To the Graduate Council:

I am submitting herewith a thesis written by April Ann Lewis entitled “Enterprise Users and Web Search Behavior”. I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Information Science.

Peiling Wang, Major Professor

We have read this thesis and
recommend its acceptance:

Dania Bilal

Lorraine Normore

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures on file with official student records)

Enterprise Users and Web Search Behavior

A Thesis Presented for
The Master of Science
Degree
The University of Tennessee, Knoxville

April Ann Lewis
May 2010

Acknowledgements

I would like to thank Oak Ridge National Laboratory (ORNL) for supporting my graduate education and encouraging me to pursue my newly found interests in applied information science research. ORNL's Chief Information Officer (CIO) and the web server support team graciously provided me with a very robust data set and answered all questions I had regarding its format.

I would also like to acknowledge the extensive amount of work that Dr. Peiling Wang has done in the area of Web data mining and analysis. Dr. Wang's relational database model for web queries was fundamental to my data mining efforts. I have learned much as a graduate student from her research work as well as from her classroom instruction. I am honored that Dr. Wang agreed to chair my thesis committee.

I am also very grateful to have had Dr. Lorraine Normore and Dr. Dania Bilal serve on my committee, both very accomplished in complementary areas of Information Science research. Dr. Normore first introduced me to human-computer interaction (HCI) relevant to information search. One of the motivations for this thesis was characterizing the corporate user's interaction with the ORNL intranet search environment. Dr. Bilal provided me with the basic understanding of search environments, tasks related to information searching and the theory of cognitive motivation for successful search.

Abstract

This thesis describes analysis of user web query behavior associated with Oak Ridge National Laboratory's (ORNL) Enterprise Search System (Hereafter, ORNL Intranet). The ORNL Intranet provides users a means to search all kinds of data stores for relevant business and research information using a single query. The Global Intranet Trends for 2010 Report suggests the biggest current obstacle for corporate intranets is "findability and Siloed content". Intranets differ from internets in the way they create, control, and share content which can make it often difficult and sometimes impossible for users to find information. Stenmark (2006) first noted studies of corporate internal search behavior is lacking and so appealed for more published research on the subject.

This study employs mature scientific internet web query transaction log analysis (TLA) to examine how corporate intranet users at ORNL search for information. The focus of the study is to better understand general search behaviors and to identify unique trends associated with query composition and vocabulary. The results are compared to published Intranet studies. A literature review suggests only a handful of intranet based web search studies exist and each focus largely on a single aspect of intranet search. This implies that the ORNL study is the first to comprehensively analyze a corporate intranet user web query corpus, providing results to the public.

This study analyzes 65,000 user queries submitted to the ORNL intranet from September 17, 2007 through December 31, 2007. A granular relational data model first introduced by Wang, Berry, and Yang (2003) for Web query analysis was adopted and modified for data mining and analysis of the ORNL query corpus. The ORNL query corpus is characterized using Zipf Distributions, descriptive word statistics, and Mutual Information. User search vocabulary is analyzed using frequency distribution and probability statistics.

The results showed that ORNL users searched for unique types of information. ORNL users are uncertain of how to best formulate queries and don't use search interface tools to narrow search scope. Special domain language comprised 38% of the queries. The average results returned per query for ORNL were too high and no hits occurred 16.34%.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and General Information

Introduction

Many companies are adopting internet search practices for their intranets. While the underlying search process is the same for both the Internet and the intranet, the search needs of the respective users and their environments are very different (Fagin et al., 2003). The Internet consists of users who have individualized information needs and share no understanding with the information providers. Internet users have access to an unbounded document set that may include advertisements and spam.

Conversely, ORNL intranet users search for information individually, but they share contextual understanding of the information space with the providers. The document set or search corpus available to ORNL users is controlled and limited. Users are not exposed to advertisements or spam within the search environment. Much more is known about internet search as many studies have been published that include search success statistics. The number of unsuccessful Internet searches reported by college students in a recent library user internet search survey was nearly 50% of all internet search submissions. (Mann, 2005). It is difficult to find any similar qualitative results measured relative to intranet search.

There are two very distinct environments when it comes to web search 1) the internet and 2) the intranet. The way these environments are viewed from both users and researchers are very different. There are only a handful of published studies regarding intranet search, but internet search reports are published nearly every three months. The most recent internet statistics published was in February (Nielson, 2010), which reported that Google is the most preferred search engine (65.2% of all searches). That same report listed Yahoo as second, losing 18% more of its previously reported search share to Google. The percentage of typical daily users has grown to nearly 50%, with users extremely positive about search engines and their search experiences (Fallows, 2008). However, in that same report users are described as generally unsophisticated about how and why they use search.

In contrast, there are no free regular web based reports available to the public on intranets statistics. When in-depth reports or studies are available they typically must be purchased. On average, intranets workers spend about 25% their time

searching for information (Feldman, &Sherman, 2004). Feldman and Sherman (2004) also report that a company with 1,000 knowledge workers may waste well over $6M dollars a year looking for information that doesn't exist, failing to find information that does, or recreating information that "could" have been found. The search experience for intranet users is not pleasant. A recent enterprise intranet search survey by Ward (2005, Sept. 7) found that "web-rage" was experienced after 12 minutes of fruitless search, although nearly 7% of the 566 people surveyed said they felt irritated after only three minutes.

Not only is there a difference in the internet and intranet search environment, there are key unique distinctions in search engine performance and query vocabulary requirements. For example, indexing and ranking of search results on the internet can be impacted by organic linking and spam. The intranet is not affected by spam and cross linking is not typically practiced in corporations. The way search results are "stitched" together as a product of federated search is different. The intranet has special rules for stitching like security access, duplication, etc. Tagging of information is not implicit within the intranet, which affects indexing. This is not to say implicit tagging of items associated with the internet always results in improved search performance. Intranets tend to have a smaller or narrow search vocabulary due to special domain language.

The functional capability of a dynamic search is also critical for intranets. It is estimated that intranets as enterprises have tens or even hundreds of times larger data collections (both structured and unstructured) than internets (Li, Cao, Hu, Xu, Li, & Meyerzon, 2005). The recent intranet study done by Li, et al.(2005) demonstrated that an intranet search does not just focus on search of relevant documents; it includes special types of information such as definitions, persons, experts, homepages and applications. Another unique challenge to search inside the intranet is dealing with secure content, when it is not included the value for the searcher is greatly diminished (Valdez-Perez, 2007). David Hawking (2006) aptly describes the enterprise as a complex information environment which makes measuring the quality of search results difficult. While this study does not offer a solution to this problem, it is characterizing the ORNL intranet which could provide a framework for evaluating corporate web search environments. Clearly this is a motivating factor for comprehensively analyzing ones corporate intranet, specifically measuring general search behavior exhibited by users, examining trends in query submission and reformulation, as well as results of search both successes and failures.

Successful search equates to optimized “findability”. Measuring findability means characterizing the enterprise search environment. This typically involves analyzing query logs to identify what topics users are searching for, query formulation which is characterizing query submissions, and the percentage of search failure (no hits or too many results). It is a presumption of this study that it is not enough to understand query level results. It is also necessary to analyze information related to the general search behavior which describes how and when users search. Only when we understand both search behavior and search results can we improve overall efficiency within intranet search systems. General search behavior can be determined by analyzing access logs and usage reports. It complements search analysis by helping us understand the unique characteristics of our web users.

It is because of these fundamental differences that organizations must evaluate their intranet search solution; simply applying best practices found with internet search is not practical. A successful organization must make sure that users can actually find information on their unique systems in a reasonable amount of time. Efficient search engines must be configured to match the characteristics of the users and the special information they seek. The most common way to characterize the users and the information they seek is to gather statistics on intranet usage and to evaluate user search logs.

Transaction Log Analysis (TLA) typically focuses on the interaction behaviors occurring among the users, the search system, and the information (Jansen, 2009). Content analysis of server log files describes user interaction as it relates to internet usage statistics/reports, and search queries. Several studies have been done in this area with only Stenmark (2005, 2006) focusing on corporate intranets (Beitzel, Jensen, Lewis, Chodury,& Frieder, 2007; Wang, 2006; Baeza-Yates, Calderon-Benavides & Gonzalez, 2006; Wang, Berry, &Yang, 2003 ; Jansen & Spink, 2006; Wolfram, Wang, & Zhang, 2008. This study will contribute to TLA by applying the (Wang, et al., 2003) method of mutual information analysis to intranet queries. It also implements the Wang, (2006) method of topic identification, complemented by general transaction analysis of ORNL user search usage statistics. In addition to contextual analysis, this study includes indirect analyses of access logs and usage reports to better characterize general ORNL search behavior. Unlike narrowly focused published intranet studies, this study will comprehensively analyze a corporate intranet’s user web query corpus for the purpose of improving the overall Intranet search experience. Along with query logs it evaluates access and usage

logs in order to gain a holistic view of the ORNL search enterprise. A literature review suggests this may also be the first study to perform TLA on an intranet site using the Microsoft Office SharePoint Search Engine.

This thesis will add to the growing body of literature associated with web query transaction log analysis for Intranets by providing methodology to other intranet users and managers who may want to holistically analyze their search environment. It combines log analysis used by search system administrators to measure search engine performance and interaction along with traditional query log analysis which measure users search performance and interaction. The thesis is organized as follows. The next section discusses the research questions associated with this study. Chapter 2 summarizes the public extent of research related to intranet and web search. Chapter 3 characterizes the ORNL enterprise search environment, the transaction log files used in the study and the research methodology. Chapter 4 presents results and discussion while Chapter 5 summarizes the study results and discusses implications of the study.

Research Questions

This study employs mature scientific internet web query transaction log analysis (TLA) to better understand how intranet users at ORNL search for information. The focus of the study is examining general search behaviors and identifying unique trends associated with query composition and vocabulary. The goals of the research are three-fold and include answers to the following research questions (RQ):

RQ1. What general search behaviors do ORNL searchers exhibit when searching the intranet?

a. What is the size of the ORNL search audience?
b. What interfaces do ORNL users employ most when they search?
c. What types of pages do ORNL users click most often when results are available?
d. What topics do ORNL users commonly search for?
e. When do ORNL users search and what are their search results?

RQ2. How do ORNL users formulate their queries?

a. What are the most frequently submitted queries?

b. How many ORNL user queries are unique?
c. How many ORNL user queries are blank?
d. What are the lengths of ORNL user queries?
e. What are the distribution of ORNL queries relative to length and time?

RQ3. What are the characteristics of the ORNL user vocabulary?

a. What is the length and distribution of ORNL unique terms?
b. With what frequency do ORNL user queries contain acronyms, abbreviations, and misspelled words?
c. What is the frequency of common stop words?
d. Are there terms that occur together frequently ("term co-occurrence")?

RQ4. How do ORNL results compare to the published studies?

Chapter 2

Literature Review

TLA Theory and Methodology

This chapter provides a brief overview of mature Transaction Log Analysis (TLA). The overview contains two major sections, the first TLA theory and the second TLA methodology. The overview is then followed by a short discussion of literature review objectives. Following review objectives are the discussions of each work and what impact they had on developing methodology for this study.

TLA Theory

The use of data stored in transaction logs of web search engines, intranets and web sites can provide valuable insight into understanding the information searching process of internet searchers(Jansen 2006). Many researchers(Jansen, Spink, & Taska, 2009) feel transaction log data can provide feedback into what users are looking for in search architectures. Although there is a body of literature on empirical studies of TLA, few provide detailed methodological clarifications on data models used and the underlying rationales for these models (Wang, Wolfram, Zhang, Hong, &Wu, 2007). While TLA is emerging as a viable research methodology, it is not without its critics. Critics feel that TLA doesn't go far enough and that the logs don't record the user's perceptions of the search and therefore don't measure the real needs of the information searcher (Kurth, 1993).

TLA Methodology

Many studies have examined transaction log analysis (TLA) of web based search engines. Researchers have used transaction logs for analyzing a variety of applications from internet search to library information retrieval (IR) systems (Croft, Cook, &Wilder, 1995; Jansen, Spink &Sarajevic, 2000; Jones, Cunningham, &McNab, 1998; Wang, et al., 2003; Wang, Wolfram, &Wu, 2008). In "Search log analysis: What it is, what's been done, and how to do it", Jansen reviews the fundamental research motivation for TLA and describes a methodology for conducting succesful TLA research. A recent tutorial published by Wang, Wolfram, and Wu (2008) entitled "Web Search Log Analysis and User Behavior Modeling" focuses specifically on the technical process for conducting

web transaction log analysis using the best tools developed by researchers over the last decade.

In all of these studies, TLA methodology is commonly described as a three-stage process. The first stage is data collection which includes the process of collecting the interaction data for a given period of time using transaction logs. The second stage is cleaning and parsing the log files to make them suitable for analysis. The third and final stage is analysis which requires selecting a specific research methodology. Of course, the research questions define what can be answered by the default data in typical transaction logs (Jansen, 2006). Fortunately, today's search logging software easily allows for expanding unobtrusive data collection of additional variables to meet analysis needs.

Data Collection

Transaction logs come in different formats, but more recent commercially available search tools produce standard World Wide Web Consortium (W3C) extended or Internet Information Services (IIS) format log files. Inherently, all data logs vary in content. The data format and fidelity should be addressed along with any predefined assumptions (Jansen, &Pooch, 2001). In "Privacy Concerns for Web Logging Data" Kirstie Hawkey (2009) suggests researchers should anonymize or otherwise transform any sensitive or personal data before receiving, working with or publishing it. Most private or government organizations have policies related to sensitive information management. Researchers should consult with the Chief Information Officer (CIO) in their organization to discuss proper handling and dissemination of search log related information.

Most log files contain data that can be used to analyze users search behaviors with IR systems whether internet or intranet by discerning attributes of distinct search processes and their resulting components. Jansen and Pooch (2001) establish the framework terminology for analyzing the search process describing three distinct components 1) *Session*, 2) *Query*, and 3) *Term*. Session analysis is focused on discrete entries entered by single users. This is the most difficult of the three as the researcher must determine what constitutes a session. Session boundary detection is difficult as users search for multiple topics on a single computer or a single computer may be shared by multiple searchers (Wolfram, et al., 2008). Sessions can be comprised of single or multiples queries.

A query is defined as a string of characters or word(s) entered into an information retrieval system. A query can contain multiple strings of characters or words (Korfhage, 1997). Query level analysis usually involves examining query length, query complexity, and failure rate. Query length represents the number of words or unique character strings in a query. Query syntax looks at specific components comprising the words or strings. This can range from the use of special symbols like hyphens to Boolean operators, even examination of capitalization and spelling. Failure rate quantifies how often a searcher receives no information matches for their character string submission. Today's search logs usually report failure rate as "number of hits". When searches receive no results matching their query, the number of hits equal zero.

A term is defined as a string of characters separated by some delimiter such as a tab, space, comma, or colon. It is up to the researcher whether they should include special syntax or delimiters in the queries or terms. There are impacts to the analysis whether you keep them or remove them like defining unique semantic terms. Term analysis involves evaluating the number of characters in a term, the frequency of the term and its tendency for it to appear with other terms in queries or the corpus. High usage terms are those terms that occur most often in a search corpus and are easily identified by tokenizing queries (splitting multiple term queries into single terms) and counting identical terms. Mutual information or term co-occurrence measures the occurrence of term pairs. In "Mining Longitudinal Web Queries: Trends and Patterns", Wang, Berry, and Yang (2003) examine co-occurrence with queries extracted unobtrusively from the website of the University of Tennessee, Knoxville (UTK). To promote statistical consistency in the ORNL search model, the present study employs these authors methodology for queries and word pairs.

Data Preparation

Data preparation is the most important and time consuming component of TLA. Cleaning of the raw log files usually require identification of format and data record errors through visual inspection of the file. Depending on the size of the file and the type of errors, a single editing script might be sufficient. More likely the search file will contain hundreds if not thousands of records, many requiring a "unique" editing solution, an instance in which manual edits and multiple scripts are required.

Typically, the percentage of corrupted data is small relative to the overall data set (Jansen, et al. 2009). Data preparation also includes identifying exclusion data.

Exclusion data are special instances of data that are excluded from analysis like addresses, or phone numbers because they will negatively impact the search log analysis objectives. The last step in data preparation is importing the clean TLA data into a relational database or log analysis software tool and calculating standard interaction metrics that will serve as a basis for further analysis (Jansen, 2006).

Data Analysis

The best way to manage search log queries for multiple types of analysis is through a robust relational database management system (RDBMS). Importing and tracking each query as a unique event affords traceability from derived characterization data. It is simpler in a RDBMS to attach additional attributes to each record and to correlate across a diverse population of records. Statistical analysis should include at least the mean, standard deviation and median wherever possible if you intend to compare results across studies. All data should be presented with the lowest unit of measure, avoiding aggregation category values at all cost (Jansen, and Pooch, 2001). Lastly, the RDBMS method for storing quantitative data is optimal for secondary analysis.

Literature Review Objectives

In support of this study, an extensive literature search was conducted using online sources, conference proceedings, technical articles and two significant reference books "Web Search: Public Searching of the Web" by Jansen, and Spink(2005) and "Handbook of Research on Web Log Analysis" edited by Jansen, Spink and Taksa(2009). The latter is a must for anyone considering TLA research.

The criteria for related work in this study was that it must be focused on context analysis using TLA methods and involve an intranet or an academic web site. This study presumes academic web sites qualify as an "intranet like site" as they do have limited access (password protected accounts) and employees use the same enterprise search site. It also presumes role based access that is staff has access to more information than students. Qualifying studies were placed in one of three context analysis categories 1) Session Analysis, 2) Longitudinal Analysis, and 3) Visual Presentation of Information Needs.

Session study usually involves analyzing query information specific to individual measures like length of session, average number and length of sessions per user. Sometimes it will involve analysis of click-through behavior, which is done to see where the searcher has been or to predict where they are going next.

Longitudinal analysis is temporal query analysis and is usually focused on analyzing query trends for a single search site across multiple time increments, usually across months and or years. These types of studies (Stenmark & Jadaan, 2006; Wang, et al., 2003) look at query and token frequencies to identify popular queries (top 100 and top 25), words, word pairs and triples. Most include characterizing words in the corpus using Zipf distributions. Only one evaluates term co-occurrence using mutual information statistics.

Visual presentation of information needs focuses on research methods used to identify what users are looking for and ways to visually represent the results in a topic map. These studies usually involve quantitative analysis of queries resulting in the clustering or aggregation of query information into topics.

Session Analysis

A literature review suggests Dick Stenmark's article "Searching the intranet: Corporate users and their queries" (2005) is one of the first intranet studies on web sessions. The study was done for SwedCorp a commercial vehicle manufacturing company using the UltraSeek search engine by Verity. Session analysis is difficult because there is no variable in the UltraSeek log file that indicates when a user begins and ends a search. The single item that varies across these studies is the time threshold defining a search session. This study chose 13 minute session boundaries. After determining the threshold Stenmark analyzed the data to determine session length in terms of interaction per session, the elapsed time of each session, and distribution of the sessions. The study also involved query analysis, reporting number of queries, zero term queries, and repeat queries. Single term queries dominated with no query containing more than 9 terms. Stenmark's study (2005) is relevant to the ORNL study because it too looks at intranet queries. Some of the results from the ORNL study can be compared to the SwedCorp results with the following caveats: the SwedCorp study involves the UltraSeek search engine which limits indexing of intranet information to URL's only. This limits the search study to page results that link to text documents, not real enterprise multimedia or applications search. UltraSeek is also an "anonymous" search engine and because it doesn't know who you are and what you can have access to, it restricts you from "all" sensitive intranet information. This is a good example of why intranet studies are needed on the newer search engines like Microsoft Office SharePoint Server (MOSS). MOSS logs do give indications as to when a user starts and stops a session. MOSS does not limit what is counted, for example access to all media in

all pages is counted, not just single URL page access. MOSS knows who the user is because it employs password protected access. Lastly, MOSS is able to index not just filter, which means it indexes more than URL's.

