# Web Benefit Utilizations with K-means Clustering Approach for Efficient Clustering

Priya B. Pandharbale[1], Sasmita Choudhury[2], Sachi Nandan Mohanty[3], Alok Kumar Jagadev[1]

*1 School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Orrisa, India.*
*2 Department of Computer science Engineering, Mckv Institute of Engineering, Liluah, Howrah, West Bengal, India.*
*3 Vardhaman College of Engineering, Hyderabad, India.*

**Abstract**

Clustering is the process of identifying similar groups in a dataset based on some characteristics of the data. This work uses the k-means clustering algorithm for finding the numerous cluster formations of various parameters in the weblog dataset. The clusters are formed and are examined for finding the various status responses generated while accessing the web data as well as the popular methods the users are using for accessing the web. The work concentrates on the optimal k value finding using the Elbow method showing the formation of the number of clusters as the value of k varies.

**Keywords:** k-Means, clustering, web service, weblog, access methods

## 1. Introduction

Clustering is essentially depicted as a division of information into bunches of identical articles. Each cluster includes objects that are comparable among themselves and various checked out of various packs. We should contemplate among various sorts of packs. The assessments under talk about are: k-means clustering, distinctive leveled out gathering assessment, self-masterminding maps assessment, and need expansion bundling computation. Assessment Metrics are selected like calculations, dataset size, programming utilized, execution, precision, and nature of calculations [7].

In this work, we are advancing to zero in on the quality and the execution of the web information. The irksome web URLs are the filling the job in off-track check of the data. The superfluous Data is causing inconveniences in page arranging. The work centers on finding the practical reactions from the net laborers by purging the immaterial information from the net log dataset. The k-means calculation could be an extraordinarily normal assessment differentiated and the wide scope of different clustering assessments to the extent the time complexities similarly as information preparing. The work focuses on the formation of various clusters of web information depending on various parameters like web information access date, the status of the web service, various access methods which can be utilized by the maximum users, etc. The work finds the optimal value of k for the k-means clustering approach applied for various web information parameters clusters.



ACI'22: Workshop on Advances in Computation Intelligence, its Concepts & Applications at ISIC 2022, May 17-19, Savannah, United States
EMAIL: priyasathe123@gmail.com (A. 1)

## 2. Literature Survey

Clustering is the process of identifying similar groups in a dataset based on some characteristics of the data. In clustering, no class information is needed. Hence it is an unsupervised learning technique. It has many applications like text clustering. It is generally divided into two categories: hierarchical and partitioning. Partitioned clustering algorithms are suitable for clustering large datasets.
The creators attempted to apply the k-Means bunching technique from the corn crop information of the most recent 2 years to deliver achievability data from each sub-district [1]. The conveyance of harvests is typically done dependent on the name of the corn-creating sub-district. A gathering of potential corn-delivering locales is needed to know which regions produce huge or modest quantities of corn.
The paper proposes a boundary profile-based gradual grouping (BPIC) technique to find self-assertively molded bunches with powerfully developing datasets [2]. This technique addresses the current bunching results with an assortment of limit profiles and disposes of the internal places of groups as opposed to keeping all information.

The work showed another social occasion approach named CluStream [3]. It had a web part that inconsistently put missing incorporate summary pieces of information and a disconnected piece that used these assessments. The internet-based part was the quantifiable information assortment piece and the disconnected part was the legitimate area. The CluStream can deal along arising and evaporating packs anyway can't administer changing information things and their portrayal.

D-stream gathering approach used thickness-based systems [4,8]. This had an on the web and disconnected section. The web-based part maps every data information thing into a structure and a disconnected area which shapes the framework thickness. The exceptional changes of the information stream were overseen using a rotting technique. It also perceived the inconsistent organizations organized through the exclusions. It will in general be used for social event constant flow information. The advantages of this procedure are that it can productively make packs progressively, can track down lots of emotional shapes, and can unequivocally perceive the creating sharpens of nonstop information streams.

