Active and Passive Utility of Search Interface Features in Different Information Seeking Task Stages

Hugo C. Huurdeman
University of Amsterdam
The Netherlands
huurdeman@uva.nl

Max L. Wilson
University of Nottingham
United Kingdom
max.wilson@nottingham.ac.uk

Jaap Kamps
University of Amsterdam
The Netherlands
kamps@uva.nl

ABSTRACT

Models of information seeking, including Kuhlthau's Information Search Process model, describe fundamentally different macro-level *stages*. Current search systems usually do not provide support for these stages, but provide a static set of features predominantly focused on supporting micro-level search interactions. This paper investigates the utility of search user interface (SUI) features at different macro-level *stages* of complex tasks. A user study was designed, using simulated work tasks, to explicitly place users within different stages of a complex task: pre-focus, focus, and post-focus. Active use, passive use and perceived usefulness of features were analysed in order to derive *when* search features are most useful. Our results identify significant differences in the utility of SUI features between each stage. Specifically, we have observed that *informational* features are naturally useful in every stage, while *input, control* features decline in usefulness after the pre-focus stage, and *personalisable* features become more useful after the pre-focus stage. From these findings, we conclude that features less commonly found in web search interfaces can provide value for users, without cluttering simple searches, when provided at the right times.

Keywords

information seeking, stages, user interfaces, information retrieval

1. INTRODUCTION

Research into Search User Interfaces (SUIs) [10, 24, 38] has proposed many different interactive features, from search suggestions [22] to facets [29] to personal spaces to collect useful results [6]. Although their usefulness has been proven in micro-level studies of complex and exploratory tasks, many of these features have not been adopted by search engines, perhaps because they can impede search during simple lookup tasks [5]. In contrast, information seeking theory [16, 30] often highlights the existence of *stages* of search *within tasks* involving learning and construction, suggesting that we should consider *when* SUI features might be useful within tasks, rather than whether they are useful for tasks. Different categories of features, such as *input*, *control*, *informational* and *personalisable* features [38], might support users in different ways, both actively and passively. An understanding of the utility of features at different stages may help to overcome the apparent divide between the dynamic stages documented in macro-level information seeking models and the more static SUIs currently available online.

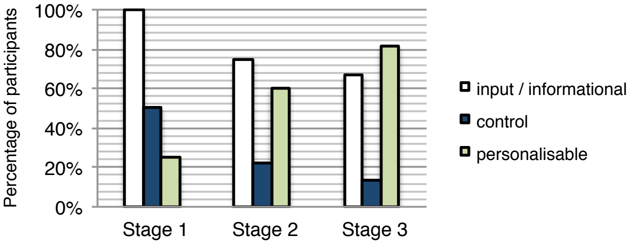


Figure 1: SUI feature categories perceived most useful by stage

This work aims to directly examine how different SUI features can support distinct macro-level task stages, through a user study using a custom search system called SearchAssist (see §3.2). Tasks were designed to take users through pre-focus, focus, and post-focus task stages [30] in order to gather active, passive, and subjective measures of when SUI features provide most value and support. More specifically, we have three research questions.

RQ1 *How does the user's search stage influence active behaviour at the interface level?*

For RQ1, we looked at *active* behaviour, the behaviour which can be directly and indirectly determined from logged interaction, such as clicks and submitted queries. Our main finding is that some features such as *informational* features (providing information about results) are used frequently throughout, while *input* and *control* features (for refinement of results) are used less frequently after the first stage.

RQ2 *How does the user's search stage influence passive behaviour at the interface level?*

For RQ2, we looked at *passive* behaviour, i.e. behaviour not typically caught in interaction logs, such as eye fixations and mouse movements. Our main finding is the difference with the active results: evidently, users look often at actively used features, but other features that are less actively used (such as the recent queries feature) are more used in a passive way, suggesting a different type of support offered by these features.

RQ3 *How is active and passive behaviour reflected in the perceived usefulness of features?*

For RQ3, we were interested in the subjective opinions of users about the usefulness of features; this data also formed a reference point for interpreting other observed data from the previous research

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHIIR '16, March 13 - 17, 2016, Carrboro, NC, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3751-9/16/03 ...$15.00.
DOI: http://dx.doi.org/10.1145/2854946.2854957 .

questions. Our main finding is that the perceived usefulness of features differs radically per search stage, as summarised in Figure 1. First, the most familiar *input* and *informational* features (the search box and results list) were perceived as very relevant overall, but declined after the initial stage. Similarly, a set of assistive *control* features (search filters, tags and query suggestions), less commonly included in SUIs were also perceived as most useful in the beginning, but less useful in consecutive stages. Third, *personalisable* features (query history and a feature to save results), are considered as less useful in the beginning, but their usefulness significantly increases over time, even surpassing the value of common SUI features. Hence, our results indicate that the macro-level process has a large influence on the usefulness of SUI features.

2. RELATED WORK

This section discusses related work in the context of task-based information seeking and searching, search user interfaces, and the utility of SUI features over time.

As Toms [28] has indicated, the "*raison-d'être* of information retrieval systems is to deliver task-specific information that leads to problem resolution." Tasks may have different levels: a *work task* may be composed of several *search tasks*, set in a particular *environment* [28]. Categorizations of tasks may include complexity and specificity [35, 36]. For instance, tasks can range from simple lookup tasks, to exploratory and open-ended tasks [21]. Past research has shown that search behaviour varies significantly by task type [20]. Complex tasks may involve learning, and "understanding, sense-making and problem formulation are essential" [3]. In this paper, we use the often-used paper writing task, as employed by [30] and [16], to study information seeking and information searching.

2.1 Information Seeking & Searching

Wilson [40] has differentiated between information behaviour, information seeking and information searching. Information behaviour, encompasses the "totality of human behavior in relation to sources and channels of information" [40]. Information seeking is related to "searching or seeking information using information sources and (interactive) information retrieval systems" [12]. Finally, information searching, as a subset of information seeking, looks at the *interaction* between information users and the information system.

At the level of information *seeking*, various models exist which describe the information seeking process from a macro perspective. These models include for instance Wilson's problem-solving model [40] and Foster's non-linear model [8]. Ellis [7]'s model includes behavioural patterns of information seeking, which are not necessarily linear. Carol Kuhlthau, in her Information Search Process (ISP) model [16], describes a more sequential and temporally-based set of stages. Based on a number of longitudinal studies, Kuhlthau found "common patterns" in tasks involving learning and construction, going through six phases: *Initiation*, *Selection*, *Exploration*, *Formulation*, *Collection* and *Presentation*. The thoughts, feelings, uncertainty, and actions of a user rise, fall, and evolve as the users pass through different stages. Vakkari [30] later refined Kuhlthau's model and summarized its stages into *pre-focus*, *focus formulation* and *post-focus* stages. By studying students at three stages during a semester-long project, Vakkari found changes in relevance judgements, search tactics, terms and operators across stages.

Whilst information seeking models may inform the general design of IR systems, information *search* models (or information retrieval interaction models) often times focus on the *means* to improve design, specifically the interaction between users and information systems. This includes Spink's model of the IR interaction process [27], which describes cycles of interaction with IR systems, including user judgements, search strategies, tactics and moves. Saracevic' Stratified model of Information Retrieval Interaction [25] views IR interaction as a dialogue between user and computer, and includes different levels (strata) of interactions. Finally, Marchionini's Information-seeking Process Model consists of various sub-processes and their relationships (e.g. 'define problem', 'select source' and 'formulate query'). These are *micro*-level models, which can help us to *design* novel SUI features.

2.2 Search User Interfaces

Search user interfaces (SUIs) serve as an intermediary between the user and the underlying data in an information retrieval system. As Hearst [10] indicates, SUIs aid "users in the expression of their information needs, in the formulation of their queries, in the understanding of their search results, and in keeping track of the progress of their information seeking efforts." SUIs may be designed in vastly different ways, though designing effective SUIs with a high usability is a complex process, as Shneiderman and Pleasant [26] argue, and it often involves finding trade-offs in simplicity and functionality. This difficulty in designing effective SUIs has led to a growing number of guidelines and theories [26].

