The Value of Multistage Search Systems for Book Search

Huurdeman, H.; Kamps, J.; Koolen, M.; Kumpulainen, S.

Publication date
2015
Document Version
Final published version
Published in
Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum

The Value of Multistage Search Systems for Book Search

Hugo Huurdeman[1,2], Jaap Kamps[1,2,3], Marijn Koolen[1,2], and Sanna Kumpulainen[1,2]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[3] ISLA, Faculty of Science, University of Amsterdam

Abstract. Often, our exploratory quests for books are highly complex endeavors which feature activities such as exploration, searching, selecting and comparing various books. Current systems for book search may not provide optimal support for this wide range of activities. The interactive Social Book Search Track investigates how users utilize different access interfaces in the context of two types of tasks, and evaluates a streamlined *baseline* interface and a rich *multistage* interface, potentially supporting different stages of search. In this paper, we analyze how these two types of interfaces influence user behavior, in terms of task duration, book selection and interaction patterns. Furthermore, we characterize the use of the different panels of the experimental multistage interface, as well as user engagement. We find initial evidence for the additional value of providing stage-based search support in the context of open-ended and focused book search tasks.

1 Introduction

The interactive Social Book Search (iSBS) Track studies how searchers use professional and user-generated metadata during different stages of complex search tasks. The iSBS track uses two experimental interfaces (a baseline and multistage interfaces), combined with open-ended and focused search tasks. Research groups participating in the iSBS track had to recruit at least 20 users for the shared study to gain access to the collected data from the experiment. In 2015, the second iteration of the iSBS track took place, and 7 teams recruited 192 participants for the study, resulting in a rich dataset.

This paper describes the University of Amsterdam's participation in this track, and we analyze the influence of task and interface on user behavior, in terms of task duration, book selection and interaction patterns. In addition, we characterize user engagement with both experimental interfaces.

2 Related Work

Previous work related to the iSBS track has been carried out in the INEX Interactive Retrieval Experiments (2004-2010) [6], the Cultural Heritage in CLEF

(CHiC) Interactive Track 2013 [8], and the INEX 2014 Interactive Social Book Search (ISBS) Track [3]. In these tracks, a standard procedure for collecting data was being used by participating research groups, including common topics and tasks, standardized search systems, document corpora and procedures. The system used for the iSBS track is a modified version of the one used for CHiC and is based on the Interactive IR evaluation platform developed by Hall and Toms [1], where different search engines and interfaces can be plugged into fully developed IIR framework that runs the entire user study [2].

3 Experimental Setup

In this section we describe the tasks, system interfaces and our pre-processing of the data generated by the experiment.

3.1 Tasks

The experiment includes two search tasks, a *focused* task and an *open* task, and each participant performs both. During the focused task, users were asked to compile a list of books, each matching a specified criterion. The *focused* task contains five sub-tasks, some of which are specific and some are more open:

> Imagine you participate in an experiment at a desert-island for one month. There will be no people, no TV, radio or other distraction. The only things you are allowed to take with you are 5 books. Please search for and add 5 books to your book-bag that you would want to read during your stay at the desert-island:
> - Select one book about surviving on a desert island
> - Select one book that will teach you something new
> - Select one book about one of your personal hobbies or interests
> - Select one book that is highly recommended by other users (based on user ratings and reviews)
> - Select one book for fun

> Please add a note (in the book-bag) explaining why you selected each of the five books.

The *open* task is derived from the non-goal task used in the iCHiC task at CLEF 2013 [9], which allows participants to come up with their own goals and sub-tasks. During the open task users could explore the collection based on their own interests, for as long as they wanted:

> Imagine you are waiting to meet a friend in a coffee shop or pub or the airport or your office. While waiting, you come across this website and explore it looking for any book that you find interesting, or engaging or relevant. Explore anything you wish until you are completely and utterly bored. When you find something interesting, add it to the book-bag. Please add a note (in the book-bag) explaining why you selected each of the books.

3.2 Interfaces

Two interfaces were developed for this study. The *baseline* interface is a standard search interface with a single screen, containing a query box, a left column with facet filters and a right column with a book bag in which users could store books they had selected. The *multistage* interface contains three screens, each with its own functionality to support different stages in the search process. It is inspired by various models of the information seeking process [5, 10], which indicate that users experience evolving stages of search in the context of complex tasks.

The *Browse* screen only allows browsing through predetermined categories (based on Amazon book categories), where the middle panel shows lists of book titles which users can click on to get detailed information on that book and the ability to save it to the bookbag.

