Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education. New Zealand: Asia-Pacific Society for Computers in Education

Using Data Analytics for Discovering Library Resource Insights – Case from Singapore Management University

LU Ning[1], SONG Rui[1], Dina HENG Li Gwek[1], Swapna GOTTIPATI[1]*& Aaron Tay[2]
School of Information Systems[1], SMU Libraries[2], Singapore Management University
*swapnag,aarontay@smu.edu.sg

Abstract: Library resources are critical in supporting teaching, research and learning processes. Several universities have employed online platforms and infrastructure for enabling the online services to students, faculty and staff. To provide efficient services by understanding and predicting user needs libraries are looking into the area of data analytics. Library analytics in Singapore Management University is the project committed to provide an interface for data-intensive project collaboration, while supporting one of the library's key pillars on its commitment to collaborate on initiatives with SMU Communities and external groups. In this paper, we study the transaction logs for user behavior analysis that can aid library admin to make operational decisions. The main challenges include the data quality and enormous datasets. Our solution not only provides the approach to data cleaning process but also suggest better visualization techniques for the user dashboard. Our experiment shows that the data cleaning process was effective in producing the insights from the library usage and the visualization techniques are efficient to summarize the big data. We used the datasets from Singapore Management University for this project.

Keywords: Library resources, students' e-resource usage, data analytics, visualization models, horizon graph techniques.

1. Introduction

Online university library resources have become integral part of several courses in education institutions for supporting education and research for all stakeholders. The key function of university library is to meet the information, research, and curriculum needs of its students, faculty and staff (Reitz, 2004; Roseroka, 2004) . The main focus is to support and improve the teaching, research and learning processes (Okunu et al, 2011). Many universities have online platforms and infrastructure in place to provide relevant information and respond to the users in an effective manner (Opoku, 2011). Continuous management and evaluation of the university libraries use is critical to ensure that the library is meeting the expected goals of the University. At the same time, adjustments should be made where necessary for effective information service delivery (Roseroka ,2004; KOUFOGIANNAKIS, 2012).

The resources collection at the Li Ka Shing Library, Singapore Management University is designed to provide resources necessary to support teaching and research by the Schools of Business, Economics, Accountancy, Information Systems and Social Sciences; and, courses taken by undergraduate and postgraduate students in all schools. The collection is interdisciplinary and covers the broad fields of business, economics and commerce with special strengths in finance, trade, accounting, management and international business.

Over years, the online collection has increased to over 250k e-book and e-journals of 80k titles. Some critical operational decisions that the library admins needs to make regarding the online resources include retaining of the resources, subscribing to new resources and unsubscribing to the resources. These massive amounts of information lead to various challenges in managing library resources. Manually studying the usage of resources is painstaking and tedious process and hence in this paper we propose an automated approach of studying the usage of library resources and thus aid in evidence based decision making process.

Data analytics has become more popular in various industries for finding insights in the business and making evidence based decisions by the business users (Sean Kandel et al., 2012; Baker et al,2009). The main aim of this project is to identify the patterns of how the university student population utilizes e-resources from their library. We devised novel solutions to distil information from millions of proxy log records. A data processing workflow and two visualization concepts – horizon chart and word cloud – has been proposed to analyse on numerous resource providers and user groups.

The remaining paper is structured as follows. Section 2 will be devoted to literature review on data analytics research in libraries. Section 3 describes our solution overview and its details. In section 4, we focus on dataset, experiments, results and discussions. We conclude in Section 5 pointing some interesting future directions of our work.

2. Literature Review

Enterprises rely on data analytics to gain insights about their customers, products, competitors, and partners to make evidence based decisions in the day to day operations (Sean Kandel et al.,2012). Recent research in education field is showing growing interest in applying data analytics on education datasets to discover insights and hidden patterns. Applying data analytics techniques in education is an emerging research field and also known as educational data mining (EDM). It involves development of methods and tools for making discoveries within the unique types of data from educational settings. The goal is to better understand the stakeholders of education and the learning settings, and to gain insights of educational phenomena (Baker et al., 2009).

