The Edge Hill Contribution to the INEX Interactive Social Book Search Task

David Walsh[1] and Mark Hall[1]

Edge Hill University, St Helens Road, Ormskirk, L39 4QP, United Kingdom,
{David.Walsh, Mark.Hall}@edgehill.ac.uk

Abstract. In our contribution we use log-analysis to investigate whether participants in the INEX Interactive Social Book Search are able to use the new *multi-stage* interface and whether it provides any benefits over the traditional IR *baseline* interface. Our initial results show that participants are able to successfully use the new *multi-stage* interface, with no significant learning effects. Additionally, for the *non-goal* task, the *multi-stage* interface actually enables the participants to collect more books than when using the *baseline* interface.

Keywords: human computer information retrieval, user study, log analysis

1 Introduction

The CLEF[1] INEX[2] track's Interactive Social Book Search task gathered data from users using one of two interfaces to complete two tasks. The *baseline* interface implemented a standard Information Retrieval (IR) interface [3] consisting of a search box, a search result list, and an individual item display. The second interface (*multi-stage*) attempted an implementation of Kuhlthau's multi-stage search process[4], filtered through Vakkari's simplification of the model [5]. Two tasks were tested, the first an *non-goal* task where participants were asked to look for any book they might find interesting, and a *goal-oriented* task where participants were asked to find books for a given topic ("laymen books on mathematics and physics"). Each participant completed both tasks in one of the two interfaces. Task order was automatically balanced to avoid ordering bias.

We investigated the following three research questions:

1. **RQ1**: Does the *multi-stage* interface enable the participants to explore and find a larger number of books?
2. **RQ2**: Does the *multi-stage* interface have an additional learning time?
3. **RQ3**: Do participants make use of all three stages in the multi-stage *multi-stage* interface?

[1] Conference and Labs of the Evaluation Forum
[2] INitiative for the Evaluation of XML retrieval

2 Time Spent in the System

The first analysis focused on how long the participants spent using the system in order to determine whether there were any differences between the two systems and tasks. The experiment system [2] automatically measured the time taken on the *non-goal* and *goal-oriented* tasks for every participant and the main analysis is based on this data. For the three stages implemented in the *multi-stage* interface, the log data acquired by the IR system [1] was processed to determine how long each participant spent using the each of the three stages ("explore", "focus", and "refine").

The first step in the analysis was to determine if the task ordering impacted the time spent on either of the tasks. Wilcox signed rank tests were used to compare the task times for all interface and task combinations, showing no significant differences in task duration for any of the combinations. For the *multi-stage* interface, the time spent in each of the three stages was also compared using a Wilcoxon rank-sum test for both ordering conditions, and also showed no significant differences in times in the two stages. For the remainder of the analysis, the task order can thus be ignored and the times for the two order conditions aggregated.

Table 1 shows the task times for the two interfaces and tasks. The data seems to indicate that participants are faster with the *multi-stage* interface for both tasks and that within the interfaces, participants are faster to complete the *non-goal* task than the *goal-oriented* task. However, for neither of these conditions does a Wilcoxon rank-sum test show significant differences. Thus for the remaining analysis presented here, we can assume that any differences in participant performance are due to the task or interface and not due to the time the participants spent on the task or interface.

Table 1. Task durations for the *baseline* and *multi-stage* interfaces for both the *non-goal* and *goal-oriented* tasks. All times are in seconds and formatted "median (1st quartile, 3rd quartile)".

Interface	***non-goal* task**	***goal-oriented* task**
baseline	217 (109, 334)	385 (142, 436)
multi-stage	160 (109, 490)	215 (148, 412)

2.1 *Modern* Interface Phase Times

In the *multi-stage* interface participants were able to switch between three stages ("explore", "focus", and "refine"). Table 2 shows the time spent in each of the three stages for the two tasks. Wilcoxon rank-sum tests were used to test for ordering effects. There are no ordering effects for the time spent in the *explore*

and *focus* stages, but there is an ordering effect in the *refine* stage. For the *non-goal* task, the time spent in the *refine* stage is longer, if it is the second task ($p = 0.012$). No ordering effect was shown for the *goal-oriented* task.

Table 2. Time spent in each of the three phrases available in the *multi-stage* interface. All times are in seconds and formatted "median (1st quartile, 3rd quartile)".

Task	***explore* stage**	***focus* stage**	***refine* stage**
non-goal	54(26.75, 73)	91.5 (52.5, 359)	0 (0, 5.75)
goal-oriented	54 (30.25, 79)	122.5 (73.25, 353)	0 (0, 18)

The times shown in Table 2 follow similar patterns for both the *non-goal* and *closed* tasks. Participants spent slightly less than a minute using the *explore* stage, and then spent between one and a half and two minutes on the *focus* stage. Only a small fraction of participants used the *refine* stage at all and those that did, did so only very briefly.

