ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(24), DOI: 10.17485/ijst/2015/v8i24/80192, September 2015* ISSN (Online) : 0974-5645

Enhanced K-Means Clustering Algorithm for Evolving User Groups

K. Selvakumar[1*], L. Sai Ramesh[2] and A. Kannan[2]

[1]School of Computing Science and Engineering, VIT University, Vellore - 632014, Tamil Nadu, India;
selvaa21@gmail.com
[2]Department of Information Science and Technology, CEG, Anna University, Chennai - 600025, Tamil Nadu, India;
sairamesh.ist@gmail.com, kannan@annauniv.edu

Abstract

To gain information about user interests in Web pages is needed to advance in Web security. An approach to pick up that information includes understanding the client's perusing conduct, examining the Web log records with the procedures of preprocessing and client clustering. Time spent on Web pages and the types of operations show the degree of a Web user's intension. The data set comprises of Web log files obtained by collecting the user logs during a six month period. A new enhanced K-means clustering algorithm proposed in this paper for grouping user based on their preferred Web content and their temporal constraints. The enhanced K-mean clustering calculates initial centroids instead of random choice and uses time intervals to heighten the security and performance. Utilizing this methodology, client access designs with comparable looking practices are assembled into a particular class amid a particular time interval. Also secured communication among the various users groups will be achieved through hill cipher technique.

Keywords: Preprocessing, Security and Hill Cipher, Temporal K-Means Algorithm, Web User Categorization

1. Introduction

A basic drawback that always arises during a nice type of fields like data processing, data revelation and example order is understood because the agglomeration drawback[1]. The significance of Web mining is expanding faster within the most recent decade and as of late once there's terribly robust powerful rivalry within the markets wherever the standard of knowledge and on time net information plays an awfully necessary role in deciding. This has attracted plenty of attention in varied organizations and data business. There's vast quantity of real Internet knowledge offered within the world and it's quite tough to access the helpful details from this vast information and supply the knowledge to users among the limit and in needed pattern. Web data processing is that the solely methodology of excavation the Web logs and finding the reasonable patterns. While not Web mining, it's not possible to look at such giant databases and to supply valuable information. However, mining the user logs is helpful to produce security.

Clustering could be a natural approach to cluster similar objects supported some similar properties, that is named similarity measures. The weather among a cluster are comparatively additional almost like one another as a result of they need similar properties or attributes. Clustering is one in every of the foremost outstanding data processing techniques that are used for varied applications like pattern discovery; knowledge analysis; prediction; visualization and personalization. The methods are often performed during a superintended, semi-superintended or unattended form[2]. Completely dissimilar procedures have thought of within the past that have taken into consideration the character of the information and therefore the input arguments so as to cluster the information. The most recent algorithms consider the quantity of clusters (K) as associate degree input that is fastened. If the fastened variety of cluster is extremely tiny then there's an opportunity that dissimilar objects are placed on that cluster and guess the quantity of fastened cluster is giant then the additional similar objects are placed into completely different teams.

** Author for correspondence*

In this research article, a replacement temporal clustering algorithmic rule known as enhanced K-means clustering algorithmic rule has been projected for effective and dynamic grouping of Web users. This projected algorithmic rule uses temporal information additionally as hill cipher to reinforce the security. The rest of the article is formed as follows. Section 2 focuses on the literature survey within the connected field. Section 3 highlights the projected methodology. Section 4 confronts the discourse. Finally, Section 5 concludes the paper with future work.

2. Literature Work

Clustering is a vital errand of information mining. Numerous methods have been utilized for grouping. Past works around there are condensed in this area. K-means bunch estimation was proposed by J. B. MacQueen in 1967, which is used to deal with the issue of data gathering, the figuring is by and large essential that it had wide effect in the exploratory field research and mechanical applications[3,4]. Wang et al.[5] received a novel system to separate the clients' enthusiasm, including Long Term Interests and Short Term Interests. In 2008, Zhijiang Wang et al. utilized a novel grouping calculation in view of a postfix tree for bunching inquiries as indicated by client logs. What's more, a magic word based comparability operation was embraced for registering the closeness between the info question and groups and the Chinese semantic connection is furthermore considered inside of the similitude computation[6]. The work[7] proposes a substance based proposal framework that uses the helpful separating procedure and positioned k-means bunching algorithmic standard to give the client to-client suggestions and then recommend new information to improve the community oriented sifting. Miao wan et al. grouped Web clients' upheld their entrance designs. This methodology may catch regular client intrigues and manufacture client profiles for cutting edge net applications, similar to Web storing and pre-getting[8]. Indrajit Mukherjee et al.[9] arranged partner degree way to deal with group clients upheld their Web use designs. They break down Web logs abuse information preparing and go up against clients with extra altered Web data. The Web bunching methods of Affinity Propagation (AP) and Streaming Affinity Propagation (StrAP)[10] were connected to group Web clients in view of the perusing practices.

