task, and two sets from one group for the Mono-Lingual IR task. The 23 sets contained 4 sets retrieved by interactive IR systems, three sets of which were for the Ad Hoc IR task and one set of which were for the Cross-Lingual IR task. We mean here that an interactive IR system was a system which used interactive method for query construction.

We experimented on the 16 sets of the search results for the Ad Hoc IR task. The relevance assessments were carried out using the three grades described above; but in this paper, we define "relevant" to include both "relevant" and "partially-relevant" rankings.

2.2.3. The test. The test was held on March 4, 1999 at the first NTCIR Workshop (Kando et al. 1999a). Twenty-three participating groups submitted a total of 121 sets of search results for 53 topics. The 121 sets consisted of 47 sets from 18 groups for the Ad Hoc IR task, 69 sets from 11 groups for the Cross-Lingual IR task, and five sets from five groups for the Mono-Lingual task. The 121 sets contained 12 sets retrieved by interactive IR systems, eight sets of which were for the Ad Hoc IR task and four sets of which were for the Cross-Lingual IR task.

Based on the results of the pre-test, the relevance assessments for the test were prepared as follows: the top X documents from each submitted search result were judged first,[3] and additional interactive searches were then carried out for the topics with more than 50 relevant documents. The new documents in the search results were judged, and then added to the original list of the relevant documents.

2.3. Definition

In this paper we refer to a search result as a “submission”, to differentiate the “submitted search results” from the general term “search results”. A submission is a file in which the top 1000 documents are listed in order of topic number for each of the 30 training search topics of the pre-test, or each of the 53 search topics of the test.

3. Pooling for the pre-test

3.1. Exhaustiveness of the pooling in NACSIS

We have discussed the previous experimental results in an earlier paper (Kando et al. 1999a) using the following pools:

V1: version 1 of the relevance assessments prepared before the pre-test;
A: the top-ranked documents of the automatic search results of more than 30 different runs by the three IR systems at NACSIS for **V1**;
I: the additional search results by recall-oriented manual searches conducted by graduate students who had majored in library and information science in consideration of recall for **V1**;
F: the final relevance assessments (version 2), which were made by adding the new relevant documents found in the pools from the submissions by the voluntary participating groups to **V1**; and
P: a pool of the top 100 documents from the 23 submissions for the pre-test.

We assume that **F** comprises the complete relevance assessments when considering the efficiency and effectiveness of the pooling method.

(1) **V1** contained 97.1% of all the relevant documents.
(2) If we performed evaluations using the lists of relevant documents based on any of **V1**, **A**, **I**, **P** or **F**, then the rankings from the mean average precision of the submissions changed slightly, but the correlation was over 0.80 and very high. Therefore, we concluded that there was little effect on the testing of the different systems.

(3) The **I** (additional interactive) searches found 17.5% of all the relevant documents in **F** uniquely, which were not found by the other methods, that is, the unique contribution of **I** to all the relevant documents was 17.5%.
(4) This was not advantageous for the interactive IR system, even if we evaluated the systems using the **I** set.

3.2. Exhaustiveness of the pools of the submissions

3.2.1. Method. Based on the above results, we now focus our attention on the exhaustiveness of pooling using only the documents from the submissions of the participating groups. We carried out an experiment using the 16 submissions for the Ad Hoc IR task of the pre-test.

For $X = 10, \ldots, 1000$, the top X documents from each submission were pooled. We refer to each of the pools as **P**X, respectively, and to the pool **P**X combined with the set **I** as **P** $\boldsymbol{XI}$.

Table 1 and figure 3 show the numbers of relevant documents for each search topic contained in the pools. In Table 1, the values rel-is (i = 0–50, 50–100, 100, all) show the average percentages of the relevant documents in each pool relative to **F** for the topics with R relevant documents (rel-all: $0 \leq R$; rel-0–50: $0 \leq R < 50$; rel-50–100: $50 \leq R < 100$; rel-100: $100 \leq R$). Table 2 shows the total number of documents in each pool. The maximal number of retrieved documents for a topic in a submission was 1000. As the number of relevant documents for topic 0028 was greater than 1000, and the number of pooled

Table 1. Number of relevant documents per topic in the pools for the pre-test.

Pool	F	I	P10	P30	P100	P1000	P100I
Min (0016)	5	5	4	5	5	5	5
Max (0014)	317	283	55	96	210	310	287
Average (0001–0027,0029–0030)	66.3	56.7	18.6	30.3	46.3	64.2	61.0
Total (0001–0027,0029–0030)	1922	1645	538	880	1342	1863	1770
rel-all (All) (29 topics) (%)	100.0	82.0	43.4	61.8	79.8	97.0	94.9
rel-100 ($100 \leq R$) (5 topics) (%)	100.0	81.9	17.5	33.8	59.7	95.3	87.6
rel-50-100 ($50 \leq R < 100$) (6 topics) (%)	100.0	92.6	35.8	57.5	79.9	98.4	95.7
rel-0–50 ($0 \leq R < 50$) (18 topics) (%)	100.0	77.6	55.4	74.1	89.2	98.8	96.3

F: the final relevance assessments (ver.2).
I: the additional interactive search results at NACSIS.
PX: a pool of the top X documents from each submission.
P$\boldsymbol{XI}$: the pool **P**X combined with **I**.
Min: the minimal number of relevant documents for the topic.
Max: the maximal number of relevant documents for the topic.
Average: the average of the relevant documents for all topics except for topic 0028.
Total: the total number of relevant documents for all topics except for topic 0028.
rel-i: the average percentage of the relevant documents in each pool relative to **F**, where rel-all is for all topics, rel-R values are for the topics with R relevant documents while rel-100: $100 \leq R$, rel-50–100: $50 \leq R < 100$, and rel-0–50: $R < 50$.

