ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 2, February 2020

Determining Top Users of Google Application Using K-means Clustering

Eugene G. Ranjo [1], Ariel M. Sison [2]

College of Engineering and Information Technology, University of Southern Mindanao, Kabacan, Cotabato, Philippines[1]

School of Computer Studies, Emilio Aguinaldo College, Manila, Philippines[2]

ABSTRACT: Monitoring of user's activity is a very significant factor to manage its online services effectively and efficiently. Every access to the server has its own logs that can be transformed into useful data. This research paper will use a clustering algorithm to analyze the data which generated using Google Server logs. K-means algorithm is by far the most popular clustering tool used in web log analysis. Logs can be analyzed to determine the user's usage in particular applications and other online services. The two main purpose of this paper is to determine the top users, and determine the applications which has the highest usage. In general, the outcome of this study can make the management aware of and carefully plan in allocating its resources effectively such as Internet bandwidth.

KEYWORDS: Data mining, k-means Clustering, Google Application, Top Users, User Access, Access Logs

I. Introduction

The rapid increase of digital technologies such as the use of email, virtual learning environment, the internet, and multimedia for instructional programs was the reason behind the growing number of academic institutions [1]. Therefore, communication between faculty, staff, students, and their alumni is increasing too, and it is considered a key role in achieving the education objectives. Mode of technology such as emails presents new ways in communication. The influence of new technologies on education is increasing proportionally on the development of technologies that offer new alternatives for communication [2].

Since, most State Universities and Colleges in the Philippines experience budget constraints, they started to find new ways to obtain and control computing resources with low expenses [3]. Cloud computing or software-as-a-service (SaaS) is a web technology. It is a computing model that IT applications are supplied as a service to enable users to reach applications from the cloud internet [4]. By using cloud computing, clients do not have to buy and own extra hardware, network equipment as well as concern about maintenance costs [5]. Academic institution are paying more attention and priority to acquire specific hardware, software, and advanced network rather than developing their information systems and academic staff skills to utilize the available technologies.

Google Apps as one of cloud computing applications has become popular and it is our belief that it can be effectively used in educational institutions for communication between academic staff and their students. Google Apps, as an online services, is easy to use, dependable, useful and productive to enhance communication in academic institution [4]. In addition, Google Apps services are cloud computing technology where users can reach Google Apps services anywhere and at any time via the cyberspace. Google Apps for Education (GAFE) is a core suite productivity applications that Google offers to schools and educational institutions for free. These communication and collaboration applications include Gmail, Calendar, Drive, Docs and Sites, and a GAFE account unlock access to dozens of other collaborative tools supported by Google [6].

DOI: 10.15680/IJIRCCE.2020. 0802001

The University of Southern Mindanao is one of the academic institutions that subscribes this online service. The account is under the education program of Google with free subscription. The university, through its University Information and Communication Technology Office task to manage the creation of accounts that are open to all faculty, staff, and students. This study aims to determine the top users and determine the applications which have the highest usage by performing a cluster analysis using K-means clustering. Data will be analyzed and the result will be the basis for possible intervention to ameliorate the use of this application to help the learning process of both faculty and students.

II. Conceptual Framework

Review of Related Literature

The heterogeneous and unstructured information available in the web reduces the analysis to a larger extent. Thus, the preprocessing phase is a prerequisite for discovering patterns and it transforms the raw click stream data into a set of user profiles. Data processing is a difficult task with a number of challenges. Hence, this work gives rise to the variety of algorithms and heuristic techniques as merging and cleaning, user and session identification, and etc. Various research works are carried in different phases of preprocessing [18][19].

Data mining [20] is the exploration and analysis of large data sets, in order to discover meaningful pattern and rules. The key idea is to find effective way to combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is designed for, and work best with large data sets. Data mining is the component of wider process called knowledge discovery from database [20]. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data mining algorithm, analyzing results and taking appropriate action. The data which is accessed can be stored in one or more operational databases. In data mining, the data can be mined by passing various process as shown in Fig.1.

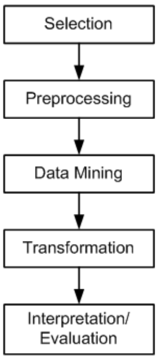


Fig. 1. The process of Data Mining [20].

Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets. It is a process of grouping data objects into disjoint clusters so that data in the same cluster are similar, and data belonging to different cluster are differ. Many algorithms have been developed for clustering [7]. One of the most popular clustering method is k-means clustering algorithm. It generates k points as initial centroids arbitrarily, where K is a user specific parameter. Each point is then assigned to the cluster with the closet centroid [8][9][10]. Then the centroid of each cluster is updated by taking the mean of the data points of each cluster. Some data points may move from one cluster to other cluster. Again, we calculate new centroids and assign the data points to the suitable clusters. We repeat the assignment and update the centroids, until convergence criteria is met i.e., no point changes clusters, or equivalently, until the centroids remain the same.

According to Shi, Peilin [15] there are three important aspects of web usage mining, namely clustering, association, and sequential analysis are often used to study important characteristics of web users. Web clustering involves finding natural groupings of web resources or web users. However, there exist some important differences between clustering in conventional applications and clustering in web mining. The patterns from web data are non-numerical, thus Runkler and Beadek [16] proposed relational clustering method to group non-numerical web data. Furthermore, due to a variety of reasons inherent in web browsing and web logging, the likelihood of bad or incomplete data is higher than conventional applications. The clusters tend to have vague or imprecise boundaries [17]. A pattern may belong to more than one candidate clusters by different degrees of the memberships.

K-means algorithm is by face the most popular clustering tool used in web log analysis [11]. In this paper, K-means algorithm is used in mining the logs generated from Google server to find any relevant groups that would relate to determining top users and determine the applications which have the highest usage as shown in Fig.2.

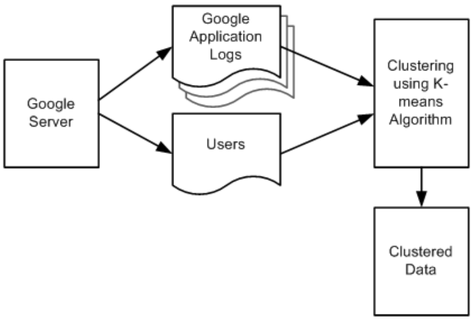
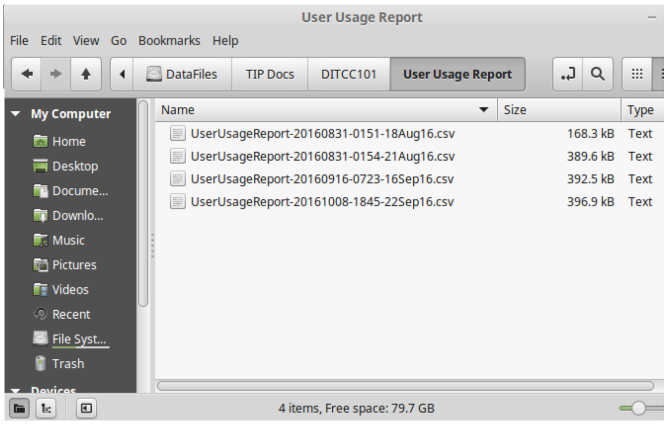


Fig. 2. Conceptual Framework.

III. Operational Framework

The study will use the user usage report generated from Goolge Apps administration dashboard which has 400 users of the University of Southern Mindanao, Kabacan, Cotabato. The user usage report was generated with specific dates (August 18, 2016, August 21, 2016, September 16, 2016, and September 22, 2016) as shown in Fig.3.

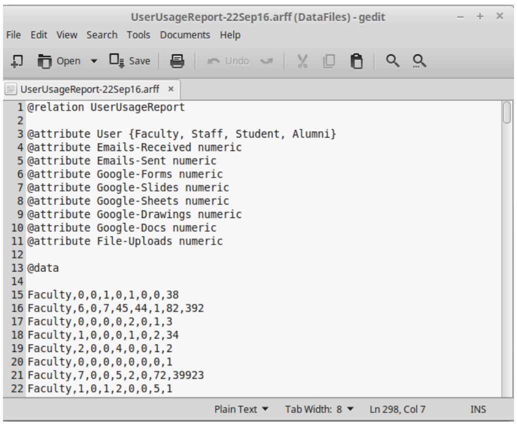
Fig.4. shows the sample data set with 400 instances and 9 attributes using Weka.