Libraries play a key role in the education and research process. The major stakeholders of the library resources are students. In one of the recent surveys, Cenage Learning's Student Engagement Survey, more than 51% of the students responded that they are at the library to use the online databases, indicating that a good portion of their research work is completed at the library (Spring, 2015). Several researchers performed usability study to understand how users choose e-resources and thus identify ways to improve the access to e-resources(Fry et al., 2011; Maryellen Allen, 2001; Heather Jeffcoat King and Catherine M. Jannik, 2005). The approach is by survey based on Likert scale and open-ended questions (Fry et al., 2011) or surveys, focus groups and task-based testing (Maryellen Allen, 2001; Heather Jeffcoat King and Catherine M. Jannik, 2005). Some works used digital traces to understand the user behaviour in accessing the library resources. Analysis of digital traces aids in understanding of user behaviour in great details. Two types of web analysis are common; web search engine log analysis and digital library systems log analysis (Agosti, 2011). Stephen et al used webserver transaction logs to study how users find information from library college websites (Stephen Asunka and others, 2009). The focus of their work is to discover the issues relating to content access, interface design and general functionality of the website. Niu et al used transaction logs and surveys to understand users' search behaviour and their preferences and perceptions of the two systems. They used transaction log analysis to assess the scope and distribution of search queries, the use of search options, as well as query construction and refinement (Niu, 2014). Digital library system log analysis is based on transaction logs which are well-organized and explicitly described library collections. In our work we use transaction logs to discover the usage of the library resources by various schools and the trends of usage.

Finding insights from the data logs is very beneficial to the library managers, as "Information derived from such analysis can thus serve as valuable input into the redesign and refinement of the library's information systems, content architecture and the website's interface" (Asunka et al. , 2009). They serve as a huge motivation to find insights and support the decision making for the librarians. Furthermore, the data models are designed on various dimensions. For example, data arranged in chronological order, each semester's record was broken up into sections by month (Asunka et al. , 2009) can aid in discovering the trends. Such data structure can also include more dimension like student user profile, for example their course of study and admission year.

Just by accessing the raw data file, it provides no constructive clues to even start the analysis. Enlighten by Jansen, "A three-stage process composed of data collection, preparation, and analysis is presented for transaction log analysis" (Jansen, 2006). We leverage the ideas from (Jansen, 2006) to prepare our datasets. Hence, a more programmatic way of identifying the domain and logging errors is needed.

Despite the plethora of studies on library analytics, most remain on the theoritical level. In particular, few provide insights on the linkage between library digital contents and user demographics, which an important missing jigsaw in providing diversified services to the end user. In our work, we not only study the transaction logs to understand the e-resources usage analysis of library users in an efficient manner but also present the findings in much user friendly manner to the library admin for ease analysis. Comprehensive time horizon analysis (i.e. horizon chart) and content level analysis (word clouds) in the dashboards in our project is novel to the library research area.

3. Data Overview

Library proxy logs are the key datasets for this project. These are the digital traces of online library users of SMU. In this project we focus only on student data. The personal information and student demographics are anonymised. The dataset captures all user requests to external databases of e-books. SMU online library website sample and e-book data is shown in Figure 1. We observe that the SMU library website is enabled with various features for library users for easy information search across various types of datasets; books, articles, journals and exam papers.

