Identifying Web Search Session Patterns Using Cluster Analysis: A Comparison of Three Search Environments

Dietmar Wolfram
School of Information Studies, University of Wisconsin, Milwaukee, P.O. Box 413, Milwaukee, WI 53201. E-mail: dwolfram@uwm.edu

Peiling Wang
School of Information Sciences, College of Communication and Information, University of Tennessee at Knoxville, Knoxville, TN 37996-0341. E-mail: peilingw@utk.edu

Jin Zhang
School of Information Studies, University of Wisconsin, Milwaukee, P.O. Box 413, Milwaukee, WI 53201. E-mail: jzhang@uwm.edu

Session characteristics taken from large transaction logs of three Web search environments (academic Web site, public search engine, consumer health information portal) were modeled using cluster analysis to determine if coherent session groups emerged for each environment and whether the types of session groups are similar across the three environments. The analysis revealed three distinct clusters of session behaviors common to each environment: "hit and run" sessions on focused topics, relatively brief sessions on popular topics, and sustained sessions using obscure terms with greater query modification. The findings also revealed shifts in session characteristics over time for one of the datasets, away from "hit and run" sessions toward more popular search topics. A better understanding of session characteristics can help system designers to develop more responsive systems to support search features that cater to identifiable groups of searchers based on their search behaviors. For example, the system may identify struggling searchers based on session behaviors that match those identified in the current study to provide context sensitive help.

Introduction

Transaction logs of Web search environments can provide investigators with a wealth of data for analysis to better understand user search behavior without subjecting users to controlled laboratory experiments. Within these logs are records of query content and search actions that may be studied for patterns. Typically, research to date into user searching of Web sources has revealed that users' interactions with search tools are brief and the users do not put much effort into their search and browsing behavior (Spink, Wolfram, Jansen & Saracevic, 2001). Research has revealed that the majority of Web queries are short and reiterated unsystematically using naïve search strategies (Wang, Barry, & Yang, 2003). Yet, the number of Internet users continues to grow. These users bring a range of search expertise. Many users still may be novices as information consumers when it comes to searching for more complex and diverse content. To assist users, system developers and search intermediaries would benefit from knowledge of patterns of usage behavior, if they can be identified.

Analysis of Web search activities can be undertaken at several different levels of granularity that provide different perspectives on searcher populations or information retrieval (IR) system usage. At the most fundamental level, one can examine the terms (or individual searchable words) entered by users and their frequencies of use (Silverstein, Marais, Henzinger, & Moricz, 1999; Wolfram, 1999). Next, one can examine query characteristics, including the relationships and patterns among terms, which can be useful for subject analysis (Jansen, Spink & Saracevic, 2000; Wolfram, 2000). More generally, one can study user session characteristics, where sessions comprise one or more consecutive queries from the same identifier and may be further bounded based on changes in the topicality of submitted queries or temporal boundaries of inactivity between query submissions (He, Göker, & Harper, 2002; Spink & Jansen, 2004). Finally,

Received May 6, 2008; revised October 6, 2008; accepted December 15, 2008

 DOI: 10.1002/asi.21034

the regularities that exist among groups of sessions that allow generalizations of search behaviors to be revealed also may be studied (Chen & Cooper, 2001; Shaaban, McKechnie & Lockley, 2003).

Transaction logs can be used to compare search patterns across different examples of similar search systems, such as public search engines that cater to broad audiences (Jansen & Spink, 2005). Logs may also be used to compare patterns across different types of IR system environments, such as bibliographic search systems, public search engines, and online public access catalogs that cater to different audiences (Wolfram, 2008). These logs can also be used to track changes in patterns over time when sufficient longitudinal datasets are available (Blecic, Dorsch, Koenig, & Bangalore, 1999).

Although, conceptually, the units of analysis (terms, queries, sessions, user groups) may be easy to comprehend, their operationalization, particularly at the session level, can be more challenging. One of the challenges associated with analyzing server-side search logs is the identification of the boundaries of a single interaction session (Silverstein, Marais, Henzinger, & Moricz, 1999; Montgomery & Faloutsos, 2000; He, Göker, & Harper, 2002; Jansen, Spink, Blakely, & Koshman, 2007). This is complicated by the fact that most transaction logs do not record uniquely identifiable information about searchers or their session boundaries, unless there is a specific log in requirement, which is not the case for most publicly available Web-based search environments.

Queries submitted in a session characterize the interactions carried out by the searcher. Session-based analysis can provide understanding of three dimensions of user behaviors: (a) interaction behaviors, by analyzing the length of a search session, the number of reiterations, and manipulation of results; (b) linguistic behaviors, by analyzing queries that represent needs, subsequent queries that revise the original query, and structural variations of the queries; and (c) underlying thoughts for the search actions. (As an example, the revision of a previous query by capitalizing names, which suggests that the searcher has hypothesized that the case mattered in the formulation of queries.) By analyzing a single user's search session consisting of reiterated queries, we are able to identify which terms or concepts were submitted to express information needs and how the terms were related. If two terms cooccurred in a query, there was an association between them. We assume that frequencies of term cooccurrence indicate the strength of associations between terms. Session-based in-depth analysis can reveal the searcher's knowledge on the topic, which can be depicted using *nodes* representing the query terms and *links* representing term associations; the strength of the association between terms can be measured by frequency of cooccurrence. This graph is known as a concept map of the searched topic (Wang, 2006).

As part of the authors' larger set of goals for a project to develop models of user search behavior in Web-based environments, this paper reports on the analysis of three Web query corpora to reveal search session patterns that may be used to identify distinct groups of sessions. Attributes associated with sessions (e.g., query characteristics, time intervals between queries, session length or number of queries per session) represent complex interactions for which patterns may not be readily apparent from descriptive summaries of individual search attributes. (Note that search patterns may also be studied using click through or pages visited data. These were not available for the datasets used.) Using exploratory analysis of transaction logs from three Web search environments, each catering to a different type of audience, the authors shed light on three research questions:

1. Are identifiable and coherent groups of session patterns evident in the logs based on limited search characteristics?

   Previous research has demonstrated that distinct groups may be identified (Chen & Cooper, 2001), but by using a rich transaction log consisting of dozens of usage characteristics. However, with some transaction logs, only a limited number (i.e. fewer than 10) of such characteristics may be available.
2. Do identified groups differ across the three investigated environments?

   To date, studies have descriptively compared search characteristics across different examples of Web-based search engines (Jansen & Spink, 2005) or different types of IR systems (Wolfram, 2008). However, no research has examined whether similar groups of session characteristics may be found across different types of IR system environments (e.g., public search engine, specialized search service, academic Web site search system).
3. Do search session patterns change over time?

   Similarly, a number of transaction log analysis studies have examined general query and session characteristics of search sessions over time (Blecic, Dorsch, Koenig, & Bangalore, 1999; Spink, Jansen, Wolfram & Saracevic, 2002; Wang, Berry, & Yang, 2003) but have not examined whether identifiable groups of session characteristics change over time.

Literature Review

The study of Web search characteristics has been undertaken for over a decade, involving different Web-based services. These have included online public access catalogs (Cooper, 2001), vendor-based bibliographic and full text databases (Wolfram & Xie, 2000), academic Web sites (Wang, Berry, & Yang, 2003), and Web search services such as AltaVista (Silverstein, Henzinger, Marais, & Moricz, 1999), Excite (Spink, Jansen, Wolfram, & Saracevic, 2003), and Fireball (Hoelscher, 1998). Early studies were largely descriptive, reporting characteristics of searches, sessions, and users. More recent studies have relied on data and text mining techniques to discover patterns of usage and to

DOI: 10.1002/asi

develop models of user behavior. The number of studies published since the mid-1990s on Web search transaction log analysis has been sizeable. The focus of this review will be on studies that reveal underlying characteristics of user search patterns using mathematical modeling or exploratory statistical methods (i.e., clustering or visualization methods) on transaction log contents.

Baeza-Yates, Hurtado, Mendoza, and Dupret (2005) proposed several mathematical models for describing user session search and usage behavior using a Chilean search engine. They presented a predictive model for the number of clicks per session, a Markov model of click behavior, and a model for the time distribution between clicks to demonstrate the feasibility of applying different methods for behavior prediction. In a similar vein, Hu, Zeng, Li, Niu, and Chen (2007) developed a model to predict Web user age and gender based on known demographics and similar Web site visitation patterns using a Bayesian network approach. The authors were able to obtain almost 80% accuracy on gender identification and just over 60% accuracy for five different age groups, indicating that demographic prediction is feasible. More recently, Sadagopan and Li (2008) used Markov models of clickstream Web usage data to identify atypical search sessions (i.e., more mechanical, illogical sequences). With a better understanding of which sessions are not representative of typical session behavior, the authors hoped to improve the robustness of data mining techniques of user sessions. The authors concluded that their method was 89% effective in identifying atypical sessions that simply contribute "noise" to the analysis of user sessions.

