Understanding Secondary School Students' Study Habits through Analysis of Web Search Query Logs

Vikas Matcha, Samuel Mann, Shijian Xu, Wilford Norales, and Jigar Jadav
Seidenberg School of CSIS, Pace University, Pleasantville, New York

***Abstract* — Some secondary schools provide students with mobile devices to give them access to the wide array of information and digital learning tools available on such devices. Students' Internet usage is monitored through web search query logs, which can give insight into their study habits. The authors of past studies have designed an algorithm that classifies query logs as school related or non-school related. A linear regression then showed a strong positive correlation between school related queries and GPA. This study considered a much larger dataset and a MySQL database was created to store it. The Levenshtein algorithm was used in a key data cleaning task.**

***Index Terms*—Data Mining, Education, Internet, School, Search Query, Student.**

I. INTRODUCTION

Technology has become an integral part of education [1]. Simultaneously, education has relied increasingly on digital methods for learning such as online courses, in-class projections, interactive exercises, and more to improve learning [2], it has also found the need to assess the effectiveness of such methods.

In a 2003 Attewell and Winston studied two groups of students and analyzed their Internet use in an educational setting [1]. Students in the first group, who came from more affluent families and attended private institutions, made effective use of computers and the Internet. For example, a fourth-grade student from this group "posted messages to bulletin boards, read political candidate speeches online, answered online polls to make his opinions heard, and even created a Website so that his school can make use of it to conduct its own class president elections online" [1]. Students from the second group, who came from poorer and working class families and who scored lower in reading tests, did not show much proficiency in making effective use of the given digital tools. These students were often quickly frustrated when they could not find what they needed for research, and used the Internet for more entertaining non-school related purposes instead. This study demonstrates that the effectiveness of digital learning tools is not always clear and calls for a better understanding of student learning habits when using digital tools.

Some school districts provide mobile tablets to all students in an effort to make learning more hands-on, and in the hope that students will use the devices to learn on their own. However, digital learning devices are not always effective and administrators need to prove their effectiveness in increasing student engagement to continue getting funding for these devices [23]. School districts are legally required to install web filter programs on every device issued to students. The data provided by these programs is valuable in understanding students' usage and habits on educational mobile devices.

This study extended past research of secondary students' search queries performed on school issued mobile devices [10] to better understand students' learning habits and determine how effective the devices were. The first goal of this study was to create a MySQL database to store the large amount of student search queries. In the previous studies [10] [12] [13], data was stored in .CSV and .TXT files. However, such formats are not practical when working with large amounts of data. A MySQL database was created, which allowed easier

access to the data [12] in addition to holding the capacity needed for the study.

The second goal was to analyze the percentage of school related and non-school related queries over given periods of time. We study first focused on 24 hour periods to discover which times of the day students were more likely to study (i.e. performing school related queries on their devices). Secondly, the study considered longer periods of time, such as one semester, or one full school year. This gave insight into students' learning behaviors over those longer periods. Providing such information to educators will empower them to know how the learning activities they implement (i.e. group projects, homework, out of class research) impacts student engagement out of the classroom, and will allow them to better format their overall teaching methods.

II. RELATED WORK

Understanding student study habits has been a great concern for educators in the pursuit of successful learning outcomes. Indeed, as Coomes and DeBars explain, knowing students' overall attitudes, beliefs, and behaviors plays an instrumental role in education [24]. Students have often been classified into different groups depending on their apparent values and behaviors. Pioneering the study of student typology in the 1960s, Clark and Throw identified four subcultures that college students belonged to: "academic, vocational, nonconformist, and collegiate" [25]. Similar groupings were made by Horowitz: "college men and women, outsiders, and rebels" [24]. In 2010, Kuh, Hu, and Vesper analyzed the responses to a College Student Experience Questionnaire from 51,155 undergraduate students at 128 universities [25]. Engagement in educational activities and reported progress towards important outcomes of university were then compared. Analysis showed that the Individualist, Scientist, Conventional and Collegiate groups were above average in terms of engagement and positive outcomes, whereas the Disengaged, Grind, Recreator, Socializer and Artists types were below average [25]. Kuh et al. further explained how whichever subculture a student belonged to had a significant impact on the student's academic expectations, and on the way the student engaged in class. Being aware of this can help educators better interact with the student and reach out in better fashions toward better learning. Another critical point mentioned by Coomes and DeBars is the generational gap between instructors and students. The beliefs and expectations from a Generation X professor can be very different to those of his Millennial students [24]. This becomes very relevant when thinking of the good use of digital learning tools. Many times, instructors are not as technologically literate as their students, making it hard to integrate technology into the classroom. Contrarily, students are already familiar with iPads, phones, computers, and it is up to the instructors to add academic value to these devices [23].

