McKay, D., & Buchanan, G. (2011). One of these things is not like the others: how users search different information resources.


Available from: http://dx.doi.org/10.1007/978-3-642-24469-8_28
One of these things is not like the others: How users search different information resources

Dana McKay1 George Buchanan2

1 Library, Institute for Social Research, Swinburne University of Technology
PO Box 218 John Street, Hawthorn, VIC 3122, Australia
dmckay@swin.edu.au

2 Centre for HCI Design, City University
Northampton Square, London EC1V 0HB, UK
george.buchanan.1@city.ac.uk

Abstract. Transaction log analyses are common practice to understand user behavior in both online databases and library catalogues. While there has been significant work done in each of these domains, there is little work comparing user queries between library catalogues and online resources. In this paper we report on an exploratory comparison between searches performed via the same interface in three different search systems: a library catalogue, an online research database, and Google Scholar.

Keywords: User behavior, search behavior, search interfaces, libraries

1 Introduction

Users’ behaviors when engaged in interactive search have been the subject of extensive study. Previous research has made clear that information seekers choose different resources for meeting different information needs; depending on the nature of their need, information seekers may look to the library for books or articles, they may search the web, or they may just ask a friend [1, 2]. Some of the selection process is based on where in the information seeking process users are: according to McKenzie, users have four stages of information seeking, only one of which is likely to find users actively engaging with information resources [3]. Similarly Kuhlthau and Marchionini defined search only as a small part of the information seeking process [4]. Ill-defined or amorphous information needs are refined by browsing [5, 6], encountering information [3] or, in a library setting, by using a reference librarian as an intermediary [7]. With increasing availability of online information, though, the opportunity to browse is less and less common, and search is becoming the dominant information-seeking paradigm.

We know that many queries are syntactically naive and limited in scope [8-12]. It is also evident from the literature that the design of information interfaces affects query formation [13, 14], and that well-designed interfaces can help even those users who have poorly defined information needs [7, 15]. Library catalogues, however,
have long stymied users in their efforts to find information [16, 17], as have research databases [18]. Consequently experienced academics and new students alike typically select Google and other internet search engines as their first choice of information resource [18, 19]. Despite choosing Google first, information seekers (at least in an academic setting) value libraries and library information resources [2, 20]. We also know from our earlier work that users of academic libraries perceive and use different types of library information (for example books and scholarly articles) differently, and are somewhat resistant to new trends to combine these resources into a single search interface [21].

While we do know that users use Google but value library resources, the literature gives little insight into the differences between the queries users enter into different kinds of information resources. Similarly, we could find no work that investigated whether any differences are due to the interfaces of such resources, or the due to the demonstrably different needs and expectations of searchers when looking for different kinds of information.

This paper reports an exploratory study of the queries entered into three different information sources via the same interface. This interface is on the homepage of an academic library website and records the initial queries users make of the library catalogue, EBSCOHost (a large library database), Gale (a smaller, more local database) and Google Scholar. It is our hope that by examining the content of these searches, we can learn about the differences in users’ perceptions of these resources.

We commence by reviewing the background literature, followed by a description of the methodology employed in our study. The study’s findings are then presented and discussed, leading to our conclusions and suggested avenues for future work.

2 Background

There is a long history of using search log analyses to understand user behavior in digital libraries [9, 10], physical libraries [12, 17], online research databases [11, 22] and on the web in general [8, 23]. These analyses have shown that users persistently accept interface defaults and use queries which average somewhere between two and three words long. This is true even where information seekers can reasonably be expected to have complex information needs. For example, a detailed study of a large academic article database shows similar query patterns by all but a very few users [22]. Not only do users accept defaults and use few words, but even in computer science databases (where searchers can be expected to be reasonably proficient with Boolean logic) Boolean operators and search limiters are not used (and where they are used, they are often used incorrectly) [10, 17].

