Overview of the Fourth NTCIR Workshop

Noriko Kando
National Institute of Informatics
Tokyo 101-8430, Japan
Noriko.Kando@nii.ac.jp

Abstract

This paper outlines the fourth NTCIR Workshop, which is the latest in a series. It briefly describes the background, tasks, participants, and test collections of the workshop. The purpose of this paper is to serve as an introduction to the research described in detail in the rest of the working notes of the fourth NTCIR Workshop.

Keywords: evaluation, information access, information retrieval, text summarization, question answering, test collections, cross-lingual information retrieval, patent retrieval, Web retrieval.

1. Introduction

The *NTCIR* Workshop [1] is a series of evaluation workshops designed to enhance research in information access (IA) technologies including information retrieval (IR), cross-lingual information retrieval (CLIR), automatic text summarization, question answering, text mining and so on by providing large-scale test collections and a forum for researchers..

The aims of the *NTCIR* project are:

1. to encourage research in information access technologies by providing large-scale test collections that are reusable for experiments;
2. to provide a forum for research groups interested in cross-system comparisons and exchanging research ideas in an informal atmosphere; and
3. to investigate methodologies and metrics for evaluation of information access technologies and methods for constructing large-scale reusable test collections.

The main goal of the *NTCIR* project is to provide infrastructure for large-scale evaluations of IA technologies. The importance of such infrastructure in IA research has been widely recognized. Fundamental text processing procedures for IA, such as indexing includes language-dependent procedures. The *NTCIR* project therefore started in late 1997 with emphasis on, but not limited to, Japanese or other East Asian languages, and its series of workshops has attracted international participation.

In *NTCIR*, a workshop is held about once every one and a half years. Because we respect the interaction between participants, we consider the whole process from initial document release to the final meeting to be the "workshop". Each workshop selects several research areas called "*tasks*", or a "*challenges*" for the more challenging tasks. Each task has been organized by the researchers of the domain and a task may consist of more than one subtask.

1.1 Information Access

The term "information access" (IA) refers the whole process from when a user realizes his/her information needs, through the activity of searching for and finding relevant documents, and then utilizing information in them. We have looked at IA technologies to help users utilize the information in large-scale document collections. IR, summarization and question answering are part of a "family", aiming at the same target, although each of them has been investigated by rather different communities.

1.2 Focus of *NTCIR*

From the beginning of the project, we have looked at both traditional laboratory-type IR system testing and the evaluation of challenging technologies, as shown in Figure 1. For the former, we placed emphasis on text retrieval and CLIR with Japanese or other Asian languages and testing on various document genres.

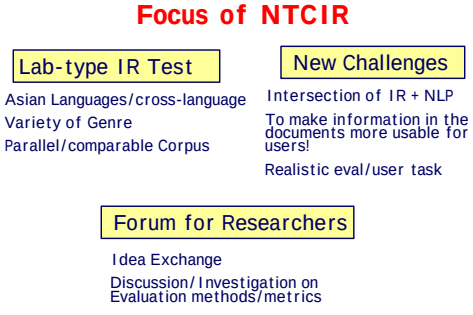


Figure 1. Focus of *NTCIR Workshops*

Table 1. Tasks of the NTCIR Workshops

	Period	Tasks	Subtasks	Test collections
1	Nov.1998-Sept.1999	Ad Hoc IR	J-JE	NTCIR-1
		CLIR	J-E	
		Term Extraction	Term Extraction/ Role Analysis	
2	June 2000-March 2001	Chinese Text Retrieval	Chinese IR: C-C	CIRB010
			CLIR: E-C	
		Japanese&English IR	Monolingual IR: J-J, E-E	NTCIR-1, -2
			CLIR: J-E, E-J, J-JE, E-JE	
		Text Summarization	Intrinsic - Extraction/Free generated	NTCIR-2Summ
			Extrinsic - IR task-based	
3	Oct. 2001-Oct. 2002	CLIR	Single Language IR:C-C,K-K,J-J	NTCIR-3CLIR
			Bilingual CLIR:x-J,x-C, x-K	
			Multilingual CLIR:x-CJE	
		Patent	Cross Genre w/ or w/o CLIR CCKE-J	NTCIR-3 PATENT
			[Optional] Alianment, RST Analysis of Claims	
		Question Answering	Subtask-1: Five Possible Answers	NTCIR-3QA
			Subtask-2: One Set of All the Answers	
			Subtask-3: Series of Questions	
		Text Summarization	Single Document Summarization	NTCIR-3 SUMM
			Multi-document Summarization	
		Web Retrieval	Survey Retrieval	NTCIR-3 WEB
			Target Retrieval	
			[Optional] Speech-Driven	
4	Apr. 2003 - June 2004	CLIR	Single Language IR:C-C,K-K,J-J	NTCIR-4CLIR
			Bilingual CLIR:x-J,x-C, x-K	
			Pivoted Bilingual CLIR	
			Multilingual CLIR:x-CKJE	
		Patent	"Invalidity Search"= Search Patents by a Patent	NTCIR-4 PATENT
			[Feasibility] Automatic Patent Map Creation	
		Question Answering	Subtask-1: Five Possible Answers	NTCIR-4 QA
			Subtask-2: One Set of All the Answers	
			Subtask-3: Series of Questions	
		Text Summarization	Multi-document Summarization	NTCIR-4 SUMM
		Web Retrieval	Informational Retrieval	NTCIR-4 WEB
			Navigational Retrieval	
			[Pilot] Geographical Information	
			[Pilot] (Search Results) Topical Classification	