Data Mining is the computational process of extracting information from a data set and transforming it into an understandable structure [3]. With the advancement of Data Mining (DM) techniques over the past twenty-years, the number of Educational Data Mining studies has progressively increased [4]. Such studies have brought valuable insight to educators, allowing them to predict student performance [5] or to better support students in need of special attention [6].

Sajadin et al. [7] conducted a research to analyze the relationships between student's behaviors and their success by developing student performance predictor using K-Means Clustering techniques [18] and Smooth Support Vector Machines (SSVM) classification [17]. They discovered that there was a strong relationship between mental condition of student and their final academic performance.

C. Marquez, et al. [8] performed research on identifying the factors that affect the low performance of students at different educational levels. They obtained middle school students' data from Zacatecas, Mexico. By using a classification algorithm and a few selected attributes, they found sociological, economic, or educational characteristics that may be more relevant in the prediction of low academic performance in school students.

R. Shanmuga Priya [9] conducted a study on improving the student's performance using Educational Data Mining based on 50 students from Hindustan College of Arts and Science, Coimbatore, India. By using a decision tree classification on eight attributes, it was found that the class test, seminar and attendance predicted the student performance. This method of predictions will help the teacher give special attention towards students who need it and improve student confidence on their studies.

Er. Rimmy Chuchra [20] applied the decision tree, clustering, and Neural network techniques to evaluate student performance by selecting the students from Sri Sai

University Engineering, Phagwara, India. In that, Chuchra found that teachers can easily evaluate student performance.

Khan [21] conducted a performance study on students comprising from the senior secondary school in India with a main objective to establish the prognostic value of different measures of cognition, personality, and demographic variables for success at higher secondary level in science stream. The selection was based on the cluster technique. The entire population was divided into clusters where a random sample of clusters was selected for further analysis. He found out that few factors affected the academic performance of the students.

III. BACKGROUND OF PAST WORK

Authors of [10] applied text analysis on student web query logs. Small samples of the raw data from the web filter database were taken to make the text classification easier. The authors then performed an intensive preprocessing of the raw data by selecting appropriate attributes. A Document-Term Matrix was created to stage the preprocessed data for analysis [11]. The authors then used various Data Mining and Natural Language Processing techniques to generate term frequencies [10]. Based on the term frequency analysis of these data, the results showed that students were using their mobile devices primarily to do school related work.

Authors of [12] performed an in-depth analysis of student web queries in a continuation of the work of authors [10]. Comma Separated Values (CSV), files containing anonymized web filter logs of each day, were created by an authorized school administrator. All available logs were merged together into one file. From the files, the authors selected 10,000 entries logged over a two-hour time period in a school day. The following attributes were taken to create a corpus for web query analysis: Suspicious, IP Address, User, User OU, User Groups, Computer Device ID, Search Query, Category, Domain, Action, Rule Set (RS), Origin, Time. Results showed that search queries tend to be short and that a significant number of queries performed by the students on school provided devices were school related [12].

Authors [13] extended the study of authors [12] by performing binary classification of student web queries as school related or non-school related. This work was conducted in three stages: first by collecting the data, secondly by specifying the classification model, and thirdly by evaluating the classification model [14]. The authors developed a new model architecture called Student Web Query Classifier (SWQC) after the traditional classifiers Support Vector Machines and Naïve Bayes yielded poor results [13].

Authors of [15] expanded the study by running a regressions analysis and found a positive correlation between school related search queries and GPA. For each student in the dataset, the percentage of school-related search queries was calculated from the total amount of search queries logged by that student and then mapped to that student's GPA. A regression analysis model was modified to include a search query threshold which represented a cutoff of what the minimum number of search queries of a student could have in order to be included in the model.