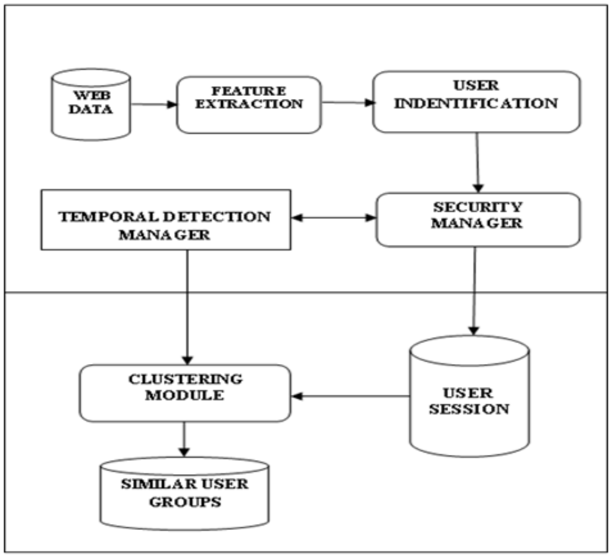


Figure 1. System architecture.

The work in[11] proposes a way to deal with Cluster Web clients in view of route examples of clients. The produced groups can catch clients with comparative hobbies with a superior precision. Cao Yi et al.[12] outlined a framework to cluster the clients by their intrigues and proposed the investigation of the clients' mining so as to gather premiums the Web skimming history. In[13], creators proposed a framework to group clients as far as their skimming hobbies on the specific page is distinguished as intrigued page and clustering clients from Web logs helps to accomplish suggestions and to customize administrations. In 2010, Jin Hua Xu et al.[14] grouped the Web clients in view of conduct trademark and they give valuable learning to customized Web administrations.

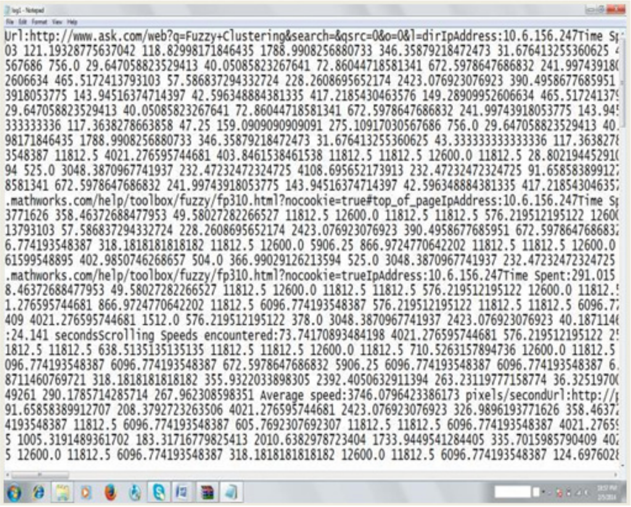
The studies gave in[17] utilized a programmed client division upheld the comparable client conduct on the online site. The client groups and their picture were discovered upheld the similitude. An unpleasant set bundle recipe[18] was wont to group the fundamental sessions upheld the most extreme pages went by. By narrowing the Weblog, it diminishes the nature of the PPE (Prediction Pre-bringing Engine-living on intermediary server). A fascinating methodology of fluffy unpleasant estimate[19] clusters the client access designs with comparative surf riding practices. Ling winged creature class et al.[20] arranged a COWES structure to group net clients by examining basic and organic procedure examples shared by clients. Amid this arranged work, the crawler data is utilized for developing client group's abuse expanded K-implies equation that is presented inside of the accompanying area. The most commitment of this paper is to deliver security through gathering bolstered hobbies and transient requirements. The insurance is expanded any abuse keys[15,16] and hill cipher.

3. System Architecture

In the proposed system, the pre-processing of Web log files is done to obtain the user sessions. The user sessions are then clustered using enhanced K-means algorithm to group the users into similar groups or categories. The designed approach contains the two phases: Data Pre-processing and Web user clustering.

