INFORMATION SEEKING BEHAVIOUR ON AN ACADEMIC LIBRARY SEARCH ENGINE

OLUBUKOLA ODUNTAN

This dissertation was submitted in part fulfilment of requirements for the degree of MSc Information Management

DEPT. OF COMPUTER AND INFORMATION SCIENCES
UNIVERSITY OF STRATHCLYDE

SEPTEMBER 2014

DECLARATION

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself. Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

(please tick) Yes [✓] No [ ]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices is **16370**

I confirm that I wish this to be assessed as a Type 1 2 3 [✓ 4] 5
Dissertation (please circle)

Signature:

Date: 29 August 2014

ABSTRACT

Transaction log analysis was conducted on the University of Strathclyde library search engine (SUPrimo) to provide insight into the usage patterns, query characteristics and search patterns on the information retrieval system in a holistic academic context with the goal of identifying areas of improvement for the system.

The research revealed that SUPrimo usage increased during exam period, users' queries were short and users' generally employed key terms search. It was discovered that a relationship exists between the query length and results. Furthermore query length, system problem and content organisation were discovered as the causes of query failure; thus leading to functional, technical and operational recommendations to improve SUPrimo to better support users' information seeking behaviour and presenting a foundation for future research.

ACKNOWLEDGEMENTS

Sincere gratitude to my supervisor; Prof Ian Ruthven for his inestimable guide through the process of establishing the dissertation topic to conducting the research and to the library information systems team; Judith Smart and Helen Cooper for making the logs available and their insights.

Immeasurable thanks to my husband for believing in me and supporting me through it all and to my children for their understanding during those special times.

CONTENTS

FIGURES

TABLES

1. INTRODUCTION

Information retrieval systems accessed over the internet are now globally used for providing access to information (Bates, 2012). Their effectiveness for information retrieval has extend their use beyond general-purpose systems to domain specific systems within an organisation or industry including traditional online retrieval systems (Jansen, 2006). Finding information on these systems is usually done through a search. Marchionini (1997) describes a search as the behavioural manifestation of humans engaged in information seeking and also describes the actions taken by computers to match and display information objects. In order to evaluate these systems on how well they support people information seeking, there is a need to understand information seeking behaviour of users on the system (Ruthven and Kelly, 2011).

Information retrieval systems (search engines) are not just a means to access information, they can provide information on the interaction between a user and a system through logs that are kept which can be analysed (Jansen and Spink, 2006). Analysing these logs through a process called transaction log analysis provides information on system performance, information structure and measurements of user interactions (Jansen, 2006).

Extensive research has been carried out to understand user and system interaction on general purpose search engines; however little is understood about users and system interaction on domain specific purpose search engines. Whilst there has been research on domain specific search engines particularly health, research into academic information retrieval systems has been limited, though search behaviours of different characteristics can be anticipated. Bates (2010) expresses this “Among the professions, it is almost certainly the health sciences where the largest body of information behaviour research has been done—probably due to abundant funding—while the education profession, despite the importance of information seeking for teachers, seems, mysteriously, to have drawn very little attention”.

1.1 Aims

The purpose of this research is to provide insight into how users' find information on a University Library's search engine in the context of an academic domain information retrieval system by answering the following questions:

- What are the usage patterns on the search engine?
- What are the characteristics of queries issued on the search engine?
- What are the patterns in users' searching?

1.2 Objectives

The objectives of the research are to examine and analyse the transaction (search) log of a university library's information retrieval system for users' information seeking behaviour, thereby providing an evaluation of the system and identifying areas of improvement and also extending current knowledge of users' information seeking behaviour on information retrieval systems. The research will specifically achieve these objectives:

- Discover the usage patterns on the system.
- Determine the length of queries used in searching on the system.
- Discover the query terms used on the information retrieval system.
- Discover the patterns used for searching on the engine.
- Discover commonly searched terms.
- Discover relationship between variables of user-system interaction.
- Discover reasons for search failure.
- Discover the most frequently used search option.

1.3 Significance of Study

Primarily, the research will extend current knowledge of user-system interactions on domain specific information retrieval systems in the context of an academic domain thereby extending the body of knowledge on information behaviour.

Furthermore, carrying out the research will potentially provide the University library authorities with an evaluation of the existing system to either improve or design better systems that will enhance users' search experience. Specifically the following:

- Discovery of usage patterns in terms of examination and non-examination period will provide explicit information on expected usage.
- Classification of users from query terms and commonly search terms will provide information to optimise the search engine to handle common requests.
- Discovery of reasons for search failure will evaluate the information retrieval system in terms of its information architecture specifically content organisation.
- Knowledge of the most used tab on the system (as information in the system can be found through four search tabs: library collections, Course materials, articles + databases and Strathclyde research) will result in identification of redundant system components.

Ultimately the researcher will gain practical knowledge of web log analysis, deeper understanding of the concepts of information seeking, behaviour and retrieval. The researcher will also indirectly gain practical knowledge of statistical analysis, entity relationship modelling as transaction log analysis involves some entity modelling and content management in the context of how information is organised.

2 LITERATURE REVIEW

2.1 Chapter Overview

This chapter explores information seeking in an electronic environment, Academic Libraries - its transition from traditional records systems, to Online Public Access Catalogue (OPAC) systems and recently more sophisticated information architecture and Transaction Log analysis for understanding information seeking behaviour in electronic environment. These interwoven topics were explored to establish the need for this research. Information seeking in electronic environment as broad description, academic libraries as an electronic environment in the domain specific context and transaction log analysis for understanding user behaviour in electronic environment.

Furthermore, an extensive amount of related work was examined for this research to reveal the current depth/extent of researches on transaction log analysis of web information retrieval systems thereby highlighting work yet to be done on current library information retrieval systems. This provides foundation for further research not only in academic field as a domain but also other domain search engines.

2.2 Information Seeking in Electronic Environment

Advancements in technology enabled dealing with information in new forms especially electronic forms that are more abstract, more dynamic and more malleable than conventional print forms and online retrieval systems. Information in electronic forms provides the advantage of easy access from anywhere in the world but on the other hand may require additional levels of learning and cognitive effort to use and acquire information (Marchionini, 1997).

Information seeking is described as the conscious effort to acquire information in response to a need or gap in knowledge (Case, 2012). Information is a valuable resource in this current age of information society where acquiring and using information are critical activities especially in a learning environment. The process of information seeking is therefore becoming more fundamental and strategic for intelligent citizenship (Marchionini, 1997), not only with respect to the world but in order to belong to an industry or organisation.

Information Seeking to bridge a gap in knowledge (human perspective) typically involves Information Retrieval (system response i.e. technical perspective) (Ruthven and Kelly, 2011). The relationship between the two is **“user behaviour”** otherwise referred to as **“information behaviour”** to acquire information from the system.

Bates (2010) defines information behaviour as the preferred term to describe the many ways in which humans interact with information, in particular, the ways in which people seek and utilize information. Therefore understanding this behaviour will provide evaluation of the systems which help improve the existing online systems or design better ones to support information seeking (Ruthven and Kelly, 2011). This behavioural information is indirectly presented in the transaction log of information retrieval systems (Peters, 1993).

2.3 Academic Library Information Retrieval Systems

Academic libraries have traditionally served as the main repositories and intermediaries for the acquisition of published information over the years. The impacts of technological change and rationalization upon the world brought about new changes related to the expanding information universe and its now electronic means of production (Carson, 2004).

In libraries, the book shelves that also served as reading desk and the traditional card catalog gave way first to the COM (Computer Output on Microform) catalog, then to the Online Public Access Catalogue (OPAC) and now even to the more sophisticated information retrieval systems (Wang, 1985). Bates (2012) describes this transition as “the use of sophisticated technology is now adopted in current library information retrieval systems which makes it more complex than just being digital libraries”.

The world is challenged to keep up with these fast paced changes by creating better web pages, to mine the rich fund of information resources on the Internet, or to develop more capable and user-friendly information retrieval systems (Carson, 2004). As a result, the evolution of academic libraries from traditional record systems through Online Public Access Catalogue systems (OPAC) to its current information architecture. To this end, this sophisticated academic libraries information retrieval systems are now defined as “a networked information systems consisting several databases and also provide location of books within the library” (Bates, 2012).

New academic libraries still operates with the fundamental principle of libraries but with some enhancement especially the convenience of offsite access as a result of the advancement in technology. Historically library founding principles was for the purpose of learning and knowledge sharing by collection of books and making it accessible through reading at a reading desk as it was in the ancient and medieval times, then to borrowing and lending and now to modern times, it is accessed electronically (Cubitt, 2006).

Modern times libraries are now being accessed through web based retrieval system and as it is with general purpose web based retrieval systems that are continually reviewed, evaluated and re-designed through the study of user behaviour to better support the needs of users, there is the need to understand how users search these new library information retrieval systems to measure the effectiveness of the systems.

This user behaviour for measuring system effectiveness have been studied for decades through surveys, controlled experiments, group interviews, protocol analyses, transactional analyses, measures of the effectiveness in retrieval, and assessments of user satisfaction (Jansen 2006). The implicit objective of all proposed methods to date is to understand user behaviour which is also the objective of this research with focus on the transaction logs.

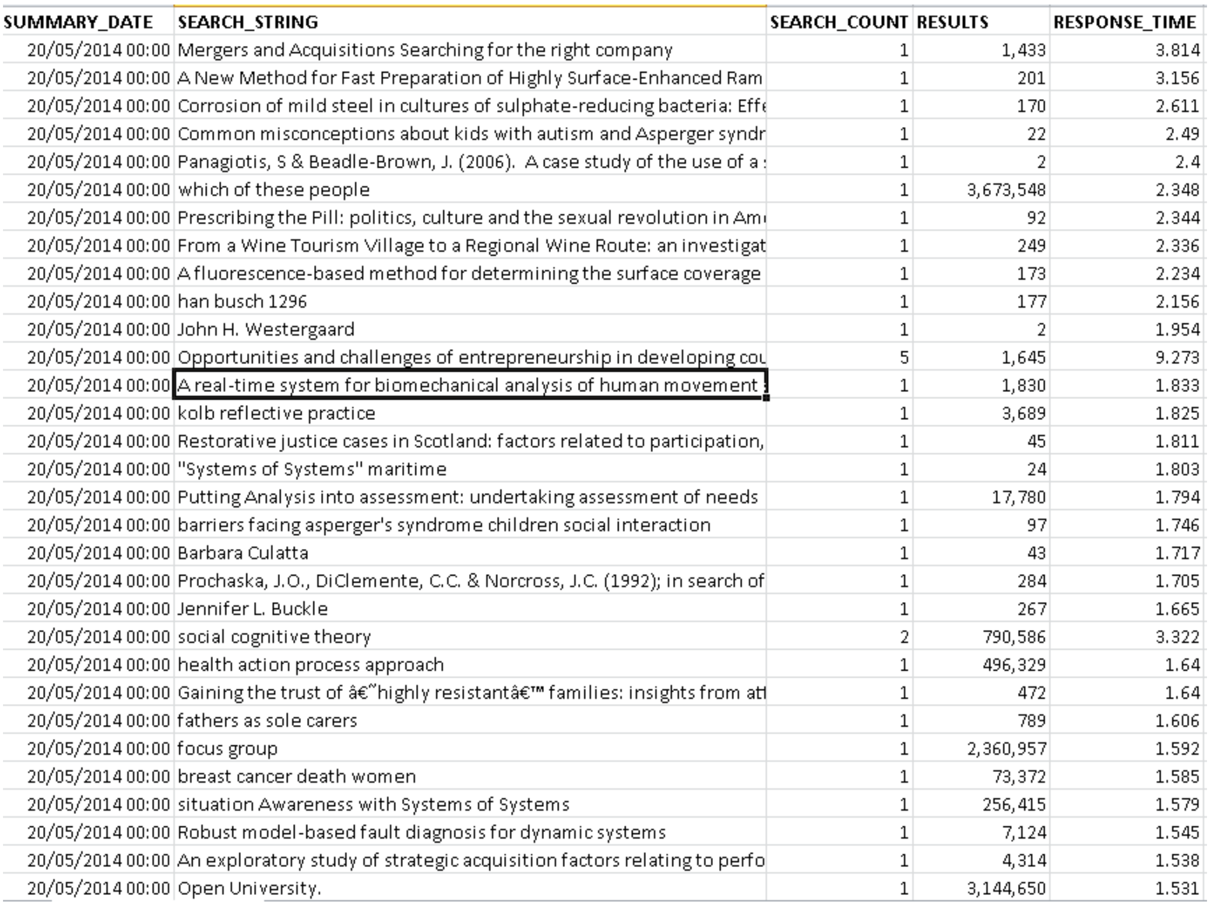
2.4 Transaction Log Analysis

Transaction Log Analysis is a method for system monitoring and a way of observing, usually unobtrusively, human behaviour in the pursuit of real information needs in a complex web information environment (Peters, 1993, Jansen et al., 2009). It provides insight to the information seeking process of searchers in an electronic environment; this understanding enlightens the design of information retrieval systems, development of interface and information architecture for content (Jansen and Spink, 2006, Jansen, 2006).

Transaction log analysis is founded on the grounded theory approach (Glaser, 1999) in that characteristics of searches are examined to isolate trends that identify typical interactions between searchers and the systems. It is a broad term that includes Web log analysis (i.e. analysis of Web system logs), blog analysis and search log analysis (analysis of search engine logs) which is the focus of this study. It uses transaction logs to discern the attributes of search process such as the searchers action, interaction between user and the systems and the delivery of result (Jansen and Pooch, 2001).

Transaction log records automatically records the interactions that have occurred between a system and users of that system. The elements recorded in these log files depends on the type of software and the options set by the provider/administrator. Most logs generally include identity of the computer, date and time of request, the request (called search strings), number of results returned, response time etc. These records of individual request-response can be retrieved over any given period of time as text files (Jansen 2006).

A typical log file contains the date, the request known as the search string, search count, number of results returned and response time (figure 2.1).



SUMMARY_DATE	SEARCH_STRING	SEARCH_COUNT	RESULTS	RESPONSE_TIME
20/05/2014 00:00	Mergers and Acquisitions Searching for the right company	1	1,433	3.814
20/05/2014 00:00	A New Method for Fast Preparation of Highly Surface-Enhanced Ram	1	201	3.156
20/05/2014 00:00	Corrosion of mild steel in cultures of sulphate-reducing bacteria: Effe	1	170	2.611
20/05/2014 00:00	Common misconceptions about kids with autism and Asperger syndr	1	22	2.49
20/05/2014 00:00	Panagiotis, S & Beadle-Brown, J. (2006). A case study of the use of a	1	2	2.4
20/05/2014 00:00	which of these people	1	3,673,548	2.348
20/05/2014 00:00	Prescribing the Pill: politics, culture and the sexual revolution in Am	1	92	2.344
20/05/2014 00:00	From a Wine Tourism Village to a Regional Wine Route: an investigat	1	249	2.336
20/05/2014 00:00	A fluorescence-based method for determining the surface coverage	1	173	2.234
20/05/2014 00:00	han busch 1296	1	177	2.156
20/05/2014 00:00	John H. Westergaard	1	2	1.954
20/05/2014 00:00	Opportunities and challenges of entrepreneurship in developing cou	5	1,645	9.273
20/05/2014 00:00	A real-time system for biomechanical analysis of human movement	1	1,830	1.833
20/05/2014 00:00	kolb reflective practice	1	3,689	1.825
20/05/2014 00:00	Restorative justice cases in Scotland: factors related to participation,	1	45	1.811
20/05/2014 00:00	"Systems of Systems" maritime	1	24	1.803
20/05/2014 00:00	Putting Analysis into assessment: undertaking assessment of needs	1	17,780	1.794
20/05/2014 00:00	barriers facing asperger's syndrome children social interaction	1	97	1.746
20/05/2014 00:00	Barbara Culatta	1	43	1.717
20/05/2014 00:00	Prochaska, J.O., DiClemente, C.C. & Norcross, J.C. (1992); in search of	1	284	1.705
20/05/2014 00:00	Jennifer L. Buckle	1	267	1.665
20/05/2014 00:00	social cognitive theory	2	790,586	3.322
20/05/2014 00:00	health action process approach	1	496,329	1.64
20/05/2014 00:00	Gaining the trust of â€˜highly resistantâ€™ families: insights from att	1	472	1.64
20/05/2014 00:00	fathers as sole carers	1	789	1.606
20/05/2014 00:00	focus group	1	2,360,957	1.592
20/05/2014 00:00	breast cancer death women	1	73,372	1.585
20/05/2014 00:00	situation Awareness with Systems of Systems	1	256,415	1.579
20/05/2014 00:00	Robust model-based fault diagnosis for dynamic systems	1	7,124	1.545
20/05/2014 00:00	An exploratory study of strategic acquisition factors relating to perfo	1	4,314	1.538
20/05/2014 00:00	Open University.	1	3,144,650	1.531

Figure 2.1 Typical transaction log

Transaction logs collects large scale unobtrusive data to a degree which overcomes the critical limiting factor in laboratory settings which are typically restricted in terms of sample size; where there are not enough participants to generalise findings, location, duration and participants' behaviour; which is likely to be altered when participants have the knowledge that they are being observed. These large scale data presented allows inference testing which highlights statistically significant relationships (Jansen 2006).

The use of transaction logs in terms of scope allows researchers to investigate the entire range of user system interactions or system functionality in a multi variable where variables are one or more elements recorded in the logs; by queries terms, query frequency, session, and clicks amongst others. These variables depend on the aims of the research (Jansen et al., 2009). The variables in the context of this study includes query because the aim of the study is to characterize the behaviour of users of the university of Strathclyde library systems and to give insight into the content of their searches.

However, apart from generic problems of research methods such as abstraction, selection, reduction, context and evolution problems (Hilbert and Redmiles, 2001); Transaction logs specifically are limited because data on individual identities is typically not recorded in a transaction log. It also does not record the reasons for the search, the searcher's motivations, or other qualitative aspects of the user (Phippen et al., 2004). In addition, client-side caching may result in incomplete data logging of the number of identical web queries from users (Jansen, 2006, Jansen and Spink, 2006).

Furthermore, there are difficulties with unobtrusive methods of data collection: Firstly these logs are not trivial to prepare, clean and analyse. Secondly data is collected by third parties hence there might be need to make assumptions. Thirdly is the issue of ethics; (Page, 2000) points out a growing distaste for unobtrusive methods due to increased sensitivity towards the ethics involved. Nonetheless transaction log analysis has long since been extensively used to gain insight into users search behaviour on information retrieval systems.