IV. METHODOLOGY

A. SWQC ALGORITHM

The algorithm first takes each search query of the student and runs the search on the Bing search engine using the Bing Web Search API [13]. The title and description (or "snippet") of each of the first 10 results are stored in the database, processed by removing stop words and punctuation marks [13], and remain associated to the initial query. The expanded queries are called "enriched queries". This step is critical: most original queries are two or three words long [12], which is not enough for the classifier to be accurate. Enriching each query by increasing its term count allows the classifier to better understand the meaning of the query and to make a more accurate classification. The enriched query is then compared with two corpora, one containing a list of school related words and the other containing a list of non-school related words. A few words in those lists are given more weight. The comparison determines the classification: if the weight of school related words in the data is more than the weight of non-school related, the search query is classified as school related, otherwise, it is classified as non-school related. There are two exception cases: 1) there are no matches between the words in the enriched query and the words in the school and non school related corpora, and 2) the number of

matches between the enriched query and the school related corpus, and between the enriched query and the non school related corpus are equal. The SWQC had been programmed to classify these ambiguous cases as non school related. This study devised a method to better handle these cases.

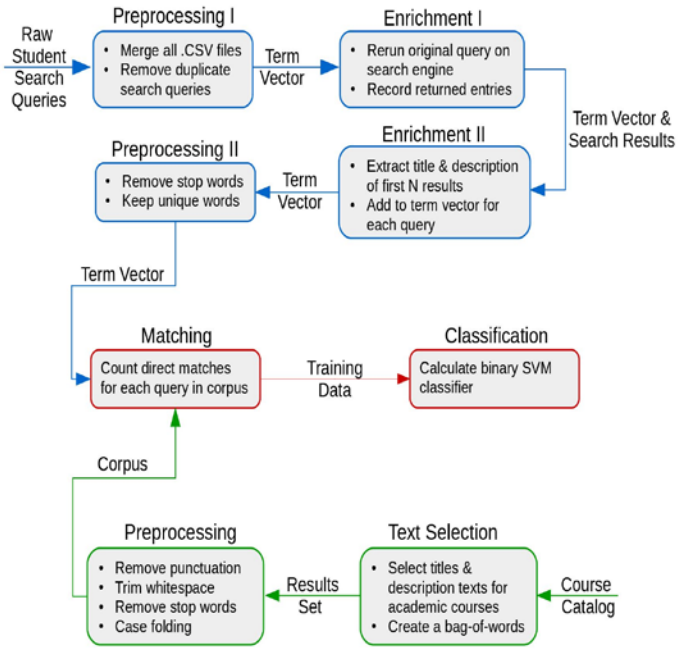

Figure 1. SWQC Algorithm

B. DATA CLEANING

1. PRELIMINARY CLEANING

This study considered two large data sets in the form of Excel files totaling approximately 1.14 million records. This data was raw and would have yielded poor results if fed to the classifier as it was, therefore an important preprocessing stage was necessary. Through careful examination of the data, the following cases which would have lowered the performance of the SWQC and skewed later analysis were identified.

EMAILS: The two problems with emails were: 1) the SWQC would always classify emails as non-school related because their special structure (e.g. name@domain.edu) would never match any word in the school related corpus, and 2) emails do not provide insight into the study habits of students being as an email lookup may be a consequence of school related work, but is not a research related query *per se*.

DOMAINS: The presence of domain lookups, for example "www.turnitin.com" or "http://www.nytimes.com/pages/politics/index.html" was identified. The SWQC was not designed to handle these cases and would automatically classify these queries as non-school related. While one could argue that certain domain lookups could be classified as school related (like the NY Times Politics link might be), many cases were very ambiguous and unperceptive. One potential expansion of the SWQC project would be to interpret domain lookups.

BLANKS: Blank rows and queries made of seemingly random sets of characters were present in the data, for example "CGMSBGj3IW4Y5MXHwAUiGQDxp4NLFZS44ZLRrjPWC_o_Z6YyL9y0r1c". As for domains, such cases would lower the performance of the SWQC because it was not designed to handle them.

