Overview of the VisEx task at NTCIR-9

Tsuneaki Kato
The University of Tokyo
3-8-1 Komaba Meguro-ku
Tokyo, Japan
kato@boz.c.u-tokyo.ac.jp

Mitsunori Matsushita
Kansai University
2-1-1 Ryozenji, Takatsuki
Osaka, Japan
mat@res.kutc.kansai-u.ac.jp

Hideo Joho
University of Tsukuba
1-2 Kasuga, Tsukuba
Ibaraki, Japan
hideo@slis.tsukuba.ac.jp

ABSTRACT

Interactive Visual Exploration (VisEx) is a pilot task at NTCIR-9 for establishing an efficient and effective framework for objectively evaluating interactive and explorative information access environments. It aims to acquire more useful and richer evaluation data based on empirical user studies, by adopting a common framework for the environments and conducting sophisticated experiments. Four teams participated in this task. Although it was harder to understand the results and draw a clear conclusion than expected, we learned much and have made useful progress.

Categories and Subject Descriptors

H.3.0 [**Information Strage and Retrieval**]: General

General Terms

Experimentation

Keywords

evaluation, empirical user studies, interactive information access

1. INTRODUCTION

Interactive Visual Exploration (VisEx) is a pilot task at NTCIR-9 for establishing a framework for evaluating interactive and explorative information access environments. It evaluates environments in which users interactively refine or elaborate their information needs, and through various activities, accumulate proper information. It is important in such environments to employ information visualization techniques for showing access results and to allow interactions with visualized information. It is also crucial to allow reformulation of queries and relevance feedback. The purpose of VisEx is to evaluate such comprehensive information access environments.

VisEx postulates a common framework for explorative information access environment systems (IAESs). The participants submit a core of an IAES, which works in the common framework. IAESs with submitted cores are evaluated through laboratory experiments with human subjects, who are requested to perform experimental tasks in given environments. The use of a common framework of IAESs and evaluation of the IAESs through common tasks are expected to help eliminate factors that may affect evaluation measures and thus produce more useful and richer data.

The tasks tackled by the subjects in the experiments are to compile a report on a given topic by collecting relevant events or facts using a given IAES. Data obtained through the experiments include reports, which the subjects make as direct products of the tasks, log records of the information access behaviors of the subjects, and subjective evaluations collected by questionnaire surveys of the subjects. A synthetic analysis of the data is expected to yield new insights on temporal aspects of information access behaviors, and relationships between objective measures and subjective impressions.

This report describes the design details of the VisEx task, a summary of the submitted systems, and experimental results and their analysis. In section 2, the background and the policy of VisEx are discussed. In section 3, the experiment design is explained in detail. In section 4, the submitted systems are briefly introduced. Then, in section 5, the results of the experiments and their analysis are reported. Finally, in section 6 some conclusions are drawn and future work is discussed.

2. BACKGROUND AND POLICY

It is difficult to evaluate interactive and explorative IAESs, because interactive and explorative information access is a complicated human activity. Such activities do not proceed in a straightforward manner, as the user trying to access information frequently performs trial and error and changes her mind during the process. The actions are varied, including browsing a list of selected documents and inspecting documents, and many of the actions are creative and intelligent, involving analysis, understanding, aggregation, and integration of information. This is especially true when IAESs employ interactive information visualization, when both the information provided by the environment and actions taken by the user become more diverse and complex.

Due to such difficulties, an evaluation of IAESs must take one of two approaches. One is empirical user studies in which subjects are requested to accomplish a given task in a controlled situation, and through observing the process and quantifying the degree of achievement, IAESs are evaluated as a whole. Provided the task is adequately designed, this approach is very helpful to obtain data in a real-world situation, but it takes significant time and resources, especially when comparing different systems. The TREC interactive tracks [1] are representative of this approach. The other is benchmark tests, in which components of IAESs and their specific functions are evaluated separately. It is relatively cost effective, but to be convincing, it is necessary to show

that the results properly reflect the system's utility or quality in real settings.

The final objective of the VisEx task is to bridge these two approaches to evaluation, and to establish an efficient and effective methodology for objectively evaluating interactive and explorative IAESs. The current attempt in NTCIR-9, however, limits itself to a sophisticated evaluation based on empirical user studies, as the first step to the final objective. The results of the experiments in this term are expected to reveal some relationships between holistic evaluation of IAESs as a whole and the benchmark evaluation of their components.

With this background, the design of the VisEx task considers the following points. First, all activities of interactive and explorative information access should be observed in the experimental study. Second, the task should be able to elicit explorative behaviors from users. Third, factors not relevant to the evaluation should be excluded as much as possible. And finally, not only the behavior of the interactive and explorative information access as a whole but also their component actions should be observed. The former two are important for the empirical study to evaluate properly the IAESs as a whole, while the latter two are needed for the results to reveal a relationship with the benchmark evaluation of their components.

