CRES at LogCLEF 2010: Towards Understanding the User Behaviors through an Analysis of Search Sessions, Search Units and Click Ranks[1]

Masao Takaku[2,7], Yuka Egusa[3,7], Hitomi Saito[4,7], Noriko Kando[1] ,
Hitoshi Terai[5,7] and Makiko Miwa[6,7],
[1] National Instititue of Informatics,
Tokyo 101-8430, Japan
{kando, cres}@nii.ac.jp
[2] National Institute for Materials Science
Ibaraki 305-0047, Japan
TAKAKU.Masao@nims.go.jp
[3] National Institute for Educational Policy Research
Tokyo 100-8951, Japan
yuka@nier.go.jp
[4] Aichi University of Education
Aichi 448-8542, Japan
hsaito@auecc.aichi-edu.ac.jp
[5] Nagoya University,
Aichi 464-8601, Japan
terai@cog.human.nagoya-u.ac.jp
[6] The Open University of Japan
2-11 Wakaba, Mihama,
Chiba, 261-8586, Japan
miwamaki@ouj.ac.jp
[7] Collaborative Researcher,
National Instititue of Informatics,
Tokyo 101-8430, Japan

Abstract. This paper describes the participation of Cognitive Research on Exploratory Search (CRES) collaborative research group at National Institute of Informatics (NII) for LogCLEF 2010. Analysis of multilingual search logs from two separated time periods was conducted with the purposes to investigate the users search behaviors as processes which consisting of sequences of actions and duration. We extended our methodologies investigating the users' search behavior using the laboratory user experiments and the user-side log analysis to the TEL's action logs. For the first, we cleaned up the log data by discarding the records from the periods without "recordPosition" or "timestamp". Secondly we did mapping the analytical framework for the Web search to the TEL by comparing the actions recorded in the TEL's log files and those recorded in our client-side logs for web search, and the page transitions in TEL and the web search. Thirdly, we have analyzed the TEL's action logs in terms of search sessions, search units and click ranks. As results the numbers of the actions in search sessions and search units were generally short and any

[1] The work is supported by MEXT grand-in-aid (#21300096) and NII Collaborative Research Program

particular groups of the uses were not found so far. But we could see rather small number of search sessions and search units had different tendencies of the number of the actions and their durations. For the future works, investigating the differences of the users' behaviors across different language and/or cultural background is considered. The scripts that we have developed and used for analysis are available through sourceforge.

Keywords: log analysis, users' search behavior, search sessions, search units, click ranks, search tasks

1 Introduction

LogCLEF 2010 is one of the workshops of CLEF 2010 Labs at Cross Language Evaluation Forum (CLEF). In LogCLEF 2010, a common data set was distributed to the participants, and in coordination with the organizers, participating groups were devoted to different tasks in exploring and understanding the data. Our group has devoted to a task to investigate the users' behavior as processes which consisting of sequences of actions and duration through the analysis of the action logs from the European Library (TEL), which is a digital library with a single user interface to search across the contents provided from many national libraries in Europe. This paper reports the results of the analysis in terms of the search sessions, search units and click ranks.

This participation is conducted as part of ongoing research activities within the Cognitive Research on Exploratory Search (CRES) collaborative research group[2] of National Institute of Informatics (NII). CRES is investigating the users' information seeking behaviors of the exploratory search on the Web for different tasks, with different levels of expertise about the search strategies, about the topic or subject domain, and about the tasks, through the analysis of the data collected from the user experiments and the client-side logs. The over all purposes are to understand the users problem solving process during the search, to evaluate the exploratory search, and to propose novel search user interfaces and search functionalities based on the investigation of users' behaviour. In this participation we intended to extend our analytical frameworks to the action log of TEL to understand the users' search behaviors on a multilingual digital library.

