Identifying Search Patterns in an Image-based Digital Library

Hyejung Han
University of Wisconsin-Milwaukee
hanh@uwm.edu

Dietmar Wolfram
University of Wisconsin-Milwaukee
dwolfram@uwm.edu

ABSTRACT

Three months of clickstream transaction log data for an image-based digital library were analyzed for patterns in user session behavior. After extensive log cleaning of the data files, k-means cluster analysis was applied using selected session-related statistics. The outcomes reveal largely uniform session behaviors (97.37%) consisting of a limited number of actions, with a higher number of queries than other session types, but little browsing of individual results. Given the specialized nature of the digital collections within the studied system, most searchers appear to engage in purposeful, directed searching. The remaining session clusters represent higher degrees of browsing over a longer time with fewer queries per session. The session behaviors for this environment are markedly different than those revealed by transaction log studies of other information retrieval environments.

Keywords
Digital libraries, Image searching, Transaction log analysis, User search behavior.

INTRODUCTION

The wider availability of transaction logs has made it possible to study how users interact with digital libraries and other forms of information retrieval (IR) system on a larger scale than is possible through direct observation. Transaction log analysis has been frequently adopted for research to identify users' behavior on information systems including web search engines and digital libraries. For example, Park, Lee and Bae (2005) analyzed a week-long transaction log data from NAVER, a Korean Web search engine and found that users engage in simple behaviors: they enter short queries with a few query terms, seldom use advanced features, and view few results' pages. Agosti, Crivellari and Di Nunzio (2012) reviewed about 80 studies with transaction log analysis over the past 10 years, emphasizing the distinction between research on digital libraries and that of general web sites. According to the authors, digital libraries differ from the Web because they are organized and managed by experts, they have specific user groups, and they have much more structured collections. Recent transaction log studies with digital libraries have highlighted unique characteristics of user behaviors when using specialized collections, such as a folktale database (Trieschnigg, Nguyen & Meder, 2013), a scholarly science digital library (Park & Lee, 2013) and a nanoscience and technology digital library (Shiri, 2011).

Similar to Web search studies, much of the research that has examined image retrieval, whether as user studies or transaction logs (e.g., Choi & Rasmussen, 2003; Jansen, 2008; Pu, 2008), has focused on query-related actions and not the full range of user search and browsing-related activities. This is also largely true of studies that have examined user interactions with digital library transaction logs (e.g., Jones, Cunningham, McNab, & Boddie, 2000); however, Sfakakis and Kapidakis (2002) did examine clickstream data, but only reported descriptive results of general search and browsing actions.

The present research examines the following research questions:

1) Are distinct types of session behavior observed for an image-based digital library based on transaction log data?
2) If so, what search and browsing features characterize session types?

METHOD

The present study relies on clickstream data from the digital collections of the University of Wisconsin-Milwaukee libraries. The collections, mounted using the CONTENTdm® content management software, represent dozens of digitized image collections totaling more than 50,000 images that include formal metadata. The collections may be searched by keywords or browsed by identifying collections and then using available entry points and provided subject terms. Three months of clickstream data (i.e. Web server logs of all system and user-initiated actions)

ASIST 2014,November 1-4, 2014, Seattle, WA, USA.

were collected from September to November 2013. The log entry fields containing vital content for each user action included: the requesting Internet Protocol (IP) address, date and time of the action, and the referring URL. Extensive cleaning of the data was needed to remove system-generated actions and duplicate user actions at the same time. User actions were embedded in the referring URL and included, query submission, requests for pages of results and specific page requests.

By analyzing individual URLs requested on the transaction log data and comparing them with the interface and the navigational scheme of the digital library, we identified the 11 most frequently used types of actions. A similar scheme was identified by Chen and Cooper (2001). What began as millions of clickstream data records was initially condensed to 1.3 million user actions. More than 50% of the remaining records were associated with two IP addresses. Because of the high likelihood that these addresses represent a shared entry point to the collections, with overlapping actions from concurrent sessions, records for these IP addresses were excluded. Search session boundaries were estimated using a temporal cutoff point based on the distribution of time intervals between adjacent actions for the same IP address submitted on the same day. The cutoff time for the dataset was 1585 seconds, which is longer than earlier studies of public search engine query logs (He, D., & Göker, 2000; Spink, Wolfram, Jansen, & Saracevic, 2001). Most identified sessions (73.9%) consisted of one or two actions. To provide more meaningful analysis of session-level behavior, only sessions consisting of three or more user actions were further analyzed. Sessions consisting of more than 1000 actions were also eliminated. A closer analysis of these very lengthy sessions revealed they were bot-generated and extended over many hours or across multiple days.