3. TASK DESIGN

Figure 1 shows the framework of the evaluation in VisEx. IAESs to be evaluated are assumed to have the architecture shown in the center of the figure. The participants submit a core of an IAES, while other modules are provided by the organizers and shared by all participants. The organizers also provide a baseline system as one of the IAES cores for comparison. IAESs with submitted cores are evaluated through laboratory experiments with human subjects.

The IAES architecture specifies the backend information retrieval (IR) engine and the editor for compiling and recording collected information as the shared parts. It also specifies that all modules work under a given web browser, that is, users conduct everything through a browser interface, which includes information access and editing reports. The IAES core exploits the IR engine and other function modules such as those for displaying documents and constructing snippets through a defined interface protocol. Its role is to obtain the user's information needs and send them to the IR engine and to display the results in a form that the user can understand and pursue the task easily.

We provided this specific architecture to the IAES because we wish to look at information access activities as a whole, that is, not only collecting proper information but also compiling it into knowledge. This specification allows us to prevent differences in IR engines and editors affecting the evaluation even when looking at such a broad range of phenomena. In addition, the architecture allows us to obtain uniform and richer data on users' information access activities. Since protocols between the modules are specified, we are able to log requests and responses conveyed between the modules uniformly, and to record the behaviors of users and the IR engine.

The laboratory experiments are designed with reference to those of the TREC interactive tracks. The task asks the subjects to collect information on a given topic and compile it into a report using a given IAES. The subtasks used in the experiments with their topics are shown below.[1] These subtasks request the subjects to collect as many nuggets as possible in a given time period, which are fundamental units of information that constitute a requested report. It is also important that the users achieve this activity comfortably with little stress. These characteristics are similar to the tasks adopted in the TREC interactive tracks, which request subjects to achieve high aspect/instance recall in a given time period.

In contrast to usual evaluation workshop tasks, the possible topics of the subtasks, ten topics for each, were informed to the participants before the experiments, so the participants could refer to these topics when designing and constructing IAES cores. Those are new for the subjects of the experiments, which must be enough for proper evaluation.

Event Collection Subtask This requests the subjects to make a report on events specified as a topic by collecting the characteristics of the events such as times and places of occurrence. This subtask is an interactive version of complex question answering, especially event-list questions, and the topics were selected from the NTCIR-7 AQLIA test set [5].

- **E0** Please tell me about cases where dinosaur fossils have been excavated in Africa. I would like to know when and where they were dug up, the teams who did it, the type of fossils, when the dinosaurs were alive, and the value of the excavation.
- **E1** Please tell me about airplane crashes that have happened in Asia. I would like to know where and when they occurred, the circumstances, the type of airplane including the name of the airline company, and number of casualties.
- **E2** Please tell me about incidents that have happened at Japanese nuclear power plants. I would like to know where and when they occurred, an outline of the incidents, and the damage caused.
- **E3** Please tell me about nuclear tests that have been conducted in different countries around the world. I would like to know when and where they were conducted, a summary of the tests, which country planned them, and their purpose.
- **E4** Please tell me about incidents that NATO has recognized as being cases of friendly fire. I would like to know where and when they happened, the facilities fired at, and the number of casualties.

Trend Summarization Subtask This requests the subjects to make a report summarizing trends related to time-series statistical information given as a topic by collecting not only the values and changes of the statistics but also the reasons for the changes and their influences. This subtask is similar to the MuST workshop, which deals with several aspects of summarizing trend information, and constructed research resources for this theme [3]. The topics were selected from the corpus constructed in the workshop.

[1] We call these *subtasks* to distinguish them from *tasks* in the NTCIR workshop, though the subtasks are what the subjects are asked to carry out.