Although many existing studies on log analysis have focused on the queries and the click-through, we have placed emphases to capture the users' search behaviors as a search process, or a series of actions and duration. This is partially because that substantial part of the information needs cannot be fulfilled by a single iteration of a search and the users often gradually specifying or clarifying the focus of the search, learning through the search interaction to have a better insight and to cumulating the understanding, and the users' interest may be developed or shifted during the interaction. Our investigations so far have indicated that such interactive search processes vary according to the types of the users' information seeking tasks or the purposes of the search, and the users' expertise. And it is also important to propose a

[2] http://cres.jpn.org/

functionality to support such interactive and/or exploratory information seeking. This paper however reported an initial analysis and general tendencies seen in the dataset, the method can be extensible to investigate the differences of the user groups with different languages or cultural backgrounds, which we hope to report some at the workshop in September 2010.

For the rest of this paper is organized as follows; Section 2 briefly describes the data preparation. Section 3 explains the basic ideas for the analysis. Section 4 describes the methods. Section 5 reports the results. Section 6 is conclusion. And appendix provides brief description of the tools we have developed and used for analysis, which are available from *sourceforge.*

2 Data Preparation

The raw data sets used for our analysis are the action log files of TEL, *logclef.zip* (Jan 2007 - June 2008) and *logclef2.zip* (Jan 2009 - Dec 2009). These two log files contain the same set of the elements with different time period of the records and different separators.

The former consists of *semicolon*-separated value data and contains 1,866,330 lines (283,993,551 bytes). Among them, the data before 2007-03-16 09:33:04 are without the "recordPosition" (the 10th column, *i.e.* "click rank"), and about half of the data after that time also did not record the "recordPosition".

The latter consists of *comma*-separated value data and contains 762,485 lines (128,119,174 bytes). Among them, all the value of the timestamp (access date) was "00:00:00" before 2009-09-16 13:10:27.966.

Table 1 showed the *dataset1* and *dataset2* that we used for the analysis described in the next section. The dataset1 was constructed by discarding the records for the period from 2007-01-01 to 2007-03-16 09:33:04 from the logclef.zip as none of the records of the time contained "recordPosition". The dataset2 was constructed by discarding the records for the periods from 2009-01-01 to 2009-09-06 from logclef2.zip as none of the records for the period contained the timestamp. The numbers of the records used for the analysis were 1,560,682 and 300,323, respectively.

Table 1. Summary of the Data Used

File name	Data name	Period of date	No. of records	Action name	
				record-Position*	timestamp*
logclef.zip	-	2007-01-01 -- 2007-03-16	305648	0.00%	100.00%
	dataset1	2007-03-16* -- 2008-06-30	1560682	49.95%	100.00%
logclef2.zip	-	2009-01-01 -- 2009-09-16	462136	100.00%	0.00%
	dataset2	2009-09-16*-- 2009-12-31	300323	100.00%	100.00%

*recordPosition: ratio of valid recordPostion
*timestamp: ratio of valid timestamp
*2007/03/16 09:33:04
*2009/09/16 13:10:27.966

As a pre-processing of the data, records in the dataset1 and dataset2 were converted into tab-separated value data files. We defined the „search session“ and re-assigned the session ID. The ID was called as „cres_sesid“ hereafter. The definition of the search session will be described later in the sectin of 3.3.

The logfiles of dataset 1 and dataset2 were analyzed using the Ruby scripts. The analytical tools that we have developed and used were made available through the Sourceforge (http://en.sourceforge.jp/projects/cres/svn/view/logclef/).

3 Methods

3.1 CRES Frameworks to Analyze the Users' Behavior in the Search Processes

To investigate the users' behaviors on the exploratory search on the Web, we have proposed the various frameworks including *i)* "*Web Action Categories*" [1] and "*Link Depth*" [2] for users' actions, *ii)* "*Lookzone*" [1] for eye movement and a visualizing tool "*Scanpath2SVG*" [7], *iii)* "*Taxonomy of Knowledge Modification and Knowledge Utilization Patterns*" for content-analysis of the qualitative data like think-aloud and interview [4-6], *iv)* "*COPATT*" as a tool integrating the above mentioned data [1], and *v)* "*Concept Map*" and its visualizing tool "*VizCMaps*" to measure the changes in the user's knowledge between pre- and post search [3]. We have also developed a client-side logging tool called "*QT-Honey*" to capture all the users actions defined by the Web Action Categories, Link Depth, click ranks, and duration of each action from the users' point of view. And then we have analyzed the differences of the users' search processes by the types of the users' search tasks and the users' expertise about search, the topics and the tasks.