n-m: n=query language, m=document language(s), J:Japanese, E:English, C:Chinese, K:Korean, x:any of CJKE

For the challenging issues, the target is to shift from document retrieval to technologies that utilize "information" in documents, and investigation of methodologies and metrics for more realistic and reliable evaluation. For the latter, we have paid attention to users' information-seeking tasks in the experiment design because they are deeply related to the appropriate types of documents, topics of the users' search requests and relevance judgment criteria. These two directions have been supported by a forum of researchers who are interested in cross-system comparison and by their discussions.

2. The Fourth NTCIR Workshop

2.1 Tasks

For the *Fourth NTCIR Workshop* (NTCIR-4) [2], the process started from April 2003 and the meeting will be held on 2-4 June 2004 [3], at National Institute of Informatics (NII) in Tokyo.

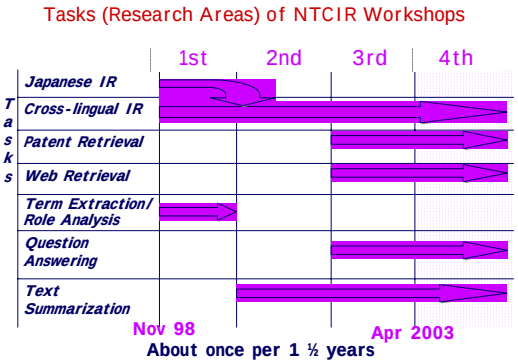
It is sponsored by the NII and Japan's MEXT Grant-in-Aid for Scientific Research on Informatics (#13224087). Question Answering Challenge's Subtask 3 was supported by NII Collaborative Research Grant Type B.

The Patent Retrieval task was organized in cooperation with the Japan Intellectual Property Association (JIPA) and NII, and the *CLIR* task was organized in cooperation with the National Taiwan University and the Korean Institute for Scientific and Technological Information (KISTI).

The *NTCIR-4* selected five areas of research as "tasks":

1. Cross-Lingual Information Retrieval Task (*CLIR*),
2. Patent Retrieval Task (*PATENT*),
3. Question Answering Challenge (*QAC*),
4. Text Summarization Challenge (*TSC*), and
5. WEB Task (*WEB*).

Since WEB was organized within somehow different management and run by its own schedule, this overview includes mainly CLIR, PATENT, QAC, and TSC.



Figirue 2. Tasks at *NTCIR Workshops*

As shown in **Table 1** and **Figure 2**, at the NTCIR-4, all of the tasks were some kind of continuation or enhancement from the previous NTCIR. Each of them increased the size of the test collections. *PATENT* proposed experiments within the different information seeking task of "invalidity search" task and challenging topic of "automatic patent map generation" as a feasibility task of a long-term research project which will last for two consecutive NTCIRs, it means for three years until NTCIR-5.

TSC included automatic evaluation of summaries and building a re-usable test collection for summarization. *CLIR* and *QAC* basically continued with minor changes in task design to remedy the major problems found in the third workshop. For CLIR, for every languages, documents were collected from multiple sources of the same publication years in somewhere in East Asia and the collection size balance between different languages was much improved by increasing the document collection size. Also TSC and QAC used the document collections collected from multiple sources.

2.2 Participants

Table 2 is a list of the active participating research groups in the *NTCIR-4*. A hundred and four groups registered, and seventy-four groups from ten different countries and areas submitted task results.

