"Mirror, mirror on my search...": Data-Driven Reflection and Experimentation with Search Behaviour

DOI:
10.1007/978-3-030-29736-7_7

Document Version
Accepted author manuscript

"Mirror, mirror on my search...": Data-Driven Reflection and Experimentation with Search Behaviour

Angela Fessl[1], Aitor Apaolaza[2], Ann Gledson[2], Viktoria Pammer-Schindler[1,3], and Markel Vigo[2]

[1] Know-Center GmbH, Inffeldgasse 13, 8010 Graz, Austria
{afessl,vpammer}@know-center.at
[2] School of Computer Science, University of Manchester, Manchester, UK
{aitor.apaolaza,ann.gledson,markel.vigo}@manchester.ac.uk
[3] Graz University of Technology, Institute for Interactive Systems and Data Science

Abstract. Searching on the web is a key activity for working and learning purposes. In this work, we aimed to motivate users to reflect on their search behaviour, and to experiment with different search functionalities. We implemented a widget that logs user interactions within a search platform, mirrors back search behaviours to users, and prompts users to reflect about it. We carried out two studies to evaluate the impact of such widget on search behaviour: in Study 1 (N=76), participants received screenshots of the widget including reflection prompts while in Study 2 (N=15), a maximum of 10 search tasks were conducted by participants over a period of two weeks on a search platform that contained the widget. Study 1 shows that reflection prompts induce meaningful insights about search behaviour. Study 2 suggests that, when using a novel search platform for the first time, those participants who had the widget prioritised search behaviours over time. The incorporation of the widget into the search platform after users had become familiar with it, however, was not observed to impact search behaviour. While the potential to support un-learning of routines could not be shown, the two studies suggest the widget's usability, perceived usefulness, potential to induce reflection and potential to impact search behaviour.

Keywords: Search Behaviour · Reflective Learning · Activity Log Data Analysis

1 Introduction

Searching the Web has become a routine behaviour for workers and learners. However, users still experience problems in finding the information they are looking for [4]. Explanations put forward for this are that people typically use simple search strategies like using only a couple of query terms, or do not spend much time on the search or only check the first result page [3]. In addition, people are creatures of habit to the extent that their usual search behaviour is

independent of the information they are looking for, or how successful they are in finding it [4]. Users tend not to use other or new functionalities, even where these might be more efficient [6].

From the perspective of technology enhanced learning, we focus in this work on reflective learning as a learning mechanism that serves to learn from experience. The experience is in our case the past search behaviour that should be improved by users (who are seen at the same time as learners). Therefore, in this paper we present research that aimed to motivate users to reflect on their search behaviour, and to experiment with different types of search functionality. To this purpose, we developed a widget for data-driven reflective learning. The widget uses low-level activity log data to mirror back past search behaviour in terms of the used search functionalities to users. In combination with reflection prompts, this is expected to trigger reflection [18]. In this work we ask the following research questions with respect to the widget:

- RQ1. Users' reaction to the widget: How do participants use the widget in the search environment and engage with it? Is the widget perceived as useful?
- RQ2. Reflection: Do users generate meaningful insights about their own search behaviour in response to reflection prompts?
- RQ3. Search behaviour: Does the widget induce users to experiment with further search functionalities?

2 Related Work

The goal of a search on the web is to satisfy users' information needs and search behaviour indicates how these needs might be fulfilled. Search behaviour is influenced by a number of factors including the users' search expertise, the information needs, the search engine used and the search task itself. Although searching the web is a routinised behaviour [3], people often struggle to find what they are looking for [4]. This costs people significant time as they spend on average more than 10 minutes before they give up their search task [8]. And, when their information needs are not satisfied, people are not sure about how to change their search behaviour, or whether and how to use othersearch features [4].

