Finding information in books: characteristics of full-text searches in a collection of 10 million books

Craig Willis , Miles Efron
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St., Champaign, IL 61820
{willis8, mefron}@illinois.edu

ABSTRACT

Searching large collections of digitized books is a relatively new area in information-seeking and retrieval research, made possible by initiatives such as Google Books and the HathiTrust Digital Library. The availability of large full-text book collections is transforming how users search and interact with information in books, but the characteristics of these changes are unknown. This paper aims to provide insight into the characteristics of full-text searches in a large collection of digitized books and is the first step in a broader research agenda intended to improve book retrieval. To better understand the types of queries that users are issuing to full-text-book collections, we analyzed a full year of anonymized query logs from the HathiTrust Digital Library full-text search engine. We also manually classified a random sample of 600 queries to develop a taxonomy of book search query types. We found that users are beginning to search *for information in books* instead of searching *for books*. Searches still largely follow bibliographic models, but, as expected, new types of searches are beginning to take advantage of full-text capabilities. Additionally, comparing the results of our query log analysis to searches in other domains, we found similar search patterns including short queries, sessions with only a few queries, and users viewing only a few pages of results per query. We discuss how these findings can be used to characterize users of large full-text book collections.

Author Keywords

Information retrieval, information behavior, book search, digital libraries, digitized books

INTRODUCTION

What does it mean to search books today? Increasing numbers of books are becoming available digitally through electronic publishing or large scale digitization efforts, such as Google Books[1], the Internet Archive[2], and the HathiTrust[3]. It is estimated that over 20 million volumes have been digitized and continue to be digitized at a rate of over 1,000 volumes per day (Howard, 2012; Internet Archive, 2013). The resulting full-text collections present new ways for users to search for and interact with information in books and new challenges with respect to the organization, representation and retrieval of the information they contain.

The HathiTrust Digital Library (HTDL), a partnership of over 70 research institutions and libraries, offers users the ability to search the full-text of over 10 million volumes in a single index, making it one of the largest research collections in the world (York, 2012). The collection presents a unique opportunity to further our understanding of what users are searching for in these large full-text book collections.

While a number of previous studies have explored user search behaviors in digital libraries (Kim, Feild, & Cartright, 2012), library catalogs (Slone, 2000), and a variety of full-text environments (Morris, Teevan, & Panovich, 2008; Teevan, Ramage, & Morris, 2011), as of yet no work has looked into specific behaviors of users searching full-text book collections. To date, information retrieval research in this area has focused more on techniques for retrieval in book collections (Kazai & Landoni, 2012; Koolen et al., 2012), not user behavior.

Through an analysis of query logs provided by HTDL, we explore two questions related to book search:

1. What are the types of queries that users issue to full-text book collections?

2. What are the characteristics of full-text book searches as compared to other domains?

Answering these questions is the first step in addressing the broader research question: what does it mean to do book search well?

AIMS AND MOTIVATIONS

In 2008, Hearst, Hurst and Dumais asked *what should blog search look like*? Today, we face a new but similar question: *what should book search look like*? Book search has long

[1] `http://books.google.com/`
[2] `http://archive.org`
[3] `http://www.hathitrust.org/`

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

been studied in the domain of bibliographic theory and library catalogs, which concerned primarily with searching *for* books. Full-text environments now offer users the ability to search *for information* in books.

As a preliminary contribution, in this paper we ask: *what does book search look like in its current form?* In asking this question we are motivated by the larger and more fundamental problem of what it means to support book search well. We hypothesize that the expectations and needs that people bring to book collections differ from both traditional bibliographic search and those needs that have been well-studied in the literature of *ad hoc* information retrieval.

Because corpora of the scale and breadth of these collections are new, we find ourselves at a valuable moment. The expectations and strategies of users are in the early stages of development. As the community designs information retrieval and discovery systems for book corpora, we are in a position that is exciting and pivotal to book search success. The tools we build will shape users' ideas of when and how book corpora are useful. But we also risk failing to meet users' needs. Without a sustained analysis of these needs, the risk of failure is high.

For example, assumptions that guide fundamental results in IR may not suit digitized book repositories. Consider a person who is interested in the use over time of the proverb *a drowning man grasps at straws*. Before a retrieval can support this query, we must address at least four questions: (1) What is the proper unit of retrieval? (2) What objective function should an IR system optimize during search? (3) What evidence suggests that a document is useful with respect to a given query? (4) What should the system present to the user as results?

Answers to these questions are not obvious. In traditional catalog-based book search, the unit of retrieval is typically the book. Queries to an OPAC usually result in a set or ranking of matching books. But data in HTDL take the form of individual page scans. Thus a logical indexing unit is the page, not the volume. A phrase query such as this may be well-served by page-level results. Yet a topical query such as *world war ii industrialization* may require book-level results, or even higher levels of abstraction, such as the work or expression.

Question 2 addresses the notion of relevance. Though it has become more multidimensional recently, a great deal of IR research simplifies the retrieval problem by using topical relevance as a proxy for document utility. The probability ranking principle gives an objective function that is optimized by a particular retrieval strategy (Robertson, 1977). While topical relevance is a sensible thing to aim for in response to *world war ii industrialization*, the searcher who submits the query *a drowning man grasps at straws* is likely to be interested in exemplars of this saying, not necessarily documents "about" it.