2.5 Related Work

Information science has a long history of transaction-log analysis for studying user behaviour on information retrieval systems. Its use has extended beyond conventional full text search engines to multi-media information search systems and across domains. Although Agosti et al. (2012) classified transaction log analysis into two main themes: Web search engine log analysis and Digital Library System log analysis, log analysis efforts of other domains should be recognised.

2.5.1 General Purpose Search Engines

Extremely popular are the numerous studies of the logs of general purpose web search engines which was earlier forecasted at its incipient stage (Jansen 2006). By 2006, researchers were comparing findings from transaction log analysis of different general purpose search engines also comparing it to blog searches: Jansen and Spink (2006) compared existing results of web searching behaviour from nine different studies of five general purpose web search engines and found similarities in session length, query length which were short, and number of results pages viewed but differences in the usage of advanced web-query operators. Mishne and De Rijke (2006) compared the user behaviour on blogs to web and found that user behaviour were related to web with short sessions and users were only interest in first few results however blog searches have different intents than general web searches, suggesting that the primary targets of blog searchers are tracking references to named entities, and locating blogs by theme.

Bendersky and Croft (2009) analysed long queries on general purpose MSN log to find characteristics of the most commonly occurring queries in order to understand the information need behind them. Carman et al. (2009) examined the difference between the vocabularies of queries, social bookmarking tags, and online documents to understand how useful tag data may be for improving search results and conversely, query data for improving tag prediction.

Researches on these general purpose web search engines have extended beyond general search characteristics to relating user behaviour with emotions, causative actions and so on in order to improve information systems responses: Kato et al. (2012) analyzed data obtained from a Microsoft Bing search engine for what circumstances cause the user to turn to query suggestion in order to improve systems to effectively assist the user depending on situations. Ruthven (2012) studied how users use search engines in times of grief and bereavement to understand their information goals to help provide tailored support for different types of queries.

2.5.2 Multi-Media Search Engines

On multimedia web search engines; Christel (2007) analyzed transaction logs and questionnaires to study the behaviour of professional users—government intelligence analysts—who use an experimental, content-based video-retrieval system to answer a set of predefined information needs. Jansen et al. (2004) found that multimedia web searching was relatively complex as compared to general web searching, with a longer average query length and higher use of Boolean operators. In a later study of transaction logs; Tjondronegoro et al. (2009) found that multimedia searches used relatively few search terms. Huurnink et al. (2010) analysed logs of an audio-visual broadcast archive to understand media professionals' information seeking behaviour.

2.5.3 Domain Specific Search Engines

Quite popular are transaction log analysis for users behaviour on domain specific search engines in particular the medical/health domain which have been done in different contexts. On medical databases Crowell et al. (2004), Zhiyong et al. (2009) and Herskovic et al. (2007) analysed logs of PubMed an interface to MEDLINE, the largest biomedical literature database in the world with focus on queries in order to improve how systems supports user queries.

Health queries on general purpose commercial search engines have also been studied; Spink et al. (2004) and Ginsberg et al. (2009) analysed health-related queries submitted to commercial search engines, the former to understand how queries are issued to the engines and improve them and the latter to detect medical condition.

In the context of electronic health record retrieval systems; Natarajan et al. (2010) and Yang et al. (2011) analysed the logs and found that users behaviour on the search engine was substantially different from general web search as the average query length was 5 terms and concluding that information need in the medical domain are more sophisticated than web search queries.

On other domain specific engines Bajracharya and Lopes (2012) analysed the logs of Koders search engine for software engineers (commercial but domain specific). They provide a statistical characterization of search behaviour and report many similarities with general purpose web search behaviour although usage behaviour is unique to Koders.

2.5.4 Academic Library Search Systems

Transaction logs for user behaviour on library systems have long since been studied to improve library systems online information retrieval especially since its transition to the use of OPAC. However searching current library information retrieval systems and searching library catalogue differs because searching a library catalogue is not done to retrieve a material electronically but to find out the availability of sources on a particular topic including bibliography (Hewson et al., 2002). Although some indexes to external journal but cannot be compared to the sophistication offered in current library information retrieval systems.

At the incipient stage of OPAC Systems, subject searching was identified as the type of search presenting most problems for the users and hence the heavy researches exist to monitor its use and improve these systems. The first major large scale study of OPACs was conducted by U.S. Council on Library Resources in 1983 to understand the uses of online catalog in libraries and revealed that users use subject searches in the online setting despite the fact that they were the ones considered as most problematic for the user, thereby recommending that the online catalogue environment be better directed at subject searching (Villén-Rueda et al., 2007).

Moving decades through into the new millennium most researches focused on single databases, OPACs, archives and so on instead of the holistic system. This was appropriate for their research specific objectives however the holistic picture would enable a generalisation that can be further expanded and researched which is the purpose of this research.

Consequently selecting a single database for analysis cannot represent holistic user behaviour in academic institution that consists of hundreds of staff and tens of thousands of students. Jones et al. (2000) analysed the log of New Zealand Digital library collected over a period of 61 weeks for user behaviour on a single database of Computer Science Technical Reports Collection focusing on computer science researchers (a selection from all students in the institution). They reported similarities of user behaviour to general purposes search engines as the average number of terms in a query in the database was 2.43 terms and also reported that users experience many of the same difficulties in searching and dealing with query languages that had been reported on general purpose search engines.

Chen and Cooper (2001) analysed the transaction logs of a web-based University of California's (UC) MELVYL® on-line library catalog system collected over a period of two months with focus on a bibliographic database (single database) in the context of traditional library catalogue for possibility of detecting usage patterns and found six distinct patterns of use of the system. Chen and Cooper (2002) found that a third-order Markov model explained five of the six clusters on further analysis of same database logs.

Park and Lee (2013) also studied user behaviour on the logs of a science and technology retrieval system of the Korea Institute of Science and Technology with focus on science and technology users (focus on specific users). They found similarities in behaviour with general purpose search in terms of query but differences in session length; it was longer on the science and technology system.

On OPACs; Lau and Goh (2006) analysed the log of Nanyang Technological University (NTU) online public access catalog collected for a semester institution and found that strategies employed by OPAC users have not changed from earlier results by Jones et al. (2000) amongst others which indicated that query’s on OPACs were on the average between 2-3 terms. The average query length was 2.86 terms and the use of keyword searches contributed to 68.9% of all queries while other options such as title, author and subject accounted for 16.5%, 8.2% and 6.4% of all searches respectively even with the ubiquity of the Internet and search engines. They find that users continue to enter simple queries of one to three terms and that users employ Boolean operators only slightly more than 11 percent of the time.

Villén-Rueda et al. (2007) analysed logs of online catalog of Library of the University of Granada collected over a period of one year to determine how users effect queries on the OPACs and found a correlation between search for information and the desire to locate a document in the OPACs which are what the OPACs are for. They further reported characteristic use in the University of Granada of a strong preference for searching by title (49 percent), followed by searches by author (37 percent), and finally, by subject search (14 percent). Thereby highlighting the declining interest surrounding subject searches against it earlier researches which found it used more despite being problematic.

Moulaison (2008) examined transaction logs of online catalog of the College of New Jersey Library collected for a month to determine if data about searching behaviours could be used to improve the catalog interface and inform plans to update the library's web site. The results show that library users employed an average of 2.6 terms in a search and 31.7 percent search by Title. They also found failed queries remained problematic, as a full one-third of searches resulted in zero hits.

Niu and Hemminger (2010) analysed the University of North Carolina library catalog and Phoenix Public library catalog to investigate people's searching behaviour with focus on use of faceted catalog in multiple search systems - an academic library and a public library. They found people do incorporate facets when they are searching through a faceted catalog for either the academic or the public library.

It is quite obvious that user behaviour on OPAC’s have been extensively studied since it was the first major transition from traditional retrieval ways and considering the problems it faced at the early stage, there was the need to improve the systems (Jansen, 2006). However, as it is with technology that keeps improving OPAC’s user behaviour cannot be generalised for current information architecture of library information retrieval systems.

On Archives, Zhang and Kamps (2010) analysed log from National Archive Netherlands archives for user behaviour focusing on archive database of a library and its users. They categorised users into types and found that both the experts and the novices are best served by the same type of system despite significant differences in search episodes reflected by their specific information requests and choice of results to inspect in detail.

Although there exists a kind of relationship between the library and the archive, primary functions are different as archives are used to preserve records which is just a part of library functions hence user behaviour on this systems cannot generalised as for the library (Zhang and Kamps, 2010).

Some researchers have used transaction log analysis but have not focused on user behaviour but on usage, the use of logs was to assess the level and extent to which the resources are being patronized, and which particular resources and services users find most beneficial and reported that access to the library's scholarly material is the predominant reason why users visit the website (Asunka et al., 2009).

Similarly, Jin Young et al. (2012) analysed logs from the Open Library a globally accessible digital library that can also be accessed via external web search engines for user behaviour with focus on cross sectional user behaviour on multiple search systems. They found significant difference between internal and external searchers i.e. those who accessed the digital library directly and those who accessed the library from external system.

Finally, some research focus on comparison of user behaviour across electronic environments Wolfram (2008) revealed that although web-based search facilities may appear similar, users do engage in different search behaviours after analysing query logs from four different environments (Bibliographic database, OPAC system, Search Engines and Specialised Search Service) where the bibliographic databank search environment resulted in the most parsimonious searching similar to a search engine.

2.6 Summary

It is evident that understanding user behaviour on general purpose search engines have been extensively studied and researches on these search engines have extended beyond general search characteristics to relating user behaviour with emotions and comparing results. Multimedia search engine have also been studied which includes image, audio and also video retrieval systems. Domain specific search engines have also been relatively studied for user behaviour particularly health information retrieval systems although other domains have been studied but not as much as the medical domain.

For academic libraries in general, transaction logs analysis to understand user behaviour only became possible at the time of OPACs and hence lots of transaction log analysis researches exist to understand user behaviour to improve the systems. However the new information architecture has only recently been adopted for information retrieval and no transaction log studies has been done on it holistically.

There is an understanding that user behaviour across these electronic environments could be similar but it is not consistent. Knowledge exists for user behaviours on various parts of library information retrieval systems through its evolution however the holistic user behaviour does not exist for current library information systems in a domain specific context - academic. This research will extend existing knowledge on information behaviour on web based retrieval systems and provide a foundation for further study of user-system interaction in an academic domain context.

3. METHODOLOGY

3.1. Chapter Overview

The aim of this study is to provide insight into information seeking behaviour of users on an academic library search engine through transaction log analysis. Jansen (2006) describes this approach as a method that discovers theories or models from data, that are grounded in observations of the "real world," rather than being abstractly generated and presents multiple variables for research purposes and these variables depends on the research aims and objectives.

To further establish the suitability of transaction log analysis for this study, the chapter starts with a rationale and description of the process, proceeding to description of the research settings – the University of Strathclyde library, its users, its electronic information retrieval system called Strathclyde University Primo (SUPrimo) and its transaction logs in order to facilitate a clearer understanding of the process in the context of this study. This description is based on publicly available technical documents and private internal reports generated through authorised system access for the research purpose.

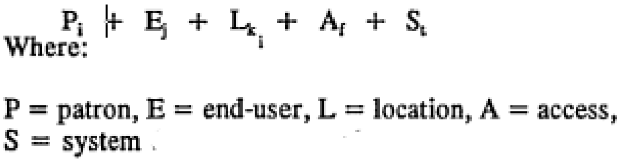
Data was analysed through an iterative process of quantitative and qualitative methods as Jansen et al. (2009) encourages the use of hybrid research designs that combine highly quantitative with qualitative approaches to arrive at a deep understanding of what is actually going on with the information seekers. The limits and limitations of the method are also highlighted.

3.2 Transaction Log Analysis (TLA) Process

Information seeking behaviour has been intelligently studied through various methods - questionnaires, interviews, directed observations, transaction log analysis etc. either singly or in combination, particularly transaction log analysis in combination with other methods as is recommended by researchers (Hancock-Beaulieu et al., 1990). However when a transaction logging software package that included online questionnaires which was able to gather searcher responses to enhance transaction log analysis of browsing behaviours was developed in an effort to address these issues, it took away the unobtrusiveness (one of the strengths of the method) of the transaction log approach (Jansen, 2006).

Transaction log analysis has since become an accepted and popular research method for understanding user behaviour on the web. However a consistent procedure has not evolved, or been adopted or accepted as the standards for this method that can be replicated (Jansen, 2006). This may be attributed mainly to the large volumes and complex data sets of interactions contained in these logs, a conceptual model for analysing dependent variables will be difficult to develop (Kaske, 1993), whilst terms and metrics may not be defined in sufficient detail to enable effective communication of results (Kurth, 1993). It could also be attributed to the varied problems or research questions that each study sought to answer (e.g. request issued, system response time, error rates, use of help facilities etc.) (Peters, 1993).

Efforts to create a procedure for conducting transaction log analysis as a research methodology includes Kaske (1993) and Jansen (2006). Kaske (1993) in the early years of transaction log analysis on library OPAC systems proposed a model of variables described as a logical map to be used specifically on library logs based on the physical variables involved in the interaction - the human interface and system factors (figure 3.1) however it does not includes actions which result from the interactions among the variables which is the critical measure for understanding user behaviour on the web.



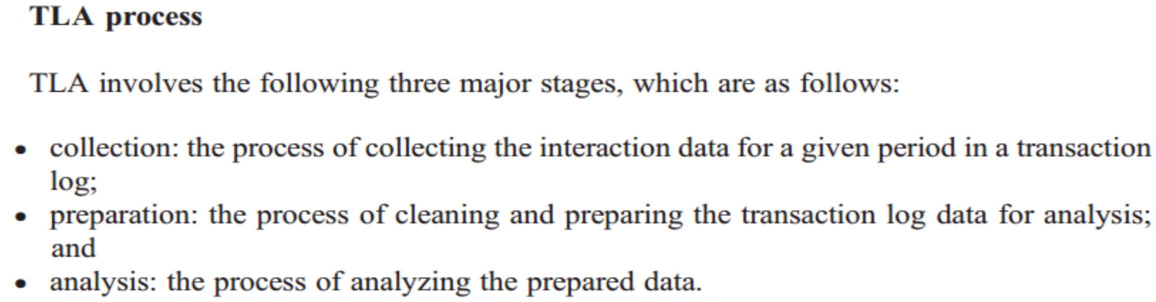
$$P_i + E_j + L_{k_i} + A_f + S_t$$

Where:

P = patron, E = end-user, L = location, A = access,
S = system

Figure 3.1 Kaske proposed model

Jansen's (2006) transaction log analysis three stage approach of "data collection, data preparation and data analysis" is based on standard transaction log format of a relational database (figure 3.2). He argued that the analysis on this format can be done in incremental portions and additional analysis steps can be easily added building off what has already been done which records more advantages than other formats which includes text files and text processing scripts formats where the analysis can be done in one pass however, if additional analysis needs to be done, the whole data set must be re-analyzed. Furthermore that this process can be modified to suit research aims as long as research questions are articulated and data collected from logs are in a standard format.



TLA process

TLA involves the following three major stages, which are as follows:

- collection: the process of collecting the interaction data for a given period in a transaction log;
- preparation: the process of cleaning and preparing the transaction log data for analysis; and
- analysis: the process of analyzing the prepared data.

Figure 3.2 Jansen three stage process

Jansen's (2006) three-stage process for transaction log analysis is the commonly used procedure as have been reported in the literature and is adopted for this study primarily because of the large scale data it presents as shown in SUPrimo transaction logs (that cannot be achieved with surveys and interviews) which enable macro analysis of aggregate user data and patterns and micro analysis of individual search patterns (Jansen et al., 2009). More so adopting the method will enable a comparison to findings from research on general purpose information systems and to allow for cross-validation, which are criticisms of transaction log (Markey, 2007).

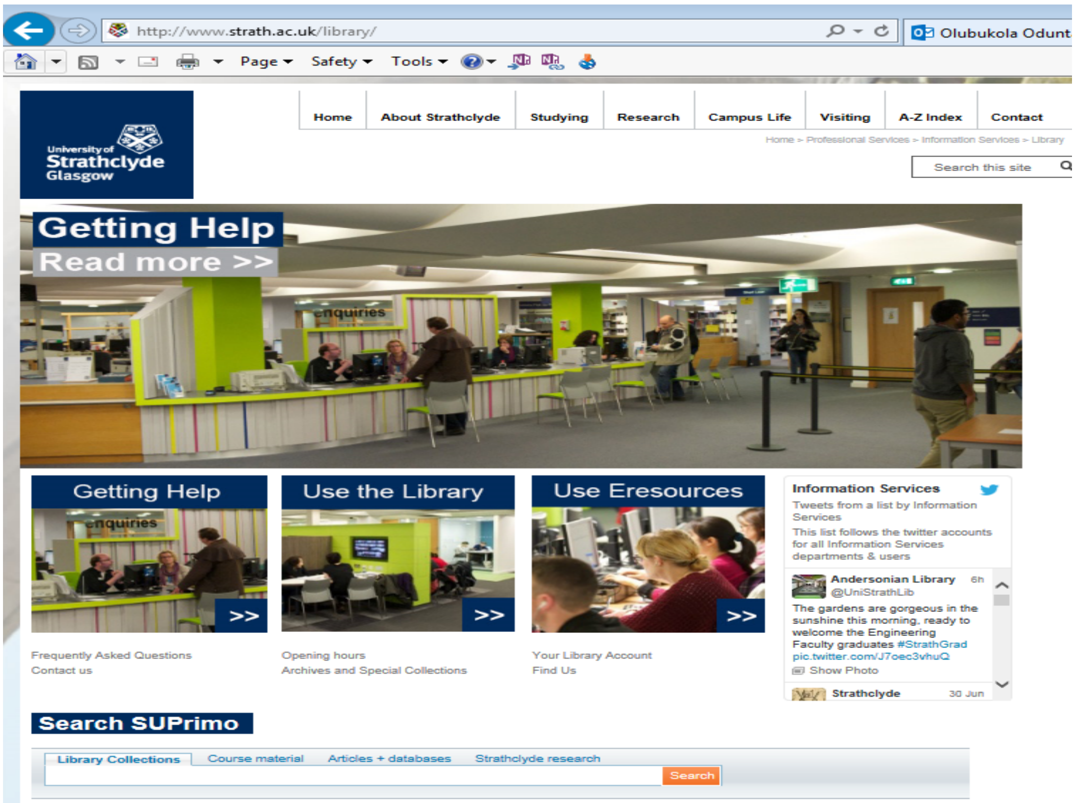
Furthermore Jansen's (2006) process is appropriate because the aim of this study to characterize the behaviour of users and to give insight into the content of their searches on the university of Strathclyde library information retrieval systems by answering research questions which includes how users issue queries to the SUPrimo search engine; this research questions can be answered through analysis of data collected from SUPrimo transaction log, more so SUPrimo transaction logs are in the relational database format on which Jansen process was based.