In this participation, we intended to investigate the users' behavior on TEL by analyzing the action logs using the frameworks extended from the ones we have used and to characterize the users' behavior in the search sessions through the comparison between the previous works on Web search using user experiments and client-side logs

In this purpose, we define *1)* the correspondence between TEL's action logs and the client-side logs captured using QT-Honey focusing on the users' actions defined by the Web Action Categories (Table 2) , and *2)* the analytical units.

Table 2. Web Action Categories

Search: searching with a search engine
Link: clicking on a page link
Next: going forward to the next page
Return: going backward to the previous page
Jump: going to a page in the Bookmark or History
Browse: browsing the next search
Submit: clicking a submit button
Bookmark: adding bookmarks
Change: changing from one tab to another
Close: closing a tab or window

3.2 Mapping the Frameworks

Figure 1 shows the framework that we have used for the analysis of the users' behavior on the web search, and the correspondent actions in the TEL's action logs with typical usages in both of the Web search and TEL.

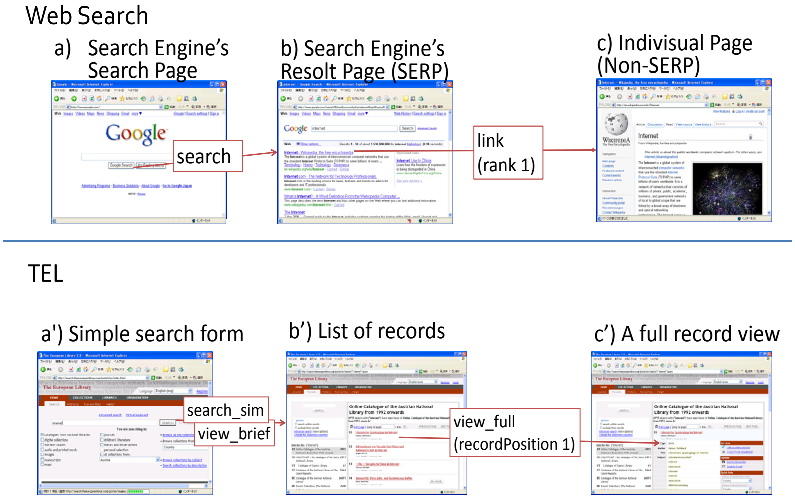

Web Search
a) Search Engine's Search Page
b) Search Engine's Resolt Page (SERP)
c) Indisivual Page (Non-SERP)
search
link (rank 1)
TEL
a') Simple search form
b') List of records
c') A full record view
search_sim
view_brief
view_full (recordPosition 1)

Figure 1: Comparison of the Frameworks for Web search and TEL.

In Figure 1, the actions captured by QT Honey and TEL's action logs were in the red rectangles. For TEL, b') is a list of the abstracts (title, author, type, language) of the top 10 retrieved items in the first collection after sorted by the collection names. c') is a full record view of the TEL and contains a detailed information about each item. We could see the correspondences between each of the pairs of a) and a'), b) and b'), and c) and c').

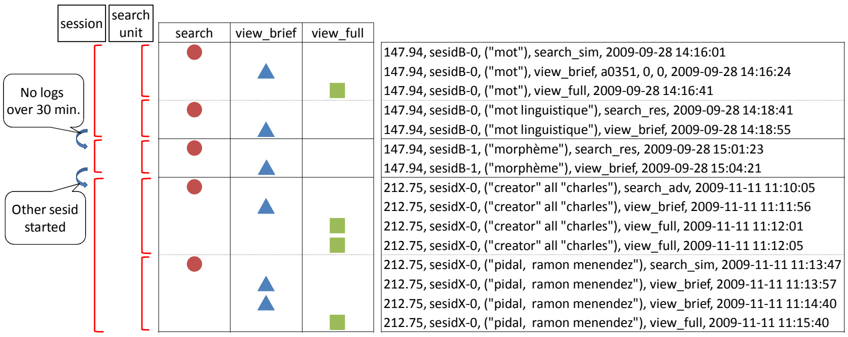
3.3 Unit of Analysis: Search Sessions and Search Units