As shown in **Figures 3** and **4**, the number of participants has gradually increased. Different tasks attracted different research groups. Many international participants enrolled in *CLIR*. The *PATNET* task attracted participants from company research laboratories and "veteran" *NTCIR* participants. The *WEB* task had participants from various research communities such as machine learning and DBMS.

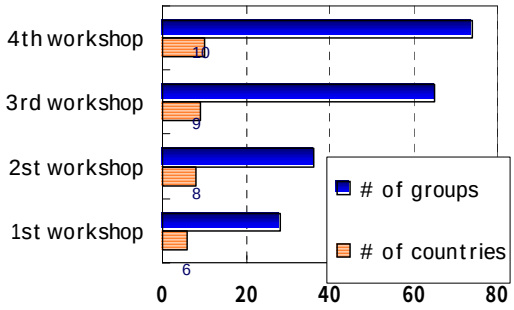


Figure 3. Number of Participating Groups

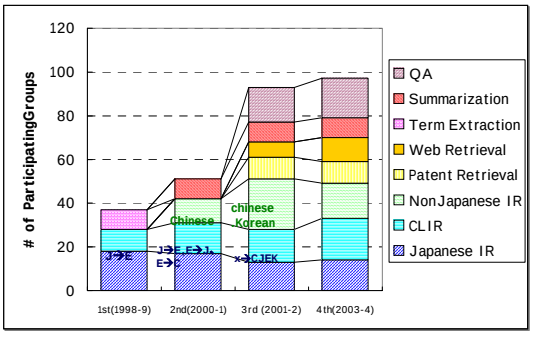


Figure 4. Number of Participating Groups, by Task

Table 2. Active Participating Groups of the Third NTCIR Workshop

[CLIR]
Chinese Academy of Sciences (China PRC)
Clairvoyance Corporation and Justsystem (USA)
Communications Research Laboratory-1 (Japan)
Fu Jen Catholic University (Taiwan ROC)
Hong Kong Polytechnic University (Hong Kong, China PRC)
Hummingbird (Canada)
Institute of Inforcomm Research (Singapore)
Korea University (Korea)
Nara Institute of Science and Technology-1(Japan)
National Institute of Informatics-1 (Japan)
National Taiwan University (Taiwan ROC)
Oki Electric-1 (Japan)
PATOLIS (Japan)
Pohang University of Science and Technology (Korea)
Queens College City Univiersity of New York (USA)
Ricoh-1 (Japan)
Royal Melbourn Intitute of Technology (Australia)
Thomson Legal and Regulatory (USA)
Tianjin University (China PRC)
Toshiba (Japan)
University of Arizona (USA)
University of California Berkeley (USA)
University of Chicago (USA)
University of Neuchatel (Switzerland)
University of Tsukuba (Japan)
Yokohama National University (Japan)

[PATENT]
Fujitsu Laboratories (Japan)
IBM Research (Japan)
Japan Patent Information Organization / Hitachi (Japan)
Nagaoka University of Technology (Japan)
NTT DATA (Japan)
Osaka Kyoiku University (Japan)
PATOLIS (Japan)
Ricoh-2 (Japan)
Tokyo Institute of Technology (Japan)
University of Tsukuba (Japan)

74 groups from 10 countries & areas

[QAC]
AIST/University of Nagoya/Univeristy of Tsukuba (Japan)
Communications Research Laboratory-1 (Japan)
Iwate Prefectural University (Japan)
Keio University (Japan)
Matsushita Electoric Industiral-1 (Japan)
Mie University (Japan)
Nagaoka University of Technology (Japan)
Nara Institute of Science and Technology-2 (Japan)
New York University (USA)/Communication Research Lobaratory-2 (Japan)
NTT Communication Science Laboratories-1 (Japan)
NTT DATA (Japan)
Oki Electric-2(Japan)
Pohang University of Science and Technology (Korea)
Ritsumeikan University (Japan)
Toshiba (Japan)
Toyohashi University of Technology-1 (Japan)
University of Tokyo-1 (Japan)
Yokohama National University (Japan)

[TSC]
Communications Research Laboratory-2 (Japan) / New York University (USA)
Graduate University for Advanced Studies (Japan)
Hokkaido University (Japan)
Pohang University of Science and Technology (Korea)
Ritsumeikan University (Japan)
Toyohashi University of Technology-1 (Japan)
University of Electro-Communications (Japan)
University of Tokyo-1 (Japan)
Yokohama National University (Japan)