A plethora of works explore how search behaviour is exhibited on the Web. However, it is not clear yet, if classifying users into novices or experts [23, 20], or using the task completion speed [3] to model the search success are meaningful approaches to understand what is good, to-be-imitated search behaviour. Therefore, we are looking at reflective learning as means for every searcher to individually develop own search competence. Reflecting on one's search behaviour could be a mechanism by which users can become better researchers in that reflection enables individuals to critically question their own behaviour, with the goal to learn from it to improve relevant aspects [5]. When it comes to online search, Edwards and Bruce [7] showed that students who are search novices do not reflect when looking for information. In contrast, experienced students not only reflect but are also aware of their own changes in their search strategy. Activity log data can be an important basis for reflective learning: Bateman et al [4]

developed a search dashboard to mirror back search history including the clicks per query, the time to click a result, or the search terms used, also in comparison to others. They showed that reflecting on search behaviour can lead to change with respect to behaviour and attitudes about search. In line with this, Malacria et al. [16] showed that a reflective widget was helpful to incite reflection on learning to use shortcuts in software. Pammer et al. [17] have shown that reflection on time log data incited users to generate insights about time management, and experiment with different time management strategies. Prior research has also shown that automatic reflection prompts can support reflective learning based on data: Fessl et al. [9] implemented and evaluated reflection prompts that were embedded both directly within action, and with a larger temporal separation from action in informal and workplace learning contexts. The authors' reflection prompts reminded users to reflect, and pointed out salient data to users. Kocielnik et al. discussed reflection prompts in private life settings (i.e. physical health [13]) as well as in a workplace setting (i.e. time management [12]). These authors' prompts were based on users' self-set goals for behaviour change.

Literature therefore suggests that online search can get difficult. One of the salient features that distinguishes experienced searchers from novice searchers is their capacity to reflect on their search behaviour and strategies. In parallel, we can build on past known successful designs for data-driven reflective learning and reflection guidance technologies based on data collected within informal learning settings. Both the design of our widget for reflective search (description below) and research question as stated above, are based on this understanding.

3 A Widget for Reflective Search

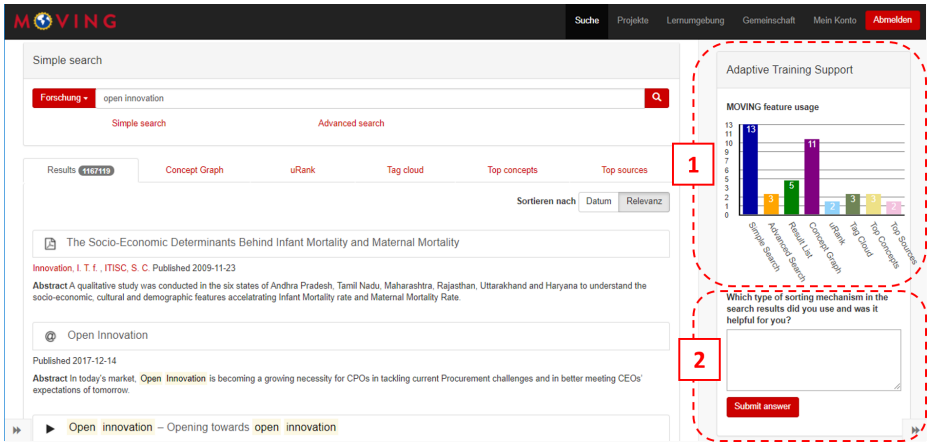

Fig. 1. Widget for behaviour change embedded in the search platform.

The widget for reflective search that we have developed is embedded into a newly developed search platform [24] that offers multiple search interfaces, such

as the typical text search, a graph visualisations of search lists, an interactively ranked visualisation of search results based on keywords according to di Sciascio et al. [22], a tag cloud visualisation based on keywords' frequency, and a bar chart visualisation presenting properties of the retrieved documents. While using this custom search platform constrained the available content for searching, it enabled us to track user interaction with the widget in a fine-granular manner.

The widget consists of two parts: First, it visualises search behaviour in terms of which functionalities are used, inspired by Malacria et al. [16]. Second, the widget prompts users to reflect on whether and in what sense the used search functionalities were used, and on overall search behaviour. These prompts constitute generic reflection prompts [10] in the sense of not directing users towards particular solutions. While directed prompts in principle have advantages especially for novices (ibid), as it is unknown what exactly constitutes good search behaviour, it is known that reflecting and adapting search behaviour to the search task is a characteristic of experienced searchers, generic reflection prompts were assumed to be the best approach in this work. The search behaviour visualisation (see Fig. 1, component 1) shows how often a user used a search feature.
The reflective prompts (see Fig. 1, component 2) are phrased as questions. Many of them refer directly to the user's way of using search functionalities, such that used features, and the number of times a feature has been used are variables that are inserted into template sentences. Examples are *"You have not tried the 'Tag Cloud'. Why haven't you tried it out before?"* or *"What did you learn by using the 'Concept Graph' feature?"*. Some reflective prompts overarch wider issues, like *"Which of the features listed above do you find the most useful, and why?"*[4]
On the server-side, we have implemented an activity tracking tool that collects all events a user is performing on the platform. The captured events include all mouse and keyboard interactions, browser window events, changes to the state of the elements on the page, and other system information. The captured data is analysed to calculate how often a user used the features on the platform.