For the unit of the analysis, we can define the four levels of the search processes as shown in the Table 2

Table 3. Four Levels of Search Processes

Levels	Definition
Search Task	The overall process to complete the search task. The concept of the search task is similar to the search trail concept of White and Drunker (2007). The range of the search trail is broader than the search task.
Intent Unit	Continuous process while searching for the same target. The concept of the intent unit is similar to that of the search mission by Guo and Agichtein (2009).
Search Unit	Continuous process while searching a single query. A search unit ends when users submits new query.
Link Unit	Continuous process while linking non-search results pages. A link unit starts when the user click a link in SERP and ends when he or she returns to SERP.

In this paper, we focused on the **search session**, which is closed to the **task unit** in Table 3 and the **search unit**. A search session is a unit for search [12] and a series of queries by a user [13] for a task. In this analysis, a series of log data containing the same session ID (*sesid*) was regarded as a series of the actions by a same user in a that the different session started. When sesid was null, the session was identified



```
147.94, sesidB-0, ("mot"), search_sim, 2009-09-28 14:16:01
147.94, sesidB-0, ("mot"), view_brief, a0351, 0, 0, 2009-09-28 14:16:24
147.94, sesidB-0, ("mot"), view_full, 2009-09-28 14:16:41
147.94, sesidB-0, ("mot linguistique"), search_res, 2009-09-28 14:18:41
147.94, sesidB-0, ("mot linguistique"), view_brief, 2009-09-28 14:18:55
147.94, sesidB-1, ("morphème"), search_res, 2009-09-28 15:01:23
147.94, sesidB-1, ("morphème"), view_brief, 2009-09-28 15:04:21
212.75, sesidX-0, ("creator" all "charles"), search_adv, 2009-11-11 11:10:05
212.75, sesidX-0, ("creator" all "charles"), view_brief, 2009-11-11 11:11:56
212.75, sesidX-0, ("creator" all "charles"), view_full, 2009-11-11 11:12:01
212.75, sesidX-0, ("creator" all "charles"), view_full, 2009-11-11 11:12:05
212.75, sesidX-0, ("pidal,  ramon menendez"), search_sim, 2009-11-11 11:13:47
212.75, sesidX-0, ("pidal,  ramon menendez"), view_brief, 2009-11-11 11:13:57
212.75, sesidX-0, ("pidal,  ramon menendez"), view_brief, 2009-11-11 11:14:40
212.75, sesidX-0, ("pidal,  ramon menendez"), view_full, 2009-11-11 11:15:40
```

Figure 2; Example of Analysis Units

search session. And if no action was recorded more than 30 minutes, it was regarded using the IP address (userip) . A search unit is a series of the actions, which starts by an action of "*search*" and ends by the next "search" or by the end of the search session. Figure 2 showed the examples of the search sessions and search units on TEL action logs.

In Figure 2, the original session IDs like "sesidB" or "sesidX" (on the second column) are replaced by the "cres-sesid" such as "sesidB-0" and "sesidB-1". The different numbers like "-0" or "-1" are added to the sesidB and divided into two different sessions because there was a 42-minutes gaps between the 5th and 6th lines.

3.4 Click ranks

TEL's action log item, "recordPosition" along with a "view_full" action was identified and analyzed the click ranks.

4. Results and Discussion

4.1 Search Sessions

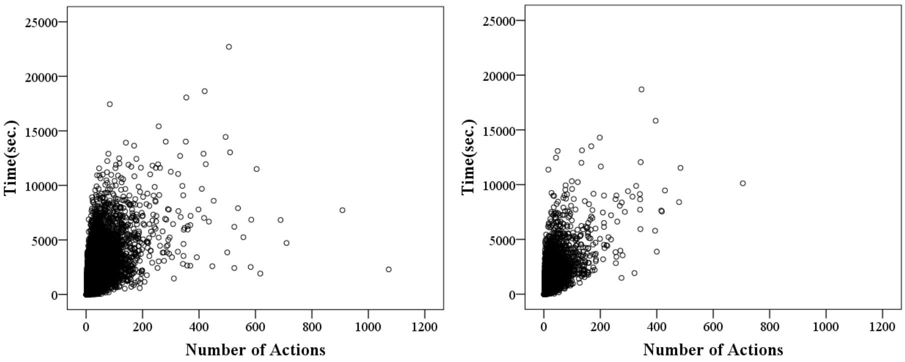
The number of actions in each search sessions and the time duration of it are shown in Table 4. No significant differences were found between the two log files of dataset1 and dataset2. As both of the numbers of actions in a search session and its time duration had lower numbers even for the third quartiles and were indicated that most of the search sessions have a few actions in the rather short time duration. And rather small number of the exceptional cases of longer sessions.

