ISSN: 2473-3857
v2 n4062

Analysis of Student Web Queries

Jigar Jadav
jj03171w@pace.edu

Andrew Burke
ab73017n@pace.edu

Pratik Dhiman
pd46199n@pace.edu

Michael Kollmer
mkollmer@live.com

Charles Tappert
ctappert@pace.edu

Department of Computer Science & Information Systems
Pace University
Pleasantville, NY

Abstract

Search engines have become a vital part of everyday life. For informational, navigational, or transactional needs, most people utilize a search engine of their choice to quickly find what they are looking for. The number of mobile devices used in the K-12 learning space has been growing steadily. Most school-issued devices are required to have web filters installed on them to filter inappropriate content. These web filters log student activity online. This study investigates student web query logs and analyzes their web queries which may provide valuable information on student learning.

Keywords: Mobile learning, data analytics, term frequency, big data, information retrieval.

1. INTRODUCTION

The Internet is the primary source of information worldwide. Millions of web queries are formed daily. Numerous search engines exist in different countries and different languages that serve this need. Various proprietary search algorithms are used to retrieve relevant information to users. Not only has the use of search engines grown, but the number of studies related to it has also grown due to its demand (Spink, 2001). Since schools are starting one-on-one mobile device initiatives, we have little understanding of how students are using these devices. Through the application of data mining techniques and text analysis on web filter logs, we attempt to analyze student web queries to potentially understand the efficacy of these devices.

The K-12 learning space is evolving in both the United States and internationally. Students are given increasingly frequent access to the Internet through various platforms such as desktop computers, laptops, tablets, and other mobile devices. Some schools are distributing mobile

ISSN: 2473-3857
v2 n4062

devices to students in order to facilitate the integration of technology in the classroom. These devices have a web filter installed on them to filter inappropriate content irrespective to the Wi-Fi network to which they are connected. These web filters collect logs of student activities on the Internet. To date, however, this data has not been systematically analyzed.

This study analyzes real data collected from web filter of K-12 school issued mobile devices, in particular student search queries. As the use of mobile devices becomes more prevalent in K-12 education communities worldwide, the analysis from this study may reveal useful information to teachers, administrators and parents. Data are explored through exploratory text analysis to generate term frequencies and word clouds from student performed web queries. Later, these web queries are analyzed similar to the study done by Spink, et. al. to possibly gain insight into the efficacy of school distributed mobile devices.

Big data analytics is common in advertising, finance, medicine, and marketing due to its monetary potential. In recent years, there has been a spike in data mining related to students' academic performance at higher education institutions (Strecht, 2015). However, such quantitative research may also be beneficial to K-12 education.