The *Search* screen has a search box and search results in the middle panel, search filters based on the Amazon book categories and on user-supplied tags from LibraryThing users. By default, the detail-view of a search result shows a thumbnail and a publisher-supplied description of each book and four tabs that allow the user to switch between (1) the publisher-supplied description, (2) publication metadata, (3) Amazon user reviews and (4) LibraryThing user tags.

The *Book-bag* interface shows the bookbag in the left panel and per book a number of buttons allowing the user to search for similar books. There is a separate button for books by the same author, one for books with similar titles and one for books with the same subject categories. When clicking one of these buttons, the right panel shows the search results.

3.3 Data

The transaction log contains transactions of 192 users who completed both tasks, with 97 users using the baseline interface for both tasks, and 95 using the multistage interface.

The dataset collected during the iSBS study consists of questionnaire and logging data. In this paper, we focus on the system logs in section 4, while we look at the questionnaire data in section 5.

The log data includes the duration of each task, which we used to calculate task duration for each task and experimental interface. In the multistage interface, the user starts in the *browse* panel of the interface. Each time a user switches between interface stages (*explore*, *search* and *book-bag*), this is logged as an action. Using these switches, we can reconstruct all actions per interface stage. We found a very small number of impossible combinations (168 out of 22,152, such as adding a search filter in the *explore* stage, which has no filters). We surmise this is either because some stage switches were not logged or because a switch was logged but did not take place.

4 Analysis of Results

We compare the two interfaces based on a number of aspects: 1) task duration, 2) difference in book bag content between the two interfaces, and 3) difference

Table 1: Distribution of task length

Task	N	Min	Max	Median	Mean	St.dev.
focused	192	4.0	3916.4	696.5	855.5	625
open	192	4.0	8243.8	369.0	579.2	784
focused						
baseline	97	4.0	2510.5	630.9	751.5	553.0
multistage	95	112.2	3916.4	772.5	961.7	674.7
open						
baseline	97	4.0	2090.8	333.6	436.1	394.5
multistage	95	54.2	8242.8	439.1	725.4	1020.9

in the types of actions performed. Finally, we zoom in further on the multistage interface and look at the use of the different panels in the multistage interface.

4.1 Task Duration

First, we examine the duration of the included tasks. Differences can be expected at both the task and interface level. The *focused* task is more complex, so users might take longer to complete that task than the *open* task.

The distribution of task lengths in seconds is shown in Table 1. The majority of users spent less than 15 minutes on the task—the median is just under 12 minutes for the focused task (696.5 seconds) and just over 6 minutes (369 seconds) for the open task—but a few spent an hour or more (1 for the focused task and 3 for the open task). Due to such outliers the mean is higher than the median. The higher median and mean for the focused task is probably due to the higher complexity of the task, as it consists of 5 sub-tasks.

Also, differences in the task time between the baseline and multistage interface can be observed: the median task time for both the focused and the open task is higher for the multistage interface. Hence, participants spend more time in the multistage interface, regardless of the nature of the task.

4.2 Bookbag

Next, we analyse the content of the bookbag at the end of each task, to determine whether users show different book selection behaviour across the tasks and the two interfaces. Given the very different natures of the tasks, we expect to see clear differences between the bookbags after each task. The *focused* task asks the user to select books given a list of five criteria, which may steer the user to select five books. In the *open* task, users are instructed that they can select as many or as few books as they like. Therefore, we expect the number of books in the bookbag in the *open* task to be more widely distributed.

Indeed, Table 2 indicates that the number of books in the book-bags are higher for the focused task: on average, participants choose 4.75 books in the

Table 2: Statistics on the number of books in the bookbag at the end of each task

Task	N	Total	Min	Max	Median	Mean	St.dev.
Focused							
Baseline	97	442	0	6	5	4.6	1.2
Multistage	95	470	0	8	5	4.9	0.6
Open							
Baseline	97	309	0	13	2	3.2	2.7
Multistage	95	357	0	19	3	3.8	3.3

focused task, and 3.45 for the open task. Also, the standard deviation is substantially higher for the open task, so there is more variation in the number of books that participants selected.

Some of the sub-tasks of the focused task may be interpreted as similar to the open task, e.g. books about *one of your personal hobbies or interests* (third sub-task) and books *for fun* (fifth sub-task). If that is the case, user may simply add some of the same books to the book-bag in both tasks. We checked the overlap between the books in the book-bag for the *focused* and the *open* task and found that only 9 users have some overlap in the book-bags, with 7 only having a single book in both bags. From this we conclude that user treat the sub-tasks of the *focused* task as different from the *open* task.