Research into Search User Interfaces (SUIs) [10, 24, 38] has suggested many different interactive features, from search suggestions [22] to facets [29] to personal spaces to collect useful results [6]. Though their usefulness has been proved in various studies, most of these features have not been adapted in common search engines. Hearst suggests some underlying reasons for the lack of adoption of advanced features: searching is used as a means to achieve a broader aim, search is mentally intensive and search systems should be understandable for people with different knowledge and experience [10]. Hence, overly complex search engines may distract from a user's core task. Furthermore, the usefulness of features may depend on the task type and complexity. Some work ties the need for advanced features to different types of tasks, such as Exploratory Search tasks [33]. Although Diriye [5] argued that in the context of known-item search tasks, excessive search features may impede people's information searching, most tasks involve at least some exploratory elements [32].

Given the multitude of features which could potentially be integrated in SUIs, it may be useful to divide the types of features in different groups, based on their functions. Wilson [38] proposed a taxonomy, which distinguishes four groups of interface features. *Input* features aid users in expressing their needs, *control* features allow users to restrict or modify their input, *informational* features provide results or information about results, and *personalisable* features are tailored to the search experience of a user. In this paper, we use this categorization to analyze the usefulness of features in different stages of search.

2.3 Utility of SUI Features Over Time

A number of user studies have looked at the utility of search system features across stages, but most authors consider 'stage' a temporal segment of a singular session, and have retrospectively identified stages in people's search. Very few authors, however, have used an explicit multistage task design. Liu and Belkin [19], for example, used one motivating work task but performed during three distinct sessions. They looked at the influence of task stage, type and topic knowledge on the interpretation of dwell time over multiple task sessions. While not directly looking at the use of SUI features, they found that task stage and topic knowledge could help to interpret time as an indicator of usefulness. Similarly, Wilson and Schraefel [39] conducted a longitudinal study of keyword and faceted search, finding that the latter only occurred after the second visit of an online video archive (likely due to confidence and interface understanding).

Other authors did not perform longitudinal studies, but used various simulated work tasks, performed during one session. Kules and Capra [17] looked at searchers' interactions with a faceted library catalog. Using a number of assigned exploratory search tasks, they examined differences in gaze behaviour on four SUI features (query box, results, facets, and breadcrumbs). Users were asked to retrospectively assign stages to segments of their search sessions, using a customly defined set of stages (most similar to the micro-analysis of search discussed before). Kules and Capra found differences in the use of Facets, which were used more in 'decision making stages.' Similarly, White et al. [34] have looked at implicit and explicit Relevance Feedback (RF), and divided search sessions into equal parts ('beginning', 'middle' and 'end') to look at stage differences. Their results indicate that implicit RF is used more in the middle stages, while explicit RF is used more towards the latter stages, and there is also an influence of task complexity.

In their user study, Niu and Kelly [22] also divided search sessions into temporal segments, rather than explicit stages, and found that query suggestions were frequently used for difficult topics and during later task stages. They suggest that participants, in the latter parts of a task, "may be exploring the various facets of the topic and/or looking for specific information". In addition, they may have exhausted their original ideas and "need alternative queries." Similarly, Diriye et al. [5] performed a user study using a rich experimental search interface. By looking at the temporal distribution of the use of four interface features during a task, they found that certain features (starter pages and search box) were search stage sensitive and other features search stage agnostic (facets and filters).

Finally, Huurdeman and Kamps [11] looked at conceptual ways to bridge macro and micro-level information seeking models, and based on changes in gaze behaviour of a small-scale user study involving book search, found evidence for differences in the use of *input* and *personalisable* features over time, such as the query box and book basket, while other features were used throughout the task.

As opposed to previous literature, mainly studying singular tasks, we use an explicit multistage approach to look at the passive and active utility of a different SUI features across macro-level stages, to provide richer insights into exactly when different types of features become more useful.

3. USER STUDY SETUP

This section details the experimental setup of the user study. To study the active and passive use of interface features in different search stages, we conducted a within-participants user study with task stage as the independent variable. For dependent variables, active system interactions were logged, passive mouse and eye movements were tracked, and questionnaires were used to collect data on perceived usefulness. Participants made use of *SearchAssist*[1], an experimental search system similar to a regular Web search engine, with different categories of SUI features potentially useful for each stage.

3.1 Task Design and Participants

While some prior work has inferred task stage, we constructed 3 task descriptions to explicitly represent three key *stages*, inspired by previous literature on tasks involving learning and construction [16, 30]. Stage 1 was modeled after the initial stages of Kuhlthau's ISP model (initiation, topic selection, exploration), summarized by Vakkari [30] as the pre-focus stage. Stage 2 was aimed to make users formulate their specific topic (focus formulation), and a question about this topic. Finally, Stage 3 was based on the final stages of Kuhlthau's model (collection and presenting), summarized by Vakkari as the *post-focus* stage. In this stage, users had to collect sources relevant to their focused topic, and to provide a draft answer to the formulated question about their topic.

Table 1: Assigned multistage tasks

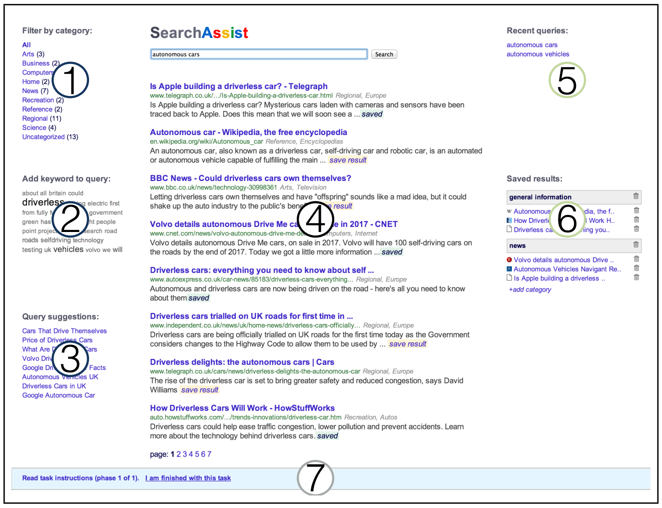
Introduction: For a class called "Computers in Society", the professor has given you the assignment to write a 5-page essay on some aspect of [topic]. Having a good grade for this essay is critical for passing the course. The essay is due in a week, but you have yet to decide on an exact topic. In a deliverable due tomorrow, you have to define your topic, a specific question about the topic and a list of sources.
Stage 1: Prepare a list of at least 3 ideas for a topic to write about in the context of [topic]. They should cover many different aspects of the topic, and unusual or provocative ideas are good. Search the web using the SearchAssist system to find out what information is available. Write down your ideas for topics in the text field below. Save any webpages you encounter via the SearchAssist system which are useful for writing on these topics (utilizing the "save result" feature).
Stage 2: Select one of the topics which you defined in the previous task. Choose the topic which interests you most, about which you are able to find enough information, and which you think you are able to finish in the allotted time. Use the SearchAssist system to find information to help you to decide on the topic, and save sources if needed using "save result". Write down the topic in the text box below. After having selected a topic you ask yourself the question "what is it that I want to find out about this topic?" Search the web using the SearchAssist system and formulate a specific question you would like to ask about this topic. You can save any pages you encounter which are useful for answering this question.
Stage 3: To be able to start writing your essay, take the specific question you have formulated in the previous step, and gather as much useful information as you can by searching the internet using SearchAssist. Find around 20 additional pages. Select the 5-10 pages that you could cite in your essay, and which are most relevant for answering the question you formulated in the previous step. If you have time left, formulate a draft answer to your question based on the information you have encountered (max. 300 words) and write it in the text box below.