Question 3 is closely related to Question 2. Consider the common query likelihood IR method (Ponte & Croft, 1998; Zhai & Lafferty, 2004), which ranks documents against a query Q on:

$$\begin{aligned} P(D|Q) &\propto P(Q|D)P(D) \\ &\propto \prod_{i=1}^{|Q|} P(q_i|D)P(D) \end{aligned} \quad (1)$$

where $P(q_i|D)$ is typically estimated by smoothing the maximum likelihood estimator with a reference collection C via a smoothing parameter λ:

$$\hat{P}(q_i|D) = (1-\lambda)\frac{c(q_i,D)}{c(D)} + \lambda\frac{c(q_i,C)}{c(C)} \quad (2)$$

where $c(.)$ is a counting function. With respect to Question 3, the usefulness of Eq. 2 and thus of Eq. 1 is problematic for book search. Eq. 2 rewards documents that contain many instances of query terms. This is sensible in many settings, and the intuition is shared by all major IR models. But in book search, it is not necessarily the case that more query-document matches suggests higher document utility. For a searcher interested in *a drowning man grasps at straws*, *many* query terms are less likely to indicate document quality than are *properly ordered* query terms. In fact, Efron found that documents in digitized book collections with extremely high numbers of query terms tended not to be useful, often consisting of scans of book indexes, tables of contents, etc. (Efron, 2013)

Finally, Question 4 treats the perennial problem of result presentation. For *a drowning man grasps at straws* or *world war ii industrialization*, a ranked list of documents (of every level of granularity) may be appropriate. But other query classes present a different scenario. For example, a searcher who types *"as it pertains to"* may need to perform post-search analysis to make his or her results useful. Instead of a ranked list of documents, a query such as this might be better supported by supplying tools such as the Google ngram viewer[4] to allow searchers to contextualize their retrieved data.

RELATED WORK

A important step in understanding book search in full-text collections is to explore the range and types of searches carried out by users. There are four primary areas of work related to this topic: analysis of searching in libraries and library catalogs, studies of user search behavior on the web, studies on techniques for query classification, and book retrieval research.

Book search behavior in libraries and library catalogs have been discussed in the library science literature for well over a century. Research has focused on the design of library catalogs and metadata schemes, conceptual models of bibliographic information (Tillett, 1991, 1992; Suar, 1998), models of research in library collections (Mann, 1993), and types of searches in libraries and related environments (Buckland, 1979; Specht, 1980; Larson, 1991; Kilgour, Moran, & Barden, 1999; Jones, Cunningham, McNab, & Boddie, 2000; Slone, 2000; Lee, Renear, & Smith, 2007).

[4] `http://books.google.com/ngrams/`

More recent research has focused on the characteristics and behaviors of users searching online library catalogs and digital libraries. Slone studied user search behavior on online public library catalogs (2000). Kim *et al.* (2012) compared user book search behavior between web search engines and an online library catalog.

These studies join a growing body of research exploring search behavior in a variety of specialized environments, including web search engines (Spink & Wolfram, 2001; Broder, 2002; White & Morris, 2007), blogs (Hearst, Hurst, & Dumais, 2008), social networks (Morris et al., 2008), and microblogs (Teevan et al., 2011). Each of these environments has proven to present unique challenges for information retrieval research.

In many of these environments query intent classification has been demonstrated to improve overall retrieval effectiveness. Query classification techniques can be used to assign queries to categories based on topical information (Beitzel, Jensen, Chowdhury, & Frieder, 2007; Broder et al., 2007), click-graphs (Li, Wang, & Acero, 2008), or the form of the query (Kang & Kim, 2003). This is particularly useful in environments with multiple *verticals* or domain-specific subcollections (e.g., news, images, videos) (Arguello, Diaz, Callan, & Crespo, 2009). Query intent classification may also prove useful in book search, as users search for different types of information (e.g., author, title, subject).

With the availability of large collections of full-text books, book retrieval research emerged as an area of interest in the information retrieval community. Since 2007, the annual conference for the INitiative for the Evaluation of XML Retrieval (INEX) and the International Conference on Information and Knowledge Management (CIKM) have included tracks focused on techniques for retrieval in book collections (Kazai & Landoni, 2012; Koolen et al., 2012). Researchers have explored techniques for focused retrieval in a collection of 50,000 books, techniques for finding pages that confirm/refute factual statements, and the effect of social information on book recommendation.

The current study contributes to each of these areas of research through the examination of user search behavior in a very large collection of scanned books.

APPROACHES TO STUDYING SEARCH BEHAVIOR

Studies of user search behavior employ a variety of techniques, including surveys, questionnaires, interviews, and laboratory studies. Another common approach, and the one used in this paper, is query log analysis. Jansen discusses many of the benefits of query log analysis (Jansen, Zhang, Sobel, & Chowdury, 2009). The method is unobtrusive, requiring no response from participants. Data is collected in a natural context as users interact directly with the system. Query logs also offer more data over longer periods of time than can be reasonably collected using other approaches.

Jansen also notes a number of weaknesses with this approach. The data are not collected to answer specific research questions. It is therefore difficult to link fields in log files to the concepts being studied. Additionally, query logs are often captured in the context of production systems and are prone to errors caused by changes or failures in associated software.

Other techniques such as surveys, interviews, and laboratory studies offer dimensions that are not possible to study using query logs alone. In this study, we combine query log analysis supplemented by the manual classification of queries to explore the range and types of searches issued by users in full-text book collections.