Authors have characterized an entropy-based objective capacity for the instatement interaction, which is superior to other existing introduction techniques for k-implies grouping. Additionally planned a calculation to ascertain the right number of bunches of datasets utilizing some group legitimacy records [5].
The calculation uses Fair-Lloyd, a change of Lloyd's heuristic for k-implies, acquiring its straightforwardness, proficiency, and solidness. Fair-Lloyd displays fair-minded execution by guaranteeing that all gatherings have equivalent expenses in the result k-grouping, while at the same time bringing about an irrelevant expansion in running time, accordingly making it a reasonable choice any place k-implies is as of now utilized [6].
A variety of k-implies grouping called round k-means bunch for report bunching [7]. It partitioned the tall dimensional unit circle through infers of social affair of great hyper circles. The estimation played out a disjoint allocating of the document vectors, and, for each package, figured a centroid using cosine resemblance. The standardized centroid was called 'idea vectors' which contain significant semantic data around bunches. The most benefit of this computation is that it meets quickly and it can deal with the sparsity of content data. Moreover, it tends to be parallelized quickly.

This article endeavors to foster a numerical model for designating the assignments to the processors to accomplish the ideal expense and ideal unwavering quality of the framework [9].

The author has introduced the review on different grouping techniques in their work [10]. Table 1 shows the introduced review for different grouping calculations by thinking about the boundaries classification, bunching calculations overviewed, and their time intricacies. Creator guarantees that K-means give a higher outcome for gigantic information than SOM and progressive grouping calculation.
Our previous works in the area of web services clustering help find better recommendations using k-means clustering [12-15].

The work deals with effective bunching strategies, for example, K-implies grouping, Hierarchical agglomerative bunching, and Balanced Iterative Reducing and Clustering utilizing Hierarchies (BIRCH) bunching are presented for web administration bunching [16].

A K -means sort of clustering to be specific Pioneer Supporter calculation is utilized here [17,18]. In this approach for an unused thing ‘i', a closest cluster middle 'c' is recognized. In the event that separates between things 'I' and cluster middle is over the edge, at that point a modern cluster is made. Something else the information thing is included to the cluster spoken to through 'c'. Rehash this handle until there are no more information things.

ICECPG clustering using extended condensation point and grid clustering algorithm which was based on fast density-based clustering techniques This algorithm used a heuristic search method to form sub-clusters. A cluster is formed by uniting all the sub-clusters reachable from one another. A steady grouping utilizing expanded build-up point and lattice for continuous bunching of dynamic information approach [4]. As the new information showed up, it was appointed to existing groups. This calculation catches the state of the information base through expanded build-up focuses. Then, at that point, for bunching the information things, it utilized a network-based and thickness-based grouping approach that utilizes slope-based climbing ideas. This strategy enjoys the benefit of thickness-based and matrix-based strategies. It has straight time intricacy and can be utilized for mining huge datasets. It decreases I/O costs.

A couple of utilization of stream bunching is interference affirmation, environment insights, E-business, crisis counter structures [19], site assessment, etc. In-stream grouping each exceptional data thing is considered as the advanced info data thing. Stream grouping approaches don't deal with lively data since they don't store the data. Gradual grouping doesn't deal with the time of unused bunches and updating a group for a thing that changes over time. Both gradual and stream grouping approaches are less sensible for enthusiastic applications like the Web. In Web-based applications, features of a data thing might modify quite a while since of an adjust inside the preferences and loathe of end clients. Also on the net, dealing with creating and evaporating groups is furthermore indispensable. To gain ground on the nature of electronic applications grouping strategies used should have the option to deal with enthusiastic circumstances. The survey of various clustering algorithms for finding out the complexities is discussed in [3].

## 3. Methodology

The web log data is pre-processed. The data set used here is available at https://www.kaggle.com/shawon10/web-log-dataset. The work focuses on the step-by-step analysis of the weblog data to find the clusters. The work uses k-means clustering [11] for the creation of the initial cluster’s formation CF1 using the User data U and the most frequently accessed URL's FA. The website utilization information parameters like date D and status S are used to form CF1. The status parameter used for the HyperText Transfer Protocol (HTTP) are identified as 400 is used to indicate a Bad Request

reaction status code it shows that the server can't or won't handle the solicitation because of something saw to be a customer mistake (e.g., contorted solicitation language structure, invalid solicitation message outlining, or beguiling solicitation directing).