Written as simulated work tasks [1] (see Table 1), the stage descriptions used elements of exploratory work tasks from previous studies [18, 19], focused on the often used 'essay writing' task. Following Borlund's guidance [1], the simulated work tasks were designed so that participants could relate to them, that they were topically interesting, and would add 'enough imaginative context'. After pilot tests and discussions with staff, two topics were selected: 'virtual reality', and 'autonomous vehicles'. The participants were undergraduate students of the School of Computer Science of the University of Nottingham (UK campus). The participants were recruited via posters, the Facebook page of Mixed Reality Lab, e-mails, and via *callforparticipants.com*. Upon completing the experiment, participants received a £10 Amazon voucher, and an additional £25 Amazon voucher was awarded to the participant with the best task outcome. In total, 26 participants joined the experiment. Two participants, however, were excluded from our analysis, where one was unable to complete all three stages, and the eye tracking data was not sufficiently accurate for the other. Of the remaining 24 participants, 18 were male and 6 were female; 22 participants were aged 18-25, and 2 were between 26-35.

3.2 Data and Interface

For this study we designed *SearchAssist*, an experimental search system based on PHP, Javascript and MySQL, depicted in Figure 2. Search results, query corrections and query suggestions were retrieved in the JSON format via the Bing Search API and displayed as a familiar Web interface, similar to common Web search engines. The use of the Search API allowed participants to access a variety of sources, including scholarly, encyclopedic and news sources.

[1] The source code of SearchAssist and eye tracking software used in this study is available via: https://github.com/timelessfuture/SearchAssist



Filter by category:

All
Arts (3)
Business (2)
Computers
Home (2)
News (7)
Recreation (2)
Reference (2)
Regional (11)
Science (4)
Uncategorized (13)

Add keyword to query:

about all britain could driverless ing electric first from fully f government green has ht people point projec search road roads selfdriving technology testing uk vehicles volvo we will

Query suggestions:

Cars That Drive Themselves
Price of Driverless Cars
What Are D Cars
Volvo Driv
Google Dri Facts
Autonomous Vehicles UK
Driverless Cars in UK
Google Autonomous Car

SearchAssist

autonomous cars

Is Apple building a driverless car? - Telegraph
www.telegraph.co.uk/…/Is-Apple-building-a-driverless-car.html *Regional, Europe*
Is Apple building a driverless car? Mysterious cars laden with cameras and sensors have been traced back to Apple. Does this mean that we will soon see a ... *saved*

Autonomous car - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Autonomous_car *Reference, Encyclopedias*
An autonomous car, also known as a driverless car, self-driving car and robotic car, is an automated or autonomous vehicle capable of fulfilling the main ... *save result*

BBC News - Could driverless cars own themselves?
www.bbc.co.uk/news/technology-30998361 *Arts, Television*
Letting driverless cars own themselves and have "offspring" sounds like a mad idea, but it could shake up the auto industry to the public's bene *e result*

Volvo details autonomous Drive Me ca e in 2017 - CNET
www.cnet.com/news/volvo-autonomous-drive-me-de *mputers, Internet*
Volvo details autonomous Drive Me cars, on sale in 2017. Volvo will have 100 self-driving cars on the roads by the end of 2017. Today we got a little more information ... *saved*

Driverless cars: everything you need to know about self ...
www.autoexpress.co.uk/car-news/85183/driverless-cars-everything... *Regional, Europe*
Autonomous and driverless cars are now being driven on the road - here's all you need to know about them *saved*

Driverless cars trialled on UK roads for first time in ...
www.independent.co.uk/news/uk/home-news/driverless-cars-officially... *Regional, Europe*
Driverless cars are being officially trialled on UK roads for the first time today as the Government considers changes to the Highway Code to allow them to be used by ... *save result*

Driverless delights: the autonomous cars | Cars
www.telegraph.co.uk/cars/news/driverless-delights-the-autonomous-car *Regional, Europe*
The rise of the driverless car is set to bring greater safety and reduced congestion, says David Williams *save result*

How Driverless Cars Will Work - HowStuffWorks
auto.howstuffworks.com/…/trends-innovations/driverless-car.htm *Recreation, Autos*
Driverless cars could help ease traffic congestion, lower pollution and prevent accidents. Learn more about the technology behind driverless cars. *saved*

page: **1** 2 3 4 5 6 7

Recent queries:

autonomous cars
autonomous vehicles

Saved results:

general information
Autonomous dia, the f..
How Driverl l Work H..
Driverless ca ning you..

news
Volvo details autonomous Drive ..
Autonomous Vehicles Navigant Re..
Is Apple building a driverless ..
+add category

Read task instructions (phase 1 of 1). I am finished with this task

Figure 2: Screenshot SearchAssist. *Left column* (1, 2, 3): control features. *Middle* (4): input and informational features. *Right Column* (5, 6): personalisable features. (7): task bar

The SearchAssist interface consisted of the following elements: *1. Category filters*. Using the category filters, searchers could filter the set of results. The categories were derived from the top-level categories of the Open Directory Project (DMOZ). Retrieved results were matched against all DMOZ categories using the hostname of each result, and the top-level categories could be used to filter the result set. *2. Tag cloud*. Using the tag cloud, it was possible to add one or more keywords to a query. The tag cloud was generated based on the most frequently occurring words in the snippets of the first 50 retrieved results. *3. Query suggestions*. Query suggestions were retrieved from the Bing Query Suggestions API, and they could be clicked to perform a new search. *4. Search box and results*. The SearchAssist interface featured a standard search box and results were retrieved from the Bing Web Search API; the Bing Spelling Suggestions API was also used. Each resultset item contained the title of the page, a URL, the DMOZ category, the snippet and a button to save a result. To better facilitate eye tracking, 8 search results were displayed at a time, similar to e.g. [13]. *5. Recent queries*. The recent queries feature showed the last 15 queries performed across all tasks of the experiment, and allowed them to be resubmitted to the search engine. *6. Saved results*. The saved results feature allowed users to view (and remove) saved webpages, to reorder collected webpages by dragging and dropping, and to add (or remove) category labels to the gathered results. *7. Task bar*. The task bar contained task-related material, including a link to open the task instructions and a link to finish the current task, after which a user was prompted to fill out the corresponding questionnaire. The task instructions were shown in a Google Doc, which was also used to collect their responses.

3.3 Protocol

The experiment started with signing the consent forms and a pre-questionnaire, asking for demographics and ratings for knowledge about the potential task topics. As domain experts would behave differently than domain novices, participants were assigned the topic that they knew least about. Participants were then introduced to the features of the experimental system via a structured Powerpoint presentation, and given a training task (approx. 5 min.), which was used to mitigate the familiarity affects in the study, and to check the calibration of the eye tracker. The task stages were performed in sequence; the stage order could not be counter-balanced without losing the cumulative learning required from stage to stage. Participants were given 15 minutes for each stage, including a one minute warning, however participants were allowed to continue after this final minute passed. After each stage, users filled out a questionnaire about the perceived usefulness of features. After the final stage, participants also completed the post-questionnaire and a short debriefing interview (taking 5-10 minutes), focused on their experiences with the system. The total time to participate in the experiment varied between 55 and 90 minutes.

3.4 Logging and Eye Tracking

The system logged the active and passive interactions in three ways: via system logging, browser history and eye tracking, and the experiment was carried out using the Chromium browser. After each experiment, the browser history was exported in JSON format using the "Export history" browser extension, and the local browser history was deleted. All user actions were saved in a database via MySQL, and as plain text files using Log4Javascript. The logged data included all clicked interface features, all entered text (in the query box), and which page was active in the browser (the search interface, a webpage or the task page). In addition, all results items, query suggestions and query corrections retrieved via the Bing API were saved in their original JSON format.

For passive behaviour, the system logged the position of the mouse cursor, and for context, took a screenshot of the user's screen four times per second. Eye tracking was performed using the EyeTribe eye tracker, calibrated using the included software. The Python-based PyGaze framework [4], and the PyTribe toolbox (a Python wrapper for the EyeTribe eye tracker) were customized to our needs and tightly integrated with the experimental interface. For the eye tracking data, the fixation counts and durations were calculated. Fixations were considered as sequences of eye tracking measurements within a 25 pixel radius; within a timeframe of at least 80ms (similar to e.g. [2]). We defined bounding boxes for each SUI element of the *SearchAssist* interface to detect the Area of Interest (AoI) of the fixation. In addition, to derive the depth of results list items inspected, we defined a bounding box for each results list item. The same methods were used to calculate the counts and duration of mouse movements in each AoI.