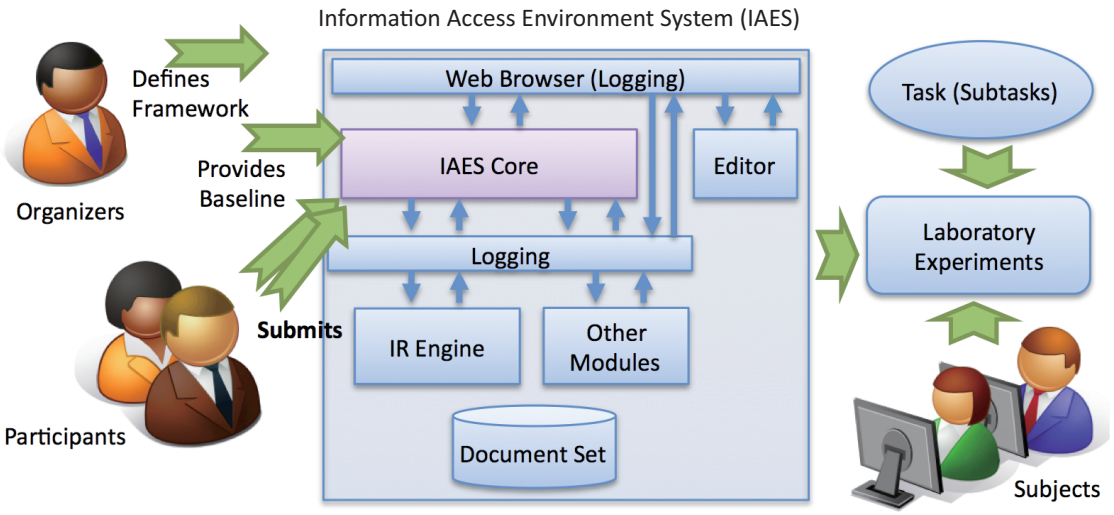

Figure 1: The Framework of VisEx

T0 Please examine the diffusion of communication devices. I would like to know the changes in the numbers of subscribers of land phones, cellular phones, and PHS.

T1 Please examine the situation about gasoline. I would like to know the changes in price of Dubai crude oil and regular gasoline.

T2 Please examine the evaluation of the Cabinet. I would like to know the changes in the approval and disapproval rating for the Cabinet.

T3 Please examine the employment situation. I would like to know the changes in the number of unemployed and the unemployment rate.

T4 Please examine the demographic composition. I would like to know the changes in the elderly population, young population, and birth rate.

The document set used in both subtasks was Mainichi newspapers in Japanese from 1998 to 2001.[2] Many events and situations are described in more than one article, and we expect the subject to gather those pieces of information and compile them in order to make a report. Such aggregation is a key feature of explorative information access. Using articles from ten years ago, the subjects may find unexpected facts and need new interpretation of information. Such new findings, in turn, may lead the subjects to change their behavior, which is another key feature of explorative information access. The task setting of VisEx is expected to derive those features of interactive and explorative information access activities.

Data obtained through the experiments include reports, which the subjects compile as direct products of the tasks, log records of the information access behaviors of the subjects, and subjective evaluations collected by questionnaire surveys of the subjects. A synthetic analysis of the data is expected to yield new insights on temporal aspects of information access behaviors, and relationships between objective measures and subjective impressions, and so on.

[2]The framework of VisEx is language independent. In fact, we conducted preliminary experiments of the event collecting subtask using Xinhua newspaper articles as the document set.

3.1 Modules of IAES

We use the Apache Solr full-text search server [6][3] for the backend IR engine of the IAESs and the Firefox browser [4] for the web browser. The Solr full-text search server also provides other function modules such as for displaying documents and constructing snippets. We adopted a bigram tokenizer, CJKTokenizerFactoy provided by Apache Solr, for indexing documents in Japanese.

The editor was developed by ourselves as an add-on of the Firefox browser, which we named Quick-Edit. Quick-Edit is not just an html document editor working under a tab of the Firefox browser, but also a logger of users' actions in the browser and the editor itself. In addition to editing documents in rich text format, the editing features allow users to copy any part of html documents including text and figures from any tab, and to paste it keeping its style. The logging starts automatically upon starting up the browser, and log records are written to a file with a name containing the start-up time. The following actions are logged. Each action data constitutes one line of the log record with the time of its occurrence and proper parameters.

- Starting up and quitting the browser
- Opening, closing and selecting a window
- Maximizing, minimizing and resizing a window
- Entering characters into the location bar and query bar
- Opening, closing and selecting a tab
- Going forward and backward
- Changing a URL
- Scrolling a page

[3]http://lucene.apache.org/solr/
[4]http://mozilla.jp/firefox/

- Mouse clicking
- Starting and stopping mouse dragging
- Cutting, copying and pasting an element or part of an html document
- Inputting and deleting a character
- Opening, closing and saving a file in the editor

Figure 2 is an example of a log record. It shows the log of a sequence of actions of document retrieval, going to a document page by clicking somewhere on the result list page, copying a part of the document, moving to the editor, pasting the copied text into the editor, moving back to the document page, and going back to the result list page by pressing the back button.

3.2 Baseline system

In order to provide some reference with which to compare the submitted systems, a baseline system with basic retrieval features was developed, which is similar to ordinary web search engines. This system was also used in training sessions for the experiments to help the subjects to understand the task. This system allows users to retrieve documents by entering keywords. Boolean expressions of keywords and phrases are also allowed. The retrieved articles are shown as a list of their headlines and snippets generated by Apache Solr. The articles in the list can be arranged in ascribing or describing order of their relevancy score or publication date. By clicking a headline, the user can examine the main text of the article. Figure 3 shows a snapshot of the system used for the leaflet which was produced to explain the system.

3.3 Experiment design

Submitted systems may choose one or two subtasks. One unit of the experiment is a combination of a system and a subtask. Five subjects carry out the following procedure in each unit.