Search Query	Domain	Time
depth echolocation learnt by novice sighted people	google	10/27/16 7:20
depth echolocation learnt by novice sighted	google	10/27/16 7:20
depth echolocation learnt by novice sighted	google	10/27/16 7:20
depth echolocation learnt by novice sighte	google	10/27/16 7:20
depth echolocation learnt by novice sigh	google	10/27/16 7:20
depth echolocation learnt by novice sig	google	10/27/16 7:20
depth echolocation learnt by novice si	google	10/27/16 7:20
depth echolocation learnt by novice s	google	10/27/16 7:20

Figure 2. Sample of "search as you type" entries

SEARCH AS YOU TYPE: Popular search engines like Google or Bing have the option to fill in what the engine thinks the user is looking for as the user is typing. "Search as you type" enabled search engines logged queries for every suggestion the engine made. Figure 2 shows that the desired query is "depth echolocation learnt by novice sighted people", but that the engine logged many subsets of that string. Hence, it was determined that only the longest string in "search as you type" cases should be kept and used for analysis.

Preprocessing was done in two stages. First, using the semi-manual method of Excel filters, emails, domains,

blanks and random sets of characters were eliminated from the two datasets. The total amount of queries was reduced from 1,140,049 to 1,131,071. The second and most effort intensive preprocessing task reduced the size of the dataset to 984,427 queries using the Levenshtein algorithm.

2. LEVENHSTEIN ALGORITHM

Comparing and finding the similarity of queries is the main part of removing "search as you type" in data cleaning. In order to find duplicated queries and keep the longest one, comparing each pair of search queries is necessary in this section.

It's not hard to process small amounts of data, but manually deleting similar queries will cost most of the resources when processing millions of items. Considering the structure of search data and calculating string metric, removing search as you type queries requires an efficient algorithm for comparing each pair of data. Based on the Levenhstein algorithm, we built a distance calculating function.

```
public static int distance(String a, String b) {
    int[] costs = new int[b.length() + 1];
    for (int j = 0; j < costs.length; j++)
        costs[j] = j;
    for (int i = 1; i <= a.length(); i++) {
        costs[0] = i;
        int nw = i - 1;
        for (int j = 1; j <= b.length(); j++) {
            int cj = Math.min(
                1 + Math.min(costs[j], costs[j - 1]),
                a.charAt(i - 1) == b.charAt(j - 1) ? nw : nw + 1
            );
            nw = costs[j];
            costs[j] = cj;
        }
    }
    return costs[b.length()];
}
```

Figure 3. "distance" function based on Levenshtein algorithm

C. REGRESSION ANALYSIS

Part of the cleaned data was selected to run a regression analysis. The tracked dataset was chosen so that a comparison with the results of [13] would be possible. This set comprised of 316,499 queries performed by 917 students between September 12th, 2016 and November 22nd, 2016. The queries were cleaned, enriched and classified with the processes explained above. Two students were removed from the analysis because of missing data (no GPA, no Queries). Using the same method as described in [13], a Pivot table was created to map each student identifier to the amount of school related searches it had performed, and to the student's GPA. The percentage of school related queries performed by each student was then calculated. A second Pivot table was created, showing the amount and classification of queries at each of the 24 hours of a day, and the average GPA of students who had done queries at that hour.

V. RESULTS

Summary statistics of the first Pivot table (Figure 4) showed that the mean percentage of school related searches was 43.1%, with a standard deviation of 1.99%, suggesting that usage habits were similar for most students. Total Queries, however, had a mean of 345.9 queries and a standard deviation of 401.07 queries, showing important variations in the amount of queries performed by students on their iPads. The distribution of Total Queries was skewed to the right (skewness = 4.313), which indicated most students were towards the lower end in terms of Total Queries performed.

	Total Queries	*School Related*	*SR %*	*GPA*
Mean	345.896	132.176	0.431	89.095
Standard Error	13.259	4.450	0.007	0.219
Median	239	94	0.408	90.944
Standard Deviation	401.072	134.622	0.199	6.630
Skewness	4.313	2.372	0.417	-1.196
Range	5234	1202	1	38.9333
Minimum	1	0	0	60.4
Maximum	5235	1202	1	99.3333

Figure 4. Summary Statistics of the dataset used for analysis

A regression analysis between percentage of school related queries and GPA was then performed. Authors of [13] had stressed the importance of using a threshold when running the regression analysis, which would exclude data from student having done less than a given

number of queries. Hence, a similar procedure was taken. At threshold T=0, the equation for the regression line was y = 3.1056x + 87.757, with a p value of $4.804*10^{-3}$ and a R-square value of $8.677*10^{-3}$. While the p value was higher than with the same threshold in [13], it was still low enough to reject the null hypothesis that percentage of school related queries has an impact on GPA. The R-square was lower than in [13], suggesting an even weaker correlation.