An additional question is whether the supplied interface makes a difference. Our analysis shows that the number of gathered books is slightly higher for the multistage interface, especially in the case of the open task.

We also look at the overlap between the book bags of users, that is, whether different users find and select the same books or different books. The ratio between the size of all book bags combined as a bag (with repetition) and as a set (without repetition). The ratio of number of overall books selected over the number of distinct books selected. For the *open* task, the overlap ratio of the baseline interface (309 book selections of 290 distinct books) is 1.07 and the overlap ratio of the multistage interface is also 7% (358 book selections of 335 distinct books). The overlap is low, which is not surprising given the open nature of the task, and the type of interface seems to have little effect. For the *focused* task, the overlap ratio of the baseline interface is 1.23 (442 selections of 360 distinct books) and that of the multistage interface is 1.14 (470 selections of 412 distinct books). The overlap for the *focused* task is thus higher than for the *open* task, probably because all users are constrained in their selection by the more specific sub-tasks. Here the type of interface has a larger effect. The users of the baseline interface more often select the same books. Perhaps the multistage interface encourages users to explore the collection in more different ways than the baseline interface.

Table 3: Mean frequency of different action types per user for each stage in the interface. The frequencies for the focused task are shown in the top half, for the open task in the bottom half.

Action type	Baseline	Multistage				Both
		Browse	Search	Book-bag	Total	
focused						
paginate	2.2	1.3	1.9	0.1	3.3	2.8
add-to-bookbag	5.0	1.8	3.5	0.1	5.4	5.2
browse		7.6	0.2	0.1	7.9	
similar-books				0.3	0.3	
add-facet	3.7		2.4		2.4	3.0
remove-facet	0.7		0.6		0.6	0.7
remove-from-bookbag	0.3		0.0	0.4	0.5	0.4
show-item		6.4	0.3	0.3	7.0	
query	8.6	0.1	6.9	0.1	7.0	7.8
show-layout	0.0	2.3	3.2	1.6	7.0	
metadata	5.8	1.3	3.9	0.0	5.3	5.5
open						
paginate	2.2	2.7	0.9	0.2	3.8	3.0
add-to-bookbag	3.2	2.5	1.4	0.2	4.0	3.6
browse		7.5		0.0	7.6	
similar-books				0.5	0.5	
add-facet	2.5		0.4		0.4	1.4
remove-facet	0.6		0.1		0.1	0.4
remove-from-bookbag	0.1			0.2	0.2	0.1
show-item		8.8		0.6	9.4	
query	3.0	0.1	2.1	0.1	2.3	2.7
show-layout		1.0	1.2	1.3	3.5	
metadata	3.0	2.3	1.5	0.1	4.0	3.5

Table 4: Screen views in the multistage interface

Task	N	Total	Min	Max	Median	Mean	St.dev.
Test	95	292	1	7	3	3.1	0.9
Open	95	383	1	15	3	4.0	3.1
Focused	95	674	1	26	5	7.1	5.1

4.3 Users' actions in the different interfaces

Each interface allows users to perform certain actions, with some overlap between the available actions across the baseline interface and the three stages of the multistage interface. The mean number of actions of each action type per user is shown in Table 3, split between the *focused* task (top half) and the *open* task (bottom half). Certain actions are only available in the multistage interface, like *show-layout*, which corresponds to a switch between stages in the interface, and *browse*, which allows users to browse through the Amazon hierarchy of book categories without providing a search box.

The Table outlines some differences between the baseline and multistage interface. First of all, the users of the multistage interface utilize 'paginate' more, suggesting that the interface encourages users to explore a larger part of the collection. On the other hand, they use fewer filters and queries (both available in the *search* panel of the multistage interface). This is perhaps due to the fact that the users had more elaborate options to explore (via the *browse* panel) and to review results (via the *book-bag* panel) in the multistage interface, hence did not have to rely on querying and filtering alone, as in the case of the baseline interface. Finally, a difference can be observed in the use of book metadata: in the open task, participants view more book metadata using the multistage interface than via the baseline interface. It is possible that participants are triggered to check more books by the distinct functionality of the different panels of the multistage interface, especially since the open task allows users to explore freely. The same difference cannot be observed for the focused task, however, where users review slightly more metadata via the baseline interface than via the multistage interface.