3.1 Data Pre-Processing

The stages of pre-processing, shown in Figure 1, which include data cleansing, user identification as well as their session identification.



```
Url:http://www.ask.com/web?q=Fuzzy+Clustering&search=&qsrc=0&o=0&l=dirIpAddress:10.6.156.247Time S
03 121.19328775637042 118.82998171846435 1788.9908256880733 346.35879218472473 31.676413255360625
567686 756.0 29.647058823529413 40.05085823267641 72.86044718581341 672.5978647686832 241.99743918
2606634 465.5172413793103 57.586837294332724 228.2608695652174 2423.076923076923 390.4958677685951
3918053775 143.94516374714397 42.596348884381335 417.2185430463576 149.2890995260634 465.51724137
29.647058823529413 40.05085823267641 72.86044718581341 672.5978647686832 241.99743918053775 143.94
333333336 117.3638278663858 47.25 159.0909090909091 275.10917030567686 756.0 29.647058823529413 40.
98171846435 1788.9908256880733 346.35879218472473 31.676413255360625 43.333333333333336 117.363827
3548387 11812.5 4021.276595744681 403.8461538461538 11812.5 11812.5 12600.0 11812.5 28.80219445291
94 525.0 3048.3870967741937 232.47232472324725 4108.695652173913 232.47232472324725 91.65858389912
8581341 672.5978647686832 241.99743918053775 143.94516374714397 42.596348884381335 417.2185430463
.mathworks.com/help/toolbox/fuzzy/fp310.html?nocookie=true#top_of_pageIpAddress:10.6.156.247Time S
3771626 358.46372688477953 49.58027282266527 11812.5 12600.0 11812.5 11812.5 576.219512195122 1260
13793103 57.586837294332724 228.2608695652174 2423.076923076923 390.4958677685951 672.59786476868
6.774193548387 318.1818181818182 11812.5 12600.0 5906.25 866.9724770642202 11812.5 11812.5 12600.0
61599548895 402.9850746268657 504.0 366.99029126213594 525.0 3048.3870967741937 232.47232472324725
.mathworks.com/help/toolbox/fuzzy/fp310.html?nocookie=trueIpAddress:10.6.156.247Time Spent:291.015
8.46372688477953 49.58027282266527 11812.5 12600.0 11812.5 11812.5 576.219512195122 12600.0 11812.
1.276595744681 866.9724770642202 11812.5 6096.774193548387 576.219512195122 11812.5 11812.5 6096.7
409 4021.276595744681 1512.0 576.219512195122 378.0 3048.3870967741937 2423.076923076923 40.187114
:24.141 secondsScrolling Speeds encountered:73.74170893484198 4021.276595744681 576.219512195122 2
1812.5 11812.5 638.5135135135135 11812.5 11812.5 12600.0 11812.5 710.5263157894736 12600.0 11812.5
096.774193548387 6096.774193548387 672.5978647686832 5906.25 6096.774193548387 6096.774193548387 6.
8711460769721 318.1818181818182 355.9322033898305 2392.4050632911394 263.2311977158774 36.325197
49261 290.1785714285714 267.962308598351 Average speed:3746.0796423386173 pixels/secondUrl:http://
91.65858389912707 208.3792723263506 4021.276595744681 2423.076923076923 326.9896193771626 358.4637
4193548387 11812.5 6096.774193548387 605.7692307692307 11812.5 11812.5 6096.774193548387 4021.2765
5 1005.3191489361702 183.31716779825413 2010.6382978723404 1733.9449541284405 335.7015985790409 40
5 12600.0 11812.5 6096.774193548387 318.1818181818182 12600.0 11812.5 6096.774193548387 124.697602
```

Figure 2. Before preprocessing.

3.1.1 Web Data

When the user accesses a Web page, the user will perform a series of clicks in that particular page. This information is collected for six months period which is stored as text documents with 50 users. Each access log entry consisted of: URL (Uniform Resource Locater), IP Address, Time Spent (Seconds), Scrolling-Speeds encountered and Average Speed (Pixels/Second) as shown in Figure 2. Web data is the input to the Data cleaning phase.

3.1.2 Data Cleaning

Superfluous records are dropped amid information cleaning. As the objective of Web Usage Mining (WUM) is to accomplish a traversal example, taking after two sorts of records is not required and ought to be taken out[22].

3.1.2.1 The Records of Illustrations, Feature and Organization Data

The records with filenames additions of GIF, JPEG, CSS thus on can be discovered and these fields ought to be wiped out.

3.1.2.2 The Records with Fizzled HTTP Status Codes

By looking at the status field of each record in the crawler information set, the record with status code more than 299 and less than 200 are took out.

3.1.3 Client Recognition

The point of the client recognizable proof procedure is to figure out diverse clients from the client log data. Clients are recognized by their Internet Protocol (IP) addresses. The technique utilized for this procedure is a referrer-based system. Client recognizable proof is troublesome because of the vicinity of inhabitant stores, firewalls and intermediary servers. To arrangement this issue, the WUM strategies were utilized that depend on client collaboration. On the other hand, it's troublesome as a result of high security and protection. The accompanying heuristics were utilized as a part of testing the proposed procedure to distinguish the client: For each IP location speaks to one client;

- If the IP address is same for more logs, but the agent log displays a change in Web browser or Operating systems, the IP location represents some other user.
- If there is a same IP address, browser and operating system, the referrer information can be considered. If a user requested page is indirectly accessible by a link from any of these pages, hence with the same IP there is different user[23].

The clients are related to the assistance of IP location. Since the IP locations of clients are same, the clients' searching conduct was utilized to recognize the client. The client conduct was recognized by utilizing the URL/question. For this reason, the title tag of every page went by the client was extricated from the entrance way. Title labels of all pages went by the 50 clients were put away in the Titlelog.txt record.

