Grid-based Interaction for NTCIR-9 VisEx Task

Hideo Joho
Faculty of Library, Information and Media
Science, University of Tsukuba
1-2 Kasuga, Tsukuba, Ibaraki 305-8550 JAPAN
hideo@slis.tsukuba.ac.jp

Tetsuya Sakai
Microsoft Research Asia
No.5 Danling Street, Haidian District, Beijing,
10080 P.R.China
tetsuyasakai@acm.org

ABSTRACT

This paper presents a grid-based interaction model that is designed to encourage searchers to organize a complex search space by managing $n \times m$ sub spaces. A search interface was developed based on the proposed interaction model, and its performance was evaluated by a user study carried out in the context of the NTCIR-9 VisEx Task. This paper reports findings from the experiment and discusses future directions of the research on the proposed model.

Keywords

Grid-based interaction, Exploratory search, User study

Team Name

JLTKB

Subtasks/Languages

Event Collection in Japanese
Trend Summarization in Japanese

External Resources Used

None

1. INTRODUCTION

Exploratory search is designed to help people with ill-defined information needs. People's information need tends to be ill-defined when they engage in a complex task. There are several models of task complexity, but Campbell [1], for example, proposed to look at a level of uncertainty in an input, process, and output of a given task to determine task complexity. To put it in the context of search, people might find it difficult to formulate effective queries, to develop search tactics and strategies, or to make appropriate relevance judgements of retrieved documents. Therefore, exploratory search systems should support multiple aspects of a complex search task.

Exploratory search systems can be categorized into two groups based on an accessibility to an entire document collection. When one has an access to an entire document collection, some level of content analysis can be performed and a set of common properties can be identified. These common properties are then used to organize the collection for users to explore. Examples of this type of exploratory search systems are Flamenco [7] and mSpace [5]. Both platforms extract a set of common properties (which are often called *facet*) from a data collection in advance. An advantage of the facet-based exploratory search systems is that searchers rarely get a zero result. Also, an overall structure of a data collection is visible to users which is helpful for exploration. An disadvantage is that a developer needs an access to an entire collection, which may not be possible in commercial search engines. Extraction and selection of facets are also often a manual or semi-automated process, and not trivial.

Another type of exploratory search systems does not assume an access to an entire collection. Therefore, support is often given to users dynamically based on their search history and some analysis of search results. Examples of this type of exploratory search systems are Aspectual browser [6] and Slice n' Dicer [3]. Both systems are designed to support exploratory search by finding properties common to a local document set (e.g., search results). Since these properties are not necessarily common to an entire collection, they are called *aspect* or *pseudo facet*. This type of systems is not designed to provide an overall structure of an entire collection, but has potential to be used in any document collections.

This paper proposes a new interaction model for the latter type of exploratory search systems, and studies its effect on task performance, information seeking behavior, and user perceptions.

2. GRID-BASED INTERACTION MODEL

An interaction model we propose in this paper is instructive. In other words, our interaction model is not based on how people naturally search but how people *should* search. Instructive design is not uncommon in consumer products. For example, spread sheet applications such as Microsoft Excel expect users to behave in a certain way to get a task done. A similar attribute can be found in search interfaces. Many search engines have a single query box for users to express their information need. This is not necessarily a natural way for us to get information that we seek, given that we often ask a question to get information. To put it in a positive way, many products, which are designed to enable people to perform a complex task, tend to *guide* users to do the task in a particular manner.

The interaction model we propose has a similar attribute. More specifically, the model is designed to guide people to express a complex search space using a grid-like metaphor. A cell in a grid is expressed by two dimensions, namely, *instances* and *aspects*. The rest of this section discusses the motivation behind the model, followed by information seeking behavior that is guided by the model. Finally, we introduce a query syntax to express a grid-based information need.

2.1 Motivation

The proposed interaction model was derived from two lines of research in interactive information retrieval (IIR). One is an instance finding task that has been studied in a series of Interactive Tracks in TREC [2]. In this task, searchers were asked to find instances of events, achievements, countries, technologies that matched a particular condition. The idea was to go beyond a conventional document retrieval which was often not sufficient to study interactive aspects in IR. For example, searchers received no reward to find duplicated instances in the task. A key element of this task was that it used instances as an axis of information seeking process. Instances can be a generic yet powerful property to search, organize, and analyze a given topic and its information space. For example, Barack Obama, David Cameron, and Hu Jintao are an instance of world leaders. Similarly, iOS, Android, and Windows Mobile are an instance of mobile operating systems. Instances tend to co-occur in documents of relevant documents since they are often contextually related. Therefore, we consider that the notion of instances plays an important role in supporting exploratory search tasks.