DATA AND METHODS

In this section, we describe the data set, preprocessing steps, and methods used in this study. We begin with a description of HTDL, which is the source of the query logs. We then describe the query logs, data preprocessing steps, and methods of analysis.

HathiTrust Digital Library (HTDL)

HathiTrust is a partnership of more than 70 major research institutions and libraries. The HathiTrust Digital Library (HTDL), one of the world's largest research collections, is comprised of OCR full-text and bibliographic metadata for over 10 million volumes representing over 5 million unique titles spanning 5 centuries (York, 2012). The majority of volumes in the collection come from the Google Books project, with additional content digitized by partner institutions.

HTDL supports three types of searches: traditional catalog search, full-text search, and search-within-the-book. Catalog search relies solely on bibliographic metadata (e.g., author, title, subject, ISBN, etc). Full-text search supports a combination of OCR and bibliographic metadata. Search-within-the book relies on individual volume full-text. Users can view and navigate the full content of out-of-copyright texts. Due to copyright restrictions, users can search within in-copyright texts, but can only view snippets and are unable to access the full content.

Full-text search in HTDL is implemented using the Apache Solr search engine (Burton-West, 2012). Users can search using Boolean operators, quotes, and the plus and minus operators. Due to performance limitations, other Lucene search features, such as truncation and wildcards, have been disabled. Using the advanced search interface, users can search combinations of metadata and OCR.

Users searching the HTDL catalog and full-text search are presented with different user interfaces. By default, full-text search interface displays 25 results per page. Each result includes basic bibliographic information (title, author, date) and links to both the catalog record and the full-text view.

Full-text search results can be refined using facets derived from the bibliographic metadata. Facets include author, subject headings, language, country of publication, date of publication, format, and the source location of the digitized text. Users can also select whether to include all results or only those works supporting full-text view (i.e., public domain).

Burton-West describes the two-tier indexing architecture employed in HTDL full-text search (Burton-West, 2012). The first tier indexes all 10 million volumes along with metadata

Field	Description
Host	Anonymized remote user IP address
Date time	Request date/time
URL	Requested URL
Status	HTTP status code
Bytes	Response size in bytes
Referer	Referring URL
User agent	Client user agent

Table 1: W3C common log format.

Field	Description
IP Address	Anonymized remote user IP address
Session ID	Anonymized user session ID
Time	Request time
Qtime	Amount of time taken to process query
Numfound	Number of results returned for query
Query	Lucene query (EDISMAX format)
CGI-URL	Web server request URL

Table 2: HTDL Solr log format.

at the individual book level. The second tier, used for search-within-the-book, indexes volumes at the page level, but only on-demand. For the full-text search the unit of retrieval is the book or volume. For search-within-the-book, the unit of retrieval is the page.

The demographics of HTDL users are currently unknown. Although HTDL service is publicly accessible, users are likely from one of the 70 partner institutions. In this case, HTDL differs from other publicly-oriented book search services, as the user base is likely predominantly academic users in the United States.

As a first step, for the purpose of this study, only queries to the full-text index are considered. Future work will also include search-within-the-book queries.

Query logs

The data used for this study consists of one year of anonymized web server and Solr query logs provided by HTDL. The query logs cover the period from January through December 2012 and contain over 6 million requests representing over 1.1 million users queries.

The log files were anonymized by HathiTrust staff by replacing the IP address and session ID with random hashes. Requests are load balanced across multiple servers and random seeds used for anonymization are server and date specific. This means that the same value (e.g., IP address) will have a different hash value depending on the date and server handling the request.

The web server logs are in the W3C common log format. Log fields are shown in Table 1 and include the anonymized remote IP address, date and time of request, and requested URL, among others. The Solr logs are shown in Table 2 and include the anonymized remote IP address, anonymized session ID, the time of the request, the amount of time taken to process the query (qtime), the number of results returned (numfound), and the query in the Solr Extended DisMax format. Beginning in November 2012, log entries also include the web server request URL to facilitate cross-referencing of Solr log entries with web server log entries.

The webserver logs contain several different request types, identifiable only by information in the request URL. Two request types are used in this study: full-text queries (identified by request path “cgi/ls") and full-text views (identified by path “cgi/pt"). Because full-text searches are executed first through the webserver, there is a one-to-one relationship between searches in the web server log and searches in the Solr logs. Full-text views are only captured in the web server logs.

Although the web server and Solr logs contain much of the same information about individual query requests, the web server logs are required for identifying associated full-text view requests. The Solr logs contain two unique pieces of information used in this analysis: the number of records returned for a query and user session ID. Additionally, the web server logs contain requests for full-text views that originate from HTDL catalog search and external sources (e.g., OCLC WorldCat). These records are not used in the current analysis.

Data Pre-processing

This section describes the preprocessing steps taken to prepare the web server and Solr log files for analysis.

A custom program was developed to parse the web server and Solr query logs, extract full-text query and full-text view requests, and merge attributes from both sources. Full-text queries were extracted from both the web server and Solr logs and merged based on a combination of query attributes and request timestamps. Full-text view requests were extracted from the web server log only. The referring URL field was used to associate full-text view requests with each query request.

Although the web server and Solr logs contain information about every request, individual requests do not equate to unique queries. For example, each time a user pages through the results displayed in HTDL interface, a record is logged that contains the current query information with a single parameter indicating the current page of results. The combined log entries were preprocessed to identify and aggregate information based on three different concepts: sessions, queries, and full-text click-throughs.