The design of specific types of systems seems to be one factor upon user behavior. Studies of library users show they find library search services hard to use [2, 24], and that finding journal articles is particularly difficult [25]. One investigation of university students’ academic library searching revealed that only library science students were prepared to search a range of resources or use advanced searching techniques; all other users preferred basic search [18]. Similarly, given the choice,
most users of an academic library prefer federated searching to searching individual library databases, even though they know that they get higher quality results by the latter strategy [25]. Finally, library users strongly associate libraries with books. Even though, in academic libraries at least, usage of electronic information resources far outstrips book loans [26], libraries are strongly associated with “books d’uh” [2] in users’ minds. This causes problems when libraries ask users about their information seeking experiences, as most users report their experiences with books (c.f. [27]), depriving libraries of patrons’ experience with what might be termed ‘the online library’.

Earlier work by the authors in this area [21] asked users to comment specifically on their experiences with a wide range of information types, focusing particularly on books and scholarly journal articles. The results of this preparatory work suggest that academic library users, at least, view different types of library resource differently, and that books and scholarly articles in particular are used in different ways. Similarly, different academic disciplines use resource types differently, some focusing more heavily on journals, others on books [28, 29].

In summary, we can confidently generalize about library users’ searching behavior: they do short, simple searches; they use a wide range of electronic resources; and they find library resources hard to use. The literature shows that books and scholarly articles are used in different contexts and to meet different information needs. What is unclear is whether users’ search strategies change with resource type, and if so, whether these changes reflect something about the information sought, or are caused by the search interfaces. We set out to examine these significant questions through the study we now report.

3 Methodology

As noted above, research has faced the challenge that different types of resources are typically accessed through different interfaces. Integrated interfaces, if common in practice, have not received scrupulous scientific study. We exploited an integrated search interface that gives access to four different catalogs and its query log data.

We first describe the search interface in Section 3.1, and in Section 3.2 we report the analysis methods applied to the searches performed using the interface.

3.1 Search the library

The Swinburne Library homepage has a single search box with radio buttons allowing users to pre-select among a range of information sources (see Figure 1).

![Fig. 1. The Library search interface](image-url)
The default search is of the library catalogue (‘Books and more’) that includes physical books, ebooks, and a small number of articles made available through a course readings system. Three other resources can be searched directly: EBSCO, a large aggregator of journal content, Gale, a smaller journal content aggregator and Google Scholar. Upon executing a search, this search box transfers the searcher to the native interface of the resource they search, and as such the searches captured by this box (and analyzed in this study) are all initial searches rather than search refinements. We cannot tell from the logs alone how many results each search returned, though number of results could be approximated by repeating the search.

The predominant association of this interface is with the library: it is located on the library home page, and the default search target is the library catalog, neither of which are first choices for information seekers. While this association may affect the type of searcher using this interface, the information resource chosen can reasonably be assumed to be responsible for any differences in search strategy.

3.2 Analysis

The searches analyzed in this study were from two randomly selected weekdays, each during the second half of an academic term so as to represent practiced searching. A brief analysis of further days was conducted, but as they were not significantly different from the two days presented here, and as this study was intended to be exploratory, no further investigation of these days was undertaken.

The searches from the two selected days (3743 in total) were first analyzed in their entirety for query length and distribution over information sources. As it was neither realistic nor worthwhile in an exploratory study to manually examine these searches for content, following the gross analysis, a selection of 100 searches per day was taken from each of the three most used resources (the catalog, EBSCO, and Google Scholar—more on this in Section 4.1), for a total of 600 searches. Each query was then examined with respect to a number of factors: whether the target was a known item, the metadata used, what kind of target item chosen, whether the search used advanced search operators or contained typographical errors, and how successful the search was.

4 Results

In the two days examined as part of this study, users performed some 3743 searches encompassing 14648 search terms. 13 searches were empty and thus were excluded from further analysis. This section will present an analysis of query form, based on analyzing the entire dataset (Section 4.1) and a more detailed analysis of query content based on the sample data (see Section 4.2)
4.1 Query Form

While we could not examine all 3730 non-empty searches manually in this study, some automated analysis was possible, and interesting results emerged around the information sources users selected and the number of search terms per query.

4.1.1 Search Source

Of the searches analyzed, some 2277 used the library catalog, meaning a full 1453 (38%) of searchers selected another target. This is a surprising finding, given that searchers typically change search defaults in fewer than 5% of cases [9, 17]. Because neither of the sample days was near the beginning of semester, it is likely we are seeing practiced searching, with searchers preselecting an appropriate information resource for their needs.

