Letting Users Lead: Analyzing Search Queries & Relevancy in USC's Web-Scale Discovery Tool

Elizabeth Namei, Reference & Instruction Librarian, University of Southern California Libraries
Christal A. Young, Reference & Instruction Librarian, University of Southern California Libraries

Presented at the California Academic & Research Libraries 2014 Conference
April 4-6, 2014
San Jose, California

Abstract

Many academic libraries have begun providing a single search box from which users can access the majority of their print and online collections. The demand for this came from users who are accustomed to a Google-like search experience. Yet, many librarians remain dubious. This paper presents the results of a study that re-executed a sample of *Summon* search queries in order to discern patterns regarding how users construct searches and whether or not USC's iteration of *Summon* actually delivers relevant content. By better understanding how users approach and experience a single search box, librarians will be better able to teach students to effectively use these tools. The results of this study reveal that discovery tools have a lot of work left to do to provide users with relevant results.

Introduction

In 2009 the University of Southern California (USC) Libraries began looking for a unified search solution to meet two primary goals: (1) to provide better discoverability of our vast subscription and purchased content; (2) to provide relevant results to our users—most urgently to provide relevance ranked results for items in our OPAC (SIRSI is our ILS and does not rank results by relevancy). In July 2010, *Summon* was chosen as the libraries' web-scale discovery layer and was made the default search tab on the libraries' homepage. Despite *Summon's* high usage numbers (3 million in 2013) there were a lot of complaints from library faculty and staff about getting "unexpected results" when searching *Summon*. The authors of this study wanted to gather evidence about how successful *Summon* actually was in leading users to relevant (or "expected") results. Another motivation for this study was to learn more about user search behavior when faced with a single search box.

Background

Web-scale discovery tools provide a single point of entry for accessing much (but not all) of the libraries' licensed and owned content. Key to understanding discovery tools is differentiating them from federated search tools. Discovery tools are built on a pre-harvested centralized index, which provide fast, relevance ranked results and faceted navigation to narrow down these large

result-sets. In contrast, federated search tools send out live queries to a handful of databases, which can be a slow process. Federated results are grouped according to which database they came from, which is not ideal, but was a better alternative to searching each database separately.

Discovery layers have been referred to as the "holy grail" of library search (Prescott & Erway, 2011). The enthusiasm and optimism surrounding discovery layers stem, in part, from vendors and early adopters pronouncing that they will lure users back to libraries for their information needs (Serials Solutions, 2014; Way, 2010). Project Information Literacy has found that the majority of users start their research using search engines (Head, 2013), so this claim has resonated with many academic libraries. Because library websites tend to provide numerous options to choose from, novice users are often uncertain about where/how to begin their research. Discovery tools have also been billed as providing a more intuitive and centralized (library) starting point for these users. One search box for all information needs "eliminate[s] the silo effect of an academic library's diverse databases" (Asher, Duke & Wilson, 2013, p. 475). It has also been claimed that these single search boxes maximize awareness and usage (Way, 2010, p. 12) of once "hidden collections" (Thomsett-Scott & Reese, 2012, p.123; Calvert, 2014) by revealing "everything the library owns or has licensed on a given topic, be it a print monograph, an electronic journal article, streaming video, or a collection of archival documents" (Little, 2012, p.346).

Literature Review

There has been a lot written about discovery tools since they were first introduced in 2008. Much of the literature provides qualitative data and anecdotal accounts. Many of the articles pertain to how a particular library chose and implemented a unified search solution. Many articles provide advice and best practices about teaching information literacy skills with these tools. There have also been a lot of usability studies done on discovery tools, most of which look at only a small handful of users, providing them with a set of tasks to complete and then interviewing them about their experiences.

More recently, qualitative studies have begun to emerge that analyze search queries and/or the search results of web-scale discovery tools. Meadow and Meadow (2012) were the first to analyze transaction logs from a web-scale discovery tool, but in their case they only measured the quality of searches conducted, not how well the discovery tool performed in retrieving useful and relevant results. They concluded: "By better understanding users, their expectations, successes, failures, and skill levels, libraries can offer more effective services to meet the needs of information seekers with the quality of library-provided resources" (p. 173). Lown, Sierra & Boyer (2013) gathered transaction log data from two semesters to see what types of sources users searched for in their *Summon* instance. They found that it was necessary to expand their notion of library search to go beyond article and catalog searching. Chapman et al. (2013) categorized a large sample of website search queries in order to learn more about the types of searches its users were conducting and to determine how successful their default search option was in providing results. Even though this study did not specifically analyze discovery tool searches it provided a useful rubric for categorizing search types. Another study conducted by Asher et al. (2013) compared "the efficacy of *EDS*, *Summon*, *Google Scholar* and 'conventional' library search tools on [typical] research tasks" encountered by undergraduates, with a specific focus on the quality of the results (peer reviewed/scholarly works were rated the highest). One key finding they had

