Search Engine Queries Used to Locate Electronic Theses and Dissertations: Differences for Local and Non-Local Users

Midge Coates, Auburn University Libraries
coatemi@auburn.edu

Library Hi Tech, accepted (July 2014), scheduled for publication (2014)

Abstract:

Purpose:

This study examines two research questions: (1) What search engine queries lead users to the Auburn University Electronic Theses and Dissertations (AUETDs) collection? (2) Do these queries vary for users in different locations and, if so, how?

Design/methodology/approach:

Search engine queries used to locate the AUETDs collection were obtained from Google Analytics and were separated into groups based on user location. These queries were assigned to empirically-derived categories based on their content.

Findings:

Most local users' queries contained person names, variants for thesis or dissertation, and variants for Auburn University. Over a third were queries for the AUETDs collection, while the remainder were seeking theses and/or dissertations from specific Auburn researchers. Most out-of-state users' queries contained title and subject keywords and appeared to be seeking specific research studies. Queries from users located within the state but outside of the local area were intermediate between these groups.

Practical implications:

Over two-thirds of visits to the AUETDs collection were made by search engine users which reinforces the importance of having ETD repository content indexed by search engines such as Google. The specificity of their queries indicates that full-text indexing will be more helpful to users than metadata indexing alone.

Originality/value:

This is the first detailed analysis of search engine queries used to locate an ETDs collection. It may also be the last, as query content for the major search engines is no longer available from Google Analytics.

Search Engine Queries Used to Locate Electronic Theses and Dissertations: Differences for Local and Non-Local Users

Introduction

Many institutions offering graduate degrees maintain digital collections of electronic theses and dissertations (ETDs). These documents are made available to disseminate the knowledge produced by the institution. Examination of the search engine queries that bring users to these collections can guide ETD collection managers in improving findability and provide insight into how users expect to interact with the collection.

This study examines two research questions:

(1) What search engine queries lead users to the Auburn University ETDs (AUETDs) collection?

(2) Do these queries vary for users in different locations and, if so, how?

Search engine queries used to locate the AUETDs collection were obtained from Google Analytics and separated into groups based on user location. Queries were categorized based on their content, and differences between the location groups were identified.

Literature Review

Theses and dissertations are scholarly documents that report on original research performed by students in pursuit of graduate degrees. Lee-Smeltzer and Hackleman (1995) found that graduate students were the primary users of Oregon State University’s print collection of theses and dissertations. Students used these materials for research and as format templates. Chu and Law (2007) found that theses and dissertations were important for Hong Kong doctoral students in the transition period between gathering background information from books and review articles and consulting journal articles for more specific and scholarly information. Ismail and Kareem (2011) showed that theses and dissertations were a preferred scholarly resource for novice researchers in Malaysia because of their depth, breadth, and perceived trustworthiness.

Before researchers can use resources such as ETDs, they must be able to find them. Although databases and directories exist (NDLTD, 2013; OATD, 2013; ProQuest, 2014), McKay (2007) predicted that academic researchers would turn to Web search engines to locate scholarly materials before consulting library or institutional resources. Tenopir and Rowlands (2007) and Ismail and Kareem (2011) found that researchers at all levels used Web search engines such as Google to locate research materials. Institutional repository users interviewed by St. Jean, *et al*. (2011), said they found it easier to go to Google first when starting a research project.

Search Query Analysis

One approach to understanding how search engines are used to find online resources involves classification of the queries submitted to Web search engines. Broder (2002) created a classification scheme based on the (perceived) intent of the user. This now-classic taxonomy contains three query classes: navigational queries (in which users are seeking a site they know or assume exists); informational queries (in which users are seeking information); and transactional queries (in which users wish to conduct some action, such as shopping, downloading files, viewing videos, etc.). Rose and Levinson (2004) and Jansen, Booth, and Spink (2008) created revised versions that added hierarchical layers to the original Broder taxonomy.

Using a customized scheme based on query topics, Waller (2011) classified search engine queries from Google Australia and found that the three most popular categories were popular culture (25%), e-commerce (24%), and cultural practice (15%). Waller also compared results for users in 11 Australian location groups based on the premise that people living in the same location have similar lifestyles. No significant differences were found between the location/lifestyle groups.

Another approach involves the analysis of queries used to locate a particular collection. Herrara (2011) examined queries that brought Google Scholar users to University of Mississippi Library online resources. Of the 6,363 unique search engine queries recorded by Google Analytics in 2009, 345 resulted in multiple visits. Manual classification of the latter group showed that 38% were in the sciences, 54% were in the social sciences, and 2% were in the humanities.

Ortiz-Cordova and Jansen (2012) correlated use data for a music Web site with search engine queries used to locate the site. Cluster analysis identified six customer groups based on engagement with the site and revenue produced. Low engagement-low revenue customers came to the site via queries for specific information about songs or artists. High engagement-high revenue customers came via queries of the form: “music [genre or artist]”, “songs of [artist]”, and “listen to [genre or artist]”. Customers in the other four clusters used queries of both types but were more likely to name the Web site in their queries or to specify that they were seeking free materials.

Transaction Log Analysis

Data for most query studies are obtained from transaction logs. Jansen (2005) defines a transaction log as “an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine". (For these purposes, the term "Web search engine" includes search applications on Web sites.) Transaction log data is collected by the Web server in the background as users perform their searches. Although this method does not reveal searchers' motivations or their satisfaction with the results, it is a non-intrusive way to obtain data.

Transaction log analysis has been used to investigate searchers' interactions with library systems from online public access catalogs (Tolle, 1983; Blecic, *et al*., 1998) to discovery tools (Niu, Zhang, and Chen, 2014). Agosti, Crivellari, and Di Nunzio (2012) have reviewed studies which use transaction log analysis to study the interaction of searchers with Web search engines and digital libraries. These Web search engine studies examined query structure (e.g., length, number of terms) or correlated queries with search results clicked by users, while the digital libraries studies investigated users' interactions with online collections.