3.1.4 Session Identification

Azclient session is frequently delineated as a gathering of pages went by indistinguishable client at interims the length of one particular visit to a Webpage[21]. The objective of session recognizable proof is to separate the page solicitations of each client into individual sessions i.e. each client's entrance design and continuous way is found. The most straightforward philosophy to detect a session is utilizing a timeout component. The timeout component was utilized for each and every individual client distinguishing proof. For this reason, the aggregate time spent of each client is refinement in begin time and end-time of client session. In the event that a client has invested most energy in a page then it ought to be accepted that they has ton of enthusiasm there in page.

The tenet concerning session recognizable proof is that the accompanying: 1. If there's a substitution client, then there's a substitution session. 2. In one client session, if the asked for page is invalid, then it's finished that there's a substitution session. 3. If the time between page solicitations surpasses a specific point of confinement, it's expected that the client is starting another session. 4. If asked for pages are frequently available through a connection from any of the chosen pages by the client, we tend to accept that it should be another inside of the session[24]. Once there two or more extra pages that have a superb connection to that in one session, then it should be set before the as of late gone to page.

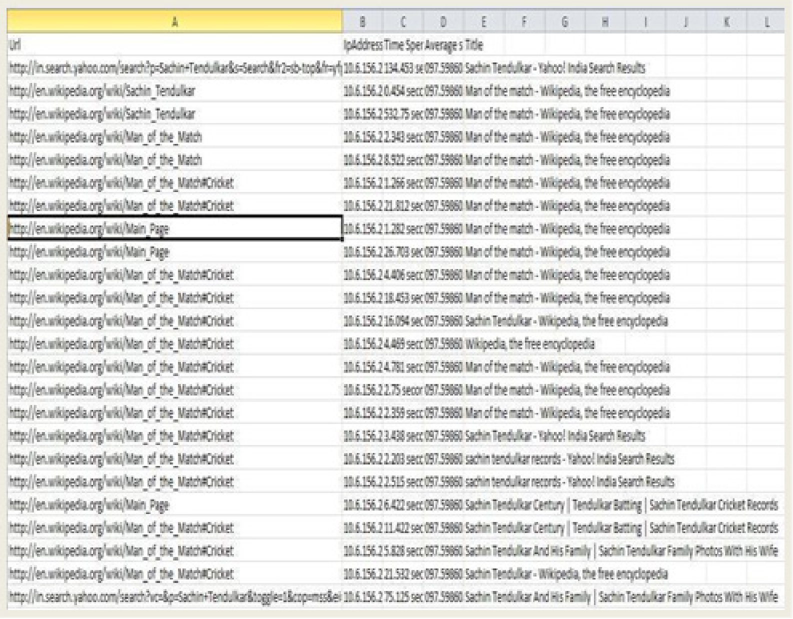
For the session distinguishing proof calculation, the accompanying log record traits are chosen: IP Address, Time Spent, URL got to and Category. For the customer side log document, the aggregate time spent for every last client is computed from the time invested and this aggregate energy spent is utilized to ascertain the session. Here 20 minutes timeout is taken as a default session timeout. Figure 3 demonstrates the pre-handled document containing just important data.

3.2 Clustering

Web User clustering, shown in Figure 1, involves clustering of user sessions obtained from pre-processing into similar user groups using enhanced K-means procedure.

3.2.1 Enhanced K-means Clustering

The enhanced K-means clustering is employed to cluster the data into similar teams supported distance among the data components. Attributable to the flexibleness in cluster roughness, extendibility and potency it helps modeling Web usage knowledge for clustering based recommender systems. Enhanced K means suggests that clustering is a concept for examining the cluster that aims to partition n knowledge objects into k clusters during which every knowledge object belongs to the cluster with the closest mean. The improved K-means (KM) clustering may be a heuristic algorithmic rule that minimize total of squares of the gaps in all samplings for agglomeration domain to clustering centers to get for the minimum k number of grouping on the idea of major function. Enhanced



A	B	C	D	E
Url	IpAddress	Time Sper	Average s	Title
http://in.search.yahoo.com/search?p=Sachin+Tendulkar&s=Search&fr2=sb-top&fr=yf	10.6.156.2	134.453 se	097.59860	Sachin Tendulkar - Yahoo! India Search Results
http://en.wikipedia.org/wiki/Sachin_Tendulkar	10.6.156.2	0.454 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Sachin_Tendulkar	10.6.156.2	532.75 sec	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match	10.6.156.2	2.343 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match	10.6.156.2	8.522 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	1.266 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	21.812 sec	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Main_Page	10.6.156.2	1.282 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Main_Page	10.6.156.2	26.703 sec	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	4.406 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	18.453 sec	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	16.094 sec	097.59860	Sachin Tendulkar - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	4.469 secc	097.59860	Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	4.781 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	2.75 secor	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	2.359 secc	097.59860	Man of the match - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	3.438 secc	097.59860	Sachin Tendulkar - Yahoo! India Search Results
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	2.203 secc	097.59860	sachin tendulkar records - Yahoo! India Search Results
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	2.515 secc	097.59860	sachin tendulkar records - Yahoo! India Search Results
http://en.wikipedia.org/wiki/Main_Page	10.6.156.2	6.422 secc	097.59860	Sachin Tendulkar Century | Tendulkar Batting | Sachin Tendulkar Cricket Records
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	11.422 sec	097.59860	Sachin Tendulkar Century | Tendulkar Batting | Sachin Tendulkar Cricket Records
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	5.828 secc	097.59860	Sachin Tendulkar And His Family | Sachin Tendulkar Family Photos With His Wife
http://en.wikipedia.org/wiki/Man_of_the_Match#Cricket	10.6.156.2	21.532 sec	097.59860	Sachin Tendulkar - Wikipedia, the free encyclopedia
http://in.search.yahoo.com/search?vc=&p=Sachin+Tendulkar&toggle=1&cop=mss&ei	10.6.156.2	75.125 sec	097.59860	Sachin Tendulkar And His Family | Sachin Tendulkar Family Photos With His Wife