Name	Size	Type
UserUsageReport-20160831-0151-18Aug16.csv	168.3 kB	Text
UserUsageReport-20160831-0154-21Aug16.csv	389.6 kB	Text
UserUsageReport-20160916-0723-16Sep16.csv	392.5 kB	Text
UserUsageReport-20161008-1845-22Sep16.csv	396.9 kB	Text

4 items, Free space: 79.7 GB

Fig.3. Downloaded logs user usage report.



```
@relation UserUsageReport

@attribute User {Faculty, Staff, Student, Alumni}
@attribute Emails-Received numeric
@attribute Emails-Sent numeric
@attribute Google-Forms numeric
@attribute Google-Slides numeric
@attribute Google-Sheets numeric
@attribute Google-Drawings numeric
@attribute Google-Docs numeric
@attribute File-Uploads numeric

@data

Faculty,0,0,1,0,1,0,0,38
Faculty,6,0,7,45,44,1,82,392
Faculty,0,0,0,0,2,0,1,3
Faculty,1,0,0,0,1,0,2,34
Faculty,2,0,0,4,0,0,1,2
Faculty,0,0,0,0,0,0,0,1
Faculty,7,0,0,5,2,0,72,39923
Faculty,1,0,1,2,0,0,5,1
```

Fig.4. Sample Dataset

K-means Clustering

The term "k-means" was first used by James MacQueen in 1967 [13]. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982. K-means is a widely used partitional clustering method in the industries. The Kmeans algorithm is the most commonly used partitional clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time [13].

The algorithm is composed of the following steps:

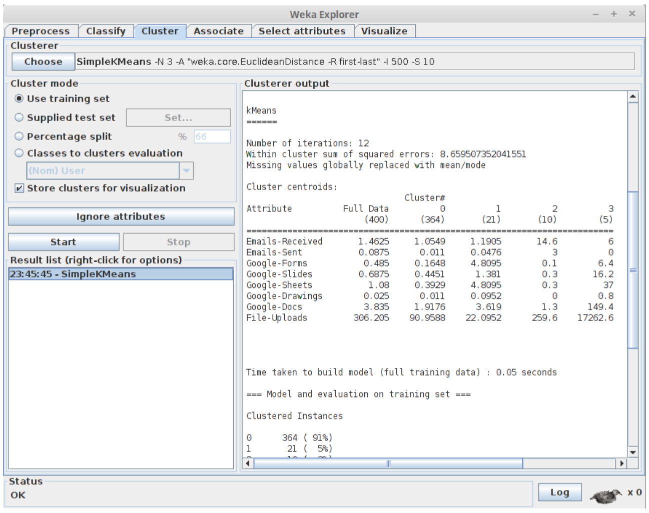
1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

The analysis using k-means clustering is being done with the help WEKA data mining software.

WEKA is a collection of machine learning algorithms for data mining tasks [13]. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka is open source software issued under the GNU General Public License thus every researcher can access the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license.

IV. RESULTS

The analysis is being performed on the basis of logs generated from the Google Server. There are 400 users involved in the logs generated. The result generated using Weka software is shown in Fig.5. It has shown that after clustering the data into 4 with a class attribute: user, cluster 0 is for the student with the highest percentage of 91%. Cluster 1 is for the faculty with 5%, cluster 2 is for the alumni with 3%, while cluster 4 is for the staff with 1%.



Clusterer
SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Nom) User
☑ Store clusters for visualization

Result list (right-click for options)
23:45:45 - SimpleKMeans

Clusterer output

```
kMeans
======

Number of iterations: 12
Within cluster sum of squared errors: 8.659507352041551
Missing values globally replaced with mean/mode

Cluster centroids:
                              Cluster#
Attribute         Full Data          0          1          2          3
                      (400)      (364)       (21)       (10)        (5)
=========================================================================
Emails-Received      1.4625     1.0549     1.1905       14.6          6
Emails-Sent          0.0875      0.011     0.0476          3          0
Google-Forms          0.485     0.1648     4.8095        0.1        6.4
Google-Slides        0.6875     0.4451      1.381        0.3       16.2
Google-Sheets          1.08     0.3929     4.8095        0.3         37
Google-Drawings       0.025      0.011     0.0952          0        0.8
Google-Docs           3.835     1.9176      3.619        1.3      149.4
File-Uploads        306.205    90.9588    22.0952      259.6    17262.6




Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      364 ( 91%)
1       21 (  5%)
```