The session represents a unique user session. For the purpose of this study, sessions are identified using the anonymized Solr Session ID. Each session can contain one or more queries and full-text views. Queries are defined as a unique combination of query terms, operators, facets, and limiters within a session. A query is considered to be a new query if the user modifies the query terms, adds/removes facets, or toggles the full-text limiter. Each query is associated with one or more pages of results viewed by the user. Full-text views are defined as the initial click-through from the query results page to the full-text viewer application. Each query can have zero or more full-text views.

In the next section, we discuss the methods used to analyze data derived from the processed logs.

METHODS

The central goal of this study is to better understand what users are looking for and how they formulate searches when searching full-text book collections such as HTDL. We used two different methods to address these questions: manual query classification and query log analysis.

For query classification, we used a simple card-sorting technique to categorize a random sample of 600 queries from HTDL full-text search logs. In the first phase, queries were sorted into logical groupings. After the groupings were created, category names were assigned to each group. To validate the categories, two information professionals were asked to code a separate set of 100 queries using the pre-defined categories with inter-rater agreement measured using Cohen's Kappa. Queries were assigned to a single category based on the interpreted intent of the user. Since the actual intent of the user cannot be determined, participants were asked to interpret the most likely intent and were allowed to use any external sources of information. Participants were also asked to identify whether the query would most likely be satisfied by the book or page as the unit of retrieval.

As described in the Data Pre-processing, for the query log analysis a custom program was developed to parse the web server and Solr query logs and extract full-text query and full-text view information. Information derived from the logs files were aggregated at the session and query levels for analysis in the R statistical package [5].

Since users can modify existing queries or issue entirely new queries within the same session, query attributes can be used to detect transitions within a single session. A custom program was developed to identify these transitions in the aggregated query logs.

RESULTS

In the following sections, we report the results of the manual query classification process and query log analysis.

Types of book searches

As described in the Methods section, card-sorting was used to classify a random sample of 600 queries from HTDL query logs. Queries were grouped together and category labels determined after the sorting was complete. To validate the categories, two participants classified a separate sample of 100 queries (Cohen's $\kappa = 0.661$).

Six broad categories of full-text book searches were identified, as listed in Table 3. Users of HTDL searched for book titles; people and organization names; topics and keyword phrases; locations and place names; unique book identifiers; and quotations.

Title searches are primarily multi-term queries, sometimes as quoted phrases (e.g., "'rubaiyat of omar khayyam'"), complete titles including capitalization (e.g., "The Desperate Outlook in Macedonia"), or title words with qualifiers, including dates and/or parts of author names (e.g., "essay of a delaware english spelling book 1806 by zeisberger"). Full-text search enables users to also search for books that mention or reference the searched title. Additionally, full-text book collections such as the HTDL may include duplicates of the same title, either as duplicate scans of the same manifestation or multiple manifestations of the same expression or work, as defined in the Functional Requirements for Bibliographic Records (FRBR) (Suar, 1998). The question remains open as to whether full-text book search needs to account for user needs at these higher levels of abstraction. This is discussed further in the discussion section.

[5] http://www.r-project.org/

Category	Examples
Titles	1859 "rubaiyat of omar khayyam" It happen'd on a Whitsunday psychology in action ninth edition "marcia militare" schubert
People Organizations	Alexander, George L. "american car co" "George Dorris" and "St. Louis" "bienville parish" reese 1784 puccini "Marguerite Clark" and "drawn by"
Topic	fracking "steam engine" electric railroads ohio
Locations Place names	"duluth Minnestota" colombia 1902 "nassau boulevard"
Book identifier	0083-3401
Quotation	"the ends justify the means" "30 seconds per week"

Table 3: Proposed query categories and example searches.

Searches for the names of people or organizations include simple one-term queries (e.g. mozart), multi-term queries, quoted phrases, and authority controlled names (e.g., "United States. Congress. House."). Name searches are sometimes qualified by locations (e.g., "'James Steele' Illinois"), dates (e.g., "1784 puccini"), or topical keywords (e.g., "bodansky who killed zia"). In full-text search without pre-selected metadata fields specified, it is unclear whether the user is searching for books written by, about, or mentioning the name in the query. Many of the queries analyzed in the sample appear to be genealogical, as users are searching for family names often qualified with specific locations.

Locations and place name queries refer to entities that are primarily geographic. Location names in the sample were generally free text, but sometimes enclosed in quotes. Location searches sometimes include qualifiers, such as dates, but are more often used as qualifiers for other query types, such as personal or organization name searches.

Quotation searches are a special type of quoted phrase search, where the user is looking for a specific quotation or passage in a book. These queries are interesting because they represent a way to use of information contained in books. The

Full-text views	4,250,843
Full-text searches	1,096,910
Search-within-the-book	947,423
Collection builder	427,847
Other	33,841
Total	6,756,864

Table 4: Summary of web server log requests.

Sessions	220,088
Queries	690,386
Click-throughs	529,457

Table 5: Summary of sessions, queries, and click-throughs extracted from the query logs.

user may be searching for a specific book where a quote or passage is known, but author and title are not. Users might also be searching for the first use of a particular phrase or a distribution of publication dates for books in which the quotation appears. Information of this sort is the basis for the Google ngram viewer (Michel et al., 2011).

Topical queries are also referred to as subject searches (Buckland, 1979) or informational searches (Broder, 2002). These queries do not fall into any of the other identified categories. With further refinement of the query categories, it is likely that new categories can be derived from topical queries, such as events or other concepts.