Techniques for summarizing regularities of search behavior observed in large datasets using agglomerative techniques have been developed by several researchers at the query level. Beitzel et al. (2007) have proposed that categorization and classification of user queries can lead to increased effectiveness and efficiency in general-purpose Web search systems. They investigated properties of a very large query log from AOL over a 6-month period and were able to identify and examine topical trends over different time periods. Most topical categories exhibited flat behaviors when averaged over the length of the study. Some topics demonstrated diurnal variability (e.g., personal finance was found to be more popular in the morning). Other topics, like holidays, demonstrated longer swings on a monthly scale. An agglomerative clustering algorithm was developed and applied by Beeferman and Berger (2000) to both search queries and relevant Web pages in a transaction log to discover potential query clusters. They demonstrated the feasibility of using their developed agglomerative clustering for Web mining at the query level but acknowledged the potential computational burden associated with the method. The ultimate application proposed for the technique was to develop a clustering method for Web pages that could then be matched to user queries.

Similarly, Wen, Nie, and Zhang (2001) studied the relationship between a search query and selected Web pages, in conjunction with query contents, to develop an algorithm to categorize queries in a transaction log. The method was outlined and demonstrated with several examples but was not applied to a large dataset. Ross and Wolfram (2000) used hierarchical cluster analysis and multidimensional scaling on query term pairs for multiterm queries on a query log from the Excite search engine to identify clusters of topics based on term cooccurrences. They identified popular general search topics based on the frequency of cooccurring terms at different levels of topical aggregation. High level clusters in the hierarchy, represented broad topics such as adult-oriented material, computer technology, education, products and services, formal information needs, and informal information needs. More recently, Nowick, Eskridge, Travnicek, Chen, and Li (2005) applied cluster analysis to user queries submitted to a water quality information system. The authors' goal was to develop a user-centered perspective of the resource information space so that a future system could then suggest additional search terms. Although not specifically undertaken as a transaction log study, they demonstrated an application of query characteristics to extend the functionality of a Web search environment.

Session-level analysis of transaction logs using clustering techniques has also been undertaken. Chen and Cooper (2001) used large-scale cluster analysis methods to detect usage patterns in Web-based OPAC search sessions based on 47 session variables (e.g., session length in seconds, average number of items retrieved, average number of search modifications). Their analysis revealed six clusters, which represented different types of search behaviors, including unsophisticated usage, knowledgeable usage, and known-item searching. Huang, Ng, Cheung, Ng, and Ching (2001) identified navigational patterns within sessions using cluster analysis on two government Web server transaction logs. They developed a cube (i.e., three dimensional) model of searcher browsing behavior that focused on browsing and page requests. However, the authors do not appear to have included query characteristics. Based on the analysis, the authors found that clusters with well-defined characteristics were usually quite small given the complexity of session characteristics. Similarly, Shaaban, McKechnie, and Lockley (2003) relied on cluster analysis techniques to identify groups of session behaviors on an architectural, engineering, and construction information system. They identified four clusters of behavior: exploratory, interactive sessions; knowledgeable searcher sessions; specific item searches with help-intensive behavior; and passive sessions with unsuccessful, short seeking episodes. In a pilot study to the present investigation, the authors performed a similar analysis on subsets of three transaction log datasets (Wolfram, Wang, & Zhang, 2007). This initial analysis on subsets of the data revealed similar cluster outcomes for the datasets, representing a public search engine, academic Web site, and consumer health information portal. Three to four primary clusters arose from the analysis for each dataset.

Based on the findings of earlier studies and the demonstrated applicability of clustering techniques for revealing hidden patterns in usage that may be grouped, the current

DOI: 10.1002/asi

study performs a larger scale analysis using larger datasets and includes a more longitudinal analysis.

Methodology

The transaction log datasets used in the present study represent three different types of Web search environments, representing a public search engine, a subject-specific search system dealing with consumer health information, and an academic Web site (Table 1). Other aspects of the datasets have previously been analyzed elsewhere (Spink, Jansen, Wolfram, & Saracevic, 2002; Zhang, Wolfram, Wang, Hong, & Gillis, 2008). Note that only parts of the UTK dataset have been analyzed earlier (Wolfram, Wang, & Zhang, 2007). This study represents the first full treatment of the data for 2003–2004.

Transaction log files were stored in MS Access and SQL Server databases. Queries were initially parsed for individual terms based on alphanumeric starting characters. Terms were delimited using spaces and other non-alphanumeric characters. The database structure used to house the data is depicted in Figure 1. Note that not all fields were available for each dataset. For further details, see Wang et al. (2007).

Session Identification

To study Web search behaviors, one must examine not only individual queries, but also groups of queries (sessions) submitted by users. With most transaction logs, sessions

TABLE 1. Transaction log dataset general characteristics.

Query set	Time frame	Size (queries)	Available fields	Comments
Academic Website (University of Tennessee-Knoxville)	2 years – 2003–2004 Complete dataset	3.9M	Date/Time IP address Query	This dataset contains queries entered into the UTK search engine
General Search Engine (Excite)	2 datasets representing one day each, collected in 1999 and 2001 Incomplete dataset	622K & 587K, respectively	Time Cookie Identifier Results page requests Query	Although relatively old, these datasets represent among the few that have been made publicly available to researchers
Consumer Health Information (HealthLink)	1 year – 2005 Complete dataset	377K	Date/Time IP address (encoded) Query	HealthLink is a consumer health information portal that provides access to thousands of full text articles on a range of health topics

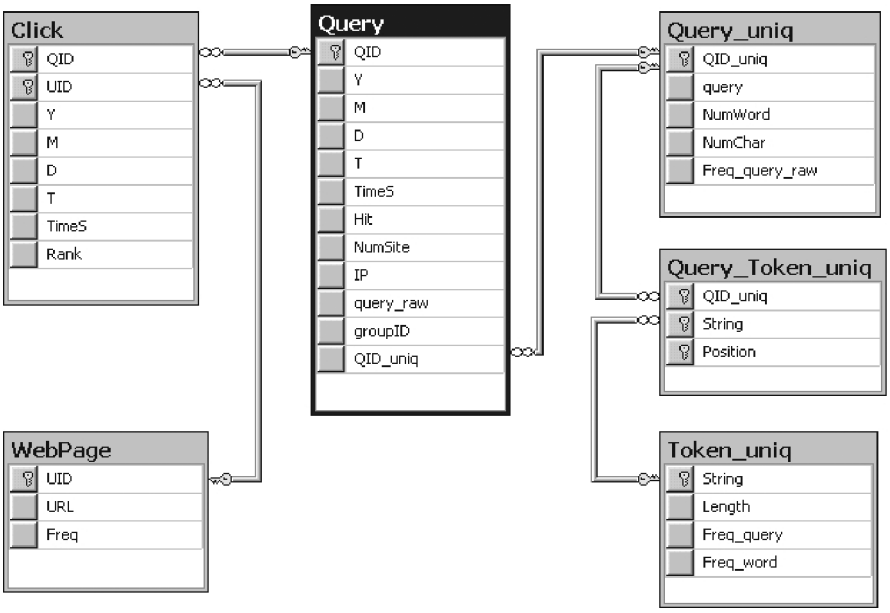

FIG. 1. Database structure used for the query data.

DOI: 10.1002/asi

are usually identified using a client-side "cookie" assigned to a given machine or an Internet Protocol (IP) address. Technically, they represent a specific computer, or group of computers, sharing an IP address and not an individual. Disambiguation of searchers in these situations is at best difficult. Different computers may share the same IP address or a public computer with multiple users may record sessions by different searchers in close temporal proximity. As a consequence, queries with the same identifier are assumed to represent queries from one searcher. For the present study, a *session*, is defined as a set of one or more consecutive queries attributed to the same identifier where the time between adjacent queries does not exceed a cutoff value. A *cutoff* value is a threshold for a query interval, or time between two temporally adjacent queries by the same identifier. Two temporally adjacent queries belong to the same session if the query interval value is less than the cutoff value; otherwise, the two queries belong to two adjacent sessions.