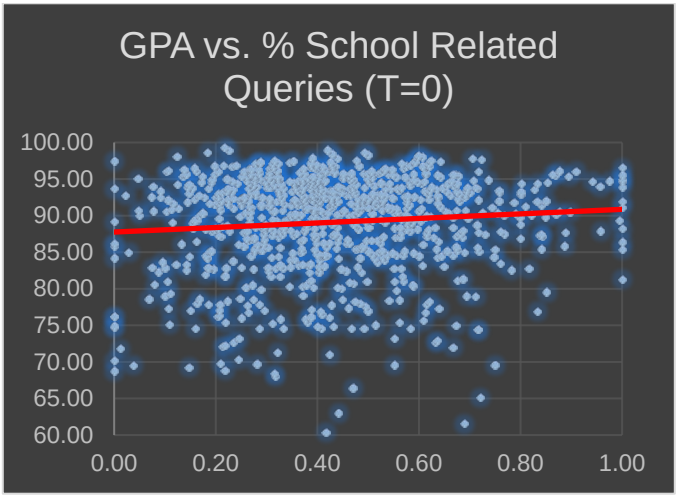

Figure 5. Regression line at threshold T=0, with equation y = 3.1056x + 87.757

Authors of [13] had found that discarding data from all students that had done less than 70 queries yielded the best correlation coefficient (R-square = 0.25) by still maintaining 13% of the student in the analysis. With the dataset used in this study, setting the threshold at 70 queries kept 84.48% of students as part of the analysis. The R-square, however, was not improved, dropping to $0.871*10^{-3}$. In fact, as shown in Figure 6, R-squared was very low for all threshold levels from 0 to 100. It only increased when setting much higher thresholds, but then too much data was being excluded from the regression: at T=500 with 20.22% of students still part of the regression, R-square was 0.025, and at T=1000 with only 5.73% of students being part of the regression, R-square was 0.132, indicating a very weak correlation between percentage of school rated queries and GPA.

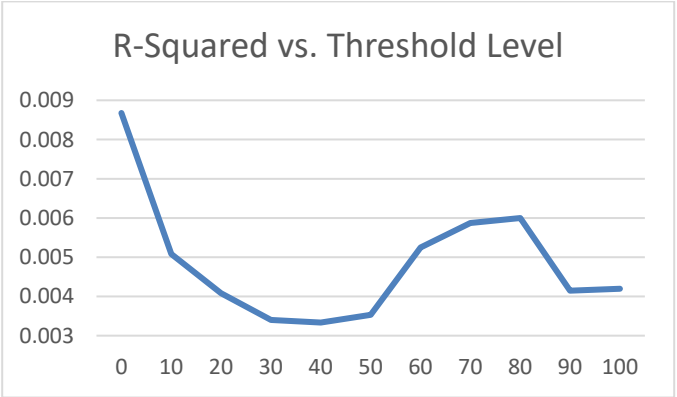

Figure 6: R-square remained low at all threshold levels

The analysis of usage over the twenty-four hours of a day gave interesting insight. First, as shown in Figure 7, usage patterns seemed to reflect the typical day of a student: usage was low from 0 to 6 in the morning, when students were sleeping. Usage rose during school hours starting at 8, peaking at 12 (lunch break) and then dropping at 3pm, when school ended and most students were in transit with no access to the internet or doing extracurricular activities. Usage then slightly increased as students got back home, and then dropped again as students went to sleep around 10-11pm.

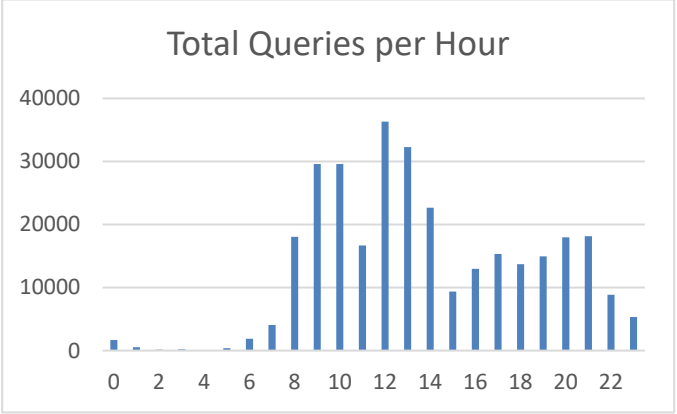

Figure 7: Total queries per hour for the period 9/12/16-11/22/16.