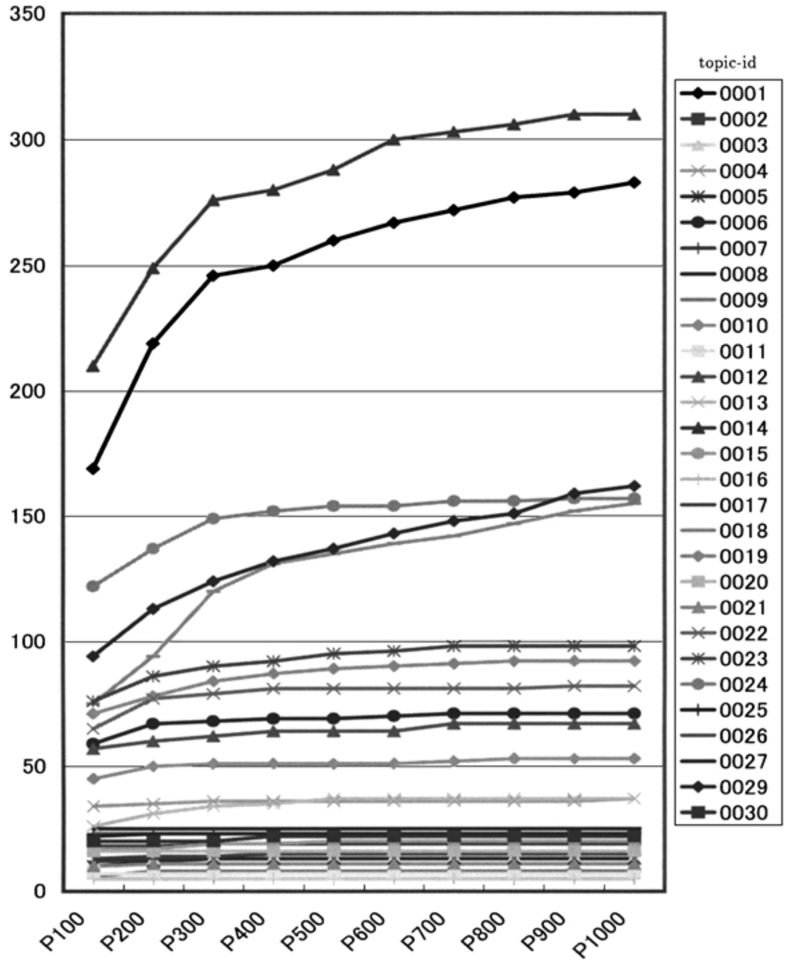


Figure 3. Number of relevant documents per topic in the pools for the pre-test.
P*X*: a pool of the top *X* documents from each submission.

documents for it may not be large enough, the results for the topic 0028 are excluded in the following tables.

3.2.2. Results. We see from rel-all in Table 1 that **P100** includes almost the same percentage of relevant documents as **I**. From rel-100, rel-50–100, and rel-0–50, we see that **I** keeps its high exhaustiveness for topics with many relevant documents, while **P100** and **P1000** are not exhaustive for such topics. For example, for topics with less than 50 relevant documents, **P100** found 89.2% of the total relevant documents, whereas **P100** found only 59.7% of the total relevant documents for topics with more than 100 relevant documents. In summary, we can conclude that it is necessary to pool many more documents for topics with more than 50 relevant documents.

Table 2. Total number of documents per topic in the pools for the pre-test.

Pool	F	I	P10	P30	P100	P1000	P10I	P30I	P100I
Min (0018)	2041	11	49	120	367	3407	62	154	413
Max (0004)	7900	5174	100	301	859	6870	5191	5251	5518
Total (29 topics)	129149	53841	2046	5774	17966	150994	54668	56720	64841
Average (29 topics)	4453.4	1922.9	70.6	199.1	619.5	5206.7	1885.1	1955.9	2235.9

F: the final relevance assessments (ver.2), **I**: the additional interactive search results at NACSIS.
PX: a pool of the top X documents from each submission.
PXI: the pool **P**X combined with **I**.
Min: the minimal number of pooled documents for the topic.
Max: the maximal number of pooled documents for the topic.
Total: the total number of pooled documents for all topics except for all topic 0028.
Average: the average of the pooled documents for all topics except for topic 0028.

In Table 1, it can be seen that the percentages of **P100I** are higher than both **P100** and **I**. Although, from Table 2, the numbers of documents in **P100I** are about a half that of **P1000**, from rel-all in Table 1 we can see that **P100I** covers 94.9% of the number of documents. Therefore, it is possible to complete the pools using an additional interactive search if the number of pooled documents is large enough.

These results show that it is not enough to simply pool the top X documents from the submissions, and that it is necessary to complete the pools by an additional recall-oriented interactive search for topics with many relevant documents.

3.3. Unique relevant documents in the pools

In order to determine how many unique relevant documents an additional interactive search can collect, we counted the number of unique documents in each of the automatic search results in version 1 **A**, version 2, **I**, and in the pool **P100**. Table 3 shows the numbers and percentages found.

A-only, I-only and **P100-only** show, respectively, the number of relevant-documents contained only in each one of the three pools. The rel-*i* values correspond to the rel-*i* values in Table 1. The “total” is the total number of the relevant documents for all topics except for topic 0028.