In this section we've reviewed the types of queries identified through the manual query classification. In the next section, we report the results of the analysis of HTDL query logs.

Query log analysis

In this section, we report the results of our analysis of HTDL query logs. As described in the Methods section above, query logs were analyzed at three levels: session, query, and full-text view or click-through. We first report general information about the log files, followed by sections on session, query, and click-through characteristics.

The web server logs provided by HTDL contain information for over 6 million requests from January through December 2012. Request types included in the logs are shown in Table 4 and include full-text views, full-text searches, search-within-the-book, collection builder, and other (e.g., authentication) requests. It is worth noting that over 71% of the full-text view requests originated from catalog searches, which are not included in our analysis.

The Solr query logs contain information for over 1.1 million full-text searches. Due to missing information in both the web server and Solr log files, we were only able to find matching information for 934,482 records between both sources. These combined records serve as the basis for the session and query analysis results reported in the following sections. Table 5 shows the number of extracted records representing sessions, queries, and click-throughs from the combined log files.

What are users looking for?

HathiTrust	Web	Twitter
biographic register	twitter	new moon
foreign service list	youtube	#knowyouruglyif
ottoman	facebook	justin beiber
cotton	google	adam lambert
geschlechterbuch	myspace	#theresway2many
rauchenecker	youtub com	taylor swift
монеты (coins)	yahoo	lady gaga
divan	ebay	modern warfare 2

Table 6: Top queries in HTDL compared to Web and Twitter.

1-grams	2-grams	3-grams
john	new york	new york city
new	united states	new york state
history	annual report	world war ii
william	world war	michigan historical collections
american	new jersey	american medical association
county	civil war	fonte dos amores
war	north carolina	first world war
james	official gazette	edgar allan poe
report	eirserne kreuz	federal energy regulatory
george	south carolina	interstate commerce commission

Table 7: Top 1-, 2-, and 3-word phrases in HTDL query logs.

We examine the top queries and phrases contained in queries to get a better understanding of the types of information that users are looking for in HTDL. Table 6 shows the top 10 queries[6] across unique sessions in the query logs compared with similar information from web and Twitter searches as reported by Teevan *et al.* (2011). In this case, the queries are only the query strings, not including facets, full-text limiter, or other operators.

The top queries in HTDL include title and topical queries. Similar to the navigational queries of web search, title queries are generally satisfied by a single result. In this case, HTDL contains 75 editions of the *Biographic Register*, a publication of the U.S. Department of State containing information about foreign affairs officials. Similarly, HTDL contains over 50 editions of the *Foreign Service List*, another source of historical information about U.S. foreign affairs personnel, and over 160 volumes of the *Deutsches Geschlechterbuch*, a popular source of information for German genealogy. Would the users' queries be better satisfied with a single result representing the multi-volume publication or with a single result per volume?

Table 7 shows the top 1-, 2-, and 3-word phrases from unique query strings in the logs. Common stopwords have been removed and terms converted to lower case for comparison. These terms suggest a focus on personal names –

[6]Due to complications caused by dynamic IP addresses, there is some difficulty determining whether some queries are from one or many users.

Characteristic	Mean	Med.	Min	Max	SD
Queries per session	2.84	2	1	12	2.48
Pages of results viewed	4.06	2	1	20	4.08
Click-throughs	1.84	1	0	15	2.97

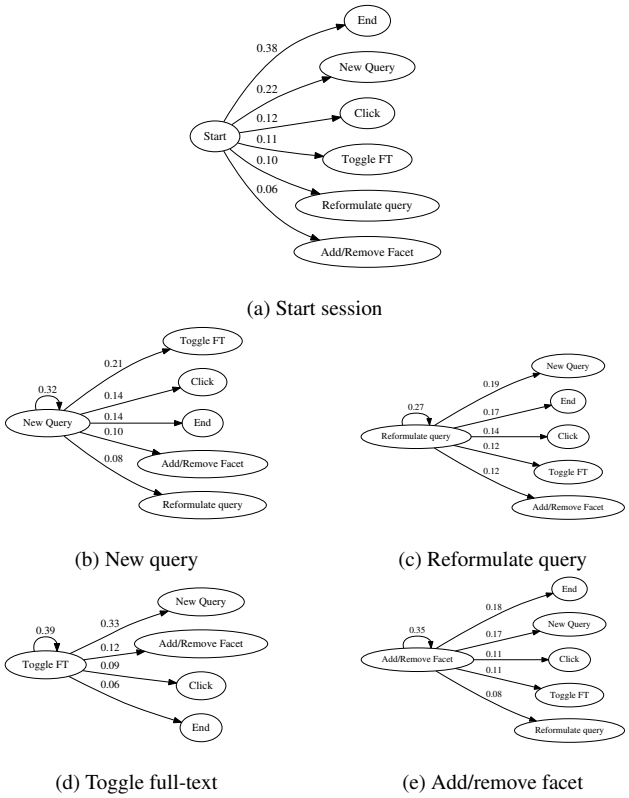
Table 8: Session characteristics.

both political and literary – as well as organization names, geographic locations, historical events, as well as literary concepts. Considering the common English names “john," “william," “james," and “george" that appear in the top one-word phrases, the most common 2-word phrases in queries containing these names include “john henry", “john smith", “william henry", “william county", “henry james", “james joyce", “george washington", and “george miller." “Eiserne kreuz" translates to the English “iron cross" and “Fonte dos Amores" from the Portuguese for “fountain of love."