<table>
<thead>
<tr>
<th>Table 1. Number of words in each query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>Library</td>
</tr>
<tr>
<td>EBSCO</td>
</tr>
<tr>
<td>Gale</td>
</tr>
<tr>
<td>Google Scholar</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

4.1.2 Search Terms

The number of terms in each query was counted and analyzed. The mean number of terms for all searches was 3.93, with Google Scholar having the highest mean (5.03) and the library catalogue the lowest (3.45) (see Table 1). Compared to the various studies reported earlier in this paper, these means are unusually high, and there are few searches with two or fewer words in comparison to other information interfaces.

Fig. 2. Number of search terms by information source
The difference between the median number of searches and the mean in all cases suggests that there is a long tail of large searches (see Figure 2). Upon examination of the queries, one reason for this long tail emerges: searchers are typing or pasting whole citations as the query (e.g. “Morsh, Joseph. (1930). Development of right-handed skill in the left-handed child. Child Development, 1(4), p311”). This behavior is a relatively new phenomenon (a citation search was used as an example in [17] but we could find no earlier reference to this behavior) but may change the face of typical searcher behavior in the future.

Finally, it should be noted that while the statistics for each search source look broadly similar, the variance between information sources was statistically significant in a one way ANOVA comparison ($F(3, 3726)=44.02, p<0.01$).

### 4.2 Query Content

A randomly selected sample of 100 queries per day was taken from searches of EBSCO, Google Scholar, and the catalogue; these 600 searches were examined manually for metadata content, known item searching, typographical errors, advanced search techniques and search success. Gale was excluded from this phase of the analysis due to its small number of queries (56 in total).

#### 4.2.1 Query Metadata

Searchers used a range of metadata, but by far the most popular were title, author and keyword. The distinction between keyword and title searches was made on the basis of matching: queries which matched three or more words in sequence in a resulting work title or any neologism in a title (e.g. “musicophilia”), or an exact title match of fewer than three words were deemed to be title searches (see Table 2). Title matches were exact in approximately 75% of all cases.

<table>
<thead>
<tr>
<th>Table 2. Metadata types used in each search source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library</td>
</tr>
<tr>
<td>Title</td>
</tr>
<tr>
<td>Author</td>
</tr>
<tr>
<td>Date</td>
</tr>
<tr>
<td>Publisher</td>
</tr>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Edition/issue</td>
</tr>
<tr>
<td>Pages</td>
</tr>
<tr>
<td>Keywords</td>
</tr>
</tbody>
</table>

Keyword searches (which can be assumed to be approximately equivalent to subject or topic searching) appeared in a variety of formats: some were traditional unconnected keyword searches (e.g. “aviation fatigue”) some were phrases relating to users’ information needs (e.g. “differences in depth perception in 2d and 3d”) and in
four cases (two each in Google Scholar and EBSCO), keyword searches were questions (e.g. “how many international students have health insurance”). The use of phrases or natural language queries occurred in about 15% of keyword searches in EBSCO and Google Scholar, and in 10% of searches of the library catalogue. This disparity is unsurprising, given that natural language searching is much more likely to be successful in online resources than in the catalogue.

When we examine the frequently used metadata types (metadata types seen in more than 5% of searches, i.e. title, author, date and keywords) there are differences in the frequency with which each metadata type is used between information resources. The differences between Google Scholar and the catalogue and EBSCO and the catalogue are significant ($\chi^2=8.546, \text{df}=3, p=0.036$ and $\chi^2=11.770, \text{df}=3, p=0.008$ respectively). The queries directed at Google Scholar and EBSCO were broadly similar; there are a number of possible reasons for this including nature of the content searched (Scholar and EBSCO both search online article-based content) and expertise of searchers (users of these information sources have demonstrated some expertise by selecting a non-default option).

A small but not insignificant number of searches contained more than one type of metadata; some were as simple as combining author and title metadata while some users entered whole citations. Citation markers (such as the words ‘et. al’, and ‘eds’, and enclosing dates in brackets—which are semantically empty and thus cannot improve search success) were seen in 1.5% of searches in EBSCO, 3% in the library catalog, and 4% in Google Scholar. There was no significant difference between search sources in the number of metadata types used, and in all sources a greater number of metadata types is strongly correlated with a greater number of search terms (as we would expect).