Considering RQ3, participants obviously do not use the final *refine* stage, either because they did not notice the stage in the user interface or because the label "Refine" did not clearly state what functionality would be available. Without looking at the participants qualitative responses it is impossible to determine the cause. However, the use pattern for the first two stages is as expected, with participants first spending time in the *explore* stage gaining an overview and then using the *focus* stage.

3 Books Collected

To investigate RQ1 we looked at the number of books participants added to their book-bag and also how quickly they added the first book. For RQ2 we also investigated which of the three stages participants added books from.

3.1 Total Number of Books Added

The total number of books each participant added to their book-bag was determine using a manual analysis of the log-data. For the *baseline* interface, the number of books added to the book-bag was counted and any books that were subsequently removed from the book-bag subtracted from that count. For the *multi-stage* interface, the same process was applied, but book counts were separated according to which of the stages the books were added from.

The resulting data-set was checked for task ordering effects using Wilcoxon rank sum tests and no significant ordering effects were found for any of the interface / task combinations. In the *multi-stage* task, the same checks were applied to the more detailed data and only for the *explore* stage in the *goal-oriented* task, was there a significant ordering bias. If the *goal-oriented* task was

the second task, then significantly fewer books were added to the book-bag in the *explore* stage (Wilcoxon signed rank, $p = 0.035$). As there were no overall ordering effects, the further analysis did not take task ordering into account.

Table 3 shows that significantly more books were added in the *non-goal* task using the *multi-stage* interface than using the *baseline* interface (Fig. 1, Wilcoxon signed rank test, $p = 0.011$). No such effect is visible in the *goal-oriented* task. This seems to indicate that the *multi-stage* interface provides significant benefit to the user when they do not yet have an explicit goal that they are searching for. At the same time, the *multi-stage* interface does not impact the performance when the user has an explicit goal in mind.

Table 3. Median number of books added to the book-bag. Numbers are "median (1st quartile, 3rd quartile)".

Interface	***non-goal* task**	***goal-oriented* task**
baseline	1 (0, 2)	3 (1, 4)
multi-stage	2 (1, 4)	3.5 (2, 5)

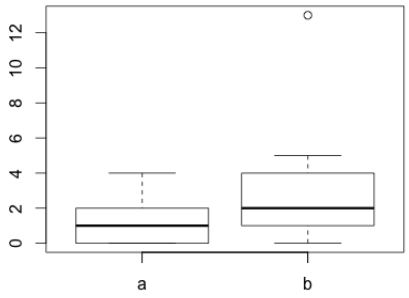

Fig. 1. Box-plot showing the number of books added to the book-bag for the *baseline* (a) and the *multi-stage* (b) interface in the *non-goal* task.

3.2 *Modern* Interface Details

To investigate RQ1 further we used the dataset from the previous section, but now looked in detail at the number of books added in the three stages of the

multi-stage interface (Tab. 4). Task ordering effects were investigated and no significant effects were found.

Table 4 clearly shows that the refine stage was not used to add any books, which is in line with the timing results that showed that the *refine* stage was essentially not used. Interestingly, although the participants spent more time in the *focus* stage, they collected more items in the initial *explore* stage. While the data seems to indicate that in the *goal-oriented* task participants collected more books in the *explore* stage, the effect is not significant.

Table 4. Median number of books added in each stage of the *multi-stage* interface. All results are formatted "median (1st quartile, 3rd quartile)".

Task	***explore* stage**	***focus* stage**	***refine* stage**
non-goal	1 (1, 2)	0 (0, 1)	0 (0, 0)
goal-oriented	2 (0, 2)	0 (0, 2)	0 (0, 0)

3.3 Time to Collect the First Book

To investigate RQ2, we analysed how quickly participants added their first book in each task. The log was manually analysed and the time between the session start and when the time at which the first book was added to the book-bag determined.

Table 5 shows the median times to collect their first book. While it looks as if the *modern* interface enables the participant to find the first book faster, the difference is not statistically significant.

Table 5. Time to add first item. All times are in seconds and formatted "median (1st quartile, 3rd quartile)".

Interface	***non-goal* task**	***goal-oriented* task**
classic	113.50 (76.25, 225.00)	112.0 (54.0, 164.5)
modern	101 (50, 271)	88.5 (58.0, 175.8)

4 Interaction Patterns

The final analysis looked at the interaction patterns, using a user-interaction bi-gram analysis. To create the interaction bi-gram distributions needed for the analysis, the log was processed in the following steps:

1. **Generate interaction string** – in the initial step each user-system interaction was mapped to a single letter. Using this mapping, for each participant and each of the participant's tasks a string representation of their interactions with the system was generated. Repeats of a single letter were reduced to a single letter;
2. **Create participant pattern distribution** – based on the interaction strings, all bi-grams were determined and bi-gram frequency distributions calculated;
3. **Aggregate distributions** – the participants' bi-gram distributions were aggregated into interface and task bi-gram distributions;
4. **Filter distributions** – the interface and task bi-gram distributions were filtered. All bi-grams that occurred fewer than three times were aggregated into a single value. This ensures that a potentially large number of interaction patterns that only occurred once or twice do not skew the results, while at the same time not completely loosing that data.