Figure 3. After preprocessing.

K-means algorithmic rule takes initial mean as input.

$$I(t)=\sum_{a=1}^{k}\sum_{b=1}^{y}\left\|y_b^{(a)}-c_a\right\|^2 \qquad t_1\le t_2 \qquad (1)$$

Every upgrade to imply that in scope of cycles makes those implies that closer to last implies that. Amid this system, improved K-implies algorithmic standard meets the information focuses when assortment of emphases. Starting implies that and last future implies that range connected to allocate information objects into bunches. At first, information items are apportioned into groups that have the nearest mean to them by exploitation introductory implies that are given to the algorithmic standard as data. When all learning articles are distributed into clusters, agglomeration proposes that are recalculated by exploitation the items inside of the groups. These recommend that protests are thought to be closer to last implies that in correlation with beginning ones. Next, all items are reassigned to clusters by exploitation new means.

Most likely, a few articles can move to entirely unexpected groups once exploitation new recommends that considering their groups with the past means. The improved K-means calculation's operation is best than existing ones. The algorithmic guideline recognizes the clusters a perception into K bunches, where K is prepared as an input parameter. After that it doles out for each perception and groups them upheld the perception's vicinity to the mean of the cluster. The group's mean is then recalculated furthermore the system starts once more. where I(t) is cluster at time t, t_1 and t_2 are the start-time and end-time of a user session separately, $||y_b^{(a)}- c_a||^2$ is a picked distance (intra) measure between an information point $y_b^{(a)}$ and the cluster centre c_a, is a marker of the separation of the *n* information focuses from the particular group focuses. The term intra is utilized to experience the reduction of the groups. The extent connection of base intracluster separation to biggest intracluster separation are frequently brought as $\boldsymbol{D_{index}}$

$$D_{index}=\frac{Cluster_{\mathrm{min}}}{Cluster_{\mathrm{max}}}$$

Where, $\mathrm{Cluster}_{\mathrm{min}}$= minimal intracluster distance, $\mathrm{Cluster}_{\mathrm{max}}$= maximal intracluster distance. The process of implementing the enhanced K-means algorithm is shown in Figure 3.

3.2.2 Distance Measure

Usually for multi-dimensional data rather than using Euclidean distance, better results are produced by cosine resemblance, because. Cosine resemblance is a measure of resemblance between two vectors by measuring the cosine of the angle between them whereas Euclidean distance may intercept all the objects to be of equal distance for high dimensional data.

- The cosine of 0 is 1, and less than 1 for any other direction.
- The cosine of the point between two vectors along these lines decides likeness between the vectors; consider the two vectors of qualities, for instance say X and Y are considered. Likeness can be spoken to utilizing a speck item and extent as

$$S(x_i, y_j)=\frac{x_i^T . x_j}{\|x_i\| . \|x_j\|} \qquad (3)$$

The process of implementing the enhanced K-means algorithm is shown in Figure 4.

```
Algorithm : Enhanced K-means Clustering Algorithm
Input : K-number of groups, D-information set
        containing n objects
Output : An arrangement of K-number of groups for
         time interim [t1, t2]
Step 1: Arbitrarily pick K objects from D as the
        starting group focuses legitimate amid [t1, t2].
repeat
Step 2: (re)assign every item to the bunch to which
        the article is most comparative utilizing eq.
        (1), taking into account the mean estimation
        of the items in the group amid [t1, t2].
Step 3: Calculate starting means
Step 4: Assign objects into groups by utilizing
        introductory means.
Step 5: Calculate $D_{index}$.
         If $D_{index}$ < $Min$ during [t1, t2].
Do-while
         Objects continue to another groups
Step 6: Re-figure method for bunches by utilizing
        items having a place with them for [t1, t2].
Step 7: Depute the articles into groups by utilizing
        computed means for the time length of time.
until no change
 Step 8: Compute intra-cluster distance using eq. (2).
 Step 9: If the new intra-bunch separation< old_intra_
         cluster separation and new_intercluster>old_
         inter_cluster distance, calculate the
         intracluster mean for [t1, t2].
```

Figure 4. Enhanced K-means clustering algorithm.