4 Methodology

4.1 Study 1 - Experimental Study

This study aimed to answer RQ1 on users' reaction to the widget, and RQ2 on whether the prompts incited reflection.

Setting: The experimental study was designed as a comparative study. It lasted for about 2 hours. Two different user groups participated in the study: the "Researcher" group, consisting of master students of "Computer Science" or "Software Engineering and Managmement" of Graz University of Technology (TUG), who were recruited during a lecture. The "Auditor" group consisting of auditors from a big auditing company in Germany and students of Software Engineering and Managmement (TUG) with a strong background in economy.

[4] All reflective prompts are listed in an online appendix published on Zenodo: https://tinyurl.com/y5wlgeyx.

Additionally, each group was divided in two subgroups, resulting in four groups: group 1S and group 1V for the researchers and group 2S and group 2V for the auditors. While the groups with "S" had to deal with search input interfaces, the groups with "V" were asked about search result visualisations.
Group 1S (researchers) and group 2S (auditors): the participants of these groups were asked to perform a search task on the search platform and to use either a typical one-line input field (simple search) or another search input page offering several input fields including domain, title, abstract, full text and person (advanced search). The screenshots of the widget were adapted to this task. For group 1S, the reflection widget screenshot showed simple search to be used more frequently than advanced search. The reflective question posed was: *"You are mostly using the 'Simple Search'. What could help to motivate you to use some other search features like the 'Advanced Search'?"*. For group 2S the screenshot showed the advanced search as the most often feature used. The reflective question was *"You are mostly using the 'Advanced Search'. What could help to motivate you to use some other search features like the 'Simple Search'?"*.
Group 1V (researchers) and group 2V (auditors): the participants of these two groups were asked to use the ranked result visualisation based on keywords in the first search task, and to use the graph visualisation of search results in the second search task. We then prepared for group 1V a screenshot of the reflection widget showing the interactively ranked visualisation as most frequently used search functionality, and the following reflective question: *"Do you think that using the 'interactively ranked result visualisation' can improve your search performance/ search skills...? And if yes how?"*. Group 2V was presented with a reflection widget screenshot that showed the graph visualisation as the most frequently used search functionality, and presented the following reflective question: *"Do you think that using the 'Graph Visualisation' can improve your search performance/ search skills...? And if yes how?"*.
Metrics and Tools: We used Google Forms to administrate the workflow of the experiment. We created a sequence/condition for each group, which provided step-by-step instructions of the tasks to perform as well as all questionnaires that needed to be filled in. While each condition followed the same structure, it differed on the search tasks, the corresponding screenshots of the widget and the reflective questions. First, all participants gave their consent to participate and were asked to provide demographic information. Then they were introduced to the search platform, and were asked to familiarise themselves with the platform and the widget. Afterwards, each of the four groups was asked to look at a screenshot of the widget and to answer a reflective question about the screenshot as well as further open questions. The questionnaire also measured constructs from the Technology Acceptance Model [19] such as perceived ease of use, perceived usefulness, attitude towards the widget, widget specific questions, learning outcome, behaviour intention, technological self-efficacy, subjective norm and system accessibility. All the questions were defined using a 7-point Likert scale where 1 indicated 'strongly disagree' and 7 'strongly agree'. Additionally, qualitative data was collected through open-ended questions.

Participants: 76 participants (61 male, 15 female) took part in the study. 42 were assigned to the research group (35 male, 7 female) and 34 participants were assigned to the auditor group (27 male, 8 female). 80% of the participants were aged between 18-27, 18.5% between 28-37 and 1.5% was aged between 48-57.

4.2 Study 2 - Field Study

This study aimed to answer RQ1 on users' reaction to the widget, and RQ3 whether the widget influenced the search behaviour.

Setting: The field study was split into two periods of one week. For each period, all participants were asked to carry out one search task per working day. The tasks followed a strict order, so if a participant missed one, they would have to carry it out the following day before they were given the next one. Hence, up to five tasks could be realised per one-week period. We kept the tasks from both periods analogous by using the same instructions, but changing the search topic. The participants were split into two groups: in group A the widget was available on the search platform during both weeks. In group B the widget was introduced at the beginning of the second week. The order of the assigned topics "Big data" and "Global warming" was randomised to counterbalance the effect of a particular topic on participants' behaviour. Henceforth we use the notation A1, A2, B1 and B2 to indicate group membership and period of the study.