Before analysing the data in any more depth, potential ordering effects were investigated and only for the *goal-oriented* task with the *multi-stage* interface is there a significant difference in the interaction pattern distributions (χ^2 test, $p = 0.047$). As the significance is border-line and there is no significant differences in any of the other metrics tested, for the purpose of this analysis, the influence of task order will be ignored.

Comparing the two interfaces shows a significant difference in the bi-gram distributions between the *baseline* and *multi-stage* interfaces on the *goal-oriented* task (χ^2 test, $p = 0.036$). Looking at the bi-gram distribution (Tab. 6) shows that the main difference is at which point participants added books to their book-bag. In the *baseline* interface this primarily happened after the participants had viewed the book's detail (IA), while in the *multi-stage* interface it happens directly from the search results list (QI). The behaviour after adding an book to the book-bag is also different. In the *baseline* interface, the next action is to run another query (IQ), while in the *multi-stage* interface it is to view another book (AI).

Table 6. Top-ten most frequent bi-grams for the *baseline* and *multi-stage* interfaces. Major differences have been highlighted in bold. Actions: A – add book to book-bag; I – view a book; L – load a page; P – paginate; Q – run a query.

Interface	**AI**	**AQ**	**IA**	**IQ**	**LI**	**LQ**	**PI**	**QA**	**QI**	**QL**
baseline	1	**47**	**71**	24	10	30	23	4	56	7
multi-stage	**45**	16	40	19	10	26	36	**20**	46	13

Comparing the two tasks within both interfaces, shows a significant difference between the *non-goal* and *goal-oriented* tasks in the *multi-stage* interface (χ^2, $p < 0.001$), but none in the *baseline* interface. From the bi-gram distribution

of the *multi-stage* interface (Tab. 7), three differences stand out. In the *goal-oriented* task, participants made more use of the pagination functionality to see more items (PI) and also viewed more items after selecting a search facet (FI). In the *non-goal* task, participants more frequently viewed different bits of meta-data after viewing an item (MI).

Table 7. Top-ten most frequent bi-grams for the *non-goal* and *focus* tasks using the *multi-stage* interface. Major differences have been highlighted in bold. Actions: A – add book to book-bag; F – add a facet; I – view a book; L – load a page; P – paginate; Q – run a query.

Task	**AI**	**AL**	**FI**	**FQ**	**IA**	**IM**	**LF**	**MI**	**PI**	**QI**
non-goal	31	23	9	22	35	**65**	27	46	6	37
goal-oriented	45	20	**19**	21	40	37	15	50	**36**	46

The difference is in line with what would be expected, due to the task differences. In the *goal-oriented* task, participants use the facetting and pagination functionality to dig into the results, a pattern that is not so relevent when the task is *non-goal*. At the same time, in the *non-goal* task, participants interact more with the books' meta-data, as the participants use the meta-data to develop the search goal.

5 Conclusions

In conclusion the goal of the initial log analysis was to determine how participants used the *multi-stage* interface. The initial question was whether there would be a learning impact into the participants' peformance. The results clearly show that participants are able to use the new *multi-stage* interface just as well as the *baseline* interface and that there are no learning effects. For the *non-goal* task, the *multi-stage* interface even outperforms the *baseline* interface. Clearly, the initial *explore* stage, designed to support open-ended exploration, enables the user to explore better and thus collect more books.

Finally, both the considering the number of books collected and the time spent in the three stages of the *multi-stage* interface, participants clearly only make use of the first two stages. For the first two stages, the behaviour is as expected, with more time spent in the second *focus* stage, compared to the first *explore* stage. Interestingly, the majority of books were collected using the first *explore* stage, a result that needs more analysis. The use of the final *refine* stage requires further analysis, as it is clearly not used much and the reasons for this need to be investigated.

References

1. Hall, M.M., Katsaris, S., Toms, E.: A pluggable interactive ir evaluation workbench. In: European Workshop on Human-Computer Interaction and Information Retrieval. pp. 35–38 (2013), http://ceur-ws.org/Vol-1033/paper4.pdf
2. Hall, M.M., Toms, E.: Building a common framework for iir evaluation. In: CLEF 2013 - Information Access Evaluation. Multilinguality, Multimodality, and Visualization. pp. 17–28 (2013)
3. Hearst, M.A.: Search User Interfaces. Cambridge University Press (2009)
4. Kuhlthau, C.C.: Inside the search process: Information seeking from the user's perspective. JASIS 42(5), 361–371 (1991)
5. Vakkari, P.: A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. Journal of documentation 57(1), 44–60 (2001)