Another line of research is called aspectual search [6]. Aspectual search emphasizes to find aspects to complete an exploratory search task. For example, writing a biography of a world leader requires to find and select what aspects of the leader should be included. Making a summary of a large event like Olympics also depends on exploration of potential angles. These aspects or angles can be relatively simple such as time and location in some cases. However, when a topic becomes complex, determining appropriate aspects is not trivial. Again, a key element of this search was that it used aspects as an axis of information seeking process. Aspects can be seen as a property of instances. For example, age, education, and political agenda are an aspect of world leaders. Similarly, price, required memory, and hardware compatibility might be an aspect of mobile operating systems. If instances are a vertical axis, aspects can be seen as a horizontal axis in exploration and organization of a search space.

As can be seen, these two lines of research are closely related, and thus, can be integrated into a single model. This was our motivation and we call it a *grid-based interaction model*. Existing studies did not make a clear distinction of the two kinds of notion. In this sense, our model was an extension of the research which looked at instance finding and aspect finding. An important consideration here is that searchers are unlikely to find relevant instances and aspects at the beginning of search. Instead, they are more likely to discover these elements of a given topic as they make progress in a search task. Therefore, we need an interaction model that can support exploration of a search space using the notion of aspects and instances.

2.2 Query Syntax

We consider that an expression of information needs is crucial for supporting exploratory search, and thus, devised a syntax to formulate a grid-based query as shown in Figure 1. Dimensions in a query space are separated by a bar sign (|), which means that this syntax can represent as many dimensions as needed. However, in this paper, we only consider two dimensions. All the terms placed before the bar sign represent the first dimension while those after the bar sign represent the second dimension. Keywords in individual dimensions are separated by a comma (,). When more than one keyword is used between commas, such term will be taken as a phrase.

```
term [, term...]  | term [, term...]
```

Figure 1: Query syntax for a grid search

For example, the query, **`'black, white | cat, dog'`**, will represent four search queries such as 'black cat', 'black dog', 'white cat', and 'white dog'. The query, **`'david cameron, barack obama | approval rate'`**, will yield two queries such as <"david cameron" "approval rate"> and <"barack obama" "approval rate">. As can be seen, the syntax is simple and intuitive. A user study suggests that participants did not have a major problem to operate a search task with the proposed syntax.

To link with the notions of instances and aspects described earlier, the first query had aspects first then instances, while the second query had instances first then aspects. Therefore, the proposed syntax does not explicitly define the order of dimensions expressed in the query. It is up to a designer (or possibly user) of search systems. The next section describes a grid search interface which can be seen as one possible case of implementing the proposed query syntax.

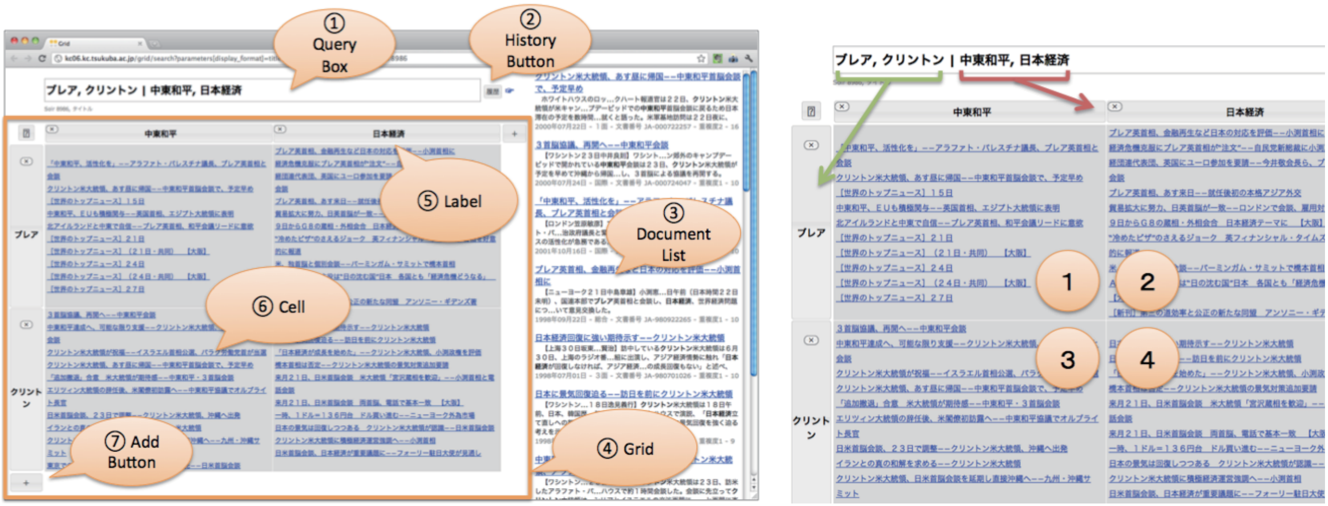
3. GRID SEARCH INTERFACE

A search interface was developed based on the interaction model discussed in Section 2. This section describes how our grid search interface was designed to work.

The interface is composed of three main areas, namely, Query Box (1), Document List (3), and Grid (4), as shown in Figure 2. When a user submits a query in Query Box using the proposed syntax, both Document List and Grid are shown with search results. There is a History button next to Query Box which allows users to revisit and rerun previous queries submitted in a search session.