Several studies reviewed here used the Google Analytics Web tracking service to obtain data (Google, 2013; Wikipedia, 2013). This can be considered a mediated form of transaction logging. To implement it, Web designers add JavaScript tracking code provided by Google to their site pages. When a browser program accesses one of these pages, Google records data about that access, provided that the browser has enabled JavaScript, caching, and cookies. If any of these are not enabled, use data cannot be collected.

Data available from Google Analytics include page URLs, access dates and times, users' locations, referring Web sites, and, for search-mediated visits, search engine queries (called "keywords" by Google) which led users to access the Web site. Google Analytics provides queries for non-secure searches but withholds them for users making secure searches. Since late 2011, searches by logged-in Google users have been secure, and, since 23 September 2013, *all*

Google searches have been secure by default (Craver, 2013). Early in 2014, the Bing and Yahoo! search engines began rolling out secure search for their users (Sullivan, 2014; Slegg, 2014). With secure search becoming the default, search engine queries will no longer be obtainable via Google Analytics. In the future, server transaction logs may be the only source of queries data.

Use of Search Engines to Locate ETD Collections

ETDs were first introduced in the 1990s (Yiotis, 2008). There has been little published use data for ETDs, and even less on the use of search engines to locate them or their content. Use data provided for the Virginia Tech ETDs collection is limited to page views with no information about user locations or referring sources (Virginia Tech, 2014). Zhang, Lee, and You (2001) provided page views data for the ETDs collection of the Korea Institute of Science and Technology Information, but location information was limited to South Korea or “Other Countries”, and no referring source information was provided.

Alemneh and Phillips (2011) showed that search engine users accounted for 62% of visits to the University of North Texas ETDs repository during a seventeen-month period. They also reported that 1.4% of the search engine queries that brought users to the collection contained the terms “thesis” or “dissertation” but provided no further query analysis. An earlier study of the AUETDs repository reported that search engine users accounted for 68% of the visits to the collection during a one-year period and that 91% of those search engine users were located outside of the state of Alabama (Coates, 2014). No search engine query data were provided at that time. The current study examines more closely the behavior of AUETDs users who came to the collection via search engines and analyzes the queries used to locate the collection.

Methodology

The AUETDs collection was created by Auburn University Libraries (AUL) as a DSpace repository (http://etd.auburn.edu/etd/) and is indexed by Web search engines such as Google, Bing, and Yahoo! and by Google Scholar. Bibliographic information pages for individual ETDs provide titles, author names, advisor names, department names, abstracts, dates, and links to full-text PDFs. About 400-500 ETDs are added to the collection each year. As of 21 August 2013, the collection contained 3,467 theses and dissertations.

Use statistics for the AUETDs collection are obtained using the Standard (free) version of the Google Analytics Web tracking service (Google, 2013). This study examined use data and search engine queries data for a six-month study period from 22 February 2013 through 21 August 2013. This period fell within the time window when Google searches were secure for logged-in searchers but not for all searchers. Thus, queries were available for non-secure Google searches and for searches made using other search engines such as Yahoo! and Bing. For this study, the assumption has been made that search behavior was similar for users of both secure and non-secure search. This study also assumes that user search behavior has not changed significantly since 23 September 2013, and that conclusions drawn from this study will continue to be valid for the near future.

The latter assumption is supported by two longitudinal studies. Spink, *et al*. (2002), compared sets of queries submitted by users to the Web search engine Excite in 1997, 1999, and 2001. Although search topics changed over the four-year period, user search behaviors such as query lengths and queries per user did not. Wang, Berry, and Yang (2003) examined search queries submitted to an academic search engine over a four-year period. They found that user behavior was consistent in terms of the lengths of the queries submitted and the most popular query terms.

Use data and queries data obtained from Google Analytics were filtered according to user location. In some cases, the data were also sorted by page type based on URL. All data sets were exported as comma-separated values documents, which were imported into Excel for analysis. Categories for search engine queries were empirically derived from the queries data and are listed in Table 1. All queries obtained from Google Analytics were manually assigned by the author to all categories that seemed appropriate.

Results and Discussion

For this study, users have been divided into four groups based on location information provided by Google Analytics. The Local group consisted of users whose region location was Alabama and whose city location was Auburn or Opelika (a city adjacent to Auburn). The Alabama–not Local group consisted of users whose region location was Alabama but excluded Auburn and Opelika users. The USA–not Alabama group consisted of users whose country location was given as United States but excluded all Alabama users, while the World–not USA group consisted of all users except those whose country location was United States.

Collection Use Data

Data for the overall use of the AUETDs collection by the four location groups are given in the upper half of Table 2. Most visits during the study period were from users in the two out-of-state groups—USA–not Alabama (37%) and World–not USA (44%). This is consistent with an earlier study of collection use by the four location groups (Coates, 2014). The percentage of visits originating at search engines was smallest for the Local group, larger for the Alabama–not

Table 1. Query categories for analysis of search engine queries that brought users to AUETDs collection during the study period 22 February 2013 through 21 August 2013.