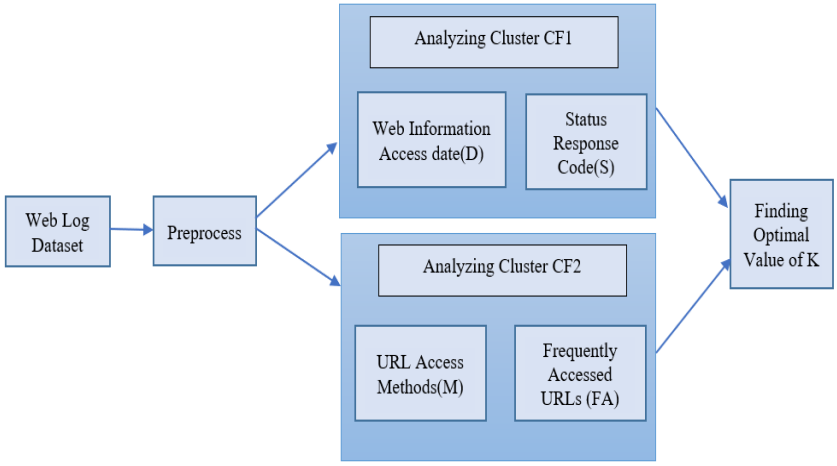


**Figure 1**: Architecture Diagram for Clustering Data on Various Parameters in Weblog Dataset and Finding Optimal Value of k.

The HTTP 300 Multiple Choices divert status reaction code shows that the solicitation has more than one potential reaction. The client specialist or the client ought to pick one of them. As there is no normalized method of picking one of the reactions, this reaction code is seldom utilized. The HTTP 200 OK accomplishment status response code shows that the sales have succeeded. A 200 response is cacheable as is normally done.

Algorithm 1:

K-Means Clustering: URL Analysis for Status Response Code

1. Input: N number of records from dataset S.
2. For each user U finds the most frequently
   accessed URLs FA.
3. cluster formation, CF1 using website utilization information date D and status S.
4. End

Algorithm 2:

K-Means Clustering: User Web URL Access Method Analysis

1. Input: N number of records from dataset S.
2. for each user web URL WU find the access
   method M
3. cluster formation, CF2 using FA and M
4. End

Reapplying the bunching calculation over the cluster formation CF1 in the boundaries for making new bunches CF2 is the client web access method M and the FA. Among the Web URL access techniques M, the GET and Post strategies are the most famous techniques utilized. The GET system requests a depiction of the predefined resource. Requesting using GET should simply recuperate data. The POST strategy is used to introduce a substance to the foreordained resource, as often as possible causing a change of state or accidental impacts on the server.

## 4. Results and Discussion

The data set used here is available at https://www.kaggle.com/shawon10/web-log-dataset.The work focuses on the step-by-step analysis of the weblog data to find the clusters for the status response code of the web services and the web URL access methods are mostly used by the users. This dataset has 16008 rows and 4 columns. Columns are IP, Time, URL, Response Status.

| | IP | Time | URL | Status |
|---|---|---|---|---|
| **0** | 10.128.2.1 | [29/Nov/2017:06:58:55 | GET /login.php HTTP/1.1 | 200 |
| **1** | 10.128.2.1 | [29/Nov/2017:06:59:02 | POST /process.php HTTP/1.1 | 302 |
| **2** | 10.128.2.1 | [29/Nov/2017:06:59:03 | GET /home.php HTTP/1.1 | 200 |
| **3** | 10.131.2.1 | [29/Nov/2017:06:59:04 | GET /js/vendor/moment.min.js HTTP/1.1 | 200 |
| **4** | 10.130.2.1 | [29/Nov/2017:06:59:06 | GET /bootstrap-3.3.7/js/bootstrap.js HTTP/1.1 | 200 |
| **...** | ... | ... | ... | ... |
| **16002** | 10.130.2.1 | [02/Mar/2018:15:47:12 | GET /showcode.php?id=309&nm=ham05 HTTP/1.1 | 200 |
| **16003** | 10.130.2.1 | [02/Mar/2018:15:47:23 | GET /allsubmission.php HTTP/1.1 | 200 |
| **16004** | 10.130.2.1 | [02/Mar/2018:15:47:32 | GET /showcode.php?id=309&nm=ham05 HTTP/1.1 | 200 |
| **16005** | 10.130.2.1 | [02/Mar/2018:15:47:35 | GET /allsubmission.php HTTP/1.1 | 200 |
| **16006** | 10.130.2.1 | [02/Mar/2018:15:47:46 | GET /home.php HTTP/1.1 | 200 |