The Grid area consists of Labels (5) and Cells (6). Labels are derived from individual terms given in Query Box. Given that a user submits a query `Blair, Clinton | Middle East Peace, Japanese Economy` to the interface, terms on the left part of a query (i.e., Blair and Clinton) are placed as a row label (Green Line), while terms on the right part (i.e., Middle East Peace and Japanese Economy) are placed as a column label (Red Line). Moreover, the terms in Query Box and Labels are synchronized, and users can edit either of them to reformulate the query. Double-clicking a label allows a user to change the keywords. Add Button (7) allows a user to append a new label to the grid. Full-text is shown in a pop-up window when the title of the result is clicked from a cell or document list.

Cells (6) show search results of the grid-based queries as shown in Figure 2. Each cell shows the result of a particular combination of terms derived from the query. Taking our earlier example query, Cell 1 shows a search result of a query, <Blair "Middle East Peace ">, Cell 2 shows that of <Blair "Japanese Economy">, and so forth. As a consequence, one grid presents the results of four sub-queries in this case. Grid does not have to be $n \times n$, and can be $n \times m$. Furthermore, the order of rows and columns can be changed by dragging a label.



(a) Interface components (b) Grid organisation

Figure 2: Screenshots of Grid Search UI (Query: Blair, Clinton | Middle East Peace, Japanese Economy)

When a cell is selected, Document List shows search results of a particular sub-query (e.g., <Blair "Japanese Economy">). When multiple cells are selected, Document List shows search results by merging selected individual cells. Currently, the merged ranking is based on redundancy and round-robin. In other words, those documents that are commonly retrieved by sub-queries are ranked higher than those are retrieved once. Tie-documents are ranked in a round-robin manner across the retrieved documents of individual sub-queries. When a query is submitted or reformulated in Query Box or via Labels, Document List shows a set of retrieved documents by merging the results of all sub-queries. This is equivalent to select all cells in Grid. A more effective way to generate the document list is of our future work. Finally, users can see the next 10 results from a navigation link available at the bottom of Document List area (not shown in the figures).

4. EXPERIMENT

The evaluation of the proposed search interface was carried out as a user study in the framework of NTCIR-9 VisEx Task[1]. Since the detail of experimental design is given in the overview paper of VisEx Task [4], this section only gives a brief summary of how a user study were carried out. It should be noted that a team who participated in VisEx Task will be referred to as *participants*, while people who participated in the user study as a subject will be referred to as *subjects* in the rest of this paper.

The task required participants to develop a user interface that can support either an event collection task or trend summarization task or both. An example of the event collection task is to find as many instances of airplane crashes that occurred in Asia as possible. An example of the trend summarization task is to find as many information that describes a prime minister's approval rate as possible. As can be seen, both tasks were exploratory and required a number of iterations of search to complete them. Subjects were given up to 50 minutes to complete the tasks. A baseline system was provided by the organizer and all systems used the API of a common backend search engine (Apache Solr).

Since we participated in both tasks, the experimental design can be seen as a 2 x 2 factorial design with task and system as the factors. Levels of the task factor consisted of event collection task and trend summarization task. Levels of the system factor consisted of the baseline system and grid-search (experimental) system. Although the VisEx framework allowed us to compare the experimental system to other participated systems, this paper only reports the findings from the comparison between the baseline system and experimental system.

Subjects were given an instruction (6 slides) of how to use the systems and a training session to familiarize with tasks and behavior of the experimental system. In the system instruction, an emphasis was made to encourage people to organize a problem space using a $n \times m$ notion such as *people* × *year*, *place* × *event*, and *things* × *attributes*.

5. RESULTS

The evaluation of the proposed system was based on three groups of dependent variables, namely, the task performance, information seeking behavior, and subjective assessments. This section presents results of our preliminary analysis on the two exploratory tasks of VisEx. It should be noted that our analysis is intentionally *qualitative* due to the experimental design. Therefore, no statistical test is performed unless otherwise stated. Finally, please note that we use the term *nugget* to refer to a unit of information required to collect in individual tasks, which might be different from traditional use.

5.1 Task Performance

Both tasks asked subjects to collect relevant as many relevant nuggets as possible within allocated time. Therefore, an overall task performance can be measured by the number of nuggets found. The results are shown in Table 1. If you look at the third row of Table 1, the performance was comparable between the baseline and experimental systems

[1] `http://must.c.u-tokyo.ac.jp/visex/hiki.cgi?FrontE`

	Event collection		Trend Summary	
	Baseline	Experimental	Baseline	Experimental
Mean	7.6	7.4	7.6	6.9
SD	3.5	2.8	3.1	2.8
Min	2.0	4.0	3.0	3.0
Max	16.0	15.0	14.0	12.0

Table 1: Number of nuggets found (N=20)

in the former task, while the difference between the two systems was found to be larger in the latter task. A similar trend was found in the maximum number of nuggets found by subjects. A grand average number of nuggets found by subjects was 7.5 (SD: 3.1) and 7.2 (SD: 2.9) for Event collection task and Trend summarization task, respectively (not shown in the table).