Interestingly, the distribution of school related queries over a day followed a similar pattern, with the percentage of school related queries rising during school hours, and then dropping during non-school hours. As Figure 8 illustrates, the highpoint was 3pm (45.07% SR) and the low point was 5am (16.67% SR). This analysis also showed that no matter what time of the day, the absolute majority of queries performed on school issued devices were not school related.

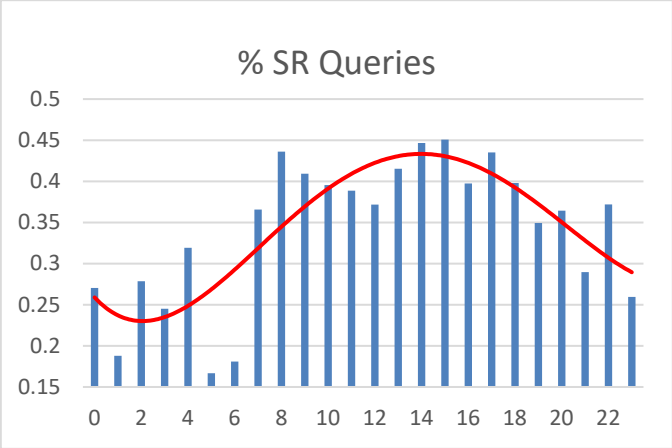

Figure 8: Percentage of school related queries at different hours of the day

Finally, an examination of average GPA of all students performing queries at any given hour showed that students who performed queries (both school and non school related) at 0am, 1am, 2am and 3am had GPAs of 83.42, 83.30, 76.36, and 74.97 on average respectively. This suggested that students using their devices in the middle of the night were weaker students. Their grades may have been impacted by lack of sleep. No other insight was found when looking at other times of the day, as the average GPA always hovered close to 89.

VI. CONCLUSIONS

Using a significantly larger amount of data than [13] did not confirm the correlation between percentage of school related queries performed by a student and GPA of that student. Even when testing various thresholds, the strength of the correlation did not exceed $8.677*10^{-3}$ (except when raising the threshold to levels where most of the students were not pat of the analysis). Analyzing time information provided interesting insight as to the habits of students. For example, the average GPA of students making queries n the middle of the night was significantly lower than the average GPA of all students. Knowing that a student is using his device instead of sleeping could be very helpful for educators to quickly identify students with bad habit, loosing motivation, or in need of special attention. Despite this insight, the results of this study call for an even better understanding of student's habits, and how they impact their performance in school.

VII. FUTURE WORK

Getting to understand student behaviors is a particularly ambiguous for teachers and school administrators. It is sometimes complicated to assess the effectiveness of methods implemented to increase students' interests and receptiveness to learning material. At a higher level, the effectiveness of educational master plans instituted by national education departments may also be ambiguous. Indeed, in the USA, public spending towards education increased by 73% from 1980 to 2005, with student to teacher ratio constantly decreasing, to reach its lowest level in 2005 [16]. But in the same period, literacy rates for 9, 13 and 17-year-old students did not change.

This study could be expanded in three possible ways. First, the data used in this study was collected only from secondary school students. But the use of digital devices is widespread at all levels of education, so an interesting expansion would be one that uses the techniques used in this study to analyze the learning behavior of students at different levels of education.

Secondly, a very useful expansion of this study would be to create a user-friendly interface for educators to collect data from their own students and to apply analysis methods discussed in this study. This study would be the most useful to schoolteachers, but most of them the lack technical skills to use our tools as it is. Therefore, building an easy to use Graphical User Interface (GUI) to operate SWQC algorithm would be valuable to educators.

This study considered tracked and untracked queries. Tracked queries are very useful because they can be traced back to a single student and can give insight about his/her habits rather than general trends of a larger cohort. A third possible expansion of the SWQC would be to classify students into groups based on their search queries. As discussed in the Related Work section, typologies are useful to understand the attitudes, beliefs, and behaviors of students. A student's type could be discovered from his online searches: a student having many science related queries may be grouped as "academic", while student with many queries in a specific topic, such as music, politics, or a professional area, might be classified in the "vocational" group. Identifying the values of a student takes time and this tool would allow instructors to understand their students better and faster, as well as allow them to keep track as they evolve.