One of the challenges for analyzing server-side search logs is the identification of the boundaries of a single session of interactions (Silverstein et al., 1999; Jansen, Spink, Blakely, & Koshman, 2007). Typical Web query log files may include interleaved searches that are the product of different searchers or multiple simultaneous search windows by the same searcher (Pass, Chowdhury, & Torgeson, 2006). Search logs by Internet providers may have subscriber's IDs in the logs. However, it is still difficult to disambiguate sessions if the searcher conducted multiple sessions for different topics or if different searchers shared the same account, which is typical in home networking settings or public access computers in libraries.

Methods for session boundary detection have been proposed based on the subject analysis of queries (Huang, Peng, An, & Schuurmans, 2004; He, Göker, & Harper, 2002) and temporal characteristics of queries (He & Göker, 2000; Murray, Lin, & Chowdhury, 2006). Subject analysis of queries can be performed manually, but it is impractical and difficult for large datasets, potentially subjective, unreliable for short queries, and does not take into account that users may engage in multiple search topics in a given session. Subject analysis can also be performed automatically using statistical language modeling and machine learning techniques. Huang et al. (2004) acknowledged that outcomes are largely defined by the parameters used by the researchers. Özmutlu and Çavdur (2005) replicated He, Göker, and Harper's method for automatic topic identification using Excite query data. They concluded the method's performance was limited because users submit few queries and search on multiple topics, for example. However, the authors also saw value in using query patterns and time intervals for this purpose.

The application of temporal characteristics has been used, for example, by Murray, Lin, and Chowdhury (2006). Their method assumed a minimum of 20 queries per identifier from which large gaps in inter-query times (i.e., a long period of time between queries submitted by the same identifier) are located to indicate session boundaries. The temporal limits used to define session boundaries have varied considerably across studies. Montgomery and Faloutsos (2000) relied on a limit of 120 minutes of inactivity among Web browsing activities to identify session boundaries for data collected nationally by over 20,000 Web users during a 30-month period. Spink and Jansen (2004) concluded that most Web search sessions lasted about 15 minutes, with a substantial percentage lasting less than 5 minutes (p. 121). Similarly, Göker and He (2002) suggested that an optimal session boundary interval is 11 to 15 minutes. Baeza-Yates et al. (2005) used several criteria for session boundaries: excluding empty query instances, excluding queries without document selection, and using a threshold value of 15 minutes to define a session; that is, if a user (IP) submitted a query 15 minutes or less after the last click, he or she started a new session. Huang et al. (2001) suggested a boundary interval of 30 minutes for the two government search systems they studied (NASA Kennedy Space Center, EPA). The rationale for recommended boundaries is not always apparent, and it is clear the interval will be dependent on the characteristics of the dataset.

More recently, Jansen, Spink, Blakely, and Koshman (2007) compared three methods for identifying sessions: (a) IP address and cookie; (b) IP address, cookie, and temporal limit; and (c) IP address, cookie, and query patterns. They noted that the third method resulted in the best match when compared with a human assessment of query session assignment. The authors also defined two concepts related to a session: *session length* as the number of queries and *session duration* as the period between first query and last query associated with the same session.

Many options, therefore, have been proposed to estimate session boundaries. The method developed by Murray, Lin, and Chowdhury (2006) appeared promising. However, based on the number of queries associated with each identifier data for all datasets in the current study, the average number of queries associated with each identifier was far below this threshold. Less than 1% of queries associated with a given identifier for the Excite and HealthLink datasets were eligible for this method.

To establish a consistent method of applying session boundaries for the current study, two assumptions were made. First, sessions were assumed not to span multiple days because few queries are submitted around midnight. This decision was concluded to be reasonable for the University of Tennessee (UTK) and consumer health information corpora because the number of queries from the same IP addresses on the same day submitted between 5 minutes prior and 5 minutes after midnight is extremely small. Each Excite dataset was collected on the same day, so this was not an issue. Second, the characteristics of each dataset were used to determine session boundaries. For the present study, the characteristics of the dataset of the different query corpora were taken into account, along with the skewing effect of long query intervals resulting from two queries submitted by the same identifier. Ideally, for the latter, the two queries should be assigned to the same session. Spink and Jansen (2004) noted that the mean session time associated with an identifier was more

DOI: 10.1002/asi

than 2 hours for some search engine datasets, but most sessions lasted less than 15 minutes (p. 121). Users may engage in search activities for a lengthy period of time, but if there is a long period of inactivity, it is reasonable to treat any subsequent query as the beginning of a new session, although it is on the same topic. Earlier studies (Jansen, Spink, & Saracevic, 2000; Spink, Wolfram, Jansen, & Saracevic, 2001) treated all queries associated with the same identifier as being part of the same session. This may be reasonable for datasets collected over a short period, as was the case for the Excite data used in those studies. However, one would be hard-pressed to conclude that a gap of considerable time in relation to all gaps for query submissions would be a result of continued browsing without additional query submissions, particularly if queries associated with the same identifier were many hours, days, or even months apart within the same log.

Another way to identify a suitable cutoff is to examine the characteristics of the query intervals to study how different cutoffs affect resulting session lengths. The identified value should not underestimate the number of queries per session, as this will result in a higher number of short sessions with brief query intervals. Similarly, the identified value should not overestimate the session length with the inclusion of lengthy query intervals that actually represent the boundaries of two adjacent sessions associated with the same identifier. This would be the case if all queries associated with an identifier were assumed to be part of the same session. Wang et al. (2007) and Wolfram (2008) found that as cutoff times increased for session boundaries, there was a logistic (S-shaped) relationship with the average session length where average session lengths increased nonlinearly but with decreasing growth at the high end of the distribution. A reasonable cutoff would be associated with diminishing changes in the average session length with increases in query intervals. Relying on a cutoff value from the region of the distribution where the rate of increase drops provides a compromise between cutoff values that would result in average session lengths that are greatly affected by small changes in cutoff times and session lengths that are too large because the high cutoff value permits the inclusion of queries from adjacent sessions. Wang et al. observed for the subsets of the Excite, HealthLink, and UTK datasets used in an earlier investigation that the 70th and 80th percentile cutoffs for the datasets represented a two- to five-fold increase in time, but only resulted in approximately a 10% to 17% increase in the average session lengths. Therefore, the differences in average session length were not largely affected by relying on a cutoff higher in the distribution. Lower level cutoff values resulted in much larger changes in average session length with smaller changes in cutoff times. Although not an ideal solution, the consideration of a cutoff point based on changes in the average session length takes into account the characteristics of each dataset. Also, this method addresses the obvious query interval outliers that extend across many hours. These outliers can skew the average session lengths and query intervals, which would result in an overestimation of the number of longer sessions. However, as with the other approaches used for session boundary detection summarized above, this method cannot address the possibility that several computers and searchers are sharing the same IP address, or that a single searcher may be relying on multiple browser windows to engage in parallel search sessions whereby interwoven queries will be logged.

Identifying Groups of Sessions

Identifying patterns inherent in large datasets is challenging. As outlined in the literature review, exploratory methods have been developed to identify common characteristics in datasets that allow observations to be grouped together. Clustering techniques, in particular, have shown promise for the identification of distinct groups of sessions (Chen & Cooper, 2001; Shaaban, McKechnie, & Lockley, 2003). Cluster analysis can be used for several purposes, including hypothesis generation and the investigation of conceptual schemes to group things (Aldenderfer & Blashfield, 1984, p. 9). The most frequent use of clustering methods is for the creation of classifications or categories. This present research uses this technique to reveal groups of sessions based on common characteristics found within the session content.

With a consistent session boundary identification method for each dataset, the investigators were able to apply a method derived from Chen and Cooper's study to cluster sessions based on session characteristic similarities across three different Web search environments. Briefly, the method involves the following steps:

1. Identify characteristics associated with each term, query, and session.
2. Select characteristics that provide a broad range of values to distinguish sessions, both those that are readily available from the transaction logs themselves and those that may be derived from the log data. These are limited by the field data of the transaction logs. Up to 10 session variables calculated from the transaction logs were available (Table 2).
3. Standardize the values for each characteristic for each session using the mean and standard deviation of each characteristic for each dataset.
4. Divide each dataset into two subsets of sessions to determine if similar clustering outcomes are produced. Similar outcomes provide a measure of internal validity for the clustering method applied to the data corpus.
5. Process each sample using exploratory cluster analysis.
6. Graph and visually inspect the outcome of each sample for consistency.
7. Interpret the nature of the session characteristics associated with each cluster.

Based on the clustering outcomes, observations may be made and conclusions drawn regarding the search patterns within sessions, and users for each system, by implication.

