Aggregated Search Interface Preferences in Multi-Session Search Tasks

Marc Bron
ISLA, University of Amsterdam
m.m.bron@uva.nl

Jasmijn van Gorp
TViT, Utrecht University
j.vangorp@uu.nl

Frank Nack
ISLA, University of Amsterdam
nack@uva.nl

Lotte Belice Baltussen
Netherlands Institute
for Sound & Vision
lbbaltussen@beeldengeluid.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

ABSTRACT

Aggregated search interfaces provide users with an overview of results from various sources. Two general types of display exist: tabbed, with access to each source in a separate tab, and blended, which combines multiple sources into a single result page. Multi-session search tasks, e.g., a research project, consist of multiple stages, each with its own sub-tasks. Several factors involved in multi-session search tasks have been found to influence user search behavior. We investigate whether user preference for source presentation changes during a multi-session search task.

The dynamic nature of multi-session search tasks makes the design of a controlled experiment a non-trivial challenge. We adopt a methodology based on triangulation and conduct two types of observational study: a longitudinal study and a laboratory study. In the longitudinal study we follow the use of tabbed and blended displays by 25 students during a project. We find that while a tabbed display is used more than a blended display, subjects repeatedly switch between displays during the project. Use of the tabbed display is motivated by a need to zoom in on a specific source, while the blended display is used to explore available material across sources whenever the information need changes.

In a laboratory study 44 students completed a multi-session search task composed of three sub-tasks, the first with a tabbed display, the second and third with blended displays. The tasks were manipulated by either providing three task about the same topic or about three different topics. We find that a stable information need over multiple sub-tasks negatively influences perceived usability of the blended displays, while we do not find an influence when the information need changes.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.5.2 [**User Interfaces**]: Evaluation/methodology

Keywords

Search interface preferences, search behavior, aggregated search, multi-session search tasks

1. INTRODUCTION

In today's information society, organizations such as census bureaus, news companies, and archives, are making their collections of high quality information accessible through individual search interfaces [22]. As these sources are rarely indexed by major web search engines, seeking for information across these sources requires users to sequentially go through each of them individually. Aggregated search interfaces are a solution to this problem; they provide users with an overview of the results from various sources (verticals) by collecting and presenting information from multiple collections. Two general types of aggregated search interfaces exist: *tabbed* and *blended*. Tabbed interfaces provide access to each source in separate tabs, while blended interfaces combine multiple sources into a single result page [20].

The focus of previous work on aggregated search interfaces has been on selecting which vertical to present given a query and where to present it in the result list based on click logs [24], judgements [2, 3, 23] or on simulations [18]. Others have investigated how users interact with aggregated search interfaces, i.e., how verticals influence user search behavior [1] or the influence of tabbed and blended interfaces on user behavior and preferences in single-session tasks of varying complexity [4, 28, 29].

A multi-session search task, such as writing a report, consists of multiple information seeking tasks, each of which might be composed of its own sub-tasks [10]. Several factors involved in multi-session search tasks have been found to influence user search behavior. Such factors include: stages in the overall task, user search experience, user knowledge of the search topic, and complexity of individual sub-tasks [5, 9, 14, 17, 31, 34, 35].

Research also shows that users apply the display which they feel will be most effective and efficient in solving their task [19]. For multi-session research tasks this means that the preference for a tabbed or blended interface might alter between sub-tasks, depending on the user's idiosyncratic needs. This potential alteration is problematic for interface designers, because the prediction of which vertical to show, where, and how, requires an understanding of the relationship between the presentation options of each interface type, the user needs, and the performed sub-tasks.

We investigate whether user preference for source presentation changes during a multi-session search task with the following research questions: (1) Do users switch between tabbed and blended display types during a multi-session search task and what is the motivation to switch between display types? (2) Do changes in users' information need across sub-tasks influence preference for a particular display type? And (3) what other factors are related to changes in display preference during a multi-session search task?

We focus our investigation on research tasks in the humanities, as the data sources investigated in this domain can be very diverse and the behavior of repeatedly re-visiting sources that one encounters here naturally requires multi-session search. Importantly, while there is a general focus on text-based media, the domain increasingly makes use of other media types too. An example here is the field of media studies, where finding answers to hypotheses as part of the paper writing process relies on the inspection of audio-visual media, (through metadata or content-based), related metadata and secondary literature [8].

The dynamic nature of multi-session search tasks make the design of a controlled experiment a non-trivial challenge. For the research presented in this paper we adopt a methodology based on triangulation [11] and conduct two types of studies: a longitudinal study and a laboratory study. In our *longitudinal* study we follow 25 students during a four week research project. We provide students with an interface that allows switching between a tabbed display, blended display, and blended display with a find-similar functionality. This allows for the study of display use and switching in a naturalistic setting. Through questionnaires and focus group discussions we elicit the motivation for using a particular display.

In a *laboratory* study we present 44 students with a multi-session search task consisting of three complex sub-tasks. Each sub-task is carried out with a different display: the first task with the tabbed display; the second task with the blended display; and the third task with the blended display with a find-similar feature. We zoom in on the influence of changes in information need associated with recurring search sessions by manipulating whether a subject is assigned three sub-tasks about the same topic or three sub-tasks about different topics. This allows us to investigate the factors associated with changes in information need and whether these influence preference for a tabbed or blended display in different stages of a multi-session search task.

In Section 2 we describe the three variants of the aggregated interface; Section 3 and 4 describe the experimental setup and results of the longitudinal and laboratory study, respectively; in Section 5 we discuss the results of both studies in light of our research questions; Section 6 describes related work; we conclude in Section 7.

2. AGGREGATED SEARCH DISPLAYS

An aggregated search interface provides access to multiple, often heterogeneous collections. In building an aggregated search interface there are three aspects to consider: (i) how to retrieve relevant information from each vertical; (ii) which verticals to show; and (iii) where to place the verticals on the screen. Below we first give details of the back-end of our aggregated search interface and then describe three types of aggregated search interface displays: a tabbed display, a blended display and a blended display with find-similar functionality.

2.1 Data and Retrieval Back-end

The theme of the course selected for our longitudinal study is television history and the research projects carried out by the students are centered around television personalities between 1900 and 2010. To provide students with relevant material we obtained six collections from several archives and libraries: (i) a television program collection (metadata records for .5M programs); (ii) a photo collection (20K photos); (iii) a Wiki dedicated to television programs and presenters (20K pages); (iv) scanned television guides (25K pages); (v) scanned news papers starting from 1900 till 1995 (6M articles); (vi) digital news papers starting from 1995 till 2010 (1M articles). Each collection was indexed using Lucene SOLR 4.0 and the retrieval model used was BM25.
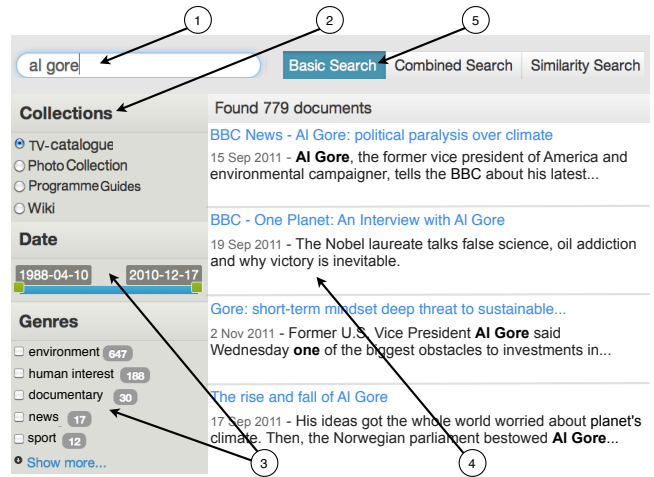

Figure 1: Tabbed display.

