Search Behaviour Before and After Search Success

Mengdie Zhuang
University of Sheffield
Sheffield, UK
mzhuang1@sheffield.ac.uk

Elaine G. Toms
University of Sheffield
Sheffield, UK
e.toms@sheffield.ac.uk

Gianluca Demartini
University of Sheffield
Sheffield, UK
g.demartini@sheffield.ac.uk

ABSTRACT
Why do users continue searching after reviewing all relevant documents with which they could have completed a work task? If we knew the answer, then a search system may be able to help users learn about their current search processes, which in turn may enable them to make the whole search process more efficient, leading to greater effectiveness and user satisfaction. This paper is a first step towards solving this problem. Using a previously collected data set, we identified the point of success and hence task completion, and investigated the search behaviour before and after users had accessed all relevant documents for answering assigned tasks. We used a set of search behaviour actions derived from Marchionini's (1995) Information Seeking Process model, and modeled the distribution of these actions throughout the entire search process, comparing actions before and after success could have been attained. Our results suggest that six defined actions, namely user-submitted query, system-suggested query, forward to items, evaluate relevant items, reflect, and answer appeared to change according to the stage of the entire search process. Also, users have notably distinct patterns before and after search success was obtained, but not realised by the user. Not all action were affected; user-submitted query and system-suggested query appeared to be unaffected by time in post-success case and pre-success case, respectively.

CCS Concepts
• Information systems~Users and interactive retrieval

Keywords
Evaluating search behaviour; Search process; Search success

1. INTRODUCTION
Do search systems know when a user achieves search success? Do users know when they have all of the raw materials, i.e., relevant documents, to complete the task? There is no simple answer. Search success has been characterised mainly by criteria from two sources: system performance and user satisfaction. The early Cranfield experiments [3] defined success according to system performance, with indicators including ranking of relevant documents, precision, recall [5]), query effectiveness [9] and accessibility to relevant documents (see examples in [1; 4]), and time spent finding relevant pages. But these did not include extracting the correct answer [1].

Search success as identified by system performance does not always correlate with real-world cases. The system does not know in advance whether the end-user could use or regard the highly ranked documents the way that systems 'think' they should be valued. To assess how users think about the result requires user response, which is measured either during search (e.g. think aloud [7]), or after the search (e.g. interview and questionnaire [5]). In the user-oriented context, the terminus of a search process is when the user feels either relieved/satisfied or disappointed [6], regardless of whether the documents retrieved are relevant to completing the task. Search success in this case is when the user identifies relief/satisfaction.

The two criteria mentioned above are highly distinctive; the point in time when the system has succeeded (i.e., delivered the relevant documents according to the user query) and the user's perception of success often differ from one another [13]. User success, which is assessed at the end of the search, comes later than system success [13], as users spend more time after retrieving relevant documents to confirm their findings. How can the system alert the user about the current status of relevant documents? Helping users learn that they have reviewed all documents needed for their current information needs or indeed that all relevant documents have been identified would reduce the delay between these two types of search success. Thus instructing users that they could stop once the system has done its job and concentrate on using the retrieved documents for the particular task. Enabling so will improve efficiency and effectiveness, and lead to user satisfaction.

In this research, we investigate search behaviour before and after the point of search success, that is, when all relevant documents needed to respond to the task have been retrieved. In this initial analysis we examine whether there are differences in search behaviour while the user is searching for relevant documents, and after all documents have been retrieved, which we call pre- and post- search success.

2. METHODS
2.1 Dataset
The dataset we used was collected in a previous study [13]. Here is a brief description (see details in [13]):

System. wikiSearch system [11] embedded into the WiIRE (Web Interactive Information Retrieval Experimentation) [10].

Participants. 381 participants were recruited via mailing list.

Tasks. Users were assigned 3 out of 12 different tasks. Based on a set of criteria, the users were asked to decide between two options in each task.

Protocol. Participants were led only by the system throughout the procedure. Before starting the task, each participant was assigned 3 of the 12 designed tasks (see details in [13]) with a unique identification number in order to track their search activities. Each of the participants first had an introduction of the study and filled in the Consent Form and the Demographics Questionnaire. Then

Search as Learning (SAL), July 21, 2016, Pisa, Italy

they were given a tutorial on how to use the system and were allowed to self-practice. After the practice, participants went into the task session. For each task, there was a pre-task questionnaire and post-task questionnaires; data from these questionnaires were not included in this paper. Once the participant had submitted their decisions on all three tasks, they were presented a "Thank you" page on the screen.

Data Preparation. Log files recorded all interactions between participants and the system. Within 1143 cases (381 participants completed 3 tasks each), 86906 behaviour records with time-stamp and participant identification number were extracted from the log files.