Results

Session Cutoff Calculations

Each of the Excite and HealthLink datasets was analyzed in its entirety. The size and extended time frame of the UTK

DOI: 10.1002/asi

TABLE 2. Available and derived session characteristics.

Name	Description	How calculated
Session length	The number of queries submitted per session	
Average number of terms used per query	The mean number of terms used per query within a session	Total terms per session divided by total queries per session
Average term popularity	The mean frequency of occurrence of query terms across all queries within the dataset	Total frequency of occurrence of each query term divided by the total number of terms in the session
Average query interval	The mean time between query submissions per session (zero if there is only one query in a session)	Total query intervals in a session divided by the total number of queries
Average term use frequency	The mean number of times a given term is used during the session	Number of query tokens in a session divided by number of query types (i.e., unique terms) in the session
Average number of pages viewed per query	The mean number of page requests made during a session for a given query	Total number of page requests made during a session divided by the number of queries in the session
Number of searches using Boolean operators	The total number of queries in a session that use one or more Boolean operators (AND, OR, NOT, +, −)	
Average number of unrecognized or nonstandard words	Each query token is compared against a standard dictionary. If it is not found in the dictionary, it's an unrecognized term	Total number of unrecognized words in the session divided by the number of queries in the session
Average number of stopwords	Each query token is compared against a standard dictionary of stopwords. If it is found in the dictionary, it's a stopword	Total number of stopwords in a session divided by the number of queries in the session
Average term number changes	Changes in the number of terms used in subsequent queries in a session are tallied	Total number of changes in a session divided by the number of queries in a session

data allowed the investigators to divide up the data into six smaller datasets representing 4-month intervals, corresponding roughly to the semesters in the academic year, to permit more of a longitudinal analysis. The frequency distributions for all query intervals were tabulated for each dataset by tallying the number of occurrences of each interval across a dataset. Cutoff points representing the times for the 65th, 70th, 75th, 80th, 85th, and 90th percentiles of the query interval data were identified based on the interval distribution. The average session length for each cutoff point was then calculated by applying the cutoff point to the queries in the dataset to identify session boundaries. As an example, if 500 seconds represented the 60th percentile of the query interval distribution, then this was used as the cutoff to calculate the average session length for this percentile. Query intervals associated with the same identifier were then compared against the cutoff with longer intervals representing boundaries between adjacent sessions associated with the same identifier. Resulting session lengths were then tallied to calculate the average session length. This was then repeated for the 65th, 70th, 75th, and so on, percentile of the data. Cutoff times and their corresponding average session lengths were plotted and visually compared for changes in the slope of the distribution at each data point. The selected cutoff value for each dataset was adopted based on the largest change in the slope of the distribution between data points.

As an example, Figure 2 shows the changes in average session lengths for different cutoff percentiles for the HealthLink dataset. The changes in the slope of the distribution are largest between the 75th (122 seconds; average session length: 2.16

DOI: 10.1002/asi

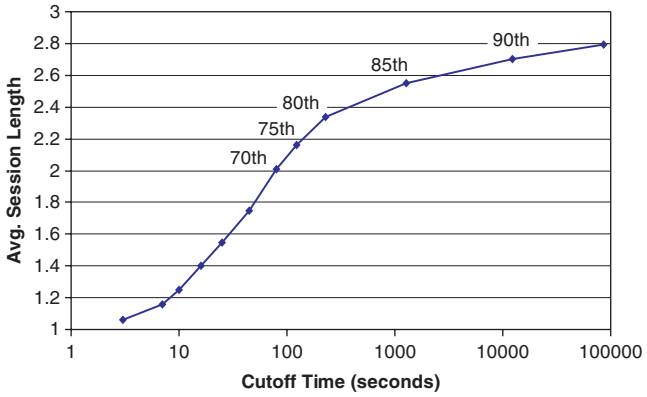

FIG. 2. Distribution of average session lengths based on cutoff values.

queries) and 85th percentiles (1274 seconds; average session length: 2.55 queries), after which there are diminishing increases in the average session length. Therefore, the cutoff time corresponding to the 80th percentile (229 seconds; average session length: 2.34 queries) was adopted as the session cutoff between queries sharing the same identifier. A summary of the adopted cutoff times appears in Table 3. The UTK datasets contain significant differences in cutoff times, beginning at 1305 seconds and dropping to 496 seconds by the last time period. Notably, however, the average session lengths remain largely unchanged over the 2-year period. These changes in cutoff time would not have been apparent had the dataset been analyzed uniformly.

Cluster Determination

Each dataset was divided into two equally sized samples of sessions, comprising every second record. Sessions within each sample were clustered to identify broad categories using the SPSS TwoStep cluster analysis feature. Unlike traditional clustering methods, which do not scale well to larger datasets (SPSS, 2001), the TwoStep analysis is useful for very large datasets and allows for an optimal number of clusters to be requested, or a specified number of clusters. The routine initially preclusters data into several dozen to several hundred subclusters using sequential clustering (Theodoridis & Koutroumbas, 1999). The resulting subclusters are then grouped into the desired number of clusters using more traditional agglomerative hierarchical clustering.

Not all of the session characteristics described in Table 2 were present in each dataset. For example, only the Excite datasets included page requests associated with each query. Also, not all characteristics contributed to stable cluster outcomes. Those variables that did not contribute to the development of stable clusters included the number of searches that used Boolean operators, the average number of unrecognized/nonstandard words, the average number of stopwords used, and the average term number changes. These characteristics were frequently zero and did not add any distinguishing features for session patterns, and so they were not used in the analysis. Those that were used and contributed to stable clusters in one or more of the datasets were the session length, average terms used per query, average term popularity, average query interval, average term use frequency, and average pages viewed per query (only available for the Excite datasets).

Analysis runs with fixed and varying numbers of clusters revealed the most coherent patterns for all three environments using three emergent clusters. Other numbers of clusters did not result in the same cluster characteristics across each group. Also, comparisons across the different environments would be more difficult with differing numbers of clusters. The resulting clusters from each environment demonstrate common session characteristics across the systems.

To provide evidence in support of the validity of the clusters, a systematic random sample of 200 sessions was drawn from each dataset. A human judge was instructed to assign each session to one of the three identified clusters based on the cluster characteristics. The level of agreement in the assignment of sessions to the identified clusters was between 66% and 70% for the datasets, indicating a reasonable match between the automatic clustering method and the human judge. One limiting factor was the human judgment of session characteristics, particularly the average term popularity. The average session length, query interval, and number of terms used per query in relation to other sessions could be assessed relatively easily. However, the human judge was not in a position to readily assess the average term popularity, because tens of thousands of query terms existed in each

TABLE 3. Dataset session characteristics.

System	Dataset	Cutoff (seconds)	Total sessions	Mean session-length (queries)
Academic	2003 Jan–April	1,305	411,014	1.81
	2003 May–Aug.	1,257	283,686	1.79
	2003 Sept.–Dec.	558	252,691	1.86
	2004 Jan–April	418	441,963	1.84
	2004 May–Aug.	415	411,396	1.83
	2004 Sept.–Dec.	496	350,934	1.83
General search engine	1999	809	385,145	1.62
	2001	1,074	327,051	1.80
Consumer health information	2005	229	161,277	2.34

DOI: 10.1002/asi

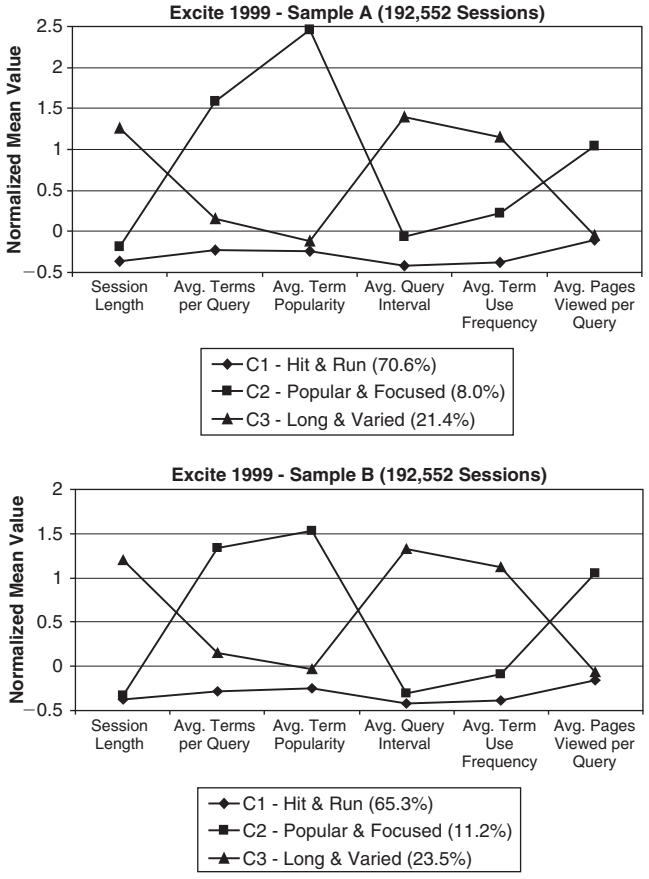

FIG. 3. Search engine—Excite 1999 cluster outcomes.