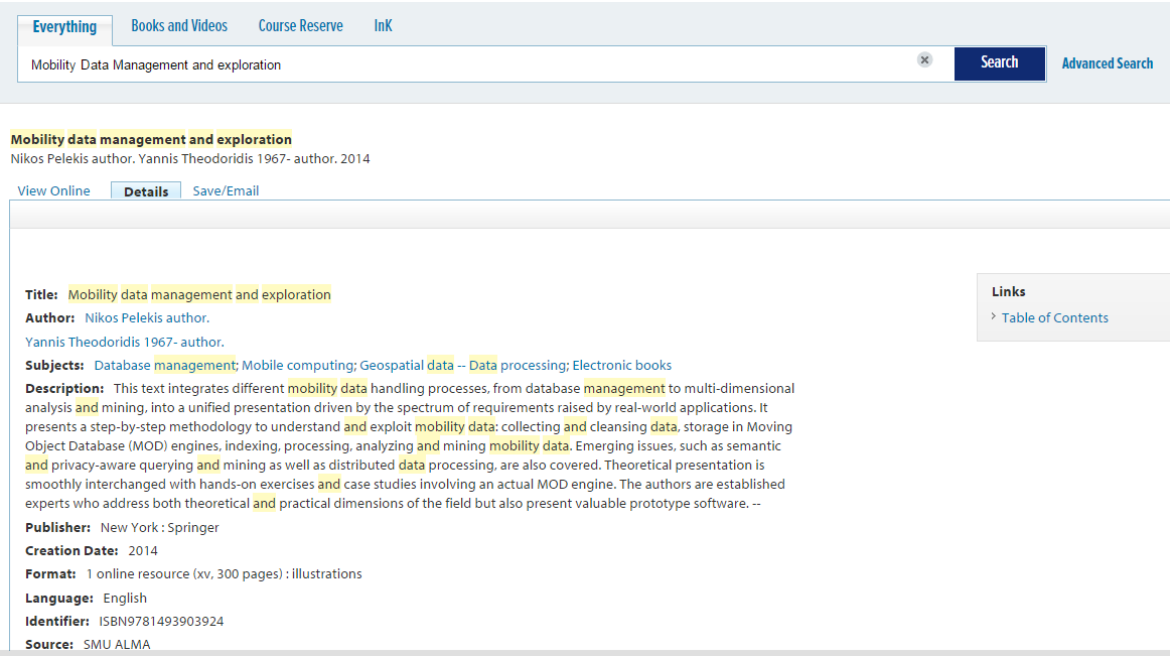


Figure 1. eBook sample from SMU library and screen shot of the SMU online library website.

The users leave digital traces whenever they use the system and this project uses the transaction logs. As with any datasets, this dataset also comes with many data quality challenges. The dataset should be pre-processed with various techniques to prepare for the visualizations and analysis. The below are the major data quality challenges in the dataset.

- *Duplicate requests* are defined as a user makes multiple identical requests with the same URL within 30-second time span Such requests are supposedly caused by auto page refresh.
- *Web assets* are used for page rendering and display. web assets do not help us understand user behavior as the requests are not generated by user and are database dependent.
- *Domain names* are not explicitly available from the proxy logs as all the users requests are directly to the external common e-resources page. They are embedded in the URLs. Domain names of the articles are critical to understand the type of the resource requested by the users.
- *E-book database names* are also not explicit in the logs and they are embedded in the URLs. To

generate the user friendly reports for library admin, the database names are critical.

In our solution, we handle these challenges using rules and pattern matching techniques. In next section we describe our solution approach with examples.

4. Solution Approach

In this section, we explain our solution model in detail. Figure 2 shows the solution design of library insights extraction. The three main stages in our solution approach are "data cleaning design process", "horizon chart design process" and "word cloud design process".