"Mining Web Search Behaviors: Strategies and Techniques for Data Modeling and Analysis" by Wang, et al. (2007) used the 80-20 empirical rule to develop an interactive web tool for exploring certain query session thresholds. The Wang, et al. (2007) study analyzed many of the same query and session issues as Stenmark's(2005) study, but the implementation was quite different. This study implemented a highly granular, comprehensive relational data model which maximized transactional data inclusion and expansion. Great detail was included in the data section describing data preparation, processing and construction of the data model. The concept of the data model was the inspiration for the ORNL data model. The data used in this analysis was from multiple sites (Excite, HealthLink, and UTK), only one of three qualifying as an "intranet like" site, UTK. The only variables that are available for comparison in this study are top queries and unique queries. Fortunately, Wang, Berry, and Yang (2003) also have an earlier longitudinal analysis using four years of UTK search data stored in a relational data model that is relevant to the ORNL study.

Longitudinal Analysis

"Intranet Users Information-Seeking Behavior: an Analysis of Longitudinal Search Log Data" by Stenmark and Jadaan (2006) is focused on temporal characterization of intranet users across three different years, comparing results to public web studies. In the 2006 study, Stenmark and Jadaan evaluated SwedCorp's query data submitted to their InfoSeek Search site. While the paper also includes some session analysis data, the bulk of the analysis focused on the search queries. His query analysis reported for each year the number of queries, empty queries, single terms, average number and maximum number of terms in a query. Again results viewed pages were analyzed with reports on number of explicit pages, the mean and maximum number of results pages viewed. Stenmark and Jadaan's (2006) study suggests intranet users engage in fewer and shorter search sessions than the public web studies. The length of intranet query submissions is significantly shorter than public searches. This study certainly gives some results that can be compared to ORNL results. Stenmark and Jadaans(2006) study tends to not discuss cleansing and processing of the data, a lack of methodology substance.

Another article by Stenmark in 2006 "What are you searching for? A content analysis of intranet search" involves a pure intranet study done using Volvo intranet

search logs. It was a longitudinal study from 2002 through 2004, although not the same months or even days across years. This study not only involved typical query analysis but included an open card sort exercise to derive topics from query terms. Zipf distributions were used to characterize the word corpus. Some analysis was done regarding term pairs and triples, as well as advanced statistics on word pairs. He also includes linguistic analysis of Boolean operators. Many of the reported results will be useful for comparison. While this study is more comprehensive in the area of context analysis, it still does not provide much substance in methodology.

"Mining Longitudinal Web Queries: Trends and Patterns" by Wang, Berry, and Yang (2003) entails the analysis of four years worth of UTK site search logs (May 1997 to May 2001). The research objectives were very user oriented, understanding their user web query behavior, identifying search problems, and developing techniques for optimizing query analysis. A comprehensive characterization of queries was done, along with word associations using Zipf distribution. What stands out for this query study is that the paper, logically presents in detail their data processing and analysis techniques. A web query entity relationship model helps describe each step in the process and how the relational data management structure was built. It was easy to see how the same measurements could be produced with the ORNL data set. This paper provides an extensive roadmap for contextual search analysis.

Visual Presentation of Information Needs

There is only one relevant publication that falls in this category, "A Dual-approach to Web Query Mining: Towards Conceptual Representations of Information Needs" by Wang (2006). This study also examines University of Tennessee, Knoxville (UTK) queries, but with an added focus of web clustering for identifying what information users are seeking. The strategy was to analyze mutual information values and similar queries of a single user session for the purpose of identifying semantically related terms. Mutual information was certainly helpful, but threshold boundaries were needed to more tightly identify sessions and thus topic branching. The visual representation of semantic networks was interesting because it helped describe the relationship between unique high frequency terms and word pairs. It also demonstrated how mutual information values can be used to help cluster words based on association strength.

Conclusion

A granular relational data model first introduced by Wang, Berry, and Yang (2003) for Web query analysis was adopted and modified for data mining and analysis of the ORNL query corpus. The ORNL query corpus is characterized using Zipf Distributions, log-log graphs and descriptive word statistics found in both Stenmark and Jadaan(2006) and Wang, et al. (2007) respectively. User search vocabulary is analyzed using frequency distribution and probability statistics (Mutual Information), a methodology both attributable to Wang, Berry, and Yang (2003). Results from both of the aforementioned studies will be used for results comparison. The ORNL study will build on visual topic identification using mutual information values similar to the study by Wang (2006).

Chapter 3

Methods

Research Environment

This research is based on analysis of web query logs from ORNL's intranet. ORNL is a multi-program science and technology laboratory managed for the Department of Energy (DOE) by UT-Battelle, LLC. ORNL is also the Department of Energy's largest science and energy laboratory. Scientists and engineers at ORNL conduct basic and applied research. Their goal is to develop scientific knowledge and technology that strengthens the nation's leadership in six key areas of science; energy science, high-performance computing, neutron science, materials science at the nanoscale, systems biology, and national security. ORNL also performs other work for DOE including isotope production, program management, and science related information management (http://www.ornl.gov/).

ORNL has over 4,600 staff and approximately 3,000 guest researchers at the laboratory every year. Staff and visitors are a mix of U.S. and foreign citizens. Educationally they represent a mix of technical professionals, degreed workers, and students at both graduate and undergraduate level.

In 2007 ORNL replaced its Verity UltraSeek search engine with Microsoft SharePoint Server 2007. SharePoint content that is shared through this tool is document libraries, picture libraries, lists, discussion boards, surveys, individual and shared web sites and web workspaces. The ORNL SharePoint search engine indexes about 200 public and internal web servers, covering close to 1,000,000 documents. This search server change netted nearly a three-fold increase in the number of documents searched by users and removed strict anonymity from the ORNL intranet search process. Now that ORNL searchers are being exposed to three times as many information sources, it is more important than ever to make sure that the results provided to users via intranet search are relevant.

The search engine unobtrusively generates several log files. Which log file is used for analysis depends on the TLA research questions and research objectives. The three types of files used in this study are usage reports, access logs and query logs. The next section provides a structural description of the different files, followed by description of how the data was prepared and processed for analysis.

The Data

MOSS uses a lot of different logging files to help in collection of user search information. Collection of transaction log files is automatic and unobtrusive, so there was no special data collection required for this study. Analyses of these incidental logs provide an effective review of the overall ORNL user search experience. MOSS provides three key sources of usage information the administrator Search usage data, as well as access and query log information. Each can be beneficial in understanding how people are generally using the intranet site and what information they are looking for. In combination they provide deeper insight into general user search behavior. All queries analyzed in this study were submitted through the MOSS search engine and occurred between September 17 and December 31, 2007.

Data Structure

MOSS 2007 Search uses Internet Information Services (IIS) standards to capture transaction information from users and stores the output in World Wide Web Consortium (W3C) extended log file format. A W3C IIS file manager utility came with MOSS and it was used by the ORNL web manager during installation to choose which information is important to regularly collect for the organization.

The first W3C IIS file we will discuss is called the access log and it contains the date and time a transaction was recorded, the address of the server which made the log, the Internet Protocol (IP) address of the requestor, the type of browser the request was made in, a query submission if one was made, the URL address of the clicked or downloaded item, the type of page the user selected, and the length of time the request took. The fields in the file are delimited by a semi-colon and maintain a strict order. The fields are the date, time stamp, the name where the search service was running, the log location, the path of the item downloaded, the query issued, the individual requesting search access, and the type of browser used to search (table 1). Here is an example of that data log from ORNL.

2007-09-17 21:45:14 W3SVC758222333 111.xx.x.xx GET /SearchCenter/_themes/Lichen/pagebackgrad_lichen.gif - 80 ORNL\ 111.xx.xxx.xxx Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.50727;+.

NET+CLR+3.0.04506;+MS-RTC+LM+8;+InfoPath.2;+.NET+CLR+3.5.21022) 200 0 0 203,2007-09-17 21:45:14 W3SVC758222333 111.xx.x.xx POST /searchcenter/Pages/Results.aspx k=mhp&s=All+Sites 80 - 160.xx.xxx.xxx

Table 1. Defines all the information fields that are available in the ORNL access log

IIS ACCESS LOG	**DEFINITION**
date	The year, month, and day entry was recorded
time	The time the log file was recorded in UTC
s-sitename	The Internet service name and instance number that was running on the client.
s-ip	The IP address of the server on which the log was created
cs-method	Command issued by the user like GET or POST or PASS
cs-uri-stem	The path of the item downloaded or posted
cs-uri-query	The query, if any, that the client submitted. A Universal Resource Identifier (URI) query is necessary only for dynamic pages.
s-port	The server port
cs-username	The name of the authenticated user who accessed the server. Anonymous users are indicated by a hyphen
c-ip	The IP address of the client
cs(useragent)	The type of browser that the client used
sc-status	The HTTP status code
sc-substatus	The substatus error code
sc-win32-status	The Windows status code
time taken	The length of time that the action took, in milliseconds

The second type of W3C IIS file used in this study is the Query log file. The query file contains data that when analyzed can provide insight on query volume trends, top queries, click through rates, queries with zero results, search topics, and various detailed information on query level statistics. For extended query analysis and reporting query log export data is provided in Excel files.

Search query logging is enabled by default in the MOSS Shared Services Provider (SSP). The information tracked in the query log includes the query terms used, search results returned for search queries and pages that were viewed from the search results. The search usage data is beneficial in understanding how ORNL users are searching and identifying the type of information they are downloading. Below is an example of a single record of that ORNL file. Each record contains 19 fields and individual fields are separated by commas.

NULL, intimal hyperplasia, 9F73D42F-7E3D-4508-B5C0-89885EFEB222, All Sites, NULL, 6, 0, NULL, 2007-09-17 06:30:14.497, 2007-09-17 06:30:40.870,0,0,https://sharepoint.ornl.gov/search/Pages/results.aspx,ORNLMOSSINDEX,0,0,0,NULL,NULL

This sample record shows that a user typed a query string in the ORNL search box “intimal hyperplasia” as indicated in field two. The search yielded six results as listed in field 6 of the record. None of the results were clicked on the results page as indicated by the term “NULL” in the first record field. This suggests the user was not satisfied with the results or was interrupted in the search process. Fields nine and ten contain a date timestamp, the first indicating when the search was submitted (9/17/2007 at 6:30 in the morning) and the second field indicates at what time the result URL was clicked. Since no URL was clicked in this instance only the date occupies this field. These fields along with number of results, the clicked URL rank and clicked URL were used in this study. A complete list of fields and their definitions in the query log can be found in Table 2.

Table 2. Information fields and definitions of ORNL collected "query log"

Field #	MSSQL QUERY LOG	DEFINITION
1	clickedurl	URI's clicked in the results page
2	query string	The query test of the search that was executed
3	site guid	The id of the site or collection from which the search query was executed
4	scope	Defines the limits of the searchable space for example All Sites(Search Center, Top-level site, sub-sites, or Lists & Libraries), this site(current site and all its sub-sites), this list of sites(Lists & Libraries, or people(on All Sites)
5	bestBet	Keyword terms as described by the administrator to enhance search results, can also be called a "synonym ring"(a glossary of names, processed, and concepts)
6	NumResults	The number of relevant results returned for the search query
7	NumbestBets	The number of bestbets returned for the search query
8	clickedurlRank	The result position of the clicked URL
9	SearchTime	The date and time when the search was executed
10	ClickTime	The time when the resulting URL was clicked
11	AdvancedSearch	In many cases, users type a keyword phrase in the search box and then click the Go Search button or press Enter to execute their query. If this technique does not produce the result they are looking for on the first few pages of search results, some users will give up. However, advanced users tend try again by using a more advanced query to target the content they are looking for.
12	Continued	Identifies the last entry corresponding to a search query
13	resultsUrl	The URI of the page where the ranked results were posted
14	queryServer	The name of the query server in which the search query was executed
15	numHighConf	The number of high confidence results returned for the search query
16	didYouMean	Indentify if spelling suggestion is returned(0=yes, 1=no)
17	ResultView	Indentifies the order in which relevant results were ordered
18	contextual Scope	The contextual search under which the query was executed
19	contextual ScopeUrl	The URI of the contextual lscope

Lastly, Usage report information was used from the search site reporting service to complement the access and query log information. Usage reporting is a service that enables intranet SharePoint site administrators to monitor high level statistics about the use of their sites. Usage reporting also includes usage reports for search queries. Items selected from that report in this study was top queries in the last 30 days of the query log data set, the average number of search requests per day and month, as well as search results of the top destination page types.

Preparation

Data preparation included developing a plan for "cleansing" and anonymizing transaction log data. Cleaning the data includes removing data errors and anonymizing the data included removing personal user information as well as ORNL descriptive network information from the logs. The query logs contain not only query requests but identifying information of the person who initiated the request. Martin(1997), an early information scientist with a legal background was the first to consider privacy issues with monitoring of online information systems for studies in user behavior. This study implements Kurth's(1993) suggestions for protecting information that may reveal searcher identity. First all personal information like the ORNL three letter user id was removed from the logs. IP addresses were anonymized by replacing all but the first three numbers of the IP address with "x". Session analysis was not performed so there was no need to track individual user session information. Permission to use the data was secured from the CIO of the organization after submitting a reasonable data security plan. Permission to publish the results was granted after review.

The access logs were mined in their native format using the Log Parser2.2 tool and therefore did not require any special processing. The usage reports also did not require manipulation. The query log however did require data cleansing, parsing, and in some cases reformatting.

Initial review of the query transaction log found structural issues amongst the files records. Queries involving names of authors were distributed across multiple record cells, thus the strings were concatenated and used to replace the partial information in the original query string field. Additionally, a small number of files had the term "efaultproperties" in the query string box and the remainder of the row data was shifted by one cell. These query strings were deleted and the remaining data moved

left by one column. The remaining query statements were at least contained inside the query column, with some exhibiting strange forms.

MOSS supports four basic types of keyword syntax for search, prefixes, phrases, and single or multiple words. Querying the system is not case sensitive and Boolean logic is not required. It was clear just from examining the first 2000 records from the query log that users did not clearly understand the query rules of the MOSS search system. Table 3 depicts some of the non-compliant and unusual queries.

Table 3. Examples of unsupported query strings submitted by ORNL searchers

Record	QueryString	Type
24		“blank”
283	10/2007 international festival	dates
22	efaultproperties	error
23	'+vascular +injury	Boolean operators
10	vascular_injury	special character
106	fmla	acronyms
333	4200000162	numbers
397	"recruiting coordinator"	quotations
437	share.ornl.gov	partial web addresses
587	https://share.ornl.gov/projects/doe_bap	URIs
606	zip-code	hyphenated terms
1817	ji*	wildcards

A cursory glance at the query string structures suggested many parsing rules were needed to establish what a qualifying query would look like. Special characters like quotations, Boolean logic operators, commas, and underscores were removed as long as it did not impact the context of the query. Special punctuation such as commas, quotes, and back slashes were also removed. The context of a query was validated by examining queries nearby the target query. Blank queries were filtered, but not removed. Since the study focused on user vocabulary most queries containing numbers were removed. The exception to the rule was when numbers gave special context to a word or phrase i.e. W-2, 401K, etc. For practical purposes of lexical study all building numbers, phone numbers, office numbers, conference room numbers and form numbers were deleted. URL addresses as queries were also removed.

Processing

A commercially available software tool called Log Parser2.2 was used to mine data within the access text file. Log parser is a powerful, versatile tool that provides universal query access to standard IIS text-based data such as log files. Using the Log Parser2.2 tool on the access log files give the first quick glimpse into the behavior of searchers. The first step is to determine what data is valuable to the study and identify it by term, for example indentifying popular “browser types”. The next step involves telling the parser to retrieve the data about browsers called “cs(useragent)” in the access log and then telling the parser “how to process” the data. The results of your query can be custom-formatted in text based output, or they can be directed to other output like SQL, or a chart. Appendix A presents the bulk of the queries that were created to mine the access log file in this study. Details of the output and how it was analyzed can be found in the Analysis section. Again, Table 1 defines all 15 of the information fields that are available in the ORNL access log. These fields are easily identifiable in the Log Parser2.2 script examples found in Appendix A.

The analysis requirements for the entire study must be considered as the data model is constructed. An ORNL data model was constructed to assist in database design. The Query data model represents the specific data needed to meet analysis requirements. It also defines processing constraints within the query corpus and depicts the relationships between data entities (Figure1).

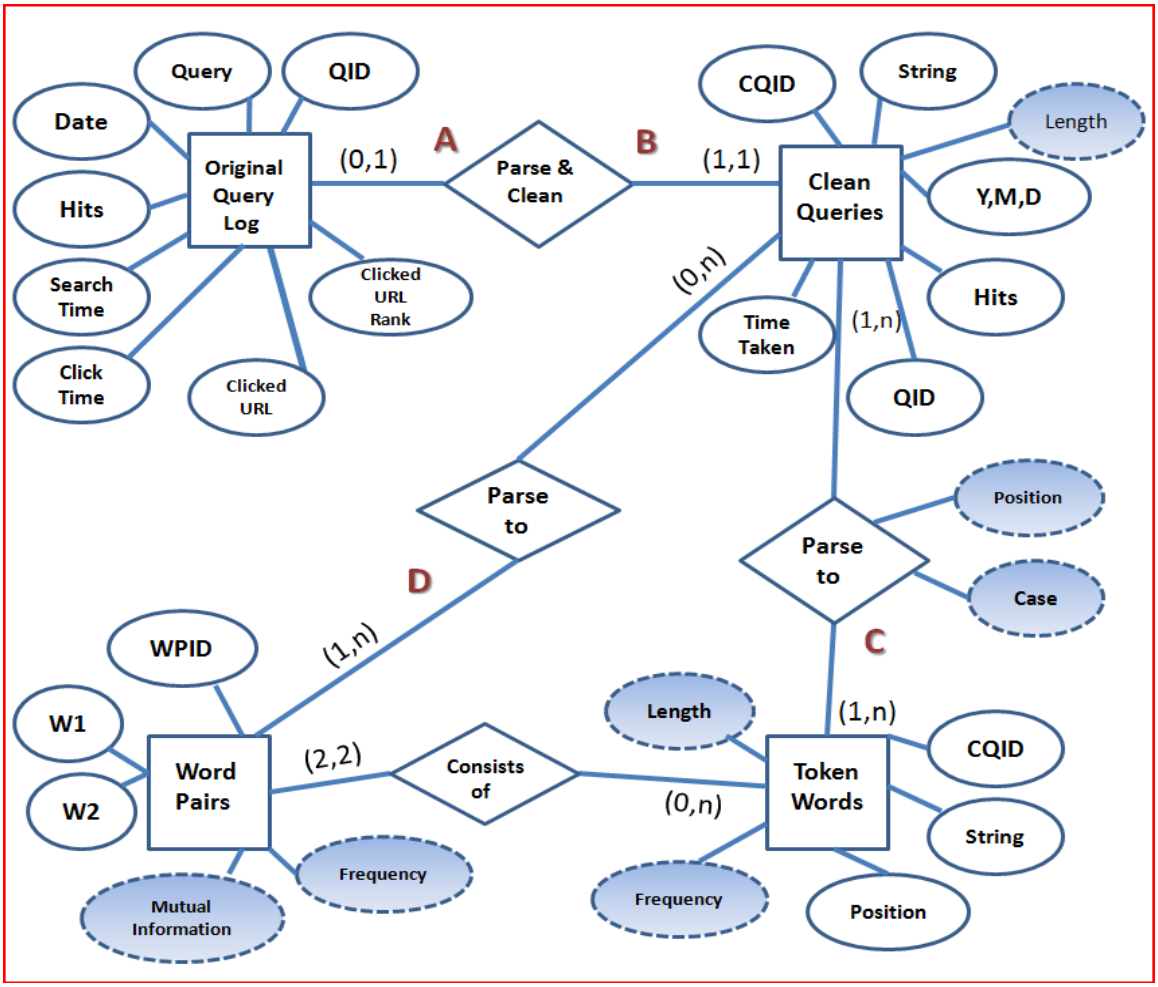


Figure 1. ORNL web query ER model for relational database

The original query log (an Excel file) was cleaned and processed prior to import into Microsoft Access 2007(**A**). Additional processing of the query log included isolating and normalizing the cleaned queries strings by converting all text to lower case. By normalizing the text we remove the distinction between “The” and “the” and “ldrd” and “LDRD”. Normalization did not include removing affixes, for example removing “ing” from parking leaving the word park. Too often this can dramatically change the context of a query. The case normalization had a positive impact on query count and determining accurately high frequency queries and terms. Spelling errors were counted as unique word occurrences. The length of the query string was derived which includes

a count of all characters in the query string to include spaces. The resultant data was imported into a data table called clean queries (**B**).