Metrics and Tools: We used three questionnaires: A pre-questionnaire was distributed to the participants at the beginning of the study. It included a consent form, a demographic questionnaire and questions about the participants' computer and Web experience as informed by [3]. The in-between weeks questionnaire, was sent out after the first study period (i.e. after a week). It captured the first impressions about the platform and the widget. The post-questionnaire was sent on completion of the study. It measured constructs of the Technology Acceptance Model [19] such as ease of use, perceived usefulness, attitude to the widget, widget specific questions, learning outcomes, search behaviour, technological self-efficacy. All questions were defined on a 5-point Likert scale, where 1 indicated 'strongly disagree' and 5 'strongly agree'.

We computed *engagement metrics* and *interactive patterns of use* from usage data logged on the search platform [14, 15]. The engagement metrics were:

- Active time: the time elapsed carrying out the task where periods that were longer than 50 seconds were not accounted for.
- Number of searches: the number of searches carried out.
- Number of selected results: the number of times a user clicks on a search result can be an indicator of search engine efficiency, but also of engagement.
- Number of episodes per task: a timeout of 40 minutes is used to split interaction into different episodes.
- Amount of scroll: measuring the scroll interaction from users is a common metric to measure engagement with a site.

The interactive patterns of use were based on pattern mining and n-gram analysis. N-grams are typically used in computational linguistics [25] and in

computational biology (e.g. protein sequencing [2]). They are a useful method for capturing low-level sequences, whilst avoiding the need for full parsing. We define a user interaction event n-gram as consisting of a time ordered sequence of n consecutive events by a single user that is fully contained within a single user *episode*. We computed n-grams of size 4 as we empirically found them to be large enough to allow patterns to be extracted for a large number of frequent n-grams in this dataset, across all users who were fully engaged in the study [1]. We visually compared the emerging patterns to look for differences between groups.

Participants: Fifteen participants (10 male, 5 female) aged between 17-46 (M = 28.8) took part in the study. On average, they had 15 years of experience with computers (SD = 7.2) and 14 with the Web (SD = 4.4). 73% use search engines and the Web on a daily basis, and 66.6% of them use a computer dailySelf-reported search skills suggest that 20% of the participants considered themselves to be very skilled, 60% skilled and only 20% reported to be neutral.

5 Results

5.1 RQ1: Users' reaction to the widget

Table 1 shows average values of users' active time, the number of selected search results, the number of episodes and searches conducted per task and the amount of scrolling. We compared whether the availability of the widget in Study 2 led to significant differences across groups. A Wilcoxon test on the metrics extracted for engagement suggest that there are no statistically significant differences: When comparing A1 and B1 (between subjects) the range of the Wilcoxon coefficient was W=2203–2384 (all p>0.33). When comparing B1 and B2 (within subjects), the range of the Wilcoxon coefficient was W=2859–3095 (all p>0.09).

Questionnaires: In Study 1 and Study 2, we conducted t-tests per study to compare the reaction on the ease of use and the usefulness of the widget of the different user groups. Yet, we found no statistical significant differences, neither in Study 1 between those who performed tasks using the search input interfaces and those who performed tasks using the graphical search result visualisations, nor in Study 2 between those who had the widget during the whole study and those who had the widget only after the first week.
We therefore, for this RQ, treat all participants for each study as one group.

Table 1. Average engagement metric per group

Metric	A1	A2	B1	B2
Active time in minutes	7.24	5.58	5.52	5.52
Number of searches	10.08	7.69	9.62	7.57
Number of selected results	2.35	2	1.61	1.62
Number of episodes per task	1.14	1.14	1.26	1.08
Amount of scroll	253.32	175.63	264.13	147.89