Status
OK

Fig.5. K-means clustering using Weka.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

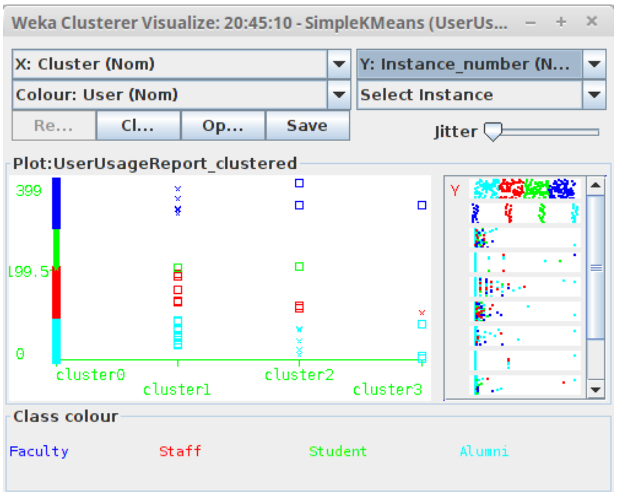
Vol. 8, Issue 2, February 2020

Table 1 presents the results of different Google Applications in terms of user usage. File Uploads gets the highest mean of 306.205 followed by Google Docs with the mean of 3.835. Emails Received with the mean of 1.463, and Google Sheets with the mean of 1.08. Google Slides gets its shares in user usage with the mean of 0.688, Google Forms with the mean of 0.485, Emails Sent with the mean of 0.088 and Google Drawings with the mean of 0.025.

Table 1. Statistics after clustering.

Statistics	Emails Received	Email Sent	Google Forms	Google Slides	Google Sheets	Google Drawings	Google Docs	File Uploads
Min	0	0	0	0	0	0	0	0
Max	104	7	18	45	101	3	309	39923
Mean	1.463	0.088	0.485	0.688	1,08	0.025	3.835	306.205
Std. Dev.	5.867	0.584	1.645	2.857	6.159	0.199	21.605	2821.437

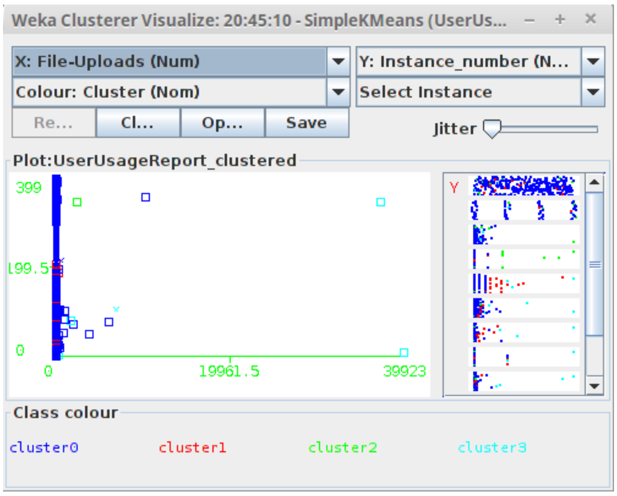
Fig.6. presents a visualization of clustered assignments with cluster 0 for the student got the highest percentage of 91%. It shows that cluster 0 has the highest user usage in terms access in the different Google Applications. Fig.7. presents a visualization of clustered assignments with cluster 0 got the highest user usage in terms of File Uploads application.


Weka Clusterer Visualize: 20:45:10 - SimpleKMeans (UserUs...

X: Cluster (Nom)
Y: Instance_number (N...
Colour: User (Nom)
Select Instance

Plot:UserUsageReport_clustered

Class colour


Weka Clusterer Visualize: 20:45:10 - SimpleKMeans (UserUs...

X: File-Uploads (Num)
Y: Instance_number (N...
Colour: Cluster (Nom)
Select Instance

Plot:UserUsageReport_clustered

Class colour

Fig.6. Visualization of clusters using Weka.