A single retrieval model may not be equally effective for each collection due to differences in term statistics and the presence of specific metadata fields in different collections [26]. To overcome this issue we provide faceted search and query preview capabilities in the aggregated search displays. These enable users to explore and learn about the characteristics of individual collections [32]. We did not optimize a retrieval model for each collection as depending on the size of the test set and type of queries a bias may be introduced towards particular types of documents, which is unwanted in a research task. The available facets depend on the collection as documents in some collections have rich metadata while others do not. The interaction model behind the facet values operates as follows: values within a single facet are combined using an OR operator, while values across facets are combined using an AND operator [30].

2.2 Tabbed Display

The tabbed display mimics the functionality and layout of a typical web search engine and is the default display presented to the user when opening the search interface. The tabbed display is shown in Figure 1; we use numbers 1, . . ., 5 to reference specific components in the display. It consists of: (1) a search box; (2) a collection selection menu; (3) values for several facets; (4) a result list; and (5) an option to select the tabbed or blended interface.

A search is initiated by submitting a query via the search box (1). The television program collection is selected by default (2). In response to a query 10 document snippets are shown for the selected collection on the result page (4). At the top of the result list the number of documents found is displayed and at the bottom of the page a pagination button enables moving to the next and further result pages. To further refine the results the top 5 values for several facets are available (3). Pressing on the *show more* button extends the list of facets up to the top 100. By selecting a different collection (2) results for this collection are displayed for the current query. Each tab displays the same number of results (ten) as a ranked list, five of which are generally visible above the fold depending on the size of the display.

A document can be viewed by clicking on a result snippet. The search result page is covered by an overlay and the content and metadata of a document are displayed. There are two special cases where the displayed information depends on the type of document, i.e., image and Wiki. Records form the photo or tv-guide collections consist of several images; in the overlay the photo and its
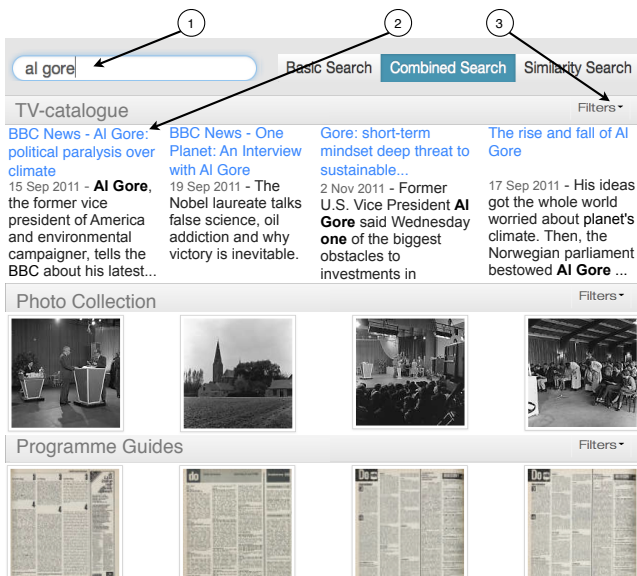
Figure 2: Blended display.

metadata are show with the additional option of viewing the next or previous photo without leaving the overlay. Wiki page content is not shown in the overlay, only the metadata and a link that opens the page in a new tab. To save a particular document a bookmark button is available, when hovering over a search results and in the document view overlay. Clicking the bookmark button saves the title of the document to a bookmark list available as a drop down list at the top of the screen. Selecting a title from the drop down list triggers an overlay with the document.

2.3 Blended Display

The blended display is based on the layout of interfaces typically used within digital libraries [25]. The blended display is shown in Figure 2; we use numbers 1, . . . , 3 to reference specific components in the display. It consists of: (1) a search box; and six horizontally oriented rectangular panes one for each collection. The vertical order in which the six collections are displayed is fixed and is the same as the order of the collection selection menu in the tabbed display, e.g., the top most pane shows results for the television program collection. Within a collection pane four results are shown ordered from left (most relevant) to right (less relevant) (2). To free up screen space for displaying results horizontally the facets for each collection are hidden. A button is available to toggle the display of facet values (3). The facets, bookmark options and document views are all the same as in the tabbed interface.

The blended display presents six by four results of which eight are generally visible above the fold depending on display size. More results are available as each source provides an independent pagination button. By scrolling down more results are available than in the tabbed display, however, only four results are available from each source. This is a trade-off between the two display types. The tabbed display supports searching through a single source, while the blended display supports searching through multiple sources.

In our blended display we use a fixed order and fixed number of results. This is in contrast to a trend seen in current web search engines, which is to display a variable number of results from a variable number of collections in a vertically oriented ranked list in a query dependent manner. The latter display method requires careful tuning, either on large numbers of users or on log data, both of which are unavailable for complex or multi-session search tasks.

2.4 Similarity Search

An additional feature was added to the blended display and provided as an additional screen, i.e., similarity search. By clicking on a document the user submits the current query and the first 100 words of the clicked document as an OR query [7, 27]. This type of explicit feedback is often implemented in a digital library context to discover related material across multiple sources [20]. However, the additional effort of using this feature and lack of understanding of how it works limits the use of find-similar like features [6, 33]. We add this feature to the blended display in order to explore its use in multi-session search tasks.

By clicking the find-similar button a query is issued and the document used for the query is placed in the query history pane appearing as a new row above the first row of search results. There are four slots, one of which is filled each time the find-similar button is clicked ordered from newest to oldest. When all slots are occupied the oldest document is removed and the others shifted to the right to free up space for the new document. Whenever a new query is issued the query history is emptied by removing the top row. Note that without clicking the find-similar button the similarity search display is exactly the same as the blended display.

3. A LONGITUDINAL STUDY

The goal of our first, longitudinal study was to investigate in which way people change between interfaces for a given work task in a natural setting and how this swapping behavior can be motivated, if it happens at all. The type of multi-session work task we make use of is a scholarly task, i.e., the task of writing a paper, which covers multiple search tasks to be carried out to collect and investigate the material that finally contributes to the paper [15]. Below we first describe the details of our experimental setting followed by an analysis of the data and discussion of the findings.