The next analysis looked at whether or not a particular topic was easy (or difficult), or a particular subject performed really well (or bad). The results are shown in Figure 3. For the topic breakdown of the number of nuggets, the X axis represents Topic 1 to 4 of the two task while the Y axis represents the number of nuggets found. Each topic line has five data points which correspond to five subjects. A horizontal bar is the median value of the five data points. Please note that Topic 1 of Event collection task has nothing to do with Topic 1 of Trend summarization task. Similarly, for the subject breakdown, the X axis represents Subject 1 to 5, and data points are the number of nuggets found by each subject across four topics. Again, Subject 1 in Event collection task is a different person from Subject 1 in Trend summarization task.

As for the topic breakdown, no obvious pattern was observed from the analysis. Topic 3 of Event collection task seems to have the best performance in both systems, but the data range is large, so this is not necessarily the easiest topic. Subjects with the baseline system appear to struggle in Topic 2 of Event collection task while Topic 1 appears to be the most difficult one by subjects of the experimental system. In Trend summarization task, fewer noticeable pattern was observed. As for the subject breakdown, it appears that subjects' performance varies over the system groups as well as the task groups. All groups seem to have a good performer and poor performer, some had a large difference across topics, while others had a similar performance.

Finally, we looked at whether or not we had order effects on subjects' performance. There were at least two reasons for us to investigate the effect. First, since this was the first time for subjects to carry out a search task with the grid search interface, their performance might increase as they got used to the system. Second, since individual tasks lasted 50 minutes, we suspected that there might be fatigue effect on their performance towards later topics. The results are shown in Figure 4. Each data cell represents a subject's task performance in the order of topics.

Again, no pattern was found to be noticeably frequent. Average number of nuggets were 5.9, 5.8, 9.4, and 8.8 for the first to fourth topics of Event collection task, while they were 7.5, 7.6, 7.3, and 6.5 for Trend summarization task. Therefore, there might be fatigue effect on Trend summarization task. Pearson's correlation shows a significant correlation ($p \leq .003$) between the number of nuggets and topic order, but coefficient was $r = .44$ which means the contribution ratio is below 20%. Therefore, an order effect appears to be generally weak.

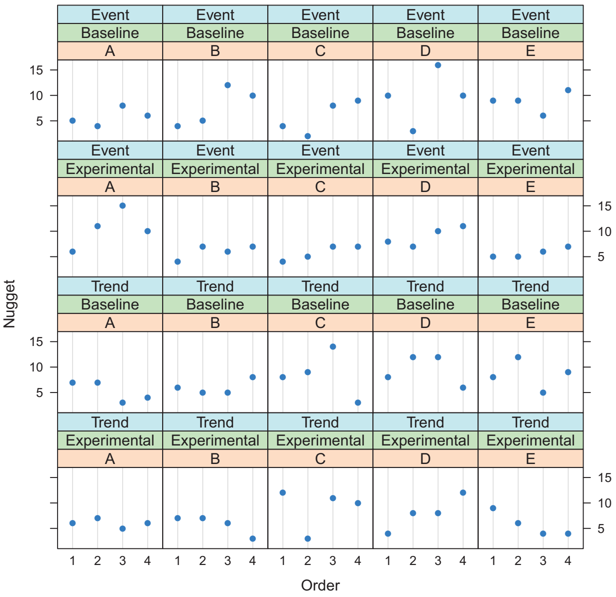


Figure 4: Order effect on the number of nugget

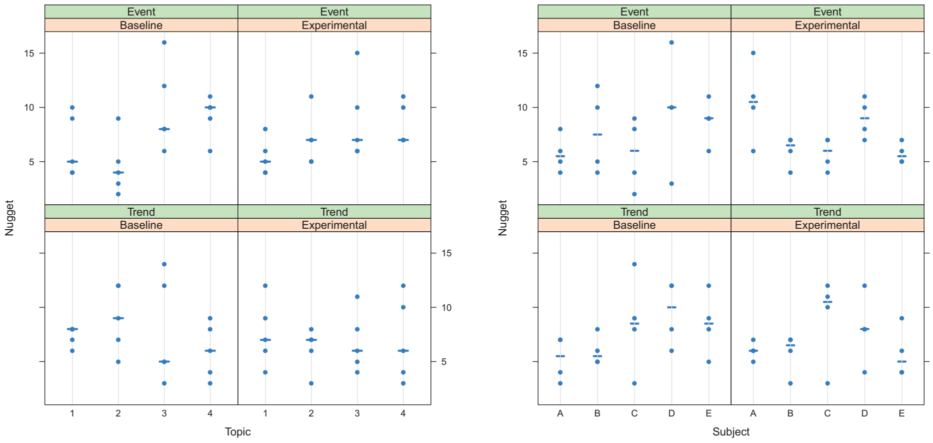
#	Left	Right
1	nuclear test	carried out
2	nuclear test	France
3	nuclear test	North Korea
4	nuclear test	Russia
5	nuclear test	nuclear-capable
6	nuclear test	UK
7	nuclear test	France
8	nuclear test	China
9	nuclear test	UK