16007 rows × 4 columns

**Figure 2:** An Example of the Information Extraction for the Status Response Code of the Web Services

Figure 2 shows the information extraction for the status response code of the web services from the weblog dataset.

Figure 3 shows the plot of the web URLs frequently utilized by many customers. It is observed that the customers like to visit some web URLs frequently making them their favorite websites based on the frequency of accessing the URL.

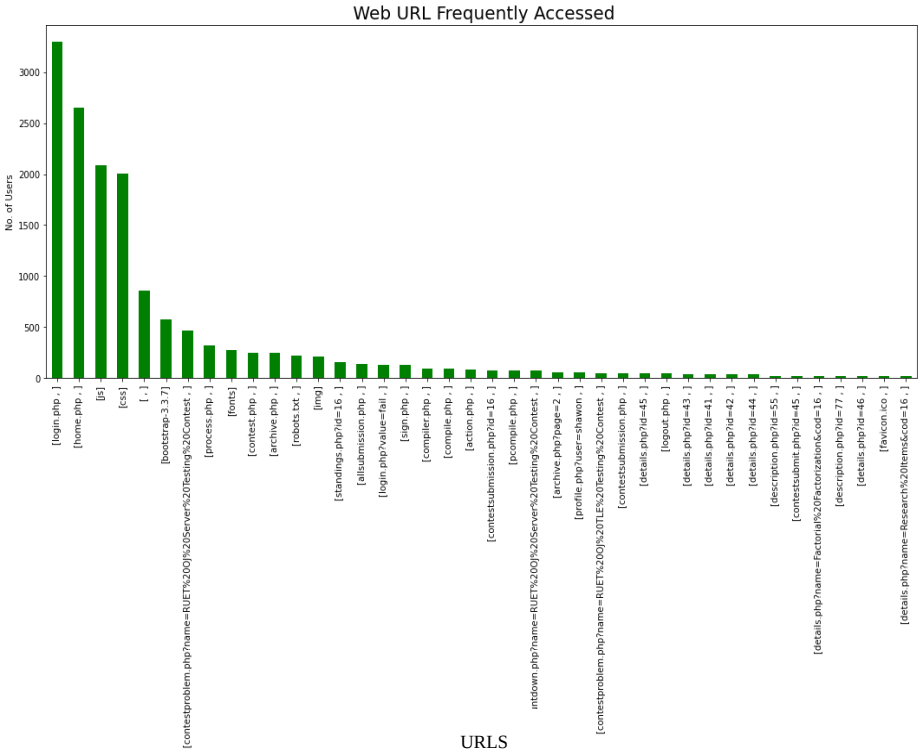


**Figure 3**: Frequently Accessed Web URLs

In figure 4 we can find the metrics for the calculation of the mean values for the creation of the initial clusters. As depicted in the methodology section the web URLs are clustered using the criteria status response code.

| | Unnamed: 0 | Status | Day |
|---|---|---|---|
| count | 15789.000000 | 15789.000000 | 15789.000000 |
| mean | 7974.763506 | 230.194693 | 23.209703 |
| std | 4622.058252 | 50.058535 | 8.346775 |
| min | 0.000000 | 200.000000 | 1.000000 |
| 25% | 3962.000000 | 200.000000 | 17.000000 |
| 50% | 7950.000000 | 200.000000 | 29.000000 |
| 75% | 11989.000000 | 302.000000 | 29.000000 |
| max | 16006.000000 | 404.000000 | 30.000000 |

**Figure 4**: Calculation of the Mean Values for the Creation of the Initial Clusters

Figure 5 shows the optimal value for k here is 4. Hence, we can observe the four clusters are formed for the status response code for various status responses.