3.3 University of Strathclyde Library (Andersonian Library)

The study was carried out on University of Strathclyde library also known as the Andersonian Library was created with the aim to provide access to materials and information resources which will help student studies and research in a supportive physical and with technology advancement, virtual learning environment (Strathclyde Library, 2014). The Andersonian library houses more than 2,000 reader places, 450 computer places and extensive Wi-Fi zones for laptop use. It has around one million print volumes as well as access to over 540,000 electronic books, 239 databases and over 38,000 ejournals that can be used 24/7 from any suitably enabled computer (Strathclyde Library, 2014).

The users of the library are primarily the staff and student of the University of Strathclyde who can access the library online services on and off campus with access to e-services and resources through designated user ids and passwords or in some cases remote access is available only via the proxy server. It also allows walk-in users in varying categories but access to electronic resources is only authorised within the physical premises of the library. The University of Strathclyde records tens of thousands of full-time and part-time students yearly - undergraduate and post graduate across four departments – Engineering, Humanities and Social Sciences, Science and Strathclyde Business School. It has thousands of staff across a range of occupations (University of Strathclyde, 2014).

The Andersonian library and its resources can be accessed with a web based interface. The interface comprises of homepage which contains a search function called Strathclyde University Primo (SUPrimo) (figure 3.3), links to numerous internal and external resources of the library and information and services offered by the library. It also includes link to social media pages of the library and related departments of the University (Strathclyde Library, 2014).



Information Services

Tweets from a list by Information Services

This list follows the twitter accounts for all Information Services departments & users

Andersonian Library @UniStrathLib 6h

The gardens are gorgeous in the sunshine this morning, ready to welcome the Engineering Faculty graduates #StrathGrad pic.twitter.com/J7oec3vhuQ

Show Photo

Strathclyde 30 Jun

Figure 3.3 Andersonian library homepage

3.4 Strathclyde Library Information Retrieval System (SUPrimo)

Strathclyde University Primo (SUPrimo) is the University of Strathclyde library information retrieval system for the discovery and delivery of local and remote resources, such as books, journal articles, and digital objects; in the transformation from the traditional library catalogue towards a new information architecture for libraries (Ex-Libris, 2014).

SUPrimo launched in July 2009 was a migration to new information architecture from an Online Public Access Catalog with the objective of enhancing usability and accessibility to the library's collections, resources and services through modern information architecture (Library, 2014, Strathclyde, 2014, Strathclyde Library, 2014). This new architecture allows student and staff worldwide to search for books, e-books, print and electronic articles, digital media, and other types of resources using a single interface.

SUPrimo includes a search box for inputting queries and supports Boolean operators (e.g. and, or, not), proximity operators (e.g. /w, /n), wildcard operators (e.g. *), and exact word match operators (e.g. “”). Its searches are based on the exact keywords a user enters in a query. Capitalization is ignored with the exception of logical commands AND, OR, and NOT (Joint, 2008).

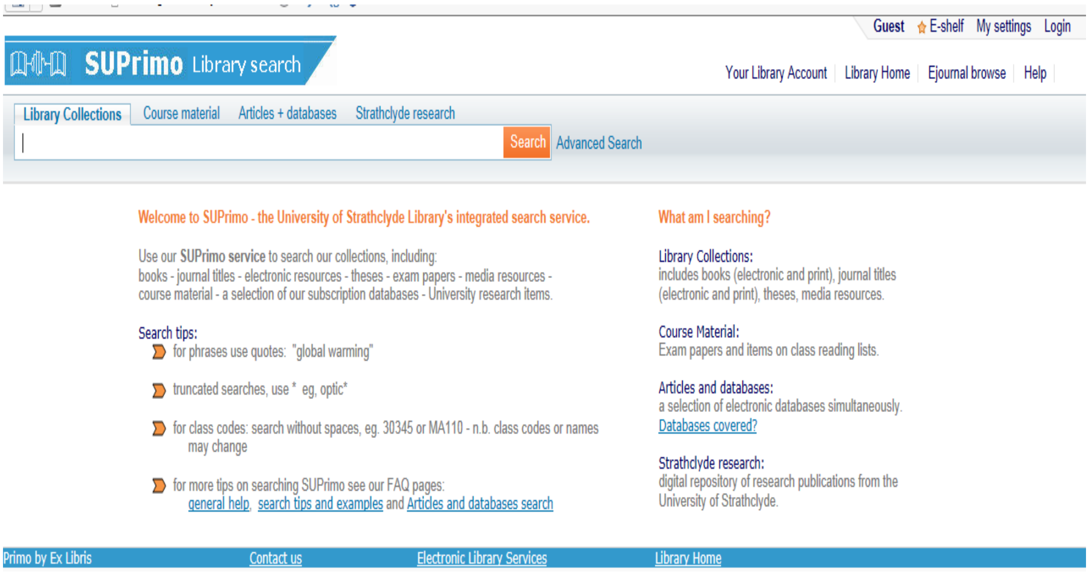
SUPrimo allows user to create virtual shelf where viewed journals or books are recorded for reference purpose, these can be seen on the top right corner of the search interface (Strathclyde Library, 2014). There are four search tabs for SUPrimo simple search (figure 3.4), where library-collections is the default that search the university library for its local resources and remote resources:

Library collections searches books (electronic and print), journal titles (electronic and print), database titles, theses, exam papers, media resources etc.

Course materials searches digital exam papers and materials from student reading lists.

Articles + databases searches a variety of online databases and services simultaneously within a specific subject grouping, a drop down box provides options for specific source selection. Some databases can be searched individually with advance search features.

Strathclyde Research searches research publications from the University of Strathclyde.



Welcome to SUPrimo - the University of Strathclyde Library's integrated search service.

Use our SUPrimo service to search our collections, including:
books - journal titles - electronic resources - theses - exam papers - media resources - course material - a selection of our subscription databases - University research items.

Search tips:

- for phrases use quotes: "global warming"
- truncated searches, use * eg, optic*
- for class codes: search without spaces, eg. 30345 or MA110 - n.b. class codes or names may change
- for more tips on searching SUPrimo see our FAQ pages: general help, search tips and examples and Articles and databases search

What am I searching?

Library Collections:
includes books (electronic and print), journal titles (electronic and print), theses, media resources.

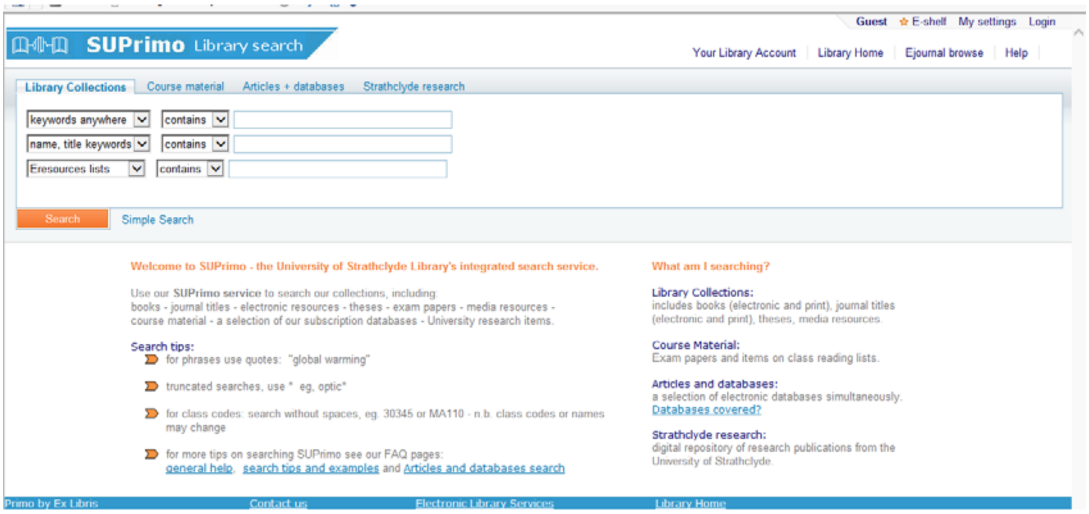
Course Material:
Exam papers and items on class reading lists.

Articles and databases:
a selection of electronic databases simultaneously.
Databases covered?

Strathclyde research:
digital repository of research publications from the University of Strathclyde.

Figure 3.4 SUPrimo interface

SUPrimo has the ability to handle known-item searching, single-word titles, phrase searching and its advanced search features (figure 3.5) offers keyword searching for author, title, subject and shelf mark thereby providing quick access to local content as a result (Strathclyde Library, 2014).



Welcome to SUPrimo - the University of Strathclyde Library's integrated search service.

Use our SUPrimo service to search our collections, including: books - journal titles - electronic resources - theses - exam papers - media resources - course material - a selection of our subscription databases - University research items.

Search tips:

- for phrases use quotes: "global warming"
- truncated searches, use * eg. optic*
- for class codes: search without spaces, eg. 30345 or MA110 - n.b. class codes or names may change
- for more tips on searching SUPrimo see our FAQ pages: general help, search tips and examples and Articles and databases search

What am I searching?

Library Collections:
includes books (electronic and print), journal titles (electronic and print), theses, media resources.

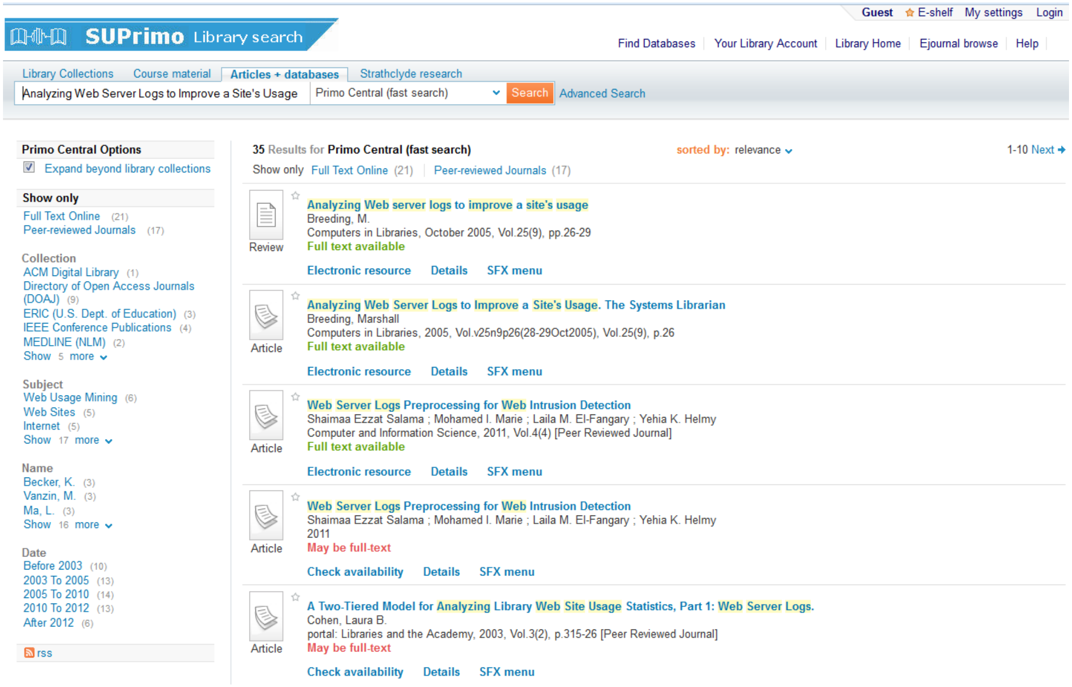
Course Material:
Exam papers and items on class reading lists.

Articles and databases:
a selection of electronic databases simultaneously.
Databases covered?

Strathclyde research:
digital repository of research publications from the University of Strathclyde.

Figure 3.5 SUPrimo advanced search interface

The results page (figure 3.6) includes faceted navigation that help users refine their search and quickly zero-in on the most relevant results. Furthermore the results are presented by content type with the availability and location and also the total number of result is displayed (Strathclyde Library, 2014).



Primo Central Options
☑ Expand beyond library collections

Show only
Full Text Online (21)
Peer-reviewed Journals (17)

Collection
ACM Digital Library (1)
Directory of Open Access Journals (DOAJ) (9)
ERIC (U.S. Dept. of Education) (3)
IEEE Conference Publications (4)
MEDLINE (NLM) (2)
Show 5 more

Subject
Web Usage Mining (6)
Web Sites (5)
Internet (5)
Show 17 more

Name
Becker, K. (3)
Vanzin, M. (3)
Ma, L. (3)
Show 16 more

Date
Before 2003 (10)
2003 To 2005 (13)
2005 To 2010 (14)
2010 To 2012 (13)
After 2012 (6)

35 Results for Primo Central (fast search) sorted by: relevance 1-10 Next

Show only Full Text Online (21) | Peer-reviewed Journals (17)

Review
Analyzing Web server logs to improve a site's usage
Breeding, M.
Computers in Libraries, October 2005, Vol.25(9), pp.26-29
Full text available
Electronic resource Details SFX menu

Article
Analyzing Web Server Logs to Improve a Site's Usage. The Systems Librarian
Breeding, Marshall
Computers in Libraries, 2005, Vol.v25n9p26(28-29Oct2005), Vol.25(9), p.26
Full text available
Electronic resource Details SFX menu

Article
Web Server Logs Preprocessing for Web Intrusion Detection
Shaimaa Ezzat Salama ; Mohamed I. Marie ; Laila M. El-Fangary ; Yehia K. Helmy
Computer and Information Science, 2011, Vol.4(4) [Peer Reviewed Journal]
Full text available
Electronic resource Details SFX menu

Article
Web Server Logs Preprocessing for Web Intrusion Detection
Shaimaa Ezzat Salama ; Mohamed I. Marie ; Laila M. El-Fangary ; Yehia K. Helmy
2011
May be full-text
Check availability Details SFX menu

Article
A Two-Tiered Model for Analyzing Library Web Site Usage Statistics, Part 1: Web Server Logs.
Cohen, Laura B.
portal: Libraries and the Academy, 2003, Vol.3(2), p.315-26 [Peer Reviewed Journal]
May be full-text
Check availability Details SFX menu

Figure 3.6 SUPrimo results page

Each and every action carried out by the user on the interface is recorded in web server transaction log at the back end of the interface. The search, the results page view, the facets, if query suggestion was used, if advanced search tab was used, if it was a local or remote transaction etc., how many users used the e-shelf facilities, how many users looked at the next page of results amongst others are recorded in transaction logs (Ex-libris, 2011) and they form the multi-variable that were described in the Literature.

These transaction log files are often very large and are not easily interpreted that software products have been developed for the purpose of summarizing log data and generating simple statistical reports of visits, responses, page views from the database etc. The contents of these summaries logs are dependent on the interface owner and the provider (Jansen, 2006). However, some researchers agree that it is useful to manually study the raw logs themselves, as this can potentially yield a lot more information regarding the user activities on web interfaces (Breeding, 2005). Therefore for the purpose of this study, the generic transaction logs and summary transaction logs were retrieved.

3.5 SUPrimo Transaction Logs

SUPrimo back end comprises of a structured system of logs where every transaction that occurs between the SUPrimo web server and any networked computer through websites hosted by the server are recorded and retrieved (figure 3.7). SUPrimo transactions are generically recorded in the server log from where two summary logs - search log and click events log (contents decided by the University authorities and the provider) are populated into a relational database where simple statistical reports are drawn for site usage analysis by the University (Strathclyde Library, 2014).