Session characteristics

In this section, we report the results of our analysis of user sessions in the query logs. As noted above, sessions are identified by an anonymized session ID in the Solr log files. Information about queries, pages of results viewed, and click-throughs are aggregated at the session level and reported here.

As shown in Table 8, half of all sessions contain no more than two queries, with an average of 2.8 unique queries and one 1.8 click-throughs per session. More than half of users view two or more pages of results in a single session.



(a) Start session

(b) New query

(c) Reformulate query

(d) Toggle full-text

(e) Add/remove facet

Figure 1: Search session transition probabilities.

Characteristic	Mean	Med	Min	Max	SD
Terms per query	2.8	2.0	1	7	1.5
Pages of results viewed	2.7	1.0	1	3	0.5
Click-throughs	2.4	0	0	3	0.7

Queries with facets	0.21
Queries with quoted phrases	0.23
Queries with Boolean operators	0.08

Table 9: Query characteristics.

After the initial query, users can take one of six options in HTDL: 1) abandon the query and end their session; 2) reformulate the query through addition/removal of terms, quotes, or operators; 3) add or remove facets; 4) toggle the full-text limiter; 5) enter a new query with no terms in common with the previous query; or 6) click-through to the full-text of a result. Figure 1 shows the transition probabilities between different query options within a session. After starting a query, 38% of sessions are ended, 22% initiate a new query, 12% click through to the full-text record, 11% toggle the full-text limiter, 10% reformulate the existing query, and 6% add a facet.

Query characteristics

Here we present the results of our query-level analysis. We first report overall query statistics.

For this study, queries are defined as a unique combination of terms, operators, facets, and limiters in a session. Sessions may contain multiple queries. Table 9 shows the mean and median values for key query characteristics including the number of terms per query, the number of pages of results viewed, and the number of full-text click throughs. Users on average view 2.71 pages of results and view 2.41 full-text results for a single query.

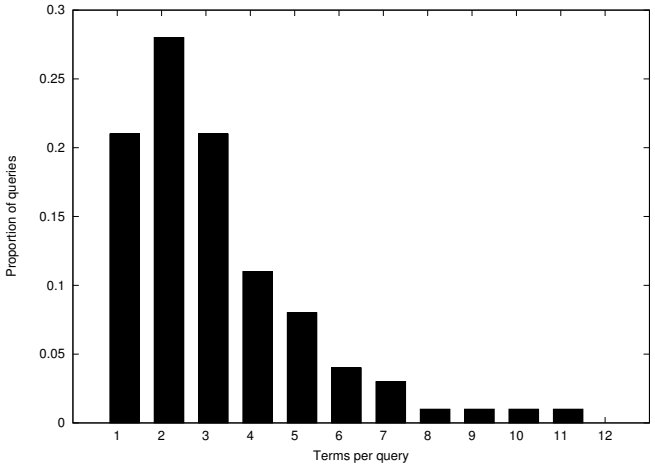


Figure 2: Terms per query.

Figure 2 shows the distribution of query lengths. Half of all queries contain only one or two terms, with an average of 2.81 terms per query.

Facet	Queries
No facet	79%
Date	9.97%
Language	6.90%
Topic	4.89%
Format	2.96%
Country	2.03%
Author	1.68%
Source	0.22%

Table 10: Facet use in queries.

HTDL supports six different facets to help users constrain the result space. Facets were used in 21% of queries analyzed. Details of facet use are shown in Table 10. Date was the most used facet, present in 9.97% of all queries. The language facet was used in 6.90%, topic (or subject) in 4.89%, format in 2.96%, and country in 1.68% of queries. The author facet was used in only 1.68% of queries.

Full-text characteristics

Of the over 3.3 million requests to view the full-text of books in HTDL, only 17% originated from the the full-text search results, with over 65% from the catalog search interface. Table 11 summaries the number of distinct volumes and queries associated with full-text views based on full-text searches. Of the over 500,000 unique volumes viewed from full-text search results, less than 20% were associated with more than one distinct query. If the over 350,000 unique full-text queries, just over 30% viewed more than one volume.

Distinct volumes viewed from full-text searches	554,403
Volumes viewed for more than one query	104,570 (18.86%)
Distinct full-text queries	352,080
Queries with more than one volume viewed	109,471 (31.10%)

Table 11: Fulltext requests.

Discussion

The term "book" conceals a great deal of complexity. The information in books is nearly is heterogenous as information on the web. Books are simply storage and presentation tools, capable of supporting a wide variety of information structures, access methodologies, and purposes (Carden, 2008). One essential distinction is that we often talk about searching *for books* while searching *for information* on the web. Massive collections of digitized books now enable us to apply techniques from decades of research in information seeking and retrieval to better enable searching *for information* in book collections.

The results of this analysis suggest that book search is similar in many ways to retrieval in other full-text environments, such as the web. However, searches still largely fit the traditional bibliographic mode. In this section, we discuss how these results can be used characterize full-text book searches.

Characterizing full text book searches

The results of this analysis suggest that book search is similar in many ways to full-text search in other environments. The session and query characteristics are generally consistent with those reported in other retrieval environments, as summarized by Markey (2007). Sessions are comprised of a small number of queries. Users view only a few pages of results per query and the majority of queries have three or fewer terms. Boolean searches are also uncommon.