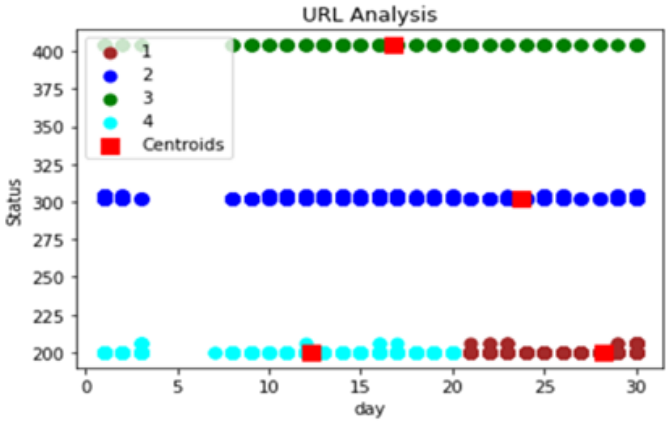


**Figure 5**: URL Analysis for Status Response Code

According to figure 6, the analysis of weblog data shows that among the Web URL access techniques the GET and Post strategies are the most famous techniques utilized by the customers. The access methods popular amongst all the other access methods are GET and POST.

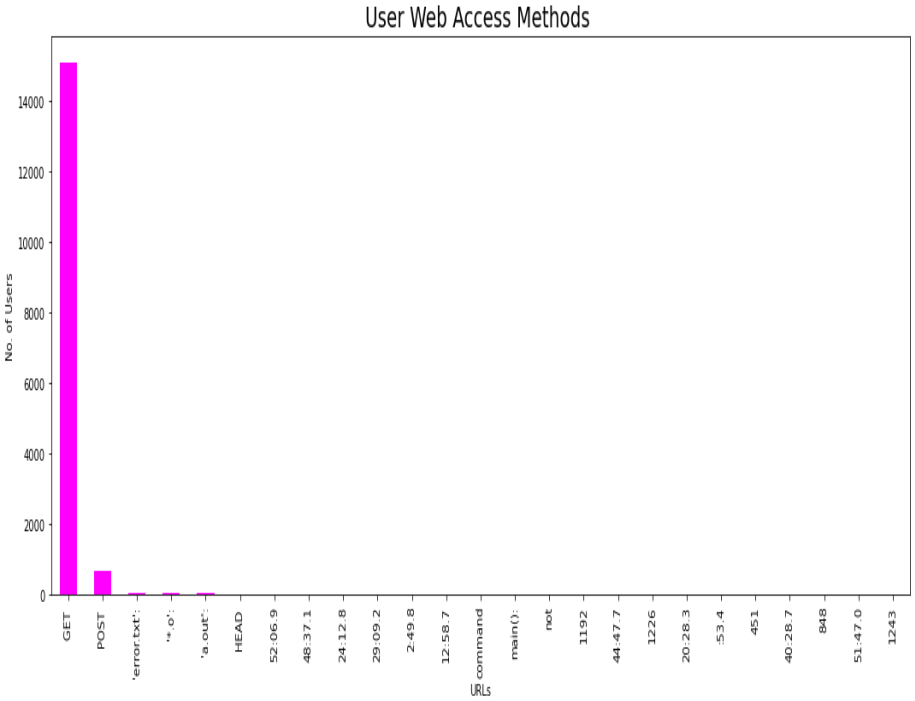


**Figure 6**: Analysis of the Web Access Methods

From figure 6 it is observed that these methods are mostly used by the customers for the invocation of the URLs. On applying the k-means clustering for the web URL access methods the optimal value for k=2. The clusters formed for the most popular Web URL access methods have two clusters.
Figure 7 shows the selection of the value of k as 2 using the Elbow method it is very easy to predict the optimal value of k at an elbow point in the graph.