3.5 Data and Task Validation

First, we sought to confirm that the two topics, 'virtual reality' (VR) and 'autonomous vehicles' (AV), were comparable. No significant differences were found between overall task time, number of queries, results viewed, nor in the majority of usefulness ratings.

Only one significant difference was found, using the Mann-Whitney test, in the post-stage usefulness ratings for the 'saved results' feature for the first (U=30, p=0.01) and second stage (U=34, p=0.02), although logged usage of this feature was not significantly different. Informal observations indicate that there may have been a higher number of *relevant* results that *could* be found in the AV topic, but that these have not affected the majority of behaviour. Overall, however, we conclude that the topics invoked comparable behaviours and continue to analyse the data from both topics as a single set.

Second, we examined the validity of our task descriptions in terms of invoking correct stages. In post-stage questionnaires users selected the activities they had conducted from a randomized list derived from Kuhlthau's model[2]. For the first stage, the most commonly selected activity was 'exploring' (N=17), followed by 'gathering' (16); corresponding to the initiation and exploration activities associated with the initial stages of Kuhlthau's model. After the

[2] Specifically: *exploring*, *focusing*, *formulating*, *collecting*, *gathering*, *becoming informed*, *choosing*, and *getting an overview*

second stage (focus formulation) users most often chose ‘focusing’ (16) and ‘collecting’ (12) as words representing their activities. The common use of focusing corresponds to the focus formulation activity, while collecting may refer to the collected documents in that stage. Finally, for the third task also ‘focusing’ (17) and ‘collecting’ (14) were the most common words. We conclude that even though the separations between stages are not always dichotomous, our experiment correctly invoked the main activities in each stage.

4. RESULTS

This section examines the results of the study, and whether the participants showed distinct behaviour in the different stages of their overall task. Together, the 24 participants issued 502 queries and clicked on 684 results. Participants spent an average of 32:56 minutes to complete the 3 task stages. Of this time, 36.8% was spent in the SUI, 33.0% on the task screen, and 30.2% on the webpages. Participants spent, on average, 11:32 minutes on the first stage, 8:24 minutes on the second stage, and 12:59 minutes on the third stage.

4.1 Search Stage & Active Behaviour

This section focused on our first research question (**RQ1**): How does the user’s search stage influence active behaviour at the interface level? We define active behaviour as the behaviour that can be directly and indirectly derived from the logged interactions, such as clicks, queries, and pages visited.

SUI features Table 2 summarizes the main interaction with each available SUI feature. The use of the *Query Box* (counted as the clicks on the ‘search’ button) is most frequent in the first stage, and decreases in the second and third stage. Using the within-participants, repeated measures ANOVA, we found a significant difference in the use of the search button (p<0.01, F(2)=13.6). Post hoc tests, using the Bonferroni correction, showed that there is a significant difference between the first and second (p<0.01), and the first and third stage (p<0.01). Hence, users use the search button more in the first stage, most likely to explore the assigned topic [37].

The clicks on retrieved results items, via the *Results List* feature, remain more or less stable without significant differences per stage. The number of times a result is saved using the adjacent ‘save result’ link, however, is decreasing after the first stage. Users also appear to examine the results beyond the first page more frequently in the third stage (by clicking ‘next page’) but these differences, like the differences in result clicks and result saves, are not significant.

The *Category Filters* feature is used significantly less frequently after the first stage, and thus seem to be most useful in the initial task stage (p<0.01, F(1.2)=8.6, Greenhouse-Geisser correction). The differences, with Bonferroni correction, are most prominent between the first and third stage (p<0.01), but also occur between the other stages (1->2: p=0.03, 2->3: p=0.03). Similarly, the clicks on the *Tag Cloud* feature are significantly different (p<0.01, F(1.4)=8.5, Greenhouse-Geisser correction). Again, the first stage features the highest number of clicks, and using a pairwise comparison, with Bonferroni correction, we found significant changes in clicks between the first and second stage (p=0.02), and between the first and third stage (p=0.01).

Compared to the other features, the *Query Suggestions* and *Recent Queries* features are not frequently used, and a slight decrease in use of the Query Suggestions and a slight increase in the use of the Recent Queries feature is visible in the data, but are not significant.

Although the differences in the use of the ‘Save result’ link in the Results List were not significant, the statistics for the *Saved Results* feature indicate that users add categories to these items mostly in the first stage (p<0.01, F(2)=8.1). Pairwise comparisons, with Bonferroni correction, show that the significant differences occur between the first and second (p=0.02), and the first and third (p<0.01) stage. Hence, participants save and categorize items in the saved results list most frequently in the first stage. The clicks on the saved results (bookmarks), on the other hand, are clearly most frequent in the last stage (p<0.01, F(1.1)=18.8, Greenhouse Geisser correction). A pairwise comparison shows significant differences between the first and third stage (p<0.01), and the second and third stage (p<0.01). Finally, the last stages show a slight increase in the removal of categories and saved items, as opposed to the additions in the first stage, but no significant differences were found.

Table 2: SUI active interaction (clicks), from system logs

mean	stage1	%	stage2	%	stage3	%
Query Box						
search clicks**	8.4	*24.3*	4.5	*19.8*	4.6	*14.9*
Results List						
result clicks	7.3	*20.9*	5.5	*24.2*	7.8	*25.2*
result saves	6.1	*17.5*	4.3	*19*	3.7	*11.8*
next page clicks	0.8	*2.4*	1.2	*5.1*	1.7	*5.5*
Category Filters						
clicks**	2.9	*8.3*	1.1	*4.7*	0.6	*2*
Tag Cloud						
clicks**	1.6	*4.7*	0.7	*3.1*	0.5	*1.7*
Query Suggestions						
clicks	0.8	*2.3*	0.4	*1.7*	0.5	*1.7*
Recent Queries						
clicks	0.3	*1*	0.6	*2.5*	0.8	*2.4*
Saved Results						
clicks**	0.7	*2*	0.9	*3.9*	6.3	*20.4*
add category**	2.2	*6.4*	0.9	*3.9*	0.6	*1.9*
move item	3.5	*10*	2	*8.8*	2.4	*7.7*
remove category	<0.1	*<0.1*	0.3	*1.1*	0.3	*0.9*
remove item	<0.1	*<0.1*	0.5	*2.2*	1.2	*3.8*
Total	34.7	*100*	22.8	*100*	31.1	*100*

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

Queries & Page Visits As we observed in Table 2, participants used the Query Box feature most often in the first stage. Now, we look at the queries in more detail, summarized in Table 3. The total number of queries submitted, including tag cloud suggestions and use of the Recent Queries feature, is significantly different per stage (p<0.01, F(2)=8.9). A pairwise comparison, with Bonferroni correction, indicates that the differences are significant between the first and second (p<0.01), and between the first and third stage (p<0.01). Likewise, the unique queries are significantly different (p<0.01, F(2)=7.9), again with a significant difference between the first and second (p<0.01) or third stage (p<0.01). Most queries performed were unique, though there is some overlap in the queries between the first, second and third stage, meaning that participants reuse queries in latter parts of the experiment (i.e. by re-entering the same query or using the Recent Queries feature). In the first stage, the majority of queries are initiated from the Query Box. However, subsequent stages show an increase of the relative use of the Recent Queries feature, and a stable share of the Query Suggestions.