Query categories
• Person name(s) (unless clearly literary or historical, e.g., Mark Twain) • Variant for Auburn University • Name of another university • Variant for thesis or dissertation (including ETD) • Variant for MS or PhD degree • URL(s), full or partial • Date(s) (unless clearly historical, e.g., 1825-1838) • Thesis or dissertation title, full or partial • Non-dissertation document title, full or partial (Any string of words with a colon was deemed to be a document title, as were long phrases lacking verbs.) • Citation (usually name(s), date, article title, journal or book title, sometimes volume and page numbers) • Title keyword(s) (terms found in titles of documents in AUETDs collection; checked against title metadata) • Subject keyword(s) (terms not found in AUETDs titles; checked against title metadata) • Question (query phrased as a question, e.g., What are the causes of …; Discuss the causes of …) • Free (query containing the word “free”) • PDF(s) (query containing the word “PDF” or “PDFs”) • Article(s) (query containing the word “article” or “articles”) • Book(s) (query containing the word “book” or “books”) • Video(s) (query containing the word “video” or “videos”) • Download (query containing the word “download”) • Upload (query containing the word “upload”) • Submit (or submission) (query containing the word “submit” or “submission”) • Format (query containing the word “format”) • Template (query containing the word “template”) • Log in (or ldap) (query containing the word “log in” or “login” or “ldap”)

Table 2. AUETDs user visit data for the study period 22 February 2013 through 21 August 2013, filtered by user location. Top portion of the table contains data for all user visits, while bottom portion contains data for visits from search engine users.

		Local Users	Alabama–not Local Users	USA–not Alabama Users	World–not USA Users	Collection Total
All Users	Overall visits (% total visits)	6,110 (15%)	1,494 (4%)	15,058 (37%)	17,804 (44%)	40,466 (100%)
	Search engine visits as % of overall visits	34%	44%	72%	78%	68%
	Avg. pages viewed/visit	8.18	5.77	2.16	2.00	3.13
	% Overall visits landing on home page	53%	32%	7%	4%	14%
	% Overall visits landing on bibliographic information pages	24%	47%	81%	89%	75%
Search Engine Users Only	Search visits (% total search visits)	2,103 (8%)	662 (2%)	10,897 (40%)	13,803 (50%)	27,465 (100%)
	Avg. pages viewed/ visit	5.45	3.60	1.92	1.79	2.16
	% Search visits landing on home page	39%	14%	4%	2%	6%
	% Search visits landing on bibliographic information pages	32%	65%	87%	93%	85%
	Visits for which search engine queries were available (i.e., non-secure searches)	947	378	7,444	10,586	19,355
	Visits for which search engine queries were available as % of all search visits	45%	57%	68%	77%	70%

Local group, and largest for the two out-of-state groups.

As the percentage of search engine visits increased, the number of pages viewed per visit decreased, and the percentage of visits beginning at the home page decreased while the percentage beginning at an individual item’s bibliographic information page increased. This observation is consistent with Mahoui and Cunningham’s (2001) study of the ResearchIndex digital library. Transaction logs showed that 47% of ResearchIndex users bypassed the collection’s search page. Mahoui and Cunningham postulated that these users had located collection documents directly via external search engines.

The trends with respect to search engine visits, pages viewed per visit, and visits beginning at the home page vs. an individual item page are probably related. Users outside of the Auburn University community were less likely to be aware of the collection and thus more likely to find its content via a search engine rather than a University Web site. As noted in an earlier study, University Web sites point users to the collection home page, while search engines may point either to the home page or to individual item pages (Coates, 2014). Users who land on the home page must use internal search and browse pages to locate documents and will probably view more pages per visit than users who land directly on bibliographic information pages.

The lower half of Table 2 contains data for just the search engine users in the four location groups. As expected, the search engine users viewed fewer pages per visit than the overall location groups. Also as expected, the search engine users in each location group were less likely than the overall group to land on the home page and more likely than to land on a bibliographic information page.

This section of Table 2 also lists the number of visits for which Google Analytics provided queries data—visits from users making non-secure searches. The remainder of this study will focus on this sub-group of searches for which queries were available. The percentage of non-secure searches was smallest for the Local group, larger for the Alabama–not Local group, and largest for the two out-of-state groups. It is difficult to conjecture why this pattern occurred.

Query Analysis Using Empirically-Derived Categories

All search engine queries obtained from Google Analytics were manually assigned by the author to the applicable query categories listed in Table 1 based on the terms contained in them. Query terms corresponding to many Table 1 categories (e.g., person name, date, PDF, article) were easily assigned. Terms not easily assigned to a specific category (e.g., catfish, consumer, hospital) were designated as keywords and searched against a database of AUETDs document titles. Those found in the database were assigned to the title keywords category, and the rest were assigned to the subject keywords category.

For assignment purposes, differences in singular vs. plural nouns were ignored, as were differences in verb endings such as -ed and -ing. Article words (e.g., a, an, the) and prepositions were also ignored. Words with obvious misspellings and/or transposed letters (e.g., finnance, disertation, htesis, aubrun) were treated as if they were spelled correctly.

Table 3 shows the most common categories for each location group and for all queries. During the study period, the most frequently used categories were title keywords, subject keywords, person name, variant for thesis or dissertation, variant for Auburn University, full or partial title of a thesis or dissertation in the AUETDs collection, and title for a research document not in AUETDs. Because queries were assigned to all appropriate categories, percentages add up to more than 100%.

Table 3. Most common categories for search engine queries (supplied by Google Analytics) for AUETDs collection during the study period 22 February 2013 through 21 August 2013, filtered by user location. Search engine queries were assigned to all applicable categories, so percentages total more than 100%.