Table 2: Query history of the best session in Event collection task with the experimental system (Topic 3)

To summarize the results of task performance, subjects with the experimental system appear to achieve a comparable performance to the baseline system in Event collection task, while there seems to be some factor in the experiment system that caused a performance loss in Trend summarization task. These observations did not seem to be influenced by a particular topic, subject, nor order, although their interaction effect might exist. The following sections look at information seeking behavior and subjective assessments to gain a further insight into these results.

5.2 Information Seeking Behavior

An advantage of the proposed query syntax was that it allowed a user to express a complex search space as they made progress in exploratory search tasks. Therefore, we first looked at how subjects formulated and reformulated their queries during the VisEx tasks. Query history of the *best performing session* (i.e., the session with the largest



(a) Topic Breakdown

(b) Subject Breakdown

Figure 3: Breakdown of the number of nuggets found

#	Left	|	Right
1	Dubai crude oil, WTI	|	gasoline, price
2	Dubai, WTI	|	gasoline, price
3	Dubai, crude oil	|	gasoline, price
4	crude oil, gasoline	|	1999, 2000
5	Dubai, gasoline	|	1999, 2000
6	Dubai, gasoline	|	1999, 2000
7	Dubai crude oil, gasoline	|	1999, 2000
8	Dubai>crude oil, gasoline	|	1999, 2000
9	Dubai, crude oil, gasoline	|	1999, 2000
10	crude oil, gasoline	|	1999, 2000
11	crude oil, gasoline	|	february
12	Dubai, crude oil	|	gasoline, price
13	Dubai, crude oil	|	gasoline, price, 1999
14	Dubai, crude oil	|	gasoline, price, 1998
15	decline	|	2000
16	decline	|	2000, Dubai

Table 3: Query history of the best session in Trend summarization task with the experimental system (Topic 1)

number of nuggets found) using the experimental system was shown in Table 2 and Table 3 for Event collection task and Trend summarization task, respectively. Queries are a translation from the original language used in the experiment (i.e., Japanese).

Table 2 shows the queries for the topic which asked for event information (e.g., time, location, people) about nuclear tests all over the world. The first query appears to try to retrieve documents that includes a text like "carried out a nuclear test". In fact, the subject found several nuggets from this query. Then, except Query #5, the rest is a combination of "nuclear test" and a country name. In other words, the grid did not really expand as the task developed. The number of nuggets found by those queries was mixed. More importantly, those queries with country names could have been expressed as `nuclear test | France, North Korea, Russia, UK, China`.

	Event collection		Trend Summary	
	Baseline	Experimental	Baseline	Experimental
Mean	.79	.89	.97	.96
SD	.17	.27	.30	.34
Min	.44	.36	.43	.31
Max	1.14	1.33	1.75	1.71

Table 4: Successful click-through rate (N=20)

Table 3 shows the queries for the topic which asked for trend information about the price of crude oil and regular gasoline in Dubai. This topic was more complex than the previous example, since it explicitly asked for three dimensions: time, price, and oil type. Using the proposed syntax, users would have to somehow divide them into a set of two dimensions. Such attempt can be found in Query #4 and #11 where the subject focused on oil type and year. However, it seems that the results were not satisfactory without a location in the query, and thus, other queries contains Dubai in either side of the query.

The next analysis looked at a successful click-through rate (SCTR) which is a proportion of click-through that led a subject to find relevant nuggets in all click-through documents. In other words, SCTR indicates to what extent one's initial relevance judgement based on document surrogates of search results were indeed correct. Finding relevant nuggets requires a more precise examination than finding relevant documents, and thus, one would like to maximize SCTR to complete a task efficiently.

The results of SCTR are shown in Table 4. As can be seen from the third row, subjects with the experimental system appeared to have a higher level of SCTR than the baseline system in Event collection task, while they were comparable in Trend summarization task. A similar trend was found in the best SCTR in individual conditions, shown at the bottom of the table. Note that $SCRT \geq 1$ means that subjects found on average more than one nuggets per click-through. Given that the number of nuggets found in the experimental systems was lower than the baseline system in Trend

	Event collection	Trend summary
Baseline	10	3
Experimental	5	4

Table 5: Number of sessions which used external sources (N=20)

	Event collection		Trend Summary	
	Base.	Exp.	Base.	Exp.
Satisfaction	3.0	3.0	4.0	3.0
Difficulty	3.5	4.0	4.0	3.5
Time	4.0	3.0	4.0	3.0
Exhaustive	4.0	3.0	4.0	4.0
Resource	2.0	3.0	4.0	3.0

Table 6: Perception of tasks (N=20) (1: Strongly Agreed, 4: Either, 7: Strongly Disagreed)

summarization task yet the SCTR was comparable, subjects might have spent more time on query re/formulation with the experimental system.