While the number of queries decreases after the first stage, the number of *words per query* increases. The highest mean number of query words occurs in the second stage (4.5), and an almost equally high value during the third stage (4.4). The higher number of queries may be related to exploration activities in the first stage, which require various queries to explore various topics. The increasing number of query words, on the other hand, may occur because a person is searching for a more specific topic, and may have built a conceptual representation of a topic [30]. For example, one user (P.02) started with the query “virtual reality” in the first stage, but queried for “the impact of virtual reality on society art and gaming

Table 3: SUI active interaction (queries and page visits)

mean	stage1	stage2	stage3
Queries**	9.5	5.5	5.9
*via Query Box***	*88%*	*81%*	*78%*
via Recent Queries	*3%*	*11%*	*13%*
via Query Suggestions	*8%*	*7%*	*8%*
Unique queries**	8.1	5.1	5.3
Overlap queries prev. stages	0	1.4	1.8
Mean num. query words**	3.2	4.5	4.4
Levenshtein distance (query diversity)	13.2	13.9	17.0
Visited pages**	8.0	6.4	14.2
via Results List	*91%*	*86%*	*56%*
*via Saved Results***	*9%*	*14%*	*44%*
Unique visited pages**	7.3	5.9	10.8
Overlap visited pages prev. stages	0	0.8	2.8
Mean rank visited pages	3.1	5.1	6.4

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

culture" in the third stage. Or, the queries from another participant (P.06) evolved from short queries such as "autonomous vehicles" to longer queries like "autonomous vehicles costs insurance industry". The differences in the number of query words are significant (p<0.01, F(2)=5.3), specifically between the first and second stage (p<0.01, Bonferroni correction). Finally, we calculated the query diversity, based on the Levenshtein distance between all pairs of unique queries of a user in a certain stage. The query diversity is similar during the first and second stage, but is highest in the third stage, meaning that the edit distance between users' queries is greater; although these differences are not significant.

Participants in our experiment visited the highest number of pages in the third stage (p<0.01, F(1.3)=11.6, Greenhouse-Geisser correction), when collecting materials. The differences are significant between the first and third stage (p=0.02) and between the second and third stage (p<0.01). This variance seems to be explained primarily by the revisiting of pages from the Saved Results feature (p<0.01, see previous section), as page visits from the Results List were not significantly different. This finding is reflected in the uniquely visited pages (p<0.01, F(1.5)=8.1, Greenhouse-Geisser correction), but here the only significant changes occur between the second and third stage (p<0.01). Further, the result is also reflected in the mean dwell times on the webpages, which are highest in the first (12.9 sec.) and second stage (14.4 sec.), but lower in the third stage (8.9 sec.). The dwell times are significantly different (p<0.01, F(2)=7.8), between the first and second (p<0.01), and between the second and third stage (p<0.01). Participants also explored further down the result set in the later task stages, with the mean visited rank increasing from 3.1 to 5.1 and 6.4 respectively, but this was not significant within our current sample.

Summarizing, this section has focused on the active interaction with the system during the experiment. Utilizing the categorizations of Wilson's framework for SUI features [38], the results show various tendencies: *input* features (the Query Box) and *control* features (Category Filters, Tag Cloud and Query Suggestions) are clearly used less often in subsequent stages, while the use of the *informational* (Results list) features remains stable. The results for the *personalisable* features (Recent Queries and Saved Results) differ. The Recent Queries feature is scarcely used, but an increasing tendency can be observed across stages. Similarly, users mostly click on their saved results in the last stage, but save the actual results and add categories most frequently in the first stages. Hence, the Saved Results feature is initially used to store and categorize important results, but later to revisit previous results. Also, users start out with a significantly higher number of queries, as compared to later stages, while the number of page revisits substantially increases in the last stage. Evidence for the learning aspects of the used tasks are found in the increase of the number of query words and query diversity [16, 30], as users seem more able to express their needs in queries.

Table 4: Passive use: mouse hovers *not* leading to a click

mean	stage1	%	stage2	%	stage3	%
Query Box**	344.7	*16.6*	250.2	*19.5*	210.2	*14.6*
Results List**	1226.8	*59.1*	701.9	*54.7*	872.9	*60.7*
Category Filters**	124.6	*6.0*	57	*4.4*	67.7	*4.7*
Tag Cloud*	165.9	*8.0*	73.1	*5.7*	47.2	*3.3*
Query Suggestions	91.3	*4.4*	58.5	*4.6*	56.6	*3.9*
Recent Queries	17.6	*0.8*	18.3	*1.4*	21.3	*1.5*
Saved Results	103.7	*5.0*	123.5	*9.6*	163.3	*11.3*
Total	2074.6	*100*	1282.5	*100*	1439	*100*

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

Finally, another contrast can be observed, namely between commonly used features and scarcely used features. Together, the Query Box, Results List and Saved Results features take up over 80% of all clicks, while the remaining set of features takes up less than 20% of all clicks (see Table 2). The infrequent *active* use of certain features, in particular the Query Suggestions and Recent Queries features, lead to the question whether some features are perhaps used in a *passive* way, which we will further examine in the next section.

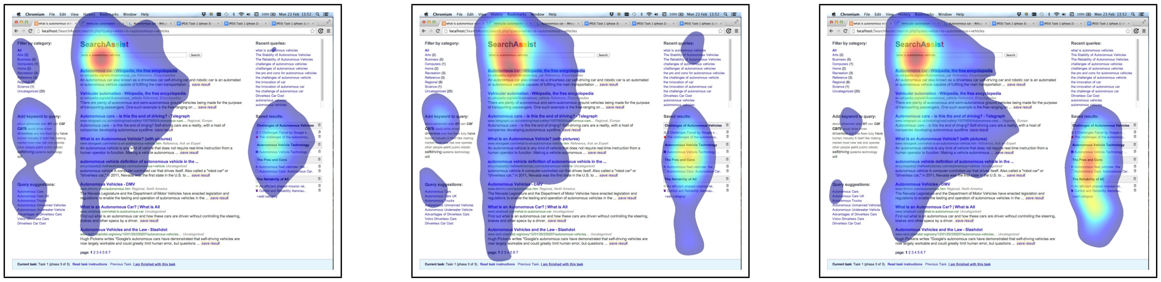
4.2 Search Stage & Passive Behaviour

In this section, we focus on the following research question (**RQ2**): How does the user's search stage influence passive behaviour at the interface level? We examine the user's mouse position and eye fixation data to look at the passive behaviour in each search stage.

Mouse hovers Participants' mouse movements can shed more light on the use and utility of SUI features in different stages. Mouse moves in a particular area can be simply movements to reach or click a SUI feature, but may also indicate different types of usage, i.e. mouse moves aiding users in processing the contents of results pages [23]. In our analysis, we look at the *passive* mouse movements: the mouse hovers in a SUI feature area that did *not* lead to a click.

Table 4 shows the mean count of mouse movements over time. We counted mouse hovers (defined as a change in the coordinates of the mouse pointer) within each SUI feature's Area of Interest. There are significant differences for the following features: the Query Box (p<0.01, F(2)=6.4), the Results List (p<0.01, F(2)=6.9), the Category Filters (p<0.01, F(2)=7.0) and the Tag Cloud feature (p=0.03, F(1.5)=4.5, Greenhouse-Geisser correction). Mouse hovers in these SUI areas are most common in the first stage, and significantly decrease in the second or third stage. The other features do not show significant changes over time. The results for this measure show overlap with the active interaction measure of the previous section, with the exception of a "dip" in mouse hovers on the Results List in the second stage, and a higher relative amount of hovers over the Query Suggestions, especially in the second and third stage. The higher and more stable degree of mouse hovers around the Query Suggestions may indicate that users use this feature passively in all three stages, as opposed to the decreasing use tendency visible in the active use measure. To gain further insights, we next look at passive use, not even involving the mouse, using eye tracking.

Eye tracking fixations To gain an initial overview of eye movements within the SearchAssist interface, we generated heatmaps for each stage across all participants. These heatmaps (Figure 3) show the spatial distribution of the fixations. A visual inspection reveals a consistent focus on the Query Box and Results List SUI features in each stage (middle pane). The Category Filters, Tag Cloud and Query Suggestions features (left pane) are most intensively used in



(a) Stage 1

(b) Stage 2

(c) Stage 3

Figure 3: Eye tracking heatmaps, based on fixations (over 80ms)

the first stage, while the Saved Results feature (lower right panel) appears to be most intensely used in the last stage.