	Search engine queries, total	Queries containing title keywords	Queries containing subject keywords	Queries containing person name	Queries containing thesis*	Queries containing AU*	Queries containing thesis title*	Queries containing other title*	Categories per query, avg.
Local Users	947	183 (19% of Local queries)	75 (8% of Local queries)	404 (43% of Local queries)	455 (48% of Local queries)	552 (58% of Local queries)	43 (5% of Local queries)	19 (2% of Local queries)	2.1
Alabama–not Local Users	378	197 (52% of Alabama–not Local queries)	66 (17% of Alabama–not Local queries)	97 (26% of Alabama–not Local queries)	64 (17% of Alabama–not Local queries)	113 (30% of Alabama–not Local queries)	17 (4% of Alabama–not Local queries)	12 (3% of Alabama–not Local queries)	1.7
USA–not Alabama Users	7,444	4,800 (64% of USA–not Alabama queries)	2,800 (38% of USA–not Alabama queries)	1,098 (15% of USA–not Alabama queries)	450 (6% of USA–not Alabama queries)	603 (8% of USA–not Alabama queries)	546 (7% of USA–not Alabama queries)	378 (5% of USA–not Alabama queries)	1.6
World–not USA Users	10,586	7,166 (68% of World–not USA queries)	3,318 (31% of World–not USA queries)	1,147 (11% of World–not USA queries)	796 (8% of World–not USA queries)	448 (4% of World–not USA queries)	854 (8% of World–not USA queries)	863 (8% of World–not USA queries)	1.6
All User Queries	19,355	12,346 (64% of all queries)	6,259 (32% of all queries)	2,746 (14% of all queries)	1,765 (9% of all queries)	1,716 (9% of all queries)	1,460 (8% of all queries)	1,272 (7% of all queries)	1.6

* Abbreviations: thesis = variant for thesis or dissertation; AU = variant for Auburn University; thesis title = full or partial title of thesis or dissertation in AUETDs collection; other title = title for document not in AUETDs collection.

Search engine queries that brought users to the collection frequently contained terms corresponding to multiple query categories. The average number of categories per query is reported for each location group and for all queries in the last column of Table 3. Differences between the location groups were smaller than expected, with values ranging from 1.6 to 2.1 categories per query. Tables 4-8 present data for the most common combinations for the five most used categories: title keywords, subject keywords, person name, variant for thesis or dissertation, and variant for Auburn University.

Search engine queries assigned to the title keywords and subject keywords categories ranged from 19% and 8%, respectively, for users in the Local group to 68% and 31%, respectively, for users in the World–not USA group (Table 3). Table 4 shows that the most popular combinations for the title keywords category were title keywords alone and title keywords + subject keywords. For Local users, the combination of person name + title keywords +/- other term(s) was also popular. Other combinations were used less frequently. The subject keywords category was used most often in combination with terms corresponding to title keywords (Table 5). All other combinations, including subject keywords alone, were used significantly less frequently.

Search engine queries assigned to the person name category ranged from 43% for users in the Local group to 11% for users in the World–not USA group (Table 3). Table 6 shows the most popular combinations for the person name category. In-state users were most likely to use person name alone, person name + variant for Auburn University +/- other term(s), and person name + variant for thesis or dissertation +/- other term(s). Out-of-state users used person name alone, person name + variant for Auburn University +/- other term(s), person name + title keywords +/- other term(s), and person name + date +/- other term(s), although differences were less marked.

Table 4. Queries containing terms categorized as “title keywords” for the study period 22 February 2013 through 21 August 2013, filtered by user location.

	Search engine queries, total	Queries containing title keywords alone	Queries containing title keywords + subject keywords alone	Queries containing person name + title keywords +/- other terms	Queries containing thesis* + title keywords +/- other terms	Queries containing title keywords + AU* +/- other terms	All queries containing title keywords +/- other terms
Local Users	947	59 (6% of Local queries)	45 (5% of Local queries)	61 (6% of Local queries)	1 (<1% of Local queries)	19 (2% of Local queries)	183 (19% of Local queries)
Alabama–not Local Users	378	102 (27% of Alabama–not Local queries)	50 (13% of Alabama–not Local queries)	12 (3% of Alabama–not Local queries)	1 (<1% of Alabama–not Local queries)	30 (8% of Alabama–not Local queries)	197 (52% of Alabama–not Local queries)
USA–not Alabama Users	7,444	2,083 (28% of USA–not Alabama queries)	2,141 (29% of USA–not Alabama queries)	250 (3% of USA–not Alabama queries)	95 (1% of USA–not Alabama queries)	96 (1% of USA–not Alabama queries)	4,800 (64% of USA–not Alabama queries)
World–not USA Users	10,586	3,509 (33% of World–not USA queries)	2,469 (23% of World–not USA queries)	356 (3% of World–not USA queries)	383 (4% of World–not USA queries)	16 (<1% of World–not USA queries)	7,166 (68% of World–not USA queries)
All User Queries	19,355	5,753 (30% of all queries)	4,705 (24% of all queries)	679 (4% of all queries)	486 (3% of all queries)	161 (<1% of all queries	12,346 (64% of all queries)

* Abbreviations: thesis = variant for thesis or dissertation; AU = variant for Auburn University.

Table 5. Queries containing terms categorized as “subject keywords” for the study period 22 February 2013 through 21 August 2013, filtered by user location.

	Search engine queries, total	Queries containing subject keywords alone	Queries containing title keywords + subject keywords alone	Queries containing person name + subject keywords +/- other terms	Queries containing thesis* + subject keywords +/- other terms	Queries containing subject keywords + AU* +/- other terms	All queries containing subject keywords +/- other terms
Local Users	947	5 (<1% of Local queries)	45 (5% of Local queries)	18 (2% of Local queries)	0 (0% of Local queries)	7 (<1% of Local queries)	75 (8% of Local queries)
Alabama–not Local Users	378	3 (<1% of Alabama–not Local queries)	50 (13% of Alabama–not Local queries)	8 (2% of Alabama–not Local queries)	0 (0% of Alabama–not Local queries)	5 (1% of Alabama–not Local queries)	66 (17% of Alabama–not Local queries)
USA–not Alabama Users	7,444	380 (5% of USA–not Alabama queries)	2,141 (29% of USA–not Alabama queries)	135 (2% of USA–not Alabama queries)	29 (<1% of USA–not Alabama queries)	45 (<1% of USA–not Alabama queries)	2,800 (38% of USA–not Alabama queries)
World–not USA Users	10,586	431 (4% of World–not USA queries)	2,469 (23% of World–not USA queries)	145 (1% of World–not USA queries)	127 (1% of World–not USA queries)	4 (<1% of World–not USA queries)	3,318 (31% of World–not USA queries)
All User Queries	19,355	819 (4% of all queries)	4,705 (24% of all queries)	306 (2% of all queries)	156 (<1% of all queries)	61 (<1% of all queries)	6,259 (32% of all queries)

* Abbreviations: thesis = variant for thesis or dissertation; AU = variant for Auburn University.