3.1 Study Setting

Through inquiries among staff at the humanities department of our university we searched for a course suitable for our study, one in which students complete the cycle of a research project including the development of a research question, searching for and interpreting materials, and reporting about their findings. The course that offered this structure and on which we finally settled is entitled: "Reception of media in historical perspective." The lecturer is experienced in teaching the course and used the same course schedule and assignments as in previous years. The course consists of two parts, we only focus on the first five weeks in which students conclude the following research project: "reconstruct the historical context of the 1950's (start of television) or 1920's (start of film) in order to explain the emancipatory role of a famous female television/film personality. Write a photo essay in which you incorporate primary and secondary sources that place the photos in context." The research project is split into four assignments, one due each week: (i) "familiarize yourself with the television/film personality of your choice and compose a list of five additional television/film personalities that fit your theme;" (ii) "start collecting images centered around your theme and collect material that motivates choosing these images;" (iii) "select ten images and add keyword descriptions to create a coherent story;" (iv) "prepare a presentation explaining the theme of your project, using the collected material." These assignments provide structure for the students as well as a natural separation of the research project into four parts.

Subjects. In total twenty-five students participated in the course and all were at the postgraduate level in the area of media studies. The sample consists of 12 men and 13 women, aged in terms of

median (MD) and interquartile range (IQR) around twenty-three years ($MD = 23$, $IQR = 22$–24). We asked subjects background questions using a 5 point Likert-type scale, where a one indicates no agreement and a five indicates extreme agreement. Subjects reported high levels of experience in general computer use ($MD = 4$, $IQR = 4$–5) and using online search tools ($MD = 4$, $IQR = 4$–5),

Procedure. At the end of the first lecture the experimenters presented the aggregated search system, explained how the three displays work and described the available data sources. After the presentation subjects were invited to sign up for the experiment, fill in a consent form and create a login. Subjects who signed up were not required to use the interface but were encouraged to use the system as a supplement to the sources normally used in the class. The incentive for using the system was the availability of unique sources otherwise unavailable. The experiment lasted 4 weeks.

Data collection. In conducting a naturalistic study there is a tension between collecting as much data as possible and affecting the environment through this data collection process. After negotiating with the lecturer we settled on three points of interaction with the subjects: at the end of the first lecture, once during the project and at the end of the final lecture (of the first part of the course). Participation of the students in all parts of the study was optional and no requests were made to use any specific features or to complete specific tasks with the aggregated search system, e.g., bookmarking. It was not deemed appropriate to collect the project grades of the students for use in the experimental analysis.

Two methods were used to collect qualitative data: open question surveys, and focus group discussions. In preparation of the focus group, subjects were asked to complete two online questionnaires, one before the second class and one before the fourth class. Questions focussed on the motivation to use a particular type of display. At the end of the second lecture and the final lecture a 15 minute focus group discussion was conducted. The discussion focussed on the motivation for using a particular type of display and switching between displays. The discussion was tape recorded and transcribed for later analysis. In addition, we collected quantitative data by logging all actions with the aggregated search system. See [12] for further details about subjects' browsing behavior.

3.2 Analysis

To investigate what type of display subjects use and whether they switch between display types in a multi-session search task, we first analyze the log data and then place our findings in the context of the qualitative data collected.

Log analysis. We first investigate whether there is a change in the amount that each of the three displays was used during the project. We considered three alternative indicators for the amount of use of a display. First, the amount of time that a subject spent logged into the system may overestimate the time a subject actually spent using the system. Similarly, clicks may underestimate the use of the system as inspection of result pages does not necessarily lead to clicks. We finally settled on the number of mouse hovers within a particular display as an indication for the amount of use of a display. A mouse hover is recorded when the cursor remains in the same position for 40ms and is only recorded again if the position changes [13]. Figure 3 shows the cumulative number of hovers with each of the three displays per day. We observe a sharp increase in the number of hovers recorded for each display before the 7th, 14th, 21th, and 28th day of the project, followed by a plateau of inactivity. These days coincide with the lecture and assignment deadline for each week and provides a natural separation of the course project into four stages. We further observe that most hovers
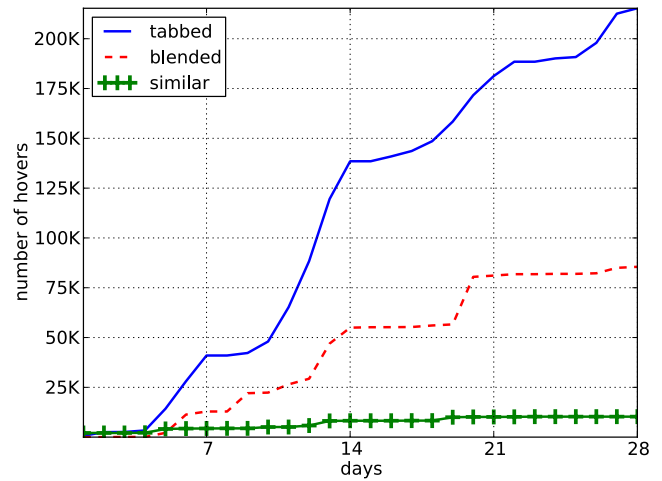

Figure 3: Cumulative number of mouse hovers recorded per day for the tabbed display (solid line), blended display (dashed line) and similarity display (cross marked).

are recorded for the tabbed display (solid) throughout the 28 days of the project. The blended display (dashed) receives less hovers, and the similarity display (crossed) receives the least hovers. In the second week more hovers are recorded for the tabbed and blended display compared to the first and last week. This may be due to the assignment of that week or the focus group after the second lecture. We return to this issue in the discussion of the qualitative analysis.

Next, we investigate whether individual subjects differ in the amount of use of each display. In the log data we find that not all subjects used all of the displays provided by the aggregated search interface. For all 25 subjects, hovers are recorded for the tabbed display, for 18 subjects hovers are recorded for the blended display, and for 11 subjects hovers are recorded for the similarity display. Figure 4 shows the total number of hovers recorded per user during the project on the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar) ordered by the total number of hovers. We observe that for 23 out of 25 subjects most hovers are recorded for the tabbed display, while for 2 subjects most hovers are recorded for the blended display. The
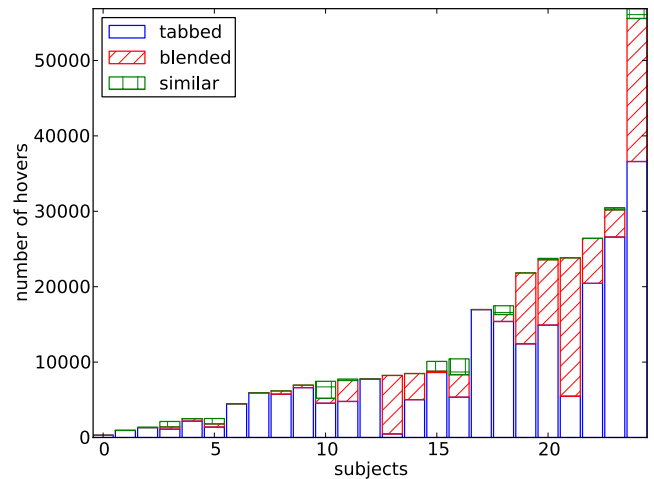

Figure 4: Total number of hovers received per user during the project on the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar).

blended interface is used notably by 8 subjects for which 25% or more of their total hovers occur with the blended display.