The next processing step was to tokenize the query strings by parsing the words into word tokens. Each token word retains its Clean Query ID (C_QID) number and is parsed into a single record with an assigned string position number which identifies which position the word occupied in the clean query(**C**). For example clean query number 145 is "business operations calendar". It is split on white space into three tokens (145, business, 1), (145, operations, 2), and (145, calendar, 3). Unique tokens are found by removing all duplicate tokens. Tokens are then spell checked and if required categorized as misspelled words (designated as a 1 or 0 in the attribute "case"). Spell check was used in the first pass through the data. Human review was also required to check for acronym and abbreviation spelling.

The last processing step parsed single word tokens into word pairs (**D**). This processing was necessary to calculate co-occurrence values or mutual information statistics. Mutual information statistics define the relationship between two words, the higher the value the tighter the relationship. Mutual information results can help drive the construction of a next word index or can assist in clustering web queries. If web queries can be clustered and classified into an information category, we can determine what topics searchers are looking for.

RDMS Development

A granular relational data model first introduced by Wang, Berry, and Yang (2003) for Web query analysis was adopted and modified for data mining and analysis of the ORNL query corpus. The relational data structure is optimal computationally for large data sets that have to be repeatedly processed. Such a data structure also provides a rich environment for multifaceted analysis.

The ER model displayed in figure 1 was used to create the ORNL relational data model or schema shown in figure 2. The ORNL query relational database consists of six tables each representing a distinct data topic: "Transaction Log, "Clean Queries", "Unique Queries", "Unique Query Tokens", "Unique Tokens", and "Mutual Information". The "Transaction Log" merely represented all of the original log queries. The MSSQL query log was imported directly into the "Transaction Log" table. The table was assigned

a primary key automatically by Excel (QID). The field names remained static, except for NumResults which was changed to Num_Hits.

The next table created was Clean Queries. The relationship between tables "Transaction Log" and "Clean Query" is one-to-one. The Clean Queries Table contains information about the time each query was submitted (Year, Month, and Date), and the elapsed time which is defined as the time from when the search was initiated until a resultant URL was clicked (Time_Taken). It also contains Tsec which is just "Time_Taken" converted into seconds, numHits, and the query_string_clean. Clean Queries has a primary key of C_QID and a foreign key of QID.

The "Unique Queries" table was derived from Clean Queries and it stores only unique queries. The primary key for "Unique Queries" is C_QID. A Visual Basic (VB) Script was written to count the occurrence of the unique queries inside the Clean Queries Corpus (Appendix B). The counts are stored in a record field called "Query_Clean_Freq" and represent how many times each query occurred in the entire ORNL search corpus. Another VB script was written to determine the number of words contained in each query (Appendix B). Lastly, a field was added called CharCount. This field contains information regarding query length, which is defined as the number of characters contained in the query to include whitespace. This field was added to the table by inserting a data formula in table design mode. The relationship between "Clean Queries" and "Unique Queries" is one-to-one.

Repeat queries happen quite often in query sessions and across a query corpus. Many common queries are submitted by different searchers, and less often duplicate queries are submitted by a single user within a web search session. There are several reasons why this occur, most often the user can't understand why there were no results and in disbelief resubmits the same query. Repeat query submission can also occur by accident. Below is an example of an individual query that was very specific, three keywords, but it contains a typo. The number of hits for the first query was zero, so the user decides next to only type a single keyword, the first word of the previous query "relocation". The user received 1,024 hits on this query, plenty of information, but was it the right information? Next, the user submits the same query receiving again 1,024 results. This was obviously not the information the searcher desired, so they altered their query a fourth time and received only 256 results. The session terminates at this point which means the user finally found their information, their search was interrupted, or they just gave up on the search.

Initial Query…	relocation perdiem	0
Query Revision 1…	Relocation	1024
Query Revision 2…	Relocation	1024
Query Revision 3 and Final Query…	per diem	256

To support vocabulary analysis, the “Unique Queries” had to be broken down into word elements (see figure 3). In linguistics this is called tokenization”. Each unique query was broken down into single text segments, with each child segment retaining its mapping or position inside the “Query_String_Clean”. The relationship between “Unique Queries” and “Unique Query Token” is one-to-many, each having the same primary key C_QID.

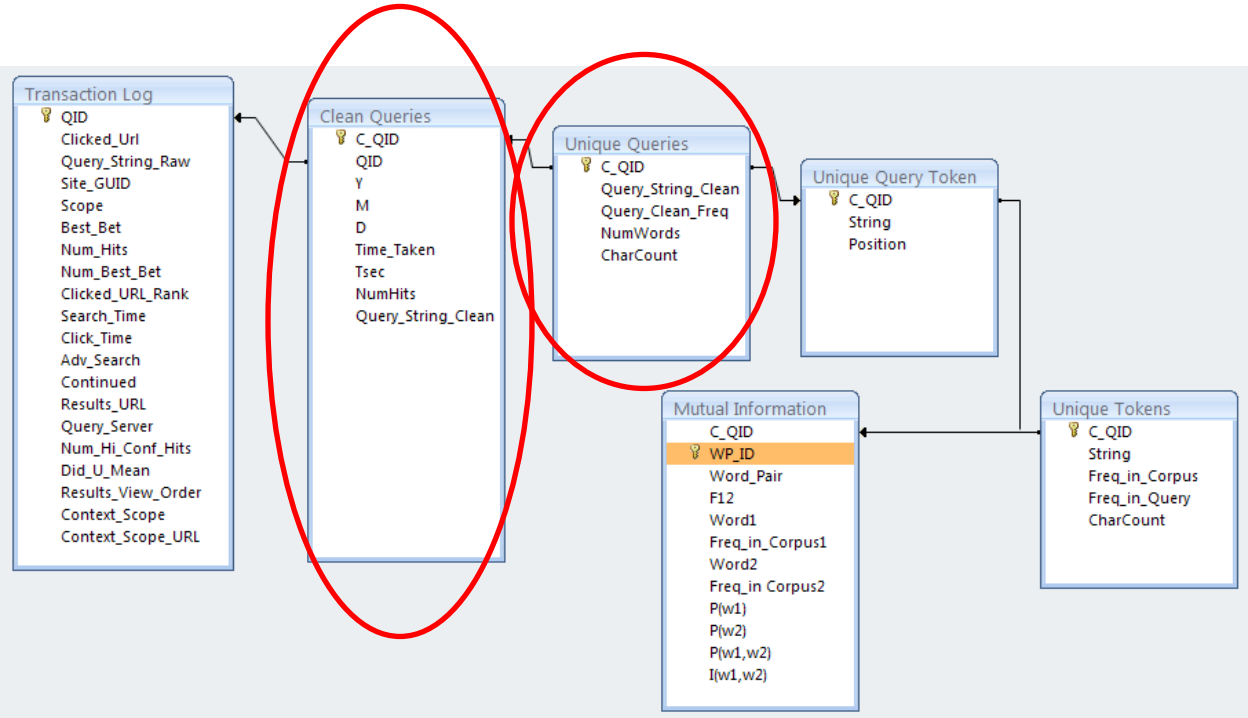


Figure 2. ORNL query database, highlighted tables supported query level analysis

The “Unique Tokens” table was designed to keep track of all the words that comprised unique tokens. It too has C_QID as a primary key. It also contains two counts, “Freq_in_Corpus” which indicates how often the word appeared in the entire corpus and “Freq_in_Query” which indicates how many time the word appeared in a query. For example, the query “Maryville College Maryville Tennessee” has three unique tokens 1) Maryville, 2) College, and 3) Tennessee. The “Freq_in_Query” value of the string Maryville for this C_QID is 2. The “Freq_in_Corpus” value is much higher. The last field in this table is “CharCount” and it describes the number of characters in the unique token string.

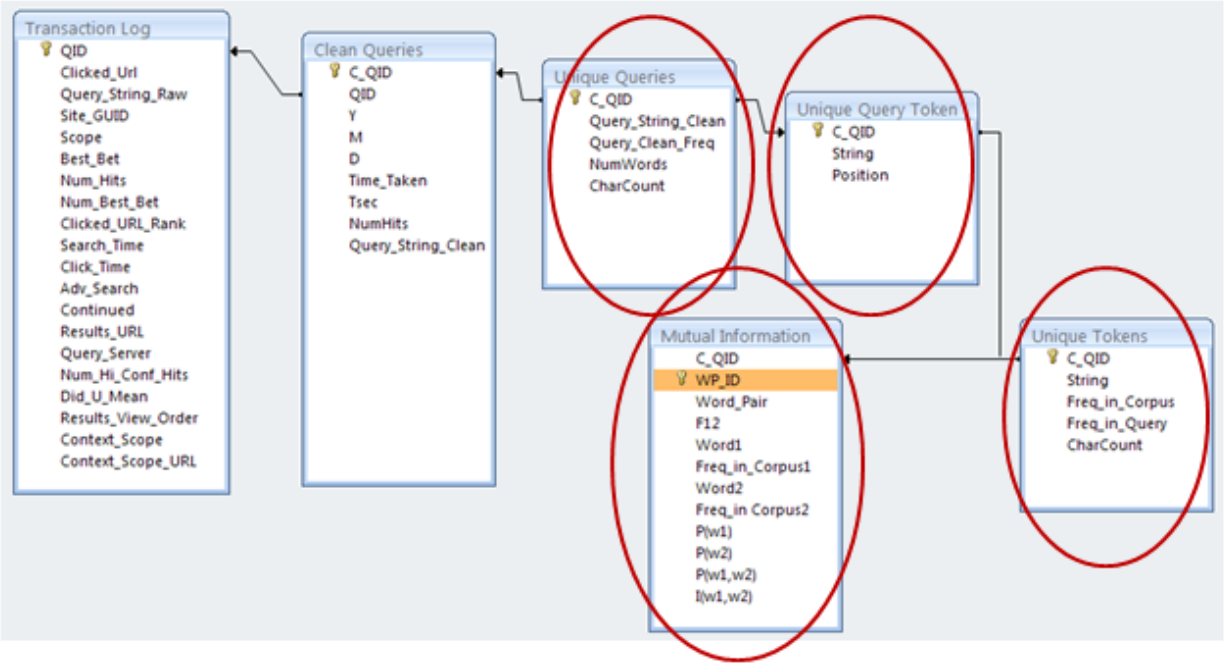


Figure 3. ORNL query database, highlighted tables supported vocabulary analysis

The last table is the mutual information (MI) table which contains information specific to unique word pairs, along with their joint frequency(F12) and a value that describes how closely word pairs are related I(w1,w2). Frequencies for each token in the word pair is also in the MI table and was imported from the "Unique Tokens" table (Freq_in_Corpus1 and Freq_in_Corpus2). The primary key for this table is WP_ID and the foreign key is C_QID. The relationship between "Unique Tokens" and "Mutual Information" is many-to-one.

Methods

Mutual Information Analysis

Word analysis is a subcomponent of Linguistics, the scientific study of natural language. Words are the smallest semantic units that comprise language and it is their patterns of occurrence in text and phrases such as intranet queries either in isolation or as pairs that can help us understand the searchers intent. This analysis focused on word pairs for queries with $2 \leq n \leq 14$, where n is the number of terms or words in the query.

Mutual Information measures the dependence that each word in a word pair has on each other. It is a common measurement used in Information Theory to quantify relationships between words found in text or queries. It is theorized that mutual information values can be used to resolve query translation and query term management. Query term translation may be cross-language or translations within language, for example translating query word pairs to key terms in a synonym index. Query translation may also be referred to as query expansion, which means the query is not replaced by new terms, but rather the query is revised to include new terms or to change the order of the terms to give it new semantic meaning based on its original interpreted conceptual intent. Mutual Information study is also sometimes referred to as collocation-based similarity or co-occurrence, word association study, and bigram analysis.

Mutual Information is also used to measure word association. Word association is very important in the area of information retrieval. Measuring the value of word associations empirically was largely developed by American psycholinguist James J. Jenkins. Psycholinguists study the psychology of language and in Jenkins case he focused on how words are combined to create meaning. A cornerstone study was conducted in

1964 by Jenkins and Palermo establishing subjective norms for measuring word association ratios. The Palermo-Jenkins word association list was subsequently adopted as a standard for testing word association.

In 1990, Church and Hanks challenged the Palermo-Jenkins standard on the grounds that is was very subjective, and proposed measuring association norms with the concept of Mutual Information. Their motivation behind establishing the new measure was increased objectivity and reduced cost.

Mutual Information is an Information Science (IS) theory term developed by Claude Shannon at Bell Labs in the 1940's. The theory is very dependent on *"entropy",* which is just a mathematical way to describe the uncertainty of a single random variable [H(x)]. Conditional entropy describes the entropy of a single random variable affected by another single random variable [H(X|Y)]. Reducing the uncertainty between these values is called Mutual Information (MI). Shannon largely used mutual information for digital communications, specifically signal data processing. He used it for data compression, which is a means to "optimally" store digital signal data.

Mutual information was later adopted for web based purposes. Web-based MI was first introduced by Turney in 2001. The method was renamed to "PointWise Mutual Information (PMI)". The PointWise MI between two words $\mathbf{w}_1$ & $\mathbf{w}_2$ can be described as the log base 2 ratio of the probability of seeing the word pair and the product of the single word probabilities (EQ1).

$$MI(w_1, w_2) = \log_2 \frac{\mathrm{P}(w_1, w_2)}{\mathrm{P}(w_1)\mathrm{P}(w_2)} \quad \text{(EQ1)}$$

In 2003 Wang, Berry and Yang adapted it to linguistic dependency of terms in web query strings. The Mutual Information (*I*) between two words w_1 & w_2 can be described as the natural log ratio of the probable word pair (relative word frequency/ number of cleaned queries (single and multi-word) and the product of the single word probabilities (relative word frequency/ number of cleaned queries). See equation two(EQ2) for definition.

Mutual Information is defined as two points (words) w1 and w2, each having probabilities P(w_1) and P(w_2)(Church, &Hanks,1990). The mutual information formula I(w_1,w_2) used in this study is defined according to Wang, Berry and Yang (2003) to be

$$I(w_1, w_2) = \ln \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \quad \text{(EQ2)}$$

Where: $P(w_1)$, $P(w_2)$ are probabilities estimated by relative frequencies of the two words (see EQ3) and $P(w_1,w_2)$ *is the relative frequency of the word pair (order is not considered, therefore* $P(w_1<w_2)$. Relative frequencies are observed frequencies (F) normalized by the size of the queries (Q):

$$P(w_1) = \frac{F_1}{Q}; P(w_2) = \frac{F_2}{Q}; P(w_1, w_2) = \frac{F_{12}}{Q'} \quad \text{(EQ3)}$$

Although both the observed frequency of a word and the frequency of a word-pair are defined as the number of queries in which the word or the word pair occurs, the total size of the queries has different bases. The size of the queries for words is the actual size of the cleaned queries, while the size of the queries for word pairs depends on the parsing algorithm. First, single-word queries do not produce pairs. Second, a two-word query produces one pair and is counted once, but a three-word query is parsed into three pairs and counted once for each of the three pairs. (That is, the query is counted a total of three times.) In the same way, a four-word query produces five pairs and is counted five times. To account for this phenomenon the adjusted size of the multi-word queries are calculated by the formula:

$$Q' = \sum_{n}^{m} (2n - 3)Q_n \quad \text{(EQ4)}$$

Where: Q_n is the actual size of all cleaned queries including single word queries (60,490), Q' is the adjusted size of all multi-word queries (80,406), and m is the maximum number of words in queries. Similar to the Wang, Berry, and Yang study of 2003, this project

observes all word-pairs, not just the most occurring word pairs in terms of strength. This ensures that the low frequency pairs are not ignored.

The protocol for parsing all qualifying queries(queries with 2 words or more)in prepration for MI analysis was to break queries into adjacent word pairs. Word order is not differented. Two word queries are natural word pairs. Three word or longer queries recieved identical adjacent pairing. For example the query “business operations calendar” is broken into 3 word pairs 1) business operations, 2) operations calendar and 3)business operations. Adjacent pairing in this fashion helped retain the queries semantic intent.

Zipf Analysis

Zipfs Law is used to generally characterize a linguistic corpus, in this case a corpus of web queries. It states given some natural corpus of language, the frequency of any word is inversely proportional to its rank in the frequency table (Kali, 2003). Zipf ‘s Law was used to characterize rank-frequency distributions of unique queries in “Searching the Web: The Public and Their Queries” (Spink, Wolfram, Jansen, &Saracevic, 2001). A double log frequency plot is normally used to plot Zipf statistics, with the x-axis representing log (frequency) and the y-axis representing the log (rank order). The corpus is considered to have “Zipf -like” distribution if when fitted with a straight line, it has a slope of m = -1. It is suggested that Web use follows a Zipfian pattern when plotted on a log-log scale (Nielsen, 1997). Zipf distribution is often used to characterize TLA components such as queries and vocabularies, page requests, and hypertext references. Starting from the upper right oval and working down to the lower left of the graph three circles describe three key word frequencies that occur in a typical corpus’s word distribution (figure 4). In a typical Zipf distribution there are a small amount of queries or words that are used repeatedly (upper left oval), another group which occurs less frequent (middle oval), and a sizeable group of words that are rarely used (lower right oval).