was that the “single most important...factor in determining which resources students will utilize is the default way in which a particular search system ranks and returns results (p. 471). They concluded that, “more in-depth investigations of how particular search tools’ relevancy ranking algorithms function and differ is warranted given the critical role they play” (p. 477).

Methods

This study utilized transaction log analysis (TLA) to examine how well USC’s discovery instance was in delivering relevant results. A transaction log is a history of actions executed in an online system. TLA involves analyzing data captured in a transaction log to investigate interactions between users and a search tool. The goals of TLA are to gain clear and objective information about “interactions among searcher, content and system” (Jansen, 2006, p. 409) in order “to improve the design of the online system” (Asunka, 2009, p. 36).

The advantages of TLA are that it is unobtrusive, inexpensive and can provide large quantities of “real world data” (Jenson, 2006, p. 409), which reveal “large-scale search patterns” of user behavior (Wolfram, 2008, p. 1290). The drawbacks of TLA are that it leaves a lot of gaps and questions regarding user demographics, satisfaction and contextual details regarding the information need and search experience. TLA only captures the actions taken by users, "ignoring all other issues such as their emotions, perceptions, background, skills etc. that can potentially influence user satisfaction" (Asunka et al., 2009, p. 37).

According to USC’s transaction logs, there were a total of 1,243,250 searches executed in *Summon* in Fall 2013. Of those, 184,076 were unique. In order to achieve a +/-5% margin of error, or 95% confidence level (Krejcie & Morgan 1970, p. 608), a random sample of 384 searches was extracted from the transaction logs and re-executed in *Summon* for this study.

Before re-executing these searches, the authors developed and normed a rubric for categorizing search types and measuring the relevance of results. “Systems-oriented relevance decisions” were used to rate the success or failure of searches. This means we rated the relevance of *Summon’s* results without knowing the context of the original query. Instead, we looked at how well the topic of the search matched the topics of the results (Maglaughlin & Sonnenwald, 2002, p. 328-9). Rating relevancy is an inherently subjective endeavor and in many cases requires making assumptions about user intentions. In order to ensure that the two raters consistently agreed we spent several weeks “norming” our search type and relevancy definitions. We each executed and rated the same 96 searches, adjusting the rubric until we unanimously agreed. We measured our inter-rater reliability using Cohen's kappa statistic, which was “designed to estimate the degree of consensus between two judges” after correcting for agreement by chance (Stemler, 2004, p.4). Our Cohen’s kappa was 0.785, which indicates a substantial level of agreement.

Search Types Defined

- *Known item searches:* had to be specific enough for us to recognize a definitive match. If the search got numerous matches but was very general, we would not categorize it as a known item—e.g., “Catherine the Great” or “Ancient America.” We looked for visual clues to confirm if a query was for a known item: capitalization of major words, punctuation (colons) or words like "Introduction to" or "Third Edition."

- *Keyword/topic/exploratory searches:* these were broad, general or ambiguous. Named persons were included in this category. These queries could be specific and have numerous search terms entered, so long as there was no perfect match (we checked *Google* to verify).

Relevancy Rubrics

Asher et al. (2013) found that 92% of students in their study only utilized sources on the first page of results (p. 474). Pan et al.'s study (2007) found that most people only click on the first or second item in the results list (p. 803). Based on this research, we examined the first 5-10 results returned, depending on the query-type. We ranked relevancy differently for known item and non-known item searches, but both types were grouped into three relevancy categories: relevant, partially relevant, or not relevant (Maglaughlin & Sonnenwald, 2002, p. 328).

Known Item Relevancy Rubric

- *Relevant*: Assigned if the first result was a match
- *Partially Relevant*: Assigned if there was a match in the 2nd-10th results
- *Not relevant*: Assigned if the known item was not listed in the first 10 results or no results were returned (this could mean USC does not own the item or there was user error).