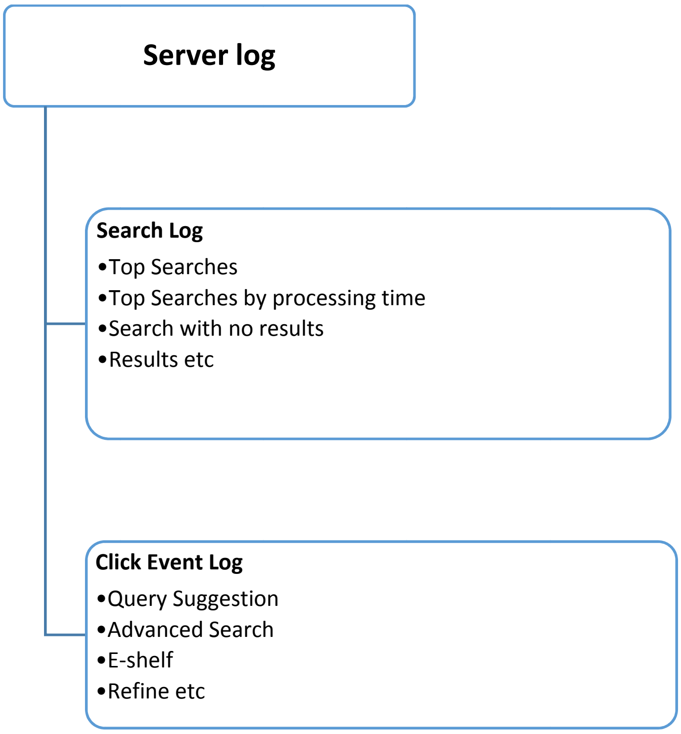
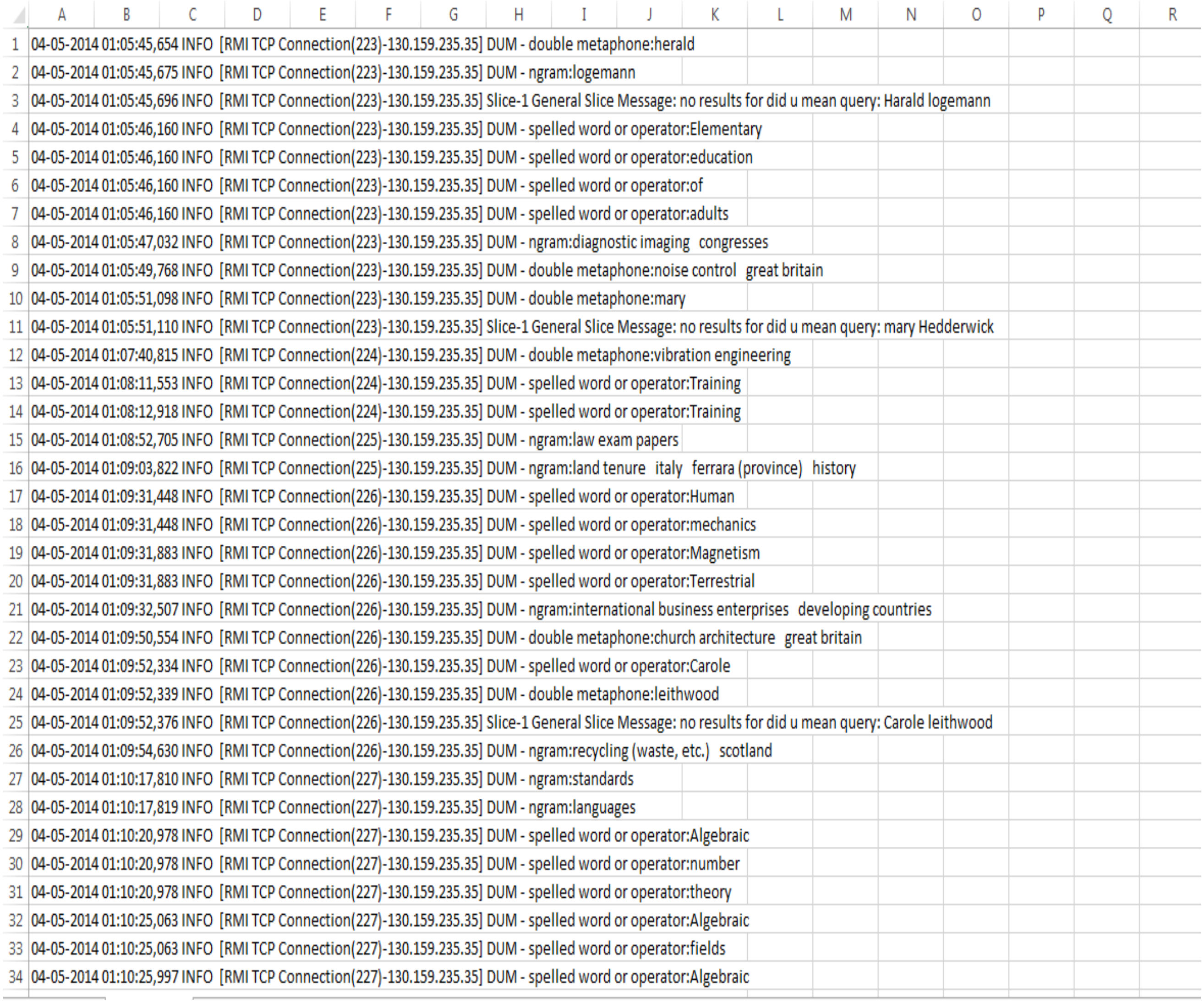


Figure 3.7 SUPrimo transaction logs architecture

3.5.1 SUPrimo Web Server Log

SUPrimo web server logs (figure 3.8) are the very large complex generic logs of all user interactions on the SUPrimo interface. They can be retrieved from the server as text files thereby providing an up-to-date chronological list of the individual request-response transactions however because of its complexity is summarised into search log and click events log in a relational database.



```
04-05-2014 01:05:45,654 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - double metaphone:herald
04-05-2014 01:05:45,675 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - ngram:logemann
04-05-2014 01:05:45,696 INFO  [RMI TCP Connection(223)-130.159.235.35] Slice-1 General Slice Message: no results for did u mean query: Harald logemann
04-05-2014 01:05:46,160 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - spelled word or operator:Elementary
04-05-2014 01:05:46,160 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - spelled word or operator:education
04-05-2014 01:05:46,160 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - spelled word or operator:of
04-05-2014 01:05:46,160 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - spelled word or operator:adults
04-05-2014 01:05:47,032 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - ngram:diagnostic imaging  congresses
04-05-2014 01:05:49,768 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - double metaphone:noise control  great britain
04-05-2014 01:05:51,098 INFO  [RMI TCP Connection(223)-130.159.235.35] DUM - double metaphone:mary
04-05-2014 01:05:51,110 INFO  [RMI TCP Connection(223)-130.159.235.35] Slice-1 General Slice Message: no results for did u mean query: mary Hedderwick
04-05-2014 01:07:40,815 INFO  [RMI TCP Connection(224)-130.159.235.35] DUM - double metaphone:vibration engineering
04-05-2014 01:08:11,553 INFO  [RMI TCP Connection(224)-130.159.235.35] DUM - spelled word or operator:Training
04-05-2014 01:08:12,918 INFO  [RMI TCP Connection(224)-130.159.235.35] DUM - spelled word or operator:Training
04-05-2014 01:08:52,705 INFO  [RMI TCP Connection(225)-130.159.235.35] DUM - ngram:law exam papers
04-05-2014 01:09:03,822 INFO  [RMI TCP Connection(225)-130.159.235.35] DUM - ngram:land tenure  italy  ferrara (province)  history
04-05-2014 01:09:31,448 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - spelled word or operator:Human
04-05-2014 01:09:31,448 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - spelled word or operator:mechanics
04-05-2014 01:09:31,883 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - spelled word or operator:Magnetism
04-05-2014 01:09:31,883 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - spelled word or operator:Terrestrial
04-05-2014 01:09:32,507 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - ngram:international business enterprises  developing countries
04-05-2014 01:09:50,554 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - double metaphone:church architecture  great britain
04-05-2014 01:09:52,334 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - spelled word or operator:Carole
04-05-2014 01:09:52,339 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - double metaphone:leithwood
04-05-2014 01:09:52,376 INFO  [RMI TCP Connection(226)-130.159.235.35] Slice-1 General Slice Message: no results for did u mean query: Carole leithwood
04-05-2014 01:09:54,630 INFO  [RMI TCP Connection(226)-130.159.235.35] DUM - ngram:recycling (waste, etc.)  scotland
04-05-2014 01:10:17,810 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - ngram:standards
04-05-2014 01:10:17,819 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - ngram:languages
04-05-2014 01:10:20,978 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - spelled word or operator:Algebraic
04-05-2014 01:10:20,978 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - spelled word or operator:number
04-05-2014 01:10:20,978 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - spelled word or operator:theory
04-05-2014 01:10:25,063 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - spelled word or operator:Algebraic
04-05-2014 01:10:25,063 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - spelled word or operator:fields
04-05-2014 01:10:25,997 INFO  [RMI TCP Connection(227)-130.159.235.35] DUM - spelled word or operator:Algebraic
```

Figure 3.8 SUPrimo web server log

3.5.2 SUPrimo Click Event Log

The click event log tables in the database is the summary of click interactions such as how many uses use the e-shelf, how many users use the query suggestion (Did you mean) etc. Although the click event log table is not the focus of this study, a brief description of the click event tables was necessary to emphasize that depth of information that can be presented in transaction logs. Tables 3.1 and 3.2 shows the description of click event fields recorded in the logs and event types respectively.

Field	**Description**
ID	System generated primary key
EVENT_DATE	Date when stats were written to the summary table
EVENT_TYPE	Type of event: Search problem and the use of facilities provided by the interface. Includes use of facets, e-shelf, did you mean (query suggestion), how many users sign in etc.
CLICK_VALUE	In some cases there is additional information: e.g. **REFINE** – the facet selected **GETIT!** – the resource type selected and record number of the record selected **ADD TO ESHELF** – the resource type selected
CLICK_COUNT	Count per event
VIEW	Primo view in which user event occurred
INSTITUTION	Active user Institution at time of the event
ON_CAMPUS	Location of user at time of search i.e. true or false
USER_GROUP	User Group of user as returned by PDS In our database values are either blank, 1 or GUEST

Table 3.1 Click event log fields

Click Event Types

Basic Search	Previous Page
Start Session	Add to eShelf
Refine	Find DataBases A-Z
Next Page	GetIt!Link2
GetIt!Link1	Display Tags and Reviews
Search	Find DataBases Link
Database	Find DataBases Simple
Data	Full Table Scan
Facets	Find DataBases Search
Did you mean	hlp
Full Display	Send an email
DS	Find DataBases Advanced
Display Details Tab	Details Print
Advanced Search	Eshelf Print
Locations	Remote More
GetIT!	Find DataBases Info
didym	Save Search
SQL Average	Create Alert
Sign-in	
eShelf Page	

Table 3.2 Click event types

3.5.3 SUPrimo Search Log

SUPrimo search log, the second type of summary log generated from the generic logs presents the variables to be examined in the context of this study. It stores information about the search of the user and response from the system including search strings, results and response time.

Individual records are logged as a search count with a unique identification number (ID) other than the IP addresses recorded in the web server logs and systematically populates four summaries tables in the relational database from which different views of fields in the database are created. These summaries are updated hourly or every 1000 queries whichever happens first. Each record includes timestamp, search string (request), the scope where was searched (if local or remote resources), the results and the response time (table 3.3).

Field	Description
ID	System generated primary key
SUMMARY_TYPE	These are the summary statistical reports generated for usage analysis by the university Individual records: **SEARCH_COUNT** Summary records*: **TOP_SEARCHES_BY_PROCESSING_TIME_SUMMARY** **SEARCH_WITH_NO_RESULTS_COUNT** **TOP_SEARCHES_SUMMARY** **TOP_SEARCHES_WITH_NO_RESULTS_SUMMARY** *these summary records are updated hourly or every 1000 queries, whichever happens first
SEARCH_STRING	Otherwise known as the **query** entered into the search engine by the user
SCOPE_NAME	Scope of the request e.g. scope: (SU) scope: (EXAM) scope : (SUREADING)
SCOPE_TYPE	Scope Type e.g. local : local search remote : remote search ds : deep search (primo central search) The default on the search engine
SEARCH_COUNT	Number of searches in the monitored period* *Period is either 1000 queries or 1 hour – whichever happens first.
AVERAGE_RESULTS	Average number of rows in the result set
SUMMARY_TIMESTAMP	If summary type = 'SEARCH_COUNT' this is the actual timestamp of the search For all other types the timestamp is the date time record saved to the table
VERAGE_RESPONSE_TIME	Average time elapsed for search response time
AVERAGE_FULL_TIME	Average total elapsed time required to process the search request, including the search response time
SOURCE_VIEW	Primo view in which search was done e.g. SUV01
SOURCE_INSTITUTION	Active user institution e.g. Strathclyde University
SOURCE_ON_CAMPUS	Location of user at time of search i.e. true or false
SOURCE_USER_GROUP	User Group of user as returned by PDS In our database values are either 1 or GUEST

Table 3.3 SUPrimo search logs fields

3.6 Data Collection

Data was collected from SUPrimo transaction logs database of interactions that occurred between April 01, 2014 and May 31, 2014, a 61 days period. The collection was conveniently and purposely chosen so as to include the time period towards before and during exams (exam start date: May 01, 2014) so as to give greater confidence in the findings and furthermore for statistical validity and completeness of understanding of user behaviour in between major academic periods.

Data retrieved contained 1,350,846 interactions of search and click event and as they were arranged in chronological order, each month's record was broken up into sections by month, and each section exported into Microsoft Excel as separate spreadsheet file that became the source file into other data processing and analysis software used throughout the process.

The variables specific to this study are described below although the focus is on “search string/query”.

Search String: The query entered into the search engine by the user

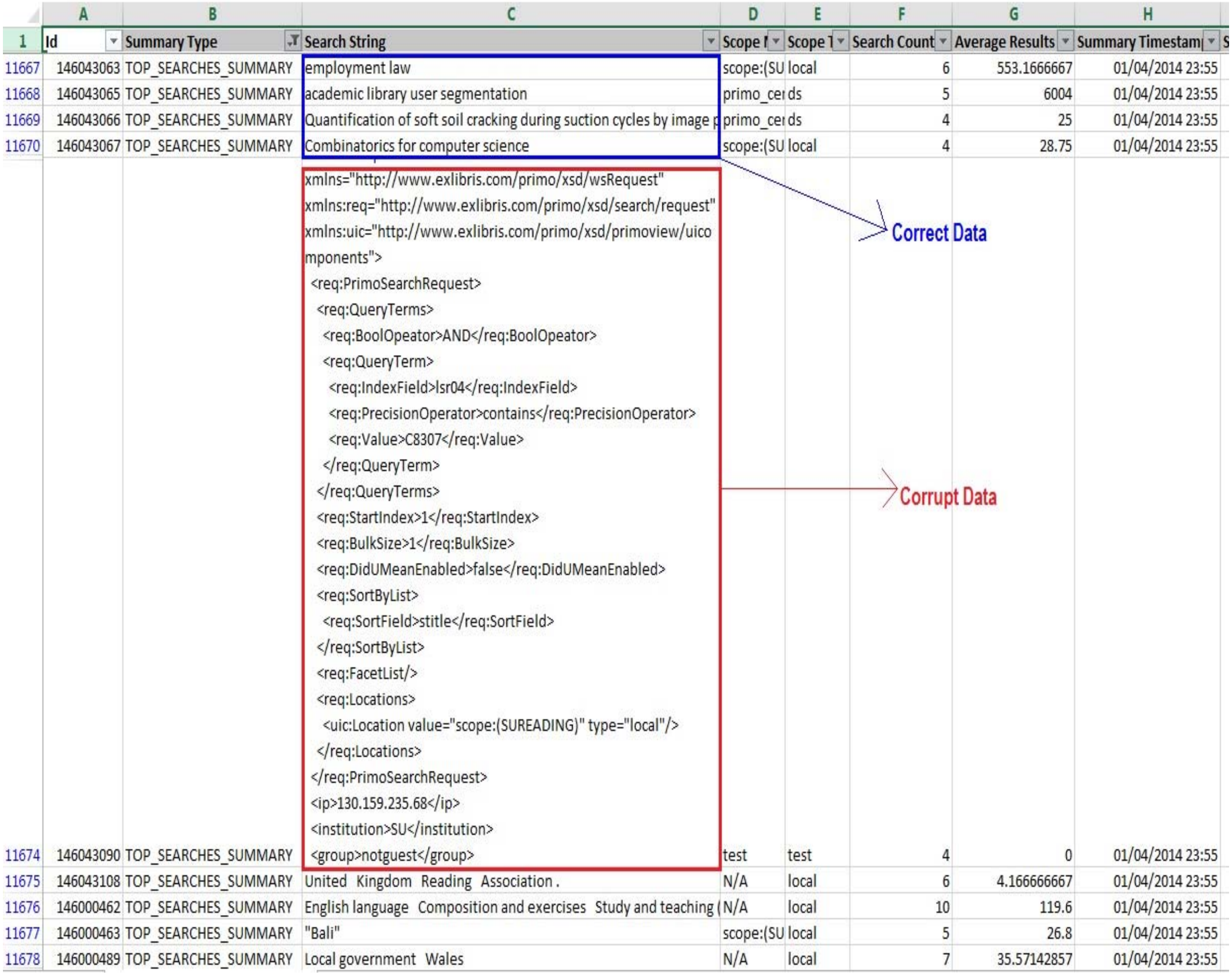
Search Count: Number of searches for each search string in the monitored period

Average Result: Average number of rows in the result set for each search count.

Summary Timestamp: Date time individual record was saved to the table.

3.7 Data Preparation

The dataset was manually filtered using Microsoft excel to remove from the logs; corrupt data (figure 3.12) which arise from software errors in logging data to the database and irrelevant data because the log records all other interactions on the interface apart from search interactions.



	A	B	C	D	E	F	G	H
1	Id	Summary Type	Search String	Scope I	Scope T	Search Count	Average Results	Summary Timestam
11667	146043063	TOP_SEARCHES_SUMMARY	employment law	scope:(SU	local	6	553.1666667	01/04/2014 23:55
11668	146043065	TOP_SEARCHES_SUMMARY	academic library user segmentation	primo_cer	ds	5	6004	01/04/2014 23:55
11669	146043066	TOP_SEARCHES_SUMMARY	Quantification of soft soil cracking during suction cycles by image p	primo_cer	ds	4	25	01/04/2014 23:55
11670	146043067	TOP_SEARCHES_SUMMARY	Combinatorics for computer science	scope:(SU	local	4	28.75	01/04/2014 23:55
11674	146043090	TOP_SEARCHES_SUMMARY	xmlns="http://www.exlibris.com/primo/xsd/wsRequest" xmlns:req="http://www.exlibris.com/primo/xsd/search/request" xmlns:uic="http://www.exlibris.com/primo/xsd/primoview/uico mponents"> <req:PrimoSearchRequest> <req:QueryTerms> <req:BoolOpeator>AND</req:BoolOpeator> <req:QueryTerm> <req:IndexField>lsr04</req:IndexField> <req:PrecisionOperator>contains</req:PrecisionOperator> <req:Value>C8307</req:Value> </req:QueryTerm> </req:QueryTerms> <req:StartIndex>1</req:StartIndex> <req:BulkSize>1</req:BulkSize> <req:DidUMeanEnabled>false</req:DidUMeanEnabled> <req:SortByList> <req:SortField>stitle</req:SortField> </req:SortByList> <req:FacetList/> <req:Locations> <uic:Location value="scope:(SUREADING)" type="local"/> </req:Locations> </req:PrimoSearchRequest> <ip>130.159.235.68</ip> <institution>SU</institution> <group>notguest</group>	test	test	4	0	01/04/2014 23:55
11675	146043108	TOP_SEARCHES_SUMMARY	United Kingdom Reading Association .	N/A	local	6	4.166666667	01/04/2014 23:55
11676	146000462	TOP_SEARCHES_SUMMARY	English language Composition and exercises Study and teaching (	N/A	local	10	119.6	01/04/2014 23:55
11677	146000463	TOP_SEARCHES_SUMMARY	"Bali"	scope:(SU	local	5	26.8	01/04/2014 23:55
11678	146000489	TOP_SEARCHES_SUMMARY	Local government Wales	N/A	local	7	35.57142857	01/04/2014 23:55

Figure 3.9 SUPrimo search log showing corrupt and correct data

3.8 Data Analysis

Data was quantitatively analysed in a continuous iterative process calculating the interaction metrics to reveal trends and pattern among interactions. The units of analysis were “query” and “term” where term is a string of characters within a query separated by white space or other separator (Jansen, 2006).

3.8.1 Usage Analysis

SUPrimo search log recorded 154638 individual queries summarised into a total 26874 queries with 12378 unique queries i.e. searched once and 14,496 repeated queries i.e. search more than once (table 3.4).