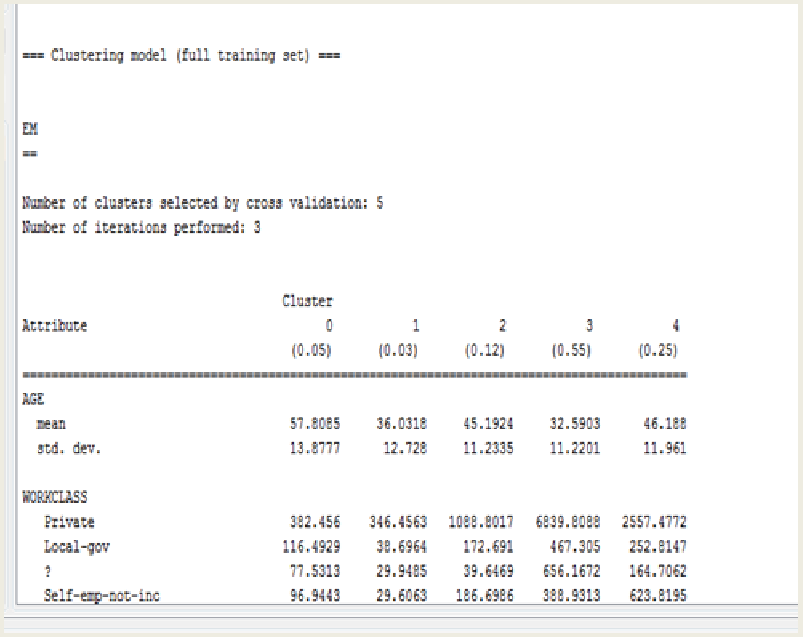
3.2.3 User Clustering using Enhanced K- means Algorithm

User grouping is the cognitive process for grouping users into a predefined set of categories based on their similarity in categories. For this purpose, k means algorithm is used. Figure 5 shows the clustering of users with categories. We take ten categories namely Education, Sports, Engineering, Computer, Entertainment, News, Travel, Finance, Music and Others. User1 (U_1) and User32 (U_{32}) come under the category EDUCATION. Similarly other users are grouped into respective categories.

3.2.4 Hill Cipher Algorithm

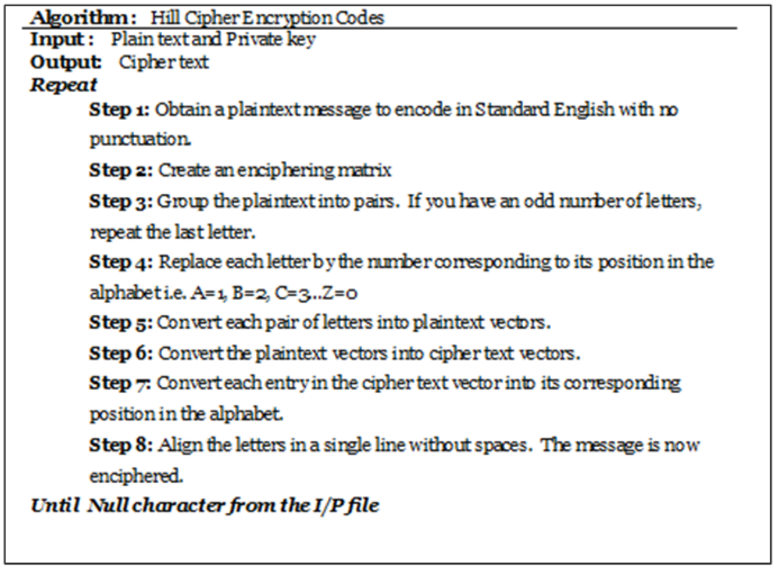
The communication between the various user cluster groups is made secure by employing encryption using hill cipher algorithm. Algorithm representation of the Hill-cipher encryption and decryption is shown in Figures 6 and 7 respectively.

By ensuring safe communication between the user groups based clusters, Web and browsing data leaks can be prevented thus eliminating security threat and privacy concern.



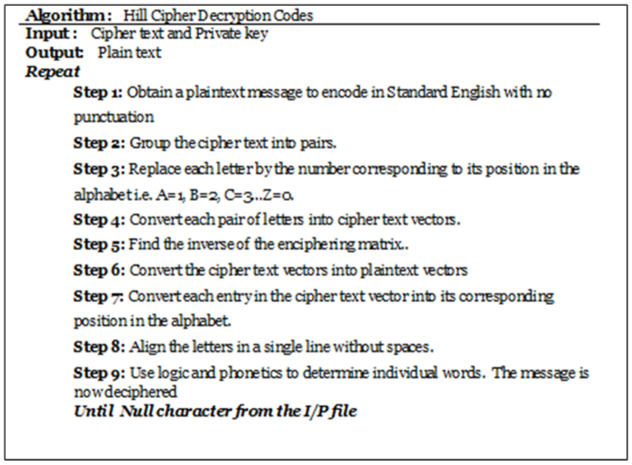
```
=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 5
Number of iterations performed: 3


                           Cluster
Attribute                        0         1         2         3         4
                            (0.05)    (0.03)    (0.12)    (0.55)    (0.25)
============================================================================
AGE
  mean                     57.8085   36.0318   45.1924   32.5903    46.188
  std. dev.                13.8777    12.728   11.2335   11.2201    11.961

WORKCLASS
   Private                 382.456  346.4563 1088.8017 6839.8088 2557.4772
   Local-gov              116.4929   38.6964   172.691   467.305  252.8147
   ?                       77.5313   29.9485   39.6469  656.1672  164.7062
   Self-emp-not-inc        96.9443   29.6063  186.6986  388.9313  623.8195
```