- Receiving an explanation of the experiment and signing a consent form
- Filling out a pre-experiment questionnaire
- Receiving an explanation of the subtask and carrying out a training session on a sample topic (E0/T0) using the baseline system (20 min)
- Receiving an explanation of the system to be used and carrying out a training session of a sample topic using the system (20 min)
- Filling out a post-session questionnaire
- For four topics, carrying out the main session (50 min each) and filling out a post-session questionnaire
- Filling out a post-experiment questionnaire

Each participant attends just one unit. A sample topic and four topics of each subtask are shown in section 3. The order of topics is the same for each subject.

The pre-experiment questionnaire asks the subjects about their experience of web search and report writing, among other matters. The post-session questionnaire asks about their familiarity with a given topic, degree of satisfaction with the results of the search activity, and how difficult the search of the session was. The post-experiment questionnaire asks about the impression of the system used and requests the subjects to evaluate it.

4. SYSTEMS SUBMITTED

Four teams participated and each team submitted one system, that is, one IAES core. Two of the teams participated in both subtasks. In addition to two units of the baseline system, six units of the experiments were conducted. Table 1 shows the teams, systems submitted and subtasks they tackled. E and T stand for the event collection and trend summarization subtask, respectively. The KN system [8] uses charts of statistical information as an interface, and allows users to retrieve relevant information by clicking a point in the shown chart. The Grid system [2] takes two sequences of keywords from users and places relevant documents in a two-dimensional matrix of keywords. The TM2011 system [7] allows the user to mark documents as having been read, and to distinguish those from newly retrieved ones. The UTLIS system [4] added two features to the baseline system: narrowing down the results by specifying a place name that the articles are related to or their date of publication; and retrieval of articles similar to a specified article. For the details of these systems, please refer to the report of each participant.

5. DATA ANALYSIS

As mentioned above, data obtained through the experiments consist of reports made by the subjects, log records, and data collected using questionnaire surveys of the subjects. Each shows a different aspect of information access activities: the reports represent the main products; the log records represent the dynamic processes; and the questionnaires represent the subjective impressions of the subjects involved. By analyzing the data individually and synthetically, it is expected to be possible to comprehensively grasp the complex activities of interactive information access, and to understand the roles of IAESs in those activities.

In fact, however, it was more difficult than expected to deduce meaningful information from the obtained data. Possible reasons for this difficulty are as follows. Since there was great diversity of behavior among the subjects and also too few subjects, the characteristics of IAESs were buried in such diversity. Since we have no clear criterion for the ideal products and processes of information access activities for the tasks, it is difficult to quantify their suitability. In addition, the mechanisms for capturing log records were not sophisticated enough.

Due to such difficulties, the results of only a fundamental analysis are discussed in this report. Although the analysis is preliminary and initial, showing no meaningful difference, it may provide a foundation for more useful analyses in future.

5.1 Report analysis

The products of the tasks are reports, which are expected to show the results of the information access activities. We had planned to discuss the quality and quantity of the reports by decomposing the contents into nuggets, and to measure the precision and recall, but our plan was found to have the following problems.

When the user finds a nugget and judges it to be relevant and important, it will be included in the user's report. However, this judgment differs among subjects. For example, an examination of the log records showed that one subject evidently found and read an article, but did not include a nugget in that article in her report, even though it looked relevant to us. Furthermore, interpretation of a given topic

```
14:55:35  browser  input         HTML:INPUT, 恐竜 化石      % input to the query box
14:55:36  browser  click         0,0,0,HTML:INPUT      % execute retrieval by clicking button
14:55:36  browser  url-change    http://xxx.ac.jp/baseline/index?query=%E6%81%90%E7%AB%9C+...
                                     % display the result list page
14:55:36  browser  window-focus  14:55:03, 恐竜 化石の検索結果
14:55:39  browser  click         0,498,502,HTML:B      % select a document by clicking
14:55:39  browser  url-change    http://xxx.ac.jp/baseline/examine?doc=JA-980318315
                                     % move to the document page
14:55:39  browser  window-focus  14:55:03,JA-980318315
14:55:42  browser  dragstart2    0,994,205,HTML:H3      % start dragging
14:55:44  browser  dragend2      0,1126,201,HTML:H3      % end draggind
14:55:45  browser  copy          HTML:#text,JA-980318315      % copy a region
14:55:46  browser  url-change    quick-edit:///      % move to the editor
14:55:46  browser  tab-select    1,Quick Edit
14:55:46  browser  window-focus  14:55:03,Quick Edit
14:55:47  browser  click         0,638,123,xul:tabbrowser#content      % set cursor by clicking
14:55:48  editor   paste         JA-980318315      % paste
14:55:49  browser  url-change    http://xxx.ac.jp/baseline/examine?doc=JA-980318315
                                     % move back to the document page
14:55:49  browser  tab-select    0,JA-980318315
14:55:49  browser  window-focus  14:55:03,JA-980318315
14:55:49  browser  click         0,376,123,xul:tabbrowser#content
14:55:51  browser  back          % go backward
14:55:51  browser  click         0,270,67,xul:toolbarbutton#back-button      % push the backward button
14:55:51  browser  url-change    http://xxx.ac.jp/baseline/index?query=%E6%81%90%E7%AB%9C+...
                                     % display the result list page
14:55:51  browser  window-focus  14:55:03,JA-980318315
14:56:05  browser  window-focus  14:55:03, 恐竜 化石の検索結果
```