Table 4: Nnumber of Actions and Duration Per Search Session

(n)	Mean.	SD	Min.	Q1	Median	Q3	Max.
No. of actions							
dataset1 (225,590)	6.92	13.34	1	2	4	8	1,072
dataset2 (41,003)	7.32	16.25	1	1	3	7	705
Time(sec.)							
dataset1 (225,590)	282.14	653.94	0	10	73	241	22,706
dataset2 (41,003)	315.94	750.48	0	0	65	263	18,697

Q1:1st qurtile, Q3:3rd qurtile

The correlation were found between the duration and the number of the actions of each search session (dataset1: r =0.665 , p < .01, dataset2: r =0.688 , p < .01) for Figure 3a and 3b. There are still rather small number of the cases with larger number of the actions in the shorter duration and the opposite cases.



(a) dataset1 *(b) dataset 2*

Figure 3: Scatter-grams of Number of Actions and Duration Per Search Session.

In the previous studies on Web search, the duration of the search sessions are related to the types of the search tasks and the purposes of the search. The search conducted by expert users (*i.e.*, the experts in search techniques, in the topics, and the problem solving of the type of the tasks) on the tasks with well-defined problems to the users getting larger number of actions in the same durations. In contract, the novice users or the users who search on the unfamiliar topics or on the ill-defined problems for the user did rather small number of the search and other actions per times and required longer durations. Providing the supports to the users for the terms of search tactics, for topics or subject domains, or for the task expertise is preferable.

4.2 Search units

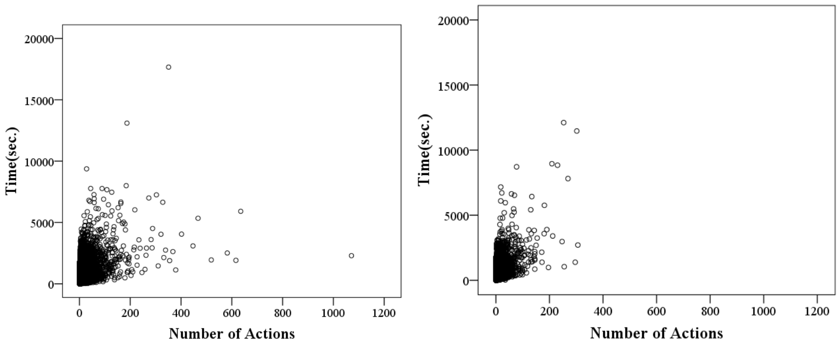
The number of the search units, the time duration of each search unit, and the number of actions in it are shown in Table 5. No significant differences were found between dataset-2 and -1 and dataset2. For the most search units, are only one or two action(s) are followed by a search and 75% search units are shortended in 2 minutes.

Table 5: Number of Actions and Duration Per Search Unit

	(n)	Mean.	SD	Min.	Q1	Median	Q3	Max.
No. of actions								
dataset1	(492,313)	2.63	5.27	1	1	2	3	1,071
dataset2	(79,830)	2.23	6.73	1	1	1	3	306
Time(sec.)								
dataset1	(492,313)	121.22	259.50	0	12	44.00	109	17,661
dataset2	(79,830)	150.78	326.75	0	6	43.70	129	12,112

Q1:1st qurtile, Q3:3rd qurtile

The correlation between the duration and the number of the actions of each search unit is found in (Figure 4a and 4b) for both datasets. (dataset1: r = 0.410, p < .01, dataset2: r = 0.508, p < .001). The duration of the search units increased according to the number of the actions in them. But there are rather small number of the cases with larger number of the actions in the shorter duration and the longer duration with smaller number of actions. Qualitative analysis of these exceptional cases shall conducted later.