Firstly, participants tended to perceive the widget to be easy to use (Study 1 (7-point Likert scale): $M = 4,82$, $SD = 1.08$; Study 2 (5-point Likert scale): $M = 3.68$, $SD = 0.58$ and useful (Study 1: $M = 4.26$, $SD = 1.42$; Study 2: $M = 3.13$, $SD = 0.92$). In Study 2, this is supported by comments we received when asking an open question about the ease of use and the usefulness of the widget: *"The widget is quite useful. I like the design and that it helps me to use the search engine more efficiently"* and some other neutral *"For me using the widget didn't make much of a difference. The system's bunch of functions is easy enough to overlook, so you rather quickly find what helps you search best and what not with or without the widget"*.
In both studies, we also asked the participants if they thought the widget would raise their engagement with the different platform functionalities. Participants' answers were varied, with no clear tendency overall (Study 1: $M = 4.04$, $SD = 1.81$; Study 2: $M = 3.27$, $SD = 1.03$). Furthermore, we asked all participants if they thought the widget would be useful to explore different search functionalities (Study 1: $M = 4.26$ $SD = 2.06$; Study 2: $M = 3.57$, $SD = 1.10$), which participants were again hesitant about, with a large variance in answers. One participant of Study 2 highlighted that whether the widget would, or wouldn't, encourage exploration of different search functionalities was highly dependent on whether their information needs were met in any given search task: *"It depends on how satisfied I am with the results I got with the usual methods. For some searches it could be useful to use other tools and the widget suggests them. As for which one: I would try them all to see which one could be useful."*.

5.2 RQ2: Reflection

In order to investigate if learning occurred when answering the reflective questions in Study 1, we textually analysed all answers given by study participants in response to reflection questions. We coded answers according to a coding schema for reflective content [21] with whichreflective expressions can be characterised according to three levels of depth of reflection, namely low, medium and high. For example, answers that describe an experience without interpretation count as low depth of reflection; answers that contain an interpretation or justification count as medium depth; and answers that describe gained insights count as high depth. One rater coded all 58 answers (given by participants in Study 1 to the four reflective questions). In case of doubt, the coding was discussed with a second coder. Agreement could be reached for all quotes. 48 answers were identified as reflective. 10 answers didn't contain any reflective content like for example *"No"* or *"I don't think so"*. Altogether 81% of the answers were assigned to the lowest level and 66% to the medium level of reflection. Some of the answers given belong to more than one category. Table 2 presents the number of answers per category. Categories, to which no answers could be assigned to, were omitted from Table 2 (hence, e.g., the missing category number 2 in the table). Table 3 presents coded examples of answers by participants from group 1V to the question *"Do you think that using 'interactively ranked result visualisation' can improve your search performance/search skills...? And if yes how?"*.

Table 2. Number of answers per coding category.

Categories of coding schema	**Number of codes**
Low-level reflection	
1. Description of an experience	41
Medium-level reflection	
3. Interpreting or explaining behaviour in the experience	24
4. Linking an experience explicitly to other experiences	5
5. Linking an experience to knowledge	7
6b. Responding to the explanation of an experience by challenging or supporting assumptions	2
Non-reflective answers	10

Table 3. Examples of analysed answers given

Categories	**Example**
1: experience	I think it can, using key words makes a huge difference.
1,3: interpretation:	If I know for what keywords I'm looking for, I'm quite sure to find relevant papers very quickly.
1,5: linking experiences to experience	I think it can help me with searching because it simplifies finding the right results for some more complicated queries.
1, 3, 6b: supporting assumptions	Yes, because i have an overview of documents that are related to my keywords. Searching for a specific document is far easier than searching for a keyword to find an appropriate document

Besides asking the participants a reflective question about the widget, we also asked them if such a question would motivate them to reflect about the own search behaviour. The answers given were ambivalent. Many confirmed to think about the own search behaviour, but others did not. For example, participants were stating that *"Yes, I would try different methods for optimised search results."*, *"A bit yes, I never thought how I can improve my searching skills and it is a valuable asset."*, *"Yes, It helps but in real life I might not have time to try out other visualisations and just use the one I am most comfortable with."*, and *"A little bit, maybe. But I still prefer text based searches due to my habit."*. On the other hand, some said just *"No"* or *"Not really"*, *"No, because I'm happy with my current way of searching."* or *"Not really, because normally when I search I get the results that I'm looking for in a fast way, changing my behaviour therefore would cost time for doing something that is already efficient for me."*

5.3 RQ3: Search behaviour

Based on the activity log data captured in Study 2, an n-gram analysis was performed to compare the effect of the widget on the interactive behaviour exhibited on the search platform between those users who:

- Used the platform for the first time with (A1) and without the widget (B1);
- Used the widget for the first time but had already been exposed to the platform (B2) and used the platform and the widget for the first time (A1);
- Used the platform without the widget (B1) and had the widget introduced later on (B2);
- Used the platform with the widget from the beginning (A1) and continued using it in the second period (A2);
- On the second week, were already familiar with the widget (A2) and had it just introduced (B2).