[WEB]
Hokkaido University (Japan)
Ibaraki University (Japan)
Matsushita Electoric Industiral-2 (Japan)
NEC (Japan)
NII-2/Univ. of Tokyo-2/KYA Group (Japan)
NTT Communication Science Laboratories-2 (Japan)
Osaka Kyoiku University (Japan)
Tokyo Metropolitan University (Japan)
Toyohashi University of Technology-1 (Japan)
Toyohashi University of Technology-2 (Japan)
University of Tsukuba/University of Nagoya

3. Test Collections

3.1 Documents

Table 3 shows the test collections constructed through the series of *NTCIR workshops*. In the *NTCIR* the term "*test collection*" is used for any kind of data set usable for system testing and experiments. One of our interests is to prepare realistic evaluation infrastructures and efforts include scaling up the document collection and increasing variety of document genres and languages. Both patent and scientific document collections have *parallel corpora* of English and Japanese abstracts. For CLIR, we prepared the enlarged, well-balanced collections of Chinese, Korean, Japanese and English news article document collections -- the size of each language collection increased and consisted of the documents from multiple sources. The Patent document collection increased the size to 10 years.

The task (experiment) design and relevance judgment criteria were set according to the nature of the document collection and of the user community who use this type of document in their everyday life.

3.2 Topics and Questions

The structure of the topic in the IR test collections is similar to that used in TREC [5] and CLEF [6]. These topics are defined as natural language statements of "users' requests" rather than "queries", strings submitted to the system, so that both manual and automatic query construction can be done.

Table 3. Test collections constructed by *NTCIR*

NTCIR Test Collections; IR and QA

Collection	Task	Documents						Task data		
								Topic/		Relevance judge
		Genre	Filename	Lang.	Year	# of docs	Size	Lang.	#	
NTCIR-1	IR	Sci. abstract	ntc1-je	JE	1988-1997	339,483	577MB	J	83	3 grades
			ntc1-j	J		332,918	312MB			
			ntc1-e	E		187,080	218MB		60	
	TE*5		ntc1-tmrc	J		2,000	-		-	-
CIRB010	IR	News	CIRB010	C_t	1998-1999	132,173	132MB	C_tE	50	4 grades
NTCIR-2	IR	Sci. abstract	ntc2-j	J	1986-1999**	400,248	600MB	JE	49	4 grades
			ntc2-e	E		134,978	200MB			
NTCIR-3 CLIR	IR	News	KEIB010	K	1994	66,146	74MB	C_tKJE	30	4 grades
			CIRB011	C_t	1998-1999	132,173	870MB	C_tKJE	50	4 grades
			CIRB020			249,508				
			Mainichi	J		220,078				
			EIRB010	E		10,204				
			Mainichi Daily			12,723				
NTCIR-3 PATENT	IR	Patent full	kkh *3	J	1998-1999	697,262	18GB	C_tC_sKJE	31	3 grades
		Abstract	jsh *3	J	1995-1999	1,706,154	1,883MB			
		Abstract	paj *3	E	1995-1999	1,701,339	2,711MB			
NTCIR-3 QA	QA	News	Mainichi	J	1998-1999	220,078	282MB	J*	1200	exact answer
NTCIR-3 WEB	IR	Web (html/text)	NW100G-01	multipl e*4	crawled in 2001	11,038,720	100GB	J*	47	4 grades + relative
			NW10G-01			1,445,466	10GB			
NTCIR-4 CLIR	IR	News	CIRB011	Ct	1998-1999	132,173	ca.3GB	CtKJE	60	4 grades
			CIRB020			249,203				
			Hankookilbo +	K		149,921				
			Chosenilbo +			104,517				
			Mainichi	J		220,078				
			Yomiuri +			373,558				
			EIRB010	E		10,204				
			Mainichi Daily			12,723				
			Korea Times +			19,599				
			Hong Kong			96,683				
			Xinhua +			208,167				
NTCIR-4 PATENT	IR	patent full	Publication of unexamined patent application	J	1993-2002	ca. 3,500,000	ca.45GB	CtCsKJE	Main: 34, Add: 69	3 grades
		Abstract	Patent Abstracts of Japan (PAJ) +	E	1993-2002	ca. 3,500,000	ca.10GB			
NTCIR-4 QA	QA	News	Mainichi	J	1998-1999	220,078	ca.776MB	J*	197	exact answer
			Yomiuri +			373,558			199	
									251	
NTCIR-4 WEB	IR	Web (html/text)	NW100G-01	multipl e*4	crawled in 2001	11,038,720	100GB	J*		