From Table 3 we see that **I-only** contains a higher percentage of uniquely found relevant documents than either **A-only** or **P100-only**. In particular, the average of **I-only** for the topics with more than 100 relevant documents was 16.0%, and the total number of relevant documents for **I-only** was 256.

3.4. Conclusion of the pre-test

According to the above results, for the test, we carried out an additional recall-oriented interactive search to collect additional candidates for relevant documents for topics with more than 50 relevant documents.

Table 3. Number of unique relevant documents in the pools for the pre-test.

		A, I, P100		
Topic	F	A-only	I-only	P100-only
rel-all (All) (29 topics) (%)		2.3	7.0	1.7
rel-100 ($100 \leq R$) (5 topics) (%)		6.5	16.0	0.6
rel-50–100 ($50 \leq R < 100$) (6 topics) (%)		2.3	10.5	0.9
rel-0–50 ($0 \leq R < 50$) (18 topics) (%)		1.2	3.3	2.2
Total (29 topics)	1922	68	256	18

F: the final relevance assessments, **A**: the automatic search results at NACSIS.
I: the additional interactive search results at NACSIS.
P100: a pool of the top 100 documents from each submission.
rel-i: the average percentage of the relevant documents in each pool relative to **F**, where rel-all is for all topics, rel-Rs are for the topics with R relevant documents while rel-100: $100 \leq R$, rel-50–100: $50 \leq R < 100$, and rel-0–50: $R < 50$.
Total: the total of relevant documents for all topics except for topic 0028.

4. Pooling for the test

4.1. Pooling and the results

4.1.1. Pooling. We expected that pooling for the test would be the same as the pooling for the pre-test. We wanted to verify this expectation, and focused our attention on the exhaustiveness of the pooled documents obtained from the submissions by the participating groups alone, and on the reliability of the relevance assessments as a tool for system testing. We therefore performed an experiment using the 47 sets of submissions for the Ad Hoc IR task of the test.

We referred to each of the pools as follows.

PX: the pool in which the top X documents from each submission were pooled ($X = 1, \ldots, 1000$);
I: the additional search results by recall-oriented manual searches conducted by graduate students who had majored in library and information science in consideration of recall for the ten search topics with more than 50 relevant documents in **P100**;
PXI: the pool **P**X combined with **I**;
W/C100: contained the top 100 documents from the submissions by systems using ⟨CONCEPT⟩s in the topics;
WO/C100: contained the top 100 documents from submissions by systems not using ⟨CONCEPT⟩s in the topics (**W/C100** plus **WO/C100** is **P100**); and
F: the final relevance assessment.

From the earlier TREC results, it was known that the searches using ⟨CONCEPT⟩s of the topics obtained a much higher search effectiveness than searches without them. Therefore, *Table 4*. Number of relevant documents per topic in the pools for the test.

Pool	F	I	P10	P30	P100	P1000	P10I	P30I	P100I
Min (0077, 0078)	6		6	6	6	6	6	6	6
Max (0054)	584	504	58	123	253	523	510	521	568
Average (0031–0083)	44.2		18.7	27.7	35.9	43.1	35.5	40.0	43.2
Total (0031–0083)	2345	1256	989	1468	1904	2282	1879	2118	2288
rel-all (All) (53 topics) (%)	100.0		64.3	84.0	95.6	99.7	73.8	89.6	97.8
rel-100 ($100 \leq R$) (4 topics) (%)	100.0	91.2	28.8	49.3	76.4	97.4	92.3	94.5	98.0
rel-50–100 ($50 \leq R < 100$) (6 topics) (%)	100.0	84.0	46.2	73.4	92.2	99.8	88.2	93.1	97.2
rel-50 ($50 \leq R$) (10 topics) (%)	100.0	86.9	39.2	63.8	85.9	98.8	89.8	93.7	97.6
rel-0–50 ($0 \leq R < 50$) (43 topics) (%)	100.0		70.1	88.7	97.8	99.9	70.1	88.7	97.8

F: the final relevance assessments, **I**: the additional interactive search results at NACSIS.

PX: a pool of the top X documents from each submission.

P$\boldsymbol{XI}$: the pool **P**X combined with **I**.

Min: the minimal number of relevant documents for the topic.

Max: the maximal number of relevant documents for the topic.

Average: the average of the relevant documents for all topics.

Total: the total number of relevant documents for all topics.

rel-i: the average percentage of the relevant documents in each pool relative to **F**, where rel-all is for all topics, rel-Rs are for the topics with R relevant documents while rel-100: $100 \leq R$, rel-50–100: $50 \leq R < 100$, rel-50: $50 \leq R$, and rel-0–50: $R < 50$.

we also wanted to know the level of contribution of ⟨CONCEPT⟩s in finding unique relevant documents.

Table 4 and figure 4 show the number of relevant documents for each search topic contained in the pools, and Table 5 shows the total number of documents in the pools. The rel-is (i = 0–50, 50, 50–100, 100, all) show the average percentages of relevant documents in each pool relative to **F** for topics with R relevant documents (rel-all: $0 \leq R$; rel-0–50: $0 \leq R < 50$; rel-50: $50 \leq R$; rel-50–100: $50 \leq R < 100$; rel-100: $100 \leq R$).

4.1.2. Exhaustiveness of the pools. We can see from rel-50 in Table 4 that **P100** contains almost the same number of relevant documents as **I**, thus matching its exhaustiveness. For **P10**, **P30**, and **P100**, the exhaustiveness for topics with less than 50 relevant documents was higher than for those with more than 50 relevant documents. In summary, it is necessary to pool many documents for topics with numerous relevant documents if the number of submissions is large.