Table 6. Queries containing terms categorized as “person name” for the study period 22 February 2013 through 21 August 2013, filtered by user location.

	Search engine queries, total	Queries containing person name alone	Queries containing person name + date +/- other terms	Queries containing person name + AU* +/- other terms	Queries containing person name + thesis* +/- other terms	Queries containing person name + title keywords +/- other terms	Queries containing person name + subject keywords +/- other terms	All queries containing person name +/- other terms
Local Users	947	98 (10% of Local queries)	14 (1% of Local queries)	177 (19% of Local queries)	104 (11% of Local queries)	61 (6% of Local queries)	18 (2% of Local queries)	404 (43% of Local queries)
Alabama–not Local Users	378	32 (8% of Alabama–not Local queries)	3 (<1% of Alabama–not Local queries)	39 (10% of Alabama–not Local queries)	20 (5% of Alabama–not Local queries)	12 (3% of Alabama–not Local queries)	9 (2% of Alabama–not Local queries)	97 (26% of Alabama–not Local queries)
USA–not Alabama Users	7,444	316 (4% of USA–not Alabama queries)	134 (2% of USA–not Alabama queries)	278 (4% of USA–not Alabama queries)	103 (1% of USA–not Alabama queries)	250 (3% of USA–not Alabama queries)	135 (2% of USA–not Alabama queries)	1,098 (15% of USA–not Alabama queries)
World–not USA Users	10,586	241 (2% of World–not USA queries)	334 (3% of World–not USA queries)	174 (2% of World–not USA queries)	64 (<1% of World–not USA queries)	355 (3% of World–not USA queries)	145 (1% of World–not USA queries)	1,147 (11% of World–not USA queries)
All User Queries	19,355	687 (4% of all queries)	485 (3% of all queries)	668 (3% of all queries)	291 (2% of all queries)	678 (4% of all queries)	307 (2% of all queries)	2,746 (14% of all queries)

* Abbreviations: thesis = variant for thesis or dissertation; AU = variant for Auburn University.

In Herrara's study (2011), the latter combination was interpreted as an abbreviated citation.

Search engine queries assigned to the variant for thesis or dissertation and variant for Auburn University categories ranged from 48% and 58%, respectively, for users in the Local group to 8% and 4%, respectively, for users in the World–not USA group (Table 3). Table 7 shows data for combinations containing terms corresponding to variant for thesis or dissertation, while Table 8 shows data for combinations of variant for Auburn University. The most significant query combination for these categories was a combination of the two. This group of users was clearly seeking the collection itself rather than a thesis or dissertation in the collection. This combination was especially popular with in-state users, many of who may have been ETD submitters rather than end-users (Coates, 2014). Other significant combinations for in-state users were person name + variant for thesis or dissertation (Table 7) and person name + variant for Auburn University (Table 8). Out-of-state users showed little preference for any particular combination of these terms.

A number of queries consisted of full URLS for individual theses and dissertations, i.e., http://etd.auburn.edu/etd/[collection number]/[item number] (Tables 7 and 8). It may seem odd to use a URL as a search query, when one could navigate directly to the document by pasting that URL into the browser's address bar. Some searchers may have been using the Google Chrome browser which has merged the browser address bar with a search bar (Wikipedia, 2014). However, Lee and Sanderson (2010) have shown that searchers attempting to re-find Web documents viewed earlier sometimes use partial URLs as search engine queries and take advantage of the auto-fill feature to obtain full URLs for these documents. Teevan, *et al.*, (2007) have postulated that as many as 40% of all search engine queries may be attempts to re-find previously-viewed documents.

Table 7. Queries containing the term “variant for thesis or dissertation” for the study period 22 February 2013 through 21 August 2013, filtered by user location.

	Search engine queries, total	Queries containing URLs for specific ETDs (queries seeking individual collection items)	Queries containing thesis* + AU* +/- other terms (queries seeking the collection itself)	Queries containing person name + thesis +/- other terms	Queries containing thesis + title keywords +/- other terms	Queries containing thesis + subject keywords +/- other terms	All queries containing thesis +/- other terms
Local Users	947	4 (<1% of Local queries)	406 (43% of Local queries)	104 (11% of Local queries)	1 (<1% of Local queries)	0 (0% of Local queries)	455 (48% of Local queries)
Alabama–not Local Users	378	5 (1% of Alabama–not Local queries)	52 (14% of Alabama–not Local queries)	20 (5% of Alabama–not Local queries)	1 (<1% of Alabama–not Local queries)	0 (0% of Alabama–not Local queries)	64 (17% of Alabama–not Local queries)
USA–not Alabama Users	7,444	78 (1% of USA–not Alabama queries)	187 (3% of USA–not Alabama queries)	104 (1% of USA–not Alabama queries)	95 (1% of USA–not Alabama queries)	29 (<1% of USA–not Alabama queries)	450 (6% of USA–not Alabama queries)
World–not USA Users	10,586	176 (2% of World–not USA queries)	102 (1% of World–not USA queries)	64 (<1% of World–not USA queries)	389 (4% of World–not USA queries)	126 (1% of World–not USA queries)	796 (8% of World–not USA queries)
All User Queries	19,355	263 (1% of all queries)	693 (4% of all queries)	292 (2% of all queries)	486 (3% of all queries)	155 (<1% of all queries)	1,765 (9% of all queries)

* Abbreviations: thesis = variant for thesis or dissertation; AU = variant for Auburn University.