Figure 5. User clustering using Enhanced K-means.



```
Algorithm:  Hill Cipher Encryption Codes
Input:  Plain text and Private key
Output:  Cipher text
Repeat
        Step 1: Obtain a plaintext message to encode in Standard English with no
        punctuation.
        Step 2: Create an enciphering matrix
        Step 3: Group the plaintext into pairs. If you have an odd number of letters,
        repeat the last letter.
        Step 4: Replace each letter by the number corresponding to its position in the
        alphabet i.e. A=1, B=2, C=3..Z=0
        Step 5: Convert each pair of letters into plaintext vectors.
        Step 6: Convert the plaintext vectors into cipher text vectors.
        Step 7: Convert each entry in the cipher text vector into its corresponding
        position in the alphabet.
        Step 8: Align the letters in a single line without spaces. The message is now
        enciphered.
Until Null character from the I/P file
```

Figure 6. Hill Cipher Encryption Code.



```
Algorithm: Hill Cipher Decryption Codes
Input : Cipher text and Private key
Output: Plain text
Repeat
        Step 1: Obtain a plaintext message to encode in Standard English with no
        punctuation
        Step 2: Group the cipher text into pairs.
        Step 3: Replace each letter by the number corresponding to its position in the
        alphabet i.e. A=1, B=2, C=3..Z=0.
        Step 4: Convert each pair of letters into cipher text vectors.
        Step 5: Find the inverse of the enciphering matrix..
        Step 6: Convert the cipher text vectors into plaintext vectors
        Step 7: Convert each entry in the cipher text vector into its corresponding
        position in the alphabet.
        Step 8: Align the letters in a single line without spaces.
        Step 9: Use logic and phonetics to determine individual words. The message is
        now deciphered
        Until Null character from the I/P file
```

Figure 7. Hill Cipher Decryption Code.

4. Discussion

Information grouping is one of the broad KDD procedures, which is increases significance in streamlining examination as it were. In the previous years, different enhancement procedures were utilized to enhance different part of breaking down bunches to accomplish ideal results. A log file was collected for a six months period which was stored as text documents with 50 users. The searched path of users was compared with ODP (Open Directory Project) Taxonomy. The root word (category name) of the path was extracted and the users were clustered under the corresponding category. The user interest as well as their temporal session time based dynamic user clusters are shown in Figure 8.

5. Conclusion

In this paper temporal based clustering analysis is proposed to group Web users into identical clusters based on their browsing behaviors during a specific time interval. The process cannot be performed unless Web Usage Mining data is passed through sophisticated pre-processing steps. The pre-processing steps include Data Cleansing, User Identification and Session Identification. Thus the pre-processed Web usage data is clustered using enhanced K-means intelligence based optimization approach called Enhanced K-means clustering algorithm. Hence providing the contribution of intelligent based clustering in developing intelligent security systems with the help of ODP.