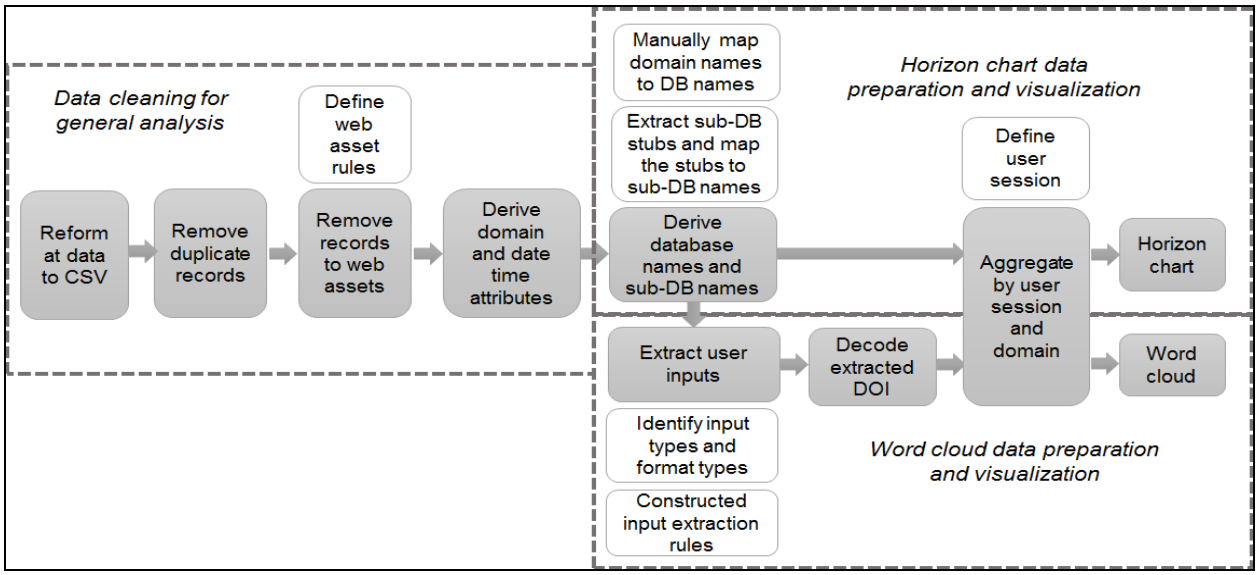


Figure 2. Solution design with three main processes- Data cleaning design, Horizon chart design and Word cloud design.

Data cleaning design process: The main tasks include domain, date and database names extraction from the web assets and URLs. The example domain name extraction is depcited in Figure 3. We apply pattern matching techniques to extract the domain names.

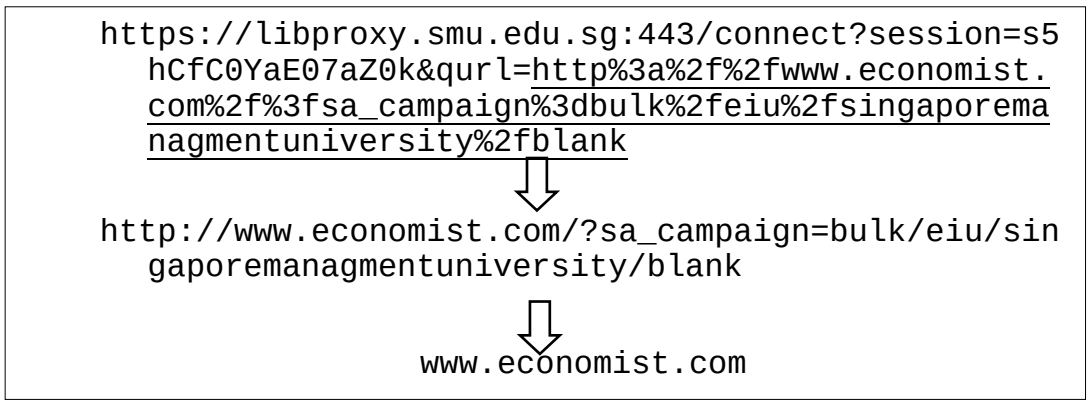


Figure 3. Example of domain name extraction from URL.

Similarly, rule matching techniques are applied to discover the e-book database names. Table 1 shows the sample e-book databases and domain names from our dataset.