We conducted a correlation analysis between the frequencies of the top-100 n-grams on the above users groups. Next we provide a guide to interpret Table 4, where coefficients around 0.4 and above are considered to be moderate correlations, and those above 0.6 are strong correlations for the following statistical tests: a high Kendall τ and Spearman ρ correlation indicates that the rankings of two vectors of n-grams are similar. The former is considered more strict and will typically produce a lower correlation coefficient. When in doubt, the p-value of Kendall's test is known to be more reliable. A high Pearson r suggests that the frequencies of the n-grams are associated (despite their ranking in their respective vectors). The results on Table 4 and an observational analysis of the top-10 n-grams suggests that:

- **A1 vs B1**: a high Pearson correlation and low Spearman suggest that behaviours are exhibited a proportionately similar number of times but their rankings are not the same (i.e. the frequency based order changes). Using the search functionality, exploring the results after searching and interacting with visualisations are within the top-5 behaviours exhibited by those who had the widget, while they are ranked in positions 6–8 for those who did not.
- **A1 vs B2**: low correlations tending toward moderate correlations indicate slightly different behaviours on first exposure to the widget, which suggests that having the widget from the outset may make a difference in that we do not observe search activity patterns on the top-10 n-grams of B2 users.
- **B1 vs B2**: high correlations that are consistent across rankings and frequencies suggest that there was no behaviour change when the widget was introduced. On the first week the participants without the widget (B1) carried out simple search activities, while in the second week (B2), we observe more interaction with visualisations and exploratory search behaviours through the use of the scroll.
- **A2 vs B2**: low correlations suggest different behaviours between those who have been exposed equally to the platform but get the widget later. While both groups show exploratory search activity patterns and interaction with visualisations, the group using the widget for a second week (A2) shows interactions with advance search features (i.e. use of filters).
- **A1 vs A2**: low correlations across the tests we run indicate that behaviours changed over time probably due to the learning effect, and exposure to the platform and the widget. As we say above, we observe the emergence of sophisticated search functionalities on the second week.

The conclusion derived from these findings suggests that the widget does not make users exhibit *new* behaviours, but *makes users prioritise other behaviours that are already in their repertoire* (A1 vs B1). The effect of the widget is particularly noticeable for those who interact with the search platform for the first time as once users get familiar with the platform (B1 vs B2), *the posterior incorporation of the widget does not lead to using further search functionalities.* This indicates that support for training is more effective when the learning gap is perceived to be large, i.e. the first time one is exposed to such system (A1 vs B2). We do not know how long it would take to make the two groups similar as one week does not seem to be enough time (A2 vs B2).

Table 4. Widget user group vs period: correlations of top-100 n-grams, where N=4.

	Kendall τ	p value	Pearson r	p value	Spearman ρ	p value
A1 vs B1	0.16	0.02	0.62	0.00	0.27	0.007
A1 vs A2	0.08	0.28	0.23	0.02	0.11	0.27
A1 vs B2	0.21	0.005	0.38	0.00	0.27	0.006
B1 vs A2[5]	0.11	0.15	0.17	0.08	0.14	0.17
B1 vs B2	0.39	0.00	0.58	0.00	0.50	0.00
A2 vs B2	0.15	0.06	0.10	0.32	0.12	0.07

Questionnaires: In Study 2, we asked participants about their search behaviour and a possible change of it. Most of the participants (especially group A) supported the idea that the widget encouraged reflection about their search behaviour (Group A: $M = 3.71$, $SD = 1.11$; Group B: $M = 3.38$, $SD = 1.06$). Whether the widget enabled search behaviour change was less clear as participants leaned toward being neutral (Group A: $M = 3.29$, $SD = 0.76$; Group B: $M = 3.25$, $SD = 0.89$), and event the intention to change it (Group A: $M = 3.14$, $SD = 0.69$; Group B: $M = 3$, $SD = 1.07$). This was supported by a participant: *"I didn't learn from using the widget – it just made me more aware of how I'm usually doing my search without wanting to change that behaviour"*.