Once cleaned, 14 session characteristics for the dataset were extracted directly from the data or were derived based on combinations of observed data. Due to the potential for high correlations between some of the session characteristics, as observed by Chen and Cooper (2001), a factor analysis was run in SPSS v. 20 using principal component extraction and varimax rotation to identify groups of related session characteristics. The session characteristic with the highest loading value for each component was selected as the representative characteristic for that component. The session data for these representative characteristics were then used to conduct a k-means cluster analysis to identify groups of session behavior. K-means clustering has been used in transaction log analysis studies for web search behavior (Kathuria, Jansen, Hafernik & Spink, 2010; Stenmark, 2008; Belk, Papatheocharous, Germanakos, & Samaras, 2012) and digital library search behavior (Xu & Recker, 2012; Frias-Martinez, Chen, Macredie & Liu, 2007). According to Kathuria et al., the use of k-means as a classification technique yielded positive results and fared better than a binary tree classification algorithm. Moreover, Frias-Martinez et al. believed that the k-means clustering is popular because it is easy to understand user behavior of digital libraries and it's easy to implement.

RESULTS AND DISCUSSION

A summary of the characteristics of the cleaned dataset appears in Table 1. The 11 identified actions and their frequencies are summarized in Table 2.

Dataset Feature	Value
Total Sessions	15186
Total User Actions	325059

User Actions per Session	Mean = 21.4
	s.d. = 45.4
	Median = 8

Table 1. Summary table of dataset Features

Type of Action	Frequency	%
Single item (image/document) views	120007	36.9%
Page of results request	91801	28.2%
Simple search	69335	21.3%
Landing page	25652	7.9%
Advanced search	12613	3.9%
Printing/Viewing an image	2098	0.6%
Information page request	1391	0.4%
Metadata search	1274	0.4%
Video player (multimedia item)	358	0.1%
Assisted search	356	0.1%
Timeline (date information was available for some collections)	174	0.1%
Total Actions	325059	100.0%

Table 2. Typology of session actions

A total of 15186 sessions containing at least three actions were identified. The factor analysis identified six components that accounted for a total of 74% of the variance. The representative session characteristics with the highest loading values for each component appear in Table 3. Only two of the six characteristics represent observed values (Session Actions, Print/View). The remaining characteristics were derived. Note that the Print Views characteristic represents a more detailed examination of an individual image over a Single Item View and/or request for a print out of the image. Although it represents a relatively infrequent action, it was the singular high loading session characteristic within the sixth component.

Session Characteristic	How Calculated	Comp. Loading
Session Actions	The total number of actions for a session	.970
Proportion of Browsing Actions	The total actions related to browsing divided by the Session Actions	.823
Proportion Landing Page Visits	Total Landing Page Visits divided by the Session Actions	.831
Average Time Between Queries	Time between the first and last query action divided by the Queries per Session	.734
Average Single Item Views Per Query	Total Image Page Visits per session divided by Queries per Session	.796
Print/View	The number of Print /Views per session	.837

Table 3. Summary table of dataset features

The k-means cluster analysis was run with different numbers of clusters. In each case, the vast majority of sessions fell into one cluster. With four or more clusters there was very small membership for some clusters (1-4 sessions). Three clusters revealed distinct session characteristics (Table 4). More than 97% of the sessions fell into a cluster that can be characterized by comparatively few actions with the shortest time between queries and little browsing of individual results. The second cluster, representing 2.52% of sessions, consisted of mid-range values for the session characteristics, with searchers less likely to use landing pages, perhaps indicating focused searches within given collections. The last cluster, consisting of the smallest number of sessions (0.11%) contained the highest levels of browsing actions (except the proportion of landing page visits) and time taken between queries. Interestingly, the longest sessions consisting of the most actions fall primarily within the first two clusters, indicating that the number of session actions, although playing a part in characterizing user search behavior, is not singularly influential. Given the greater support for browsing in a digital library environment in comparison to other types of IR environments such as public search engines and bibliographic databases, the high level of browsing or viewing-related actions (viewing individual items, requesting a page of results, requests for information pages) across all three clusters is not surprising. They represent approximately 74% of recorded actions over the complete dataset, indicating that users engage in more exploratory searching (Marchionini, 2006) in this environment. The average number of individual images viewed per query is relatively small, so users may be initially evaluating results based on the thumbnails of the image along with provided metadata and then are selective about the individual images they view. The fact that the average time between queries is relatively long for all clusters supports the idea that users are spending time evaluating content--although for the vast majority of sessions users are not interested in a large number of images. The results for the studied environment, with one overwhelmingly dominant cluster, were very different than those observed by Chen and Cooper (2001) for an online public access catalog system and Wolfram, Wang and Zhang (2009) for the three public Web systems they studied, perhaps highlighting the more uniform nature of search behavior in the studied environment. One implication arising from the present findings is that interfaces for image-based digital libraries need to facilitate rapid browsing of groups of images. Detailed browsing of individual images was less important for searchers. Further research with other image-based digital library environments is needed to confirm whether this is also observed with other similar image-based collections.