Figure 2: An Example of Log Records

Table 1: Participants and Subtasks Tackled

Team ID	Organization	Submitted System	E	T
JLTKB	University of Tsukuba	the Grid system	◯	◯
KUTC	Kansai University	the KN system		◯
TOTLA	Tokyo Metropolitan University	the TM2011 system	◯	
UTLIS	The University of Tokyo	the UTLIS system	◯	◯

depends on the subject. While some interpretations were apparently wrong, the cause of such misinterpretation was not the IAES that was used. As a consequence, the precision and recall of the nuggets in reports reflect merely a combination of the efficiency of the IAES and judgment and interpretation of the subjects; it is difficult to separate those factors. This problem was already pointed out in earlier studies, and was confirmed to occur in the task we used in our experiments in which we had believed that such variations of judgment were relatively small.

The approach to the task taken by the subject also influences the reports. For example, in the event collection subtask, one subject may try to obtain many piece of information on one event, while another subject may try to collect many events. These two approaches have different difficulty, and either approach may lead to a situation in which the subject tries in vain to find information that does not exist. Such failed activities could not be identified from the reports.[5]

Since the initial analysis revealed such problems and suggested that calculating nugget-based precision and recall based on relevance judgment might not be fruitful, we decided to start by conducting easier analysis using articles as the unit, and to grasp the outline of the contents of the reports compiled by the subjects. It is possible to automatically extract the articles referred to in each report as we instructed the subjects to write the article's ID when including the contents of an article in their report. We did not check the relevancy of the articles and their contents but only paid attention to the existence of references to those articles in the reports.

Table 2 shows the minimum, maximum, average, and median of the numbers of articles referred to in the reports for each system and topic. There is no clear difference in the quantities.

In order to show the characteristics of the submitted systems, Table 3 summarizes the tendency of the articles retrieved by those systems. A system is considered to have retrieved an article when at least one subject who used that system referred to that article in her report. For each topic, the numbers of types of all articles retrieved and articles retrieved by all four systems, three systems, and one system are shown. For articles retrieved by three systems and one system, the numbers of articles that each system retrieved are appended. The Grid system in the event collection subtask and the KN system in the trend summarization subtask frequently missed articles that the other three systems retrieved, showing a slightly different tendency from the oth-

[5]Such failed activities might be detected by observing the log records, but we could not conduct such analysis this time.