2.2 Search Action Extraction

The set of behavioural action cases were grouped based on the pair of participants' id and task id. Each set of grouped behaviour records represents all the search actions one participant conducted in one task, and we refer to it as the search process of {participant id, task id}. As defined by the experimental interface [13], 15 scenarios (Table 1) were generated to describe the actions participants performed during the search sessions. In order to gain an understanding from a conceptual level and to reduce the noise caused by multiple action types, we generated 9 actions (Table 1) based on the Information Seeking Process model [8]. In the present study, we extracted features of the sub-process "examining results" into two backward motions, two forward motions [14], and evaluated relevant items: "formulating and executing query" into two query actions; and "extracting information and reflect" into answer and reflect.

In the system interface [13] for the in-lab experiment, all functions were presented and contained in only one single window to control the search process, and eliminate conflicting variables, such as the execution of multiple simultaneous user tasks. Thus the experimental control using a single task: 1) gave users clear instructions; 2) provided a simple user interface and eliminated the labyrinth typically introduced by search system, allowing users to focus on actions of interest, which would lead to a simplified, holistic model.

2.3 Time points

In this paper, we used two points in time within the search process of a single task:

- **System success:** we manually identified the system success point in each search process (see a detailed description in [13]). First, "Answer Sets" were created; each set contains all the relevant documents that are needed to be used to provide a possible answer for each task. No set is a subset of another. Each task had 1-5 relevant answer sets, representing the different approaches that could be taken to find an answer, and no single document could provide a complete answer. The point at which a participant has reviewed all documents in one answer set for the current search task is considered the point of "system success". In this paper, we define the time period from when user starts a task to

Table 1. Actions and Characters assigned to scenarios with descriptive statistics of the time spent on related actions. Time spent on each action was calculated by the time between the start of such action and the start of the subsequent action, measured in seconds. Q1 and Q3 represent the first and third quartile of the sample, respectively (N=1143).

Action	Char	Description	***Mean***	Q1	Q3	***SD***	Scenarios
Answer tasks	A	Generating and submitting an answer.	71.9	3	109	115.3	• Click/type into the answer box
Backward to items	B	Checking items previously visited but not considered as relevant	20.2	0	0	72.1	• Click a item in history section
Backward to pages	b	Checking pages previously visited but not considered as relevant	9.6	0	0	45.6	• Click a previous viewed page
Evaluate relevant items	E	Re-examining items previously visited and considered as relevant again	48.3	0	31	109.3	• Click a item in bookbag
Reflect	R	Reflecting on items/pages.	124.5	22	185	129.2	• Add /delete/rate item to bookbag • Add/delete page into bookbag • Block/unblock an item or page
Forward to items	F	Checking items not previously visited	25.1	6	21	48.6	• Click on a item in results list
Forward to pages	f	Checking pages not previously visited	4.2	0	0	18.5	• Click on a new page in result list
User-submitted query	Q	Executing a new user-submitted query using query box	91.6	14	138	104.8	• Submit a query
System-suggested query	q	Executing a system-suggested query and get a related result list	48.9	2	56	85.0	• Click a query in history section, suggested by system. • Click a hyperlink in item article, which leads to a result list.

the point in time when system success was achieved as the pre-success case.

- **Search Terminus:** Search terminus in this study refers to the time point at which a user submitted an answer for a task, and stopped searching further. It is also the task terminus. We define the point in time from when the user reached system success point to the terminus as the post-success case. Participants completed a post-task questionnaire at the end of each task about their search experience, but these data are outside the scope of this analysis.

3. RESULTS AND DISCUSSION

3.1 Overview of Search Action

The 381 participants completed three tasks each, which resulted in 1143 action cases. In 39 cases (3.4%), participants did not review at least one relevant answer set. These were excluded, resulting in 1143 pre-success cases and 1104 post-success cases in total. The {participant id, task id} pair uniquely identifies each case, which contains nine possible actions. Table 1 presents descriptive statistics of the possible actions.

The interquartile range (IQR) is the difference between the 3rd quartile and the 1st quartile. We considered outliers the values that are either 1.5 IQR above the 3rd quartile, or 1.5 IQR below the 1st quartile [2]. We observed a large number of outliers in all backward motion and forward motion actions (B, b, F, and f). However, these four actions all account for a very small percent of the total time (2-6%). Particularly, the values of the 3rd quartile for B, b, and f are 0, indicating that most of the participants did not perform any action of these types at all. However, a few participants spent a very long time in both backward motion actions and forward to items (Max(B) = 939, Max(b) = 654, Max(F) = 566). The large number of outliers and the high maximum value suggest that general statistics used to measure time of these four actions, typically relying on mean value, may inaccurately represent the data sample. On the other hand, Action Query (Q), query (q), action answer tasks (A), action reflect (R), and evaluation of items (E) account for 20% (M = 91.6), 10% (M = 48.9), 16% (M = 71.9), 28% (M = 124.5), and 10% (M = 48.3) of the total time, respectively.