4.4 The use of interface panels in the multistage system

In the case of the multistage interface, users had the ability to switch between interface screens, or *stages* in the interface. In this section, we look at the time spent in each screen, the number of switches between screens and the transition probabilities for switching between screens.

First of all, Table 4 shows the number of screens a participant viewed, as the participant has the possibility to switch multiple times between the *browse*, *search* and *book-bag* screens. For comparison purposes, we initially look at the training task, in which we expect users to test out all three interface screens. This is reflected in the table, as the mean and median of switches is close to three. Of the 95 users of the multistage interface, 75 (79%) went through the

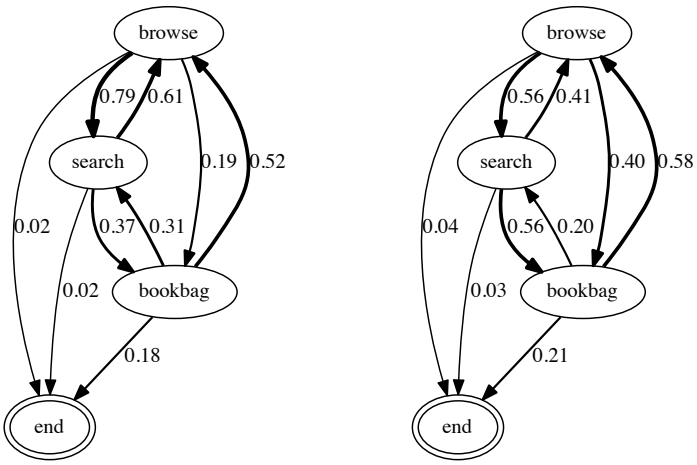

Fig. 1: Transition probabilities for the focused task (*left*) and open task (*right*).

screens in order—i.e. from *browse* to *search* and finally to book-bag to finish the training task. During the open task, participants may use more different interface units. This is reflected in the mean number of screens used, and a higher standard deviation in the second row of the table. Finally, in the focused task, participants have to carry out five sub-tasks, so we would expect that participants switch interface units more frequently. Indeed, the focused task results in a substantially higher median, mean and standard deviation. Hence, we have found evidence for frequent switching between interface units.

Next, we analyze the probabilities of switching between these interface units. Figure 1 shows the transition probabilities for the focused task, i.e. the probability that a user switches from a certain interface screen to another (or ends the task). Participants frequently switch between the *browse*, *search* and *book-bag* screens, and most commonly end the task from the *book-bag* screen. The higher probability of switching between search and explore in the focused task can be explained again by the task properties: having five sub-tasks to complete, the participants frequently move from one interface screen to another. The figure also shows the transition probabilities for the open task. Here, we see that in the open task, users more frequently switch between the *browse* and the *book-bag* stage, and less often from the *browse* to the *search* stage. Hence, the browse screen may be more important than the search screen in the open task, while the search screen is more important in the focused task.

To derive more insights into the importance of each screen in both focused and open tasks, we look at the time spent in each interface unit. We measured the time spent in an interface screen after initiating the task or switching to the *browse*, *search* or *book-bag* screen. Table 5 provides a summary. It shows that

Table 5: Total time spent in a panel

Task	N	Min	Max	Median	Mean	St.dev.
Focused						
Browse	95	0	2012	189	277.5	322.8
Search	95	23	1832	413.5	499	358.5
Review	95	0	653	162	199	154.7
Open						
Browse	95	0	7701	190.5	434.7	1116.1
Search	95	0	1453	107	205.6	265.6
Review	95	11	3211	128.5	222.25	459.3

for the focused task the *search* screen is used for the longest total duration by far, reflected in the highest median and mean duration, followed by the *book-bag* and *browse* screen, which are used substantially shorter. The open task features a different emphasis: the *browse* screen is used most frequently, as shown by the median and mean duration. The *search* and *book-bag* screen are comparatively used less often, both having similar values, but the difference for the median is less clear as in the case of the focused task. Finally, the standard deviation of total usage duration of the *browse* screen is a lot higher. Even without taking one outlier into account (most likely caused by a specific user's long period of inactivity), the variation in the use of this screen in the open task is the highest.

Summarizing, we found evidence for frequent switching between interface units, especially in the focused task. The willingness of users to switch between screens does provide positive indications for the usefulness of novel multistage interfaces and the enrichment of existing search options.