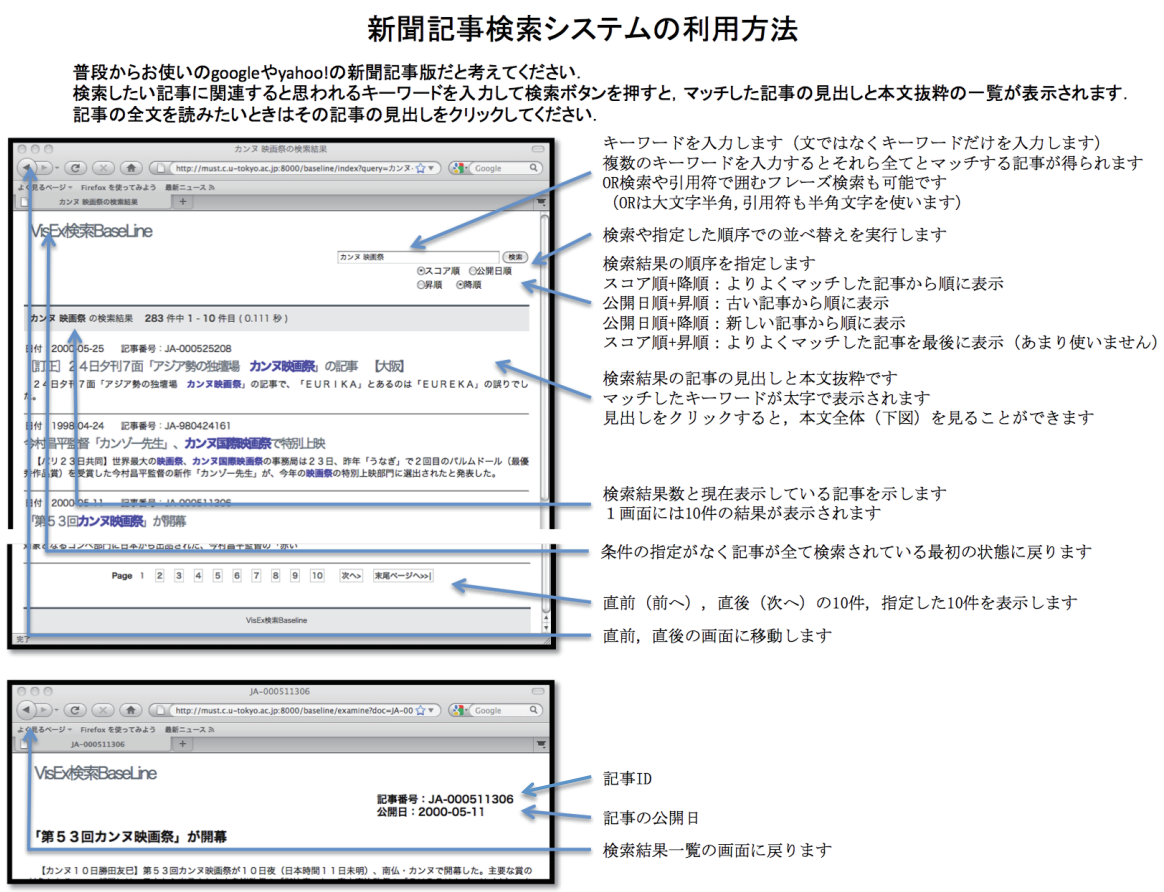

Figure 3: Baseline System

ers.

This table also shows some characteristics of the task. In most of the topics, more than half of the articles were retrieved by only one system. Articles retrieved by all systems account for only 20% at most, and sometimes less than 10%. From the perspective of the task design, this suggests there were too many relevant documents. In practice, different users would surely retrieve different articles using different systems, and the results would contain many different relevant articles.

Table 4 shows another measure of the variance of articles retrieved by the systems: the cosine measure between characteristic vectors of systems' retrieval for a given topic. Each dimension of the vector corresponds to an article, and its value is the number of subjects who used a given system and retrieved the corresponding article. For each topic, the cosine measures between the systems and the average of all systems and those between the systems and the baseline system are shown. The tendency is the same as shown in Table 3. In addition, there are slightly smaller variations among the systems in topic E4 and T4 compared to the other topics.

For the event collection subtask, as each event could be identified by a human assessor, we conducted a similar analysis based on events reported instead of on articles referred to. The results are shown in Tables 5, 6, and 7, which correspond to Tables 2, 3, and 4, respectively.

5.2 Log analysis

As mentioned above, the subjects' actions in the browser including those on the editing tab could be logged in detail and precisely. Some problems, however, became apparent. The action units of the log can be too small for some actions. For example, one character input or deletion is recorded as one action and each constitutes one line of the log record. Screen scrolls with which we can intuitively see one grouped action are recorded as a long sequence of actions. Although we could summarize and interpret those log records by writing a computer program, it would not be easy since we would need to discuss and agree the specifications of such summarization and interpretation. On the other hand, enough information could not be logged by the current mechanism in cases where the system conducted asynchronized communication and dynamic redrawing of pages using Ajax. There is another reason why it is difficult to understand log records: the log record does not show the semantics of actions explicitly. For example, a mouse click by the user is recorded as the time and coordinates of the window where clicked. It is not explicit only from the log record whether that click actually selected an article or moved to the next page.

Since such semantics can be understood by the designer of the system, the designer could conduct a detailed analysis that takes the semantics into account. An example of such analysis is shown in a report of the UTLIS system, which

Table 2: The Number of Articles Referred to in Reports

		1				2				3				4			
		min	max	avg	med	min	max	avg	med	min	max	avg	med	min	max	avg	med
E	Baseline	4	12	8	8	3	10	6.4	6	9	14	11	11	9	18	12.6	12
	Grid	5	11	7.2	6	4	13	8.2	8	6	15	8.6	8	7	14	11.2	12
	TM2011	5	15	9.8	10	4	12	8.4	9	8	18	12.4	13	11	18	14	14
	UTLIS	4	12	9.2	10	7	12	8.8	8	6	11	8.6	9	8	13	9.8	9
T	Baseline	4	8	6.2	6	5	12	9	9	5	14	8.8	7	5	9	7.6	8
	Grid	4	8	6.2	6	3	16	8.2	7	4	13	8.2	8	3	13	9.2	12
	KN	4	10	7.8	10	2	23	10.6	10	8	22	12.4	9	9	13	10.6	10
	UTLIS	5	10	7	6	4	9	7.4	9	6	12	8.8	8	6	12	8.2	8

Table 5: The Number of Events Mentioned in Reports

		1				2				3				4			
		min	max	avg	med	min	max	avg	med	min	max	avg	med	min	max	avg	med
E	Baseline	4	10	6.4	5	2	9	4.6	4	6	16	10	8	6	11	9.2	10
	Grid	4	8	5.4	5	5	11	7	7	6	15	8.8	7	7	11	8.4	7
	TM2011	4	7	5	4	3	9	5.2	4	4	17	10.2	12	7	13	10	11
	UTLIS	2	7	4.6	4	3	9	6	6	4	10	7	6	5	11	8.8	9