Table 8. Queries containing the term "variant for Auburn University" for the study period 22 February 2013 through 21 August 2013, filtered by user location.

	Search engine queries, total	Queries containing URLs for specific ETDs (queries seeking individual collection items)	Queries containing thesis* + AU* +/- other terms (queries seeking the collection itself)	Queries containing person name + AU +/- other terms	Queries containing title keywords + AU +/- other terms	Queries containing subject keywords + AU +/- other terms	All queries containing AU +/- other terms
Local Users	947	4 (<1% of Local queries)	406 (43% of Local queries)	177 (19% of Local queries)	19 (2% of Local queries)	7 (<1% of Local queries)	552 (58% of Local queries)
Alabama–not Local Users	378	5 (1% of Alabama–not Local queries)	52 (14% of Alabama–not Local queries)	39 (10% of Alabama–not Local queries)	30 (8% of Alabama–not Local queries)	5 (1% of Alabama–not Local queries)	113 (30% of Alabama–not Local queries)
USA–not Alabama Users	7,444	79 (1% of USA–not Alabama queries)	187 (3% of USA–not Alabama queries)	278 (4% of USA–not Alabama queries)	96 (1% of USA–not Alabama queries)	45 (<1% of USA–not Alabama queries)	603 (8% of USA–not Alabama queries)
World–not USA Users	10,586	176 (2% of World–not USA queries)	102 (1% of World–not USA queries)	174 (2% of World–not USA queries)	16 (<1% of World–not USA queries)	4 (<1% of World–not USA queries)	448 (4% of World–not USA queries)
All User Queries	19,355	263 (1% of all queries)	693 (4% of all queries)	668 (3% of all queries)	161 (<1% of all queries)	61 (<1% of all queries)	1,716 (9% of all queries)

* Abbreviations: thesis = variant for thesis or dissertation; AU = variant for Auburn University.

Some search engine queries consisted of full or partial titles for documents in the AUETDs collection (Table 3). In other instances, queries consisted of titles or full citations for documents *not* in the collection. The fact that these latter queries brought users to the collection anyway suggests that these documents had been cited in ETD bibliographies and that Google had indexed the entire texts, rather than just the metadata.

Few queries were used more than once or twice. The most common exceptions were terms or combinations corresponding to the collection itself, e.g., http://etd.auburn.edu, AUETD, variant for Auburn University + variant for thesis or dissertation, variant for Auburn University + variant for thesis or dissertation + online, variant for Auburn University + variant for thesis or dissertation + electronic. That some users were seeking the collection itself is consistent with data in Table 2 which show that some search engine users landed on the collection home page rather than on individual item pages.

Location Group Search Behavior

Local Group

The Local group consisted of users living or working in the Alabama cities of Auburn and Opelika. Table 2 shows that 34% of these users came to the AUETDs Web site via search engines. While 43% of the searchers in this group were seeking the overall collection (Tables 7, 8), the rest were using search engines to locate individual ETDs directly without using the collection's search and browse pages. This is consistent with the observation that search engine users viewed fewer pages per visit than the overall location group and were more likely to begin their visits at individual items' bibliographic information pages (Table 2).

Most searchers in this group composed their queries from person names, variants for thesis or dissertation, and variants for Auburn University (Table 3). Title and subject keywords were used less frequently by this group than by other location groups. These results suggest that many searchers in the Local group were seeking specific theses and dissertations from specific degree candidates at Auburn University.

Alabama–not Local Group

The Alabama–not Local group consisted of users located within the state of Alabama but outside of Auburn and Opelika. Table 2 shows that 44% of these users came to the Web site via search engines. As with the Local group, search engine users viewed fewer pages per visit than the overall location group and were more likely to land on bibliographic information pages, which suggests that search engines made it possible for many in this group to locate individual ETDs directly. Only 14% of the searchers in this group were looking for the overall collection (Tables 7, 8).

Most searchers in this group composed their queries from title keywords, person names, variants for Auburn University, subject keywords, and variants for thesis or dissertation (Table 3). These results suggest that fewer searchers in this group were seeking specific theses and dissertations from specific degree candidates at Auburn University, as compared to users in the Local group. Instead, many seemed to be seeking articles documenting specific research studies.

USA–not Alabama Group

The USA–not Alabama group consisted of users located within the United States but outside of the state of Alabama. Table 2 shows that 72% of this group came to the collection via search engines. Differences were slight between the sub-group of search engine users and the overall location group, probably due to the large proportion of search engine users in this group. Only 3% of the searchers in this group were looking for the overall collection (Tables 7, 8). The remaining 97% were using search engines to navigate directly to individual ETDs.

Most searchers in this group composed their queries from title and subject keywords (Table 3). Person names were used infrequently, and other query categories were used even less often. These results suggest that most searchers in this location group were seeking articles documenting specific research studies rather than specific theses and dissertations.

World–not USA Group

The World–not USA group consisted of users located outside of the United States. Table 2 shows that 78% of this group came to the collection via search engines. As with the USA–not Alabama group, differences were slight between search engine users and the overall location group. Only 1% of the searchers in this group were looking for the overall collection (Tables7, 8). The remaining 99% were using search engines to navigate directly to individual ETDs.

Most searchers in this group composed their queries from title and subject keywords (Table 3). All other query categories were used infrequently. These results suggest that most searchers in this location group were seeking articles documenting specific research studies rather than specific theses and dissertations.