REFERENCES

[1] M. Warschauer. “The paradoxical future of digital learning.” Learning Inquiry, vol. 1, pp. 41-49, Mar. 2007.

[2] J. Watson, L. Pape, A. Murin, B. Gemin, L. Vashaw. “Keeping Pace with K-12 Digital Learning, 11th Edition” Evergreen Educational Group, 2014.

[3] G. Piatetsky-shapiro and W.J. Frawley, Knowledge Discovery in Databases, AAAI/MIT Press, 1991.

[4] A. Dutt, M. Ismail and T. Herawan, "A Systematic Review on Educational Data Mining", IEEE Access, 2017.

[5] T. Devasia , Vinushree, V. Hegde. “Prediction of Students Performance using Educational Data Mining”, IEEE Access, 2016.

[6] B. Kumar and S. Pal, "Mining Educational Data to Analyze Students Performance", International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, 2011.

[7] Sajadin Sembering, M.Zarlis, “ Prediction of student academic performance by an application of data mining techniques”, International conference on management and Artificial Intelligence- 2011

[8] C. Marquez-Vera, C. Romero and S. Ventura.” Predicting School Failure Using Data Mining” -2011

[9] K. Shanmuga Priya” Improving the student’s performance using Educational data mining”, International Journal of Advanced Networking and Application, Vol.4,pp-1680- 1685 (2013)

[10] J. Jadav, C. Tappert, M. Kollmer, A. Burke, and P. Dhiman, "Using text analysis on web filter data to explore k-12 student learning behavior," in UEMCON, IEEE Annual, 2016, pp.1-5

[11] “Basic text mining in r,” https://goo.gl/UMk8UF, accessed: 2016-02-13.

[12] J. Jadav, A. Burke, P. Dhiman, M. Kollmer, and C. Tappert, “Analysis of Student Web Queries,” in Proceedings of the EDSIG Conference ISSN, 2016, p. 3857.

[13] J. Jadav, A. Burke, P. Dhiman, M. Kollmer, and C. Tappert, "Classification of Student Web Queries," in Proceedings of the CCWC, IEEE, 2016.

[14] E. E. Services, Data Science and Big Data Analytics: Discovering,Analyzing, Visualizing and Presenting Data. John Wiley & Sons, 2015.

[15] J. Jadav, A. Burke, G. Goldberg, D. Lindelin, A. Preciado, C. Tappert, and M. Kollmer, "Correlation Discovery Between High School Student Web Queries and their Grade Point Average," in Proceedings of the CCWC, IEEE, 2016.

[16] McKinsey & Co. “How the World's Best-performing School Systems Come out on Top.” New York: McKinsey, 2007.

[17] Y.J. Lee. And O.L Mangasarian, “ A Smooth Support Vector Machine for classification”,Journal of Computational Optimization and Applications.20, 2001, pp.5-22

[18] Romero, C. and Ventura, S.,” Educational Data mining: A survey from 1995 to 2005”, Expert systems With Application” (33) 135-146. 2007

[19] James Rumbaugh, "The Unified Modeling Language Reference Manual, ” 2nd editon, Boston, Pearson Ed Inc, 2004.

[20] Er.Rimmy Choura “Use of Data Mining Techniques for the evaluation of student performance: A Case Study”, Vol 1 Issue 3 October 2012.

[21] Z. N. Khan, “Scholastic achievement of higher secondary students in science stream”, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[22] Gonzalo Navarro, “A Guided Tour to Approximate String Matching,” Dept. of Computer Science, University of Chile, Blanco Encalada 2120 - Santiago – Chile.

[23] O. Hlodan, “Mobile Learning Anytime, Anywhere,” *BioScience*, vol. 60, no. 9, pp. 682–682, 2010.

[24] M. D. Coomes and R. Debard, “A generational approach to understanding students,” *New Directions for Student Services*, vol. 2004, no. 106, pp. 5–16, 2004.

[25] Kuh, George D; Hu, Shouping; Vesper, Nick. “They Shall Be Known by What They Do: An Activities-Based Typology of College Students.” *Journal of College Student Development*; Baltimore41.2 (Mar/Apr 2000): 228.