2. BACKGROUND

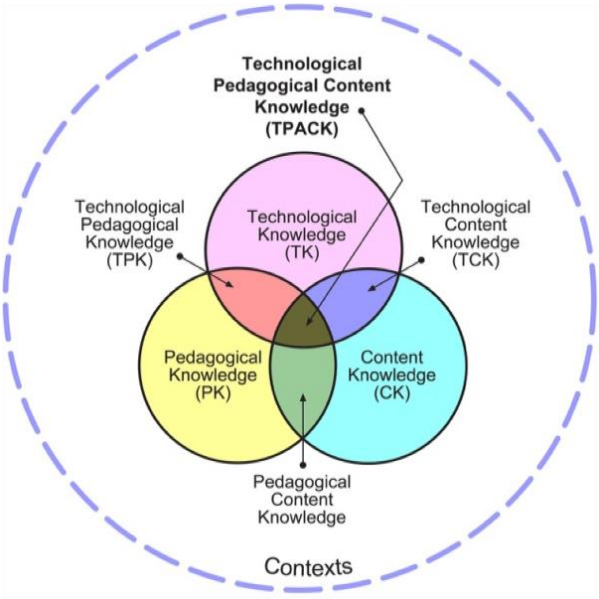
Mobile Learning

Mobile learning as a concept dates back to the 1970s when organizations began releasing recordings of dialect lessons that could be listened to at the learner's comfort (Sharples, 2010). There are three aspects to mobile learning: mobility of the technology, mobility of the learning, and mobility of the learner. Recent advances in smartphones and tablet devices have resulted in devices that meet these aspects of mobile learning. Behaviorally, this is relevant to why educational technology has moved toward the use of cellular phones, tablets, and other mobile devices as a medium for mobile learning. Mobility of the learner depends upon the individual and what kind of device they are using for the learning process. Researchers and practitioners of mobile learning are engaged in pioneering experiments for transmitting content to students by means of mobile cellular devices (El-Hussein, 2010).

TPACK Model

The Internet is the primary source of information worldwide. Millions of web queries are formed daily. Numerous search engines exist in different countries and different languages that serve this need. Various proprietary search algorithms are used to retrieve relevant information to users. Not only has the use of search engines grown, but the number of studies related to it has also grown due to its demand (Spink, 2001). Since schools are starting one-on-one mobile device initiatives, we have little understanding of how students are using these devices. Through the application of data mining techniques and text analysis on web filter logs, we attempt to analyze student web queries to potentially understand the efficacy of these devices.

TPACK builds knowledge from interactions between content, pedagogy, and technology. In order to enhance instruction by the addition of technology, TPACK requires an understanding of how to represent concepts with technology, knowledge of pedagogical techniques that employ technology constructively, understanding of how to use technology to support students' learning differences, and knowledge of how technology can be used to build on students' existing knowledge. This creates a foundation for effective teaching with technology (Koehler, 2009). The TPACK framework is widely accepted in the K-12 community. However, such educational frameworks have been evaluated mostly using qualitative data.



ISSN: 2473-3857
v2 n4062

Figure 1: TPACK Framework (Koehler, 2009)

There are several studies evaluating student learning with mobile devices (Falloon, 2014, Barbour, 2014, Jahnke, 2013, Fisher, 2013, Saine, 2012, Peluso, 2012, Jahnke, 2012, Hutchison, 2012). Chou et al. identified the positive impact of mobile devices on student learning. The area of interest is active engagement, increased time for projects, improved digital literacy, and digital citizenship. These authors suggest future research could also include quantitative data such as student performance, the amount of app usage by a teacher, and data on the technical skills of the participants for comprehensive data analysis (Chou, 2012).

Related Work

In 2001, Spink, et. al. did an extensive analysis of public web queries using search engine logs. They concluded that the number of terms used in search queries is small and that users tend to keep the subsequent search queries more or less the same. The number of results used per page is small compared to the number of results provided by the search engine. The language of web query is unique. In this study, student web queries are initially explored in a similar fashion.

3. METHODOLOGY

Raw data were extracted from the database of the web filter service deployed by the local district. Over 10,000 entries were collected from a two-hour time period in a school day. The following attributes were extracted: Suspicious, IP Address, User, User OU, User Groups, Computer Device ID, Search Query (SQ), Category Domain, Action, Rule Set (RS), Origin, Time. Of these attributes, SQ was used to create a corpus for text analysis.

The data were sorted based on attribute RS which segments the data into four user groups – staff, administrator, teacher, and student. Four different CSV files were created to store the queries (SQ) for each user group. All other attributes were removed from the files including the header.

Student Web Query Analysis

Following this study's methodology, raw data were extracted from the database of the web filter service deployed by the local school district. Over 10,000 entries were collected from a two-hour time period in a school day. The extracted attributes are described in Table 1. Of these attributes, SQ was used to create a corpus for web query analysis (Services, 2015).

Attribute	Description
Suspicious	A flag variable, true if the query's Category is in a prohibited set.
IP Address	The IP Address of the computer where the user logged in.
User	The logged-in user's network login name.
User OU	The logged-in user's Organizational Unit.
User Groups	The logged-in users groups.
Computer Device ID	The device name where the user logged in.
Search Query (SQ)	The word or phrase in the Search Query.
Category	The Content Database category assigned to the site at the time it was blocked.
Domain	The search engine used to make the Search Query.
Action	Shows whether the query was allowed or blocked.
Rule Set (RS)	The Rule Set assigned to the user.
Origin	One of "Internal Network" or "Proxy Server," indicating whether the user is on-campus or connecting remotely.
Time	The date and time the User made the Search Query.