5 Participants' perceptions of multistage interfaces

The User Engagement Scale (UES) [7] is a multidimensional scale that contains six sub-scales: Aesthetics, Novelty, Felt Involvement, Focused Attention, Perceived Usability, and Endurability. Its purpose is to assist researchers in reaching a holistic understanding of users perceptions of technology. According to O'Brien and Toms [7] the scale seeks to measure multiple aspects of engagement and understand their relationships to one another.

To prepare the data for analysis some items were reverse coded. An initial examination of the data showed that there were no missing variables for any of the items. The 31 items were comprised into the 6 sub-scales. Table 6 shows the sub-scale means with both interfaces. The multistage interface seems more engaging in all sub-scales. However, we tested the differences using the Mann-Whitney test and found that only the differences for Endurability (p.= 0.006) and Felt Involvement (p. = 0.041) were statistically significant.

We also grouped the participants in three age groups (Group 1: age range 18-25, N=80; Group: 2: age range 26-35, N=80; Group: age range 35+, N=40)

Table 6: Mean engagement with baseline and multistage interfaces

	Aesthetics	Novelty	Felt Involvement	Focused Attention	Perceived Usability	Endurability
baseline	1.74	1.89	1.84	1.47	2.30	1.86
multistage	1.89	2.07	2.09	1.64	2.37	2.19

and examined whether the engagement varied between the groups, but the differences were not significant (Krutska-Wallis test). Also engagement did not vary significantly whether the open task or the focused task was performed first.

6 Conclusions

The analyses performed in this paper lead to various insights into the value of multistage interfaces in the context of complex book search tasks. First of all, the task duration in the multistage interface is substantially higher for both focused and open-ended tasks, suggesting that users are more involved in searching for books. This is also reflected in the significant differences for user engagement in the multistage interface, in terms of Endurability and Felt Involvement. Second, users viewed more result pages and collected more books in the multistage interface as compared to the baseline interface. In addition, the collected books had less overlap between participants. Hence, the longer task time also seems to result in a larger and more varied set of collected books. Finally, the frequent screen switching in both tasks suggests that the different screens encourage different types of activities. This can also be seen in the time spent in each screen: the *browse* screen is used more in the open-ended task, while the *search* screen is used for a longer total duration in the focused task.

The results suggest that the multistage interface encourages users to explore the collection in more different ways than the baseline interface. Further analysis is needed, however, since there may be personal differences between users, for example in terms of common patterns of interactions with the multistage interface. We plan to analyze these aspects in future work. Similar to [4], we also plan to look at the differences at different points in time of the tasks.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (WebART project, NWO CATCH # 640.005.001).

Bibliography

1. M. Hall and E. Toms. Building a common framework for iir evaluation. In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume

8138 of *Lecture Notes in Computer Science*, pages 17–28. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40801-4. doi: 10.1007/978-3-642-40802-1_3.

2. M. Hall, S. Katsaris, and E. Toms. A pluggable interactive ir evaluation work-bench. In *European Workshop on Human-Computer Interaction and Information Retrieval*, pages 35–38, 2013.
3. M. Hall, H. Huurdeman, M. Koolen, M. Skov, and D. Walsh. Overview of the INEX 2014 interactive social book search track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Labs and Workshops, Notebook Papers*, CEUR Workshop Proceedings (CEUR-WS.org), 2014.
4. H. C. Huurdeman and J. Kamps. From Multistage Information-seeking Models to Multistage Search Systems. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 145–154, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2976-7. doi: 10.1145/2637002.2637020. URL http://doi.acm.org/10.1145/2637002.2637020.
5. C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(199106)42:5⟨361::AID-ASI6⟩3.0.CO;2-#. URL http://dx.doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#.
6. R. Nordlie and N. Pharo. Seven years of inex interactive retrieval experiments - lessons and challenges. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, and G. Santucci, editors, *CLEF*, volume 7488 of *Lecture Notes in Computer Science*, pages 13–23. Springer, 2012. ISBN 978-3-642-33246-3.
7. H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2010.
8. V. Petras, T. Bogers, E. Toms, M. Hall, J. Savoy, P. Malak, A. Pawowski, N. Ferro, and I. Masiero. Cultural heritage in clef (chic) 2013. In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 192–211. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40801-4. doi: 10.1007/978-3-642-40802-1_23.
9. E. Toms and M. M. Hall. The chic interactive task (chici) at clef2013. http://www.clef-initiative.eu/documents/71612/1713e643-27c3-4d76-9a6f-926cdb1db0f4, 2013.
10. P. Vakkari. 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.