6 Discussion

RQ1: Users' reaction to the widget. The widget was perceived to be easy to use and useful by participants in both studies, and via both questionnaires and engagement metrics. We understand this to be a necessary prerequisite for supporting learning and behaviour change (cp. Kirkpatrick's [11] hierarchical model of evaluating learning interventions).
RQ2: Reflection. From the analysis of the answers given to the reflective questions we can show that reflection took place mostly on the lowest level (81%) and the medium level (66%) of reflection (dual coding, hence the sum is larger than

[5] B1 vs. A2 is added for completeness reasons but the comparison is not meaningful.

100%). This could be explained by the following two facts. First, it is easier to describe (low-level reflection) or interpret an experience (medium level reflection) than to derive insights from reflection and put them in writing (high level reflection) [9]. Second, the experimental study (about 2 hours) may have been too far outside participant's real search practice for them to be able to derive deeper insights search behaviour. Additionally, we received further thoughts from participants when asking them if the reflective question motivated them to reflect on their search behaviour. The thoughts of some study participants include on the one hand that they would like to improve their search skills to receive optimised search results. On the other hand, others mentioned after becoming aware of how they search, that they are happy with the way they currently search. They still prefer using the one-input line they are used to and do not want to un-learn or change their search behaviour due to time reasons. As a consequence this shows that people are creatures of habit, thus, changing internally operationalised behaviour is difficult as it requires a significant investment of time, effort and motivation on the user's side [4, 16]. This is explained by the *active user paradox* in that users tend not to use other or new functionalities, even where these might be more efficient [6].

RQ3: Search behaviour. The n-gram analysis suggests that the widget influenced the activity patterns of those participants who were introduced to the new search platform and widget together (group A). This group of users were more active searchers than those who did not have the widget (group B). Interestingly, on the second week of use, they (group A) exhibited activity patterns that signaled search behaviours that were beyond the traditional search box. However, we observed that users did not exhibit those search behaviours when the widget was incorporated on the second week (group B). This may indicate that having the widget from the beginning might have facilitated the initial prioritisation of search behaviours upon which, more sophisticated behaviours were exhibited in the second week.

7 Conclusions

In this work, we focused on reflective learning as a learning mechanism that serves to learn from experience to drive future search behaviour. We have presented two studies that investigate if a widget that mirrors back users' current search behaviour in terms of search features used is able to stimulate reflective learning and experimentation with different search behaviours. In Study 1, we could show that reflective learning took place, and that the improvement of own search skills was thought of. However, a search behaviour change is still refused due to being a creature of habit. In Study 2, we could show that there was an effect on the search behaviour in the second week on those participants (group A) that had been exposed both to the novel search platform and the widget from the study outset. We didn't see an effect on those users (group B), however, that used the novel search platform without the widget in week 1 and with the widget in week 2 of the study. We suspect that there are two reasons: First, unlearning

behaviour is harder than exploring a novel technology, especially in the presence of technology that aims to incite reflection and exploration. Second, learning the widget, reflecting on search behaviour, and experimenting with novel search behaviours may take longer than a week; which was all the time that study participants had with the widget in group B.
While the two studies therefore show the widgets usability, perceived usefulness, potential to induce reflection, and potential to impact search behaviour; the potential to support unlearning of routines could not be shown. The immediate outlook to future work is a longer-term experimental field study. Beyond this, this work shows that there are knowledge gaps in existing research with respect to evidence for best search practices; and with respect to designing for reflective search practice.

Acknowledgements The project "MOVING - TrainiG towards a society of data-saVvy inforMation prOfessionals to enable open leadership iNnovation" is funded under the Horizon 2020 of the European Commission (project number 693092). The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. Apaolaza, A., Gledson, A., Fessl, A., Bienia, I., Pournaras, A., Blume, T., Saleh, A., Simic, I., Maas, A., Gnther, F., Mutschke, P., Lorenz, R., Collyda, C., Vigo, M.: MOVING Project, Deliverable 1.4: Final implementation of user studies and evaluation (2019)
2. Asgari, E., Mofrad, M.R.: Continuous distributed representation of biological sequences for deep proteomics and genomics. PloS one **10**(11), e0141287 (2015)
3. Aula, A., Nordhausen, K.: Modeling successful performance in web searching. J. of the american society for inform. science and technology **57**(12), 1678–1693 (2006)
4. Bateman, S., Teevan, J., White, R.W.: The search dashboard: how reflection and comparison impact search behavior. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1785–1794. ACM (2012)
5. Boud, D., Keogh, R., Walker, D.: Reflection: Turning Experience into Learning, chap. Promoting Reflection in Learning: a Model., pp. 18–40. Routledge Falmer, New York (1985)
6. Carroll, J.M., Rosson, M.B.: Interfacing thought: Cognitive aspects of human-computer interaction. chap. Paradox of the Active User, pp. 80–111. MIT Press, Cambridge, MA, USA (1987)
7. Edwards, S.L., Bruce, C.S.: Panning for gold: Understanding students information searching experiences. Transforming IT education: Promoting a culture of excellence pp. 351–369 (2006)
8. Evans, B.M., Chi, E.H.: An elaborated model of social search. Information Processing & Management **46**(6), 656–678 (2010)