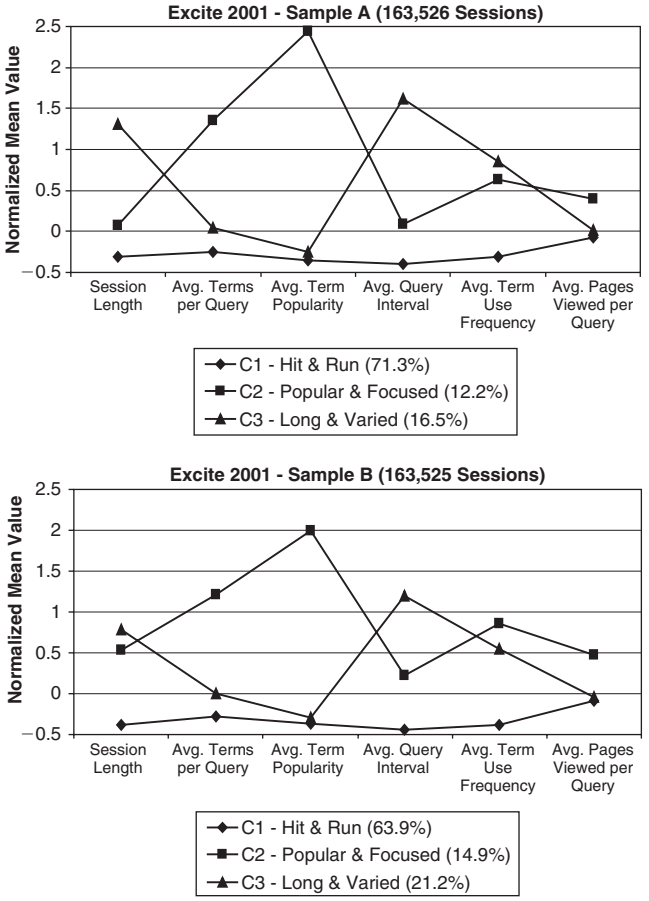

FIG. 4. Search engine—Excite 2001 cluster outcomes.

dataset and this assessment would need to be made for each of the hundreds of terms in the session samples.

General Search Engine

The availability of two logs taken approximately a year and a half apart presented the opportunity to determine if shifts in session behavior could be determined between the datasets. Figures 3 and 4 summarize the cluster outcomes for the 1999 and 2001 datasets, respectively. Although the data for these figures can be represented in a tabular format, the use of line graphs allows the similarities between the samples and the differences between the clusters to be more readily interpreted. The y-axis of each figure represents the average normalized value for members of that cluster. The content of the data logs allowed six search characteristics to be used as part of the analysis. Because the search characteristics are measured in different scales, the normalized outcomes are presented, where zero represents the average value for each characteristic across all sessions.

In the pilot study for this research, the investigators allowed the number of clusters to be determined by SPSS. The resulting number of clusters was four, unlike the other two environments that produced three stable clusters. It is possible that the additional variables for cluster formulation available in the Excite data permitted finer distinctions between clusters. To provide a more direct comparison between the systems, the number of clusters for the Excite datasets was fixed at three. The three clusters generated for both Excite datasets demonstrate distinct types of session behaviors. There are slight variations in the normalized mean values between the samples and two datasets, but the "shape" of the search characteristics for each sample is largely the same, producing very similar figures. The largest cluster for both datasets (C1), representing from 63.9% to 71.3% of sessions per sample is characterized by below average standardized values representing brief sessions (usually one query), with few terms per query, usage of relatively unique terms, and short query intervals. Essentially these are "hit and run" sessions, usually incorporating infrequently used terms, in which little effort and follow-up is undertaken in the form of modified queries. They may represent highly effective searches where relevant results were found quickly, or possibly ineffective searches where the session was quickly abandoned. Cluster C2 represents the smallest session category. It is characterized by mid-range length sessions comprising long queries using popular vocabulary, average re-use of terms in subsequent queries, average length query intervals, and above average numbers of results page requests. This cluster appears to represent more focused searches on popular topics. Finally, cluster C3, representing from 16.5% to 23.5% of sessions, contains the longest sessions and query intervals with above average re-use of terms in subsequent queries but consisting

DOI: 10.1002/asi

TABLE 4. Sample sessions representative of each cluster.

	Time	Query	Total words	Pages viewed
Academic Web site				
Cluster 1	1/1/2003 3:00:00 P.M.	death records	2	
Cluster 2	2/5/2003 2:37:45 P.M.	office of student conduct	4	
Cluster 3	2/12/2004 4:29:22 P.M.	human rights	2	
	2/12/2004 4:29:30 P.M.	human rights	2	
	2/12/2004 4:32:41 P.M.	definition of human rights	4	
	2/12/2004 4:32:54 P.M.	acts of omission	3	
General search engine				
Cluster 1	06:29:16 A.M.	"insightful quotes"	2	1
Cluster 2	07:14:31 A.M.	russian financial shock	3	1
	07:20:56 A.M.	russia financial shock	3	2
	07:24:46 A.M.	russia financial crisis	3	1
Cluster 3	3:17:36 P.M.	allentown	1	1
	3:18:21 P.M.	allentown map	2	1
	3:27:19 P.M.	milford square, pa	3	2
	3:36:08 P.M.	milford	1	1
	3:36:34 P.M.	milford pa	2	2
	3:44:34 P.M.	spinnerstown, pa	2	2
Consumer health information				
Cluster 1	1/27/2005 2:25:29 P.M.	uterine + fibroids	2	
Cluster 2	4/10/2005 11:42:21 A.M.	too + much + vitamin + c +	4	
	4/10/2005 11:42:32 A.M.	risks + of + too + much + vitamin + c	6	
Cluster 3	8/6/2005 2:36:37 P.M.	stiff + neck	2	
	8/6/2005 2:36:53 P.M.	neck	1	
	8/6/2005 2:38:47 P.M.	cigarettes	1	
	8/6/2005 2:39:22 P.M.	cigarettes	1	

of infrequently used terms and a mid-range average number of terms per query. Between the two datasets, the average proportional representation of cluster C1 remains largely the same whereas the size of clusters C2 and C3 increase and decline, respectively. The only notable difference between the two time frames is the decrease in difference in the average number of pages viewed per query for cluster C2, indicating fewer differences in page browsing behavior over time among the clusters.

Consumer Health Information Portal

Generated clusters for HealthLink exhibit the same characteristics as the Excite data, although with only five search variables used in the cluster generation (Figure 5). The proportion of sessions in cluster C1 is similar to the Excite datasets, but the percentage of more focused searches, represented by cluster C2, is higher (between 17.3 and 19.8%), with a lower percentage in cluster C3 (between 12.7% and 12.9%). Notably, the normalized average query intervals for cluster C3 are even higher for HealthLink, indicating an even larger difference between the other clusters than observed for the Excite datasets.

Academic Web site

The 2 years of academic Web site search log data permitted a longitudinal investigation of session clustering. This allowed us to address the third research question. The data log was divided into 4-month intervals, corresponding roughly to winter/spring, summer and fall semesters at UTK. Only four search characteristics resulted in stable clusters (session length, average terms used per query, average term popularity, average query interval). The average term use frequency characteristic, although available, did not produce enough variability to be suitable for differentiating session clusters. Similar outcomes were observed for the cluster characteristics as in Wolfram, Wang, & Zhang (2007), although these varied over time. Instead of plotting figures for each of the six samples, each variable is plotted across the 2-year timeframe for comparison.

The mean session length does not vary for any of the clusters (Figure 6). Cluster C3 represents the longest sessions, as it did for the Excite and HealthLink datasets. Both C1 and C2 contain relatively short sessions. The average number of terms used per query does not vary over time for cluster C3, but it does vary somewhat for clusters C1 and C2 (Figure 7). The mean values for the two clusters are very close; thus, they can only be differentiated by other variables such as average term popularity (Figure 8). For cluster C2 the plot demonstrates an overall upward trend (slightly longer queries) over this period, whereas for cluster C1, the plot displays an overall downward trend (slightly shorter queries). The average term popularity for clusters C1 and C3 does not change appreciably over the time frame, but it does decline for cluster C2 (Figure 8). Queries in cluster C2 relied on less frequently used terms.