Finally, we investigate whether individual subjects differ in the amount of use of each display per week. Figure 5 shows the number of hovers recorded for each subject per week of the project with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar). We observe that during the first week (bottom row in Figure 5) out of the 25 subjects 11 make no or limited use of the displays. Of the remaining subjects' most have at least 75% of their hovers in the first week recorded with the tabbed display (12/25). For the remaining subjects (2/25) one predominantly uses the blended display, the other the find-similar display. In the second week (second row Figure 5, use of all displays increases. Out of 25 subjects 10 have around 100% of their hovers recorded for the tabbed display. The blended displays are used more than in the first week as for 8 subjects about 50% or more of their hovers are recorded with the blended display and two with the find-similar display. The remaining 5 subjects have no or hardly any hovers recorded for any of the displays. In the third week no or a limited number of hovers are recorded for most subjects (17/25) with the aggregated search system. The remaining subjects (8/25), however, continue to use the system with an intensity comparable to that of week 2. In the fourth week use drops to a level similar to that of week 1 with (13/25) active subjects. Of these subjects, one uses the tabbed and blended display equally, the remaining subjects predominantly use the tabbed display.

We perform a Chi square test to investigate whether these differences in display usage per week are due to chance and find that there is significant association between project week and use of display ($\chi^2(df = 6, N = 311061) = 14458.8, p < 0.001$). Table 1 shows a cross-tabulation of display type and project week. We find that in the second and third week relatively more attention is spent to the blended display (50% and 30%) than the tabbed display (45% and 20%), which receives attention more throughout the project. The standardized residuals of the hover values that most contribute to the significant effect are highlighted in boldface. Especially the amount of use of the similarity display in the first week is surprising, as well as the increased use of the blended display in the third and diminished use in the fourth week. Use of the tabbed display dominates the fourth week.

The absolute number of hovers provides an indication of the amount of use. Whether subjects switch between displays, how-

Table 1: Cross-tabulation of display type and project week, in percentages of both the week marginals and the interface marginals. The number of hovers are distributed as: tabbed $N = 215284$, blended $N = 85488$, and similarity $N = 10289$. Standardized residuals are show in brackets, the largest are highlighted in boldface

	tabbed	blended	similarity	
week 1	70.4% 19.0% (3.5)	22.1% 15.0% (-24.7)	7.5% 42.4% (**55.5**)	100%
week 2	67.7% 45.3% (-6.7)	29.5% 49.7% (14.8)	2.7% 38.3% (-11.8)	100%
week 3	60.5% 19.9% (-27.9)	36.9% 30.5% (**47.8**)	2.6% 17.9% (-10.2)	100%
week 4	89.1% 15.8% (**46.7**)	10.5% 4.7% (**-63.2**)	0.4% 1.4% (-31.7)	100%
	100%	100%	100%	

ever briefly, and in what sequence remains unclear. Figure 6 shows the same data as Figure 5—the number of hovers recorded for each subject per week of the project with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar)—on a log scale. We observe that 6 out of 25 subjects only use the tabbed display and 4 out of 25 subjects only switch once to another display during the 4 weeks of the study. The remaining 15 subjects switch 2 or more times between displays and switching occurs in all 4 weeks. The tabbed display is the default whenever subjects login to the system and switching between displays takes a conscious effort. It is to be expected that in the first week subjects explore the system and try the various options. That switching between displays occurs in any of the weeks and for some subjects several times within a week indicates that subjects have a function for each display during the project.

Figure 6: Number of hovers log scaled per user for each of the four stages with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar).

Qualitative analysis. The above analysis shows that a majority of the subjects switches display during the weeks of the research project. In the first week subjects predominantly use the tabbed display, while a minority prefers to use another display. In the second and third week of the project usage of the blended display increases
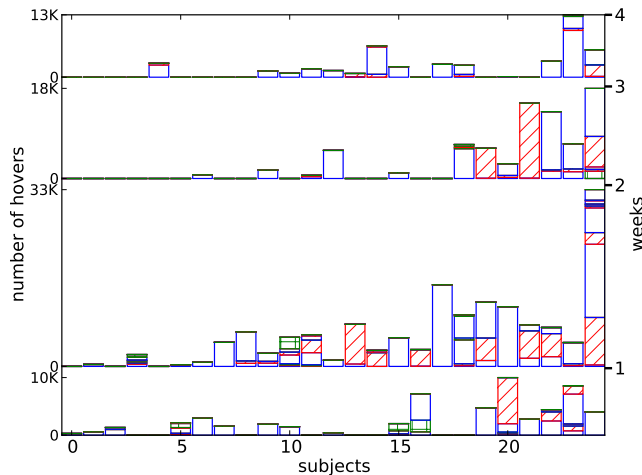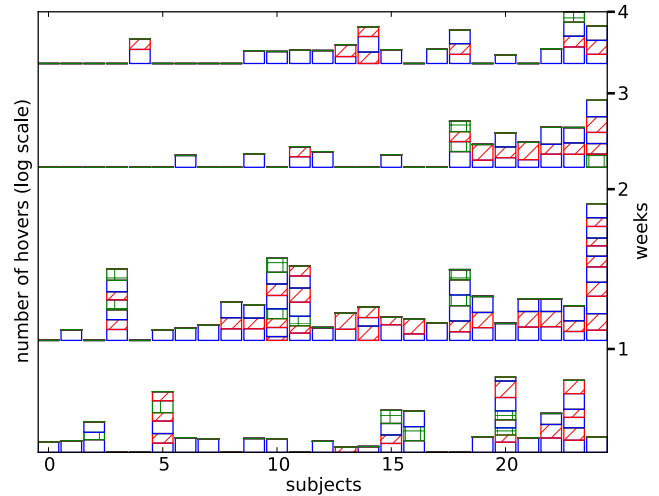

Figure 5: Number of hovers per user for each of the four stages of the project with the tabbed display (solid bar), blended display (dashed bar), and similarity display (crossed bar).

over that of the first week. The number of subjects using the system, however, decreases. Below we summarize the qualitative data gathered from two surveys and two focus group discussions and discuss subjects' motivation to switch between displays.