These differences can be inspected in more detail using the absolute and relative fixation counts (with a minimum duration of 80 ms), summarized in Table 5. For the most part, the results for the passive use of SUI features confirm the results regarding active use. The number of fixations on the Query Box is significantly decreasing after the first stage (p=0.01, F(2)=4.9), which is comparable with the lower number of unique queries performed in the second and third stage observed in the active interactions. In particular, the difference is significant between the first and second stage (p=0.02). In addition, the less frequent active use of the Tag Cloud and Category Filters is reflected in a decrease in the number of fixations in the second and third stage, and this difference is significant for both Category Filters (p=0.01) and Tag Cloud (p<0.01). A pairwise comparison reveals significant differences between the first and second stage for the Category Filters (p=0.03), and between the first and second (p=0.01) or third stage (p=0.02) for the Tag Cloud.

The fixations on the results list decrease significantly after the first stage (p=0.02, F(2)=4.4). A significant difference for the fixations on the Results List feature exists between the first and second stage (p<0.01, Bonferroni correction), though the relative degree of fixations changes less. Table 2 in the previous section, however, did not show a significant difference for the number of clicks on resultset items in any stage. Similarly, the decreasing number of clicks on the Query Suggestions features are coupled with a lower number of fixations on this feature. These differences are significant (p=0.04, F(2)=3.5) between the first and second stage (p=0.01). As in the case of the active interactions, the Recent Queries feature does not show a significant difference, but the relative values for the fixations increase in the second stage. Finally, the fixations on the Saved Results feature rise during the stages, which is similar to the measured increase in the previous section, but the difference for the fixations is not significant (p=0.09).

The previous section showed some features which were used frequently, in particular the Query Box, Results List and Saved Results feature, and other features which were used infrequently, such as the Query Suggestions and Recent Queries. We would expect that the often-used SUI features also have a high degree of fixations. The results confirm this: the Results List takes up more than half of the fixations, and also the relative degree of the fixations on the Query Box and Saved Results is high. There is a difference, however, for features that were little used in an active way, such as the Query Suggestions and Recent Queries features. The percentage of fixations on the Query Suggestions over all three stages is 3.6% instead of 1.9% of clicks, and the fixation percentage for the Recent Queries is 3.01% instead of 1.97%. While the difference is relatively small, it does provide evidence that participants may use these features more passively than actively. Another difference can be observed for the Category Filters and Tag Cloud: participants look more at the Tag Cloud (5.8%), than they click on it (3.1%), while a contrary situation exist for the Category Filters (5% of all clicks, but 3.8% of all fixations).

Table 5: Passive SUI use: mean eye tracking fixation count

mean	stage1	%	stage2	%	stage3	%
Query Box*	58.08	*14*	35	*13.1*	41.42	*11.8*
Results List*	224.88	*54.3*	139.83	*52.5*	187.17	*53.5*
Category Filters*	17.63	*4.3*	10.46	*3.9*	11.42	*3.3*
Tag Cloud**	31.71	*7.7*	14.58	*5.5*	15.5	*4.4*
Query Suggestions*	16.88	*4.1*	9.71	*3.6*	10.83	*3.1*
Recent Queries	10.92	*2.6*	9.79	*3.7*	10.13	*2.9*
Saved Results	54.38	*13.1*	47.17	*17.7*	73.63	*21*
Total	414.48	*100*	266.54	*100*	350.1	*100*

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

Summarizing, on the one hand, the fixations and mouse movements by users validate the active behaviour, showing similar tendencies. The significant differences in the use of the Query Box, Category Filters and Tag Cloud confirm the findings from the previous section, while the eye tracking data also suggests significant changes in the use of Query Suggestions. On the other hand, subtle differences exists in the passive use of less often used features, such as the Query Suggestions and Recent Queries features. This suggests that some features may not be used often in an active way, but that they are still used passively. In Section 4.3, we validate and contextualize these findings with subjective ratings of usefulness and qualitative feedback from participants.

4.3 Search Stage & Perceived Usefulness

Our final research question looks at the potential influence of search stages on the *perceived* usefulness of SUI features: (**RQ3**): How is active and passive behaviour reflected in the perceived usefulness of features? Findings from questionnaires after each stage, after finishing the whole experiment, and brief post-experiment interviews are used to contextualize the findings so far.

Usefulness ratings After each task stage, participants were asked to rate the usefulness of each SUI feature of SearchAssist on a Likert scale of 1 to 7, as shown in Table 6. Somewhat expectedly, the most highly rated features are the Search Box and Results List features. As it turns out, however, this is closely followed by the Saved Results feature, which was also deemed to be very useful by most of the participants. Conversely, the least popular features among the participants were the Tag Cloud and Category Filter features. The most useful features were also rated most consistently among participants: the standard deviation values for the Search Box, Results List and Saved Results are substantially lower than for the other SUI features. Conversely, the most "controversial" feature was the Tag Cloud, with a standard deviation of 1.71, suggesting a relatively high variability of user ratings: some participants found it useful, and others did not perceive it as useful.

Table 6: Mean usefulness ratings, gathered after each stage (s.dev.). *Bold: stage with highest rating for feature.*

mean	stage 1	stage 2	stage 3	total
Search Box/Results*	**6.67 (0.7)**	6.33 (0.9)	6.08 (1.1)	6.36 (0.9)
Category Filters	**4.08 (1.5)**	3.79 (1.6)	3.46 (1.7)	3.78 (1.6)
Tag Cloud	**3.92 (1.7)**	3.54 (1.5)	3.63 (1.9)	3.70 (1.7)
Query Suggestions	**4.80 (1.4)**	4.00 (1.7)	4.00 (1.6)	4.26 (1.6)
Recent Queries*	3.46 (1.6)	4.13 (1.7)	**4.71 (1.6)**	4.10 (1.6)
Saved Results	5.83 (1.2)	6.17 (1.1)	**6.30 (0.9)**	6.08 (1.0)

*Non-parametric Friedman test: * significant (p<0.05)*

Table 7: Mean post-experiment usefulness ratings – at which moment were the SUI features most useful (% of participants).

perc	stage1	stage2	stage3
Query Box/Results List**	**100.00%**	75.00%	66.67%
Category Filters**	**54.17%**	20.83%	12.50%
Tag Cloud**	**41.67%**	16.67%	8.33%
Query Suggestions*	**54.17%**	29.17%	20.83%
Recent Queries*	12.50%	54.17%	**70.83%**
Saved Results **	37.50%	66.67%	**91.67%**

*Chi-square test: * significant (p<0.05); ** significant (p<0.01)*

Comparing the stages, the Search Box and Results, Category Filters, Tag Cloud and Query Suggestions are all rated most highly in the first stage, which generally corresponds with the results for the active and passive interaction in the previous sections. The inter-stage differences for the Search Box and Results List (non-parametric Friedman test, p<0.01, $\chi^2(2)$=13.3) are significant. The Query Suggestions feature has significance ratings close to 0.05 (Friedman, p=0.07, $\chi^2(2)$=5.4); and is deemed most useful in the first stage. While the previous features are rated slightly lower in successive stages, the opposite holds true for the Recent Queries and Saved Results features, which both have their highest rating in the third stage. In the case of the Recent Queries feature, the differences are significant (Friedman, p<0.01, $\chi^2(2)$=15.2). Here, we note that the Recent Queries feature did not show any significant differences using the previous active and passive interaction measures, though a general increase of use could be observed.

Table 7 summarizes the users' ratings after the *whole* experiment, which are also visualised in Figure 1. Participants were asked to indicate in which stage or stages a feature was *most* useful, and were allowed zero to multiple answers for each feature. This table shows similar tendencies as Table 6, but the differences are more pronounced. Hence, participants judged the usefulness of interface features slightly more explicit after completing the full experiment, perhaps at that moment having an overview of the stages involved in it. A chi-square test indicates that the differences are significant for all SUI features (p<0.01, $\chi^2(10)$=33.5). The feature ratings show a clear division: the Search Box/Results List, Category Filters, Query Suggestions and Tag Cloud were most useful in the first and second stage. The opposite is true for the Recent Queries and Saved Results, which were deemed more useful in the latter stages.