J:Japanese, E:English, C:Chinese (C_t:Traditional Chinese, C_s: Simplified Chinese), K:Korean;
"+" indicates the document collection was newly added for NTCIR-4
* English translation is available
** gakkai subfiles: 1997-1999, kaken subfiles: 1986-1997
*3: kkh : Publication of unexamined patent application, jsh: Japanese abstract, paj: English translation of jsh
*4: almost Japanese or English (some in other languages)

NTCIR Text Summarization

Collection	Task	Documents					Summaries		
		Genre	Filename	Lang	Year	# of doc	Types	Analysts	total#
NTCIR-2 SUMM	Single doc	News	Mainichi	J	1994.1995.1998	180 doc	7	3	3780
NTCIR-2 TAO	Single doc	News	Mainichi	J	1998	1000 doc	2	1	2000
NTCIR-3 SUMM	Single doc	News	Mainichi	J	1998-1999	60 docs	7	3	1260
	Multi doc		Mainichi	J		50 sets	2	3	300
NTCIR-4 SUMM	Multi doc	News	Mainichi	J	1998-1999	30 sets	2	1	60*
			Yomiuri						

Table 4. Topic Fields in NTCIR Test Collections

Topic Structure of NTCIR IR Test Collections

	NTCIR-1	NTCIR-2	CIRB010	NTCIR-3 CLIR	NTCIR-3 PATENT	NTCIR-3 WEB	NTCIR-4 CLIR	NTCIR-4 PATENT
Task	ad hoc, CLIR	ad hoc, CLIR	ad hoc, CLIR	CLIR	Cross-genre, CLIR	ad hoc	CLIR	invalidity
Mandatory Run *	D-only	D-only	N/A	D-only	S+A	T-only, D-only	T-only, D-only	CLAIM-only
Topic Field								
TITLE **	very short	very short	very short	very short	very short	query	query	very short
DESC	yes	yes	yes	yes	yes	yes	yes	yes
NARR (unstructured	yes	yes	yes	yes	yes			yes
NARR (structured)						yes	yes	
NARR. BACK *10						yes	yes	
NARR. RELE *10						yes	yes	
NARR. TERM *10						yes	yes	
PURPOSE *7								yes
CONC	yes	yes	yes	yes	yes	yes	yes	
FIELDS	yes	yes						
TLANG / LANG *3				yes			yes	
SLANG *3				yes			yes	
RDOC *4						yes		
PI *4					yes			
USER *5						yes		
ARTICLE *6					yes			
DOC *9								yes
SUPPLEMENT *6					yes			
CLAIM *8								yes
COMP *8								yes
COMP. CNUM *8								yes

*: D-only=DESC only, T-only=TITLE only, A+S= run using ARTICLE and SUPPLEMENT only

**: "very short"=very short description of search request; "qeury"=comma separated term list

*3: TLANG/LANG=target language, the language of the topic; SLANG=source language, the language the topic originally constructed.

*4: RDOC=known relevant documents; PI=the patent for the invention mentioned in the news articles.

*5: USER=users' attribute

*6: ARTICLE=a news article reporting an invention; SUPPLEMENT=memorandam to focus the issues in the article relevant to the user's needs; if a human knowledgeable searcher reads ARTICLE and SUPPLEMENT, he/she understand the user's search request as specif

*7: Purpose of search (only "invalidity search" for NTCIR-4 PATENT)

*8: CLAIM=Target claim in the query patent. It was used as qeury of the search and may consists of multiple components; COMP=Component of a claim; CNUM=Claim component ID

*9: Query patent fulltext (fulltext of a patent that is used as a query of the search)

*10: BACK=Background knowledge/purpose of search; RELE=relevance judgment criteria; TERM=term de

In NTCIR, *Mandatory Runs* are defined for each IR-related task, and every participant must submit at least one mandatory run using the specified topic field only. The purpose of this is to enhance cross-system comparisons by basing them on common conditions, and to judge the effectiveness of the additional information. Mandatory runs were originally designated "<DESC> only" because <DESC> is the basic description of the users' search requests, but from NTCIR-4, CLIR was designated both "<TITLE> only" and "<DESC> only". It was partially because short queries like <TITLE> only runs are more realistic and partially because that to test the effectiveness of the disambiguation mechanisms, which is one of the critical components in CLIR, shorter queries is more preferable. Any combination of topic fields may be used in experiments for research purposes.