Before the second lecture 15 subjects completed a survey with open questions about the use of the three displays during the first week. On the question why subjects did or did not use the tabbed display all subjects indicated to have used the system: 8 subjects indicated to use the display to explore the collections and to get an impression of the available material; 5 subjects used the display to try out the system; 1 subject used the display but provided no motivation; and 1 subject liked the ability to specify in which collection to search. On the question why subjects did or did not use the blended display 10 subject indicated not to use the blended display: 5 subjects indicated not to have decided on a specific topic for their project yet and that they preferred the tabbed display to explore available material; and 5 subjects did not provide a motivation. Of the remaining 5 subjects that did use the blended display: 3 subjects indicated to like the blended display as it provided them with a better overview than the tabbed display; 1 subject found the blended display confusing; and 1 subject was trying out the display. On the question why subjects did or did not use the similarity display 8 subjects indicated not to use this display: 4 subjects indicated not to have decided on a specific topic; and 4 provided no motivation. Of the remaining 7 subjects that did use the similarity display: 3 subjects expected to find related television personalities; 2 subjects found that the similarity display did not provide relevant information; and 2 subjects were trying out the display.

In a 15 minute focus group discussion with 25 participants conducted at the end of the second lecture we focussed on the motivation for using or not using the system with a particular display. The consensus among participants is that the information provided by most of the sources is too specific for the current stage of the project: "in this phase I want general information, now when I search for photos of her I would get very specific information." Due to the stage of the project the tabbed display, with which a single collection can be explored, was preferred: "I found the Wiki interesting, because of our assignment this week" and "next week I will search differently, as this week was for exploration." In the similarity and blended display all collections are presented at the same time which subjects found less useful: "you get results with which you can not get an overview of the topic." Subjects also noted to prefer other resources, e.g., Wikipedia, and web search engines in order to get an overview of the possible topics for their project.

In summary, the above qualitative data suggests several reasons why subjects preferred the tabbed display during the first week and why use of the system was lower compared to the second week. First, the search behavior in this stage was of an exploratory nature as subjects tried to get an overview of available information and to decide on a topic for their project. Second, the information in most sources other than the Wiki was too specific. For example, a photo of a presenter and a guest during a specific broadcast of a television show provides little background information necessary to get oriented on a topic, for this purpose Wikipedia and similar sources are more useful. Finally, subjects preferred to use other resources such as Wikipedia and web search engines.

Before the fifth lecture 15 students again completed a survey with open questions about the use of the three displays during the second third and fourth week. Regarding the use of the tabbed display 6 subjects remarked that they liked to use it to find specific information in a specific collection, while 3 subjects preferred it to get an overview, and 1 provided no motivation. Subjects that did not use the tabbed interface favored the use of external sources (3

subjects) or preferred the overview of multiple collections (2 subjects). Regarding the blended display 7 subjects remarked that it was useful to find additional information: "to find next to images, more specific information about a person or event" or "get an as complete as possible overview of what is available about a specific program." To search for images 2 subjects preferred the blended display and 1 subject preferred the overview without searching for anything specific. The remaining 4 subjects did not find the blended display useful as it was confusing or did not provide relevant material. The similarity display was used by 2 subjects that mentioned being curious about the material that would be found. The remaining 13 subjects did not use the similarity display because they did not get around to using it or because results were irrelevant.

We conducted a second 15 minute focus group discussion with 25 students at the end of the fourth week. We focused again on the motivation for using a particular display and centered the discussion around three points: (i) did becoming more familiar with the search system affect display choice; (ii) did increased knowledge of the research topic affect display choice; and (iii) did the assignment each week affect display choice. Subjects did not experience difficulties in using the displays: "it is a reasonably intuitive system, I understand it." They did not associate any changes in display use with an increase in familiarity with the search system. Both increasing knowledge of the research topic and type of assignment are factors that subjects associate with changes in display use. Two groups emerged in this discussion. One group that changed from using the tabbed display to using the blended display: "I started to use the combined display more, because I know what I need, and then I want to see everything that is available about her." The other group changed from using the blended display to using the tabbed display: "I felt the same, that my search started to become more focussed, but then I preferred using the tabbed display because in the first week you do not know how to use photos and program guides, and now I wanted to know what was said in those sources about a specific person." Some subjects remarked that they did not use the system in later weeks of the project as it did not contain any relevant material for the theme they had chosen. There were two options for a project theme, i.e., female role models in television or film, see section 2. The material in the repositories is predominaltly focussed on television and does not contain as much relevant material for the second theme.

In summary, the second survey and focus group suggests that after the first week subjects' use of the system increased as their search topic became more concrete and the various sources became more relevant. We find several motivations for the switching behavior between displays throughout the project. The use of the tabbed display is motivated in the first week by an exploratory information need, where the Wiki was the only suitable source. In later weeks searching a single source is motivated by a more specific information need, e.g., photos for the essay. The use of the blended display is initially motivated by a need to explore the content of the various sources. Later the blended display provides a way to explore multiple sources simultaneously for material about a specific entity. We observed several switching patterns in Figure 6 that can be explained using the above motivations. Whether subjects engaged in a particular pattern seemed to depend on the individual subject. Some subjects indicated to have searched with the blended display in the first week and preferred to continue searching with the tabbed display in the second week. In this case the blended display is only used for an initial exploration of the sources. Other subjects repeatedly switched between the tabbed and blended display during the project. In these cases the blended display is likely used to search for additional information about a specific entity across sources.

Limitations. The setup of the longitudinal study involved that subjects were asked several times, i.e., focus group and surveys, about their use of and switching between displays. This may have influenced the behavior of the subjects. To avoid influencing subjects two measures were taken. First, subjects do not get a reward and have no obligation to use the tool. Their priority is to finish the assignment for each week and pass the course. To this end subjects choose to use the display which they feel will be most effective in solving their task [19]. As one subject remarked: "I have to search longer before I find something [with the aggregated search system] and I notice that I want to be efficient and then I quickly switch to a web search engine. I want to use the [aggregated search system] because I know you [the experimenters] need that, but at some point I think [insert expletive] I need to find my material and then I choose for something that quickly gives me results..." Although use of the system diminished in the third and fourth week, a number of subjects used the system and switched displays throughout the project. This suggests that the tabbed and blended displays are both found useful by some of the subjects and that each display supports a recurring need. Second, subjects were encouraged to use the system and explore all of its functionality with equal emphasis. The find similar and blended displays were both mentioned in the first focus group as being used less than the tabbed display in order to elicit a discussion about why this is. If the focus group discussion had any influence on the use of the find-similar display than this had a diminishing effect on its use.

4. A LABORATORY STUDY

In our first, longitudinal study we observed that the majority of subjects switched between displays and that use of the blended display was motivated by a need to explore the content of the sources available in the aggregated search system. Initially subjects desired an overview of the sort of information generally available in the sources, later the blended display was used to obtain an overview whenever a new specific information need arose, i.e., what is available in each source about a specific entity. We are interested in whether user preferences change in later stages of the multi-session search task depending on whether there are changes in the users' information need. We recreated this process in a laboratory study by assigning subjects three tasks to complete: the first task with a tabbed display, the second with a blended display, and the third with the similarity display. We compare two conditions: (i) one in which the three tasks are about the same topic allowing subjects to become increasingly familiar with the topic; and (ii) one in which tasks are about different topics simulating changes in information need. In this setting we investigate whether subjects preferences for displays deviate as some become more familiar with a topic, while others encounter new topics.