Questionnaire and interview data The data from the questionnaires and interviews were collected to provide insight into the utility of features at different moments of the task, and to contextualize our measurements. Here, we focus mainly on the *control* and *personalisable* features, which may support a user, but are not commonly included in regular search engines.

The general tendency for the *control* features (the Category Filters, Tag Cloud and Query Suggestions), as visualized in Table 7, is that their usefulness decreases over time. In particular, the Tag Cloud feature is deemed less useful in the second and third stage, and is a 'controversial' feature with a considerable variation in user ratings. In the post-stage questionnaires, some participants emphasize the usefulness: *"the tag cloud really aids exploring the topic"* (P.6), and *"the tag cloud came up with words that I hadn't thought of using that were very useful"*. However, especially after the second and third stage, a number of participants (P.05, P.16, P.18, P.21, P.27) indicated that the tag cloud is not so useful, saying that it *"contributed little during this task"* (P.05), that *"the tag system doesn't help to narrow the search much"* (P.18) and that it *"in the end seemed to be too general"* (P.07). P.12 summarizes this in the interview after the experiment: *"The Tag Cloud, I think, was good at the beginning, because when you are not exactly sure what you are looking for, it can give inspiration"* [14]. This can explain the fluctuations in use and perceived usefulness: the Tag Cloud is mainly useful in the beginning of the task, when users are exploring the topic, since provides basic vocabulary to the user (using frequent words in the retrieved snippets), and it may provide inspiration. In another interview, P.15 emphasizes the support of the Tag Cloud feature in generating ideas: *"it was nice to look at what other kinds of ideas [exist] that maybe you didn't think of. Then one word might spark your interest"*. However, once a user had built up a certain level of background knowledge about the topic, the value of the Tag Cloud seemed to diminish, because the user may already be familiar with the words that it displays. A similar situation exists for the Category Filters, as P.11 suggests: *"Category Filters, [those were] good at the start (...) but later I wanted something more specific"*. Hence, the refining of search results using general categories may be useful in the initial stages, but later users have more specific ideas of what they want to search for, and wish for more specific categories. For example, P.16 indicated in the questionnaire after the second stage that *"Category Filters could be more specific in its categories"*, and P.26 ideally wanted to choose a custom set of categories.

The Query Suggestions also have a similar variation in perceived value. While deemed more useful than both the Category Filters and Tag Cloud in the initial stage, the usefulness ratings for the Query Suggestions decrease in the subsequent stages. As in the case of the Category Filters, users ask after the second and third stage for improved precision, and quality (P.2, P.19), and indicate that the suggestions were *"not relevant"* for the current task (P.6, P.8). Again, this can be further contextualized using the interview data: P.11 suggested that the Query Suggestions feature *"was good at the start, but as soon as I got more specific into my topic, that went down"*. P.23 provided a suggestion for design improvement of such a feature, and indicates that over time, the Query Suggestions should take into account previous searches and *"tailor to the kinds results"* he was visiting. Still, some users mentioned, similar to the Tag Cloud, that the Query Suggestions may provide inspiration, but also serendipity: *"I clicked the query suggestions a few times. They gave me sort of serendipitous results, which are useful."* (P.24)

As opposed to the previously discussed *control* features, the *personalisable* features, the Recent Queries and Saved Results features, were increasingly highly rated. Except for some small usability issues (e.g. indicated by P.03 and P.17), users were enthusiastic about the Saved Results feature, and 13 participants wrote down positive comments in the questionnaires (P.04,P.07,P.13-16,P.18,P.21,P.23-27). P.15 remarked: *"I really found the save results feature useful, very easy to use, I wish my search engine had this!"*. The ability to categorize results was also seen as useful: *"The way that I can categorize all the pages I get is useful"* (P.27), and *"I just felt I was organizing my research a little bit."* (P.18). One participant (P.07) also indicated that the Saved Results feature helped to lay out the plans for his search. It also encouraged participants that normally do not use bookmarks to save results. Regarding the usefulness over time, various participants (for example P.12) indicate that the Saved

Results *"are most useful in the end"*. One of the participants also provided feedback in the interview that can explain the previous findings that the highest number of links were saved in the first stage (P.20): *"at the start [I was] saving a lot of a general things about different topics. Later on I went back to the saved ones for the topic I chose and then sort of went on from that and see what else I should search."* Hence, users may search and save many items in the beginning, but, if they formulated a focus, will save more specific sources later. Similarly, P.26 said *"I guess in the end I was looking for a more specific search, while my search in the beginning was just simple – [I] just searched virtual reality [and] didn't do anything on top of that"*.

Some participants indicated that the Recent Queries feature, like the Saved Results feature, was more useful later in the experiment: *"Recent queries were more useful in the end because I had more searches from before"* (P.26). The fixation data analyzed in the previous section has shown some evidence that users look more at this feature than that they actually click on it, or hover over it with the mouse. P.23 provides insight into this finding: *"the previous searches became more useful 'as I made' them, because they were there and I could see what I searched before. I was sucking myself in and could work by looking at those."* Thus, the continuous display of recent queries may aid users in their process by providing feedback about the previous paths followed; this may be the case especially in the context of complex tasks.

Summarizing, this section has looked at the perceived usefulness of features. The user ratings of different SUI features largely confirmed the findings from the previous sections, in that certain *input* and *control* features were deemed highly useful in most stages (Query Box and Results), while *personalisable* features become increasingly useful (Recent Queries, Saved Results) and *control* features decreasingly useful (Category Filters, Tag Cloud and Query Suggestions). The changes in ratings after each stage are significant for the Recent Queries and Search Box/Results feature. The increasing use of the Recent Queries feature could be observed in both active and passive interactions, but the significant difference in the usefulness ratings provides more substantial evidence for when this feature provides most value. Finally, the questionnaires and interviews provided contextualization to the active and passive interactions: the variations in the use of certain features, like the Tag Cloud and Query Suggestion feature, are caused by a user's increasing domain knowledge. As the participants indicated in the questionnaires and interviews, the features useful at the start do not provide the specific information needed in later stages, hence do not take into account a user's growing understanding of a topic.

5. DISCUSSION AND CONCLUSIONS

This section discusses the results of the study, the answers to our research questions, and broader implications for search systems. The main aim of this study was to directly examine how different SUI features can support distinct macro-level stages. By looking at active, passive and perceived utility of SUI features during different stages of a complex task, we have observed that *informational* features are naturally useful in every stage, while *input*, *control* and *personalisable* features varied by stage.

At a user's initial pre-focus stage, as Vakkari and Hakala [31] have indicated, thoughts of users are "general, fragmented and vague." Searchers are unable to express "specifically what information is needed," and their "relevance criteria are vague." At this stage, the uniqueness of encountered information is high, while the redundancy of found information is low [16]. The first stage of our experiment represented pre-focus user activities, and at this stage the *input*, *informational* and *control* features are most useful. Naturally, *input* features, are needed at this stage to express users' needs in terms of queries, while users retrieve results via the *informational* features. The user's vague understanding, the trouble in expressing her need, limited domain knowledge, but also the large amount of new information can explain the prominent role of *control* features in this initial stage: users may utilize them to explore different kinds of information, and to control their result set.

As Vakkari and Kuhlthau [16, 31] suggest, the subsequent focus formulation stage is crucial in the process. During this stage, "the search for information becomes more directed", and a better understanding drives persons to seek relevant information, using differentiated criteria. This stage was represented by the second task of our experiment. Our experimental results show that the *control* features become less essential at this point, likely caused by user's improved understanding and emerging focus. This even causes provided categories, suggested tags and searches to be "not specific enough" anymore. The *personalisable* features, on the other hand, become more important during the focus formulation stage and beyond. Contrary to control features, personalisable features may continosly support users in their process, providing feedback on the paths followed in their information journey. These features "grow" with the emerging understanding of a user. For example, users in our experiment repeatedly updated their categorizations and saved results along the way, and one participant even indicated that these features helped him to lay out the plans of his research.