As shown in Table 4, emphasis has been shifted towards the topic structure to allow more realistic experiments and to gauge the effect of background information on the topic. For example, the narrative <NARR>, longer natural language explanation for each topic, can be structured using tags indicating subfields in <NARR>, such as <BACK> for "Background/Purpose of Search", <RELE> as "Relevance Judgment Criteria", or <TERM> for "Term Definition". Most *NTCIR* collections contain a list of concepts <CONC>, but they are not heavily used by participants. The topics in the *PATENT* collections are various according to the information seeking tasks each of the tasks set upped.

For TSC, both the documents themselves and the topics of each of the document sets were given to the participants. These topics are very simple expression typically a few terms, but this can be seen as users' initial search requests and the set of documents were produced as retrieval results for the requests.

3.3 Relevance Judgments and Evaluation

In IR-related tasks, relevance judgments were graded using a scale similar to previous *NTCIR* workshops: highly relevant, relevant, partially relevant and irrelevant. For the *Patent Retrieval* task, professional patent intermediaries conducted judgments on the pooled documents consisting of the documents listed in the higher ranks in each submitted run, together with intensive interactive search and judgments using several commercial patent retrieval systems and the system provided by the task organizers. Such integration of the two different strategies to collect relevant documents was found to improve the completeness of the relevance judgments for a large-scale document collection with longer documents.

For the *QAC*, exact answers were used for evaluations. They were prepared before the runs by assessors, then also all the submitted answers were reviewed and revised answer sets were released after. For the evaluation, the mean reciprocal rank (MRR) is used for subtask 1, in which the participating systems were requested to return five possible answers with no penalty for wrong answers, and the modified mean F-measure is used for subtask 2, in which the participating systems returned one set of all the answers with penalties given for wrong answers. For subtask 3, a series of questions are used for either of in the user's information seeking tasks of "information gathering" in which a user supposes to raise a series of questions on a particular topic, and "browsing", in which users questions are keep drifting through the interaction with systems.

For *Text Summarization*, two types of summaries, short and long, were produced by analysts as gold standards, and then each sentence in those summaries are related to the sentences in the source documents to be summarized. The analysts asked to check all the possible relations between sentences in the human created summaries and the source documents. Using these greedily annotated human produced summaries, the effectiveness of system produced summaries can be automatically evaluated as an extract in the aspects of "number of sentences should be extracted", "precision", and "coverage" as the intrinsic evaluation of extraction. For intrinsic evaluation for abstract, content and readability were tested using a set of quality questions. For extrinsic evaluation of abstracts, system produced summaries were evaluated by question answering.

5. Discussion

A brief overview of the *fourth NTCIR Workshop* is reported here. The details of the achievements from each task and those of each participant are reported in the reports from each task in this issue, the papers in this volume [4].

To enhance the research in the fourth workshop, special attentions were paid (1) to provide longer time period for experiments, and (2) to enhance the document collections. It was because that, in the *NTCIR-3*, lots had to be done by the participants for the new tasks and new task components. As results, participants could only implement some of their research ideas, but generally such new task components had not been fully investigated nor analyzed because of tight schedule of the workshop.

Moreover, in the *NTCIR-3*, for some of the new components like "passage-level relevance judgments" *PATENT*, *QAC* and *WEB* tasks and "Search Results Classification" at the *WEB* task, none of the participants fully accomplished. Then,

6.

in the *NTCIR-4,* in order to obtain sufficient time to think of the task and original//unique idea for the experiment strategies, we released the document collection as early as possible and omitted the dry runs for the tasks in which the *NTCIR-3 collection* were usable for training. For much more challenging issue, we set the "*feasibility study*" subtask, in which the investigation is performed through the two consecutive workshops, i.e. for three years. In such ways, we expected that each participant could spend sufficient time for experimentation and implementation.

Evaluation must adapt to technological evolution and the change in social needs. We are working towards this goal, and suggestions are always welcome.

References

1. NTCIR Project: http://research.nii.ac.jp/ntcir/
2. NTCIR Workshop 4 (2003-2004) : http://research.nii.ac.jp/ntcir-ws4/work-en.html
3. NTCIR Workshop 4 Meeting (2-4 June 2004) : http://research.nii.ac.jp/ntcir/ntcir-ws4/
4. Kando, N. and Ishikawa, H. (eds): NTCIR Workshop 4: Working Notes of the Fourth NTCIR Workshop Meeting on Evaluation of Information Retrieval, Question Answering and Summarization, Tokyo Japan, June 2-4 2004., NII, Tokyo (2004) (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/)
5. TREC: http://trec.nist.gov/
6. CLEF: http://clef.iei.pi.cnr.it/