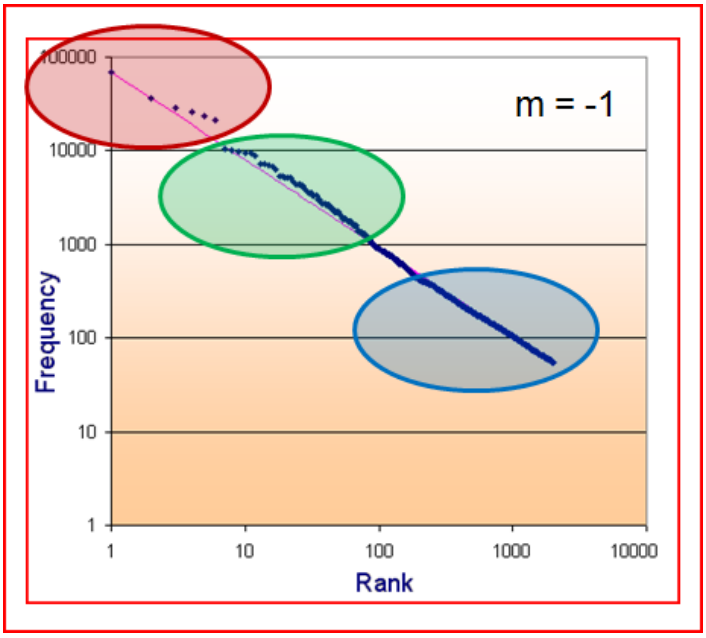


Figure 4. Typical Zipf distribution plot

Approach to Spell Check Query Vocabulary

The misspelled query words were identified using a custom spell checker application. The reference tools in Microsoft Word specifically the dictionary, grammar guide,Thesaurus, and spell checker are very useful in application development. There are two key objects denoted by the read bracket in the Word spell-check procedure (figure 5). The first is "ProofReadingErrors" collection which is a range object containing the proof errors which can be any form of text (word, sentence, paragraph, or an entire document). After SpellCollection is declared as a set of range objects, then SpellCollection is populated with the list of corpus misspelled words. The procedure loops through the word list comparing them with Word Reference Tools. Microsoft provides examples of Visual Basic for Application procedures of Word on the web. It does not take much programming experience to download and invoke the spell-checker procedure, but it does require programming skills to manage the output in a project specific user interface.

```
Private Sub Button1_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Han
dles Button1.Click
    Dim rngRange As Word.Range
    Me.Text = "Starting Word ..."
    WordApp.Documents.Add()
    Me.Text = "Checking text..."
    rngRange = WordApp.ActiveDocument.Range
    rngRange.InsertAfter(TextBox1.Text)
    Dim SpellCollection As Word.ProofreadingErrors
    SpellCollection = rngRange.SpellingErrors
    If SpellCollection.Count > 0 Then
        ListBox1.Items.Clear()
        ListBox2.Items.Clear()
        Dim iword As Integer
        Dim newWord As String
        For iword = 1 To SpellCollection.Count
            If SpellCollection.Item(iword).Text <> TextBox2.Text Then
                newWord = SpellCollection.Item(iword).Text
                If ListBox1.FindStringExact(newWord) < 0 Then
                    ListBox1.Items.Add(newWord)
                End If
            End If
        Next
    End If
    Me.Text = "Word Spelling-Checker"
End Sub
```

Figure 5. Spell-Check Procedure

Chapter 4
Results and Discussion

RQ1: What general search behaviors do ORNL searchers exhibit when searching the intranet?

Most intranet search software automatically collect basic usage information in files called access and query logs. Information that can be mined from the log files describes general organizational search behaviors. The least basic information includes: what browsers the users prefer to search with, times when they search, which external search engines they are more likely to use when searching outside the intranet, what topics they are seeking via page views, how often they are receiving no hits on their requests, and what types of page results are clicked most often. The next few pages describe how this study used information contained in access and query log files to characterize "the general search behaviors of ORNL searchers.

Four distinct SQL queries were developed (Appendix A) to extract data that identified general user search behaviors exhibited by ORNL intranet searchers.

a. What is the size of the ORNL search audience?
b. What interfaces do ORNL users employ most when they search?
c. What types of pages do ORNL users click most often when results are available?
d. What topics do ORNL users commonly search for?
e. When do ORNL users search and what are their search results?

The first and second of the queries focused on determining the approximate size of the ORNL search audience and characterizing what tools they are using for search inside the enterprise. The results showed that ORNL had 8,640 total distinct users between September and December of 2007, with the average visitor staying 12.2 minutes (a). The average unique daily visitor is 4,966. The number of users is defined as the number of distinct IP addresses that submitted search queries. This assumption does not take into consideration that one user may actually submit queries from multiple computers. The latter is highly probable as most ORNL users have at least one desktop computer and one laptop computer. The data is useful as an estimate of the "general size" of our search audience for this study.

The browser chosen by a user for search often has much to do with the platform and operating systems. Examining page hits by browser type shows the top two operating systems for ORNL is Windows XP and Vista, followed by Mac OS (table 4). Windows clearly represents the bulk of computer platforms at ORNL. Since the top platform OS is by Microsoft, one might assume the most popular browser is by Microsoft.

Table 4. Top ORNL computer platforms

		Hits	% of Total Hits	Visits	% of Total Visits
1	Windows XP	41,784,932	83.54%	356,627	78.05%
2	Windows Vista	3,990,070	7.98%	44,565	9.75%
3	Mac OS	2,196,299	4.39%	34,517	7.55%
4	Windows 2000	747,356	1.49%	6,325	1.38%
5	Windows 2003	661,293	1.32%	6,702	1.47%
6	Linux RedHat	263,638	0.53%	3,114	0.68%
7	Linux	108,984	0.22%	1,350	0.30%
8	Linux Ubuntu	32,748	0.07%	476	0.10%
9	Linux Mandriva	11,074	0.02%	165	0.04%
10	Windows NT	4,650	0.01%	53	0.01%
11	Windows 98	1,196	0.00%	8	0.00%
12	Windows ME	624	0.00%	0	0.00%
13	Linux CentOS	388	0.00%	5	0.00%
14	Windows CE	292	0.00%	2	0.00%
15	OS/2	43	0.00%	1	0.00%
16	Windows 3.x	3	0.00%	2	0.00%
17	Windows 95	1	0.00%	1	0.00%
	Subtotal	**49,803,591**	**99.57%**	**453,913**	**99.34%**
	Total	**50,020,657**	**100.00%**	**456,935**	**100.00%**

Browsers are software residing on computer platforms that allows users to access and search a web based search environment like the Internet or an intranet. ORNL employees need to necessarily access the intranet to use applications, to do research, to share information, or to order equipment. The range of intranet based tasks by user is great so browser developers have been diligent in creating browsers with distinct performance characteristics. Two commonly utilized browsers are Internet Explorer, made by Microsoft and Firefox, developed by Mozilla. Other browsers emerging in the search environment is Chrome by Google. Browsers have different levels of speed, reliability, ease of use, information organization, data presentation and formatting, search engine plug-ins, etc. Understanding what browsers your user audience prefers may impact how the intranet information should be organized and presented for search.

The number of distinct browsers reported for search in this study using the logparser2.2 query was 494. The browser count seemed high for the data, but that was because the browser is reported as a brand (Firefox, Internet Explorer, etc.) plus as a specific version, for a specific operating system, for a specific Operating System (OS) version, etc. This was not surprising as most lab workers have at least two computers (a laptop and a desktop) and each likely has a different OS, with varying OS version numbers, browsers, version numbers, etc. Independent of the high number of distinct browsers reported, the results showed that ORNL searchers overwhelmingly use Internet Explorer 7.0(b) as their connection to the SharePoint search server (table5).

The third SQL query selected the top 20 most viewed web links or URL's. The most popular URL's requested were pulled from the Access Log (Table 6). URL queries were removed from the query logs to focus solely on vocabulary analysis. The top ranking URL is the default MOSS page site. The second URL is a handler for managing web site administration requests. The third URL listed simply as "/" is another reference (URL shortcut shows sophistication of users) for the MOSS default page. URL's 6-15 are about SharePoint themes, navigation designs, headers, etc. URL's 16 through 20 are specific to designing web portals and include links on page "layouts" using JavaScript and Cascading Style Sheets (CSS) style pages. The demand for these types of material was not unusual as the MOSS search environment was just implemented by ORNL. Demand was high for SharePoint web design, web administration and general SharePoint information.

Table 5. Browser breakdown for ORNL users

		Hits	% of Total Hits	Visits	Pageview
1	Internet Explorer 7.x	40,588,433	81.14%	349,843	7,787,300
2	Internet Explorer 6.x	4,397,061	8.79%	35,600	801,922
3	Firefox	2,898,732	5.80%	36,454	800,025
4	Safari	780,794	1.56%	17,424	363,135
5	Netscape 7.x	493,210	0.99%	5,829	141,316
6	Netscape 8.x	326,239	0.65%	3,135	86,056
7	Mozilla 1.x	241,007	0.48%	2,598	73,477
8	Microsoft-WebDAV-MiniRedir/6.0.6000	132,486	0.26%	728	93,100
9	Internet Explorer 5.x	45,064	0.09%	260	2,226
10	Avant Browser	20,424	0.04%	251	4,164
11	Windows-RSS-Platform/1.0+(MSIE+7.0;+Windows+NT+6.0)	13,889	0.03%	1,710	13,889
12	MS Frontpage	11,982	0.02%	6	5,134
13	Mozilla/5.0+(compatible;+Google+Desktop)	10,947	0.02%	1,752	10,250
14	Maxthon	9,534	0.02%	123	1,458
15	Konqueror	6,164	0.01%	13	151
16	Microsoft+Office+Protocol+Discovery	3,772	0.01%	2	3,772
17	Microsoft+Office+Existence+Discovery	3,443	0.01%	0	268
18	Opera	1,968	0.00%	26	576
19	OUTLOOK+STS	1,508	0.00%	0	1,334
20	Microsoft+Data+Access+Internet+Publishing +Provider+Protocol+Discovery	1,504	0.00%	5	1,208

The most popular URL's give us an idea of what information is most sought after by users. It also helps to identify whether a single source of information is preferred over multiple sources. Reducing search access to unpopular sites can reduce the overall search time and improve relevant query results. There are many ways to improve the search experience for users within MOSS. The most effective way is to allow users to narrow the search scope (sites that can be searched). As part of scope the search service administrator can establish scope rules that would direct keyword queries to search only certain "Page Types", within the search scope. For example if I am looking for retirement planning applications, I would enter the string in the search box and that would prompt MOSS to search in the scope of "Human Resources" looking only for

interactive web “page types” like those ending in“.js” (java script executable program) or “.dll” (dynamic link library) page type that have the term retirement in them. It would also be prudent to limit page type to those with “.aspx” endings which are active server web page applications.

Table 6. Top 20 URL's requested

		Hits	Visits	% of Total Visits
1	/Pages/	6,554,989	373,830	81.81%
2	/WebResource.axd	3,084,536	311,898	68.26%
3	/	2,062,082	369,149	80.79%
4	/_themes/Lichen/Lich1011-65001.css	1,953,907	408,765	89.46%
5	/_styles/core.css	1,942,545	409,222	89.56%
6	/_themes/Lichen/topnavunselected_lichen.g if	1,897,639	409,644	89.65%
7	/_themes/Lichen/topnavselected_lichen.gif	1,832,020	408,521	89.40%
8	/_themes/Lichen/siteactionsmenugrad_liche n.gif	1,721,644	408,056	89.30%
9	/PublishingImages/ORNL-logo.gif	1,716,469	407,691	89.22%
10	/_themes/Lichen/siteTitleBKGD_lichen.gif	1,647,719	406,810	89.03%
11	/_themes/Lichen/pageTitleBKGD_lichen.gif	1,638,811	406,656	89.00%
12	/_themes/Lichen/navshape_lichen.jpg	1,620,273	406,260	88.91%
13	/PublishingImages/lstbulet.gif	1,609,541	406,660	89.00%
14	/_themes/Lichen/quickLaunchHeader_lichen. gif	1,576,529	403,903	88.39%
15	/_themes/Lichen/pagebackgrad_lichen.gif	1,562,818	406,093	88.87%
16	/common/home_page_styles.css	1,443,211	403,126	88.22%
17	/_layouts/1033/init.js	1,146,511	307,198	67.23%
18	/_layouts/1033/styles/ controls.css	1,082,292	308,363	67.49%
19	/_layouts/1033/core.js	1,036,131	314,232	68.77%
20	/_layouts/portal.js	1,035,712	306,341	67.04%
	Subtotal	48,358,440	N/A	N/A
	Total	50,020,657	N/A	N/A

A fourth query was issued to select data from the access file that helped break down the distribution of page types. The top twenty page types were requested. The output was a ranking of page types and their associated hit count (Table 7). Page types were categorized into clusters based on the functionality associated with the file type. Files types in MOSS are identified by the file extension. Definitions of page file types can be found in Appendix C. Four distinct clusters emerged. They are labeled as Images, Applications, Web Pages and Documents (Figure 6). Figure 6 clearly shows that image pages were selected the most, followed by "application" page types, Web pages and documents respectively(c). Assuming clicked pages contained the information the user was looking for, images seem to be sought most by users and documents sought the least. Page type views are counted each time a visitor actually views or clicks on a web page. The count is irrespective of hits generated, Pages are comprised of sub-files, images, media, etc. Every image in a page is a separate file. When a visitor looks at a page they may see numerous images, graphics, etc. If a web page contains five images, a view on that page will generate six hits on the server (one page view and 5 images). This explains why the images category count is so high. What is so surprising with these results is that documents are not what users look for most. Because image results are statistically over reported, applications likely actually share ranks very close to images in topic seeking importance. Clearly this is a limitation of using page view file types alone for topic identification.

Table 7. Distribution and categorization of page types

IMAGES	*Count*	APPLICATIONS	*Count*	WEB PAGES	*Count*	DOCUMENTS	*Count*
gif	21,578,202	js	5,597,728	css	6,635,513	xlsx	17,826
jpg	2,034,285	axd	3,084,615	htm	11,941	doc	14,587
GIF	613,186	aspx	1,458,828	html	3,721	ppt	8,804
ico	110,293	asmx	75,613			pdf	2,538
png	15,930	dll	14,290			vsd	1,785
		ini	1,263			docx	1,556
	24,351,896		**10,232,337**		**6,651,175**		**47,096**

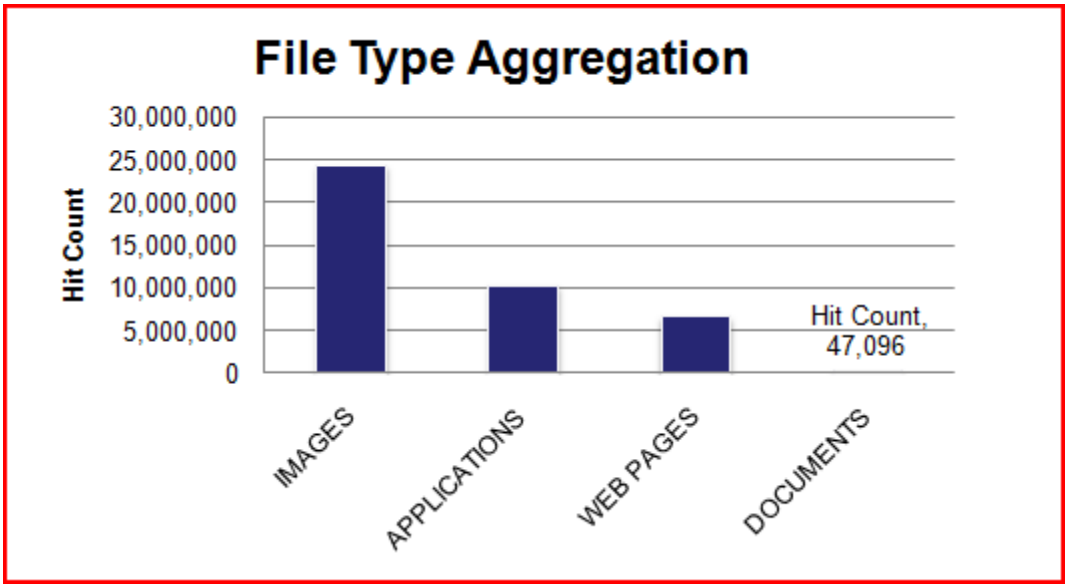


Figure 6. ORNL aggregated page types most clicked

Additional aggregation of information was undertaken to further study what topics were sought by ORNL searchers. A strategy was developed to categorize search requests first using the top 100 "frequent query terms" contained in the query log and second by the top 100 Mutual Information Values. Mutual Information (MI) values indicate how closely word pairs are related. Higher MI values suggest the two appear together more often than not and that their co-occurrence is semantically significant. In nearly every one of the subcategories, 50% of the contributions are in both frequent queries and high MI values.

Fourteen subjective categories emerged from this analysis (Figure 7)(d). The interactive cluster map reveals the term and word pair contribution from both frequent queries and mutual information. Figure 8 shows contributions for the "Business" category. The top 100 frequent query contributions are located inside the diamond and the top 100 MI value terms are located inside the oval. The strongest topic descriptors found in both shapes for the business category are listed inside the tear drop.

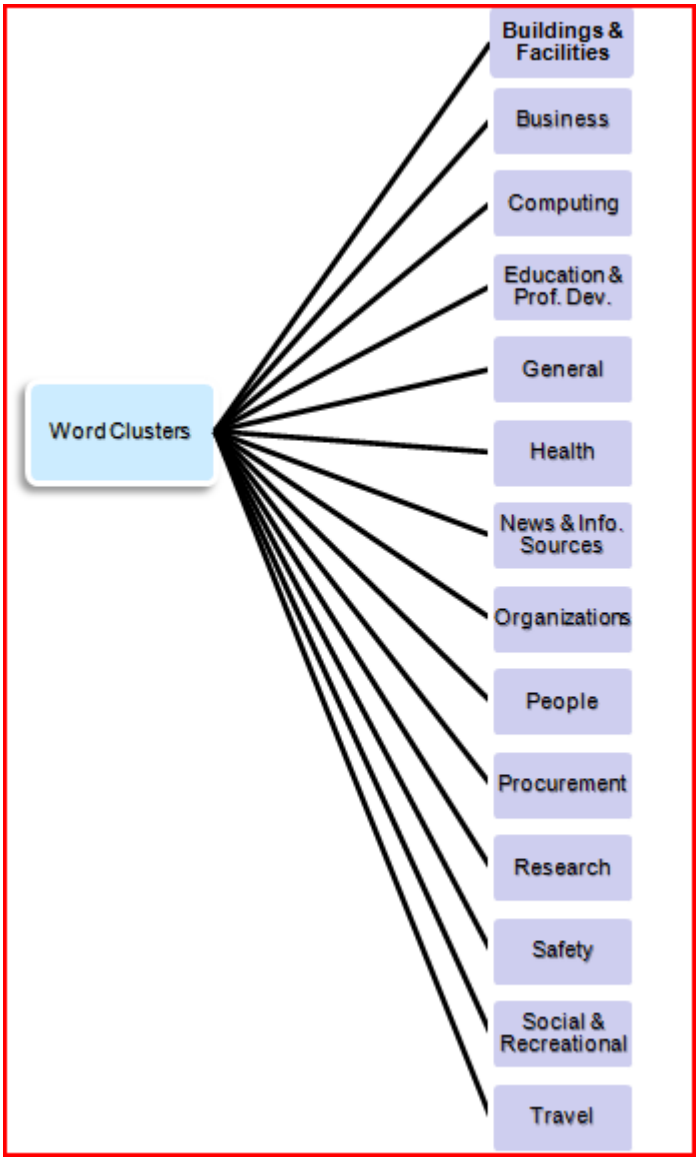


Figure 7. ORNL topic categories

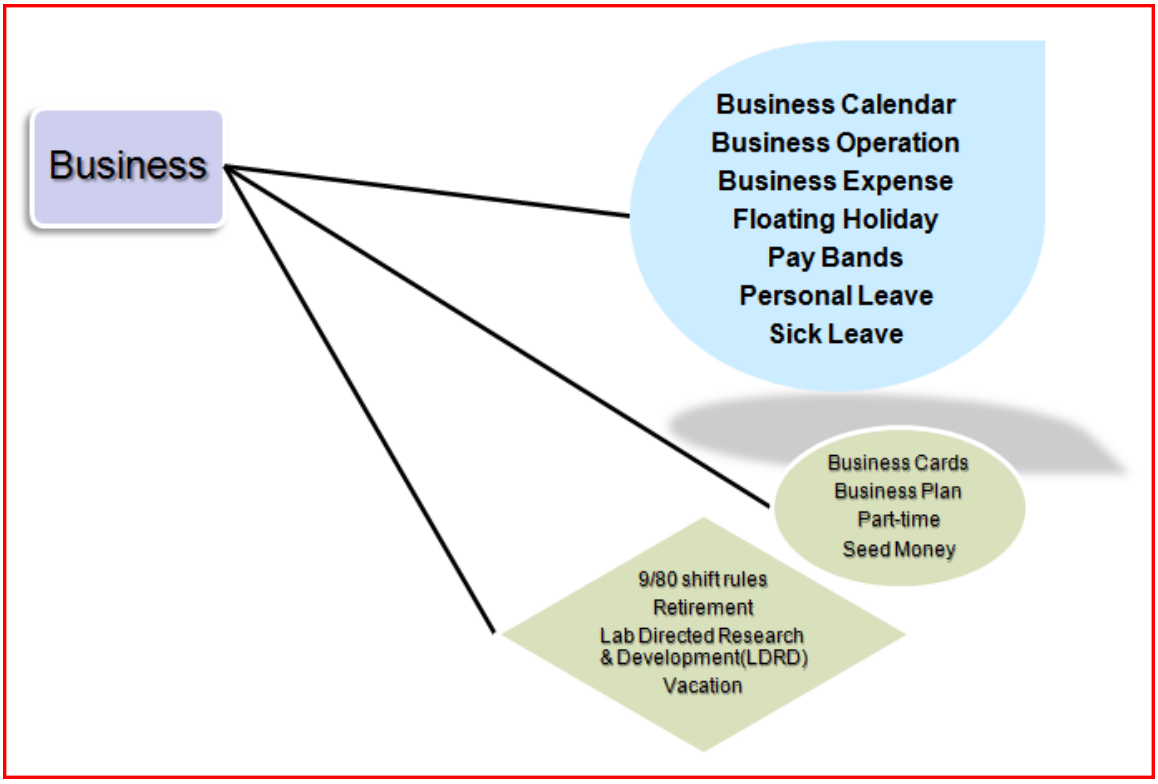


Figure 8. ORNL business category queries

Lastly, general search analysis involved characterizing how often users sought information external to the enterprise, when ORNL users searched, and how often they were getting no hits in response to their requests (e). All of the data required for analysis in this section is contained in the query log file. Knowing when users search allows search server administrators to schedule crawls, indexing and updates when intranet search is the slowest. If users tend to seek a lot outside of the intranet, it may indicate that users are not finding the information they need within the enterprise. Lastly, if users are experiencing a high number or no hits, users will use the enterprise search system less often.