General Queries Summaries	
Total number of Individual queries	154638
Unique queries	12378
Repeated queries	14496
Successful queries	22253
Failed queries	4621
Total Number of Queries examined	26874

Table 3.4 Queries summary

About 83% (22253) of the total queries were successful while 17% (4621) of the queries failed (figure 3.13), where a successful query is defined as one that returned non-zero results and failed query returned zero results according to (Jones et al., 2000).

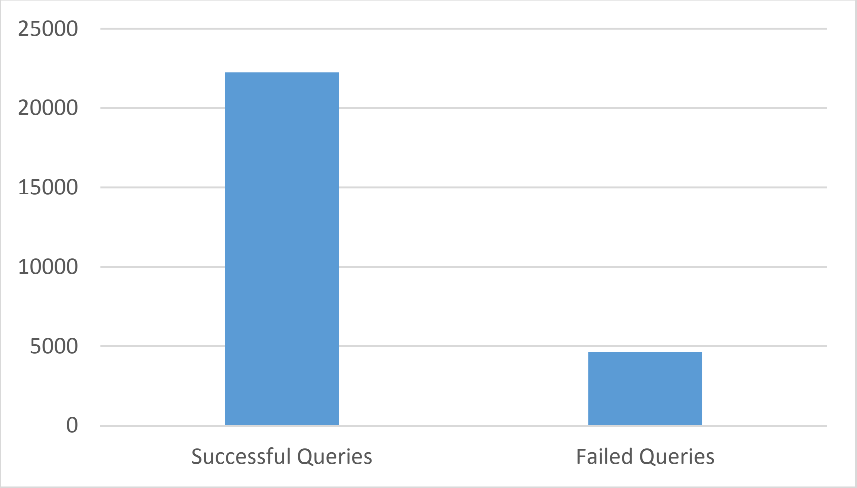


Figure 3.10 Query success and failure comparison

On further breakdown, about 63% of the individual queries were recorded in the month of May during the exam period (May 06 – May 31 2014) with 54% increase in the number of queries when compared to the previous month (table 3.5).

Period	Individual Queries	Percentage
April 01 – April 15	30069	19.4%
April 16 – April 30	32695	21.1%
May 01 – May 15	47762	30.8%
May 16 – May 31	49006	31.6%

Table 3.5 Queries usage analysis

Furthermore, it can be seen that 46% of the increase occurred during the first half of the month (figure 3.14) – the scheduled start date for most exams. This could be attributed to increase in student usage in preparation for the exams and confirms the intuition that student access more information in preparation for exams as it is expected in an academic library setting.

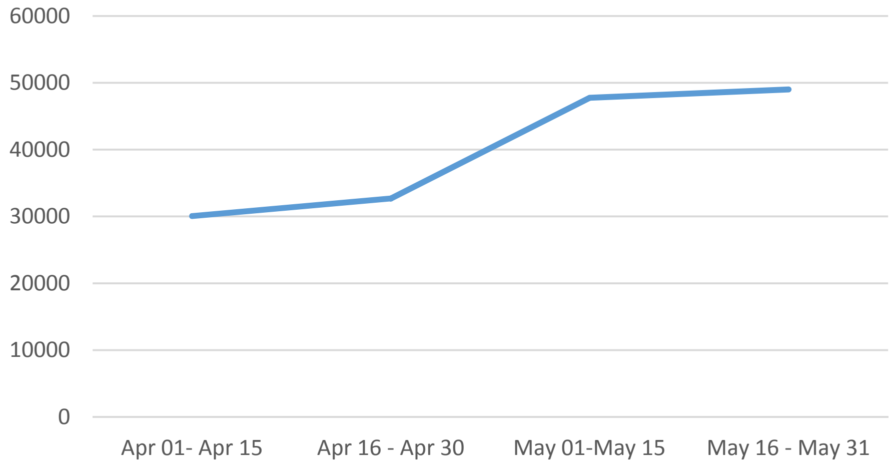


Figure 3.11 Queries usage analysis graphical representation

3.8.2 Queries – Length and Distribution

The overall average length of a query on library information retrieval system was 6.58, even though the most frequently used queries had 2 terms query length. This average is significantly longer than average length of queries on general purpose search engines of between 1 and 3 terms (Jansen and Spink, 2006, Silverstein et al., 1999), on medical search engines of 5 terms (Yang et al., 2011) also on OPAC's and other web information retrieval systems of between 2 and 3 terms (Jones et al., 2000, Lau and Goh, 2006, Moulaison, 2008).

Query length	Frequency	Percentage
1	2926	10.89%
2	5207	19.38%
3	4336	16.13%
4	2915	10.85%
5	1829	6.81%
6	1319	4.91%
7	878	3.27%
8	563	2.09%
9	505	1.88%
10	694	2.58%
11	671	2.50%
12	668	2.49%
13	564	2.10%
14	584	2.17%
15	481	1.79%
16-410	2734	10.17%
Total	**26874**	**100.00%**

Table 3.6 Query length frequency distribution

Interestingly, 57% of the queries had between 1 and 4 query length and about 70% between 1 and 6 query length which can be interpreted that 7 out of 10 queries have between 1-6 terms (table 3.6) and only about 30% of queries have more than 6 terms (Appendix A shows the complete query lengths and their frequency distribution).

As a result, queries with length between 1 and 6 terms were examined and the new average length was 2.97 which although is slightly higher but is then consistent with query length on general purpose search engines, OPAC's and other web information retrieval systems including medical domain. Since these queries constitute the larger percentage of queries on the system, it can therefore be assumed technically that queries on the SUPrimo search engine are short.

Further statistical analysis confirmed the disparateness in the data as the distribution was a positively skewed distribution with a very high skewness of 8.922 (table 3.7). The standard deviation of 7.431 (table 3.7) resulted in the range between "-5.431 to 9.431" as 68% of the total queries. Therefore confirms the mean of the query lengths is not a true representation of a typical query on the SUPrimo search engine.

General Statistical Analysis	
Mean	6.58
Median	4.00
Mode	2
Std. Deviation	7.431
Skewness	8.922

Table 3.7 Statistical analysis results

However the distribution follows a Zipf distribution remarkably well - a power law distribution which holds for languages amongst other things. The frequency of shorter queries on SUPrimo was higher the than the frequency of longer queries (figure 3.15). The result complements previous studies of language distribution and reveals SUPrimo query patterns are similar to Jansen and Spink's (2006) patterns on general multi-purpose search engines, Lau and Goh's (2006) OPAC's query patterns and Bajracharya (2010) query patterns on code search engine, as SUPrimo short queries constituted the fathead of the curve and the long queries constituted the long tail in the distribution (figure 3.15).

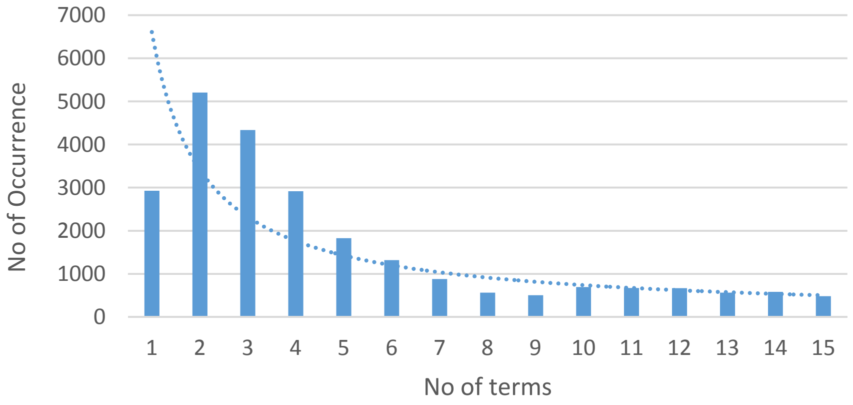


Figure 3.12 Zipf query distribution

3.8.3 Terms – Frequency and Distribution

The total number of terms in the queries were 181420 terms, 73.43% i.e. 133,210 terms constitute unique terms occurrences while 26.57% i.e. 48210 terms constitute common terms occurrences. This culminated into a total of 18583 unique terms, 18082 unique terms and 501 common terms (table 3.8).

Term Occurrence Summaries		
Total occurrence of Unique terms	133210	73.43%
Total occurrence of common terms	48210	26.57%
Total number of terms	**181420**	**100%**
Unique terms	18082	97.35%
Common terms	501	2.69%
Total unique terms	**18583**	100%

Table 3.8 Summary of term occurrence

It can be seen from the percentages that there were more unique terms than common terms (table 3.9) even though the top 10 commonly occurring terms constituted about one-fifth of the total terms (table 3.9). Furthermore these top common terms fits into the Standard English language stop words list defined as a list of high frequency words that describe too many objects and are useless for information retrieval.

Term	Frequency	Percentage
and	7835	4.32%
Of	7028	3.87%
the	5689	3.14%
In	3975	2.19%
A	2703	1.49%
for	1562	0.86%
To	1487	0.82%
On	1206	0.66%
With	889	0.49%
An	736	0.41%
Total	**33110**	**18.25%**

Table 3.9 Top 10 commonly occurring terms

There were 501 commonly occurring terms that included undefined terms that were removed resulting into 18082 unique terms which covered a broad range of field.

The top 100 unique terms were identified and are visually represented below (figure 3.13) in the order of terms with highest frequency the boldest. These terms constituted about 18% of the total number of terms i.e. about 2 out of 10 users use these terms. (Appendix B shows the frequency of occurrence of the top 100 terms).

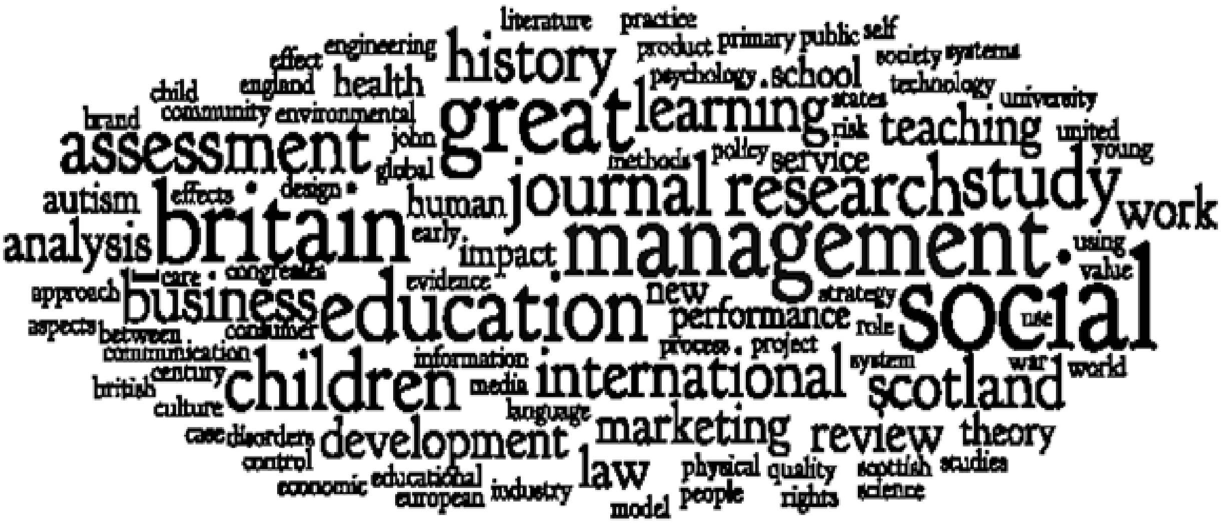


Figure 3.13 Top 100 unique terms

Summarily, this data analysis stage presents quantitative descriptive analysis of the queries. It was established that the usage of the library online system increased during exam period and was particularly concentrated during the first few weeks. The failed queries on the system were about 17% of the total queries on the logs. Queries patterns implied that users on the SUPrimo search engine may use key terms as about 70% of the queries had between 1-6 terms and the unique terms constituted about 90% of total terms.

These findings presents a foundation to further extend the analysis to finding relationships between variables and examining queries by results for definite inferences and possibly categories of users' ultimately resulting in clearer and deeper understanding of users' information seeking behaviour on an academic library.

3.9 Limitations

Transaction log analysis is an inexpensive and non-obtrusive method for collecting reasonable amount of user–system interaction data for understanding information seeking process, considering the current nature of the web. However, as with any method it is not without its limitations and in the context of this study includes:

Firstly, the size and complexity of the data; which required significant processing achieved with various computing resources. Additionally, the logs contained a large amount of undesired entities which had to be identified and removed before further processing.

Secondly, the inherent characteristics of transaction logs such as inaccurate transaction counts where some interactions' events are masked from these logging mechanisms, defining what constitutes a successful transaction etc. More so current Universities libraries offer virtual learning and unrestricted access. Consequently these figures are an underestimation of the total queries but still provides substantial queries required in the context of this study.

Thirdly, on-site public terminals logs do not usually contain information to differentiate one search session from another easily, it usually requires reading the log line by line (Kurth, 1993). SUPrimo log is one such and consequently it is not automatically able to provide information to isolate and characterize individual users search session on the online system, in order to describe the patterns of their use. Furthermore the dates in the summary logs are not discriminatory enough to establish, without a doubt, the query time for each user, thus limiting the analysis to exclude time and invariably session.

Fourthly, demographic data are not collected in SUPrimo logs as other transaction logs in accordance with ethical and legal standards particularly the privacy law. Researchers are in unanimous agreement that there are numerous ways to support individuals’ information seeking on web based information retrieval systems without identifying individuals (Kaske, 1993). As a result, this study excludes any demographic analysis.

Finally, transaction logs have been severely criticised for its cognitive deficiencies which includes its inability to record users’ perception of their searches and user’s satisfaction. SUPrimo log is no exception because data collected only deals with queries entered by users of the online systems, whilst ignoring the information needs that could not be expressed in the search statements and users' satisfaction with the results of their searches. However this intrudes the unobtrusiveness; an inherent strength of the method for unbiased user behaviour and as a result the variables provided in the logs are the limits of this study.

4. ANALYSIS

4.1 Chapter Overview

The research aim was to provide insight into users' information seeking behaviour on SUPrimo search engine by finding the characteristics of queries and the patterns of searching on the academic search engine.

The descriptive analysis from the previous chapter revealed characteristics of SUPrimo users' queries as short since 70% of the search engine users issue queries with 1-6 terms and 97% of total terms were unique terms implying key term search for information. However, the content of these queries and how it affects results are not known. This presents the need to dive into the content of user's queries in order to find the patterns of searching for definitive inferences on user behaviour and evaluate SUPrimo search engine responses.

This chapter thereby extends the analysis in a continuous iterative process to find patterns of searching by classifying users and examining repeated queries using a purposively selected sample from the dataset. The complete dataset was further statistically tested for possible existence and measure of strength of relationship between variables using scatter plots and Pearson correlation coefficient statistical analysis and finally failed queries were examined with respect to correlation results and for other possible characteristics.

Patterns discovered in addition to query characteristics are the "user behaviour" the research sought to understand that will provide evaluation of the systems which help improve the existing online systems or design better ones to support information seeking (Ruthven and Kelly, 2011).

4.2 Queries – Classification

To find patterns of searching on SUPrimo as an academic library search engine, questions such as how do users search the Information retrieval System arises; and following from its description and applicability of SUPrimo as an academic library web based search engine with capacities to search by author, title, subject and shelf mark (Strathclyde Library, 2014), this questions includes: is it by book title? Subject? Author? Consequently a classification of possible queries issued on SUPrimo was necessary.

All queries with 2 and 3 terms were purposively selected for manual examination primarily because a culmination of this queries is over 35% of the total queries recorded during the period (table 3.4) i.e. about 4 in 10 queries have 2 – 3 terms.

Furthermore because 70% of the total queries had an average length of between 2 and 3 terms as it had been reported in the method analysis, as a result they represent good source of data for insight into user behaviour on SUPrimo.

Therefore, from SUPrimo capabilities and content analysis of queries; where content analysis involved making the content of messages manifest through identification of characteristics in as objective way as possible (Ruthven, 2012); three possible categories of queries emerged and are defined as follows:

Topic Search: A string of words separated by space that is specifically not a name of person. This includes book titles and subject search as separating book titles from a user subject queries is not feasible. Queries with name and time period for which information is being sought about and are not authors such as “Paul Chambers 1935-1969” are also included.

Author Search: A string of words separated by space identified as names validated on the SUPrimo search engine. Further included in this category are author queries supported with year and occasionally with topics such as “Fisher 2008. Childcare”. However there is the possibility of omission of names not easily identified and inclusion of names that are not authors.

Shelf-mark Search: These are set of alphanumeric characters together or separated by space used by the university library for identification and location of books in the library. This includes international identification ISBN & ISSN for books and journals respectively.

The results showed that 78% of the users employed topic search, 21% employed author search while less than 1% used the shelf mark search (table 4.1 and figure 4.1) which further emphasizes the likelihood that users of SUPrimo employ key term search. This results is similar to findings from studies on OPAC’s – Lau & Goh 2006 OPAC found that 68% of users employ key word search and Villien Rueda 2007 found that the 49% employ title search and 37% carry out author search amongst other search types.

Categories	Frequency	Percentage
Topic Search	7509	78.69%
Author Search	2016	21.13%
Shelf-mark Search	18	0.19%
Total	9543	100.00

Table 4.1 Summary of query classification

The result from the findings (table 4.1) could be further interpreted as:

1. Users of SUPrimo search engine are looking for topical information to gain available information about a topic not specific books. Intuitively this might be and understandably so considering it as an academic setting for learning; to get as many resources as possible that have information about the topic.
2. SUPrimo queries cannot be classified into the 3 Broder's classification of web queries "navigational", "transactional" and "informational" as with other domain specific search engines (Yang et al., 2011, Bajracharya and Lopes, 2012), however the "topic search" category on SUPrimo can be likened to the "informational" category of general purpose web search engines as the search appears to gain available information about the topic furthermore, 78% of users were in the category and on Broder's classification, the informational category records the highest percentage.
3. On the other hand, it can also be implied that users employs online books more than hard copy books since only 15 users use the shelf mark search which is less than 1% of the sampled population. This result emphasizes Cubitt's (2006) statement about new academic libraries being accessed electronically as a result of advancement in technology. Although there is a possibility that users are not aware of this hard copy specific search option and use the information retrieval system as other general purpose search engines such as google to gain available information-online materials or hard copy.