Topic/Keyword/Exploratory Relevancy Rubric

- *Relevant*: Assigned if all search terms appeared in the item's title or record. The first 5 items had to be "perfect" matches or clearly be about the topic.
- *Partially Relevant*: Assigned if all search terms were not visible in the title or record, but at least 3 of the first 5 results were somewhat related to the topic, even if only tangentially.
- *Not Relevant:* Assigned if 4-5 of the first 5 results appear to be false hits. If one or more of the search terms (or synonyms of those terms) appeared in the title or record, but results appear to be only about a portion of the search terms entered and not about all of them combined. Items that appeared to be false hits were checked to see if the search term(s) appeared in the record or full text.

Since vendors market discovery tools as the library's answer to *Google*, we wanted to see how *Summon*'s relevancy measured up to *Google* and *Google Scholar.* The same relevancy rubric was used for topic/keyword/exploratory searches in *Google* and *Google Scholar.* For known item searches in *Google* and *Google Scholar* a relevant or partially relevant result did not have to be a full-text match. Links to *Amazon*, *Google Books, WorldCat*, imdb.com, and *YouTube* were all considered matches.

Results

User Search Behavior

Out of 384 searches, 236 (61.5%) were topic/keyword/exploratory searches and 139 (36.2%) were for known items. Nine (2.3%) of the searches were unusable due to the presence of unrecognizable characters. The most commonly searched for known items were books (56) and articles (55); other searched for item types included: book chapters (8), videos (6), databases or journal titles (6), sound recordings (3), conference proceedings (3), and multiple formats (2). The

number of words entered in topic/keyword/exploratory queries ranged from 1-13, with the median being 3 words. The types of topic searches executed varied, but the most common, 139 (59%) were basic/simple searches, 26 (11%) included quotation marks, 24 (10%) used Boolean operators, 22 (9%) searched for a named person, 20 (9%) used natural language, 5 (2%) used command language or the advanced search form. These results correspond with Asher et al.'s (2013) finding that the majority (81.5%) of the searches observed were "simple keyword searches" (p. 473).

Relevancy Analysis

Summon returned "relevant" results in over half (124/53%) of the 236 topic/keyword/ exploratory searches. This compares negatively to *Google,* which had "relevant" results for 79% (187) of the queries. Surprisingly, *Summon* outperformed *Google Scholar*, which had "relevant" results for 45% (107) of the searches.

Summon returned "relevant" results in 82 (59%) of the 139 known item searches. Again, *Summon* compared negatively to *Google,* which had 131 (94%) known item queries ranked as "relevant." *Summon* fell slightly behind *Google Scholar* for known item searches, with *Google Scholar* returning "relevant" results for 96 (61%) searches. In order to see how well *Google Schola*r would fair on an even playing field we decided to remove items from our known item sample that *Google Scholar* was not designed to retrieve (16 queries were excluded: videos, sound recordings, database titles, and journal titles). We also excluded items that USC did not own (14), reducing the sample to 109 queries. The number of *Summon*'s relevant results went up 13% (to 72%-78/109), *Google* went up 1% (to 95%-104/109) and *Google Scholar* went up 15% (to 78%-85/109).

Conclusions

Summon has some work to do in providing more relevant results and thus does not yet provide a true Google-like search experience. Because discovery tools are still in their infancy, it is imperative that public service librarians advocate for improving them. This requires working "with database vendors to make sure users' needs are well-represented so vendors can create products that work well for students" (Rempel et al., 2013, p. 382). We must be proactive in discovering and fixing problems before users encounter them. "If we think like users (instead of as librarians) it is easy to understand the frustration. Our tools must seem broken or outdated to them" (Matthews, 2013). It is also important to empathize with all levels of users and "work with, not against, students' actual searching behavior….acknowledge and honor the students' expectations for easy search and access, and then lead them into the process of deeper thinking about their research assignments and search results as a path toward developing a set of generalizable research skills" (Rose-Wiles & Hofmann, 2013, p.156, 161).

In order to teach users how to effectively use these increasingly commonplace tools, librarians must change what and how we teach. This means moving away from teaching "advanced search features or Boolean logic if students purposefully choose not to use them or fail to use them correctly. Rather than teaching students more effective search syntax, more attention should be placed on developing critical thinking and evaluative skills" (Holman, 2011, p. 24). Discovery tools are designed to be success engines. Users no longer need to learn complicated search skills

to find information. Instead, more time can be spent helping students form mental models of online search, how to troubleshoot failed searches, explaining the publishing landscape and the politics and economics of information production.