Comparing Queries for AUETDs with Other Repositories

It is difficult to make direct comparisons of queries used to locate AUETDs to those used to locate other ETD repositories because of the scarcity of data reported in the literature. No search engine queries data have been provided for the Virginia Tech ETDs repository (Virginia

Tech, 2013) or for the Korea Institute of Science and Technology Information ETDs collection (Zhang, Lee, and You, 2001). To date, Alemneh and Phillips (2011) have provided the only search engine query analysis available for ETD collections. They reported that 1.4% of the queries bringing users to the University of North Texas ETDs collection contained the words “thesis” or “dissertation”. Most of the query examples they showed were similar in pattern to those that brought users to the AUETDs collection, e.g., variant for University of North Texas + variant for thesis or dissertation, person name + variant for thesis or dissertation, keywords + variant for thesis or dissertation.

Implications for Research and Practice

This study was concerned with how users found the AUETDs collection using Web search engines. Analysis of the queries used to locate the collection suggests that most users who came to the collection via search engines were looking for the collection itself, for individual theses and dissertations, or for articles documenting specific research projects. It would be interesting to see if this holds true for other scholarly collections.

One limitation of this study is that the queries studied here were necessarily those that were *successful* in locating the collection. There seems to be no convenient way to examine *unsuccessful* queries, i.e., those relevant to AUETDs documents which did *not* drive traffic to the collection. However, the specificity of the successful queries suggests that repository managers wishing to increase user traffic should ensure that their collections are indexed by the major search engines and provide metadata as complete as practical (e.g., topical keywords, document abstracts, documents’ full-text if possible).

Another limitation is that Google Analytics data, like transaction logs, cannot determine users’ actual intents. One way to obtain this information might be a pop-up survey asking users who come to the collection via search engines what kinds of documents they were hoping to find when they formulated their queries.

A more serious limitation is the unavailability of queries from users of secure search. This study is the first detailed analysis of search engine queries used to locate an ETDs collection. It may also be the last, as the default mode for the major search engines has become secure search, and queries from these will no longer be available from Google Analytics.

Conclusions

The first research question this study addressed was: What search engine queries lead users to the AUETDs collection? During the study period, the most frequently used search engine queries included, in descending order, title keywords, subject keywords, person names, variants for thesis or dissertation, variants for Auburn University, and full or partial titles for theses and dissertations and for other research documents.

The second question addressed in this study was: Do these queries vary for users in different locations and, if so, how? Table 3 shows that there were location-based differences in query construction. Users in Auburn and Opelika were more likely to use person names, variants for Auburn University, and variants for thesis or dissertation in their queries, while users in the two out-of-state groups were much more likely to use title and/or subject keywords. Alabama users not located in Auburn and Opelika exhibited behavior in-between that of users in the Local group and users in the out-of-state groups.

Many users in the two in-state groups appeared to be seeking specific theses and dissertations, while others were seeking the collection itself (Tables 7, 8). Users in the two out-of-state groups rarely searched for specific theses and dissertations or for the AUETDs collection by itself. It is reasonable to assume that most members of these two latter groups had no prior knowledge of the collection. However, these groups used title keywords in their search queries twice as often as subject keywords (Table 3). This suggests they did have prior knowledge of the research studies for which they were seeking articles.

The fact that over two-thirds of visits to the AUETDs collection were made by search engine users reinforces the importance of having repository content indexed by search engines such as Google. The specificity of the queries that brought users to the collection indicates that full-text indexing will be more helpful to users than metadata indexing alone.

McKay (2007) said that, without end-user research, repository managers could not know (1) whether users were local or were located outside the institution; (2) whether users found the repository via institutional sources or external search engines; (3) what kind of information they sought and used; and (4) how they used the functionality offered by the repository. An earlier study of the AUETDs collection showed that 82% of its end-users were located outside of the state of Alabama, and that 76% of out-of-state users and 33% of in-state users found the collection via search engines (Coates, 2014). This follow-up study supports those findings and shows that most search engine users who came to the collection were looking for specific theses and dissertations and/or articles documenting specific research studies. Future research will address the kinds of documents that collection users view and download and how the collection's functionalities are used.

References

Agosti, M., Crivellari, F., and Di Nunzio, G. M., (2012), "Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction", *Data Mining and Knowledge Discovery*, Vol. 24, No. 3, pp. 663-696.

Alemneh, D. G., and Phillips, M. E. (2011), "Assessing the usage of electronic theses and dissertations: An overview of ETD statistics and metrics in the UNT libraries", paper presented at Texas ETD Association Annual Conference, 31 March 2011, Arlington, Texas, available at: http://digital.library.unt.edu/ark:/67531/metadc32969/ (accessed 9 May 2013).

Blecic, D., Bangalore, N. S., Dorsch, J. L., Henderson, C. L., Koenig, M. H., and Weller, A. C., (1998), "Using transaction log analysis to improve OPAC retrieval results", *College and Research Libraries,* Vol. 59, No. 1, pp. 39-50.

Broder, A., (2002), "A taxonomy of web search", *SIGIR Forum,* Vol. 36, No. 2, pp. 3-10.

Chu, S., and Law, N., (2007), "Development of information search expertise: Research students' knowledge of source types", *Journal of Librarianship and Information Science*, Vol. 39, No. 1, pp. 27-40.

Coates, M., (2014), "Electronic theses and dissertations: Differences in behavior for local and non-local users", *Library Hi Tech,* Vol. 32, No. 2, pp. 285-299.

Craver, Thom, (2013), “Goodbye, keyword data: Google moves entirely to secure search”, *Search Engine Watch*, available at: http://searchenginewatch.com/article/2296351/Goodbye-Keyword-Data-Google-Moves-Entirely-to-Secure-Search (accessed 20 December 2013).

Google, (2013), “Get the power of Google Analytics”, available at: http://www.google.com/analytics/premium/features.html (accessed 12 January 2013).