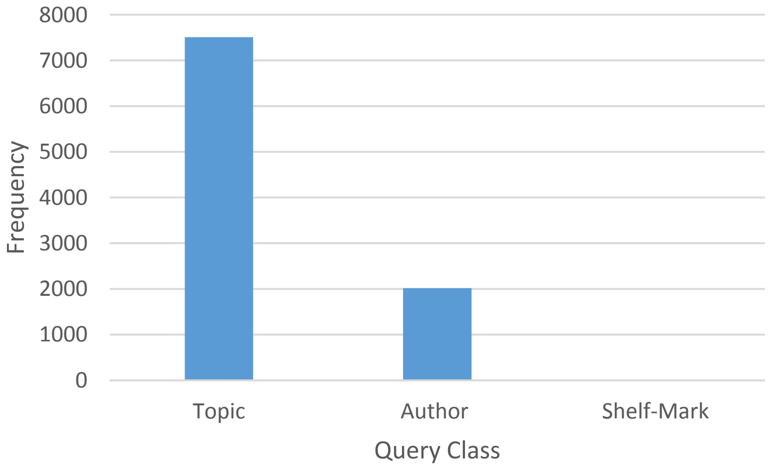


Figure 4.1 Query classification graphical representation

4.3 Queries – Repeated Query

Query occurrence is used to examine the frequency of repeated queries i.e. with more than one search frequency on a search engine. On SUPrimo logs they are identified by the search count variable. They are examined in this quest for understanding user behaviour on SUPrimo because knowledge of the most frequent queries can be further exploited in structuring indexes for information retrieval on web based search engines as they can be reordered to provide more rapid response for common queries i.e. are optimized to handle common requests (Bendersky and Croft, 2009).

The initially selected sample of 2 and 3 terms query was manually analysed in excel through the search count variable and results revealed a high amount of repeated queries totaling 6798 which represented 70% of the sample while unique queries were only 2745 which represented about 29% of the sample (table 4.2). Intuitively, this high occurrence of repeated queries could be expected on an academic library search engines because although there are thousands of students and staff, each students and staff belongs to a particular department thereby potentially resulting in users' groups by department seeking for information which includes books, titles or subjects and authors related to the departments which they belong to consequently a high amount of repeated queries.

Queries	Occurrence	Percentage
Repeated queries	6798	71.24%
Unique	2745	28.76%
Total	9543	100.00%

Table 4.2 Summary of repeated queries

The query occurrence frequency (repeated queries) ranged between 2 and 1305 times with queries containing 2 terms constituting the larger percentage of 56% while queries with 3 terms was 44% and on the whole dataset queries which occurred between 2 and 50 times constituted over 98% over the total queries (table 4.3).

Repeated Query Occurrence	Frequency	Percentage
02-10	4947	72.77%
11-20	1324	19.48%
21-30	281	4.13%
31-40	94	1.38%
41-50	61	0.90%
51-1305	91	1.34%
Total	**6798**	**100.00%**

Table 4.3 Repeated queries occurrence frequency distribution

It had been earlier discovered that the query length frequency distribution of SUPrimo users follows remarkably well the Zipf language distribution law (figure 3.12), this query occurrence distribution further validates the findings as repeated queries decreased with increasing number of occurrence.

Previously in order to understand the length characteristics of query the focus was on the fat head of the distribution curve and this were the short queries as this constituted the largest number of users', however to understand patterns of searching to optimize SUPrimo to handle common requests, the focus shifts to the tail part of the curve (Figure 4.2).

This is because although this group (51-1305) consists of very few (91 queries (1.34%), they have the highest occurrences i.e. a high amount of users search for them thereby making these queries the part of user behaviour on SUPrimo that presents data that can be used for optimizing SUPrimo to handle common request to better support users' search following from (Ruthven and Kelly, 2011).

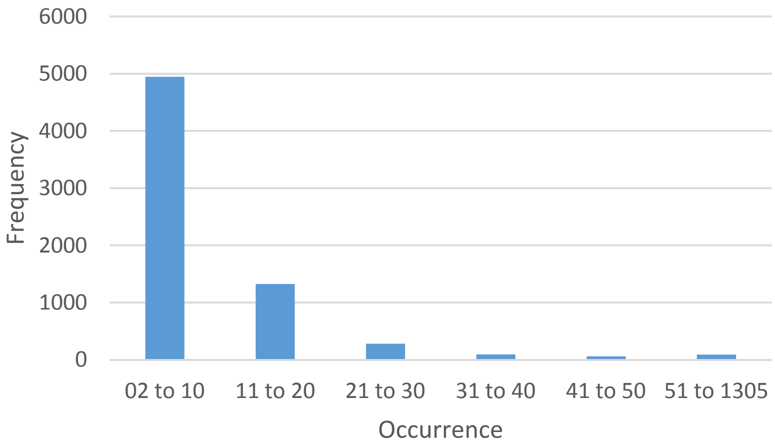


Figure 4.2 Graphical representation of repeated queries distribution

Therefore a sample of the top 20 queries (table 4.4) which constituted over 40% of the queries from the group were closely examined. Queries with 3 terms constituted 55% (11 queries) while queries with 2 terms constituted 45% (9 queries).

Top Repeated Queries	Frequency	Percentage	Category
the fifth discipline	1305	37.82%	T
Richard S Kay	217	6.29%	A
environmental policy	181	5.24%	T
Alan D Lopez	128	3.71%	A
social psychology	126	3.65%	T
Shakespeare, William, 1564-1616	114	3.30%	A
Critical path analysis	109	3.16%	T
J. C Murray	107	3.10%	A
Robert Aitken	107	3.10%	A
J. J Fawcett	106	3.07%	A
Sandy Brownjohn	103	2.98%	A
J Richard Ashcroft	102	2.96%	A
julia donaldson	98	2.84%	A
Penny Gay	97	2.81%	A
Jack D. Douglas	96	2.78%	A
Housing England London	94	2.72%	T
Arthur Charles	92	2.67%	A
Finland. Tilastokeskus	90	2.61%	T
The Andersonian Library	90	2.61%	T
kevin waldron	89	2.58%	A
Total	**3451**	**100.00%**	

Table 4.4 Top 20 repeated queries

Note: T = Topic Search and A = Author Search

Interestingly, despite the high percentage (70%) of “topic search” on SUPrimo as discovered in the previous section, the top 20 very high occurrence queries were dominated by “author searches” with 13 searches which represented 65% while topic searches records 7 which represented 35%. This implies that about 3 in 5 of these searchers are seeking for information either by authors or about individuals on SUPrimo. However as the number of queries increased, the title search increased while author search decreased so it cannot be generalised for users’.

Either way this findings presents valuable information for SUPrimo optimisation and a foundation for further research because it will be more interesting to further analyse these high occurrence queries to possibly find the fields for which the names and topics belongs and invariably the departments and can then be a measure for knowing the university departmental usage of the library, however this is not within the scope of this study.

4.4 Queries – Correlation Analysis

The characteristics and content of SUPrimo queries have been established however their effects on results are not known. As a result the variables "query length and average results" from SUPrimo log were correlated to find potential relationship between query and results which formed the foundation for subsequent directed observations on samples.

Scatter plots are generally used to test possible relationships between variables while correlation analysis measures the strength of the relationship between the variables. These knowledge can be used to present potential patterns, groups or clusters for closer examination in a data set (Silverstein et al., 1999) and as a results they were applied on the dataset.

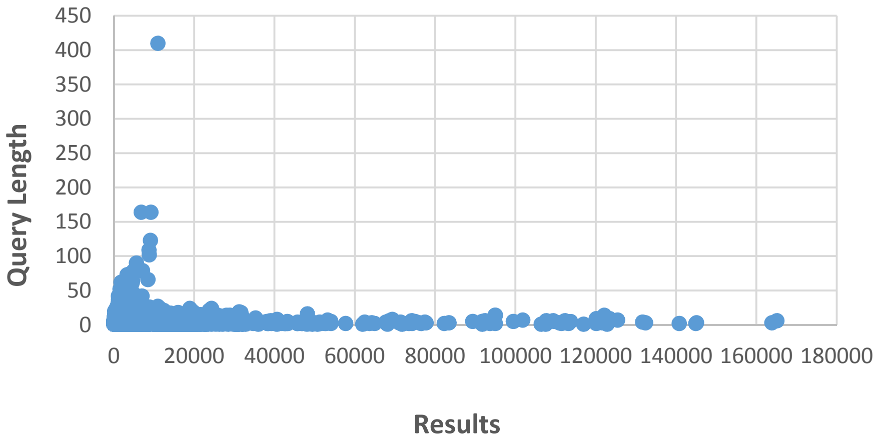


Figure 4.3 Query length-result scatter plot

It can be perceived from the scatter plot (figure 4.3) that there is a possibility of a relationship between the two variables because although the majority of the points are concentrated at the bottom part, some few points appears potentially scattered about an underlying straight line (at the top left hand and bottom right); implying the likelihood of a relationship.

Pearson's correlation coefficient was applied to measure the strength of the possible relationship. Result showed a negligible negative correlation between query length and results, with correlation coefficient of r=-.054, p<0.001 for the sample containing query length between 1-410 and results between 0-22361545.

This correlation result implies that there is a possibility that as the query length increases the number of results reduces indicating that the query length is inversely proportional to the results i.e. shorter queries returns more results than longer queries and are likely to fail, otherwise explained as shorter queries are more successful than longer queries.

However the correlation is insignificant because of the population that falls into this category is very small compared to the sampled total population and as had been earlier highlighted with points on the scatter plot (figure 4.1). On the other hand this small percentage can also be inferred as outliers in the total population sample.

The new findings presented the need to analyse failed queries for their characteristics for a conclusive inference that shorter queries are more successful than longer queries on SUPrimo even though there is knowledge that correlation does not mean causation.

4.5 Queries – Failure

There are numerous definitions of query failure but for this study the applicable definition was its definition as a search that matches no data or documents in a web based search engine collection ("zero results")(Jones et al., 2000). This can be for varied reasons and as with all web based search engines includes ''incorrect use' of the system such as - mistakes in spellings, wrong scope (using wrong search options), queries vocabulary that were not supported by the search engine (Jansen et al., 2000) and occasionally impatience of the users (Lau and Goh, 2006).

It had been reported in the previous chapter that 17% (4621) of the total queries in the period examined in the context of this study failed out of which 3076 (66%) were repeated queries while 1541 (33%) were unique (table 4.5). This implies that about 2 out of 10 queries on SUPrimo fail which is quite significant.

Failed Queries	Frequency	Percentage
Unique queries	1545	33.43%
Repeated queries	3076	66.57%
Total	**4621**	**100.00%**

Table 4.5 Summary of failed queries

SUPrimo queries were closely examined for possible reasons for query failure answering questions such as; does the length of a query affects the results as insinuated by results of statistical correlation analysis? Could search option be a cause of query failure and what can the time reveal about query failure?

4.6 Queries – Failure and Query length

The number of terms in a failed query on SUPrimo ranged between 1 and 410 terms. Queries with 1-10 terms constituted 72% of the total failed queries (table 4.6). Surprisingly the result implies that query failure decreases with increasing number of terms in the query on SUPrimo. Although the result is consistent with Jones et al. (2000) findings that the fewer the terms in a query, the more likely that the query will fail but it is contrary to more recent studies such as Lau and Goh (2006) findings that shorter query lengths decrease the likelihood of search failure. It is also contrary to the pattern revealed by earlier correlation analysis and as a result, the sample was critically examined to clarify conflicting results.

Query Length	Frequency	Percentage
1	863	18.68%
2	574	12.42%
3	497	10.76%
4	381	8.24%
5	309	6.69%
6	199	4.31%
7	175	3.79%
8	121	2.62%
9	111	2.40%
10	134	2.90%
Total	**4621**	**72.80%**

Table 4.6 Failed queries frequency distribution

It had been reported from earlier findings that 70% of SUPrimo users employ short queries on the average of 2.97 terms and shown by the Zipf distribution curve (figure 3.12) which means there are few long queries. If there are few long queries, it therefore implies that measuring query failure with the frequency of query length in a failed queries may not be a fair assessment of the query failure on the retrieval system consequently a purposively selected failed queries with 1-10 terms were further analysed by total number of failed queries per query length for explicit understanding of user behaviour.

Query Length	Failed Queries	Total Queries	Percentage Failure
1	863	2926	29.49%
2	574	5207	11.02%
3	497	4336	11.46%
4	381	2915	13.07%
5	309	1829	16.89%
6	199	1319	15.09%
7	175	878	19.93%
8	121	563	21.49%
9	111	505	21.98%
10	134	694	19.31%

Table 4.7 Failed queries per length distribution

Interestingly result showed that the search failure per total queries in a query length was highest in queries with 1 term and lowest in queries with 2 and 3 terms. While there appears to be no defined pattern as the percentages varied for the other terms, percentages failure per length were significantly higher than 2 and 3 terms queries and significantly lower than 1 term query (table 4.7).

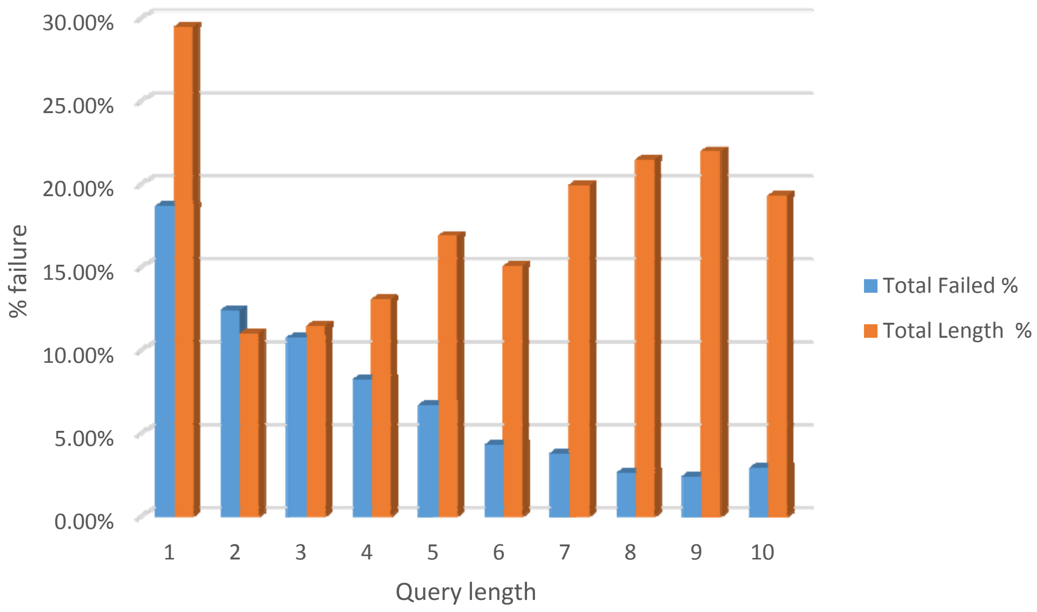


Figure 4.4 Comparison of query failure between failed queries and length queries

From figure 4.4, results can be interpreted as longer queries fail more on SUPrimo evident from the percentage of failed queries per each length with respect to the total number of queries per length. It is understandable why the percentage of failed long queries per term is low because they are relatively smaller than short queries and therefore not a fair basis for assessing query failure however when measured by the total number of failed queries per each lengths (which is the fair assessment), the failure increases much higher.

Furthermore the absence of a specific failure pattern further explains the disparateness in the point's distribution earlier highlighted on the scatter plot and the generally higher percentage of failed longer queries 100-410 terms explains the top left corner points while the low percentage of 2-3 terms emphasizes the concentration at the bottom left corner of the chart. These findings validates earlier correlation analysis findings that shorter queries return more results than longer queries on SUPrimo and therefore it can be concluded that although the correlation appears negligible, it is significant.

4.7 Queries – Failure and Response time

All failed queries with 2-3 terms; a total of 1071 queries were purposively selected 497 (46.4%) had 3 terms and 574 (53.6%) had 2 terms (table 4.6) for closer examination because these queries were more successful than other query lengths and the larger percentage of users search with this number of terms as a result they represent good data source for to understand user behaviour for accurate system evaluation. Results showed that showed that the frequency of occurrence of failed queries ranged between 1 and 22 where 916 were repeated queries and 155 were unique queries.

The 155 unique queries were separated for further examination and it was found that about 70% of this queries had a response time of over 500 seconds and some of these unsuccessful searches were in fact useful ones with no spelling mistakes amongst others (Appendix C shows the response time of the 155 failed unique queries). Intuitively it suggests a system problem which could be in vast range that includes problem with assessing the database, system overload amongst many other reasons that prevented the system from retrieving results.

On the other hand, this could also be as a result of lack of perseverance on the part of the user as explained by Lau and Goh (2006) following Drabenstott and Weller (1996) findings from the Mardigian and Lilly Libraries OPAC. In this case, the user impatiently terminates the search because as shown the response time of this queries were long and since there was zero results there is a likelihood that the user did not wait for the search to be completed and the system had already logged the search (Lau and Goh, 2006, Drabenstott and Weller, 1996).

Interestingly however on closer examination of “Project Management” an example of such useful query which had no spelling error that returned zero results and the average response time was 21352 (Appendix C). There appears to be nothing wrong with the query but at that point in time it returned zero results. This type of failure can be potentially explained by the combination of both “user perseverance” and “system problem” in which the user was impatient and did not wait for the result because the system was taking an unusual/unexpected long time for unknown reasons, which makes the user terminates the search while the system logged the expected time the system would have responded with or without results.

However, these potential reasons cannot be explicitly explained by this study primarily because the logs used are summary records and date recorded in the logs are the dates query was written to the summary log and have been earlier mentioned as a shortcoming of transaction logs.

4.8 Queries – Failure and Search Options

Another possible reason for a failed search could be the use of wrong search option. As had been described in the previous chapter, SUPrimo provides access to library materials through four (4) options and the default is library collections. This search options is recorded under the "scope type" variable in SUPrimo log. The scope types are local, Remote and DS (Deep Search) (table 3.3).

In order to ascertain if the wrong use of search options is a possible reason for query failure, there is the need to know the scope type for all queries in the context of the period examined by this study. The result shows that 54% of users search "local" collections, 40% search DS collections while only 5% search remote collections (table 4.8).

Scope Type	Frequency	Percentage
Local	14684	54.64%
Remote	1466	5.46%
DS	10724	39.90%
Total	**26874**	**100.00%**

Table 4.8 Summary of total queries by search options

From the results, the DS results suggests that about 4 out of 10 SUPrimo users have knowledge of where to find information being sought because it is not the default search option. Although not so substantial but still implies some form of training on how to use the retrieval system had been provided meanwhile it could also be that the system is easy to use.