Query counts were plotted by day and month to help determine when users searched. Figure 9 represents all query counts for the entire study period. Note that September has a "broader" distribution plot as the data only covers 14 days. Comparing the latter three months October, November and December one can see an approximate normal distribution of queries for each week. Variability of queries across weeks is small. Mean daily queries during weekdays are between 603 and 952. Weekend daily

query counts are very consistent with a mean from 42 to 47, a significant drop from weekly counts. A month to month comparison of the data shows a bi-monthly harmonic pattern in query frequency. Figure 10 represents a comparison of week three in September (light) and October (dark) of 2007, the plots are nearly identical. Note the significant drop in weekend queries. The most active day for the entire study period is Monday and the least active Saturday. The most active hour of the weekday for search is 2:00 pm – 2:59 pm and the least active is 7:00 am – 7:59 am. The most active search date for the study was Monday, October 15, 2007 and the least active was November 22, 2007(*Thanksgiving Day - a holiday for ORNL*).

ORNL users can search external to the intranet by first selecting any one of a number of commercial search engines like Google, Yahoo, MSN, etc. All qualifying queries in the original query log were classified as internal and external. External searches were grouped according to the search engine chosen (Table 8). Many semantically similar keywords were grouped as necessary, for example keywords assigned to the Google category included Google, google, google search, google web, google.com, google maps, www.google.com and even the Google misspelled ("goggle"). Of the 65,000 qualifying raw queries in the ORNL search corpus, only 197 or .3% were external to the enterprise. Google was clearly the most used commercial search engine with 154 keyword searches. Yahoo, MSN and Netscape had 23, 18 and 1 queries respectively. This statistic reveals during this period users did not often use external search engines suggesting users are largely finding what they need to do their job on the intranet. Also, like most internet users, they prefer Google as their search engine (Nielsen, 2010).

Table 8. Most popular external search engines

Search Engine	Qualifying Queries	Count
Google	Google, google, google search, google web, google.com, google maps, www.google.com and "goggle"	154
Yahoo	Yahoo, Yahoo.com, yahoo japan!, yahoo.jp	23
MSN	MSN, MSN.com, MSN Messenger	18
Netscape	Netscape, netscape	1

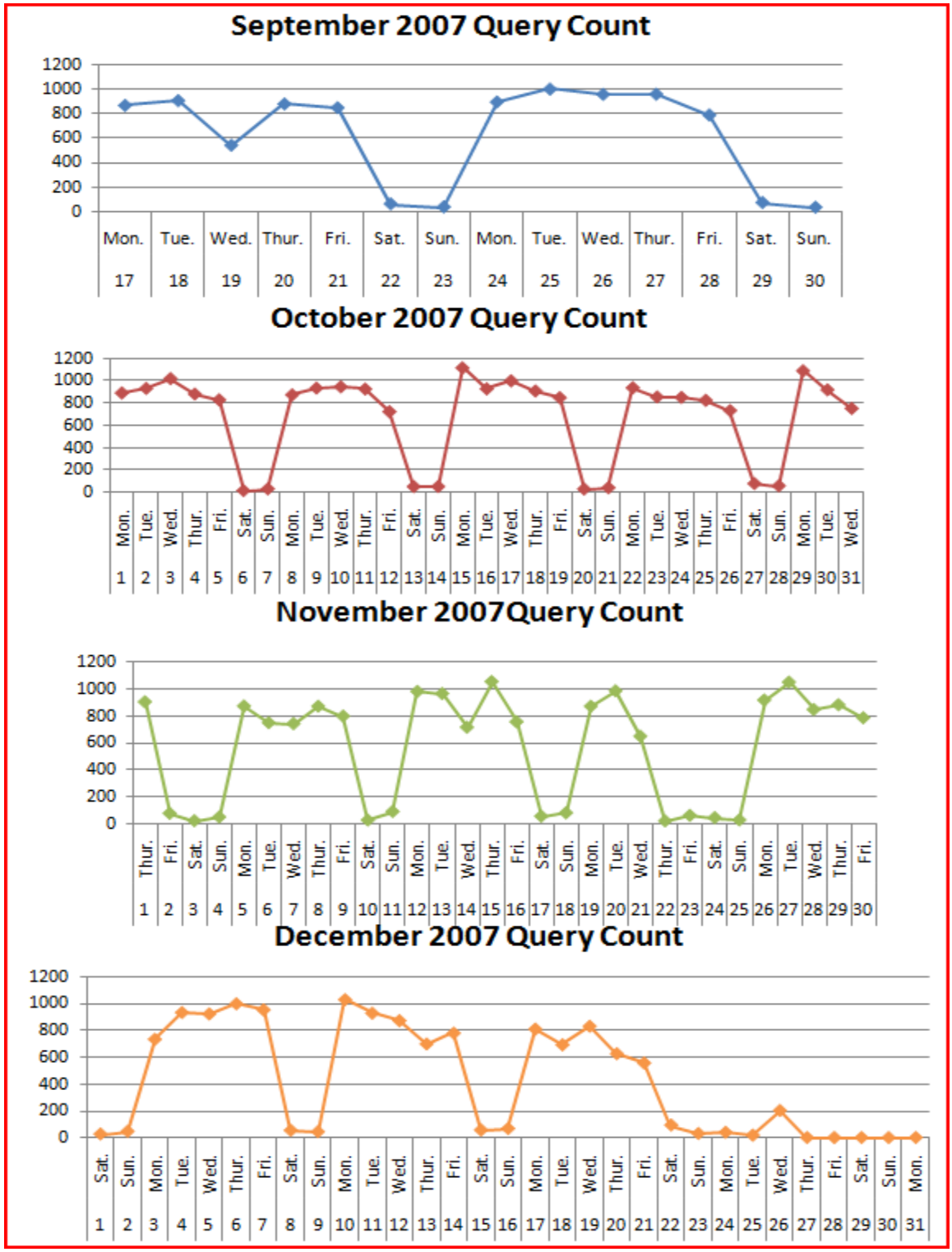


Figure 9. Query counts for each month

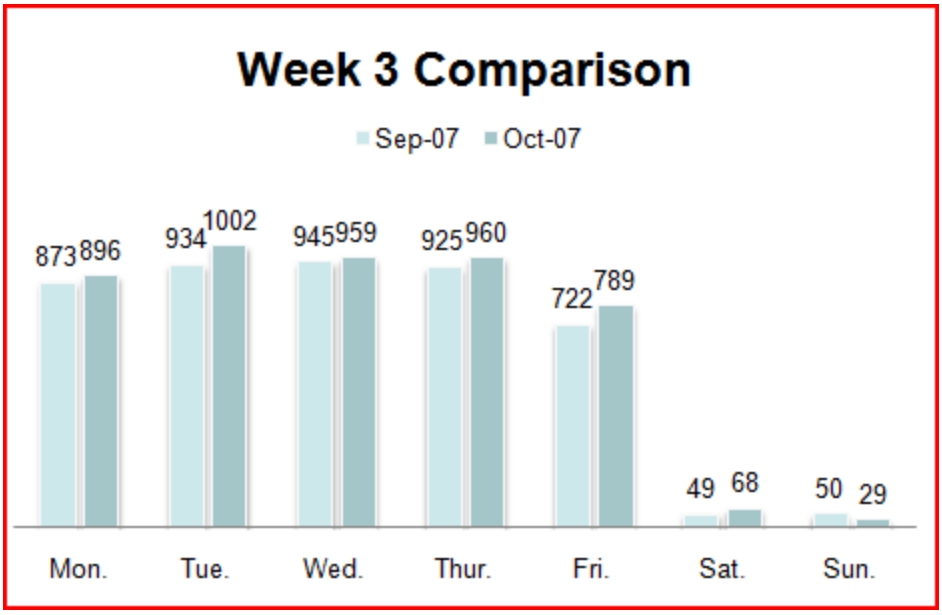


Figure 10. Bi-monthly comparison week 3, September & October 2007

Examining the number of hits associated with queries tells us much about the integrity of our search environment and the search behaviors of our users. A high number of no hits may suggest that users are submitting bad queries or that the enterprise information is not indexed well. Either way there can be a huge gap between searchers needs, their understanding of the system and the system itself. Approximately 16% of all ORNL queries submitted result in no hits. The minimum number of hits is zero and the maximum is 1,448,192, with a mean of 3,394. Clearly the results on average are high for each query. There are many reasons for no hits and most of them can be attributed to bad queries, although some are likely attributable to a non-optimized search system.

The primary reason for the no hits is the use of Boolean operators in a query, for example “+shoe +slips”. The second and third most frequent no hit result is due to misspelled words and extra special characters appearing in the query, for example (‘lost and found). The problems can be corrected by the user becoming more familiar with the search environment and spell checking of queries. The other problems for no hits must be addressed by the search server administrator. This includes making sure information

in the enterprise is indexed to include acronyms, technical jargon, and special language and that broken links are repaired as quickly as possible when found.

RQ2: How do users formulate their queries?

Not only is the search environment different for internet and intranet users, the information spaces differ by the volume, type of content and utilization (Nielsen, 15 Sep. 1997). Combining directorate and division level assets ORNL has 27 technology organizations and 36 Support organizations. Each one of these organizations contributes information to the enterprise. The types of information vary from protocol and policy information, reports, databases, forms, images, videos, to applications. Searchers must rely on these publishing organizations to categorize and index the information appropriately. This includes everything from picking smart web page titles to considering the importance of metadata tags and most importantly "keywords".

An assumption can be made that intranet search should automatically be successful because the publishers and searchers all work together for a single employer. They have a discrete knowledge landscape, so they share special vocabulary, a sense of how the company is organized and even how the organization thinks about publishing and sharing information. This portion of the study looks at *"how ORNL users formulate their queries".* Query formulation is critical to successful information retrieval (IR) within the enterprise.

ORNL user queries to the intranet can be described as one or more terms (keywords) submitted to the MOSS search engine. When an ORNL intranet user enters a query with a keyword(s), the search engine retrieves the web pages that were indexed using the same or similar keywords. The results are ranked reflecting the relevance of the keywords to the user request. If the user receives "no hits" as a query result it means either the information does not exist on the intranet or that more likely the keywords that the user submitted did not match directly or indirectly keywords that index the enterprise information.

The next few pages describe the characterization and analysis of ORNL user intranet queries. Query level analysis involves examining the size and length of the queries, number of blank and unique queries, and their distribution relevant to length and time. As a search server administrator it is important to understand how users "think" the enterprise information is indexed or tagged.

To characterize the ORNL intranet queries, SQL query expressions were submitted to the ORNL “Clean Queries” and “Unique Queries” database tables. A SQL expression is a string that makes up all or part of a parsing statement. Parsing statements were used to select, filter, aggregate, make calculations on, and count query records in the ORNL Access database. This data was analyzed and used to help answer the following questions:

a. What are the most frequently submitted queries?
b. How many ORNL user queries are unique?
c. How many ORNL user queries are blank?
d. What are the lengths of the queries?
e. What is the distribution of the ORNL queries relative to length and time?

Mining the Unique Query table enables the discovery of the most frequently submitted queries. The higher the frequency, the higher the ranking, thus the more popular a query is labeled. Table 9 describes the ORNL top queries for the study period (a). This ranking excludes blank queries which represented 1.62% of all clean queries(c). Note that misspelled words and each form of a single word in or as a query are counted as a unique occurrence. For example in Table 9 the query “pay bands” (plural form) is ranked 3rd while “pay band” (singular form) is ranked first. Similarly, it is interesting to note that “helpline” and “computer helpline”, one a bit more descriptive, actually reference the same information. Table 9 also includes the topic cluster to which the frequent query was assigned. Business and computing information are the most sought after topics. Corporate news and information is a close second.

Table 9. ORNL top 25 most frequent clean queries

RANK	QUERY	FREQUENCY	TOPIC CLUSTER
1	pay band	254	Business
2	file upload	230	Computing
3	pay bands	214	Business
4	pointsec	209	Computing(Application)
5	vpn	206	Computing(Application)
6	spuds	194	Computing(Application)
7	helpline	194	Computing
8	holidays	167	General
9	itms	151	Computing
10	ldrd	148	Business
11	map	148	General
12	awards night	135	News & Information
13	identity theft	132	News & Information
14	sharepoint	129	Computing(Application)
15	powerpoint template	128	General
16	google	114	News & Information
17	email	108	Computing(Application)
18	computer helpline	107	Computing
19	pii	106	Business
20	it training	106	Education/Professional Dev.
21	sns	105	Buildings & Facilities
22	netreg	99	Computing(Applications)
23	blog	98	Computing
24	entrust	95	Computing(Application)
25	safety shoes	93	Safety

The total number of all clean queries was 60,490 with 18,052 unique queries (b). The minimum number of terms amongst unique queries was 1 (excluding blank or empty queries which have 0 terms) and the maximum was 14(d). Interestingly enough, no queries had 13 terms. 95.4% of all queries contain four terms or less (Figure 11). Nearly 48% of the query terms can be classified as unique or special language. There were minuscule occurrences of repeat terms in a query. Repeat words appeared in only five queries:

1. “one call one”
2. “move or move me”
3. “falls creek falls conference”
4. “maryville college maryville tn”
5. “falls creek falls conference participants”

The longest query contained 14 words ("*if you could take any type of it training what you want to take*") and occurred only three times. Most queries contained two terms. The second most queries contained a single term.

The analysis also looked at the top 10 most frequent queries across single, double, triple, and quadruple word queries. The analysis examined growth of the most popular queries, not necessarily queries expanded by a single user. This analysis let us determine if any patterns exist in the way users modify queries for the most popular information requests. Table 10 describes the top 10 queries with N-words taken from the Unique Queries table. An interesting pattern emerges as you compare queries across the N-word categories. More adjectives are added to the queries and often it is the word ORNL. Single word queries are more likely to be acronyms, technical jargon, or names. As the queries lengthen they include more natural language. The corporate name qualifier increases in popularity.

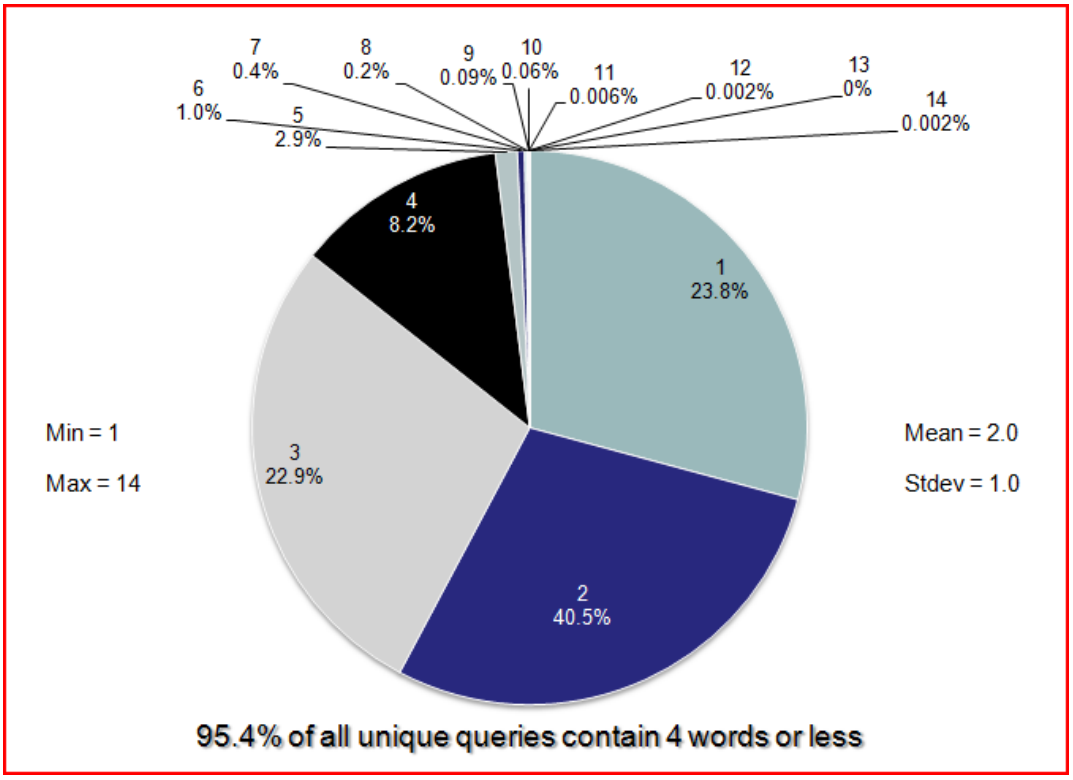


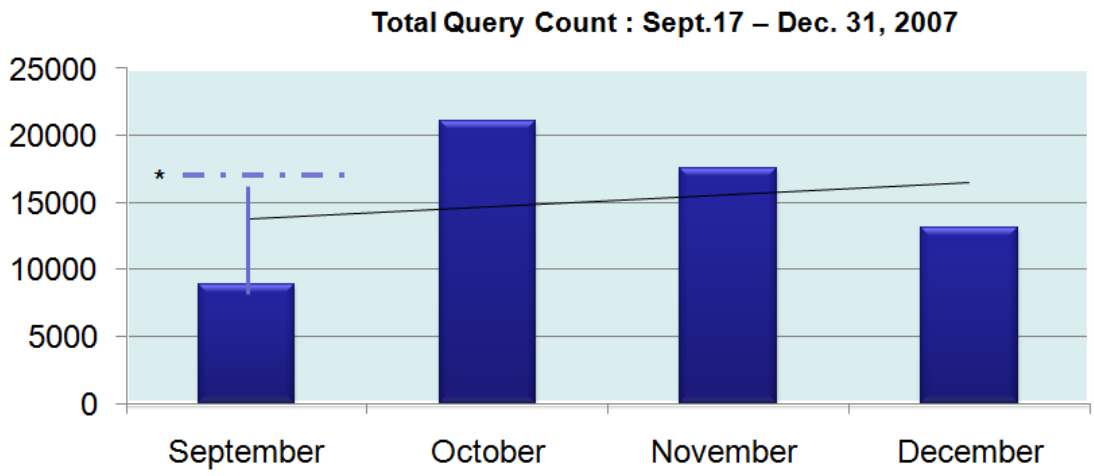
Figure 11. The distribution of words in Unique Queries

For queries of three words or longer the word "ORNL" appears approximately 8% of the time. No hits decrease amongst two word queries, and significantly decrease amongst three word queries. Four words may or may not help improve findability. The number of no hits decreases with four words, but so does the query total. Statistically it is hard to quantify as the number of longer queries occur infrequently.