Table 3: The Number of Artciles Systems Retreived

		total	4 systems	3 systems	1 system
E	1	75	7	8 [5, 7, 6, 6]	46 [13, 7, 16, 10]
	2	108	3	5 [5, 2, 4, 4]	85 [12, 28, 23, 22]
	3	75	10	11 [9, 6, 10, 8]	41 [9, 10, 13, 9]
	4	62	11	12 [9, 5, 12,10]	28 [5, 5, 5, 13]
T	1	52	6	3 [3, 3, 0, 3]	34 [6, 1, 19, 8]
	2	67	3	14 [12, 10, 8, 12]	33 [6, 8, 16, 3]
	3	93	2	14 [9, 9, 12, 12]	56 [14, 10, 21, 11]
	4	51	12	6 [3, 4, 6, 5]	27 [8, 8, 8, 3]

$[a, b, c, d]$ represents the numbers of articles that the baseline, Grid, TM2011, and UTLIS system retrieved for the event collection subtask, and the baseline, Grid, KN, and UTLIS system for the trend summarization subtask, respectively.

Table 4: System Similarity based on Retrieved Articles

Similarity to the average

		1	2	3	4
E	Baseline	0.76	0.72	0.87	0.88
	Grid	0.78	0.56	0.78	0.82
	TM2011	0.79	0.71	0.86	0.93
	UTLIS	0.83	0.67	0.84	0.81
T	Baseline	0.90	0.84	0.69	0.89
	Grid	0.88	0.75	0.66	0.88
	KN	0.67	0.68	0.77	0.92
	UTLIS	0.85	0.76	0.72	0.94

Similarity to the baseline system

		1	2	3	4
E	Grid	0.48	0.28	0.50	0.57
	TM2011	0.44	0.41	0.70	0.75
	UTLIS	0.54	0.28	0.67	0.67
T	Grid	0.76	0.60	0.20	0.70
	KN	0.50	0.35	0.41	0.78
	UTLIS	0.72	0.66	0.38	0.79

reproduced and showed the behaviors of subjects on a time line. However, it is still difficult to relate such records of actions to their evaluation.

For these reasons, the log analysis common to all participating systems remains primitive. The difference of knowledge compiling time is one example of such analysis. Knowledge compiling time is defined as the time when the editor tab is active, and is an approximation of the time taken by the user to compile information and write it into a report using the editor after finding suitable information. Table 8 shows the average proportion of the knowledge compiling time to the total time of the task for each system and topic. Clearly, the baseline system in the event collection subtask and the KN system in the trend summarization subtask show relatively small values.

Along with this analysis, we can see how such proportions change as the information access process proceeds. For example, the proportion of knowledge compilation time may be large in the beginning, and then may decrease when a given topic has a small number of relevant articles that are easy to find. This is because the subject finds all relevant documents in the beginning, writes the report, and then tries in vain to find more articles. Based on such expectation, we divided the total task time into three parts, and examined the proportions of several actions including knowledge compilation in each part. Such analysis, however, did not yield meaningful results.

5.3 Questionnaire analysis

Regarding examples of data obtained by questionnaire surveys of the subjects, the subjects' evaluations of the systems in terms of usability, functionality, and efficiency are shown in Table 9. When not mentioned by others, the results in this section are the average values of the subjects' answers to questions, each of which used a seven-point Lik-

Table 6: The Number of Events Retrieved

		total	4 systems	3 systems	1 system
E	1	33	4	3 [3, 2, 2, 2]	22 [9, 5, 6, 2]
	2	44	4	4 [2, 3, 3, 4]	30 [7, 11, 4, 8]
	3	35	14	4 [2, 2, 4, 4]	11 [3, 7, 1, 0]
	4	28	11	3 [2, 2, 3, 2]	11 [1, 3, ,1, 6]

$[a, b, c, d]$ represents the numbers of articles that the baseline, Grid, TM2011, and UTLIS system retrieved, respectively.

Table 7: System Similarity based on Retrieved Events

Similarity to the average

		1	2	3	4
E	Baseline	0.90	0.87	0.93	0.92
	Grid	0.91	0.86	0.85	0.92
	TM2011	0.94	0.94	0.92	0.95
	UTLIS	0.93	0.88	0.94	0.92

Similarity to the baseline system

		1	2	3	4
E	Grid	0.76	0.68	0.69	0.79
	TM2011	0.76	0.77	0.80	0.84
	UTLIS	0.78	0.68	0.89	0.79

ert scale, with a score of 7 meaning "completely agree with." The results for systems that participated in both subtasks show that the evaluation in the event collection subtask is slightly higher for the baseline system and the UTLIS system, while the results in the trend summarization subtask are higher for the Grid system. This suggests some compatibility between a system and task characteristics.