4.1 Study setting

The study uses a mixed methods design with 3 search tasks as within subject factor and the (in)dependency of the tasks as between subject factor. This design is similar to other work on multi-session search tasks [16] The following 3 sub-tasks emulate a multi-session search task: (i) imagine that you work at the editorial office of a current affairs program and are asked to collect information about [celebrity]. Collect at least 5 items deemed to be relevant for this collection, (ii) search for events that were key in the career of [celebrity]. Collect at least 5 items for this collection, for example articles and photographs, about these events, (iii) key in the career of [celebrity] was [event]. Collect at least 5 items about the run-up to the program, the program itself, and the aftermath. In this experiment [celebrity] stands for one of three television personalities and

[event] represents an event related to the corresponding celebrity. The tasks are modeled after the assignments subjects received in the longitudinal study, i.e., starting out broad and gradually becoming more specific. Randomizing this order would obfuscate the simulation of a multi-session search task.

A subject completes 3 sub-tasks that are either dependent or parallel [16]. In a dependent series of sub-tasks all tasks are performed with respect to the same celebrity, while in a parallel series of sub-tasks each task has a different celebrity as topic simulating changes in a users information need. In the parallel condition the celebrities are randomized across tasks and in the dependent condition celebrities are randomized across subjects. In this manner each celebrity and task combination is seen an equal amount of times in each condition. All subjects are presented with the same display for each task. To complete the first task subjects are provided with the tabbed display, for the second task with the blended display, and for the third task with the similarity display.

We used a fixed order for the displays as we are interested in the preference of the blended display in later stages of the multi-session search task. In this way subjects are able to become familiar with the repositories using the tabbed display before encountering the blended display.

Subjects. A group of 44 undergraduate students participated in the study. For two of the subjects a technical failure prevented recording the pre-experiment questionnaire data. All analyses reported in the results section are excluding these two subjects and are conducted on the remaining 42 subjects. The students (30 female, 12 male) were around nine-teen years of age ($MD = 19.0$, $IQR = 19$–21.75). Random assignment to conditions resulted in two groups. The dependent task group contained 23 students (6 male and 17 female), where the parallel condition group contained 19 subjects (6 male and 13 female). We asked subjects of both groups to answer background questions using a 5 point Likert-type scale, where a one indicates no agreement and a five indicates extreme agreement. Subjects generally reported high levels of experience in general computer use ($MD = 4$, $IQR = 4$–5).

We additionally measured topic knowledge (1 item scale) and search experience (7 item scale) to make sure that these characteristics were balanced over the two groups. We found no significant difference between the two groups in terms of topic knowledge ($Z=.68$, $p=.50$), general computer use ($Z=.18$, $p=.86$) or search experience ($Z=.20$, $p=.84$) using Wilcoxon rank-sum tests.

Procedure. The study was conducted on two separate occasions in a computer lab equipped with 30 computers. Subjects could sign up for one of two time slots to participate in the study. Subjects of the first slot were asked not to communicate about the experiment until the second run had been finished. Each occasion of the study started with an introduction to the research project and a viewing of a 5-minute tutorial video explaining the use of the search displays. After the viewing of the tutorial, subjects were invited to fill in a consent form and create a login. Subjects were not allowed to talk, use mobile phones or open any other browsers. Four experimenters invigilated each of the experiments. After completing a background questionnaire eliciting demographics subjects were randomly assigned to either the dependent or parallel condition and presented with a pre-experiment questionnaire to collect information about search experience and prior knowledge about the task topic(s). Next, subjects started on the multi-session search task. For every task the subjects were given 10 minutes after which they were redirected to a post-task questionnaire asking about the topic difficulty, the perceived usability of the display, and the search effectiveness of the display. After the final search task a post-experiment questionnaire was presented that asked subjects to order the dis-

Table 2: Search effectiveness scale.

1	The display has supported me in solving the search task.
2	The display provided enough information for every collection to solve the search task.
3	The display provided surprising search results relevant to the search task.
4	The display supported me in finding relevant search results.

Table 3: Interaction effects of the perceived usability, search effectiveness, topic difficulty, one collection is important and an overview is important responses in terms of median (interquartile range) across conditions (dependent and parallel) and across topics. The Kruskal Wallis hypothesis test is used to test for significant within subject effects, the Mann-WhitneyU hypothesis test is used to test for significance between subjects.

condition	task 1	task 2	task 3	K-W (H,p)
		perceived usability		
dependent	3.9 (3.8-4.1)	3.6 (3.3-4.1)	3 (2.7-4)	**(17.7, <.001)**
parallel	3.7 (2.9-4)	3.7 (3.1-4.1)	3.4 (2.9-3.7)	(2.90, .24)
M-W (U,p)	**(108,<.001)**	(197, .30)	(190, .24)	
		task difficulty		
dependent	3 (2-3)	2 (2-3)	3 (2-4)	(5.95, .051)
parallel	3 (3-4)	3 (2-3)	2 (2-3)	**(6.45, .040)**
M-W (U,p)	**(134, .017)**	**(151, .045)**	(155, .054)	
		search effectiveness		
dependent	3.3 (3.3-3.8)	3.3 (2.8-3.9)	3.3 (2.4-3.9)	(2.68, .26)
parallel	3.3 (3.3-3.8)	3.3 (3.0-3.6)	3.3 (2.6-3.6)	(3.21, .20)
M-W (U,p)	(194, 0.27)	(201, 0.33)	(194, 0.27)	
		overview important		
dependent	4 (3-4)	3 (3-4)	3 (3-4)	(4.02, .13)
parallel	4 (3-4)	4 (3-4)	3 (3-4)	(2.75, .25)
M-W (U,p)	(214, .46)	**(149, .040)**	(183, .19)	
		one collection important		
dependent	4 (3.5-4)	3 (2-3.5)	3 (2-3.5)	**(16.4, <.001)**
parallel	4 (3-4)	3 (2-4)	3 (2-4)	(2.70, .26)
M-W (U,p)	**(148, .038)**	(211, .43)	(218, .50)	

plays by preference, and to express any remarks they might have about the experiment. In total an experiment lasted 1.5 hours.

Data collection. We use two Likert-scales to measure subjects' preference for a particular type of display. For each item in the scales subjects indicated their agreement with the statement on a scale from one to five, where a one indicates no agreement and a five indicates extreme agreement. To arrive at a single score for each Likert-scale the mean of the responses was taken. To measure perceived usability we use an adaptation of the perceived usability sub-scale of the O'Brien Engagement Scale [21]. We modified the scale to apply to an aggregated search setting by substituting the word "shopping" and "website" by "searching" and "display" in the items. Additionally, we rephrased some of the items to a positive wording to arrive at an alternating 10 item scale.