In pooling, we used the 16 sets of submissions for the pre-test, and we used the 49 sets of submissions for the test, making three times as many submissions as for the pre-test. For

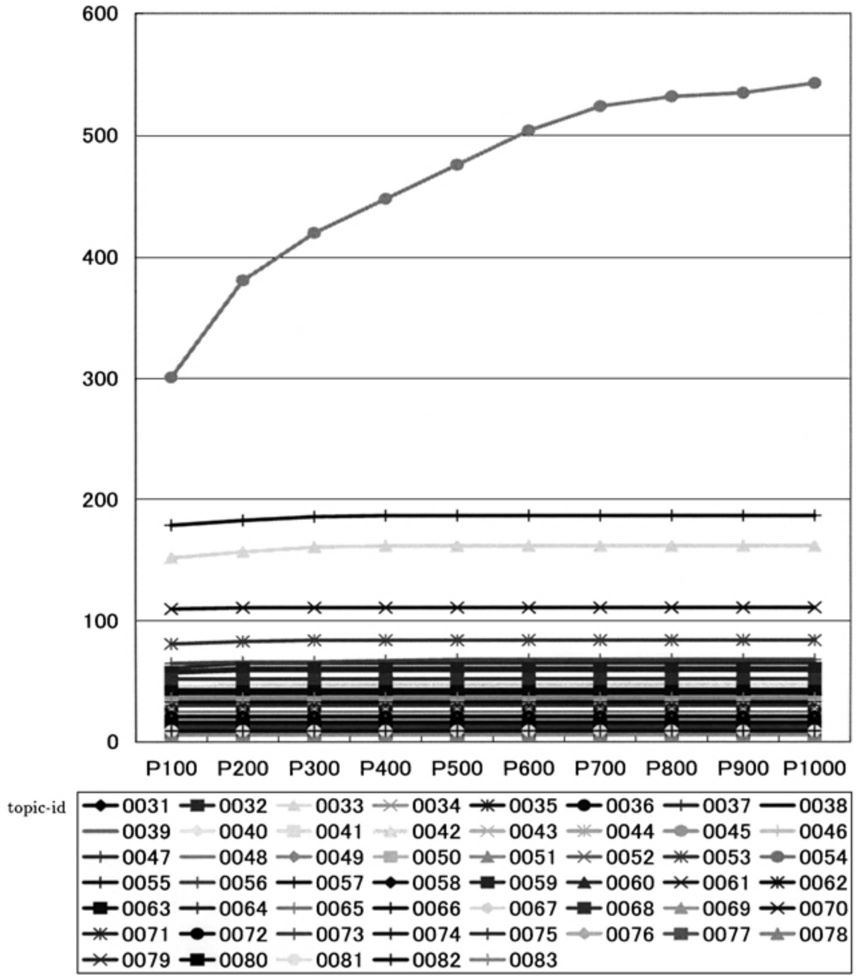


Figure 4. Number of relevant documents per topic in the pools for the test.
P*X*: a pool of the top *X* documents from each submission.

both pools, although the top-ranked documents from each submission overlapped, there were about three times as many documents in the pools for the test as for the pre-test. It is necessary to consider the effect upon the number of pooled documents as the number of submissions increases, although the probabilities of relevance for the documents in each submission are possibly correlated with the characteristics of the IR systems, and with the nature and number of relevant documents for the search topics.

It was expected that pooling would produce more exhaustive relevance assessments with a large enough number of submissions. However, it is not exhaustive for topics with many relevant documents. As the above results show, even though the test had many more submissions than the pre-test, it was not enough to pool the top *X* documents from the

Table 5. Total number of documents per topic in the pools for the test.

Pool	F	I	P10	P30	P100	P10I	P30I	P100I
Min (0055)	1554	546	77	282	805	77	282	805
Max (0059)	3943	2253	221	601	1718	2310	2478	3208
Total (53 topics)	132672	12484	7599	21656	68189	19399	32677	77543
Average (53 topics)	2503.2	1248.4	143.4	408.6	1286.6	366.0	616.5	1463.1

F: the final relevance assessments, **I**: the additional interactive search results at NACSIS.
PX: a pool of the top X documents from each submission.
PXI: the pool **P**X combined with **I**.
Min: the minimal number of pooled documents for the topic.
Max: the maximal number of pooled documents for the topic.
Average: the average of the pooled documents for all topics.
Total: the total number of pooled documents for all topics.

submissions to prepare relevance assessments. It was also necessary to complete pooling by carrying out additional interactive searches for those topics with many relevant documents. From rel-100 in Table 4, **P30I** covered 93.7% for all the topics, which was higher than the coverage of **P100**. If the number of pooled documents was not very large, the additional interactive search could be completed effectively.

4.1.3. Unique relevant documents in the pools. To determine how many unique relevant documents could be found (a) by an additional interactive search and (b) by automatic runs making use of ⟨CONCEPT⟩ fields in the topic, we counted the number of unique documents in each of the pools **I**, **W/C100**, and **WO/C100**. These numbers and percentages are shown in Table 6.

I-only, **W/C100-only**, and **WO/C100-only** show, respectively, the numbers of relevant documents only contained in each one of the three pools. The rel-i values correspond to the rel-i values in Table 4. The “total” is the total number of relevant documents for all the topics.