Horizon chart design process: The popular technique for visualizing time series data and compare the datasets is horizon graph (Few, 2008). The horizon chart reduces space use by dividing the chart into bands and layering the bands to create a nested form. The rationale to utilise the horizon chart is that it overpowers the usefulness of simple trend line in comparing data over time for

many items within a category. The traditional line chart that squeeze all the curves on a single graph is not palatable when it comes to tens of databases. However, it is still necessary to put them on a same perspective to make logical comparisons. Eventually, horizon chart becomes the dominant alternative to this situation. Converting from the trend lines to colour coded horizon chart not only helps segmenting messy trend lines into clear and straightforward heat map rows, but also insulate each element to include the trend lines with abnormal value for better user analysis (J. Heer et al., 2008; Tableau, 2016).

Table 1: Sample e-book databases and corresponding domain URLs

e-book database	Domain URL patterns
LawNet	*.lawnet.sg
WestLaw	*.westlaw.*
Ebsco	*.ebscohost.com
My iLibrary	*.myilibrary.com
Ebrary	*.ebrary.com

Word cloud design process: Another approach to derive insights with the data in common log files is via text mining on the extracted user accessed contents. Word cloud is the most popular visuals as it best illustrates the most searched topics in an easily understandable way. This will help the librarians to understand what each school searches on and cater recommendations to them later on. Two aspects been chosen for aggregation is by school and by batch. The rationale behind this course of action is to find out what each school and each batch is searching for the most.

5. Experiments

In this section, we first explain our datasets followed by results and discussions. Our experiments are designed to evaluate the implicit suggestion extraction stage and the visualization stage.

5.1 Data Processing Results

The dataset contains 22,427 records, not only limited to full-time but also postgraduates and exchange students. The request log records are filed by months. The monthly numbers of records in request log data vary from 3 million to 11 million and the file sizes range between 1GB to 3GB. Requests are produced more in March, September and October, and few in school holidays. Figure 4 shows the statistics of the log requests in 2016.

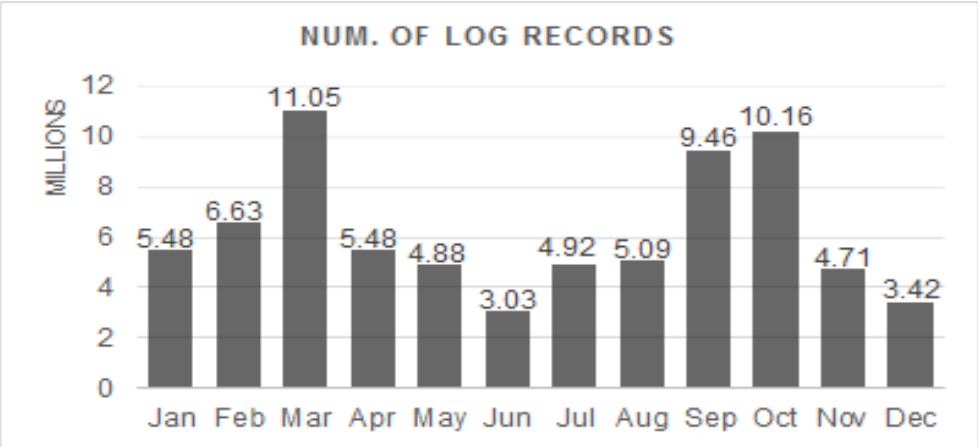


Figure 4. Number of log requests for each moth in 2016

As discussed in the previous section, the first task is the removal of duplicate request. The removed duplicate records amount to 5% of all proxy requests. Generally, the number of duplicate lines

correlates with the number of requests for each database. Only half of the requests are directed to web resources. In reality, page rendering typically requires many web resources files, which predicts a large percentage of web asset requests. The contradiction is explained by the fact that only a small proportion of non-web resources requests are requests to web pages loaded by browser. A large proportion of all requests are AJAX requests used by web pages to load parts of the web app interface.

5.2 Horizon Chart Analysis

Dashboard with filter enables the user to interactively filter and change the results being displayed. For instance, Figure 3 depicts the usage rate of e-library by 'Undergraduates' from 'School of Information System' and 'School of Accountancy' that are currently in their first year. "Daily percentage diff" is calculated from 'daily usage rate' of each domain to its own "yearly average usage rate". The values are then represented by two colour schemes. The bluish colour shows the less proportionate daily usages rate to its yearly average and on the other hand, the reddish colour shows the more proportionate end. The deeper colour represents the higher differences of the values. The horizon diagram will then only display the relevant usage rate per the criteria being selected. Furthermore, to compare the difference between the most and third popular domain used, we can click on the 'Domain' colour list.