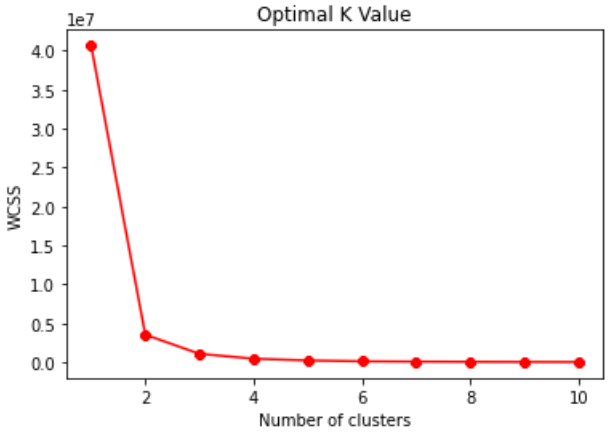


**Figure 7**: Finding the Optimal k Value

Figure 8 shows the clustering of the Web URL access methods The web server processes the data and communicates a HTTP status code. Should the solicitation find success, the server sends an information bundle to the internet browser with all the data expected for the page.

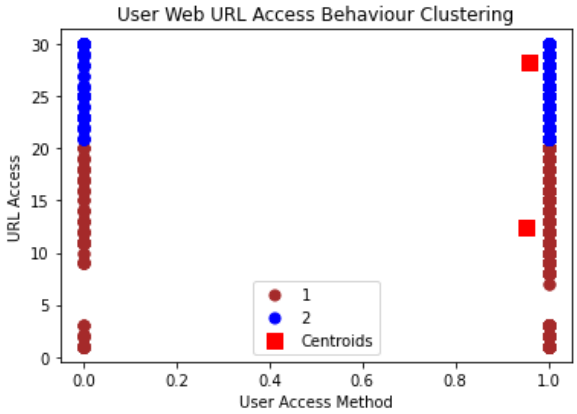


**Figure 8**: User Web URL Access Method Analysis

| | Sr.No | IP | Time | URL | Status | Month | day | Methods | URL_new |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 479 | 10.129.2.1 | 29-Nov-17 | POST /pcompile.php HTTP/1.1 | 200 | Nov | 29 | 0 | ['pcompile.php ', ''] |
| **1** | 480 | 10.131.0.1 | 29-Nov-17 | POST /allsubmission.php HTTP/1.1 | 200 | Nov | 29 | 0 | ['allsubmission.php ', ''] |
| **2** | 949 | 10.129.2.1 | 30-Nov-17 | POST /pcompile.php HTTP/1.1 | 200 | Nov | 30 | 0 | ['pcompile.php ', ''] |
| **3** | 950 | 10.129.2.1 | 30-Nov-17 | POST /allsubmission.php HTTP/1.1 | 200 | Nov | 30 | 0 | ['allsubmission.php ', ''] |
| **4** | 955 | 10.130.2.1 | 30-Nov-17 | POST /pcompile.php HTTP/1.1 | 200 | Nov | 30 | 0 | ['pcompile.php ', ''] |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **15775** | 15803 | 10.130.2.1 | 28-Feb-18 | GET /favicon.ico HTTP/1.1 | 404 | Feb | 28 | 1 | ['favicon.ico ', ''] |
| **15776** | 15820 | 10.128.2.1 | 01-Mar-18 | GET /robots.txt HTTP/1.1 | 404 | Mar | 1 | 1 | ['robots.txt ', ''] |
| **15777** | 15833 | 10.130.2.1 | 01-Mar-18 | GET /robots.txt HTTP/1.1 | 404 | Mar | 1 | 1 | ['robots.txt ', ''] |
| **15778** | 15965 | 10.128.2.1 | 02-Mar-18 | GET /robots.txt HTTP/1.1 | 404 | Mar | 2 | 1 | ['robots.txt ', ''] |
| **15779** | 15968 | 10.130.2.1 | 02-Mar-18 | GET /robots.txt HTTP/1.1 | 404 | Mar | 2 | 1 | ['robots.txt ', ''] |

15780 rows × 9 columns

**Figure 9**: An Example of Clustering Using the Web URL Access Methods.

In the event that the server can't observe the page at the mentioned address, it either sends a 404-blunder code (site page not found) or sends the guest to the new URL through divert assuming it's known. In figure 9 the example for the clustering of the web URLs is shown for the cluster formation for methods GET (0) and POST (1).