3.2 Pre-success case

From the 1143 pre-success action cases extracted from log files, we studied the distribution of each action observed during the pre-success case. Figure 1 (left half) shows the probability of an action, *P(Action)*, taking place at different time points of the pre-success case. We normalised the time scale to show the proportion of the whole pre-success case time frame. The total probability of all 9 actions is 1. Three of our defined backward motions and forward motions (f, B, b) and action E have extremely low probabilities across the entire pre-success case ($P < 0.001$), and almost no action A occurred except in those 39 tasks that did not achieve search success. We interpret this as mis-clicking (action A only happened for 1 or 2 seconds, and $P(A) < 0.02$), indicating that almost all the users who achieved search success in this task did not even attempt to generate answers before actually reviewing all the relevant documents. Therefore, these 5 actions, B, b, E, f, A, are not included in Figure 1 (left side).

The left end of Figure 1 (left half) reveals the probability of actions occurring right after a task began, while the right end of Figure 1 (left half) represents the probability of actions happening close to the search success point. As defined by the experimental design, action Q has an almost 100% probability at the beginning, when very few participants clicked the answer box. P(Q) drops as participants turned to check new items from the result lists, reflected by the sharp spike in P(F).

P(R) achieves a value above 0.4, and then keeps dropping slowly until just before the success point. P(F), on the other hand, has a sharp drop to around 0.05 from 20% to 80% of the pre-success case, and grew significantly just after 88% of the pre success case. As defined in this study, users who achieved success checked all the relevant documents, and so the action preceding the success point would be that they approached new items (F). P(q) is around 0.1 for the whole pre-success case with a slight decreasing trend until 90% of pre-success case, after which there is a slight upward trend until right before the end of pre-success case, suggesting that users accessed the last relevant but un-reviewed document from both user-submitted query (Q) and system-suggested query (q). With a shape complementary to P(F), P(Q) keeps increasing until 90% of the pre-success case.

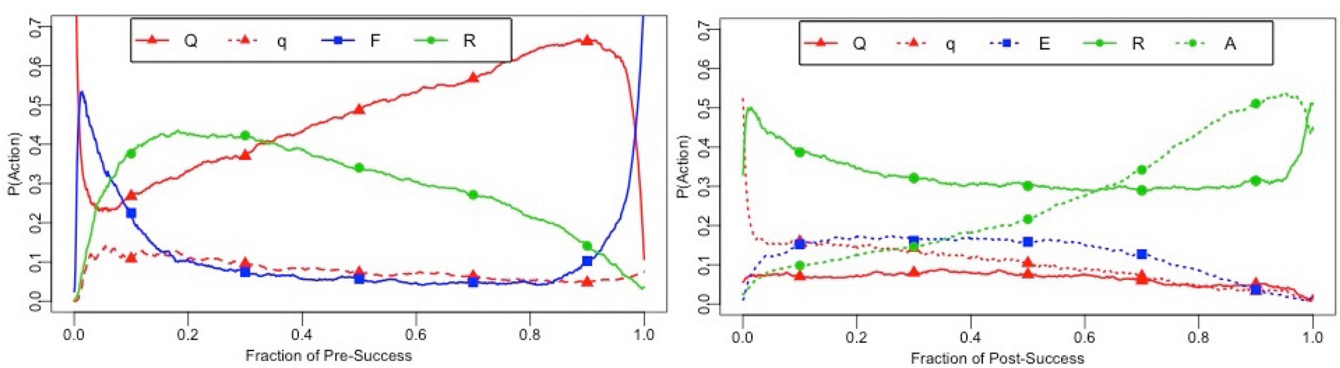

Figure 1. Temporal dynamics of search action: (left half) probability of selected actions in pre-success cases; (right half) probability of selected actions in post-success cases. The red line with triangular points represents actions Q and q of formulating and executing query; The blue line with square points represents actions F and E of examining results; The green line with round points represents action R and A of answer and reflect; The horizontal size of each plot represents fraction of pre/post-success case.

3.3 Post-success case

Figure 1 (right half) shows the probability of actions taken in 1104 post-success cases, with a normalised time scale rather than absolute time. Four of our defined backward and forward action types (B, b, F, f) have a flat and low probability across the entire post-success case ($P < 0.1$). These 4 actions (B, b, F, and f) are not included in Figure 1 (right half).

P(R) has a value above 0.28 across the entire post-success case, which is the highest among all actions, indicating that users reflected, based on their viewings, consistently more frequently than in the pre-success cases. It peaks right after the success point, and has a smooth decrease until 90% of the post-success case, followed by a rise to 0.52 at the end of the search session. This demonstrates that users tend to reflect on their findings before ending their search session. Compared to the extremely low P(A) in the pre-success case, P(A) increases steadily throughout different stages of the post-success case, before an eventual drop near the very end. This indicates that not all users waited until the very end of the case to give an answer, and they continued with the case after A.