Finally, the third, post-focus stage features specific searches for information, and re-checks for additional information [31]. Searchers may collect information pertinent to their focused topic [16]. At this stage, users are able to "express precisely what information is needed", and encounter low uniqueness, and high redundancy of information [16]. In our experiment, participants performed long, specific queries at this stage, and frequently reopened previous URLs and queries (via the personalisable features). The importance of the *control*, and to a lesser extent, *input* features further declined in the post-focus stage. *Personalisable* features on the other hand, were used relatively often. These features allowed users to keep track of their previous searches and captured material.

Besides insights into *when* SUI features may be useful, our findings have shown that some features were frequently used in an active way, while others were used more passively, but still received a high user rating. Hence, some features, like the Query Box and Result List, directly support users in performing their task, while other features, such as the Recent Queries feature, provide more indirect support, for example by providing context or help them manage their task progress.

Most web search systems have converged over fairly static and familiar designs, where some trialled features, such as the *personalisable* SearchPad [6] feature and Google's Wonder Wheel *control* feature, have struggled to provide value for searchers. This is perhaps because, at the wrong times, SUI features can actually impede search [5]. Conversely, these more novel control and personalisable features appear consistently in systems like online retail stores, where users are more likely to perform more complex tasks. The results of our work help to provide the insights needed to consider *when* SUI features might be useful during evolving search episodes, such that we could design responsive SUIs that introduce features at the times when they provide value, even on web search. This pleads for UIs that adapt to the needs of the task and task stage at hand.

Our results provide characteristics of behaviour observed as users transfer between different stages of a complex essay-writing task, and thus could be used to detect when live users are in pre-focus, focus, and post-focus stages. In future work, this may be extended to other types of complex tasks. Furthermore, this study has focused on one user population (undergraduate students in Computer Science), therefore future work could expand towards other user groups. Future work may also consider turning the analysis around, and try to train a classifier to accurately detect which stage a user is in. Our results, however, indicate that *control* features also need to evolve with the maturity of the the users knowledge level. Similarly, our results suggest that *personalisable* features provide more support after users move on from initial querying stages. These results support, for example, the premise behind Golovchinsky's work on Querium [9], which personalised control features with metadata about a users search history, to give users filters that develop with their task over time. This naturally leads to further research into task-aware search systems [15] and into additional features which could be useful at different stages (such as co-author visualizations, or user hints and assistance), as well as research into functions which could support interruptions and reinitiating complex search tasks. Thus, future work should directly test how dynamic provision of SUI features does support searchers when exhibiting behaviour indicative of different stages, without being impeded when features are not needed. This complex tension between support and impedence, however, is challenging to study.

Concluding, our findings suggest that the active, passive and perceived utility of SUI features across stages, especially in the context of complex and learning tasks, is inherently *dynamic* with different types of features being useful in different task stages. This is in line with macro-level information seeking models, describing broad changes in information behaviour across stages, and sheds light on the type of support needed in each stage. This provides new handles to overcome the largely *static* support for information seeking in current search systems, and facilitate a move towards more dynamic and responsive SUIs, providing tailored support to different information seeking stages.

Acknowledgments

This research is supported by EPSRC Platform Grant EP/M000877/1 and the Netherlands Organization for Scientific Research (WebART, # 640.005.001).

Data access statement: consent was not gained from participants to release their study data online, and so a dataset is not openly available.

REFERENCES

[1] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Inf. Res.*, 8(3), 2003.

[2] G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 2991–2996. ACM, 2008.

[3] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *IP&M*, 31(2):191–213, 1995.

[4] E. S. Dalmaijer, S. Mathôt, and S. Van der Stigchel. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behav. Res. Meth.*, 46(4):913–921, 2013.

[5] A. Diriye, A. Blandford, and A. Tombros. When is system support effective? In *Proc. IIiX*, pages 55–64. ACM, 2010.

[6] D. Donato, F. Bonchi, T. Chi, and Y. Maarek. Do You Want to Take Notes?: Identifying Research Missions in Yahoo! Search Pad. In *Proc. WWW'10*, pages 321–330, 2010. ACM.

[7] D. Ellis. A behavioural approach to information retrieval system design. *J. Doc.*, 45:171–212, 1989.

[8] A. Foster. Nonlinear information seeking. In *Theories of information behavior*. Information Today, 2005.

[9] G. Golovchinsky, A. Diriye, and T. Dunnigan. The future is in the past: Designing for exploratory search. In *IIiX*, pages 52–61. ACM, 2012.

[10] M. A. Hearst. *Search user interfaces*. Cambridge University Press, 2009.

[11] H. C. Huurdeman and J. Kamps. From Multistage Information-seeking Models to Multistage Search Systems. In *Proc. IIiX'14*, pages 145–154, 2014. ACM.

[12] P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*. Springer, 2005.

[13] J. Jiang, D. He, and J. Allan. Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time. In *Proc. SIGIR'14*, pages 607–616, 2014. ACM.

[14] D. Kelly. Query suggestions as idea tactics for information search. In *Proc. HCIR'09*, pages 9–12, 2009.

[15] D. Kelly, J. Arguello, and R. Capra. NSF workshop on task-based information search systems. *SIGIR Forum*, 47(2):116–127, Jan. 2013.

[16] C. C. Kuhlthau. *Seeking Meaning: A Process Approach to Library and Information Services*. Libraries Unlimited, 2004.

[17] B. Kules and R. Capra. Influence of training and stage of search on gaze behavior in a library catalog faceted search interface. *JASIST*, 63: 114–138, 2012.

[18] B. Kules and B. Shneiderman. Users can change their web search tactics: Design guidelines for categorized overviews. *IP&M*, 44(2): 463–484, 2008.

[19] J. Liu and N. J. Belkin. Personalizing information retrieval for multi-session tasks. *JASIST*, 66(1):58–81, Jan. 2015.

[20] J. Liu, M. J. Cole, C. Liu, R. Bierig, J. Gwizdka, N. J. Belkin, J. Zhang, and X. Zhang. Search behaviors in different task types. In *JCDL*, pages 69–78. ACM, 2010.

[21] G. Marchionini. Exploratory search: from finding to understanding. *CACM*, 49(4):41–46, 2006.

[22] X. Niu and D. Kelly. The use of query suggestions during information search. *IPM*, 50:218–234, 2014.

[23] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI'08 Extended Abstracts*, pages 2997–3002. ACM, 2008.

[24] T. Russell-Rose and T. Tate. *Designing the search experience: The information architecture of discovery*. Newnes, 2012.

[25] T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. In *Proc. of the ASIS Annual Meeting*, volume 34, pages 313–327. Learned Information (Europe) Ltd, 1997.

[26] B. Shneiderman and C. Pleasant. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education, 2005.

[27] A. Spink. Study of interactive feedback during mediated information retrieval. *JASIS*, 48(5):382–394, 1997.

[28] E. G. Toms. Task-based information searching and retrieval. In *Interactive Information Seeking, Behaviour and Retrieval*. Facet, 2011.

[29] D. Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009.

[30] P. Vakkari. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *J. Doc.*, 57:44–60, 2001.

[31] P. Vakkari and N. Hakala. Changes in relevance criteria and problem stages in task performance. *J. Doc.*, 56:540–562, 2000.

[32] R. W. White and S. M. Drucker. Investigating Behavioral Variability in Web Search. In *Proc. WWW'07*, pages 21–30, 2007. ACM.

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

[34] R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *Proc. SIGIR*, pages 35–42. ACM, 2005.

[35] B. Wildemuth, L. Freund, and E. G. Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *J. Doc.*, 70(6):1118–1140, 2014.

[36] B. M. Wildemuth and L. Freund. Search tasks and their role in studies of search behaviors. In *Proc. HCIR'09*, pages 17–2, 2009.

[37] M. L. Wilson. Keyword search: Quite exploratory actually. In *Proc. HCIR'09*, pages 106–108, 2009.

[38] M. L. Wilson. Search user interface design. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(3):1–143, 2011.

[39] M. L. Wilson and m. c. schraefel. A longitudinal study of exploratory and keyword search. In *In Proc. JCDL'08*, pages 52–56. ACM, 2008.

[40] T. D. Wilson. Models in information behaviour research. *J. Doc.*, 55: 249–270, 1999.