The degree of satisfaction with achieving the task for each topic is shown in Table 10. These results were collected by the post-experiment questionnaire after all sessions had finished. The results show that the subjects were satisfied more in E4, T2, and T3, though the figures are not conclusive. The post-experiment questionnaire asked the subjects about their impression of the task in general, also. The average agreement score with the statement that the task was difficult was 4.3, that it was complex was 3.3, and that it was time consuming was 5.0. That is, the subjects regarded the given tasks as relatively simple but time consuming, suggesting that there were too many relevant articles.

Table 8: The Propotion of Knowledge Compilation Time

		1	2	3	4
E	Baseline	0.42	0.26	0.38	0.31
	Grid	0.45	0.38	0.44	0.43
	TM2011	0.53	0.32	0.40	0.43
	UTLIS	0.44	0.33	0.31	0.36
T	Baseline	0.55	0.60	0.49	0.41
	Grid	0.58	0.56	0.55	0.53
	KN	0.41	0.36	0.31	0.34
	UTLIS	0.56	0.58	0.58	0.41

Table 9: System Evaluation through Questionnaire Surveys

		Usability	Functionality	Efficiency
E	Baseline	5.8	4	4
	Grid	2.8	3.8	3.2
	TM2011	5.6	3.8	4.2
	UTLIS	5.8	4.2	5.2
T	Baseline	4	3.4	3.2
	Grid	5.8	5	5.2
	KN	5.4	4.4	4.6
	UTLIS	5	4	4.8

Table 10: The Degree of Satisfaction with Achieving Tasks

		1	2	3	4
E	Baseline	3.6	4	2.6	4.8
	Grid	4	4.6	4.6	4.8
	TM2011	3.8	3.2	4	4.4
	UTLIS	4.6	5	4	5.6
T	Baseline	3.2	4.4	4.8	3
	Grid	4.2	4.6	3.8	5.6
	KN	2.8	4.4	5.2	4.8
	UTLIS	4.4	4.6	5.2	2.6

In the post-session questionnaire which was conducted after each session, the subjects were asked about their familiarity with the topic, satisfaction with achieving the task, and its difficulty, among other matters. Table 11 shows relationships between some of those, using Pearson's product-moment correlation coefficient. It shows that a subject's familiarity with a topic in advance is scarcely correlated to the perceived difficulty in achieving the task. It also shows that the satisfaction with task achievement is strongly correlated to the feeling of having collected sufficient relevant information.

6. CONCLUSION

The VisEx task was conducted as a pilot task at the NTCIR-9 workshop, which aimed to establish a framework for evaluating interactive and explorative information access environments. Four teams participated in the task. Although we learned much, it was harder than expected to understand the results and to draw a clear picture. The task should be made more difficult in order to derive explorative behaviors of users. It is also necessary to reduce the diversity of user behavior, and to consider the log-taking mechanism for each submitted system. The basic framework appears to be useful for obtaining data on several aspects of complex behaviors of interactive information access. The experiments conducted this time yielded valuable data which should be closely analyzed. In conclusion, these experiments marked a useful first step toward establishing an efficient and effective methodology for objectively evaluating interactive and explorative information access environments.

7. REFERENCES

[1] S.T. Dumais and N.J. Belkin: The TREC Interactive Tracks: Putting the User into Search. In E.M.

Table 11: Correlation among Subjects' Impressions

	I'm satisfied with the report	The task is difficult
The task is difficult	-0.39	-
I'm familiar with the topic	0.18	0.01
I had enough time	0.50	-0.23
I collected enough information	0.75	-0.38
Relevant articles were large	0.56	-0.31

Voorhees and D.K. Harman ed. *TREC Experiment and Evaluation in Information Retrieval*. pp. 123-152. The MIT Press, 2005.

[2] H. Joho and T. Sakai: Grid-based Interaction for NTCIR-9 VisEx Task. In *Procs. of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, to appear, 2011.

[3] T. Kato and M. Matsushita: Overview of MuST at the NTCIR-7 Workshop – Challenges to Multi-modal Summarization for Trend Information. In *Procs. of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 475-488, 2008.

[4] T. Kato: Effects of the variety of Document Retrieval Methods on Interactive Information Access -An Experiment in the NTCIR-9 VisEx Task-. In *Procs. of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, to appear, 2011.

[5] T. Mitamura, E. Nyberg, et al.: Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access. In *Procs. of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 21-25, 2008.

[6] D. Smiley and E. Pugh: *Solr 1.4 Enterprise Search Server*. PACKT publishing, 2009.

[7] Y. Takama, S. Hattori, and R. Miyake: Read Article Management in Document Search Process for NTCIR-9 VisEx Task. In *Procs. of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, to appear, 2011.

[8] K. Tanaka, D. Hasui, and M. Matsushita: How Does a User Utilize a Chart-based Interface to Conduct Exploratory Data Analysis? In *Procs. of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, to appear, 2011.