## 5. Conclusion

In this work, we have discussed various clustering techniques used efficiently for the analysis of the data and removing the barriers to accessing the huge datasets. Moreover, this work helps to elaborate k-Means clustering over the weblog dataset to analyze and utilize the weblog dataset efficiently. The algorithm utilizes various parameters of the weblog dataset for the formation of various clusters. The Elbow method is then used to find the optimal value of the k in k-means to predict the number of clusters formed for the given dataset parameters. The optimal value of k is 4 for the status response code for various status responses. Whereas the value of k=2 for the most popular methods to access the web that is GET and POST. For the future work we will be using the various width clustering algorithm for the calculation of the distance for finding the optimal value of k.

## References


[1] Aldino, A. A., et al. "Implementation of K-means algorithm for clustering corn planting feasibility area in south lampung regency." Journal of Physics: Conference Series. Vol. 1751. No. 1. IOP Publishing, 2021.

[2] Bao, Junpeng, et al. "An incremental clustering method based on the boundary profile." Plos one 13.4 (2018): e0196108.

[3] Benabdellah, Abla Chouni, Asmaa Benghabrit, and Imane Bouhaddou. "A survey of clustering algorithms for an industrial context." Procedia computer science 148 (2019): 291-302.

[4] Zhuo, Chen, Liu Xiang-shuang, and Zhuang Xiao-dong. "A fast incremental clustering algorithm based on grid and density." Third International Conference on Natural Computation (ICNC 2007). Vol. 5. IEEE, 2007.

[5] Chowdhury, Kuntal, Debasis Chaudhuri, and Arup Kumar Pal. "An entropy-based initialization method of K-means clustering on the optimal number of clusters." Neural Computing and Applications 33.12 (2021): 6965-6982.

[6] Ghadiri, Mehrdad, Samira Samadi, and Santosh Vempala. "Socially fair k-means clustering." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

[7] https://medium.com/analytics-vidhya/comparative-study-of-the-clustering-algorithms-54d1ed9ea732.

[8] Khalilian, Madjid, Norwati Mustapha, and Nasir Sulaiman. "Data stream clustering by divide and conquer approach based on vector model." Journal of Big Data 3.1 (2016): 1-21.

[9] Kumar, Harendra, Nutan Kumari Chauhan, and Pradeep Kumar Yadav. "A high performance model for task allocation in distributed computing system using k-means clustering technique." Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing. IGI Global, 2021. 1244-1268.

[10] Li, Wei, et al. "Data Stream Clustering Algorithm for Smart Site and Its Implementation Based on Flink." 2019 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2019.

[11] MacQueen, J. "Classification and analysis of multivariate observations." 5th Berkeley Symp. Math. Statist. Probability. 1967.

[12] M. P. B. P. M. S. M. B. P. Semantic Search and Social-Semantic Search as Cooperative Approach. International Journal on Recent and Innovation Trends in Computing and Communication, 5(1), 110 - 114. https://doi.org/10.17762/ijritcc.v5i1.98.

[13] Pandharbale, Priya B., Sachi Nandan Mohanty, and Alok Kumar Jagadev. "Recent web service recommendation methods: A review." Materials Today: Proceedings (2021).

[14] Pandharbale, Priya, Sachi Nandan Mohanty, and Alok Kumar Jagadev. "Study of Recent Web Service Recommendation Methods." 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2020.

[15] Pandharbale, Priya Bhaskar, Sachi Nandan Mohanty, and Alok Kumar Jagadev. "Novel Clustering-Based Web Service Recommendation Framework." International Journal of System Dynamics Applications (IJSDA) 11.5 (2021): 1-15.

[16] Parimalam, T., and K. Meenakshi Sundaram. "Efficient clustering techniques for web services clustering." 2017 ieee international conference on computational intelligence and computing research (iccic). IEEE, 2017.

[17] Reyes, Jaciel E., et al. "A Classification of Web Service Credibility Measures." 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2021.

[18] Sardar, Tanvir Habib, and Zahid Ansari. "An analysis of distributed document clustering using MapReduce based K-means algorithm." Journal of The Institution of Engineers (India): Series B 101.6 (2020): 641-650.

[19] Yeoh, Jia Ming, et al. "A clustering system for dynamic data streams based on meta heuristic optimisation." Mathematics 7.12 (2019): 1229.