Lastly, temporal query distribution was examined to see if the most popular terms changed over time and if users search frequency varied over the 3.5 months of the study period (e). For search frequency all of the clean queries were considered. Figure 12 depicts the total clean query counts from September through December of 2007. *Note, only 14 days from September are represented in the study, the dashed line over September "estimates" the normal amount of query submissions. The remaining months are fully represented. The trend line indicates that query counts gradually rise from September to December. The rise in activity can be attributed to the start of the new fiscal project year which begins October 1st, 2007(FY 2008). The decline from November to December is due to the Thanksgiving and Christmas holiday breaks. The general distribution over all months is fairly uniform.

Table 10. ORNL top 10 queries with N-words

	Single Word	Two Words	Three Words	Four Words
1	pointsec	pay band	business month calendar	ornl credit union deduction
2	vpn	file upload	day of science	ornl staff salary range
3	spuds	pay bands	business operations calendar	ornl federal credit union
4	helpline	awards night	cell phone stipend	integrated facility disposition project
5	holidays	Identity theft	ornl pay bands	race for cure captain
6	itms	powerpoint template	controlled business expense	ut battelle powerpoint template
7	ldrd	computer helpline	continuity of operations	authorized personnel only form
8	map	It training	site access training	environmental management system photo
9	sharepoint	safety shoes	payroll run times	electrical protection glove specifications
10	google	business calendar	conflict of interest	integrated facilities disposition plan



MONTH	#Queries	Min	Max	Mean	Mean(weekdays)	Mean(weekends)
September	8,836 *(15,593)	29	1002	631.14	951.8	46.5
October	21,037	13	1115	678.61	899.91	42.38
November	17,524	18	1052	562.8	750.36	47
December	13,093	0	1033	422.35	603.38	42.2
TOTAL	60,490					

Figure 12. ORNL total query count distribution from September to December 2007

Another way to examine temporal query trends is to plot the frequency of a random sampling of popular queries over the length of the study (e). Figure 13 show the frequencies of 10 unique queries (ranks 2 -8) for the ORNL searchers. The terms are listed in random order. At first glance most words follow the same popularity pattern except for map, holidays, and pay bands. The term holidays trends upward through middle December which makes sense, everyone wants to know the holiday work schedule. The query "pay bands" which describes the ORNL compensation structure rises steadily until Christmas break. Most companies that use the pay band structure system make updates at the end of November or the first part of December. Employees typically receive raises in January when available, so managers are busy submitting information to Human Resources on wage increases, allocations, and promotion which may affect band changes. The most distinct anomaly is the query "map" and I am not sure why that term rose and fell so sharply in November. It may be a result of the many construction projects taking place on the ORNL campus at that time, or it could be due

to a large influx of visitors for experimentation or even a conference event. I do know map requests typically go up sharply in the summer due to a large influx of student workers. Explanation of the true trends of temporal queries can best be explained in user search survey and searcher interviews.

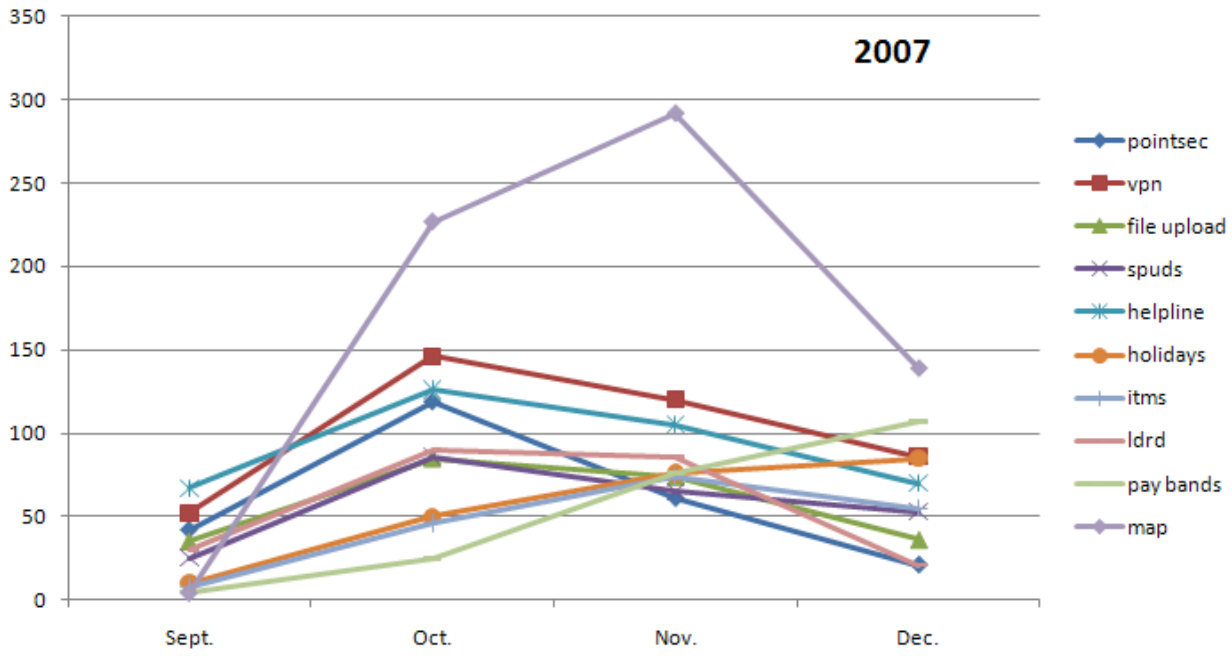


Figure 13. Temporal frequency sampling of ten unique queries

RQ3: What are the Characteristics of the User Vocabulary?

Whether we are discussing the Internet or the Intranet, incorrect search vocabulary is likely the number one cause of failed search inside intranets. Second is the knowledge gap created by enterprise publishers because they fail to index their information with terms or words with which users are most familiar. Vocabulary in this study refers to all the words that ORNL searchers are familiar with, which depending on the role of the user may not include business oriented vocabulary. While ORNL users in some measured capacity share special domain vocabulary specific to the Department of Energy, they have their own unique vocabularies. A user's special domain vocabulary grows over time, so a new employee or an employee who has only worked for a few years at ORNL will have a smaller domain vocabulary with which to choose search terms. User vocabulary varies from person to person and is influenced by language, education, profession, job function, and background.

Once ORNL users determine they have an information need, they head for the intranet site. Once there they can either navigate to a piece of information or search for it. If they are familiar with the exact vocabulary to address an information need and it resides in the web site index they can simply click on the word link item. If they are not familiar with the name or location of the information they are looking and it is not in the site index, they must search for it. Once users have selected search, they determine their search vocabulary by using words that are most familiar to them. The implication of improperly selected vocabulary is no hits or non-relevant results.

The next few pages describe the analytical characterization of the ORNL user vocabulary. The analysis includes a close examination of the type and length of terms. It discusses with what frequency ORNL queries contain vocabulary defined as abbreviations, acronyms, and misspelled words. The characterization looks at both the frequency of standard search terms and terms that are specially referred to as commonly occurring connective words like to, and the, etc. These words are also called stop words. Lastly, term pairs and their frequency are examined.

SQL queries were submitted to the "Unique Queries", "Unique Tokens", and "Mutual Information" tables. The data gathered and analyzed from these tables helped answer the following questions about ORNL user vocabulary:

a. What is the length and distribution of ORNL unique terms?

b. With what frequency do ORNL user queries contain acronyms, abbreviations, and misspelled words?
c. What is the frequency of common stop words?
d. Are there terms that occur together frequently ("term co-occurrence")?

The analysis found 9,804 unique terms in the ORNL corpus (a). Unique terms represent the extent of ORNL user vocabulary submitted as keywords for intranet search. Unique terms represented only 7.8% of all terms in the unique query corpus, including both popular query words and stop words (table 11). The relatively low number implies the extent of the ORNL vocabulary is relatively small. Approximately 80% of top ranked unique terms held string position 1, indicating it was the first term in the clean query. The remaining terms held position 2 and were always a stop word or non-discriminating term. The number of characters for each unique term was binned and plotted to see the distribution of characters (Figure 14). The minimum character was one; the maximum was 29 with a mean of 7.0 and a standard deviation of 3.0. Approximately 98% of all words contained 13 characters or less (a)

Table 11. Popular ORNL query words

Rank	Unique Term	Query Position	Frequency
1	ornl	1	1636
2	or	1	1139
3	a	1	1100
4	and	2	1000
5	plan	1	936
6	of	2	908
7	training	1	812
8	pay	1	806
9	map	1	781
10	business	1	707
11	it	1	600
12	web	1	574
13	help	1	573
14	safety	1	559
15	computer	1	548
16	phone	2	544
17	office	1	543
18	ice	1	538
19	file	1	509
20	holiday	1	499
21	security	1	486
22	report	1	475
23	calendar	1	456
24	template	1	451
25	email	1	442

* based on total cleaned queries

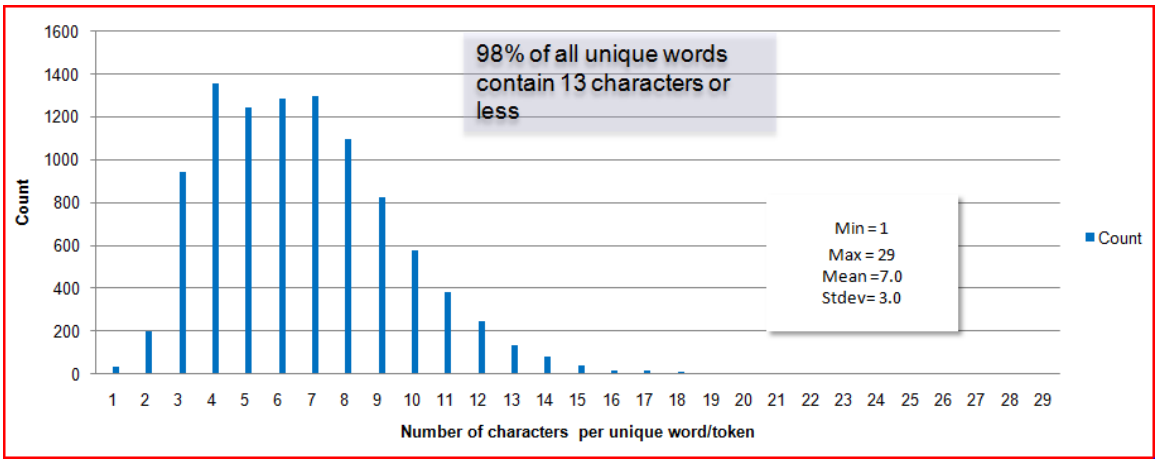


Figure 14. Distribution of word length associated with ORNL unique query words

Next, Zipfs law was applied to ORNL word frequencies of the top 100 and the top 2000 unique words. Zipfs law defines the probability of occurrence of words in a given natural language corpus. The occurrence or frequency of words starts high and then tapers off. Thus, a few words occur very often while many others occur rarely. Zipf curves follow a straight line when plotted on a double-logarithmic scale. If the frequency distribution approximates a straight line with a slope of -1, it can be characterized as a Zipfian distributed corpus. As figure 15 depicts the ORNL unique term corpus is very close to a Zipfian distribution with the first 100 words, and at -1 when using a much larger sample of words. Zipf distributions have even been applied to corpus letters. The ORNL letter corpus did not meet Zipf criteria. This is likely due to the large amount of acronyms and abbreviated text comprising ORNL queries. ORNL words were correlated with an ORNL acronyms list as well as other government acronym lists. Approximately 26 % of unique words were defined as acronyms, 12 % of were abbreviations and misspelled words accounted for 9.6% of all unique words (b). Figure 16 displays some examples of acronyms, abbreviations and misspelled terms submitted in ORNL queries.

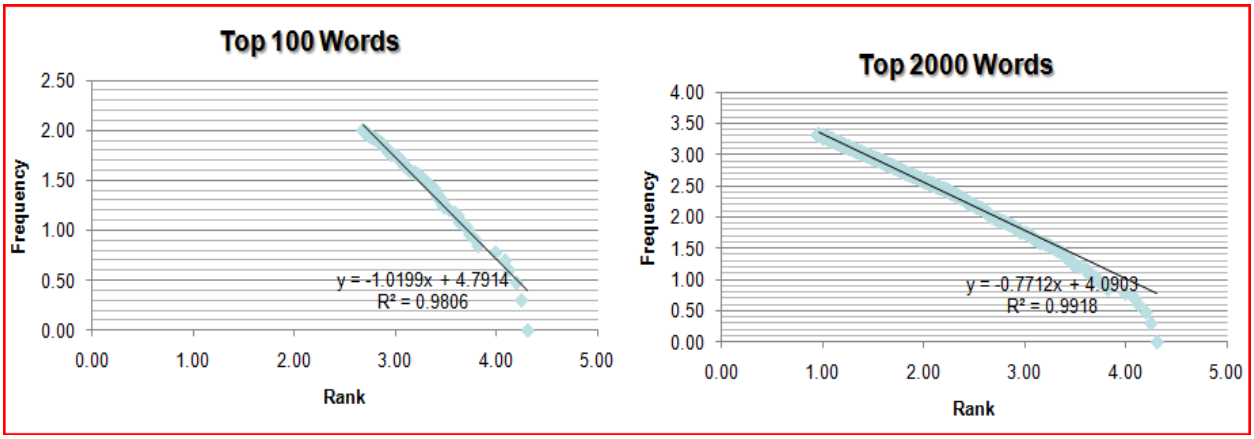


Figure 15. Zipf distribution plot of the top 100 and top 2000 words

ACRONYMS	ABBREVIATIONS	MISSPELLED
SNS	rad	licence
VPN	r&d	liscense
HFIR	mac	acess
SBMS	es&h	requirments
UCAMS	ip	occurence
LDRD	f&o	managment
UT	ig	confrence
PII	wep	modelling
ORISE	dns	buisness
EPO	postdoc	standdown
MRF	fy	absense
JICS	tru	berylium
MHP	telecon	maintainence

Figure 16. Sample of unique or irregular vocabulary

The vocabulary used by search engine users can impact a search engine in two ways. The first way is the reduction of the index by pruning useless terms (i.e. stop words, unpopular queries). The second way is through caching of frequently used terms or queries to improve performance (Zien, 2000). This section of the analysis involves stop words. Stop words are very common words that have no discrimination power. Stop words include words like “or”, “and”, “the”, “of”, etc. Stop word lists are useful in creating smaller indexes for information retrieval. They can be built by sorting the unique terms by frequency and selecting the most frequent. In a typical data base of research documents at ORNL, the term “ornl” would be considered a stop word. Unfortunately ORNL searches find the term popular, so it should really not be removed from the index. The word “or” was the highest frequency stop word with 1,139 occurrences, followed by “a”, “and”, and “of” respectively. All words had frequencies above 909(c). Again, this is another instance where it is important to understand user behavior in intranet search. There was a problem reconciling “it” and “or” as stop words because they can also represent the acronyms for information technology (IT) and Oak Ridge (OR).

The last item for analysis in this section is term co-occurrence or how often two words appear together. To determine term co-occurrence we can combine all of the unique terms in the corpus into unique random word pairs and then count the number of times the word pairs actually appear in the clean query corpus. This value is called the pairs joint frequency (F12). For computing term collocation strength which is a measured quantity to characterize how much random words occurring in pairs depend on each other, we use the joint frequency, individual word frequency (F1 and F2) and token totals(Q and Q’). In web query analysis this is also referred to as Mutual Information statistics. Ultimately, Mutual Information statistics identifies the strength of word association assuming mutual independence of terms within a pair (Wang,et al., 2003)

Table 12 lists some select mutual information values from the ORNL word pair corpus. The table is ranked by joint frequency (F12) (d). A mutual information value (I) located in the far right hand column describe the relationship strength between word one and word two. The “I” values for the ORNL corpus range between -3.61 and 10.72. The average “I” value is 4.0 with a standard deviation of 2.67. Higher “I” values infer a strong association between words. Unfortunately high “I” values in this corpus can represent either terms that often appear together with words that both have low

individual frequencies(low F1 or F2), like ranks 10954, 55, and 56 respectively in table 12 or words that appear together once in a query and has spelling mistakes like ranks 10300 and 10331 in table 12. Conversely, the top 15 ranked word pairs have lower "I" values indicating that those word pairs are weakly associated. The low "I" value is due to each word in the pair having extremely high frequencies. This implies that both the "I" value and the joint frequency (F12) contributes equally to characterizing word pair strength and thus its importance. Notice some pairs share the same F12 value, but have different "I" values (see table 12 word ranks 1902 and 1904). Pairs with low joint frequencies (F12) and high counts for F1 and F2 exhibit negative "I" values implying weak association (see table 12, ranks 1902,1903 and 1904).

Table 12. Select ORNL mutual information values

Rank	Word1	Word2	F 12	F1	F2	I
1	pay	band	254	806	376	3.64
2	file	upload	230	509	405	3.92
3	pay	bands	214	806	258	3.84
4	awards	night	187	277	199	5.04
5	cell	phone	183	332	544	3.83
6	business	month	173	707	219	3.93
7	computer	help	157	548	573	3.12
8	identity	theft	135	135	143	5.76
9	powerpoint	template	128	232	451	4.02
10	computer	helpline	115	548	368	3.25
11	it	training	113	600	812	2.35
12	ornl	logo	111	1636	305	2.31
13	credit	union	105	169	120	5.46
14	conference	room	104	356	191	4.24
15	safety	shoes	93	559	177	3.9
68	controlled	business	51	60	707	4
69	help	desk	50	573	203	2.97
1902	ornl	template	5	1636	451	-1.18
1903	ornl	awards	5	1636	277	-0.69
1904	ornl	energy	5	1636	265	-0.65
3912	mentor	protege	3	17	4	7.6
3913	carbon	monoxide	3	17	3	7.89
10300	webwide	seminer	1	2	2	9.34
10331	Populus	trichocarpa	1	2	1	10.3
10954	moon	phases	1	1	1	10.72
10955	exxon	aromatic	1	1	1	10.72
10956	fumed	silica	1	1	1	10.72

Note: F12 was calculated using Q' value F12 has 92 distinct values

After reviewing initial results, it seemed logical to look more closely at the very popular word pair “pay band” because it had a relatively low “I” value of 3.64(table13). This pair was ranked number one in unique query with 254 hits as well as the number one in word pairs with the same frequency. All of the word pairs containing the word pay or band (42 in total) as the first or second words were retrieved (table13). The word pay is combined with 43 words and the word band is combined with only 5 words. Together these 42 unique pairs represent semantically the entire set of words that users think of when they search for information regarding the central subject of “pay”. It is useful to note that none of the “I” values go above 4.03, the highest “I” value amongst the “pay band” set.