Table 1: Data set attribute descriptions.

The data were sorted based on attribute RS which segments the data into four user groups: staff, administrator, teacher, and student. All other attributes were removed from the files. The data consist of 6,477 student web queries without any preprocessing. The following analysis was performed following (Spink, 2001):

1) *Term/token*: is a complete string of alphanumeric characters that the students entered in their search query. Terms were extracted separated by space.
2) *Query*: a collection of at least one search term as defined above. Queries can further be separated into two categories:
   a) *Unique queries* are all different queries sub- mitted in this session by one student. These queries include changes made to previous queries or new queries.
   b) *Repeat queries* are all multiple occurrences of the same query that represent request for multipage viewing. (Spink, 2001)
3) *Session*: search queries performed by a student over this two-hour time period. A session maybe a single query or multiple queries with or without repeats.

Term Frequency (TF) Analysis

Figure 5 shows the term frequency analysis of student search queries (SQ attribute) before preprocessing. It includes terms like "ww1" (appearing in three separate terms) that relate to the World War I unit, as well as terms like "the"

ISSN: 2473-3857
v2 n4062

that do not provide any insight into student learning behavior.

Number of students	315
Number of queries allowed	6460
Number of queries blocked	17
Total number of queries (allowed and blocked with repeats)	6,477
Number of unique queries	1,363
Mean of unique queries per student	4.54
Number of repeat queries	5,114
Mean number of repeat queries per student	20.56
Total number of terms (including repeat queries)	18,745
Total number of terms (unique queries only)	4,320
Unique terms	2,068
Mean terms per query (excluding repeat queries)	3.17
Mean terms per query (including repeat queries)	3

Table 2: Data Set Summary.

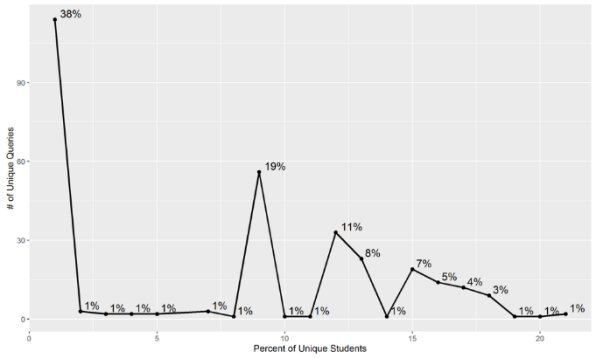


Figure 2: Number of unique queries submitted by each user.

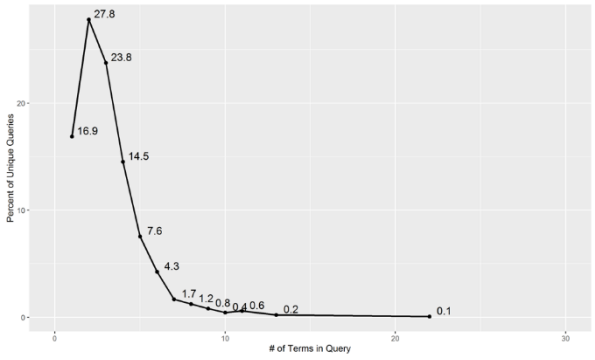


Figure 3: Number of terms appearing in each unique query.

In order to get meaningful insight from text, it is critical that the corpus undergo systematic and thorough preprocessing. The data were imported into RStudio running on a laptop with Windows 7 operating system.