We can see in Table 6 that **I-only** contains 384 documents, and covers 16.4% for all the topics and 26.8% of the topics with more than 50 relevant documents. **W/C100-only** had a higher coverage for all the topics than **WO/C100-only** did, while the percentages of **W/C100-only** and **WO/C100-only** were almost the same for the 10 topics. However, they covered much less than **I-only**. Therefore, the additional recall-oriented interactive searches are effective in collecting more relevant documents.

4.2. System testing using different pools

4.2.1. System testing. To examine whether there was some effect on the evaluation, we carried out experimental evaluations of the submissions using the lists of relevant documents based on the pools shown in the previous subsections. We selected the ten submissions with the best mean average precision, submitted by the different groups from the 47 sets submitted for the Ad Hoc IR task. We selected the ten submissions from different groups because we

Table 6. Number of unique relevant documents in the pools for the test.

		I, W/C100, WO/C100		
Pool	F	I-only	W/C-only	WO/C-only
rel-all (All) (53 topics) (%)	100.0		4.4	2.5
rel-100 ($100 \leq R$) (4 topics) (%)	100.0	21.7	1.9	2.2
rel-50–100 ($50 \leq R < 100$) (6 topics) (%)	100.0	5.0	1.6	3.2
rel-50 ($50 \leq R$) (10 topics) (%)	100.0	11.7	1.7	2.8
rel-0–50 ($0 \leq R < 50$) (43 topics) (%)	100.0		5.0	2.4
Total (53 topics)	2345	384	93	73
Total-50 ($50 \leq R$) (10 topics)	1433	384	41	46

F: the final relevance assessments.
I-only: the number the relevant documents contained only by the additional interactive searches **I** at NACSIS.
W/C100-only: the number of relevant documents contained only by the systems using ⟨CONCEPT⟩s.
WO/C100-only: the number of relevant documents contained only by the systems not using ⟨CONCEPT⟩s.
rel-i: the average percentage of the relevant documents in each pool relative to **F**, where rel-all is for all topics, rel-Rs are for the topics with R relevant documents while rel-100: $100 \leq R$, rel-50–100: $50 \leq R < 100$, rel-50: $50 \leq R$, and rel-0–50: $R < 50$.
Total (53 topics): the total number of relevant documents for all topics.
Total-50 ($50 \leq R$) (10 topics): the total number of relevant documents for the topics with more than 50 relevant documents.

supposed that the submissions from a group would be similar and would not have a large difference in their mean average precisions. We then separately selected the submission with the best mean average precision from all the submissions per group. Each submission was given a run-id: a, b, c, d, e, f, g, h, j and k.

We also evaluated the submissions using a list of relevant documents in each pool as follows.

W/C100I: the pool **W/C100** combined with **I**;
WO/C100I: the pool **WO/C100** combined with **I**;
IS100: the pool of the top 100 documents from the submissions by the interactive IR systems;
AS100: the pool of the top 100 documents from the submissions by the automatic IR systems;
IS100I, AS100I: the pools **IS100** and **AS100** combined with **I**, respectively;
P100-i: the **P100** without the submission "i", and other submissions retrieved by the group that submitted the "i", (i = a, b, c, d, e, f, g, h, j and k); and
P100I-i: the pools **P100**-i combined with **I**, respectively.

We scored the ten submissions and ranked them by their mean average precisions. Table 7 shows the results with the distinction of rankings highlighted by different shades and fonts. To rank the submissions for relative comparison, we assumed that if the difference between the average precisions was larger than 5%, then the difference was important and the rankings were different.[4]

We computed the difference of the mean average precision of each submission in Table 7, and we found that there were a few pairs of the submissions with significant differences, and two or more submissions had the same ranking in each row of the table. Thus, we could divide the submissions that had less than a 5% difference into four groups. The first group contains the top-ranked submission, the second contains the second-ranked 2 submissions, the third contains the third-ranked 3 submissions, and the fourth contains the fourth-ranked 4 submissions in each row, respectively. We now find that the grouped rankings are the same.

To examine whether or not there was a correlation between the ranking by the mean average precisions using **F** and the one using each of the other pools, we also computed the

Table 7. Mean average precisions and rankings of the ten submissions for the test.