Most of the “I” values in this range contain at least one misspelled word in the pair (table 13 ranks 6502-6504). Misspelled words are counted as unique words in the corpus and occur very infrequently. The misspelled word accounts for the low frequency of one word in the pair. The high “I” value in these pairs indicates that these low frequency words are strongly associated with a high frequency word such as “pay”. This association rule does not hold true throughout the set. Comparing other high frequency words with low frequency words can yield very low “I” values (see table 13 ranks 4314 and 6499).

Overall, table “I” values seem to make sense for this word pair subset as the more popular pairs tend to have higher “I” values. The query corpus was revisited to investigate the number of hits associated with a random sampling of the word pairs. Pairs with higher “I” values generally got fewer hits, so I am assuming these hits were more relevant results. I verified the URL Click rank that most users chose and it was the first result (the highest rank). As the “I” value decreased on the pair, the number of results grew and URL click ranks that were chosen if selected at all were much lower. This directly supports the value of characterizing word pairs using mutual information and word popularity. Generally word pairs with higher mutual information values tend to be strongly associated (Wang, et al., 2003)

Table 13. All word pairs sets involving the words "pay" and "band"

RANK	WORD1	WORD2	F12	F1	F2	I
4319	pay	stubs	2	806	2	4.03
6502	pay	bankds	1	806	1	4.03
6503	pay	dispersment	1	806	1	4.03
6504	pay	rool	1	806	1	4.03
4	pay	bands	214	806	258	3.84
48	pay	stub	62	806	76	3.83
2	pay	band	254	806	376	3.64
4318	pay	grades	2	806	3	3.62
695	pay	ranges	12	806	20	3.52
1972	band	charts	5	376	20	3.41
694	pay	scales	12	806	24	3.34
6501	pay	stab	1	806	2	3.34
193	pay	dates	30	806	63	3.29
4477	band	three	2	376	14	2.85
110	pay	date	39	806	136	2.78
412	band	chart	18	376	136	2.77
407	pay	scale	18	806	68	2.70
2451	pay	zones	4	806	24	2.24
73	pay	schedule	49	806	299	2.22
2450	pay	increase	4	806	26	2.16
4317	pay	ladders	2	806	16	1.95
3175	pay	levels	3	806	27	1.83
994	pay	rates	9	806	86	1.77
6500	pay	dues	1	806	10	1.73
1928	pay	grade	5	806	57	1.60
2449	pay	calender	4	806	50	1.50
4316	pay	period	2	806	27	1.43
4315	pay	classification	2	806	31	1.29
4476	band	ladder	2	376	76	1.15
4314	pay	guidelines	2	806	37	1.11
6499	pay	monthly	1	806	37	0.42
3174	pay	chart	3	806	136	0.22
4475	band	hr	2	376	221	0.09
1571	pay	in	6	806	430	-0.24
872	pay	day	10	806	838	-0.40
4313	pay	roll	2	806	276	-0.90
6498	pay	pal	1	806	153	-1.00
4312	pay	travel	2	806	375	-1.20
6497	pay	information	1	806	262	-1.54
6496	pay	days	1	806	280	-1.60
6495	pay	points	1	806	289	-1.64
6494	pay	calendar	1	806	456	-2.09

RQ4: How do ORNL results compare to the published studies?

This section of the study describes findings that come from comparing ORNL TLA results with a specific set of published study results. Again, the criteria for selecting TLA methods to apply to the ORNL data set was that it had to be focused on context analysis using traditional internet TLA methodology on "intranet like" search sites. Three primary publication results were compared to the ORNL results: 1) "Mining Longitudinal Web Queries: Trends and Patterns" (Wang, et al, 2003), hereafter referred to as the UTK study, 2) "What are you searching for? A content analysis of intranet search engine logs" published in 2006 by Stenmark, hereafter referred to as the Volvo study and 3) "Intranet Users Information-seeking Behaviour: an Analysis of Longitudinal Search Log Data", published in 2006 by Stenmark and Jadaan , hereafter referred to as the SwedCorp study. Summaries on the intent of each study can be found in the Literature Review section.

The first results compared will be general search patterns and topics, specifically when users searched and what they searched for. The frequency of query counts for the UTK study rose from September and fell off again in December. Query frequency was dramatically reduced during holiday breaks and school closures. Similarly, ORNL trend lines for query frequency indicate a gradual rise from September to December, however there is a granular decrease from October to December. Query frequency drops dramatically for ORNL holiday periods. For more specific usage patterns, both report Monday through Thursday had substantially more queries submitted than Friday, Saturday and Sunday. For ORNL Friday submissions were only slightly lower (- 12% on average) than preceding weekdays. The weekends were consistently much lower than weekdays (-95%). The UTK study shows highest values on Mondays and Tuesdays, ORNL on Mondays.

The Volvo study used most popular terms and a card sorting process to categorize what information users were most seeking. Nine categories resulted in the process: Human Resources, IT, Organizational Issues, Places, Products, Projects, Standards, Computer Systems and Applications, and Miscellaneous. Computer Systems and Applications were the dominating categories for Volvo searchers, while organizational issues and documents were of least interest. The ORNL study used popular page type selection and MI values to determine what types of information searchers most sought. Images and applications were sought most and documents

sought least, with business and computing information being popular topics. The UTK study did not seek to classify topics, so no comparison can be made there.

The next set of results compared were at the query level which includes no hits, empty or blank queries, average query length in terms(words) and characters. The number of no hits for UTK was greater than 30%. The primary explanation behind the large number was that the UTK Search engine used Boolean "and" for multi-word queries. The secondary cause was that users didn't understand the query system requirements. ORNL reported no hits only 16.34% of the time. The reasons for ORNL no hits were first the improper use of Boolean operators, and second misspelled and extra characters added to query terms. Other reasons ORNL no hit queries included external search mistakenly submitted to the "internal search engine", incomplete queries, and technical jargon or acronyms submitted as queries and unfamiliarity with the search engine. Misspelled words for the UTK study was 26% while ORNL queries were misspelled only 9.6% of the time. Empty or blank queries were few for ORNL (1.62%). Similarly, UTK reported even fewer at .9% and the SwedCorp study reported the least at .7% .The average query length in words for the ORNL queries was three and for the UTK study was two. The SwedCorp study also reported few queries containing more than three words. For ORNL the number of "no hits" significantly decreased among three word queries. Four words or more did not seem to improve "findability". The average length of the terms in characters was seven for both the ORNL and the UTK study. The longest query for ORNL was 14; for UTK it was nearly double that at 26.

The last set of results to be compared is vocabulary, which includes the number of unique queries and terms, top words and stop words, as well as word association trends when applicable. For UTK, most queries contained unique text strings and only a small number of queries were frequently submitted with approximately 6.7% common to all years studied. Similarly, ORNL unique terms represented 7.8% of the total term count. The small number implies ORNL vocabulary is relatively small, a few terms repeated often. The average number of unique terms reported was much higher (37.1%) in the Volvo study. For intranets, top words like corporate names were identified as possible stop words. In the ORNL study ," ORNL" was ranked number 1 in frequency and as users submitted longer queries ORNL was often added as a descriptor, 8% for three words or more and nearly double that for two words or more. In the Volvo study, "Volvo" also ranked consistently the most frequent term. The term "Volvo" also appeared in about 33% of their pairs and triple word queries. In the UTK study, the

word UT ranked 19th, while Tennessee ranked 20th and did not seem to have the same keyword impact. Word association trends were available only in the UTK study. Both the ORNL and UTK studies evaluated word association strength using Mutual Information (I) statistics. The UTK "I" values Range from -4.94 to 12.99 and the ORNL "I" values ranged from -3.61 to 10.72, ORNL having less "I" variation. For each study the more popular word pairs tended to have higher I values with the exception of artificially high values due to misspelled and infrequently occurring terms. The ORNL study found that the co-occurrence of the top 100 mutual information values and the top 100 frequent query terms is semantically significant.

It is interesting to note that many of the TLA measures are similar across studies, particularly with the UTK and ORNL study (Table 14). It is interesting that many of the ORNL results are similar to the UTK results even though the UTK was a longitudinal study covering multiple years and the ORNL study was much shorter covering only 3.5 months. The similarities are most likely attributable to the fact that nearly identical methodology was used for query analysis and that they were both U.S. based web sites. It is not clear to what extent language translation occurs in European search engines and how it may impact query length or syntax. It should be noted that all of the studies analyzed different search engines MOSS 2007 for ORNL, InfoSeek for Volvo and SwedCorp, and Verity UltraSeek for UTK. Each of these engines had distinct methods for joining query terms.

Table 14. ORNL results compared to published studies

Results	ORNL	UTK	SwedCorp	Volvo
Search Engine	MOSS	UltraSeek	InfoSeek	InfoSeek
Query frequency rises from September to October, but falls towards December	✓	✓	-	-
Query frequency is dramatically reduced on holidays & closures	✓	✓	-	-
Weekend queries much lower than weekdays	✓	✓	-	-
Top popular topic "applications" and least popular was "documents"	✓	-	-	✓
No hits	16%	> 30%	-	-
Blank queries	1.62%	.9%	.7%	-
Misspelled terms/words	9.6%	26%	-	-
Avg. Query Lengths	3	2	3	-
Avg. Length of Terms	7 characters	7 characters	-	-
Longest Query	14	26	-	-
Unique Terms	7.8% of all terms	6.7% of all terms	-	37.1%
Small Vocabulary	✓	✓	-	
Institution Name Rank	1	19, and 20	-	1
MI Values	-3.61 -10.72	-4.94 -12.99	-	-

In all cases users seem to not fully understand interaction with the search engines, especially as it relates to query submission. Basic query statistics like misspelled word, empty queries, and average lengths of queries are nearly identical. Two results are unique for the intranet studies. The first is the role that the Corporations name plays in query expansion. It is always the most popular term and the most popular descriptive term for query expansion. The second result is that intranet users look for very different topics than internet users. For example 8 out of 10 internet users have looked online for health information (Fox, 2010). Fox's survey (2010) describes shopping, health and entertainment as the top three topics that internet users search for. Intranet users only

search for information regarding their “health plans or health benefits”. Intranet users place little importance on social networking and general entertainment primarily because it is prohibited. Their topic priorities are business, computing information and corporate news respectively. Lastly Intranet users have fewer queries greater than three terms.

Discussion

General Search Behavior

By looking at top URLs one can begin to see topically what users were searching for. Aggregating and classifying the page types clicked gave us good insight into what users click most on the Intranet. The page type accessed most was images and the least sought after items were documents. Topic clusters gave a better indication of specific topics sought. I would recommend further study of the query text and its direct relationship to page types. Using top queries and mutual information we were successful at categorizing information into a simple three level hierarchy.

Transaction log analysis is very data intensive, but when done right yields a wealth of information about users general search behavior. Different aspects of general search behavior can be gleaned from both the transaction log and the usage log. The best way to start with the analysis is to mine the access log using a Log Parser. Additional information can be gotten from analyzing search usage reports. All of that data along with the analysis from query logs tell you what you need to generally know about users search behavior.

Written permission should be gathered before using corporate search logs. Caution must be taken to protect the personal information of users that may be contained in the search logs. Each type of log access, usage and query provides different types of information with varying amounts data processing and data mining. It is best to develop a data processing plan based on your analysis requirements, before you start cleaning, parsing and analyzing data.

Lastly, it is important to become familiar with the ORNL search audience. Knowledge regarding how ORNL intranet users search can assist in optimizing the ORNL intranet search environment. Providing classes to users on the new search tool interface would be most impactful.

Query Formulation

Query level analysis has proved valuable for recognizing popular content. Once popular content is identified it should be more prominently posted, displayed, and most importantly indexed. Popular content is identified by first aggregating all identical corpus queries, counting their frequency, and removing duplicates. The most popular content has the highest frequency. Popular queries identify the keywords that are most often used by employees to find content within the enterprise. These keywords should appear in the title of a web page, in the description of a page or in the file names of documents. Certainly, they should appears in Best Bets and "Did you mean" results.

Examining the temporal components of query analysis help to identify the times when users most need certain kinds of information. For example, "maps" of ORNL are insignificant in September, but highly sought after in November. Information publishers also need to take notice of how often the word ORNL appears as an additional "descriptive" keyword. A synonyms index can and should be developed to identify "contextually similar" queries. For example, the query "pay bands" and "ornl pay bands" would be indexed as synonyms and yield the same result to either search submission. Successful search is dependent on enterprise publishers understanding how "users" think the data should be tagged.

Vocabulary Analysis

It is very important to characterize intranet users search vocabulary. Incorrect vocabulary is likely the number one cause of failed search within the intranet. There are several methods available to characterize the vocabulary; this study has employed three of them. The first is evaluation of the unique terms to include determining the frequency with which they occur and the length of each term. The second method employed was categorization of terms into classes like misspelled words, acronyms, abbreviations, and stop words. Categorization helps identify which acronyms should be included in the index. Developing stop word lists can improve the speed of the search engine as filters through published material. The last method employed was using mutual information values to identify strongly associated word pairs. Word pairs can be used to create a "next word" index.

Successful query formulation involves the process of selecting keywords that will optimize retrieval of information specific to a users need. More often than not the ORNL user will be using search terms that are not indexed in the enterprise search engine. The average term length at ORNL is less than 8 letters so it is very likely that the term will not be a word at all rather an acronym or an abbreviation. The bulk of the queries submitted have two words, mutual information values can help identify which pairs are strong pairs. Enterprise publishers as well as administrators can build a much more efficient search engine if they study the vocabulary of their users. To assume ORNL searchers will search the intranet efficiently because they are an employee of ORNL, thus possessing special domain knowledge would be incorrect.

Chapter 5

Summary and Conclusion

Summary

The results of this study were limited largely by the size of the data set. The data analyzed covered only a little more than one quarter of the ORNL fiscal year. Because it lacked variation of events associated with other quarters, the study does not necessarily warrant a generalization and results should be interpreted strictly for that period. Unique events occur during each quarter that may impact query results and general user behavior dramatically. While these results are not authoritative, they are the first provided on the ORNL intranet using the Microsoft SharePoint Search. This study goes far beyond the standard usage statistics analysis and it gives merit to the value of combining standard web log analysis with TLA for intranets.

General Search Behavior

In this portion of the study I have characterized the general search behavior of ORNL intranet users. ORNL searchers have a consistent submission of queries from week to week and month to month, with the exception of weekends and holidays when query submission falls off significantly. They rarely use keyword search to access external search, but when they do they prefer Google. ORNL users overwhelmingly choose Internet Explorer 7.0 as their browser to connect to the Intranet. The most popular operating system platform upon which the browser resides is Windows XP.

The most active day for search is Monday and the least active is Saturday. The most active hours of search during weekdays are from 2pm – 2:59pm. The least active is from 7am-7:59am. The least amount of queries occurred on Thanksgiving. Approximately 16% of all ORNL search queries resulted in no hits. The mean number of hits was 3,394. The primary reason for no hits was Boolean operators in the query.

The most requested URL's during this time period all had something to do with SharePoint and web site construction. The most popular page types clicked were images with web applications close behind. The least clicked page types were documents. The topics most sought by users were business information and computing applications.

Query Formulation

In query formulation analysis, I successfully completed a query level assessment of ORNL Intranet queries. I found a total of 60,490 clean queries of which 18,052 were unique queries. Repeat queries occurred quite often, as well as blank queries. Blank or zero term queries accounted for 1.62% of all clean queries. ORNL query length varied from 1 to 14 terms, with no queries having 13 terms. Nearly 48% of ORNL terms were classified as special or unique language. There were only 5 occurrences of repeat terms in a query. 95.4% of all queries contained four terms or less. Most queries contained two terms.

As the most frequent queries got longer, they more often contained the term ornl. No hits decreased amongst two word queries and significantly decreased amongst three word queries. Few words changed popularity over time.

Vocabulary Analysis

For vocabulary analysis I was able to characterize the vocabulary of the ORNL intranet user. ORNL vocabulary is very small, consisting of 9,804 unique terms. The length of the terms varied from 1 to 29 characters, with a mean of 7 characters. 98% of all words contained 13 characters or less. 26% of the words were classified as acronyms, 12%, abbreviations, and 10% technical jargon. 9.6% of all unique words were misspelled. The ORNL unique term corpus tends toward a Zipfian distribution.

There was a problem reconciling “it” & “or” as stop words because they also represent frequently acronyms for “information technology” and “oak ridge”. Generally ORNL word pairs with higher information values tend to be strongly associated, but had to be verified using click rank results. Misspelled words and joint low frequency pairs can create false positives for word association strength.

Results Comparison to Published Studies

The average length of ORNL queries and the average length of query terms are consistent with other published studies. ORNL no hits and longest query is nearly 50%

less than reported in other studies. ORNL, like other organizations have a small vocabulary. Unfortunately, ORNL reports slightly higher blank queries.

Conclusion

This study successfully applied mature scientific internet web query transaction log analysis (TLA) to better understand how intranet users at ORNL are searching for information. It was able to comprehensively provide characterization of the intranet search experience by combining TLA with access log and usage report analysis. The direct results were compared to studies employing traditional internet TLA. Statistically the ORNL results were very similar to the published studies (Wang, et al., 2003; Stenmark &Jadaan, 2006), with just a few distinctions, thus proving Internet TLA is directly translatable to intranets. Most notable was the amount of special language used inside the ORNL intranet. Traditional vocabulary analysis results were skewed because of acronyms and technical jargon.

All the traditional log analysis methods were applied in this study, but their results were complemented by analysis results from access and usage logs. Access and usage log analysis is largely employed by intranet content system managers. This study adopted Mutual Information calculations from another study (Wang et al., 2003), but adapted its implementation for identifying topic clusters. This study altered slightly the approach to identifying topic clusters. Instead of using MI to predict likely semantic pairs, it used high MI values and high frequency queries to classify word association significance. The classification was validated by examining click results. This method did yield 14 topic clusters.

While it was revealed that users search for different types of information on intranets versus internets, their methods for seeking it are the same. Independent of the simplicity or complexity of the search engine, or whether it's on the internet or the intranet, users are uncertain of how to best formulate queries. They also tend to be unfamiliar with how the search engine works. Search Engine administrators are not familiar with the user's perceptions of how information is indexed on the intranet. The average results returned per query for ORNL were too high (3,395) and it reflects a knowledge gap for both users and implementation designers.

The implications for this study are many. The findings will be shared with ORNL IT staff and management so they may better understand ORNL web search behavior and

its impact on organizational effectiveness. The results can provide scientific feedback to ORNL SharePoint site managers for site design improvements.

Future Study Recommendations

Future considerations to improve ORNL search include building a next word index using the word pairs with high mutual information values and high frequencies to assist users in query formulation. The results may also be used to help classify searches as “descriptive” (fact finding) or “procedural knowledge” (how to do something). Lastly, the results can be used to help develop metadata taxonomies for the portal which will improve search by categorizing content type information (narrows search). Site navigation is also likely to improve with well designed taxonomies.

Some future analytic considerations include user surveys with the TLA. Results of the vocabulary analysis suggest further study is warranted regarding the use of special domain languages like acronyms and abbreviations. It would be beneficial to also study the relationship of query text to page types to improve topic classification. Even more insight could be gained on true user satisfaction by analyzing results pages clicked. This analysis would identify how soon relevant results are appearing, if at all. Lastly, interest remains to continue to periodically evaluate how users are interacting with ORNL’s intranet search environment because it will help reduce the number of hours users spend searching for critical job related information.

References

Baeza-Yates, R. A., Calderón-Benavides, L., & González-Caro. (2006). The Intention Behind Web Queries., *Lecture Notes in Computer Science* (Vol. Volume 4209/2006, pp. 98-109). Heidelberg: Springer Berlin.