(a) dataset1 *(b) dataset 2*

Figure 4: Scatter-grams of Number of Actions and Duration per Search Unit

4.3 Click Ranks

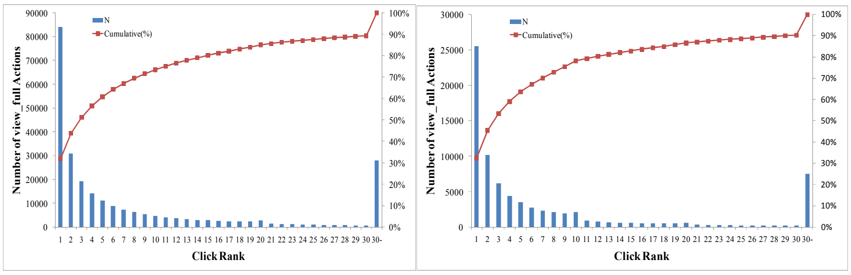
The numbers of the click ranks and their histgrams are shown in Table 6 and Figures 5a and 5b.

Table 6: Click Ranks

	(n)	Mean.	SD	Min.*	Q1	Median	Q3	Max.
dataset1	(262,883*)	17.63	66.59	0	1	3	11	9,987
dataset2	(78,520)	15.61	100.05	1	1	3	9	19,999

Q1:1st qurtile, Q3:3rd quartile

*: As explained in Section 2, about half of the records in the dataset1 missing the value of the recordPosition, which is equivalent to the click rank.



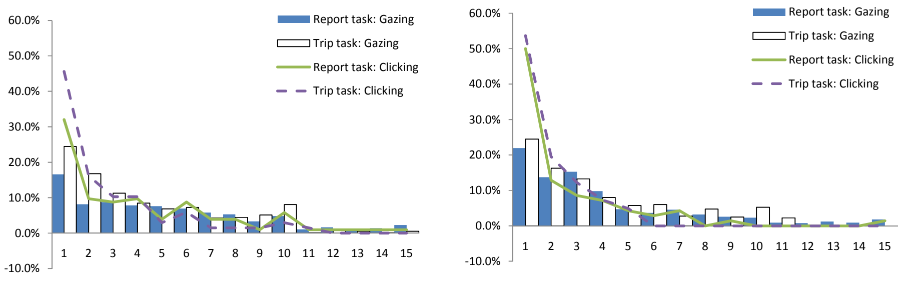
(a) *dataset1* *(b) dataset 2*

Figure 5: Histograms of Click Ranks

For both datasets, the number of the clicks on the top ranked documents on the list of the retrieved results on the brief view page was highest and a big gap between ranks of 1 and 2 were found. About 30 % of the total clicks were on the top-ranked retrieved documents.

For dataset1 has a small gap between the 20th and 21st, and dataset2 has a small gap between the 10th and 11th. The current TEL's user interface displays the top 10 documents on the first page. This affected on the users behaviors and the number of the clicks on the second pages declined. For dataset1, we guess that the TEL might have a user interface which listing the top 20 documents on the first pages sometimes during the period for the dataset1.

In the past research on the web searches, the distributions of the click ranks are suggested the relationship with the types of the tasks that the users are involved and the users' experiences in various ways.



(a) Undergraduate

(b) Graduate

Figure 6. Gaze and Click Ranks in Web Search by Different User Groups for Different Tasks [9]

As shown in Figure 6, the results of the user experiments on Web search showed the similar tendencies in the click rank distributions, and more acute concentrates [1][9]. About 50% of the clicks searches by graduate students were rank=1, and about 45% and 30% of the total clicks done on SERP in the searches conducted by the undergraduate students were rank=1 for trip-planning and for report-writings, respectively. In terms of the distribution of the click ranks, the TEL's distribution observed in the action logs were rather similar to those for the Web searches by undergraduate students who were not well-experienced in the search for report-writing. The structures of the systems of TEL and ordinary web search engines which allow the users to navigate among the web pages far from SERP are different, and the contents and users' information seeking tasks related to the searches could be also different, and then we cannot conclude the relationship between them for here. Further analysis shall be done for the different user groups and different search sessions to investigate across the search sessions using different languages, and those showed characteristic number and duration for search sessions and search units.