Herrara, G., (2011), “Google Scholar users and user behaviors: An exploratory study”, *College and Research Libraries*, Vol. 72, No. 4, pp. 316-331.

Ismail, M. A., and Kareem, S. A., (2011), “Identifying how novice researchers search, locate, choose and use web resources at the early stage of research”, *Malaysian Journal of Library & Information Science*, Vol. 16, No. 3, pp. 67-85.

Jansen, B. J., (2006), “Search log analysis: What it is, what’s been done, how to do it”, *Library and Information Science Research*, Vol. 28, No. 3, pp. 407-432.

Jansen, B. J., Booth, D. L., and Spink, A., (2008), “Determining the informational, navigational, and transactional intent of Web queries”, *Information Processing & Management*, Vol. 44, No. 3, pp. 1251-1266.

Lee, W. M., and Sanderson, M., (2010), “Analyzing URL queries”, *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 11, pp. 2300-2310.

Lee-Smeltzer, J., and Hackleman, D. (1995), “Access to OSU theses and dissertations in Kerr Library: How are they used … or are they?”, *Technical Services Quarterly*, Vol. 12, No. 4, pp. 25-37.

Mahoui, M., and Cunningham, S. J., (2001), “Search behavior in a research oriented digital library”, *Computer Science Technical Reports*, Paper 1505, available at: http://docs.lib.purdue.edu/cstech/1505 (accessed 24 May 2014).

McKay, D., (2007), “Institutional repositories and their ‘other’ users: Usability beyond authors”, *Ariadne*, No. 52, available at: http://www.ariadne.ac.uk/issue52/mckay (accessed 6 April 2013).

NDLTD: Networked Digital Library of Theses and Dissertations, (2012), available at: http://www.ndltd.org/ (accessed 25 May 2013).

Niu, X., Zhang, T., and Chen, H.-L., (2014), “Study of user search activities with two discovery tools at an academic library”, *International Journal of Human-Computer Interaction,* Vol. 30, No. 5, pp. 422-433.

Open Access Theses and Dissertations: OATD, (2013), available at: http://oatd.org/ (accessed 25 May 2013).

Ortiz-Cordova, A., and Jansen, B. J., (2012), “Classifying web search queries to identify high revenue generating customers”, *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 7, pp. 1426-1441.

ProQuest, (2014), “ProQuest dissertation publishing”, available at: http://www.proquest.com/en-US/products/brands/pl_umidp.shtml (accessed 30 January 2014).

Rose, D. E., and Levinson, D., (2004), “Understanding user goals in web search”, in *Proceedings of the 13th International Conference on the World Wide Web*, ACM Press, New York, pp. 13-19.

Slegg, Jennifer, (2014), “Bing begins rollout of secure search, say goodbye to more keyword data”, *Search Engine Watch*, available at: http://searchenginewatch.com/article/2322665/Bing-Begins-Rollout-of-Secure-Search-Say-Goodbye-to-More-Keyword-Data (accessed 17 February 2014).

Spink, A., Jansen, B. J., Wolfram, D., and Pedersen, J., (2002), “From e-sex to e-commerce: Web search changes”, *IEEE Computer Magazine*, Vol. 35, No. 3, pp. 107-109.

St. Jean, B., Rieh, S. Y., Yakel, E., and Markey, K., (2011), “Unheard voices: Institutional repository end-users”, *College & Research Libraries*, Vol. 72, No. 1, pp. 21-42.

Sullivan, Danny, (2014), “Yahoo makes secure search the default”, *Marketing Land*, available at: http://marketingland.com/yahoo-makes-secure-search-default-71308 (accessed 17 February 2014).

Teevan, J., Adar, E., Jones, R., and Potts, M. A. S., (2007), “Information re-retrieval: Repeat queries in Yahoo’s logs”, in *Proceedings of the 30th annual international ACM SIGIR conference*, ACM Press, New York, pp. 151-158.

Tenopir, C., and Rowlands, I., (2007), “Information behaviour of the researcher of the future: Age-related information behaviour”, available at: http://www.jisc.ac.uk/media/documents/programmes/reppres/ggworkpackageiii.pdf (accessed on 11 May 2013).

Tolle, J., (1983), “Transactional log analysis: Online catalogs”, in Kuehn, J.J., (ed.), *Proceedings of the 6th Annual International ACM SIGIR conference on Research and development in information retrieval, SIGIR ‘83*, ACM Press, New York, pp. 147-160.

Virginia Tech University Libraries, (2014), “Electronic theses and dissertations at VT: Electronic theses and dissertations online access data”, available at: http://scholar.lib.vt.edu/theses/data/somefacts.html (accessed 29 January 2014).

Waller, V., (2011), “Not just information: Who searches for what on the search engine Google?”, *Journal of the American Society for Information Science and Technology*, Vol. 62, No. 4, pp. 761-775.

Wang, P., Berry, M. W., and Yang, Y., (2003), “Mining longitudinal Web queries: Trends and patterns”, *Journal of the American Society for Information Science and Technology*, Vol. 54, No. 8, pp. 743-758.

Wikipedia, (2013), “Google Analytics”, available at http://en.wikipedia.org/wiki/Google_Analytics (accessed 12 January 2013).

Wikipedia, (2014), “Google Chrome”, available at http://en.wikipedia.org/wiki/Google_Chrome (accessed 17 May 2014).

Yiotis, K., (2008), “Electronic theses and dissertation (ETD) repositories: What are they? Where do they come from? How do they work?”, *OCLC Systems & Services: International Digital Library Perspectives*, Vol. 24, No. 2, pp. 101-115.

Zhang, Y., Lee, K., and You, B.-J., (2001), “Usage patterns of an electronic thesis and dissertations system”, *Online Information Review*, Vol. 25, No. 6, pp. 370-377.