Run-id	a	b	c	d	e	f	g	h	j	k	
Method,	Inter,	Inter,	Auto,	Auto,	Auto,	Auto,	Auto,	Auto,	Auto,	Auto,	
concept	w/c	w/c	w/c	w/c	wo/c	wo/c	wo/c	wo/c	wo/c	wo/c	t-stat
F	**0.5378**	**0.4426**	**0.4360**	*0.3499*	*0.3498*	*0.3484*	0.3429	0.2592	0.2587	0.2584	-
P10	**0.6166**	**0.5153**	**0.5297**	*0.4225*	*0.4241*	*0.4098*	0.4179	0.3116	0.3170	0.3226	9.876
P30	**0.5826**	**0.4790**	**0.4816**	*0.3799*	*0.3857*	*0.3812*	0.3798	0.2808	0.2864	0.2863	9.783
P100	**0.5508**	**0.4550**	**0.4472**	*0.3569*	*0.3592*	*0.3570*	0.3508	0.2647	0.2650	0.2643	9.591
P10I	**0.6069**	**0.5078**	**0.5091**	*0.4055*	*0.3996*	*0.3918*	0.3981	0.2944	0.3040	0.3046	9.775
P30I	**0.5719**	**0.4691**	**0.4674**	*0.3686*	*0.3712*	*0.3691*	0.3681	0.2705	0.2766	0.2770	9.615
P100I	**0.5444**	**0.4482**	**0.4420**	*0.3531*	*0.3537*	*0.3521*	0.3462	0.2614	0.2613	0.2615	9.419
W/C100	**0.5644**	**0.4674**	**0.4599**	*0.3637*	*0.3649*	*0.3620*	0.3588	0.2659	0.2682	0.2701	9.296
WO/C100	**0.5532**	**0.4555**	**0.4581**	*0.3650*	*0.3692*	*0.3672*	0.3597	0.2719	0.2741	0.2719	9.842
W/C100I	**0.5540**	**0.4555**	**0.4510**	*0.3574*	*0.3589*	*0.3562*	0.3524	0.2631	0.2642	0.2660	9.283
WO/C100I	**0.5489**	**0.4536**	**0.4523**	*0.3580*	*0.3578*	*0.3572*	0.3520	0.2651	0.2676	0.2664	9.628
IS100	**0.5772**	**0.4775**	**0.4630**	*0.3681*	*0.3705*	*0.3665*	0.3630	0.2691	0.2707	0.2717	9.184
AS100	**0.5541**	**0.4546**	**0.4542**	*0.3627*	*0.3664*	*0.3635*	0.3559	0.2690	0.2684	0.2688	9.777
IS100I	**0.5652**	**0.4631**	**0.4557**	*0.3622*	*0.3655*	*0.3615*	0.3581	0.2664	0.2680	0.2683	9.329
AS100I	**0.5476**	**0.4502**	**0.4460**	*0.3559*	*0.3568*	*0.3549*	0.3487	0.2633	0.2628	0.2641	9.605
P100-a	**0.5498**	**0.4573**	**0.4494**	*0.3578*	*0.3606*	*0.3585*	0.3521	0.2657	0.2661	0.2654	9.655
P100-b	**0.5532**	**0.4518**	**0.4498**	*0.3593*	*0.3620*	*0.3597*	0.3528	0.2666	0.2668	0.2658	9.706
P100-c	**0.5520**	**0.4560**	**0.4474**	*0.3579*	*0.3601*	*0.3576*	0.3516	0.2650	0.2657	0.2648	9.609
P100-d	**0.5510**	**0.4551**	**0.4475**	*0.3569*	*0.3593*	*0.3570*	0.3509	0.2647	0.2650	0.2644	9.579
P100-e	**0.5509**	**0.4551**	**0.4473**	*0.3569*	*0.3590*	*0.3570*	0.3509	0.2648	0.2650	0.2643	9.587
P100-f	**0.5512**	**0.4553**	**0.4476**	*0.3573*	*0.3593*	*0.3568*	0.3511	0.2648	0.2652	0.2645	9.590
P100-g	**0.5510**	**0.4553**	**0.4473**	*0.3570*	*0.3593*	*0.3571*	0.3506	0.2648	0.2650	0.2643	9.575
P100-h	**0.5528**	**0.4568**	**0.4489**	*0.3583*	*0.3610*	*0.3585*	0.3524	0.2648	0.2659	0.2653	9.577
P100-j	**0.5522**	**0.4555**	**0.4476**	*0.3573*	*0.3596*	*0.3575*	0.3509	0.2648	0.2642	0.2646	9.508
P100-k	**0.5511**	**0.4552**	**0.4480**	*0.3572*	*0.3596*	*0.3572*	0.3512	0.2648	0.2650	0.2642	9.573
P100I-a	**0.5453**	**0.4503**	**0.4440**	*0.3539*	*0.3546*	*0.3531*	0.3472	0.2622	0.2622	0.2625	9.465

(*Continued on next page.*)

Table 7. (*Continued.*)

P100I-b	**0.5458**	**0.4483**	**0.4431**	*0.3540*	*0.3543*	*0.3528*	0.3468	0.2618	0.2619	0.2621	9.455
P100I-c	**0.5453**	**0.4487**	**0.4425**	*0.3537*	*0.3541*	*0.3524*	0.3467	0.2616	0.2617	0.2618	9.436
P100I-d	**0.5447**	**0.4483**	**0.4422**	*0.3532*	*0.3537*	*0.3521*	0.3462	0.2614	0.2614	0.2616	9.390
P100I-e	**0.5444**	**0.4482**	**0.4420**	*0.3531*	*0.3537*	*0.3521*	0.3462	0.2614	0.2613	0.2615	9.419
P100I-f	**0.5447**	**0.4483**	**0.4422**	*0.3533*	*0.3539*	*0.3521*	0.3464	0.2615	0.2614	0.2617	9.434
P100I-g	**0.5445**	**0.4482**	**0.4420**	*0.3531*	*0.3537*	*0.3521*	0.3461	0.2615	0.2613	0.2615	9.413
P100I-h	**0.5462**	**0.4497**	**0.4435**	*0.3545*	*0.3554*	*0.3536*	0.3477	0.2621	0.2622	0.2625	9.533
P100I-j	**0.5456**	**0.4486**	**0.4422**	*0.3533*	*0.3540*	*0.3525*	0.3463	0.2616	0.2608	0.2618	9.283
P100I-k	**0.5446**	**0.4485**	**0.4428**	*0.3534*	*0.3541*	*0.3523*	0.3465	0.2616	0.2614	0.2615	9.409

a, b, c, d, e, f, g, h, j and k: run-id's of submissions, which were the ten submissions with the best mean average precision, submitted by the different groups from the 47 sets submitted for the Ad Hoc IR task.
Inter: submission from the interactive IR system.
Auto: submission from the automatic IR system.
W/C: submission from the system using ⟨CONCEPT⟩s.
WO/C: submission from the system not using ⟨CONCEPT⟩s.
P, . . . , **P100I-k**: the pools in the previous subsections.
t-stat: t-statistic, (if $\alpha = 0.05$ and $\nu = 9$, $t_{0.05}(9) = 2.262$, $t_{0.10}(9) = 2.821$) shading and bold font **x.xxxx** denotes the 1st rank.
Bold font **y.yyyy** denotes the 2nd rank.
Italic font *z.zzzz* denotes the 3rd rank (when the difference between the mean average precisions was larger than 5%, the difference is important, and the rankings are different. So two or more submissions may have the same rank in each row).