Beitzel, S. M., Jensen, E. C., Lewis, D. D., Chodury, A., & Frieder, O. (2007). Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs, *ACM 2007 Transactions on Information Systems* (Vol. 25(2), Article No. 9).

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist., 16*(1), 22-29.

Croft, W., Cook, R., & Wilder, D. (1995). *In Providing government information on the internet: Experiences with thomas*. Paper presented at the Digital Libraries Conference.

Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A., et al. (2003). *Searching the Workplace Web*. Paper presented at the 12th International Conference of the World Wide Web.

Fallows, D. (2008). *"Search Engine Use | Pew Research Center's Internet & American Life Project"*. Retrieved from http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx.

Feldman, S., & Sherman, C. (2004). High Cost of not Finding Information. *: KMWorld." KMWorld.com - from the Editors of KMWorld Magazine* Retrieved 30 Apr., 2009, from http://www.kmworld.com/articles/readarticle.aspx?articleid=9534

Fox, S., & Purcell, K. (2010). Health | Pew Research Center's Internet & American Life Project. *Pew Research Center's Internet & American Life Project* Retrieved 7 Apr., 2010, from http://www.pewinternet.org/topics/health.aspx

Hawkey, K. (2009). Privacy Concerns for Web Logging Data. In A. S. Bernard J Jansen, Isak Taska (Ed.), *Web Log Analysis* (pp. 80-98): IGI Global.

Hawking, D. (2004). *Challenges in Enterprise Search*. Paper presented at the ACM International Conference Proceeding Series: 15th Australian Database Conference (ADC 2004).

Ives, B. (2010). The FASTForward Blog » Global Intranet Trends 2010 Report: Enterprise 2.0 Blog: News, Coverage, and Commentary." Retrieved 6 Mar., 2010, from http://www.fastforwardblog.com/2010/03/02/global-intranet-trends-2010-report

Jansen, B., Spink, A., &Taksa, I. (2009). The Methodology of Search Log Analysis *Handbook of Research on Web Log Analysis* (pp. 100-123). Hershey, Pa.: Information Science Reference.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research, 28*, 21.

Jansen, B. J., & Pooch, U. (2001). Web user studies: A review and framework for future work. *Journal of the American Society for Information Science and Technology, 52*(3), 235-246.

Jansen, B. J., & Spink, A. (2005). *Web Search: Public Searching of the Web*. Information Science and Knowledge Management. [New York]: Springer Science + Business Media.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 15.

Jansen, B. J., Spink, A., & Sarajevic, T. (2000). Real life, real users, real needs: A study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 20.

Jones, S., Cunningham, S., & McNab, R. (1998). *In usage analysis of a digital library*. Paper presented at the Third ACM Conference on Digital Libraries.

Kali, R. (2003). The city as a giant component: A random graph approach to Zipf's law. *Applied Economic Letters, 10*, 717-720.

Korfhage, R. (1997). *Information Storage and Retrieval*. New York: Wiley.

Kurth, M. (1993). The limits and limitations of transaction log analysis. *Library Hi Tech, 11(2)*(42), 98-104.

Li, H., Cao, Y., Xu, J., Hu, Y., Li, S., & Meyerzon, D., (2005). *A New Approach to Intranet Search Based on Information Extraction*. Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management.

Mann, T. (2005). "Google Searching in Library." Library of Congress Professional Guild, AFSCME, Local 2910- Home Index. Retrieved 15 Apr., 2007, from http://www.guild2910.org/google.htm.

Martin, T. H. (1977). *Monitoring and Individual Rights: In information management in the 1980's.* White Plains, NY: Knowledge Industry Publications.

Microsoft. (2007). Microsoft Office SharePoint Server (MOSS). Retrieved October 14, 2008, from http://www.microsoft.com

Microsoft. (2008). [MS-SQL]-v1.01 Search Service Database Protocol Specification. from http://www.microsoft.com

Nielsen, J. (11 Nov. 2002). "Intranet Usability: The Trillion-Dollar Question (Alertbox).". Retrieved 13 June 2007, from http://www.useit.com/alertbox/intranet-usability-1st-study.html

Nielsen, J. (15 Sept. 1997). "Intranet vs. Internet Design (Alertbox)." Useit.com: Jakob Nielsen on Usability and Web Design. 15 Sept. 1997. Retrieved 14 May, 2008, from http://www.useit.com/alertbox/9709b.html

Nielsen, J. (1997). Zipf Distribution (power Law) of Website Popularity (Alertbox Sidebar). *Useit.com: Jakob Nielsen on Usability and Web Design* Retrieved 19 Apr., 2009, from http://www.useit.com/alertbox/zipf.html

Nielsen, J. (2010). "Nielsen | Internet." Nielsen | Home. . Retrieved 5 Apr. 2010: http://en-us.nielsen.com/rankings/insights/rankings/internet

Palermo, D., & Jenkins, J. (1964). *Word Association Norms*. Minneapolis: Minnesota Press.

Spink, A., Wolfram, D., Jansen, B. J., &Seracevic, T. (2001). Searching the Web: The Public and Their Queries. *Journal of The American Society for Information Science and Technology, 52*(3), 9.

Stenmark, D. (2005). Searching the intranet: Corporate users and their queries, *Proceedings of ASIS&T 2005*. Charlotte, North Carolina.

Stenmark, D. (2006). What are you searching for? A content analysis of intranet search engine logs, *Proceedings of IRIS29*. Helsingør, Denmark.

Stenmark, D., & Jadaan, T. (2006). Intranet users' information-seeking behaviour: An analysis of longitudinal search log data. *Proceedings of the American Society for Information Science and Technology, 43*(1), 1-19.

Turney, P. D. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the 12th European Conference on Machine Learning (ECML-2001).* Frieburg, Germany.

Valdez-Perez, R. (2007). Enterprise Searching to Surpass Web Searching? Search Done Right. Retrieved April 19, 2008, from http://searchdoneright.com/2008/02/enterprise-searching-to-surpass-web-searching/

Wang, Peiling. 2006. “A Dual-approach to Web Query Mining: Towards Conceptual Representations of Information Needs.” 2005 OCLC/ALISE research grant report published electronically by OCLC Online Computer Library Center, Inc. Available online at: http://www.oclc.org/research/grants/reports/2005/wang-p.pdf

Wang, P., Berry, M. W., & Yang, Y., (2003). Mining Longitudinal Queries: Trends and Patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 16.

Wang, P., Wolfram, D., & Wu, L. (2008). *Web Search Log Analysis and User Behavior Modeling*. Paper presented at the Conference on Information & Knowledge Management.

Wang, P., Wolfram, D., Zhang, J., Hong, N., & Wu, L. (2007). *Mining Web Search Behaviors: Strategies and Techniques for Data Modeling and Analysis*. Paper presented at the In Proceedings of 2007 Annual Meeting of the American Society for Information Sciences &Technology.

Ward, T. (2005, Sept. 7). "Intranet Blog :: Intranet vs Internet Search (back Issue)." *Web log post. IntranetBlog.com.* Retrieved 18 Mar., 2010, from http://intranetblog.blogware.com/blog/_archives/2005/9/8/1207102.html

Wolfram, D., Wang, P., & Zhang, J. (2008). *Modeling Web Session behavior Using Cluster Analysis: A Comparison of Three Search Settings*. Paper presented at the American Society for Information & Technology.

Zien, J., Meyer, J., & Tomlin, J. (2000). *Web query characteristics and their implications on search engines.* San Jose, Ca.: IBM Research Division.

Appendix

Appendix A

1. **Count the distinct number of users and browsers for the whole data set:**

C:\Program Files\Log Parser 2.2>LogParser -i:IISW3C -o:NAT "SELECT COUNT(DISTINCT cs(User-Agent)) AS Browsers, COUNT(DISTINCT c-ip) AS Users INTO STDOUT from ex*.log“

2. **Show most used browsers (Top 20) in 3-D bar chart:**

C:\Program Files\Log Parser 2.2>LogParser -i:IISW3C -o:chart -view:on "SELECT TOP 20 cs(user-agent) AS Browser, COUNT(*) AS Hits INTO t20browsers.gif FROM ex*.log GROUP BY Browser ORDER By Hits DESC" -charttype:Bar3D -charttitle:"Top 20 Browsers“

3. **Show the top 20 URLs' requested:**

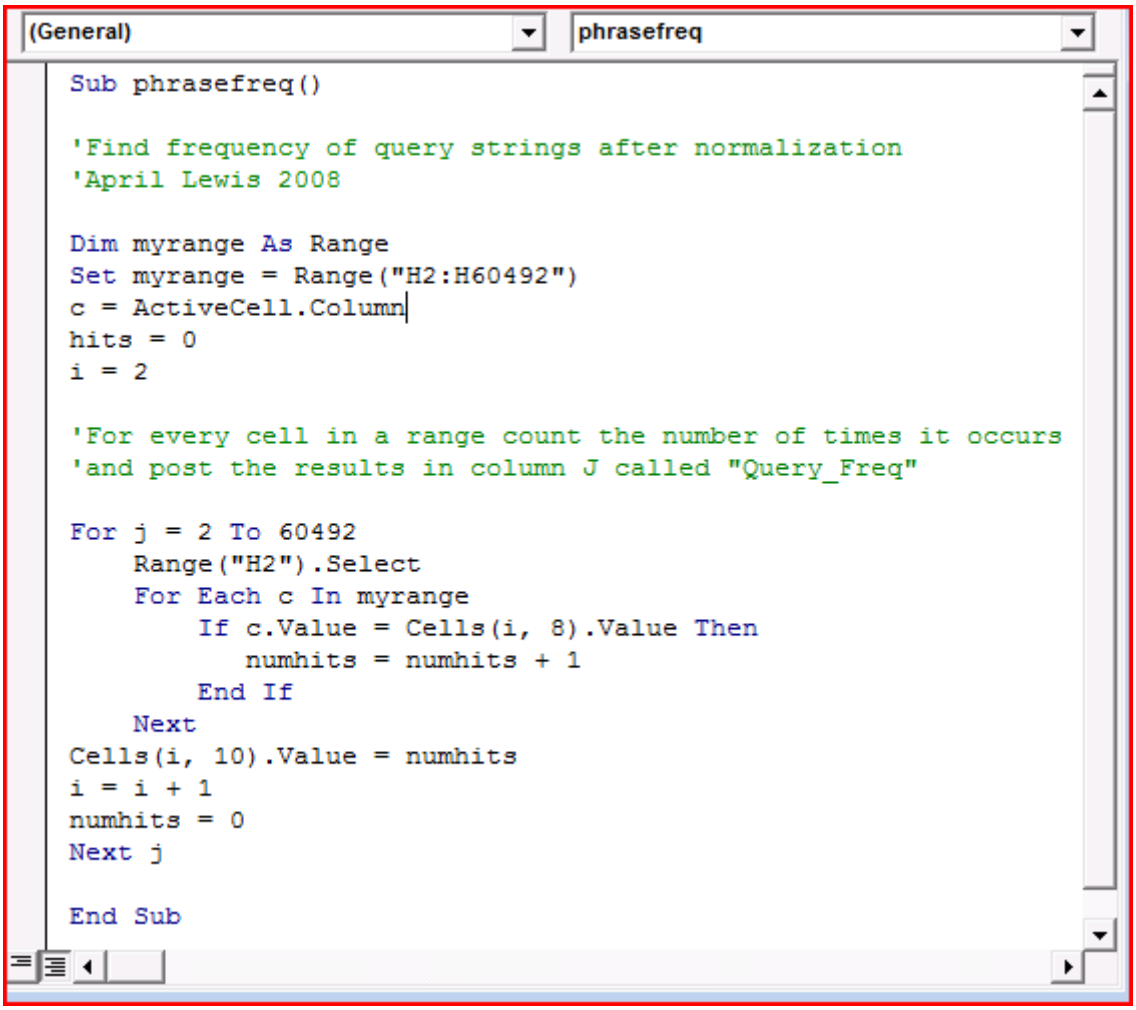
C:\Program Files\Log Parser 2.2>LogParser -i:IISW3C -o:NAT "SELECT TOP 20 cs-uri-stem AS Url, COUNT(*) AS Hits INTO STDOUT FROM ex*.log GROUP BY Url ORDER By Hits DESC“

4. **What are the top page "types" requested (to qualify the pages must get more than 250 hits)?**

C:\Program Files\Log Parser 2.2>LogParser -i:IISW3C -o:NAT "SELECT EXTRACT_EXTENSION(cs-uri-stem) AS PageType, COUNT(*) AS PageTypeHits FROM ex*.log GROUP BY PageType Having PageTypeHits >=250 ORDER BY PageTypeHits DESC" PageType PageTypeHits

Appendix B

Visual Basic Macro



(General) CountWords

```
Sub CountWords()

'Count the number of words in a query
'April Lewis 2008

Dim myrange As Range
Set myrange = Range("A2:A60492")
i = 2
c = ActiveCell.Column

'For every cell in a range read cell and
'count the number of words in the query
'post the result in a seperate column

For j = 2 To 60492
    Range("A2").Select
    For Each c In myrange
      NumWords = Len(Trim(Cells(i, 1).Value)) - Len(WorksheetFunction.Substitute(Cells(i, 1).Value, " ", "")) + 1
    Next
Cells(i, 2).Value = NumWords
i = i + 1
NumWords = 0
Next j
End Sub
```

(General) id_populate

```
Sub id_populate()

'populate cells with contiguous word pair identification
'numbers
'April Lewis 2008

'have loop start at row 2
'and start numbering from 1

 i = 2
 N = 1

For j = 1 To 23594
    Range("L2").Select
    Cells(i, 12).Value = N
i = i + 1
N = N + 1
   Next j


End Sub
```

```
Sub MakeWordList()
'Make a word list and count frequency of unique words
'orginal code by John Walkenbach

    Dim InputSheet As Worksheet
    Dim WordListSheet As Worksheet
    Dim PuncChars As Variant, x As Variant
    Dim i As Long, r As Long
    Dim txt As String
    Dim wordCnt As Long
    Dim AllWords As Range
    Dim PC As PivotCache
    Dim PT As PivotTable

    Application.ScreenUpdating = False
    Set InputSheet = ActiveSheet
    Set WordListSheet = Worksheets.Add(after:=Worksheets(Sheets.Count))
    WordListSheet.Range("A1") = "All Words"
    WordListSheet.Range("A1").Font.Bold = True
    InputSheet.Activate
    wordCnt = 2
    PuncChars = Array(".", ",", ";", ":", "'", "!", "#", _
        "$", "%", "&", "(", ")", " - ", "_", "--", "+", _
        "=", "~", "/", "\", "{", "}", "[", "]", """", "?", "*")
    r = 1
'   Loop until blank cell is encountered
    Do While Cells(r, 1) <> ""

'       Remove punctuation
        For i = 0 To UBound(PuncChars)
            txt = Replace(txt, PuncChars(i), "")
       Next i
'       Remove excess spaces
        txt = WorksheetFunction.Trim(txt)
'       Extract the words
        x = Split(txt)
        For i = 0 To UBound(x)
            WordListSheet.Cells(wordCnt, 1) = x(i)
            wordCnt = wordCnt + 1
        Next i
    r = r + 1
    Loop
  ' Create pivot table
    WordListSheet.Activate
    Set AllWords = Range("A1").CurrentRegion
    Set PC = ActiveWorkbook.PivotCaches.Add _
        (SourceType:=xlDatabase, _
        SourceData:=AllWords)
    Set PT = PC.CreatePivotTable _
        (TableDestination:=Range("C1"), _
        TableName:="PivotTable1")
    With PT
        .AddDataField .PivotFields("All Words")
        .PivotFields("All Words").Orientation = xlRowField
    End With
End Sub
```

(General) Word_Pair_Count

```
Dim Count As Integer
Dim Target As String
Dim Cell As Object
Dim N As Integer
Dim word As String
Dim i As Long

Sub Word_Pair_Count()

'The following Visual Basic procedure reads a range
'of input cells (can be words or characters,then searches
'through the currently selected cell, or range of
'cells, and displays the total number of occurrences
'of that character or character string incolumn F.
'This works for all alphanumeric characters and works
'more specifically for "partial string matches".
'To use select target column and the input strings
'you want to count
'April Lewis 2008

   Count = 0
   i = 2
For j = 2 To 2468
   word = Cells(i, 10).Value
   Target = word


   If Target = "" Then GoTo Done
      For Each Cell In Selection
         N = InStr(1, Cell.Value, Target)

         While N <> 0
            Count = Count + 1
            N = InStr(N + 1, Cell.Value, Target)
         Wend
      Next Cell
    Cells(i, 6).Value = Count
    i = i + 1
    Count = 0

Next j
Done:
End Sub
```

Appendix C

Page/File Type	Description	Count
asmx	*Active Server Method File*: A file type used by Microsoft Windows internet servers.	17322
aspx	*Active server page web applications* for dynamic web services(MS 2002)	334155
axd	Actrix Technical, a handler file used to manage Web site administration requests, typically	31835
css	*Cascading Style Sheets*, language most often used for styling web pages	184781
dll	*Dynamic Link Library* is a collection of small programs, which can be called upon when needed by the executable program (EXE) that is running	3324
doc	*Document*; most commonly associated with document created with Microsoft Word, but were originally used with WordPerfect in the 1980's.	605
docx	*Open Office* and *MS-Word*	430
gif	*Graphic interchange format*, bitmap for images (Compuserve 1987)	648886
GIF	Same as "gif"	30095
htm	*HyperText Markup Language*, is the predominant markup language for Web pages	866
html	Same as *"htm"*	444
ico	The ICO file format is an image file format used for *icons* in Microsoft Windows	2960
jpg	*Joint Photographic Expert Group*, a lossly compression method used for storing large bitmap images and displaying images on the web	68358
js	*JavaScript*, a scripting language created by Netscape to create interactive web pages	95985
pdf	*Portable Document Format* is a file format created by Adobe Systems in 1993 for document exchange	1097
png	*Portable Network Graphics* is a bitmapped image format that employs lossless data compression	9867
ppt	Associated with digital documents created in Microsoft® *PowerPoint®*	284
vsd	*VISIO Drawing "Microsoft Corporation"*	347
xlsx	*Microsoft Office 2007 Excel* spreadsheet file	569
ini	*Standard Configuration"Initialization" File* is more common in Windows, but also in other environments	282

Vita

April Lewis was born in southwestern Pennsylvania, located just south and west of Pittsburgh. She graduated with honors from Uniontown Area Sr. High School in 1980. After graduating, she entered the US ARMY serving as an Intelligence Analyst largely focused on remote sensing data analysis. The last two years of her military career was spent at Ft. Hauchuca, in Sierra Vista Arizona as an instructor of remote sensing technology. After leaving the military she spent ten years with the Environmental Research Institute of Michigan (ERIM), a non-profit research company in Ann Arbor Michigan. There she focused on developing feature extraction algorithms for synthetic aperture radar imaging systems. She attended Eastern Michigan University while with ERIM earning her Bachelor of Science degree in Physical Science in 1995. In late summer of 1998, April started working for the Department of Energy (DOE) Remote Sensing Laboratory (RSL) in Las Vegas, Nevada. RSL provided the opportunity to expand her research into multispectral and thermal imaging domains. In 2002 April began working for Oak Ridge National Laboratory (ORNL) in Oak Ridge Tennessee as a Research and Development Manager. While at ORNL, April returned to higher education at the University of Tennessee earning a Master of Science degree in 2010. Her advanced degree in Information Science was essential for her new research endeavors in Human Computer Interaction and Information Management Systems. In the future, April plans to continue information research at ORNL and completing a Doctoral degree in Information Science with the University of Tennessee.