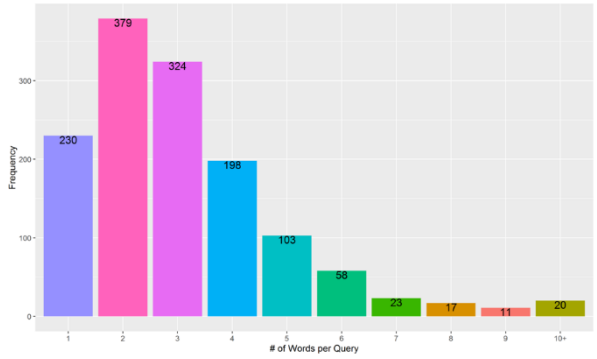


Figure 4: Frequency of terms per query.

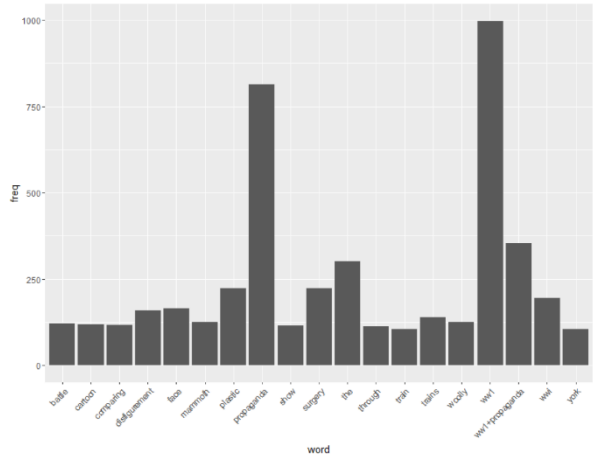


Figure 5: Frequent terms (100 or more) before preprocessing.

First, common text normalization techniques such as tokenization and case folding were applied to the data. Tokenization segments each term from the body of the text and case folding converts all the letters in the text to lower case to perform proper term counts (Services, 2015). Second, all punctuation, digits, newline, vertical tab, form feed, carriage return, and space characters were removed. Third, Porter's stemming algorithm was used to remove common terms (called stop words) such as "the", "a", "and", etc. (Porter, 2001). Finally, the data were saved as a text document and staged by creating a Document-Term Matrix (DTM) (Wang, 2011).

From the DTM, terms were organized based on their frequency. The total number of dimensions was calculated from the DTM. It was then exported into a CSV file for record keeping. Next, sparse terms were removed from the DTM. Lastly, the frequency of the least and most frequently occurring terms was explored.

ISSN: 2473-3857
v2 n4062

The following definition of term frequency was used to explore student data. Given a term t in a document, $d=\{t_1,t_2,t_3,\ldots t_n\}$ containing n terms, the term frequency of t in d is defined as the number of times t appears in d, as shown in Equation 1 (Services, 2015).

$$TF(t,d) = \sum_{i=1}^{N} f(t,t_i) \qquad t_i \in d;\ |d| = n$$

Where

$$f(t,t') = \begin{cases} 1, & if\ t = t' \\ 0, & otherwise \end{cases} \qquad (1)$$

Equation 1: Term Frequency (TF) Formula

In order to understand the term frequency analysis, a histogram of the terms appearing at least 100 times was created (Figure 6).

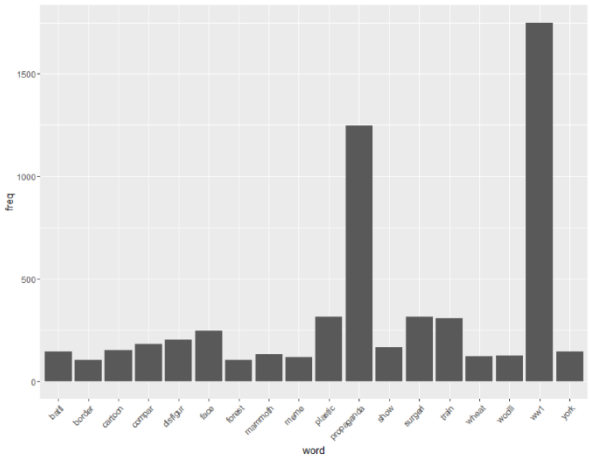


Figure 6: Frequent terms (100 or more) after preprocessing.