Kendall correlations (Kendall's tau) between the pairs that were the mean average precisions using **F** and each of the other pools. When the significance level was 1% the correlations were over 0.71. Therefore, the correlations between the mean average precisions using **F** and each of the pools are significant.

Moreover, to examine whether there were significant differences between the set of mean average precisions using **F** and the ones using the other pools, we carried out paired t-tests. The t-statistics for the differences were computed by the following equation:

$$t\text{-}statistic = \frac{mean}{\sqrt{\frac{variance}{number\ of\ submissions}}}, \text{ where}$$

$$mean = \frac{\sum(v_{Fi} - v_{POOLi})}{number\ of\ submissions},$$

$$variance = \frac{number\ of\ submissions \times \sum(v_{Fi} - v_{POOLi})^2 - \{\sum(v_{Fi} - v_{POOLi})\}^2}{(number\ of\ submissions)(number\ of\ submissions - 1)},$$

$$v_{Fi} = mean\ average\ precision\ of\ submission\ \text{“}i\text{''}\ using\ \mathbf{F},$$

$$v_{POOLi} = mean\ average\ precision\ of\ submission\ \text{“}i\text{''}\ using\ a\ pool,$$

$$i = \text{a, b, c, d, e, f, g, h, j, and k},$$

$$number\ of\ submissions = 10.$$

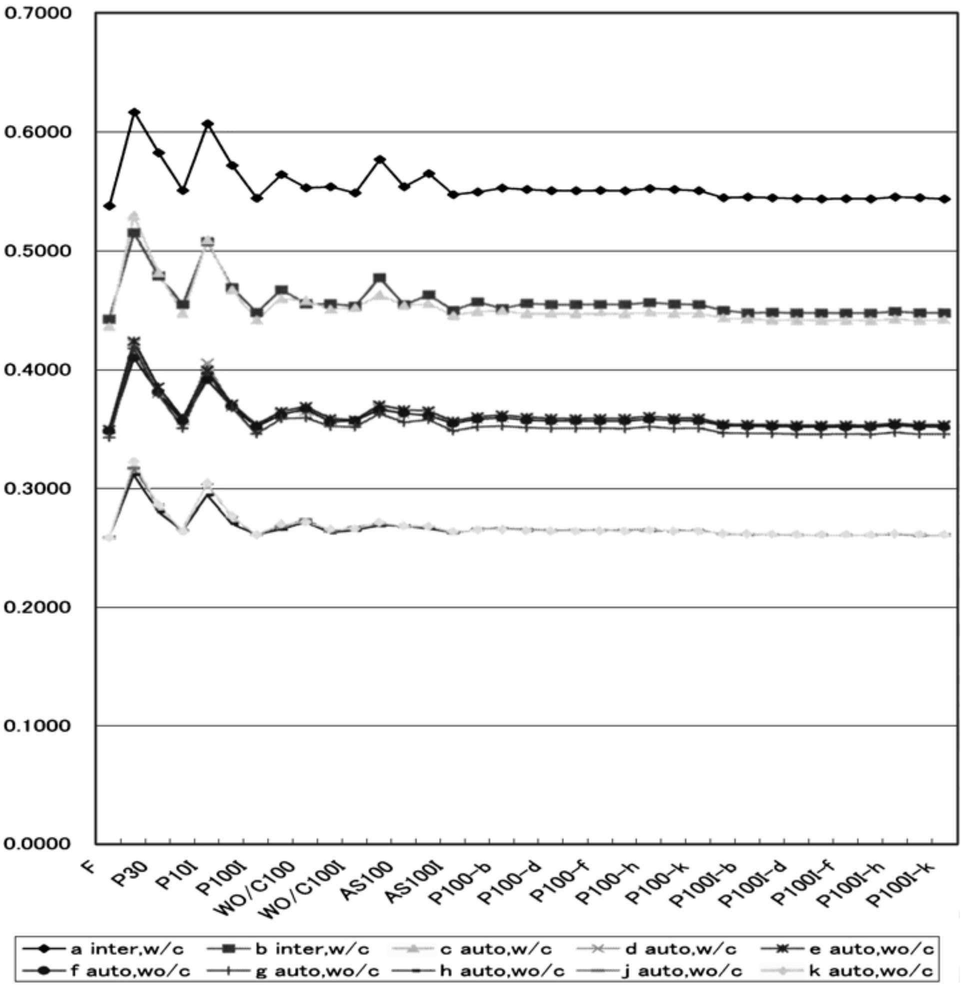


Figure 5. Graph of the mean average precisions of the ten submissions for the test: a, b, c, d, e, f, g, h, j and k: run-id's of submissions, inter: submission from the interactive IR system, auto: submission from the automatic IR system, W/C: submission from the system using ⟨CONCEPT⟩s, WO/C: submission from the system not using ⟨CONCEPT⟩s, **P**, . . . , **P100I-k**: the pools in the previous subsections.