There are, however, noteable differences. While most queries are shorter, a larger proportion of queries contain more than three terms (26%), which appears consistent with library catalog searches. This is likely due to title searches, which often contain more terms. More queries also contain phrases enclosed in quotes (23%) than in web search environments. Facets continue to be used to further constrain the search space. However, it would be interesting to further explore how user behavior might change if metadata-based facets were not available.

Query types and units of retrieval

Searches in HTDL remain largely bibliographic, as users still search for authors, titles, and subjects. Even though the HTDL offers a catalog-specific search interface, users are still likely to issue typical bibliographic searches to the full-text index.

The six query types identified in the query classification process were developed independent of the FRBR model, but are consistent with FRBR entities. Title and book identifier queries can be mapped to the FRBR Group 1 entities (*work*, *expression*, and *manifestation*). Person queries can be mapped to the FRBR Group 3 *person* and *corporate body* entities. Location and topic queries can be mapped to the FRBR Group 3 *concept*, *object*, *place* entities. Our query type taxonomy could be expanded to include the additional FRBR entities.

The FRBR model was developed based on a logical and empirical analysis of bibliographic records and is intended to address the specific needs of users of bibliographic systems. The model was not developed for use with full-text retrieval systems. It is therefore understandable that it does not account for user expectations and needs associated with full-text search. For example, searches for specific passages within books, which might be loosely mapped to the Group 2 *concept* entity; factoid searches; or searches for personal names that occur rarely in the texts, as with genealogical searches. The FRBR model also does not account for the new uses of information contained in books enabled by large-scale digitized collections, such as the Google ngrams project (Michel et al., 2011).

What do the FRBR Group 1 entities mean for the unit of retrieval for book search systems? To date, book search research has focused on two primary units of retrieval – the whole book and parts of books or pages. HTDL contains over 10 million volumes but only 5 million unique titles, indicating a fair amount of duplication. The applicability of Group 1 entities suggests that even full-text book search systems will

benefit from the process of identifying and collocating similar books (or parts of books). This could be achieved through grouping results based on the FRBR model, sometimes referred to as "FRBRization". This process has already been explored in large-scale bibliographic systems, including OCLC WorldCat (Hickey, O'Neill, & Toves, 2002; Bennett, Lavoie, & O'Neill, 2003). The book full-text may also be found to be useful in the FRBRization process, where bibliographic metadata alone is inadequate.

What is book search?

Teevan *et al.* characterize microblog search as *timely*, *social*, and *topical*, with a strong emphases on people and celebrities in particular. In their discussion of blog search, Hearst, Hurst, and Dumais emphasize the quality, style, and personality of blogs and also stress the importance of searching for people. Perhaps unsurprisingly, book search in HTDL is largely historical and literary. Users search primarily for titles, personal and organization names, and topical information including people and locations.

Searches for specific titles and book identifiers are often referred to as "known-item searches" or searches with high document specificity (Lee et al., 2007; Buckland, 1979). Full-text book search now enables users to search for known items or specific documents based on additional attributes, including specific passages in the text. Known-item searches have been widely discussed in the information retrieval literature, since these searches are usually satisfied by a single result.

Similarly, searches for people and organization names are no longer limited to information captured in metadata fields. Book search systems must distinguish between queries for books by, about, or mentioning a name. Given the query "Edgar Allan Poe", are users more likely searching for books by or about the author? Do the results change if the name is in quotes?

Many retrieval systems rely on external sources of information including query logs to improve overall retrieval effectiveness. The query click-graph, which captures the items selected by users in response to a particular query, has been applied in conjunction with learning-to-rank algorithms to improve results.

Limitations

This preliminary work is based on an analysis of anonymized query logs. The anonymization process is intended to protect the privacy of users, but also limits some of the information that can be derived from the logs. Despite these limitations, the results provide insight into a topic of growing importance and we hope can inform future work in this area.

Conclusion and future work

In this paper, we have presented an analysis of how users search for information in a large collection of digitized books using the full-text query logs from the HathiTrust Digital Library (HTDL). By analyzing a sample of queries, we developed a simple taxonomy of searches and demonstrated similarities and differences to traditional bibliographic search. Users of full-text book retrieval systems still search for traditional bibliographic entities, but the full-text capability introduces new types of searches including the ability to look for books based on quotations or passages. We concluded that full-text retrieval systems and evaluation environments might benefit from further consideration of the FRBR model. The units of retrieval found in practice, namely whole book or page, are likely inadequate, as certain queries would be better served by results at even higher levels of abstraction.

This study is intended as a first step in a broader research agenda intended to improve retrieval in large digitized book collections. We are current working to apply the results of this study to the development of a test collection intended to facilitate the evaluation of retrieval algorithms in the context of the HathiTrust collection. Future work will further supplement information derived from query logs with studies of actual users. We will also continue to explore automatic query classification techniques for book search.

ACKNOWLEDGEMENTS

We would like to thank Tom Burton-West and Jeremy York from the HathiTrust for their support and feedback on an early draft of this paper. This work was supported in part by the HathiTrust Research Center and campus bridge funding from the University of Illinois.

References

Arguello, J., Diaz, F., Callan, J., & Crespo, J. (2009). Sources of evidence for vertical selection. In *Proceedings of the 32nd international ACM conference on research and development in information retrieval (SIGIR '09)* (pp. 315–322).