These frequency counts were used to generate a word cloud (Figure 7) to better visualize the terms (Brand, 2010).

4. RESULTS

Studies have shown that although content on the Internet changes rapidly, user's information needs and behavior remain consistent (Spink, 2001).

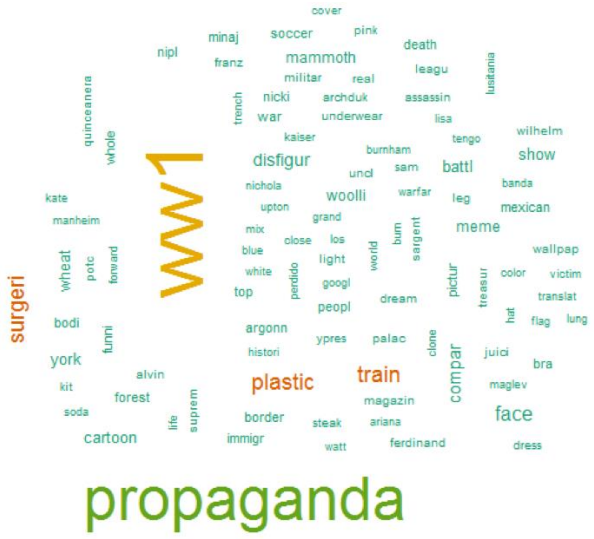


Figure 7: Word cloud of student search queries.

Word	Frequency	Word	Frequency
ww1	1750	border	102
propaganda	1249	war	79
plastic	315	world	59
surgeri	315	militar	55
train	301	ferdinand	54
face	205	archduke	53
disfigur	201	Franz	52
battle	135	assassination	48

Table 3: Frequency of terms related to World War I.

Comparable conclusion can be made where more than half the total number of students in our study performed two or less queries in a session. A total of 315 students were counted based on unique IP address. These students performed a total of 6477 queries of which 21% were unique queries, and 79% were repeat queries. The mean number for total queries in a session was 20.56, while the mean queries per student session was 4.54. Such disparity exists because when this data was collected, students were performing informational search queries related to World War I as shown in Table III. Once a query was performed, students were clicking on a number of links on the results page. See Table 3.

The number of students that performed a single unique query was 38%, two queries was 18.7%, three queries was 11% and four queries or more

ISSN: 2473-3857
v2 n4062

32.3% (Figure 2). Similar results were attained as (Spink, 2001) where the distribution of student search queries is skewed "toward the lower end of the number of queries submitted, with a long tail of very few users submitting a large number of unique queries" (Spink, 2001).

The mean number of terms in unique queries was 3.17. Figure 3 displays the number of terms in unique queries and its percentages. Based on this analysis, similar conclusions can be made as (Spink, 2001), web queries tend to be short. On the contrary, this study shows that the number of terms in web queries have gone up. Approximately 16.9% of queries had one term only, 27.8% had two terms, and 23.8% had 3 terms. However, close to 69% of all queries had either one, two or three terms. Less than 5% of the queries had more than seven terms.

The number of terms used in search queries may have gone up over the course of a decade. However, the largest distribution still hovers around two search terms. This result may be interpreted as a change in the way users perform search queries on search engines. Figure 4 shows the frequency of terms per query.

Figures 6 and 7 provide meaningful insight into the Internet search queries performed by students. After preprocessing our sample data and performing text analysis, 1877 unique terms with a total frequency of 16142 was obtained from the Document-Term Matrix. Table 3 illustrates the term frequency of the terms associated with World War I.

The sum of term frequencies in table 3 is 4973. Therefore, it can be concluded that approximately 31% of all the searches performed by students in this school district during this two-hour window were school related. However, the unit on World War I was taught only at the high school. The total number of high school students in this school district is approximately 30% of the entire student body. Therefore, it can be concluded that a significant portion of the queries performed by the students on school provided devices was related to schoolwork.