4.8.1 Local Search Options

Narrowing the analysis to the 4621 failed queries; 79% of the failed queries were local search, while only 15% of remote searches failed and 5% of DS search failed (table 4.9). The local collection search option's higher percentage of 79.25% is understandably so because it is the default search option "library-collection" is included in the "local" scope type category and suggests that when the results returned are not relevant to the users they use the "article+database" option which is included in the DS scope type, a possible explanation for its being the 2nd most used search option.

While the remote collections percentages suggest that it is not used because it is not among the 4 search options presented on the SUPrimo home page and requires more specific and technical directions to access thereby making it difficult to use.

Scope Type	Failed Queries	Percentage
Local	3662	79.25%
Remote	713	15.43%
DS	246	5.32%
Total	**4621**	**100.00%**

Table 4.9 Summary of total failed queries by search options

The very high query failure on local search and low query failure on DS search complements the earlier findings which implied that users use default setting more and suggests the possibility of queries being refined when first search hit zero results and are then another search option is selected. Naturally assuming that the DS was selected the query failure on DS is as a result drastically reduced. This assumptions can be validated when the analysis involves session level analysis which is not within the scope of this study because of the limitations of the logs.

4.8.2 Remote Search Options

However what remains unclear is why the remote collections searches records high amount of failed queries in comparison to the other scope type benchmarking with the total remote searches (table 4.10)? It had been suggested that these remote collections maybe difficult to access, so it therefore implies that a user that searches this collections apparently know the location of what is being sought and probably how to use the system but there is still high query failure.

Scope Type	Total Queries	Failed Queries	Percentage
Local	14684	3662	**24.94%**
Remote	1466	713	**48.63%**
DS	10724	246	**2.29%**
Total	**26874**	**4621**	

Table 4.10 Summary of failed queries per total queries in a scope

Previous analysis shows that the total remote search type carried out in the period examined by this study was 1466 and 713 were failed queries (table 4.10) which represented about 50% of all remote searches implying that 1 in 2 remote searches fails which is significantly high. As a result of this observation, the 155 failed unique queries from section 4.7 were closely examined.

Not so surprising, the result showed that 47% of this users search remote collections, 39% search local collections and only 12% search the primo central collection (table 4.11). The higher percentages of failed queries were remote searches implying that about 1 out of 2 remote searches failed.

Scope Type	Frequency	Percentage
Local	61	39.35%
Remote	74	47.74%
DS	20	12.90%
Total	**155**	**100.00%**

Table 4.11 Summary of unique failed queries by search options

There are two possible explanations from the observed behaviour: following from the assumption that searchers of remote collections have identified the collection as the location of information being sought, it can be implied that information in that collection is not well structured for retrieval. While on the other hand there is also the possibility the users knows the location of the information being sought but are searching wrongly.

The 74 failed queries of remote collections were closely examined. It was found that 91% of these queries were “topic search”, and 9% were “author search”. 55% consisted of 2 terms and 45% consisted of 3 terms. The response time ranged between 1008secs to 30244secs (table 4.12). Spelling errors were identified on only 5% of the queries. These spelling errors were identified by language knowledge and not based on domain or expert knowledge.

As have been earlier suggested, these results confirms a likely problem with the remote collections which might include information structure in the collection especially considering the response time and when the query success is compared to of local and DS scope type searches (table 4.10). On the hand, the percentage of spelling errors identified indicates the possibility of user problem such as how to use seemed unlikely as only 4 out of 74 had spelling mistake. However expert knowledge will be needed to verify the queries and identify if the searches were carried out on the appropriate remote collection.

Query	Search Class	Response time (sec)	Scope Name
fair trade nation	S	30244	103
ansoff matrix	S	29333	103
Lionel Mackenzie	A	24440	109
past papers V1103	S	22378	109
leadership challenge	S	21628	103
Project Management	S	21352	103
leadership challenge	S	18749	103
Embedded Intelligence	S	18348	114
Embedded Intelligence	S	15258	114
thermochim / miyagawa	S	14989	114
Embedded Intelligence	S	14268	114
fair trade nation	S	14006	103
Wind Feed Forward	S	13932	114
fair trade nation	S	13635	103
kotler marketing concept	S	13345	103
Embedded Intelligence	S	12663	114
Embedded Intelligence	S	11802	114
stockbroker / Edinburgh	S	11719	109
fair trade nation	S	11714	103
Marques / FaFFIF	S	10659	104
EIA desiign	S	9694	114
Kwanta Panthongprasert 2010	A	9106	103
Embedded Intelligence	S	9012	114
epoxy resin/hardner effects	S	8748	114
cymbopogon proximus	S	8479	116
social media	S	8111	116
tadanafil solubility	S	8025	104
sodium lauryl sulfate	S	7986	116
Embedded Intelligence	S	7873	114
disab / experience	S	7822	111
JaffÃƒÂ©, Nebenzahl 2001	A	7618	103
Embedded Intelligence	S	7401	114
board diversity	S	7182	116
business monitor international	S	7152	116
eglington lane glasgow	S	6990	116
chemical reviews	S	6962	116
constration import	S	6622	103
rectification scotand	S	6620	108
waterfront regneration	S	6575	103
soh / fauzee	S	6546	104

Edmond Bequerel	S	6521	105
Marques / FaFFIS	S	6505	104
arana / FaSSIF	S	6476	104
phoenyx dactylifera	S	6467	116
corporate governance	S	6392	116
location strategy	S	6113	103
geese imprinting	S	5992	109
tadanafil AND solubility	S	5769	104
consuming frutose-sweetened	S	5681	104
dairy / GCC	S	5663	116
influenza champion	S	5435	104
pvrc bulletin	S	5398	116
risk oversight	S	5398	116
newspaper articles	S	5380	116
"liu tianye"	A	5319	STR02890,
Environmental Impact Assessment	S	5317	STR01492,STR00637,
/ Corporate governance	S	5258	111
KAOLINITE SLURRIES	S	4689	STR01259,
project cost management	S	4681	103
soil sample	S	4658	STR01259,
dairy / GCC	S	4499	116
social media	S	4357	116
uws ayr library	S	4343	STR02695,
sconul award 2013	S	4299	STR02695,
marketing eresources	S	4271	STR02695,
marks and spencer	S	3577	103
"environmental impact assessment"	S	3298	STR01492,
Environmental Impact Assessment	S	3235	STR01492,
kissan joseph	A	2890	STR02977,
Market prenatation	S	2507	103
tribal communities	S	2467	103
kissan josepÃ¼	A	2231	STR02977,
"Environmental Impact Assessment"	S	2183	STR01492,
creative brief communication	S	1008	103

Table 4.12 Remote collections 74 failed queries

4.8.3 DS Search Options

Results have shown that DS collection searches appears more successful than the search on the other collections considering only 2% of the total queries searched on the DS collections failed (table 4.10). It had been implied that the reason for high search success on DS collections maybe because queries have been refined after the zero results from default setting of library-collections. More so there is also the possibility of the better information structure than the remote searches considering the failed query response time which was about 7 times reduced than remote queries.

Query	Search Class	Response time (sec)
diesel engine emittions	T	4657
John Black asswssment	T	4461
dylan william assesssment	T	4344
School readiness	T	3583
halpern and goldfarbv2013	A	3470
School readiness	T	3436
School readiness	T	3339
School readiness	T	3090
School readiness	T	2612
School readiness	T	2583
r2 ageloc	T	2573
brand / Fourier	T	1915
boundary spanners slaes	T	1855
low speedshipping	T	1416
DKSH operationa	T	1291
gases separation	T	1288
Zhang, Duan	A	1133
TMPMgCl.LiCl turbo-Hauser base	T	422
Joseph Ambrose Banks	T	412
cervone / pervin	T	300

Table 4.13 DS search option 20 failed queries

Nonetheless, in order to completely find the pattern of user searches to further improve system responses the 20 unique DS failed queries were examined (table 4.13). Results on closer examination showed that 90% of the queries were topic-search and 2% author-search. 55% of consisted of 2 terms and 45% consisted of 3 terms. The response time ranged between 300secs to 4657secs. 14 distinctive queries emerged out of which 21% had spelling errors majorly typographical errors identified by familiarity with language. This percentage of spelling errors shows that user problems may not be how to use but patience in typing the query correctly.

Furthermore benchmarking the query success on the collection (table 4.10), there is a likelihood that the issue of perseverance on the part of the user earlier discussed (section 4.7, Lau and Goh, 2006); where at the time of the search; system issues which might include synchronising the collection, systems overload amongst others makes the user terminates the search while it had been logged with the proposed response time. On the contrary unlike the remote collection where queries could be verified and collection vetted, verifying DS collection queries is not feasible considering time and cost however considering the query success rate this failure is not a problem.

4.9 Summary

This chapter dived into the content of user's queries in order to find the patterns of searching and its effects on results for definitive inferences following the preliminary analysis implications from the previous chapter on user behaviour and evaluate the SUPrimo search engine responses.

In order to find patterns of searching; the contents of a purposively selected queries with 2-3 terms were closely examined and it established earlier implications that users may use key term search as 78% of SUPrimo users fall under the "topic search" query classification which includes key term search. However interestingly it was observed that the top 20 searches on SUPrimo had a higher percentage (65%) of "author search" query classification.

In order to evaluate the system, the complete dataset was statistical analysed for the relationship between queries and results. Next, a purposively selected failed queries with 2-3 terms were critically examined and it was revealed firstly that query failure increased with increasing query length; Secondly that the response time of failed queries predominantly ranged from 500 seconds; And finally it revealed that the "DS" search option records the lowest query failure rate while "remote" search option which is the least used search option records the highest failure rate.

It had been reported in the literature that understanding "information behaviour" presents an evaluation of information retrieval systems on how well they support people information seeking (Ruthven and Kelly, 2011), consequently these findings thereby suggests possible improvements for SUPrimo to better support user's information seeking behaviour.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Research Overview

The ultimate goal of any information retrieval system is to support users in fulfilling their information needs and an academic library retrieval system is no exception. Information Scientists agree that understanding user behaviour will reveal the adequacy of these systems in supporting user needs. This was evident from the numerous research reported in the literature that had been done to understand user behaviour in varying categories on different types of information retrieval systems including traditional library OPAC's however, no research exists for holistic user behaviour on current academic libraries information retrieval system. Consequently the need therefore arises to understand holistic user behaviour on a University library's search engine in a domain specific context.

Transaction log analysis is a method useful for understanding the system performance and the user interactions during a search process and in addition offers in-depth pictures of user-system interactions (i.e. user behaviour otherwise known as information behaviour) over a specific period of time. As a result it was adopted for this research to understand user's information seeking behaviour on the University of Strathclyde Library information retrieval system/search engine known as SUPrimo (Strathclyde University Primo) used by students and staff of the University of Strathclyde for locating information to support learning.

Furthermore, although SUPrimo transaction logs cannot measure users' satisfaction amongst other limitations; analysing an Information retrieval system's transaction logs uncovers search patterns that yield ways to further improve the system. Such improvements may not apply only to the system being analyzed but may lead to enhancements in the design of other systems. This chapter thereby presents a summary of the dissertation research findings and ensuing recommendations.

5.2 Conclusions

The aim of this research was to provide insight into how users' find information on a University Library's search engine in the context of a domain specific retrieval system by answering targeted questions that will reveal areas of improvement for the search engine. The following sections thereby provides conclusions drawn from the results and interpretations presented during this research within the context of this research as a domain specific academic information retrieval system and with respect to the objectives of this study.

5.3 What are the usage patterns of SUPrimo?

5.3.1 Usage Analysis

Users of the University of Strathclyde run more searches during exam period. The use of SUPrimo was increased in the month of May by 63% over the preceding month with 46% of this increase occurring during the first half of the month.

5.4 What are the characteristics of queries issued on SUPrimo?

5.4.1 Query Length

SUPrimo users employ short queries (2.97terms). This conclusion was drawn from results of analysis of the total queries selected for this study that showed 70% of the queries contained between 1 and 6 terms and statistical analysis that revealed the mean of the query lengths was not a true representation of a typical query on SUPrimo.

5.4.2 Query Distribution

SUPrimo queries follows a Zipf distribution – a power law distribution. This was evident in the frequency of shorter queries being higher than the frequency of longer queries.

5.4.3 Query Term

SUPrimo queries consists of unique terms. This conclusion was drawn from the high occurrence of unique terms of 73% of the total term occurrence from the results of analysis.

5.5 What are the patterns of user's search on SUPrimo?

5.5.1 Query Classification

SUPrimo users employ key term search. This conclusion was drawn from a combination of analysis results that 73% of the total terms on SUPrimo were unique and 78% of queries were classified as topic search.

5.5.2 Query Repetition

Query repetition rate on SUPrimo is high. The conclusion was reached as 78% of the total queries examined were repeated queries. This is an expectation in the context of this research as an academic library information retrieval system because there is at least more than 1 person in each class of study.

Interestingly however, the author search classification constituted 65% of the top 20 searches.

5.5.3 Query Correlation

There is a relationship between query length and results. This conclusion was drawn from a combination of analysis results; from statistical correlation analysis and manual comparison of percentage failure between total failed queries and total length queries.

5.5.4 Query Failure

- Query failure rate on Suprimo is quite significant. This was evident from the 17% failed queries of the total queries examined.
- Spelling errors are not a significant cause of query failure on SUPrimo. Only 5% of the queries examined had spelling errors.

5.5.5 Query Failure and Query Length

- Queries with 2 and 3 terms are the most successful on SUPrimo. This was evident from their low percentages of 11.02% and 11.46% respectively with respect to the total queries in each length.
- Queries with 1 term are not quite successful on SUPrimo. This was evident from the significant 29% failed queries of the total 1-term queries.
- Longer query length increases query failure. This was evident from the increasing percentages of failed query per query length from 4-term query length.

5.5.6 Query Failure and Response Time

Queries failed on SUPrimo as a result of system problem. This was evident from results of analysis that showed 70% of failed queries had a response time of over 500 seconds and some of these unsuccessful searches were in fact useful ones.

5.5.7 Query Failure and Search Options

- SUPrimo default search option (local search) is the most frequently used. This was concluded as results from analysis showed that 54% of the total searches carried out during the period examined by this research were local search.
- Searching on SUPrimo DS search option is the most successful. This was evident from results of analysis that showed that only 2% of the total DS searches failed.

- SUPrimo remote collections are not well structured. This conclusion was drawn from results of analysis that showed high percentage of 48% query failure from the total remote searches in conjunction with a reasoning that users of the remote collection search option presumably know the location of the information being sought and were unlikely to search wrongly.

These findings suggest information retrieval systems have deeply improved with advancement in technology from the earlier years to its current information structure while users' search patterns have not changed very much across the years despite many years of experience in the use, research and development of information retrieval systems. These improvements are as a result of continuous research in the Information Science field which usually results in recommendations to further enhance systems.

5.6 Recommendations

The results obtained in this study give rise to some recommendations that could be considered for enhancing user experience and improving SUPrimo search engine which are applicable to other information retrieval systems. The recommendations are presented under four headings: SUPrimo functional recommendations, SUPrimo technical recommendations, SUPrimo operational recommendations and further research.

5.6.1 SUPrimo Functional Recommendations

These are recommendations that could be considered for improving the design of SUPrimo to better support users' search for information:

- The high query repetition on SUPrimo suggests that the search engine could be optimised to handle common requests. This could be done by including the very high frequency queries in the query suggestion which is a capability of SUPrimo. This optimisation will also address the issue of failure with regards to query length. Using only the very high frequency queries from a set benchmark will eliminate the potential of large list of queries.
- System needs to be improved to handle queries with 1 term successfully considering that queries with 1 term are not quite successful and since the current systems are the sophisticated web based information retrieval systems.
- Structuring collections for the remote search options as this might be the reason for the high failure rate since lack of knowledge of its use seemed unlikely. Furthermore it could be considered to make the remote collections accessible from the home page.

5.6.2 SUPrimo Technical Recommendations

- Developers can includes the time of transaction in the summary logs and not just the time transaction was written to logs as this will enable session analysis and other time related analysis that will provide more insight into users information seeking behaviour
- Increasing the system capacity to a handle volumes of transaction at the same time could also be considered as this might be a likely cause of query failure with respect to response time.

5.6.3 SUPrimo Operational Recommendations

These recommendations includes analyses that could be considered for inclusion to the operations processes of SUPrimo at uniform intervals such as:

- **Query repetition analysis**: Knowing repeated queries will be source of information for the optimisation to handle common requests.
- **Search options analysis:** Knowing how the search options is used could serve as mean of checking charges in situation where charges are made for each access to external sources.
- **Query failure analysis:** Knowledge of query failure rate could be used to measure system performance.

Although not an analysis, it could be considered that system maintenance are carried out for possible detection of the system problem associated with the very high response time for failed queries.

5.6.4 Further Research

The results obtained also form a foundation for further research in areas work that could be explored to the academic world. The findings can be further researched as follows:

It would be interesting to find the usage of the library with respect to time i.e. what time the library is used as this will guide the provision of out of hour's access as this was a limitation of this study as the summary logs did not record transaction time.

Commonly searched terms can be manually analysed for categorisation of users by subject which might potentially discover the usage of the library by university departments.

The longer queries can be manually analysed to further understand the information being sought by these type of queries, although there is the possibility of lack of understanding on the part of the user where it is assumed that SUPrimo works like general multipurpose search engines such as google.

The limitations of the transaction log could be addressed in subsequent researches such as the limitation of not measuring user satisfaction could be addressed by evaluating the information retrieval systems focusing on measuring satisfaction of users, ease of learning, ease of use and not on user's behaviour.

5.7 Summary

Huge sums of money are being invested by academic libraries to provide electronic access to users as result of advancements in technology and its impact on the world. Investigating the information seeking behaviour of users' is therefore essential to ensure that this web based information retrieval systems provide access to the resources effectively and efficiently.

This study have presented characterization of usage patterns, users' queries and search patterns in the use of the University of Strathclyde library information retrieval system (SUPrimo) over a two months period and suggested improvement in user interface, content organisation and system operations for SUPrimo thereby achieving the aims and purpose of the study.

In addition, the findings from this study have played a useful role in informing the management of the University of Strathclyde library information retrieval systems about the use of the library search engine; revealing lots of interesting things to consider in terms of the systems and users. Furthermore, the findings of the study also presents a foundation for further research into information seeking behaviour on academic domain information retrieval system.