	Cluster		
	1	2	3
Cluster Membership (Number of Sessions, Percentage of Total)	14787 97.37%	382 2.52%	17 0.11%
	Cluster Center		
Session Actions	21.01	35.59	47.29
Proportion of Browsing Actions	.69	.69	.83
Proportion Landing Page Visits	.14	.09	.02
Average Time Between Queries	50.70	880.84	3038.64
Average Single Item Views Per Query	1.37	3.52	7.29
Print Views	.13	.32	.41

Table 4. K-means cluster outcomes

CONCLUSIONS

The analysis of clickstream transaction log data for the studied image-based digital library reveals that although search lengths may be very different, the general search behavior is quite uniform, with the vast majority of search sessions representing focused searches with predominantly browsing-related actions, but little viewing of individual items. More detailed browsing, represented by longer intervals between actions and viewing of more source images, constituted less than 3% of the recorded sessions. System support of focused browsing that allows users to easily scan and identify relevant content is needed in these types of specialized digital library environments.

The exploratory cluster analysis represents one method by which to study patterns of search behavior. Future research will apply additional approaches such as sequence mining and network analysis of session actions to further characterize user search behavior in an image-based digital library.

ACKNOWLEDGMENTS

We'd like to thank the University of Wisconsin-Milwaukee libraries for providing access to the transaction log data.

REFERENCES

Agosti, M., Crivellari, F., & 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*, *24*(3), 663-696.

Belk, M., Papatheocharous, E., Germanakos, P., & Samaras, G. (2012). Investigating the Relation between Users' Cognitive Style and Web Navigation Behavior with K-means Clustering. In Advances in Conceptual Modeling (pp. 337-346). Springer Berlin Heidelberg.

Chen, H. M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*, *52*(11), 888-904.

Frias-Martinez, E., Chen, S. Y., Macredie, R. D., & Liu, X. (2007). The role of human factors in stereotyping behavior and perception of digital library users: A robust clustering approach. *User Modeling and User-Adapted Interaction*, *17*(3), 305-337.

He, D., & Göker, A. (2000). Detecting session boundaries from web user logs. In *Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research* (pp. 57-66).

Jansen, B. J. (2008). Searching for digital images on the Web. *Journal of Documentation*, *64*(1), 81-101.

Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. (2000). A transaction log analysis of a digital library. *International Journal on Digital Libraries*, *3*(2), 152-169.

Kathuria, A., Jansen, B. J., Hafernik, C., & Spink, A. (2010). Classifying the user intent of web queries using k-means clustering. *Internet Research*, *20*(5), 563-581.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41-46.

Park, M., & Lee, T. S. (2013). Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, *31*(1), 123-140.

Park, S., Lee, J. H., & Bae, H.J. (2005). End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research*, *27*(2), 203-221.

Shiri, A. (2011). Query management techniques and their impact in digital libraries. *International Journal of Information*, *3*(1), 9-17.

Pu, H. T. (2008). An analysis of failed queries for Web image retrieval. *Journal of Information Science*, *34*(3), 275-289.

Sfakakis, M., & Kapidakis, S. (2002). User behavior tendencies on data collections in a digital library. In *Research and Advanced Technology for Digital Libraries* (pp. 550-559). Springer Berlin Heidelberg.

Spink, A., Wolfram, D., Jansen, M. B., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, *52*(3), 226-234.

Stenmark, D. (2008). Identifying clusters of user behavior in intranet search engine log files. *Journal of the American Society for Information Science and Technology*, *59*(14), 2232-2243.

Trieschnigg, D., Nguyen, D., & Meder, T. (2013). In search of Cinderella: A transaction log analysis of folktale searchers. In the *Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, 36th Annual ACM SIGIR Conference, SIGIR 2013*, Dublin, Ireland. August 1st, 2013.

Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology*, *60*(5), 896-910.

Xu, B., & Recker, M. (2012). Teaching analytics: A clustering and triangulation study of digital library user data. *Educational Technology & Society*, *15*(3), 103-115.