9. Fessl, A., Wesiak, G., Rivera-Pelayo, V., Feyertag, S., Pammer, V.: In-app reflection guidance: Lessons learned across four field trials at the workplace. IEEE Transactions on Learning Technologies **10**(4), 488–501 (2017)
10. Ifenthaler, D.: Determining the effectiveness of prompts for self-regulated learning in problem-solving scenarios. Ed. Technology & Society **15**(1), 38–52 (2012)
11. Kirkpatrick, D.L., Kirkpatrick, J.D.: Evaluating training programs: The four levels. 3rd ed., Berrett-Koehler Publishers, San Francisco. (2006)
12. Kocielnik, R., Avrahami, D., Marlow, J., Lu, D., Hsieh, G.: Designing for workplace reflection: A chat and voice-based conversational agent. In: Proceedings of the 2018 Designing Interactive Systems Conference. pp. 881–894. ACM (2018)
13. Kocielnik, R., Xiao, L., Avrahami, D., Hsieh, G.: Reflection companion: A conversational system for engaging users in reflection on physical activity. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(2), 70:1–70:26 (Jul 2018)
14. Lagun, D., Lalmas, M.: Understanding user attention and engagement in online news reading. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 113–122. ACM (2016)
15. Lalmas, M., O'Brien, H., Yom-Tov, E.: Measuring user engagement. Synthesis Lectures on Information Concepts, Retrieval, and Services **6**(4), 1–132 (2014)
16. Malacria, S., Scarr, J., Cockburn, A., Gutwin, C., Grossman, T.: Skillometers: reflective widgets that motivate and help users to improve performance. In: Proc. of the 26th ACM symposium on user interface software and technology. pp. 321–330. ACM (2013)
17. Pammer, V., Bratic, M., Feyertag, S., Faltin, N.: The value of self-tracking and the added value of coaching in the case of improving time management. In: Design for Teaching and Learning in a Networked World, pp. 467–472. Springer (2015)
18. Pammer, V., Krogstie, B., Prilla, M.: Let's talk about reflection at work. International Journal of Technology Enhanced Learning (IJTEL) **9**(2/3), 151–168 (2017)
19. Park, S.Y., et al.: An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. Educational technology & society **12**(3), 150–162 (2009)
20. Perrone, V.: Librarians and the nature of expertise. In: Proceedings LIANZA Conference 2004. LIANZA (2004)
21. Prilla, M., Renner, B.: Supporting collaborative reflection at work: A comparative case analysis. In: Proceedings of the 18th International Conference on Supporting Group Work. pp. 182–193. ACM (2014)
22. Sciascio, C.D., Sabol, V., Veas, E.: Supporting exploratory search with a visual user-driven approach. ACM Trans. on Interactive Intel. Systems **7**(4), 18 (2017)
23. Tucker, V.M.: The expert searchers experience of information. In: Information experience: Approaches to theory and practice, pp. 239–255. Emerald Group Publishing Limited (2014)
24. Vagliano, I., Gnther, F., Heinz, M., Apaolaza, A., Bienia, I., Breitfuss, G., Blume, T., Collyda, C., Fessl, A., Gottfried, S., Hasitschka, P., Kellermann, J., Khler, T., Maas, A., Mezaris, V., Saleh, A., Skulimowski, A.M.J., Thalmann, S., Vigo, M., Wertner, A., Wiese, M., Scherp, A.: Open innovation in the big data era with the moving platform. IEEE MultiMedia **25**(3), 8–21 (July 2018)
25. Xia, Y., Cambria, E., Hussain, A., Zhao, H.: Word polarity disambiguation using bayesian model and opinion-level features. Cognitive Computation **7**(3), 369–380 (2015)