5. Conclusion

This paper reports the results of our analysis of the TEL action logs using the frameworks extended from the authors' investigations on the users' search behaviors on the web with different search tasks by the users who have different levels of expertise for search techniques, for subject domains or topics, and for the task. We have focused on the analysis of the search process and analyzed the action logs in terms of search sessions, search units and click ranks. Unfortunately the most of the search sessions and search units are shorts and we could not find any particular user groups from the analysis so far. The analysis of the click-ranks indicating the changes in the user interfaces sometime between the first log file. Also the most frequent clicks were done on the top ranked documents, but those tendencies are more moderate comparing to the Web searches.

For the further research, the analysis on the search sessions across the different languages and qualitative analysis of the search sessions and units which obtained characteristic behaviors in terms of the number of the actions in the search sessions and units and their durations. We hope to report some of the results of such additional analysis at the workshop in September.

References

1. Terai, H., Saito, H., Takaku, M., Egusa, Y., Miwa, M. and Kando, N. (20008) Differences between informational and transactional tasks in information seeking on the Web. In Proceedings of IIiX 2008, 152-159
2. Egusa, Y., Takaku, M., Terai, H., Saito, H., Miwa, M. and Kando, N. (2010). Link Depth: Measuring how far searchers explore Web, *Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS 2010)* (p.8). Kauai: IEEE
3. Egusa, Y., Saito, H., Takaku, M., Terai, H., Miwa, M. and Kando, N. (2010). Using a concept map to evaluate exploratory search, In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX 2010)* (10p.), New Brunswick, NJ.
4. Miwa, M. and Kando, N. (2006). Role of naïve ontology in search and learn processes for domain novices. In *Digital Libraries: Achievements, Challenges and Opportunities: the Proceedings of the 9th International Conference on Asian Digital Libraries, ICSADL 2006* Kyoto, Japan, November 2006, (pp. 380-389). Berlin, Germany: Springer LNCS
5. Miwa, M., and Kando, N. (2007). Naïve ontology for concepts of time and space for searching and learning, *Information Research*, 12(2), from http://informationr.net/ir/12-2/paper296.html
6. Miwa, M. and Kando, N. (2007). Methodology for capturing exploratory search processes. *Proceeding of the ACM SIGCHI 2007 Workshop on Exploratory Search and HCI : Designing and Evaluating Interfaces to Support Exploratory Search Interaction (ESI2007)*; pp.76-80
7. Egusa, Y., Takaku, M., Saito, H., Terai, H., Miwa, M. and Kando, N. (2008) Visualization of user eye movements for search result pages, In *Proceedings of the Second International Workshop on Evaluating Information Access (EVIA 2008)* (NTCIR-7 Pre-Meeting Workshop); pp.42-46
8. Saito, H., Terai, H., Egusa, Y., Takaku, M., Miwa, M. and Kando, N. (2009). How task types and user experiences affect information-seeking behavior on the web: Using eye-tracking

and client-side search, *Workshop on Understanding the User (UUIR 2009) (ACM SIGIR 2009 Workshop)*. Boston: ACM

9. Saito, H., Takaku, M., Egusa, Y., Terai, H., Miwa, M. and Kando, N. (2010) Connecting qualitative and quantitative analysis of Web search process: Analysis using Search Units. In *Proceedings of the 4th Asian Information Retrieval Societies' Conference (AIRS 2010)*, Taipei (to appear)
10. White, R.W. and Drucker, S.M. (2007). Investigating behavioral variability in web search, *Proceedings of the 16th international conference on World Wide Web* (WWW 2007) (pp. 21-30). Banff: ACM
11. Guo, Q. and Agichtein, E. (2009). Beyond session segmentation: predicting changes in search intent with client-side user interactions, *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in information Retrieval* (SIGIR '09) (pp. 636-637). Boston: ACM.
12. Jansen, B.J. and Pooch, U.W. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3), 235–24613
13. Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999). Analysis of a very large Web search engine query log, *SIGIR Forum*, 33(1), 6–12