Figure 5. Usage Pattern Dashboard on 1st year Undergraduate from SIS and SOA

From Figure 5, we observe some similarities lies within each trend lines. The usage rate of online library platform starts to surge from week 7 onwards across all three schools, and plunges during none school terms. This is in line with the school academic schedules, as well as the course schedules where most projects only kick starts in the second half of the semester.

However, exceptions are also very prominent in this case, as the students from SOB start to generate more accesses from week 5 onwards, which is much earlier than the other two schools. It may lead by a different course structure in school of business as they have an interim report due in the first half of the semester. Even though the usage rate by students from SOB and SOSS has three similar period of spikes in the second half of the semesters, SOSS students demonstrate a more consistent high demand than SOB students as the bar constantly filled with warm colours. We also observe that the students from school of business and law tend to behave differently across their term of study. It is also true that the contents they access for may not always align with their course of study.

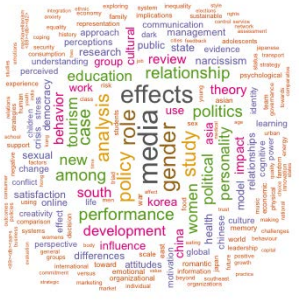
5.3 Word Clouds Analysis

Word clouds be generated on various dimensions. In our project, we collected data only about schools. Figure 6 shows word clouds aggregated by school/faculty and aids to find out the most popular search queries for different faculties.

Word cloud of Business School

Word cloud of Information Systems School

Word cloud of Accountancy School



Word cloud of Social Sciences School

Figure 6. Word Cloud for Faculties

From Figure 6, we observe that "Tourism" is a hot topic for various school. School of Business, Economics, Law and Social Sciences students search within their own major whereas School of Accountancy and Information Systems have the tendency to search outside their major. Its purpose is a preparation for the librarians to develop recommender system based on their search queries. This can also value add to the online library platform for the students by showing more relevant articles on top while they perform searching. Further, this also aids the library admin to study on subscriptions of e-books databases. This requires deeper analysis with more detailed datasets which will be our future work.

5.4 Discussions

In our solution, we adopted heuristic approaches to handle the data challenges to the specific task of extracting insights of library e-resources usage. However, they is a dire need for a conceptual library analytics framework that can be of more generalised and aid in various operational decision process in the library. Understanding the meaning of each URL is a challenging, if not impossible, task. Extracting content information from URLs that are generated by various content providers requires a large amount of manual work. However, programming can largely reduce the pain by sacrificing a small degree of accuracy for higher efficiency.

There are a few limitations in this study. Firstly, user contents are not always available from the URLs as many contents are encoded into database-specific IDs. This reduces the base of analysis and possibly led to biased results. Besides, few attributes of students were provided, this restrained in explaining certain usage patterns. If the registered courses were given, we could have obtained a more confident correlation between courses taken and e-resources accessed.

Instead of simply using trend line to represent the usage pattern of the e-databases, horizon chart that we built can give a clear overview on the library proxy log data. It is also able to interactively alter the results with various filters attached to the chart which enables it to discover new insights in the future. From the word cloud, the library can quickly recognise the most popular resources. It also helps the library to sieve out valuable information otherwise hard to identify from a pile of proxy logs. Word cloud and horizon chart implemented in this project are succinct and dense in information. These visualisations not only deliver useful information, but also serve as a proxy for users to interact with the data. The charts are still limited due to the type of data collected in our project. With additional data, the visualizations can be further improved.