DOI: 10.1002/asi

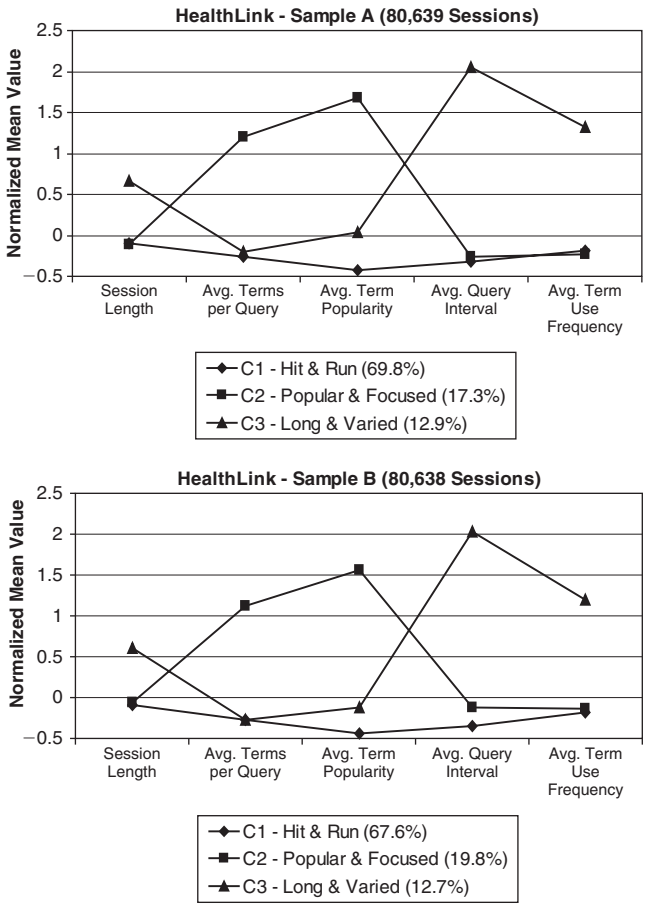

FIG. 5. Consumer health information portal cluster outcomes.

Similar to session length, the average query interval does not change across the six datasets for all clusters, with the exception of an increase for cluster C2 in the final time period (Figure 9). Cluster C3 contains, by far, the longest average query intervals. These figures demonstrate that both the average session lengths and the query intervals did not change over the 2-year time frame for the three clusters, but the average number of terms used per query and the average term popularity did change for cluster C2, with longer queries and less frequently used terms. In addition, the average number of terms per query for C1 dropped slightly. It appears that the sessions in cluster C3 represent the problematic and struggling sessions that were on popular topics and reiterated over and over again.

Figures 6 through 9 demonstrate stability and change in the session variables but do not provide any indication of proportional changes in cluster membership. The changes in cluster membership, or the percentage changes in the distribution of the clusters, are found in Figure 10. Membership in cluster C3, which comprises the longest sessions and longer query intervals, shows no changes over the 2 years. These sessions represented a relatively constant percentage of all sessions. What is particularly interesting is the 30% drop in cluster C1, "hit and run" type of searches, whereas the proportional representation of sessions within cluster C2,

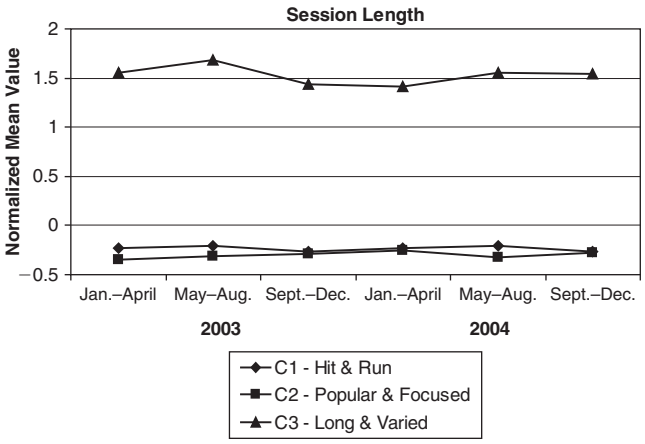

FIG. 6. Longitudinal comparison of mean normalized session length for the UTK data.

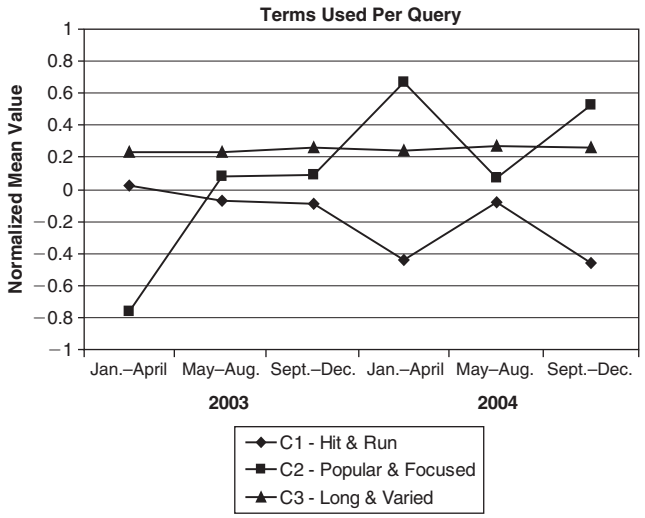

FIG. 7. Longitudinal comparison of mean normalized terms used per query for the UTK data.

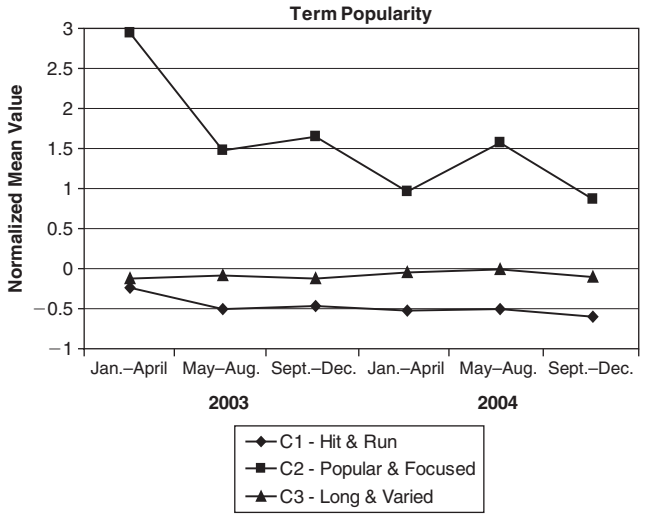

FIG. 8. Longitudinal comparison of mean normalized term popularity for the UTK data.

DOI: 10.1002/asi

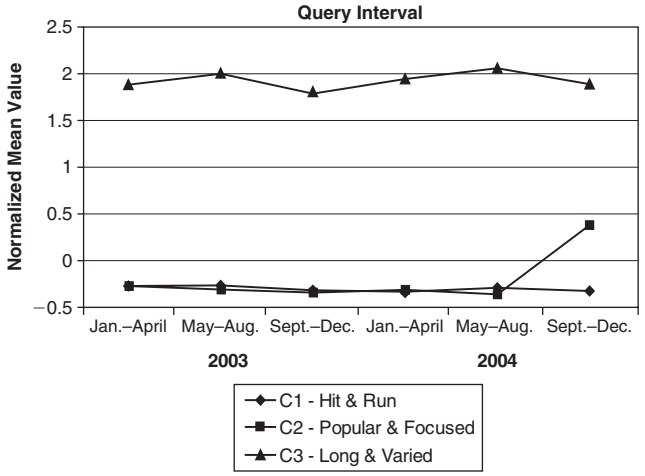

FIG. 9. Longitudinal comparison of mean normalized query interval for the UTK data.