Beitzel, S. M., Jensen, E. C., Chowdhury, A., & Frieder, O. (2007). Varying approaches to topical web query classification. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '07)* (pp. 783–784).

Bennett, R., Lavoie, B., & O'Neill, E. (2003). The concept of a work in WorldCat: an application of FRBR. *Library Collections, Acquisitions, and Technical Services*(Spring), 45–59.

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, *36*(2), 3–10.

Broder, A., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., & Zhang, T. (2007). Robust classification of rare queries using web knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '07)* (pp. 231–238).

Buckland, M. K. (1979, May). On types of search and the allocation of library resources. *Journal of the American Society for Information Science*, *30*(3), 143–147.

Burton-West, T. (2012). Practical Relevance Ranking for 10 Million Books. In *INEX 2012 workshop pre-proceedings.*

Carden, M. (2008). E-books are not books. In *Proceedings of the 2008 acm workshop on research advances in large digital book repositories.*

Efron, M. (2013). Query representation for cross-temporal information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information (SIGIR'13).*

Hearst, M. A., Hurst, M., & Dumais, S. T. (2008). What should blog search look like? In *Proceedings of the 2008 ACM workshop on search in social media (SSM'08* (pp. 95–98).

Hickey, T., O'Neill, E., & Toves, J. (2002). Experiments with the IFLA functional requirements for bibliographic records (FRBR). *D-Lib Magazine*, *8*(9).

Howard, J. (2012, March). Google Begins to Scale Back Its Scanning of Books From University Libraries. *Chronicle of Higher Education*.

Internet Archive. (2013). *Scanning Services.* Retrieved 5/24/2013, from `http://archive.org/scanning`

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Micro-blogging as online word of mouth branding. In *Proc. of the 27th international conference extended abstracts on human factors in computing systems (CHI EA'09)* (pp. 3859–3864).

Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. (2000). Human-computer interaction for digital libraries A transaction log analysis of a digital library. *International Journal on Digital Libraries*, *3*, 152–169.

Kang, I.-H., & Kim, G. (2003). Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval (SIGIR '03)* (pp. 64–71).

Kazai, G., & Landoni, M. (2012). BooksOnline'12: 5th Workshop on Online Books, Complementary Social Media and their Impact. In *Proceedings of the 20th annual conference on information an knowledge management (CIKM'12)* (pp. 2764–2765).

Kilgour, F. G., Moran, B. B., & Barden, J. R. (1999). Retrieval effectiveness of surname-title-word searches for known items by Academic Library users. *Journal of the American Society for Information Science*, *50*(3), 265–270.

Kim, J., Feild, H., & Cartright, M.-A. (2012). Understanding Book Search Behavior on the Web. In *Proceedings of the 21st acm international conference on information and knowledge management (CIKM'12)* (pp. 744–753).

Koolen, M., Kazai, G., Kamps, J., Preminger, M., Doucet, A., & Landoni, M. (2012). Overview of the INEX 2012 social book search track. In *CLEF online working notes.*

Larson, R. (1991). The decline of subject searching: Long-term trends and patterns of index use in an online catalog. *Journal of the American Society fo Information Science*, *42*(3), 197–215.

Lee, J. H., Renear, A., & Smith, L. C. (2007). Known-Item Search: Variations on a Concept. In *Proceedings of the american society for information science and technology* (pp. 1–17).

Li, X., Wang, Y.-Y., & Acero, A. (2008). Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '08)* (pp. 339–346).

Mann, T. (1993). *Library Research Models*. New York: Oxford University Press.

Markey, K. (2007). Twenty-Five Years of End-User Searching, Part 1 : Research Findings. *Journal of the American Society for Information Science and Technology*, *58*(8), 1071–1081.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Aiden, E. L. (2011, January). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–82.

Morris, M. R., Teevan, J., & Panovich, K. (2008). What Do People Ask Their Social Networks , and Why? A Survey Study of Status Message Q & A Behavior. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'08)* (pp. 1739–1748).

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. *Research and Development in Information Retrieval*, 275–281.

Robertson, S. (1977). The probability ranking principle in ir. *Journal of documentation*, *33*(4), 294–304.

Slone, D. J. (2000). Encounters with the OPAC: On-line searching in public libraries. *Journal of the American Society for Information Science*, *51*(8), 757–773.

Specht, J. (1980, September). Patron use of an online circulation system in known-item searching. *Journal of the American Society for Information Science*, *31*(5), 335–346.

Spink, A., & Wolfram, D. (2001). Searching the web: The public and their queries. *Journal of the American Society of Information Science and Technology*, *52*(3), 226–234.

Suar, K. (Ed.). (1998). *Functional Requirements for Bibliographic Records*. Munich: K.G. Saur Verlag.

Teevan, J., Ramage, D., & Morris, M. R. (2011). # TwitterSearch : A Comparison of Microblog Search and Web Search. In *Proceedings of the fourth ACM international conference on web search and data mining (WSDM'11)* (pp. 35–44).

Tillett, B. B. (1991). A Taxonomy of Bibliographic Relationships. *Library Resources & Technical Services*, *35*(2), 150–158.

Tillett, B. B. (1992). Bibliographic Relationships: An Empirical Study of the LC Machine-Readable Records. *Library Resources & Technical Services*, *36*(2), 162–188.

White, R. W., & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '07)* (pp. 255–262).

York, J. (2012). HathiTrust: The Elephant in the Library. *Library Issues*, *32*(3).

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, *2*(2), 179–214.