To measure search effectiveness we used the items in Table 2. A final 3 items were devoted to subjects' perception of the task: (i) "the task was difficult to complete;" (ii) "there was one collection most useful in solving the search task;" and (iii) "to get an overview of the results in different collections is important in solving the task." These were not combined and are analyzed separately. All actions with the aggregated search interface were logged.

4.2 Results

We first investigate whether subjects' preferences differ within conditions in terms of relations between task and the dependent variables usability, task difficulty, and search effectiveness. We use non-parametric tests as data is collected on ordinal scales, and not all variables meet the assumption of normality. Table 3 shows the median (M) and inter quartile range (IQR) for each of the dependent variables split over each topic and the two conditions, i.e., dependent and parallel. We observe that there is a significant interaction between task and perceived usability of the display in the within subject dependent condition (Kruskal-Wallis H(2, N = 42) =17.7, p<.001), top part of Table 3). Post hoc comparisons show a significant difference between task 1-task 2 (Mann-Whitney U(N =42) =155, p=.008) and between task 1-task 3 (Mann-Whitney U(N =42) =81.5, p=<.001) at the Bonferroni corrected significance level of $\alpha = .05/3$. These results show that when the topic of the search task remains the same, subjects find the tabbed display easiest to use, and that the blended displays provided in later stages are considered more difficult to use. In the parallel condition we do not find a significant interaction between topics and usability. The median and IQR values are stable for the first two topics and decreases for the third topic. With this result we are unable to determine whether users find the blended display easier to use to explore a new topic.

We further observe that there is a significant interaction between task and task difficulty in the within subject parallel condition (Kruskal-Wallis H(2, N = 42) =6.45, p=.04), second row of Table 3). Post hoc comparisons show a significant difference between task 1-task 3 (Mann-Whitney U(N=42) =94.5, p=.006). In the dependent condition we do not find a significant interaction, although in terms of the median values the first and third task are considered more difficult than the second task. This suggests that in the parallel

condition subjects find that the tasks become easier as they search material for new topics with the blended displays, while in the dependent condition task difficulty first decreases and then increases. This seems counter intuitive as subjects in the dependent condition have encountered more material on the topic and during the first two search tasks and should have a better idea of what search terms and which sources to use.

We find no relation between topics and the search effectiveness that subjects experience with the displays.

Two possible factors that moderate the preference of subjects for a certain display are: (i) whether for a particular task a subject considers a single collection as the most important source to search for material; and (ii) whether a subject considers it important to get an overview of the material available in different collections for a certain task. We do not find any interaction between the tasks and the importance of getting an overview. We do find a significant interaction between tasks and the fact that a single collection is considered to be the most important in solving a task (Kruskal-Wallis H(2, N = 42) =16.4, p<.001). Post hoc comparisons show a significant difference between task 1-task 2 (Mann-Whitney U(N =42) =109.5, p=<.001) and between task 1-task 3 (Mann-Whitney U(N =42) =103.5, p<.001). Subjects in the dependent conditions initially consider a single collection important to solve the search task, while in later tasks more collections become important.

Next, we investigate between subject effects in terms of relations between topic knowledge and the dependent and moderating variables. We find that in the first task there is a significant difference between the perceived usability reported by subjects from the dependent and parallel condition (Mann-Whitney U(N=42)=108, p<.001). We further observe that in the first and second task there

Table 4: Cross tabulation of search moves in stage 1 for the dependent and parallel conditions. Search moves are given in percentages with the standardized residuals in brackets.

move	dep.%(N=926)		par.%(N=728)	
paginate	6.48	(-1.38)	9.34	(1.55)
bookmark	11.77	(-1.37)	15.52	(1.55)
filter	17.93	(1.21)	14.29	(-1.36)
change tab	21.38	(1.14)	17.58	(-1.29)
queries	11.23	(-0.55)	12.64	(0.62)
view docs	22.14	(0.28)	21.15	(-0.32)
delete bookmark	0.32	(-0.46)	0.55	(0.52)
unique queries	8.75	(-0.08)	8.93	(0.09)

is a significant difference between the task difficulty reported by subjects from the dependent and parallel condition (Mann-Whitney U(N=42)=134, p=.017), i.e., in the dependent condition the first and second task are considered easier.

Regarding the moderating variables we find that in the second task subjects in the parallel condition consider getting an overview more important than subjects in the dependent condition (Mann-Whitney U(N=42)=149, p=.40). With respect to the importance of a single collection to solve the first search task, subjects in the dependent condition consider this more important than subjects in the parallel condition.

During the first stage subjects are assigned the same task and possess the same level of prior knowledge. Subjects were randomly assigned to the dependent and parallel conditions and no significant differences were found in terms of subjects' background, i.e., prior knowledge, search experience, and computer use. Finding significant effects in the first stage indicates the presence of additional factors that possibly interact with the dependent variables. Table 3 shows that subjects in the dependent condition consider the first task easier, a single collection more important, and find the tabbed display easier to use than subjects in the parallel condition.

We also investigate to what extend search strategy is involved in the effects between the two conditions. Table 4 shows a cross tabulation of the search moves with the tabbed display recorded within the dependent and parallel condition. We find that there is a significant relation between the moves made in a display and the conditions ($\chi^2(df$=7,N=1654)=16.2, p=.022). From the standardized residuals we find that especially the following search actions contribute to the significant effect: paginating, bookmarking, using filters and changing collection (tab). Subjects in the dependent condition tend to switch between collections and use the facet filters, while subjects in the parallel condition tend to paginate and bookmark more frequently. This suggests that subjects in the dependent condition initially explore more of the collections while subjects in the parallel condition dig deeper into material of a single source.

5. DISCUSSION

Regarding our first research question, we find that the majority of subjects switch between displays during the project. The main motivation to use the tabbed display is to zoom in on a single source as other sources are not considered relevant at that stage of the project, e.g., Wiki documents are relevant at the start of the project to gather general background information, while photos of specific events become relevant at a later point. The use of the blended display is initially motivated by a need to explore the content of various sources. Later the blended display provides a way to explore multiple sources simultaneously for material related to a specific information need.

With respect to our second research question, we find that when subjects are completing search tasks about the same topic there is a negative influence on the usability of a blended display when switching from a tabbed to a blended display. We also find that when subjects are completing search tasks about different topics there is a decrease in perceived task difficulty when switching from tabbed to blended displays. These findings suggest that subjects are more likely to switch to a blended display when their information need changes since switching is then associated with decreasing task difficulty. While switching to a blended display is less likely when subjects are engaged in a sequence of search tasks related to the same topic because of a negative influence on usability.