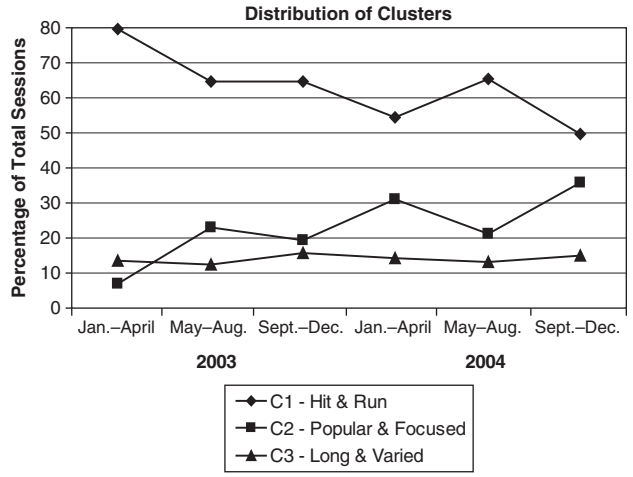

FIG. 10. Longitudinal comparison of cluster distribution for the UTK data.

representing focused searches on popular terms, increased approximately 30% during the 2-year timeframe.

Discussion

The analysis demonstrates there are common session groups within each environment, although each system caters to different audiences and very different query intervals were observed, notably, the very brief query interval associated with HealthLink, which was up to five times shorter than some of the other datasets. Unlike more general search services, HealthLink is targeted to serve only searchers interested in health-related information. One could surmise that this difference is because of the focused information needs of HealthLink's users, as well as the more limited and specific database of content to which it provides access. The UTK Web site, conversely, that caters to people interested in university-related information, although still somewhat specialized, is more general than that of HealthLink. A public search engine like Excite attracts a much broader range of users with equally broad information needs, as reported by previous analyses of general search engine transaction logs (Spink & Jansen, 2004). Despite these differences, and the fact that the average query interval varied greatly among the systems, common session clusters emerged across all three systems.

The ability to identify distinct session behaviors provides insight into user searching and a clearer picture of overall searching behaviors. Most sessions are brief, with little sustained interaction across the three environments. The largest cluster for each environment (C1), constituting the majority of sessions for most of the samples, represented brief system interactions with little effort in searching or time spent. This cluster could be argued to represent sessions that use naïve search strategies, as reported by Wang et al. (2003), or they could represent focused searches on specific topics that did not entail additional effort. Unfortunately, the transaction logs themselves cannot reveal whether users who engage in this type of session behavior found what they were looking for, or if they gave up without modifying their queries.

The remaining clusters define two different types of sessions. Sessions involving popular search topics (C2) make use of lengthier queries than other session types but do not result in lengthy sessions. The session cluster containing observed lengthy sessions and long query intervals (C3) might be indicative of purposeful, pensive searching but is also possibly an indication that searchers are struggling. By submitting short queries that do not bring about the sought-for items, searchers must engage in additional browsing and search modification to fill their information need; most notably, the UTK data are the consistency in the relative size of the C3 cluster, ranging from approximately 12% to 16%. If this cluster does represent struggling searchers, then they do not appear to be going away. For the Excite datasets, the equivalent group represented just over 22% of the search sessions for the 1999 dataset and just under 19% for the 2001 dataset. Similarly, this group represents 13% of sessions for the HealthLink dataset. Each group comprises a sizeable and persistent minority of search sessions that may represent tens of thousands of searchers. If a search system could recognize the session characteristics of struggling searchers, then systems could incorporate a proactive help feature that assists searchers with recommendations for additional terms to refine their queries. The clustering technique used represents a post hoc analysis of session characteristics, but real time assessment may be undertaken without great computational overhead. A retrieval system could compare session behaviors to characteristics observed for cluster C3 members (e.g., lengthy sessions in relation to the average session length with longer than average query intervals and relatively obscure terms in relation to other sessions) and intervene after a threshold of session characteristics has been crossed.

The longitudinal nature of the UTK data presented the opportunity to examine changes in session characteristics over time. The Excite datasets, although representing 2 separate days a year and a half apart, do not permit the same continuity of investigation. The only observed difference was the decline in normalized average number of pages

DOI: 10.1002/asi

viewed per query for cluster C2, indicating less variability among the clusters for this session characteristic. Similarly, 1 year of HealthLink data and the comparatively smaller number of sessions generated over this time, which would result in much smaller seasonal samples, are not long enough to determine if shifts are evident or because of sampling differences. The UTK analysis demonstrated consistency in search characteristics across the clusters for some variables and changes in others. The most striking example of change over time occurred with cluster C2. The relative number of terms per query increased over the time frame, while the average popularity of terms used in these sessions went down, which points to broader vocabulary use or greater specificity. One could argue that for searches to be more effective, they should be more precise, which would mean increasing the number of terms per query and using terms with greater specificity. The increase in representation of this group over the 2-year period with an equivalent decrease in the cluster C1 "hit and run" searches could be demonstrating that searchers are learning gradually to improve their search strategies, although there is still that persistent C3 group. This was also evident in the Excite datasets.

Limitations of the present research include the inability to determine searcher intentions from the transaction logs alone. The logs represent excellent objective data that record actual search activities but cannot reveal why searchers did what they did. Complete session data, including all searcher actions taken, would provide a more complete picture. But these data were unavailable. Similarly, the number of search variables available, on which the cluster analyses are based, is limited to the content of the logs. Not all variables contributed to stable cluster identification (e.g., average term use frequency did not contribute to stable clusters for the HealthLink system). The investigators were able to derive only four to six viable variables from the available datasets. Additional variables would allow for a richer analysis and possibly the identification of finer-grained clusters of session characteristics. The age of some of the data could be problematic for inferring current search behaviors. The Excite data in particular may make the search characteristics of more historical interest. However, the similarities of the findings across the datasets, which span 6 years, indicate that there are search behavior themes that continue to be relevant over time. Another potential limitation for session-level analysis is the fact the transaction logs used did not record session boundaries. These boundaries, like with many transaction log studies, must be assessed based on the dataset content. The probabilistic method developed for this study represents another way to deal with this challenge by taking into account the dataset query interval characteristics. A lower cutoff point would have resulted in more sessions with fewer queries. Without system usage monitoring when sessions begin and end, it will always be an estimate, even with the availability IP address or client side cookie data. However, not all session characteristics will be affected by the session boundaries. The number of terms used per query, term popularity, and number of pages viewed per query are a function of the queries themselves. The characteristics that are affected by session boundaries are the session length, query interval (where the query interval is reduced because lengthy query intervals are treated as session boundaries), and term use frequency within a session (which may increase as the session length increases). The small differences in average session length, as reported in the findings, even with two- to five-fold increases in the cutoff value, would suggest that any such impact would not be substantial. Finally, cluster analysis is intended to serve as an exploratory method for revealing hidden relationships within data corpora. It does not provide definitive proof of outcomes. The similar clustering patterns across samples taken from the same datasets reveal that the outcomes are not random patterns that are unique products of each sample.

Based on the findings of the research and the challenges presented, the study of session behavior identification would benefit from richer datasets to allow a more fine-grained analysis of the sessions. Chen and Cooper (2001) were able to derive 47 variables for their OPAC session data. A particular challenge that affects many transaction log studies is the estimation of session boundaries. Research would be greatly assisted with this determination by the Web servers, which could then be included in logs for researchers. Session identifiers are more readily available in some environments like OPACs and bibliographic retrieval systems (Wolfram, 2008), where time-outs or login requirements help to delineate boundaries. Public search services like search engines, specialty public databases (e.g., HealthLink), or institutionally implemented search engines (e.g., UTK Web site) by default do not record session identifiers. The convenient availability of search boxes for quick accessibility on public access machines makes it difficult to implement session identifiers. However, a reasonable time-out feature implemented on these publicly available systems could be helpful for defining sessions in these environments. The present research relied on available data fields, which were limited across most of the systems to query data. The lack of variability of several session characteristics that could be derived from the available data demonstrated that not all characteristics are useful for defining session attributes. The lack of variability in these characteristics did not help to distinguish session types. The availability of rich, clickstream data, including retrieval results and items viewed, would provide a richer set of data to study session characteristics.

Conclusion

The authors have investigated search session patterns using cluster analysis on session variables derived from transaction logs representing three different types of Web-based search environments. First, the results revealed that coherent clusters of search behaviors emerge and that these patterns are observed across the three environments studied. The results from this study corroborate findings from studies in other IR environments that search behaviors can be clustered into distinctive groups based on search session characteristics.

DOI: 10.1002/asi

Second, the longitudinal analysis of the UTK dataset revealed a shift in the number of sessions associated with each cluster whereby proportionately fewer sessions exhibited "hit and run" searching and shifted towards more popular, focused topics. Third, each system revealed a cluster consisting of sessions that appear to exhibit characteristics of struggling searchers. The ability to recognize these types of searches based on session characteristics could allow systems to provide real-time help to searchers whose searches match these session characteristics.