The last analysis on information seeking behavior examined to what extent subjects accessed external sources to complete tasks. In the VisEx tasks, subjects were allowed to access external sources to support their tasks, although none of the nuggets found in external sources counted in performance measures. The results are shown in Table 5. As can be seen, there was a noticeable difference between the two systems in Event collection task while the frequency was comparable in Trend summarization task. Given that an overall task performance of Event collection task was comparable between the two systems, subjects with the baseline system appeared to need more support outside the given system than the experimental system. The motivation for accessing external sources varied. The post-search questionnaire established that subjects sought for a detail of a particular event, definition of a technical term, or information needed to judge relevance of nuggets.

To summarize the results of information seeking behavior, we have observed cases where the proposed syntax could be effective to organize a search space. There were cases where subjects organized a query space two-dimensionally (e.g., oil types × year in Table 3). However, the frequency of effective use of the proposed syntax was relative low. As for relevance judgements on document surrogates, the experimental system appeared to increase a chance of finding relevant nuggets in click-through documents in Event collection task. Finally, subjects tended to need an access to external sources in the baseline system when compared to the experimental system.

5.3 User perceptions

The last part of our analysis looked at subjects' perceptions on tasks and systems. Subjective assessments were captured by a 7-point Likert scale where subjects indicated a degree of agreement to a given statement. For example, a statement was “The task I performed was complex” and the scale were Strongly Agree (1), to Either (4), to Strongly Disagree (7). Task perceptions were captured after every topics, while system perceptions and interaction perceptions were captured after all topics.

	Event collection	Trend summary
Baseline	1	10
Experimental	16	8

Table 7: Number of sessions where subjects discovered new knowledge (N=20)

The results of task perceptions are shown in Table 6 and its topic breakdown in Figure 5. The numbers in Table 6 are a median of corresponding data. As can be seen, we did not observe a large difference in any aspects of task perceptions. However, Figure 5 suggests that, for example, the difference in Satisfaction in Trend summarization task is likely to be due to Topic 3 and 4. On the other hand, the difference in the perception of time (You had sufficient time to complete a task) seems to be consistent over all topics in both tasks. The difference in the perception of resources (You found a sufficient amount of news articles to complete a task) is likely to be due to Topic 1 and 3 in Event collection task. Little pattern was observed in the rest of questions.

Another question we asked in the post-search questionnaire was whether or not subjects discovered new knowledge which was somehow unexpected, during the task. The results are shown in Table 7. As can be seen, there was a surprisingly large difference between the two systems in Event collection task, while the number was comparable in Trend summarization task. Examples of discovery reported by subjects include an association between two events, varied amount of information across countries, lack or bias of information in news articles, as well as topics themselves.

The next set of questions asked subjects' perceptions of the systems, and the results are shown in Figure 6. In Event collection task, subjects appeared to find the baseline system easier to learn and to operate than the experimental system. There was a clear trend in the assessments of functionality. Additional comments suggest that subjects wanted an ability to submit a standard query, to move to next 10 results within a cell, and to sort documents by date. Response speed seems to be acceptable. All subjects seemed to feel some level of frustration during the task, but the variance was much larger in the experimental system than the baseline system.

To summarize user perceptions, we did not observe a large difference between the two systems in terms of the perception of tasks although some values varied over topics. However, with the experimental system, subjects tended to encounter new knowledge during the tasks when compared to the baseline system. As for the perception of systems, subjects seemed to find it more familiar to the baseline system than the experimental system. Subjects expressed several features that they would like to have in the experimental system, which can be considered for further development.

6. DISCUSSION

As we have seen so far, the performance of the grid-based search interface was measured by three aspects in our study: task performance, information seeking behavior, and subjective assessments. This section first summarizes the major findings of the experiment and discuss their implications for further research and development of the grid-based interaction model.

6.1 Event collection task

The overall task performance of the two systems was comparable in Event collection task (Table 1). However, we observed some positive signals in the experiment. First, based on the result of successful click-through rate (Table 4), users might be more accurate in detecting the documents with relevant nuggets from search results when they use the grid-based interface. This might be due to a structured presentation of search results in the experimental system. Second, based on the frequency of external source access (Table 5) and of discovering new knowledge (Table 7), the grid-based interface might facilitate the exploration of a document collection through analytical search process during the tasks. It might be possible that the performance of the baseline system was actually due to a frequent access to external sources. We need more analysis to exploit this aspect.

We also obtained ideas to improve the current implementation of the grid-based interaction model. The query analysis of the best performing session (Table 2) suggests that there were cases where the proposed query syntax can be effective to represent a complex search space. However, subjects did not appear to take advantage of the syntax. Since our tasks were not simple like home page finding, it is possible that subjects were focusing on finding relevant nuggets than effectively leveraging the potential benefit of the syntax. Therefore, we should take more time to familiarize with the system. More examples should be given in the instruction, and a step-by-step tutorial might be needed in a training session. A comparison of query formulation process between the two systems should give us a better idea of how exactly such instruction should be formed.