P(q) appears to be above 0.5 at the very left end of Figure 1 (right half) with a quick sharp drop at 0.15, and slowly decreases until the end. In the entire search session, P(q) only has a peak above 0.5 right after the success point. This value may be exaggerated by normalising the time scale, but still shows that a portion of users clicked a query from the history list or clicked a hyperlink after they reviewed the last relevant item. The probability of system-suggested query (action q) is larger than submitting a user-submitted query (Q) for the first 77% of the post-success case. In the pre-success case, P(Q) is always larger than P(q). P(Q) in the post-success case has a relatively flat curve below 0.1.

P(E) in Figure 1 (right half) (value>0) indicates participants returned to reviewed items in the post-success case, which is rarely observed in the pre-success case.

4. CONCLUSIONS AND FUTURE WORK

The key objective of this work was to assess whether search behaviour pattern changes before and after users reach system success. We analysed nine search actions from a previously collected dataset of 381 users completing search tasks using a search system. The 15 observable behavioural scenarios were categorised into 9 search actions based on the Information Seeking Process model [8], and then the distribution of each action was observed in pre- and post- success cases with a normalised time scale.

Our results show that six defined actions: user-submitted query, system-suggested query, forward to items, evaluate relevant items, reflect, and answer, appear to change according to the portions of the search process, as well as being notably different between pre- and post- success case types. However, three defined actions, backward to items, backward to pages, and forward to pages, have extremely low probability across the entire search process. In addition, user-submitted query and system-suggested query appear to be unaffected by time in post-success case and pre-success case, respectively.

In a previous research, Toms, Villa and McCay-Peet (see Figure 2 in [13]) using the same dataset observed that the average system success was achieved before 40% of the entire search process was completed. Moreover, in 78% of the tasks, participants did not click on any new, unseen documents after they reached the system search success point. But, users committed significant time and effort after attaining system success. As mentioned above, user behaviour changed before and after the system success. However, to our knowledge, current mainstream systems do not examine, assess and exploit changes in behaviour to intervene in the search process to notify that no more new relevant information is available (based on queries submitted) so that users can stop searching and start interpreting the information retrieved. In other words, current systems do not help the user to learn about the current search; they simply exacerbate the problem by repeating search results.

To obtain further insights into why users continue to search after system success and how to design system functions to instruct users about the search process, we are currently working on extracting key sub-sequences in pre/post-success cases, and on examining the indicators. This could potentially augment the development of a system function to help user learn about their search progress, and will further improve efficiency of search process evaluation that would benefit both users and search engine developers.

5. REFERENCES

[1] Ageev, M., Guo, Q., Lagun, D., and Agichtein, E., 2011. Find it if you can: a game for modeling different types of web search success using interaction data. *Proc. SIGIR'11*, 345-354.

[2] Bamnett, V. and Lewis, T., 1994. Outliers in statistical data.

[3] Cleverdon, C.W., 1960. The aslib cranfield research project on the comparative efficiency of indexing systems. In *Aslib Proceedings*, 421-431.

[4] Hassan, A., Jones, R., and Klinkner, K.L., 2010. Beyond DCG: user behavior as a predictor of a successful search. *Proc. WSDM'10,* 221-230.

[5] Kelly, D., 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval 3*, 1—2, 1-224.

[6] Kuhlthau, C.C., 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science 42*, 5, 361.

[7] Lewis, C., 1982. *Using the" thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center.

[8] Marchionini, G., 1995. *Information seeking in electronic environments*. Cambridge University Press.

[9] Shah, C., Hendahewa, C., and González-Ibáñez, R., 2015. Rain or shine? forecasting search process performance in exploratory search tasks. *Journal of the Association for Information Science and Technology*.

[10] Toms, E.G., Freund, L., and Li, C., 2004. WiRE: the Web interactive information retrieval experimentation system prototype. *Information Processing & Management 40*, 4, 655-675.

[11] Toms, E.G., Mccay-Peet, L., and Mackenzie, R.T., 2009. wikiSearch: From Access to Use. In *Research and Advanced Technology for Digital Libraries* Springer, 27-38.

[12] Toms, E.G., Villa, R., and Mccay-Peet, L., 2013. How is a search system used in work task completion? *Journal of information science 39*, 1, 15-25.

[13] Toms, E.G., Villa, R., and Mccay-Peet, L., 2013. How is a search system used in work task completion? *Journal of information science*.

[14] White, R.W. and Dumais, S.T., 2009. Characterizing and predicting search engine switching behavior. *Proc. CIKM'09*, 87-96.