Fig.7. Visualization 2 of clusters using Weka.

V. Conclusion and Recommendation

In this research paper, K-means clustering was applied. The method can assist the office University Information and Communication Technology Office of the University of Southern Mindanao to evaluate and monitor its online services. As shown in the result, Students gets the highest percentage of user usage followed by the faculty, alumni, and staff. This study will play an important role in planning, monitoring, and management of its online services. Based on the results, the university should consider subscribing or increase its Internet bandwidth that can be allotted to students and faculty for them to create, share, and join forces in everyday learning processes. Furthermore, periodic training and workshops for faculty, staff, and students regarding Google Apps to maximize the use of free cloud computing technology.

REFERENCES

[1] L. Kagima and C. Hausafus, "Integration of Electronic Communication in Higher Education: Contributions of Faculty Computer Self-Efficacy", *The Internet and Higher Education*, vol. 2, no. 4, pp. 221-235, 2000.

[2] A. Baghestan, M. Zavareh, and M. Abu Hassan, "Communication channels used by academic staff in interacting with their students". *Pertanika Journal of Social Sciences & Humanities*, *17*(2), 167-178, 2009.

[3] 2016. [Online]. Available: http://data.gov.ph/infographics/ched-highereducation/d4. [Accessed: 30- Sept- 2016].

[4] T. Sviridova, L. Sviridova, and B. Tymoshenko. "Google Apps as solution of communication issues in educational process". In *Perspective Technologies and Methods in MEMS Design*, 2011.

[5] B. K. Bates, "Using Google Apps in Professional Learning Communities," pp. 1– 9, 2011.

[6] B. Holland, "GAFE - Using Google Apps for Education in the Classroom", EdTechTeacher, 2016. [Online]. Available: http://edtechteacher.org/gafe. [Accessed: 03-Oct- 2016].

[7] S. Na, L. Xumin and G. Yong. "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm". *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*, Jinggangshan, pp. 63-67, 2010.

[8] C. Zhang and S. Xia, "K-means Clustering Algorithm with Improved Initial Center," *2009 Second Int. Work. Knowl. Discov. Data Min.*, vol. 1, no. 2, pp. 790–792, 2009.

[9] F. Yuan, Z. Meng, H. Zhang, and C. Dong. " A new algorithm to get the initial centroids. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on* (Vol. 2, pp. 1191-1193). IEEE, 2004.

[10] R. Elmasri, and S. Navathe, (2015). *Fundamentals of Database Systems*, Pearson Education, 2008.

[11] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881-892, Jul 2002.

[12] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java",*Cs.waikato.ac.nz*, 2016. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka. [Accessed: 02- Oct- 2016].

[13] B. Chaudhari and M. Parikh, "A Comparative Study of clustering algorithms Using weka tools," *Inf. Technol. J.*, vol. 1, no. 2, pp. 154–158, 2012.

[14] M. Maurizio, "Data Mining Concepts and Techniques," 2011.

[15] P. Shi, "An Efficient Approach for Clustering Web Access Patterns from Web Logs," *Int. J. Adv. Sci. Technol.*, vol. 5, pp. 1–14, 2009.

[16] T. A. Runkler and J. C. Bezdek, "Web mining with relational clustering," *Int. J. Approx. Reason.*, vol. 32, no. 2–3, pp. 217–236, 2003.

[17] P. Lingras and C. West, "Interval set clustering of web users with rough kmeans, "*Journal of Intelligent Information Systems*, vol. 23, no. 1, pp. 5–16, Jul. 2004.

[18] V. Chitraa and A. S. Thanamani, "Web Log Data Analysis by Enhanced Fuzzy C Means Clustering," *Int. J. Comput. Sci. Appl.*, vol. 4, no. 2, pp. 81–95, 2014.

[19] D. Dong, "Exploration on Web Usage Mining and its Application," *Intelligent Systems and Applications, 2009. ISA 2009. International Workshop on*, Wuhan, 2009, pp. 1-4.

[20] K.Kameshwaran and K.Malarvizhi, "Survey on Clustering Techniques in Data Mining," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 2272–2276, 2014.