Novices are not the only ones using these prominently positioned search boxes. Yet, usability studies have found that if someone has had training about specialized tools and search strategies they do not rate discovery tools favorably (Lundrigan et al., 2012, p. 10). Trained users have adapted to an imperfect and often inefficient process for locating information, but it is a process that ultimately works and meets their needs. As libraries transition to more intuitive and streamlined search options there will be a period of adjustment for experienced users. Librarians can help to smooth this transition by gathering evidence regarding the success or failure of these search tools, with the aim of improving them so that they actually deliver on their promise, and meet the expectations of all users.

References

Asher, A. D., Duke, L. M., & Wilson, S. (2013). Paths of discovery: Comparing the search effectiveness of EBSCO discovery service, Summon, Google Scholar, and conventional library resources. *College & Research Libraries*, *74*(5), 464-488.

Asunka, S., Chae, H. S., Hughes, B., & Natriello, G. (2009). Understanding academic information seeking habits through analysis of web server log files: The case of the Teachers College library website. *The Journal of Academic Librarianship*, *35*(1), 33-45. doi:10.1016/j.acalib.2008.10.019

Calvert, K. (2014). Maximizing academic library collections: measuring changes in use patterns owing to EBSCO Discovery Service. *College & Research Libraries.* Pre-print. Retrieved from CRL's website: http://crl.acrl.org/

Chapman, S., Desai, S., Hagedorn, K., Varnum, K., Mishra, S., & Piacentine, J. (2013). Manually classifying user search queries on an academic library web site. *Journal of Web Librarianship*, *7*(4), 401-421. doi:10.1080/19322909.2013.842096

Head, A.J. (2013). Project information literacy: What can be learned about the information-seeking behavior of today's college students? *Association of College and Research Libraries (ACRL) Conference Proceedings* 2013, Chicago: ALA.

Holman, L. (2011). Millennial students' mental models of search: Implications for academic librarians and database developers. *The Journal of Academic Librarianship*, *37*(1), 19-27. doi:10.1016/j.acalib.2010.10.003.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, *3*, 607-610.

Lown, C., Sierra, T., & Boyer, J. (2013). How users search the library from a single search box. *College & Research Libraries*, *74*(3), 227-241.

Lundrigan, C., Manuel, K., & Yan, M. (2012). 'Feels like you've hit the lottery': Assessing the implementation of a discovery layer tool at Ryerson University. Retrieved from Ryerson University's Digital Repository: http://digitalcommons.ryerson.ca/

Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology*, *53*(5), 327-342.

Matthews, B. (2013, August 21). Database vs. database vs. web-scale discovery service: Further thoughts on search failure (or: more clicks than necessary?) (or: info-pushers vs. pedagogical partners). *Ubiquitous Librarian.* Retrieved from the Chronicle of Higher Education website: http://chronicle.com/blognetwork/theubiquitouslibrarian

Meadow, K., & Meadow, J. (2012). Search query quality and web-scale discovery: A qualitative and quantitative analysis. *College & Undergraduate Libraries, 19*(2-4), 163-75. doi:10.1080/10691316.2012.693434.

Pan, B., Hembrooke, H., Joachims,T., Lorigo, L., Gay, G. & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*(3), 801-23. doi:10.1111/j.1083-6101.2007.00351.x.

Rose-Wiles, L. M., & Hofmann, M. A. (2013). Still desperately seeking citations: Undergraduate research in the age of web-scale discovery. *Journal of Library Administration, 53*(2-3), 147-66. doi:10.1080/01930826.2013.853493.

Serials Solutions. (2014) Returning researchers to the library. Retrieved from Serials Solutions website: http://www.serialssolutions.com/en/services/summon

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from PARE's website: http://pareonline.net

Thomsett-Scott, B., & Reese, P. E. (2012). Academic libraries and discovery tools: A survey of the literature. *College & Undergraduate Libraries*, *19*(2-4), 123-143. doi:10.1080/10691316.2012.697009

Wolfram, D. (2008). Search characteristics in different types of web-based IR environments: Are they the same? *Information Processing & Management*, *44*(3), 1279-1292. doi:10.1016/j.ipm.2007.07.010