5. FUTUREWORK

The results of this study introduce a number of important follow-up questions:

What are the sociological and/or privacy ramifications of analyzing these data? They cannot be anonymized owing to the nature of our objectives, and there are certainly a dizzying number of studies or other intervention measures that might benefit from their analysis. Can binary classification of student web queries as school-related or non-school related provide us with more insight into student learning behavior online?

Is there a correlation between the total number of schoolwork related searches and non-schoolwork related searches to student GPA?
Can a teacher be provided with a real-time breakdown of student's schoolwork related searches to improve time on task?
Can this data analysis help teachers and guidance counselors identify students that need Response to Intervention (RTI) sooner?
Based on the usage history of paid apps, can school districts make better app investment decisions in the future? For example, is student time well-spent on the devices provided? Are certain activities more engaging than others?

The use of electronic devices in the classrooms is in its infancy. Technology, however, is an increasingly critical component in our students' educational path. Stakeholders have to make critical decisions as the benefit of such devices, and policy should leverage the available tools and data to their maximum utility.

6. REFERENCES

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

Barbour, M., Grzebyk, T. Q., & Eye, J. (2014). Any time, any place, any pace-really? examining mobile learning in a virtual school environment.

Chou, C. C., Block, L., & Jesness, R. (2012). A case study of mobile learning pilot project in K-12 schools. *Journal of Educational Technology Development and Exchange, 5*(2), 11-26.

El-Hussein, M. O. M., & Cronje, J. C. (2010). Defining Mobile Learning in the Higher Education Landscape. *Educational Technology & Society, 13*(3), 12-21.

ISSN: 2473-3857
v2 n4062

Falloon, G. (2014). What's going on behind the screens? *Journal of Computer Assisted Learning, 30*(4), 318-336.

Fisher, B., Lucas, T., & Galstyan, A. (2013). The role of iPads in constructing collaborative learning spaces. *Technology, Knowledge and Learning, 18*(3), 165-178.

Gagne, R. M. (2013). *Instructional technology: foundations*: Routledge.

Hutchison, A., Beschorner, B., & Schmidt-Crawford, D. (2012). Exploring the use of the iPad for literacy learning. *The Reading Teacher, 66*(1), 15-23.

Jahnke, I., Bergström, P., Lindwall, K., Mårell-Olsson, E., Olsson, A., Paulsson, F., & Vinnervik, P. (2012). Understanding, reflecting and designing learning spaces of tomorrow. *Proceedings of IADIS mobile learning*, 147-156.

Jahnke, I., Norqvist, L., & Olsson, A. (2013). *Designing for iPad-classrooms.* Paper presented at the European Conference on Computer-Supported Cooperative Work (ECSCW), 21-25 September, Cyprus.

Koehler, M., & Mishra, P. (2009). What is technological pedagogical content knowledge (TPACK)? *Contemporary issues in technology and teacher education, 9*(1), 60-70.

Peluso, D. C. (2012). The fast-paced iPad revolution: Can educators stay up to date and relevant about these ubiquitous devices? *British Journal of Educational Technology, 43*(4), E125-E127.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Services, E. E. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*: John Wiley & Sons.

Sharples, M., Taylor, J., & Vavoula, G. (2010). A theory of learning for the mobile age *Medienbildung in neuen Kulturräumen* (pp. 87-99): Springer.

Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., & Yang, Q. (2005). *Q 2 C@ UST: our winning solution to query classification in KDDCUP 2005*. Paper presented at the ACM SIGKDD Explorations Newsletter.

Shinas, V. H., Yilmaz-Ozden, S., Mouza, C., Karchmer-Klein, R., & Glutting, J. J. (2013). Examining domains of technological pedagogical content knowledge using factor analysis. *Journal of Research on Technology in Education, 45*(4), 339-360.

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

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance.

Wang, D., Zhu, S., Li, T., Chi, Y., & Gong, Y. (2011). Integrating document clustering and multidocument summarization. *ACM Transactions on Knowledge Discovery from Data (TKDD), 5*(3), 14.

ISSN: 2473-3857
v2 n4062