6. Conclusion

The analysis on the proxy log data has revealed that the e-resources usage patterns from different groups of students are significantly different. Despite the ability to identify the user behaviours, we realised the courses are the key factor that influences e-resources usage patterns rather than students themselves. From another perspective, students are lacks the motivation to read on their own. Future study can be done on packaging the analysis pipeline and automating the process. This will enable the library admin team to dynamically generate reports or constantly monitor for resource anomaly. Such effort will greatly improve the timeliness and accuracy of the analysis at the point of request. Expanding the time dimension would produce trend chronological patterns in a longer time frame and how students' behaviours have evolved as well as how their topics of interest have changed over time. Further, we are working on the conceptual framework for library analytics for deeper analysis of the user behaviour and aid operational decision making process of the admin from library.

References

Agosti, M., Crivellari, F., & Di Nunzio, G. M. (2011). Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Mining and Knowledge Discovery, 24(3), 663–696. doi:10.1007/s10618-011-0228-8

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, 35(1), 33–45. Retrieved from

Baker, R.S.J.d., Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1 (1), 3-17.

Few, S. Time on the Horizon. Visual Business Intelligence Newsletter (2008) Jun/Jul 2008. Online at http://www.perceptualedge.com/articles/visual_business_intelligence/time_on_the_horizon.pdf

Fry, Amy and Rich, Linda, (2011) "Usability Testing for E-Resource Discovery: How Students Find & Choose E-Resources Using Library Websites" (2011). University Libraries Faculty Publications. Paper 14.

Heather Jeffcoat King and Catherine M. Jannik, (2005) "Redesigning for Usability: Information Architecture and Usability Testing for Georgia Tech Library's Website", OCLC Systems & Services 21 (2005): p. 236,

http://journals.ohiolink.edu/ejc/article.cgi?issn=1065075x&issue=v21i0003&article=235_rfu

J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In ACM CHI, pp. 1303–1312, 2009

Jansen, B. (2006). Search log analysis: What it is, what's been done, how to do it. Library and Information Science Research, 28(3), 407-432.

KOUFOGIANNAKIS, Denise. (2012) Academic Librarians' Conception and Use of Evidence Sources in Practice. Evidence Based Library and Information Practice, [S.l.], v. 7, n. 4, p. 5-24, dec. 2012. ISSN 1715-720X.

Maryellen Allen, (2002) "A Case Study of the Usability Testing of the University of South Florida's Virtual Library Interface Design", Online Information Review 26 (2002): pp. 40-53,

Niu, Xi; Zhang, Tao; and Chen, Hsin-liang, (2014) "Study of User Search Activities with Two Discovery Tools at an Academic Library"(2014). Libraries Faculty and Staff Scholarship and Research. Paper 106. http://dx.doi.org/10.1080/10447318.2013.873281

Okunu, H. O., Akalumhe, K. O., &Monu, J. O. (2011). An evaluative study of academic library services to users in Nigerian universities: A case study of FatiuAdemolaAkesode Library, Lagos State University, Ojo, Lagos. International Journal of Research in Education. 8(1), 74 – 80.

Opoku, D. (2011). Improving service quality in academic library: A managerial approach.International Journal of Research in Education. 8 (2), 198 – 209.

Reitz, Joan (2004). Dictionary for Library and Information Science. Westport, Connecticut: Libraries Unlimited

Roseroka, K. (2004). The roles of libraries.Association of African Universities.Retrieved on January 11, 2013 at http://www.aau.org/english/documents/librole.htm

Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer (2012) Enterprise Data Analysis and Visualization: An Interview Study Proc. IEEE Visual Analytics Science & Technology (VAST), Oct 2012.

Spring 2015 Student Engagement Insights survey, we asked: What do you do when you're at your college library? https://goodsteinlibrary.org/category/library-services/

Stephen Asunka and others, (2009) "Understanding Academic Information Seeking Habits through Analysis of Web Server Log Files: The Case of the Teachers College Library Website", Journal of Academic Librarianship 35 (2009): p. 33.

Tableau horizon charts. https://www.tableau.com/learn/tutorials/on-demand/horizon-charts