Finally, the library catalogue saw the entry of a specific metadata type not seen in the other sources: In 3% of cases catalogue searches were for course-related material using an alphanumeric course code. As this type of search was not seen in the other interfaces, and it is relevant only to the catalog, this further reinforces the likelihood that users are making intentional choices about information resource selection.

### 4.2.2 Known Item Searching

Queries were considered to be for a known item if they used title searching, or if a combination of other metadata (such as author and date) identified an individual document. Non-title searches accounted for about 8% of all known-item searches, as per Table 3 below. There were significant differences ($\chi^2=8.998, \text{df}=2, p=0.011$) between the three resources with respect to the number of known item searches,

<table>
<thead>
<tr>
<th></th>
<th>Library</th>
<th>EBSCO</th>
<th>Scholar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known items</td>
<td>109</td>
<td>53</td>
<td>61</td>
</tr>
<tr>
<td>Non-title known items</td>
<td>11</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>
Known item searches were classified by item type: by definition known item searches were for specific and identifiable texts, and each known-item query was classified as to the type of material sought. We discovered users were looking for a range of materials including books, articles and DVDs (see Table 4). It is interesting to note that (as we would hope, though the literature gives us no particular reason to expect this) known item searches of the catalogue are largely for books, while in EBSCO and Google Scholar they are mostly for articles. Scholar does show a number of book-related searches; however given that it returns relevant results from Google Books this strategy is likely to be successful.

<table>
<thead>
<tr>
<th>Item type sought</th>
<th>Library</th>
<th>EBSCO</th>
<th>Scholar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>68</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>Article</td>
<td>8</td>
<td>41</td>
<td>43</td>
</tr>
<tr>
<td>Journal</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Database</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DVD</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

### 4.2.3 Advanced Searching

As in other studies (e.g. [9-11, 17]) we found that only a small number of searches (4.5% across all resources) included advanced search techniques—13 (7.5%) each in EBSCO and Google Scholar, and 2 (1%) in the catalogue. The advanced catalogue searches were both well formed, but 4 in Google Scholar and 9 in EBSCO contained errors, echoing findings of other studies [9, 10, 17]. It is possible that more search modifiers may be observed if the interface captured search refinements, but the literature suggests that users are unlikely to use advanced query formulations even when refining searches.

### 4.2.4 Unsuccessful searches

A search could be unsuccessful in one of three ways: 1) no results were returned; 2) the search was for a known item which did not appear in the top 5 results (bad results); 3) acceptable results only appeared with a ‘did you mean’ search modification. There were slight variations between interfaces in the number of unsuccessful searches (Table 5).

<table>
<thead>
<tr>
<th>Unsuccessful searches</th>
<th>Library</th>
<th>EBSCO</th>
<th>Scholar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>42</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>No results</td>
<td>13</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Bad results</td>
<td>22</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Did you mean</td>
<td>7</td>
<td>25</td>
<td>15</td>
</tr>
</tbody>
</table>
It is evident from these results that the catalogue is the least ‘forgiving’ system; this is to be expected as it indexes only relatively limited content (which may also help to explain why users consistently find catalogues unusable [6, 8]: catalogues simply do not index the material they want, and do not communicate this clearly). Google Scholar is the most ‘forgiving’ with the fewest failed searches and a minimum of searches with no or bad results.

 Searches failed for a number of reasons. Typographical errors contributed to problems with search in about half of all search failures not all of which were corrected by “did you mean” functionality.

 Another cause of failed searching (3 cases in the catalogue and 2 in EBSCO) was including an entire (unedited) citation in the search. The only search source that performed at all well when given so much information was Google Scholar; EBSCO and the catalogue both routinely failed to return any results on such searches.

 A further cause of failed searching was searching in a system that did not index the kind of content sought (e.g. looking for books in EBSCO). While there was some cross-indexing, in all three systems some users failed to find what they were looking for because of the selected information source; there were 11 (1.8% of all searches) instances in total —5 (2.5%) each in EBSCO and the catalog, and 1 (0.5%) in Google Scholar. The low rate of failure in this way in Google Scholar is largely due cross-indexing of Google Books.

 Finally, in six instances (3%) the library catalogue failed to return results because the library did not have a copy of the item users sought.