6.2 Trend summarisation task

The overall task performance of the experimental system appeared to be lower than the baseline system in Trend summarization task. We did not observe a particular topic or subject strongly affected the average performance (Figure 3). Furthermore, several aspects such as the successful click-through rate, frequency of accessing external sources, and frequency of discovering new knowledge seem to be comparable between the two systems.

We speculate that information needs often formulated in Trend summarization task require more than two dimensions to express, which was not supported well in the current implementation of the interface. We intentionally limited the dimension size to two, but this could cause an extra effort to subjects to divide a search space to a set of two dimensions. This was exemplified in the query reformulation of the best session in this task (Table 3). In short, this task was more complex than the current search interface was designed to support. It is technically possible to expand the proposed query syntax to accept more than two dimensions. However, this would require further consideration regarding how to present search results in a way that they make sense to searchers. More fundamentally, we need to study how to guide searchers to divide high-dimensional complex search into sub spaces, and how to support such tactics using the grid-based interaction model. These are all interesting research questions to pursue as future work.

Finally, it should be emphasized that participating in both of the exploratory tasks available in NTCIR-9 VisEx turned out to be beneficial, since we gained different aspects of insight into the effects of the grid-based interaction model and interface.

7. CONCLUSION AND FUTURE WORK

This paper proposed a grid-based interaction model to support exploratory work by encouraging searchers to organize a search space using $n \times m$ subspaces. The model includes a query syntax to express multidimensional information needs. A search interface was developed based as one possible implementation of the proposed interaction model. A user study was carried out to measure effects of the proposed model on performance, information seeking behavior, and perceptions in the context of NTCIR-9 VisEx task. The findings from the experiment suggest several directions for further research on the grid-based interaction model.

First, it seems worthwhile investigating an effect of a query syntax which accepts more than two dimensions on people's query formulation process. Effective ways to present search results of over 2 dimensional quires also need to be investigated. Second, we are interested in strengthen the interface's ability to support *comparative analysis* of search results, as it is missing from most exploratory search interfaces. Third, we plan to integrate a mechanism of suggesting new aspects or instances, as in interactive query expansion in IR. Finally, a method to automatically generate a grid-based query from a standard query should be developed.

8. ACKNOWLEDGEMENTS

This work was funded by Microsoft Research Asia under the MS Core 7 Project. Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] D. J. Campbell. Task complexity: A review and analysis. *The Academy of Management Review*, 13(1):40–52, 1988.

[2] S. Dumais and N. Belkin. The TREC interactive tracks: Putting the user into search. In E. M. Voorhees and D. M. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, pages 123–153. MIT Press, 2005.

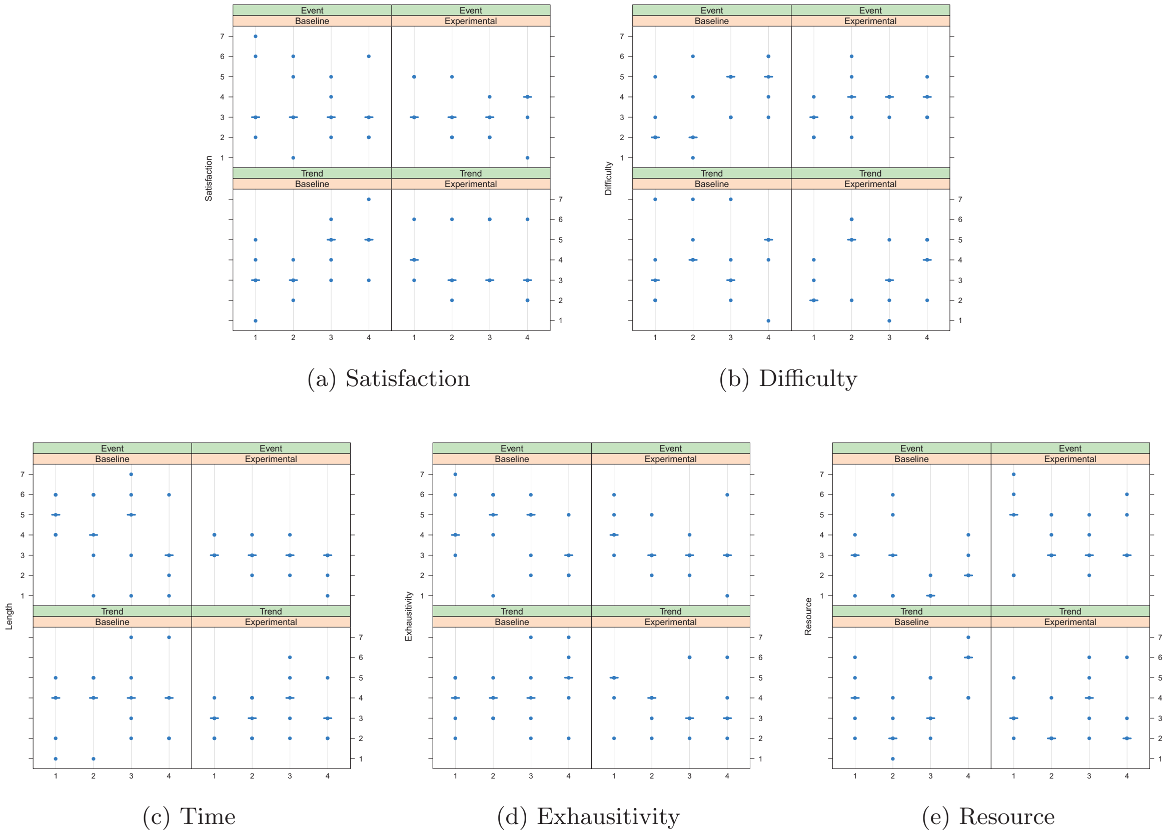
[3] H. Joho and J. M. Jose. Slicing and dicing the information space using local contexts. In *Proceedings of the First Symposium on IIiX*, pages 111–126, 2006.