BIBLIOGRAPHY

AGOSTI, M., CRIVELLARI, F. & DI NUNZIO, G. 2012. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery, 2012, Vol.24(3), pp.663-696,* 24.

ASUNKA, S., CHAE, H. S., HUGHES, B. & NATRIELLO, G. 2009. Understanding Academic Information Seeking Habits through Analysis of Web Server Log Files: The Case of the Teachers College Library Website. *The Journal of Academic Librarianship, 2009, Vol.35(1), pp.33-45,* 35.

BAJRACHARYA, S. & LOPES, C. 2012. Analyzing and mining a code search engine usage log. *Empirical Software Engineering, 2012, Vol.17(4), pp.424-466,* 17.

BATES, J. M. 2012. *Understanding information retrieval systems [internet resource] management, types, and standards,* Boca Raton, Fla., Boca Raton, Fla. : Auerbach/CRC Press.

BENDERSKY, M. & CROFT, W. B. 2009. Analysis of long queries in a large scale search log. *Proceedings of the 2009 workshop on Web Search Click Data.* Barcelona, Spain: ACM.

BREEDING, M. 2005. Analyzing Web Server Logs to Improve a Site's Usage. The Systems Librarian. *Computers in Libraries, 2005, Vol.v25n9p26(28-29Oct2005), Vol.25(9), p.26,* 25.

CARMAN, M. J., BAILLIE, M., GWADERA, R. & CRESTANI, F. 2009. A statistical comparison of tag and query logs.

CARSON, J. 2004. *Professional Practice and the labour process: Academic Librarianship at the Millennium,* Bingley, U.K., Bingley, U.K. : Emerald.

CASE, D. O. 2012. *Looking for Information : a survey of research on information seeking, needs and behavior,* Bingley, UK, Bingley, UK : Emerald Group Pub. ;.

CHEN, H. M. & COOPER, M. D. 2001. Using clustering techniques to detect usage patterns in a Web-based information system. *Journal Of The American Society For Information Science And Technology, 2001, Vol.52(11), pp.888-904,* 52.

CHEN, H. M. & COOPER, M. D. 2002. Stochastic modeling of usage patterns in a web-based information system. *Journal Of The American Society For Information Science And Technology, 2002, Vol.53(7), pp.536-548,* 53.

CHRISTEL, M. G. 2007. Establishing the utility of non-text search for news video retrieval with real world users.

CROWELL, J., ZENG, Q., NGO, L. & LACROIX, E. M. 2004. A Frequency-based Technique to Improve the Spelling Suggestion Rank in Medical Queries. *Journal of the American Medical Informatics Association, May 2004, Vol.11(3), pp.179-185,* 11.

CUBITT, S. 2006. Library. *Theory, Culture and Society, March 2006, Vol.23(2-3), pp.581-590,* 23.

DRABENSTOTT, K. M. & WELLER, M. S. 1996. Failure analysis of subject searches in a test of a new design for subject access to online catalogs. *Journal of the American Society for Information Science, 1996, Vol.47(7), pp.519-537,* 47.

EX-LIBRIS 2011. *Voyager 8.1 Technical User's Guide,* United States.

EX-LIBRIS, P. 2014. *Empowering Libraries to Address User Needs* [Online]. Available: http://www.exlibrisgroup.com/category/PrimoOverview. [Accessed 10 July 2014].

GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S. & BRILLIANT, L. 2009. Detecting influenza epidemics using search engine query data. *Nature, 2009, Vol.457(7232), pp.1012-4,* 457.

GLASER, B. G. 1999. *The discovery of grounded theory : strategies for qualitative research,* New Brunswick, N.J., New Brunswick, N.J. : Aldine Transaction.

HANCOCK-BEAULIEU, M., ROBERTSON, S. & NEILSON, C. 1990. Evaluation of Online Catalogues: an assessment of methods. London: The British Library Research and Development Department.

HERSKOVIC, J. R., TANAKA, L. Y., HERSH, W. & BERNSTAM, E. V. 2007. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *Journal of the American Medical Informatics Association, March 2007, Vol.14(2), pp.212-220,* 14.

HEWSON, C., YULE, P., LAURENT, D. & VOGEL, C. 2002. *Internet research methods : a practical guide for the social and behavioural sciences,* London, London : SAGE.

HILBERT, D. M. & REDMILES, D. F. 2001. Large-Scale Collection of Usage Data to Inform Design. In M. Hirose (Ed.), INTERACT' 01, IFIP TC.13 International Conference on Human-Computer Interaction: IOS Press.

HUURNINK, B., HOLLINK, L., VAN HEUVEL, W. D. & DE RIJKE, M. 2010. Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis. *Journal of the American Society for Information Science and Technology, June 2010, Vol.61(6), pp.1180-1197,* 61.

JANSEN, B. J. 2006. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research,* 28**,** 407-432.

JANSEN, B. J. & POOCH, U. 2001. A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology, 2001, Vol.52(3), pp.235-246,* 52.

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, 2006, Vol.42(1), pp.248-263,* 42.

JANSEN, B. J., SPINK, A. & PEDERSEN, J. 2004. The Effect of specialized multimedia collections on web searching. Journal of Web Engineering: Rinton Press.

JANSEN, B. J., SPINK, A. & SARACEVIC, T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management,* 36**,** 207-227.

JANSEN, B. J., SPINK, A. & TAKSAI, I. 2009. *Handbook of research on web log analysis,* Hershey, PA, Hershey, PA : Information Science Reference.

JIN YOUNG, K., FEILD, H. & CARTRIGHT, M. 2012. Understanding book search behavior on the web. *Proceedings of the 21st ACM international conference on Information and knowledge management.* Maui, Hawaii, USA: ACM.

JOINT, N. 2008. *Primo Presentation-SUPrimo at the University of Strathclyde,* Madrid, University of Strathclyde.

JONES, S., CUNNINGHAM, S. J., MCNAB, R. & BODDIE, S. 2000. A transaction log analysis of a digital library. *International Journal on Digital Libraries, 2000, Vol.3(2), pp.152-169,* 3.

KASKE, N. 1993. Research methodologies and transaction log analysis: Issues, questions, and a proposed model. *Library Hi Tech, 1993, Vol.11(2), p.79-86,* 11.

KATO, M. P., SAKAI, T. & TANAKA, K. 2012. When do people use query suggestion? A query suggestion log analysis. *Information Retrieval, 2012, pp.1-22*.

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

LAU, E. P. & GOH, D. H. L. 2006. In search of query patterns: A case study of a university OPAC. *Information Processing & Management, 2006, Vol.42(5), pp.1316-1329,* 42.

MARCHIONINI, G. 1997. *Information seeking in electronic environments,* Cambridge ; New York, Cambridge ; New York : Cambridge University Press.

MARKEY, K. 2007. Twenty-five years of end-user searching, part 2: Future research directions. *Journal Of The American Society For Information Science And Technology, 2007, Vol.58(8), pp.1123-1130,* 58.

MOULAISON, H. L. 2008. OPAC queries at a medium-sized academic library: A transaction log analysis. *Library Resources and Technical Services, October 2008, Vol.52(4), pp.230-237,* 52.

NATARAJAN, K., STEIN, D., JAIN, S. & ELHADAD, N. 2010. An analysis of clinical queries in an electronic health record search utility. *International Journal Of Medical Informatics, 2010, Vol.79(7), pp.515-522,* 79.

NIU, X. & HEMMINGER, B. M. 2010. Beyond text querying and ranking list: How people are searching through faceted catalogs in two library environments.

PAGE, S. 2000. Community research: The lost art of unobtrusive methods. *Journal Of Applied Social Psychology, 2000, Vol.30(10), pp.2126-2136,* 30.

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

PETERS, T. A. 1993. The history and development of transaction log analysis. *Library Hi Tech, 1993, Vol.11(2), p.41-66,* 11.

PHIPPEN, A., SHEPPARD, L. & FURNELL, S. 2004. A practical evaluation of Web analytics. *Internet Research, 2004, Vol.14(4), pp.284-293,* 14.

RUTHVEN, I. 2012. Grieving online: the use of search engines in times of grief and bereavement. *In:* FUHR, N., KAMPS, J. & KRAAIJ, W. (eds.). ACM.

RUTHVEN, I. & KELLY, D. 2011. *Interactive information seeking, behaviour and retrieval,* London, London : Facet.

SILVERSTEIN, C., MARAIS, H., HENZINGER, M. & MORICZ, M. 1999. Analysis of a very large web search engine query log. *SIGIR Forum,* 33**,** 6-12.

SPINK, A., YANG, Y., JANSEN, J., NYKANEN, P., LORENCE, D. P., OZMUTLU, S. & OZMUTLU, H. C. 2004. A study of medical and health queries to web search engines. *Health information and libraries journal, 2004, Vol.21(1), pp.44-51,* 21.

STRATHCLYDE LIBRARY. 2014. *SUPrimo Library Search* [Online]. Available: http://suprimo.lib.strath.ac.uk/primo_library/libweb/action/search.do?dscnt=1&tab=metalib&dstmp=1408397742396&vid=SUVU01&mode=Basic&fromLogin=true&fromLogin=true [Accessed 10 July 2014].

TJONDRONEGORO, D., SPINK, A. & J.JANSEN, B. 2009. A study and comparison of multimedia Web searching: 1997-2006. *Journal of the American Society for Information Science and Technology, September 2009, Vol.60(9), pp.1756-1768,* 60.

UNIVERSITY OF STRATHCLYDE. 2014. *Facts & Figures* [Online]. Available: http://www.strath.ac.uk/press/factsfigures/ [Accessed 10 July 2014].

VILLÉN-RUEDA, L., SENSO, J. A. & DE MOYA-ANEGÓN, F. 2007. The Use of OPAC in a Large Academic Library: A Transactional Log Analysis Study of Subject Searching. *The Journal of Academic Librarianship, 2007, Vol.33(3), pp.327-337,* 33.

WANG, C. 1985. THE ONLINE CATALOGUE, SUBJECT ACCESS AND USER REACTIONS&colon; A REVIEW. *Library Review, 1985, Vol.34(3), p.143-152,* 34.

WOLFRAM, D. 2008. Search characteristics in different types of Web-based IR environments: Are they the same? *Information Processing and Management, May 2008, Vol.44(3), pp.1279-1292,* 44.

YANG, L., MEI, Q., ZHENG, K. & HANAUER, D. A. 2011. Query log analysis of an electronic health record search engine. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2011, Vol.2011, pp.915-924,* 2011.

ZHANG, J. & KAMPS, J. 2010. Search log analysis of user stereotypes, information seeking behavior, and contextual evaluation. *Proceedings of the third symposium on Information interaction in context.* New Brunswick, New Jersey, USA: ACM.

ZHIYONG, L., WON, K. & WILBUR, W. J. 2009. Evaluation of query expansion using MeSH in PubMed. *Information Retrieval, February 2009, Vol.12(1), pp.69-80,* 12.

APPENDIX A – Complete query lengths frequency distribution

Query length	Frequency	Percentage
1	2926	10.9%
2	5207	19.4%
3	4336	16.1%
4	2915	10.8%
5	1829	6.8%
6	1319	4.9%
7	877	3.3%
8	563	2.1%
9	505	1.9%
10	694	2.6%
11	671	2.5%
12	668	2.5%
13	564	2.1%
14	584	2.2%
15	481	1.8%
16	435	1.6%
17	356	1.3%
18	277	1.0%
19	248	.9%
20	200	.7%
21	187	.7%
22	133	.5%
23	128	.5%
24	106	.4%
25	93	.3%
26	61	.2%
27	71	.3%
28	53	.2%
29	46	.2%
30	44	.2%
31	46	.2%
32	41	.2%
33	34	.1%
34	17	.1%
35	15	.1%
36	14	.1%
37	16	.1%
38	18	.1%

39	8	.0%
40	9	.0%
41	10	.0%
42	2	.0%
43	1	.0%
44	2	.0%
45	11	.0%
46	5	.0%
48	1	.0%
49	1	.0%
50	4	.0%
51	1	.0%
52	2	.0%
56	1	.0%
57	5	.0%
58	1	.0%
59	2	.0%
61	2	.0%
62	9	.0%
63	1	.0%
64	1	.0%
66	1	.0%
73	4	.0%
78	1	.0%
79	3	.0%
80	1	.0%
90	1	.0%
102	1	.0%
109	1	.0%
123	1	.0%
164	2	.0%
410	1	.0%
Total	**26876**	**100.0**%

APPENDIX B – Frequency distribution of Top 100 unique terms

Term	Frequency
social	1069
britain	866
management	856
great	825
education	715
research	690
assessment	680
study	679
children	660
journal	610
international	596
learning	578
history	537
business	531
scotland	496
development	484
marketing	482
work	480
teaching	440
law	425
analysis	423
review	417
impact	387
new	373
autism	367
health	351
service	342
human	341
school	340
performance	335
theory	302
based	299
design	297
language	295
brand	289
child	288
practice	284
environmental	284
communication	283
systems	280
united	273
science	273
case	272
information	270
england	252
century	249
economic	242
effects	240
primary	237
model	235
approach	234
use	233
technology	233

Term	Frequency
psychology	233
process	229
culture	229
people	226
states	223
university	221
public	221
quality	221
john	214
european	212
studies	211
literature	211
care	208
role	208
policy	207
disorders	205
rights	205
physical	204
congresses	199
young	198
consumer	194
using	189
system	189
self	188
educational	187
risk	186
engineering	186
early	185

Term	Frequency
media	183
world	181
aspects	179
british	179
control	174
effect	173
industry	172
strategy	171
society	165
scottish	165
methods	164
product	163
war	163
value	162
project	162
community	161
evidence	160

APPENDIX C – Complete 155 failed queries and their response time

Query	Response Time
fair trade nation	30244
Hospitals Scotland Aberdeen	29849
ansoff matrix	29333
Lionel Mackenzie	24440
past papers V1103	22378
leadership challenge	21628
Project Management	21352
leadership challenge	18749
Embedded Intelligence	18348
Embedded Intelligence	15258
thermochim / miyagawa	14989
Embedded Intelligence	14268
fair trade nation	14006
Wind Feed Forward	13932
fair trade nation	13635
kotler marketing concept	13345
Embedded Intelligence	12663
Embedded Intelligence	11802
stockbroker / Edinburgh	11719
fair trade nation	11714
Marques / FaFFIF	10659
EIA desiign	9694
Kwanta Panthongprasert 2010	9106
Embedded Intelligence	9012
epoxy resin/hardner effects	8748
cymbopogon proximus	8479
social media	8111
tadanafil solubility	8025
sodium lauryl sulfate	7986
Embedded Intelligence	7873
disab / experience	7822
JaffÃƒÂ©, Nebenzahl 2001	7618
Embedded Intelligence	7401
board diversity	7182
accretion in scotland	7160
business monitor international	7152
eglington lane glasgow	6990
chemical reviews	6962
constration import	6622
rectification scotand	6620
waterfront regneration	6575
soh / fauzee	6546
Edmond Bequerel	6521
Marques / FaFFIS	6505
arana / FaSSIF	6476
kissan joseph	2890
Â¿eÃŸ Â¿f sÂ¿Â¿eÂ¿Â¿e	2694
School readiness	2612
School readiness	2583
r2 ageloc	2573
john anderson's legacy	2531
Market prenatation	2507

Query	Response Time
tribal communities	2467
fill marketing communcations	2437
kissan josepÃ¼	2231
"Environmental Impact Assessment"	2183
RME I	2161
brand / Fourier	1915
boundary spanners slaes	1855
fractions ks2	1802
urban realm magazine	1672
fluid mechnics	1589
janeway's immuno	1582
janeway's immuno	1582
t guntier	1567
120 idioms	1560
low speedshipping	1416
DKSH operationa	1291
gases separation	1288
Zhang, Duan	1133
creative brief communication	1008
TMPMgCl.LiCl turbo-Hauser base	422
Joseph Ambrose Banks	412
elizabethan social problems	342
101 warm ups	327
cervone / pervin	300
stokoe cooperative	299
ANSYS14 WORKBENCH TUTORIAL	277
air energi	273

Query	Response Time
marketing communications 512	244
maddness and women	243
denton corker marshal	204
Cultural Differences	204
verionica roth	184
Multifactor Leadership Questionnaire	181
ASHRAE 15-201	178
IMC creative brief	177
McKinsey 7-S Model	175
dale and apelbee	162
classroom management postivie	160
verionica roth	184
Multifactor Leadership Questionnaire	181
ASHRAE 15-201	178
IMC creative brief	177
McKinsey 7-S Model	175
dale and apelbee	162
classroom management postivie	160
phoenyx dactylifera	6467
corporate governance	6392
location strategy	6113
geese imprinting	5992
tadanafil AND solubility	5769
consuming frutose-sweetened	5681
dairy / GCC	5663

Query	Response Time
influenza champion	5435
pvrc bulletin	5398
risk oversight	5398
newspaper articles	5380
"liu tianye"	5319
Environmental Impact Assessment	5317
/ Corporate governance	5258
KAOLINITE SLURRIES	4689
project cost management	4681
soil sample	4658
diesel engine emittions	4657
dairy / GCC	4499
John Black asswssment	4461
social media	4357
dylan william assesssment	4344
uws ayr library	4343
sconul award 2013	4299
marketing eresources	4271
School readiness	3583
marks and spencer	3577
halpern and goldfarbv2013	3470
School readiness	3436
School readiness	3339
"environmental impact assessment"	3298
Environmental Impact Assessment	3235
School readiness	3090
DeÂ´centralisation budgeÂ´taire	159
Video tape collections	133
MIR spectroscopy	132

Query	Response Time
seismic safety hospita	131
intreraction design	124
joanna baillie's plays	122
hydraulics and pn	117
dispensing errors	114
M Dearing (Maria-Denise)	111
collaborative groupw-ork	108
Jan Keppler	107
roper roushka ayudyhya	105
AG 209	101
physics relativity	99
vanadium diabetes	95
chemistry spec	92
Georgie Geyer	90
subsea hand	86
titanic lynch	80
phillipa gregory	77
Steven Lukes	75
psycology papers	74
applied hyrdology	71
propaganda exam	68
Anne Bruetsch	67
qatar tourists	66
darley 1969	62
ashoka / restaurant	62
Buildings Earthquake effects	61
E. McCune	49
Tom D.. Dillehay	43
Steven Lukes	25