Turning to our third research question, we find that there are several factors that influence the preferences of subjects for the tabbed display, i.e., whether subjects find the tabbed display easy to use, whether they find the first search task easy, and whether they find that one of the collections is most important in solving the first search task. These factors limit the generalizability of our findings as the observed interactions between changes in information need and usability (information need and task difficulty) may be specific to our sample with this particular configuration of these factors.

In general, both studies suggest reasons for subjects to change preferences and switch displays during a multi-session search task.

6. RELATED WORK

The work described in this paper expands on work from two fields of study: aggregated search and multi-session tasks. Studies in aggregated search have investigated whether users prefer tabbed or blended displays for single-session search tasks of varying complexity. Most closely related to our study is work by [28] who find that for complex tasks users prefer blended displays. A later study shows that indeed more verticals are clicked when task complexity increases, but that users do not necessarily prefer a blended or tabbed display [4]. As a possible explanation the search experience of the subjects is suggested. We find that the users actually change interface preference during a specific type of multi-session search task depending on changes in information need.

Regarding studies on multi-session search tasks, there is work that shows that increasing topic knowledge affects dwell time [16], which can be used as a predictor for document usefulness. We did not analyze the effect of increasing topic knowledge at this level of granularity and focussed on the effect of changes in information need rather then the increase of topical knowledge. Other work on understanding multi-session search task is focussed on describing the factors involved [9, 14, 31] and the search patterns of users with existing systems [17, 35]. Our work is different in that it attempts to compare the usefulness of a new type of system to an existing system within the context of a multi-session task.

7. CONCLUSION

Aggregated search interfaces are a promising way to provide users with an overview of results from various sources. In this paper we investigated the use and preferences of users for a tabbed and blended display within the context of a multi-session search tasks. In our first, longitudinal study we observed that the tabbed display is predominantly used motivated by a need to zoom in on specific sources. The majority of subjects, however, switched between the tabbed and blended displays. Use of the blended display was motivated by a need to explore the content of the sources available in the aggregated search system. Initially, subjects desired an overview of the sort of information generally available in the

sources, later the blended display was used to obtain an overview whenever a new specific information need arose.

In a laboratory study a multi-session search task was recreated, composed of three tasks with either a tabbed or blended display. The tasks were manipulated by either providing three tasks about the same topic or about three different topics. We found that a stable information need across multiple sub-tasks negatively influences perceived usability of the blended displays, while we do not find an influence when the information need changes.

The combined results from both studies suggest that subjects change display preference during a specific type of multi-session search task, i.e., seeking archival materials across multiple heterogeneous sources during a research project.

8. REFERENCES

[1] J. Arguello and R. Capra. The effect of aggregated search coherence on search behavior. *CIKM'12*, 2012.

[2] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. In *CIKM'11*, pages 201–210. ACM, 2011.

[3] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In *Adv. in Inf. Retr.*, pages 141–152. Springer, 2011.

[4] J. Arguello, W. Wu, D. Kelly, and A. Edwards. Task complexity, vertical display and user interaction in aggregated search. In *SIGIR'12*, pages 435–444. ACM, 2012.

[5] A. Aula, R. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In *SIGCHI'10*, pages 35–44, 2010.

[6] N. J. Belkin. Helping people find what they don't know. *Communications of the ACM*, 43(8):58–61, 2000.

[7] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *TPDL '11*, pages 360–371. Springer, 2011.

[8] M. Bron, J. van Gorp, F. Nack, M. de Rijke, and S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In *SIGIR '12*, pages 425–434. ACM, 2012.

[9] K. Byström. Information and information sources in tasks of varying complexity. *JASIST*, 53(7):581–591, 2002.

[10] K. Byström and P. Hansen. Conceptual framework for tasks in information studies. *JASIST*, 56(10):1050–1061, 2005.

[11] N. Denzin. *The research act: A theoretical introduction to sociological methods*. Aldine De Gruyter, 2009.

[12] J. He, M. Bron, and A. de Vries. Characterizing stages of a multi-session complex search task through direct and indirect query modifications. In *SIGIR '13*. ACM, 2013.

[13] J. Huang, R. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *CHI 2011*, pages 1225–1234. ACM, 2011.

[14] C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *JASIS*, 42(5):361–371, 1991.

[15] Y. Li. Exploring the relationships between work task and search task in information search. *JASIST*, 60(2):275–291, 2008.

[16] J. Liu and N. Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *SIGIR '10*, pages 26–33. ACM, 2010.

[17] J. Liu, N. Belkin, X. Zhang, and X. Yuan. Examining users' knowledge change in the task completion process. *Information Processing & Management*, 2012.

[18] W. Lu, Q. Wang, and B. Larsen. Simulating aggregated interfaces. *Workshop on Aggregated Search*, pages 24–28, 2012.

[19] J. McGrenere, R. M. Baecker, and K. S. Booth. A field evaluation of an adaptable two-interface design for feature-rich software. *ACM Trans. Comput.-Hum. Interact.*, 14(1), 2007.

[20] M. Melucci and R. Baeza-Yates. *Advanced topics in information retrieval*, volume 33. Springer, 2011.

[21] H. O'Brien and E. Toms. The development and evaluation of a survey to measure user engagement. *JASIST*, 61(1):50–69, 2009.

[22] S. Raghavan and H. Garcia-Molina. Crawling the hidden web. *Stanford Digital Libraries Technical Report*, 2000.

[23] R. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. In *Adv. in Inf. Retr.*, pages 250–261. Springer, 2011.

[24] J. Seo, W. Croft, K. Kim, and J. Lee. Smoothing click counts for aggregated vertical search. In *Adv. in Inf. Retr.*, pages 387–398. Springer, 2011.

[25] A. Shiri. Metadata-enhanced visual interfaces to digital libraries. *JIS*, 34(6):763–775, 2008.

[26] M. Shokouhi and L. Si. Federated search. *Found. & Trends in Inf. Retr.*, 2011.

[27] M. D. Smucker and J. Allan. Find-similar: similarity browsing as a search tool. In *SIGIR '06*, pages 461–468. ACM, 2006.

[28] S. Sushmita, H. Joho, and M. Lalmas. A task-based evaluation of an aggregated search interface. In *SPIRE '09*, pages 322–333. Springer, 2009.

[29] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. Factors affecting click-through behavior in aggregated search interfaces. In *CIKM'10*, pages 519–528. ACM, 2010.

[30] D. Tunkelang. Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–80, 2009.

[31] P. Vakkari, M. Pennanen, and S. Serola. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information processing & management*, 39(3):445–463, 2003.

[32] R. White and R. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[33] R. W. White, I. Ruthven, and J. M. Jose. The use of implicit evidence for relevance feedback in web retrieval. In *Adv. in Inf. Retr.*, pages 93–109. Springer, 2002.

[34] T. Wilson. Models in information behaviour research. *J. Doc.*, 55(3):249–270, 1999.

[35] I. Xie and S. Joo. Factors affecting the selection of search tactics: Tasks, knowledge, process, and systems. *Information Processing & Management*, 48(2):254–270, 2012.