[4] T. Kato, M. Matsuhita, and H. Joho. Overview of the VisEx task at NTCIR-9. In *Proceedings of the Ninth NTCIR Workshop Meeting*, Tokyo, Japan, 2011. NII.

[5] m. schraefel, M. Wilson, A. Russell, and D. A. Smith. mspace: improving information access to multimedia domains with multimodal exploratory search. *The Communication of ACM*, 49:47–49, 2006.

[6] R. Villa, I. Cantador, H. Joho, and J. M. Jose. An aspectual interface for supporting complex search tasks. In *Proceedings of the 32nd international ACM SIGIR conference*, pages 379–386. ACM, 2009.

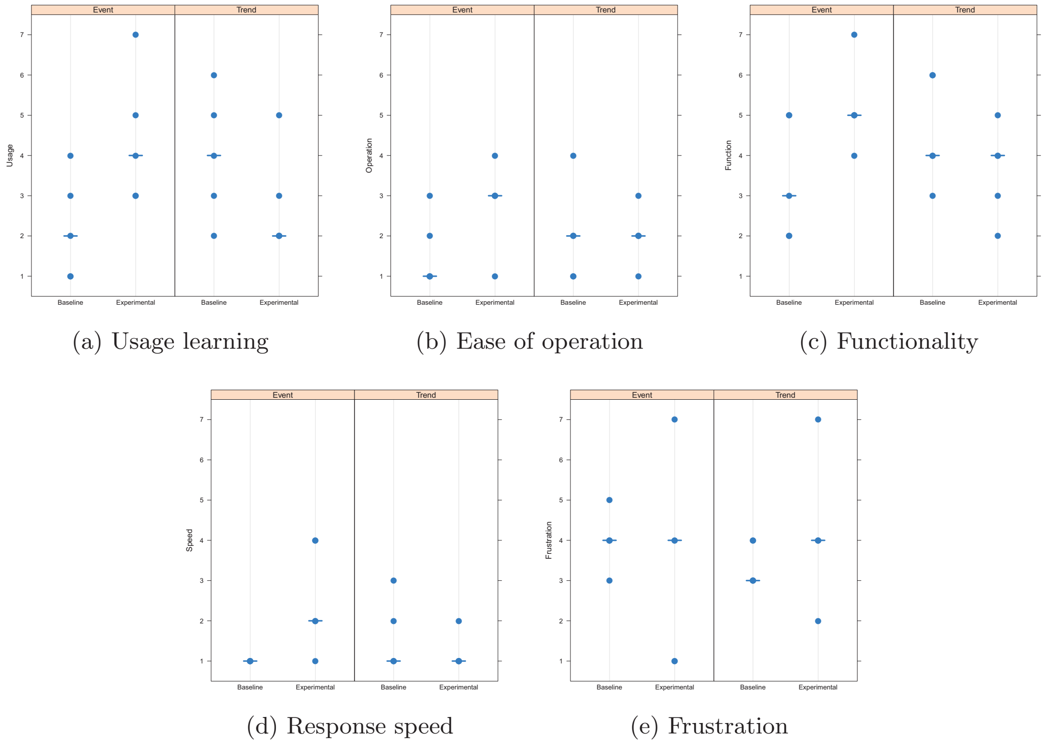
[7] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI conference*, pages 401–408. ACM, 2003.



(a) Satisfaction (b) Difficulty

(c) Time (d) Exhausitivity (e) Resource

Figure 5: Perceptions of tasks: Topic Breakdown (1: Strongly Agreed, 4: Either, 7: Strongly Disagreed)



(a) Usage learning (b) Ease of operation (c) Functionality

(d) Response speed (e) Frustration

Figure 6: Perceptions of systems (1: Strongly Agreed, 4: Either, 7: Strongly Disagreed)