At the 5% significance level ($\alpha = 0.05$) and the degree of freedom $\nu = 9$, the t-statistics were $t_{0.05}(9) = 2.262$, $t_{0.10}(9) = 2.821$). We can see that each of the t-statistics values in Table 7 is higher than the values of 2.262 and 2.821. Hence the difference between the set of the mean average precisions using **F** and the set using each of the other pools is significant. In addition, figure 5 shows a graph of the mean average precisions. Figure 5 shows that the tendencies of the rankings produced by the different pools are very similar.

4.2.2. Results. We can see from Table 7 that the rankings between the pools are the same for the 53 topics. We can conclude that there is no effect on the relative evaluation for the differences between the numbers of documents in the pools and the numbers of documents from submissions using any retrieval method.

We have not reported on the results from the system testing for the pre-test submissions in this paper. However, similar results were observed, and we reached the same conclusion.

5. Summary and conclusion

To investigate how to collect candidates for relevant documents efficiently and fairly, we have carried out experimental poolings and evaluations using the submissions for the pre-test and test of the first NTCIR Workshop. From these experiments, our conclusions relating to the construction of NTCIR-1 are as follows.

(1) In terms of exhaustiveness, pooling of the top 100 documents from each submission worked well for topics with less than 50 relevant documents. For topics with more than 100 relevant documents, although the top 100 pooling covered only 51.9% of the total relevant documents for the pre-test and 76.4% for the test, the coverage reached 89.7% and 98.0%, respectively, when combined with additional recall-oriented interactive searches.

(2) For the top X documents, if additional (plus X) documents were pooled from the submissions that obtained higher ranks when ranked by the mean average precision, we could collect many more relevant documents than with ordinary pooling without additional interactive searches. Moreover, if additional pooling with additional interactive searches was applied, it was possible to efficiently collect relevant documents with pools of more than 100 documents from each submission and additional interactive search.

(3) We considered relevance assessments based on the document pool created by collecting the top X documents from each submission. In both the test and pre-test, we found that the rankings of the different systems were not rotated by the use of pools with different coverages.

(4) In this paper we have not discussed any inter-assessor consistency and its effect on the system evaluation. The results of a study on that topic have been reported on briefly on various other occasions (see Kando et al. 1999a, 1999b, Kuriyama et al. 1999). Regardless of the inconsistency of the relevance assessments, we found a strong correlation between the system rankings produced using the relevance judgments by the primary assessor, the secondary assessor, and the final judge. The results follow the same direction as shown by Voorhees (1998).

We conclude that the test collection NTCIR-1 is reliable as a tool for system evaluation based on these analyses.

Notes

1. National Center for Science Information Systems (NACSIS), known since April 1, 2000, as National Institute of Informatics (NII).

2. This project is supported by the "Research for the Future" Program JSPS-RFTF96P00602 of the Japan Society for the Promotion of Science.
3. At first X was set 100, but this was adjusted to between 80 and 100 by ten, so that the total number of documents in each pool was not excessive. Though X varied with topics, the number of pooled documents for a certain topic from each submitted search result was fixed.
4. For comparison between two similar methods, 5%–7% is important. For two methods on the same system (tuning), 1%–7% may be important, depending on what is being changed (Buckley and Voorhees 1999).

References

Buckley C and Voorhees E (1999) Tutorial: Theory and practice in text retrieval system evaluation. In: Tutorial in ACM-SIGIR'99, Berkeley, CA, USA, pp. 1–109.

Cormack GV, Palmer CR and Clarke CLA (1998) Efficient construction of large test collections. In: Proceedings of the ACM-SIGIR'98, Melbourne, Australia, pp. 282–289.

Gilbert G and Sparck Jones K (1979) Statistical bases of relevance assessment for the 'Ideal' information retrieval test collection. BL R&D Report 5481, Cambridge, England.

Kageura K, Koyama T, Yoshioka M, Takasu A, Nozue T and Tsuji K (1997) NACSIS corpus project for IR and terminological research. In: Proceedings of the Natural Language Processing Pacific Rim Symposium 1997, Phuket, Thailand, pp. 493–496.

Kando N, Kuriyama K and Nozue T (1999a) NTCIR-1 (NACSIS Test Collection for Information Retrieval Systems-1): Its Policy and Practice. IPSJ SIG Notes, 99–FI–53–5:33–40. (In Japanese).

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H and Hidaka S (1999b) Overview of IR tasks at the first NTCIR workshop. In: Proceedings of the NTCIR Workshop 1, Tokyo, Japan, pp. 11–44.

Kando N and Nozue T (1999), Eds. NTCIR Workshop 1: Proceedings of the First NTCIR Workshop on Retrieval in Japanese Text Retrieval and Term Recognition, Tokyo, Japan. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/(visited March 24th, 2001).

Kuriyama K, Eguchi K, Nozue T and Kando N (1999) NACSIS test collection for information retrieval systems-1 (1): Analysis of the pooling and the relevance assessments. In: Proceedings of the IPSJ Annual Meeting, Morioka, Japan, pp. 3,105–106. (In Japanese).

NTCIR (NACSIS Test Collection for IR Systems) Project. http://research.nii.ac.jp/ntcir/ (visited March 24th, 2001).

Voorhees EM (1998) Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the ACM-SIGIR'98, Melbourne, Australia, pp. 315–332.

Voorhees EM and Harman D (2000), Eds. The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246, Maryland, U.S.A., Text REtrieval Conference (TREC). http://trec.nist.gov/(visited March 20th, 2001).

Zobel J (1998) How reliable are the results of large scale information retrieval experiments? In: Proceedings of the ACM-SIGIR'98, Melbourne, Australia, pp. 307–314.