The findings of the study also raise additional questions and hypotheses to be further investigated. Continued mining of the collected transaction logs will focus on sessions that are representative of the different clusters, particularly cluster C3, which represents the longest sessions in terms of the number of queries and longest query intervals. Session data from this cluster can provide insight into what problems searchers encountered and how the searchers reiterated their queries. In addition, analysis at the individual searcher's level using clickstream data, where available, can provide a better understanding of the searcher's information needs. Cooper and Chen's (2001) method for assessing the relevance of a search was based on transaction log content with access to a broader array of session variables. For public Web search environments, transaction logs data are usually limited to fewer available variables, and thus, a more finer-grained analysis of user search regularities may not be possible. However, this approach might still be combined with searcher interviews or targeted post search surveys to assess what users have been doing, whether they have been successful, and what types of difficulties they have been encountering.

Acknowledgement

The authors would like to thank Excite@home, the University of Tennessee, and HealthLink for access to the transaction log data, as well as Ningning Hong and Lei Wu for research assistance. This research was funded by the Institute for Museum and Library Service National Leadership Research grant program (LG-06-05-0100-05). Any views, findings, conclusions or recommendations expressed in this paper do not necessarily represent those of the Institute of Museum and Library Services. Thanks also go to the anonymous reviewers for their valuable suggestions.

References

Aldenderfer, M.S., & Blashfield, R.K. (1984). Cluster analysis. Beverly Hills: Sage Publications.

Baeza-Yates, R., Hurtado, C., Mendoza, M., & Dupret, G. (2005). Modeling user search behavior. In Proceedings of the Third Latin American Web Congress (pp. 242–251). Washington: IEEE Computer Society.

Beeferman, D., & Berger, A. (2000). Agglomerative clustering of a search engine query log. In Proceedings of the Sixth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (pp. 407–416). New York: ACM.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized Web query log. Journal of the American Society for Information Science and Technology, 58(2), 166–178.

Blecic, D.D., Dorsch, J.L., Koenig, M.H., & Bangalore, N.S. (1999). A longitudinal study of the effects of OPAC screen changes. College and Research Libraries, 60(6), 515–530.

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.

Cooper, M.D. (2001). Usage patterns of a Web-based library catalog. Journal of the American Society for Information Science and Technology, 52(2), 137–148.

Cooper, M.D., & Chen, H-M. (2001). Predicting the relevance of a library catalog search. Journal of the American Society for Information Science and Technology, 52(10), 813–827.

Göker, A., & He, D. (2002). Analysing Web search logs to determine session boundaries for user-oriented learning. Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (pp. 319–322). London: Springer-Verlag.

He, D., Göker, A., & Harper, D.J. (2002). Combining evidence for automatic Web session identification. Information Processing & Management, 38(5), 727–742.

Hoelscher, C. (1998). How Internet experts search for information on the Web. In H. Maurer & R.G. Olson (Eds.), Proceedings of WebNet98 World Conference of the WWW, Internet & Intranet. Charlottesville, VA.

Hu, J., Zeng, H.J., Li, H., Niu, C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. In WWW '07 Proceedings of the 16th International Conference on World Wide Web (pp. 151–160). New York: ACM.

Huang, X., Peng, F., An, A., & Schuurmans, D. (2004). Dynamic Web log session identification with statistical language models. Journal of the American Society for Information Science and Technology. 55(14), 1290–1303.

Huang, Z., Ng, J., Cheung, D.W., Ng, M.K., & Ching, W.K. (2001). A cube model and cluster analysis for Web access sessions. In Proceedings of WEBKDD 2001, pp. 47–57.

Jansen, B.J., & Spink, A. (2005). How are we searching the World Wide Web?: An analysis of nine search engine transaction logs. Information Processing & Management. 42(1), 248–263.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. Information Processing & Management, 36(2), 207–227.

Jansen, B.J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on Web search engines. Journal of the American Society for Information Science and Technology, 58(6), 862–871.

Montgomery, A.L., & Faloutsos, C. (2000). Trends and patterns of WWW browsing behavior. Retrieved December 10, 2008, from http://pages.cpsc.ucalgary.ca/~saul/personal/other_pubs/web_trends.pdf

Murray, G.C., Lin, A., & Chowdhury, A. (2006). Identification of user sessions with hierarchical agglomerative clustering. In Proceedings of the ASIS&T Annual Meeting [CD-ROM]. Medford, NJ: Information Today, Inc.

Novak, J.D. (1998). Learning, creating, and using knowledge. Mahwah, NJ: Lawrence Erlbaum.

Nowick, E.A., Eskridge, K.M., Travnicek, D.A., Chen, X., & Li, J. (2005). A model search engine based on cluster analysis of user search terms. Library Philosophy and Practice, 7(2). Retrieved December 10, 2008, from http://www.webpages.uidaho.edu/~mbolin/nowick-et-al.htm

Özmutlu, H.C., & Çavdur, F. (2005). Application of automatic topic identification on Excite Web search engine data logs. Information Processing & Management, 41, 1243–1262.

Pass, G., Chowdhury, A., & Torgeson, C. (2006). A picture of search. Retrieved December 10, 2008, from http://ir.iit.edu/~abdur/publications/pos-infoscale.pdf

Ross, N.C.M., & Wolfram, D. (2000). End-user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information Science, 51(10), 949–958.

Sadagopan, N., & Li, J. (2008). Characterizing typical and atypical user sessions in Clickstreams. In WWW '08 Proceedings of the 17th International Conference on World Wide Web (pp. 885–894). New York: ACM.

DOI: 10.1002/asi

Shaaban, S., McKechnie, J., & Lockley, S. (2003). Modelling information seeking behaviour of AEC professionals on online technical information resources. ITcon, 8, 265–281. Retrieved December 10, 2008, from http://www.itcon.org/data/works/att/2003_20.content.03818.pdf

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR Forum 33, 1 (Sep. 1999), 6–12.

SPSS. (2001). The SPSS TwoStep Cluster Component. Retrieved December 10, 2008, from ftp://ftp.spss.com/pub/web/wp/TSCWP-0101.pdf

Spink, A, & Jansen, B.J. (2004). Web search: Public searching of the Web. Dordrecht: Kluwer.

Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. Computer Magazine, 35(3), 107–109.

Spink, A., Wolfram, D. & Jansen, B.J., & 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.

Theodoridis, S., & K. Koutroumbas. (1999). Pattern recognition. New York: Academic Press.

Wang, P. (2006). Final Report on ALISE/OCLC 2005 Research Grant: A Dual-approach to Web query mining: Towards conceptual representations of information needs. Retrieved December 10, 2008, from http://www.oclc.org/programsandresearch/grants/reports/2005/wang-p.pdf

Wang, P., Berry, M.W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. Journal of the American Society for Information Science and Technology, 54(8), 743–758.

Wang, P., Wolfram, D., Zhang, J., Hong, N., Wu, L., Canevit, C., & Redmon, D. (2007). Mining Web search behaviors: Strategies and techniques for data modeling and analysis. In Proceedings of the 2007 Annual Meeting American Society for Information Science and Technology.

Wen, J., Nie, J., & Zhang, H. (2001). Clustering user queries of a search engine. In Proceedings of the 10th International Conference on World Wide Web. Retrieved January 16, 2009, from http://www10.org/cdrom/papers/368/index.html

Wolfram, D. (1999). Term co-occurrence in Internet search engine queries: An analysis of the Excite data set. Canadian Journal of Information and Library Science, 24(2/3), 12–33.

Wolfram, D. (2000). A query-level examination of end user searching behaviour on the excite search engine. In H. Olson, (Ed.). Proceedings of the 28th Annual Conference of the Canadian Association for Information Science. Retrieved December 10, 2008, from http://www.cais-acsi.ca/proceedings/2000/wolfram_2000.pdf

Wolfram, D. (2008). Search characteristics in different types of Web-based IR environments: Are they the same? Information Processing & Management, 44, 1279–1292.

Wolfram, D., Wang, P., & Zhang, J. (2007). Modeling Web session behavior using cluster analysis: A comparison of three search settings. In Proceedings of ASIST 2007 Annual Meeting (Milwaukee, WI, October 19–24).

Wolfram, D., & Xie, H. (2000). End user database searching over the Internet: An analysis of the state of Wisconsin's BadgerLink service. In M. E. Williams (Ed.), Proceedings of the 21st National Online Meeting (pp. 503–512). Medford, NJ: Information Today.

Zhang, J., Wolfram, D., Wang, P., Hong, Y., & Gillis, R. (2008). Visualization of health subject analysis based on query term co-occurrences. Journal of the American Society for Information Science & Technology, 59(12), 1933–1947.

DOI: 10.1002/asi