5 Implications for Digital Libraries

 Many of our results mirrored those of other studies [9, 10, 17]: users still do not use advanced search techniques and still make a number of typographical errors. Digital library interfaces that expect users to perform complex queries without error are still likely to fail their users.

 There were some behaviors in evidence, however, that demonstrate new patterns of behavior: e.g. users entered citations as query terms relatively frequently, increasing the average number of search terms, and often resulting in search failure. Digital library interfaces would do well to facilitate searching with entire citations.

 We can reasonably assume that even those users searching the default information source (i.e. the library catalogue) generally expected to be searching for books; libraries are strongly associated with books [2] and only 2.5% of all catalogue searches examined in this study failed because the user was looking for articles or other resources more frequently indexed elsewhere. Unusually, the users in this study selected non-default search options 38% of the time—this is a dramatic contrast to other log analyses (for example [9, 17]) which show this behavior occurring only rarely. This suggests that the searchers in our study understood the differences between information sources and were intentional in their searching, a result previously only seen in topic experts and expert searchers [19, 29-31]. It is not clear
from this study why users were so knowledgeable about their options, but it suggests that differentiation of content in an initial search can be valuable to users.

The fact that users in this study did differentiate their search strategies based on the target information resource demonstrates a level of intention in their search strategy not previously observed outside of experts in either domain or searching [19, 30, 31], and suggests that search strategies which have, in the past, looked naïve may be merely economical. This suggests that digital libraries should aim to support very simple searches rather than encouraging more complex formulations.

Despite this apparently high level of understanding of the options on the library homepage it is evident from usage statistics that far more academic library users access more electronic content than books [26] which suggests that many users are not searching for digital content directly from the library homepage. This means that the opportunity for digital libraries to be a single information point, which users would very much like [24, 25] has not yet been realized.

Finally, library resources are not a first choice of information source for many users [2, 24], so we can reasonably assume that those searching from a library website have entered the ‘active searching’ phase of information seeking [3]. Nonetheless, active searching doesn’t imply known-item searching, which accounted for just slightly more than half of all searches in the library catalogue, but less in other sources. The not inconsiderable number of non-known item searches implies users will browse search results quite heavily; digital library systems should take this into account and facilitate effective browsing.

6 Conclusions

This paper presents the results of an exploratory analysis of queries performed over four information systems: an academic library catalogue, two scholarly article databases (Gale and EBSCO), and Google Scholar. The investigation was in two parts: a high level analysis of query structure over a large number of searches, and a more detailed analysis of query content based on sample searches of EBSCO, the catalogue and Google Scholar.

We found a surprisingly large number of searchers elected to search non-default information sources. Furthermore, there were differences in both the way users searched and in the results their searches returned, between different information sources. Longer queries were used in Google Scholar and EBSCO than in the catalog, but there were no differences in the number of metadata types queried across sources.

Similarly, queries of EBSCO and Google Scholar were more likely to be keyword searches than in the catalogue. Known item searches in Google Scholar and EBSCO targeted articles, whereas catalogue queries aimed at finding books. While citation searching was seen in all information sources, it was more common in Google Scholar where it was also more likely to be successful.

Given that we can discard interface effects, we must interpret these results to mean that users are intentionally varying their search strategies based on the information source searched. While the search strategies employed in each information source are
similarly simple, they are different and we must necessarily conclude that users goals and tactics vary specifically based on the target collection. This suggests that rather than being purely naïve, searchers tactics are rather parsimonious, a result previously only seen in domain experts and expert searchers.

We can (and should) take these results to show that digital libraries must continue to provide support for error recovery, allow users to preselect between collections and develop support for searching with unedited citations. More importantly, though, we must consider the possibility that users are more sophisticated in their search strategies than we had previously anticipated; the level of intention seen in this study is logical but previously undocumented.

7 Future work

Clearly there is scope for a more complete study to supplement this work: as this work was exploratory it analyzed only a relatively small number of queries, and an analysis of a larger sample to test the hypotheses about searching formed in this paper would be beneficial. Further, this work, like all log analyses, cannot unpack the motivations of searchers for behaving the way they do. The finding of non-default searching in this study is particularly striking, and it would be valuable to determine the drivers of this behavior in interviews or users studies.

8 References