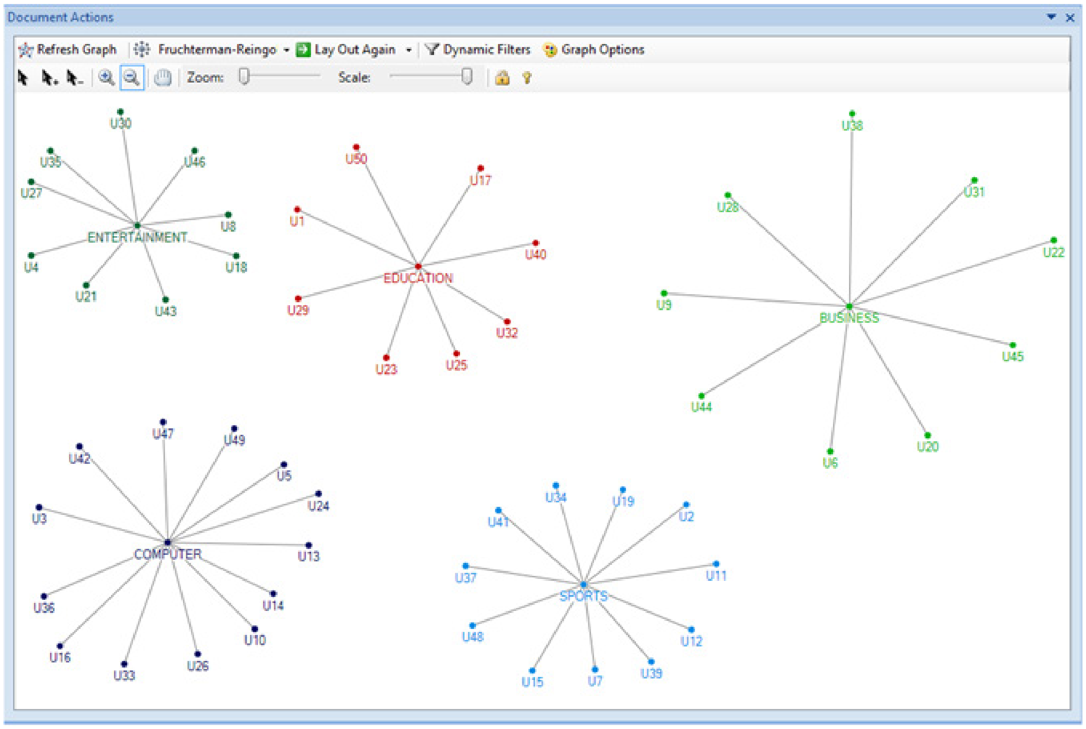


Figure 8. User grouping based on their Interest.

6. References

1. Wei Li. Modified K-means clustering algorithm. IEEE computer society Congress on Image and Signal Processing. 2008 May 27-30; 618–21.
2. Singh RY, Bhatia MPS. Data clustering with modified K-means algorithm. IEEE International Conference on Recent Trends in Information Technology. ICRTIT. 2011 Jun 3-5; p. 717–21.
3. Wang S, Dai F, Liang B. A path-based clustering algorithm of partition. Information and Control. 2011; 40(1):141–4.
4. Wu J, Li X, Sun T, Li W. A density-based clustering algorithm concerning neighbourhood balance. Journal of Computer Research and Development. 2010; 47(6):1044–52.
5. Wang S, She L, Liu Z, Fu Y. Algorithm Research on User Interests Extracting via Web log data. International Conference on Web Information Systems and Mining. 2009 Nov 7-9; p. 93–7.
6. Wang Z, Wu J. Related queries recommendation based on User Logs for Chinese Search Engines. IEEE. 2008 Oct 12-14; 1–4.
7. Chi CC, Kuo CH, Lu MY, Tsao NL. Concept-based pages recommendation by using cluster algorithm. 8th IEEE International Conference on Advanced Learning Technologies. 2008; p. 298–300
8. Ujjin S, Bentley PJ. Particle swarm optimization recommender system. IEEE. 2003 Apr 24-26; 124–31.
9. Wan M, Jonsson A, Wang C, Li L, Yang Y. A random indexing approach for web user clustering and web prefetching. PAKDD 2011 Workshops, LNA1. Springer. 2012; 7104:40–52.
10. Mukherjee I, Bhattacharya V, Banerjee S, Gupta PK, Mahanti PK. Efficient web information retrieval based on usage mining. 1st International Conference on Recent Advances in Information Technology. 2012 Mar 15-17; p. 591–5.
11. Bouras C, Tsogkas V. Clustering user preferences using W-k means. Seventh International Conference on Signal Image Technology and Internet-Based Systems. 2011 Nov 28-Dec 1; p. 75–82.
12. Yi C, Ning Z. Study of the users' interests based on the internet browsing history. International Conference on Business Management and Electronics Information (BMEI). 2011 May 13-15; p. 234–8.
13. Zheng W, Zhang M. The investigation for web user clustering based on interest. International Conference on Electronics, Communications and Control (ICECC). 2011 Sep 9-11; p. 553–6.
14. Xu J, Liu H. Web user clustering based on k-means Algorithm. International Conference on Information, Networking and Automation (ICINA). 2010 Oct 18-19; p. V2-6–9.
15. Vijayakumar P, Bose S, Kannan A. Centralized key distribution protocol using the greatest common divisor method. Computers and Mathematics with Applications. 2013 May; 65(9):1360–8.
16. Sethukkarasi R, Ganapathy S, Yogesh P, Kannan A. An intelligent neuro fuzzy temporal knowledge representation model for mining temporal patterns. Intelligent and Fuzzy Systems. 2014; 26(3):1167–78.
17. Slaninova K, Dolak R, Miskus M, Martinovic J, Snasel V. User segmentation based on finding communities with similar behavior on the web site. International Conference on Web Intelligence and Intelligent Agent Technology. 2010; p. 75–8.
18. Jyoti, Sharma AK, Goel A, Gulati P. A novel approach for clustering web user sessions using RST. International Conference on Advances in Computing Control and Telecommunication Technologies. 2009 Dec 28-29; p. 657–9.
19. Chen C.A fuzzy rough approximation approach for clustering user access patterns. IEEE World Congress on Software Engineering. 2009 May 19-21; p. 276–80.
20. Chen L, Bhowmick SS, Nejdl W. COWES: Web user clustering based on evolutionary Web sessions. Data and Knowledge Engineering. 2009 Oct; 68(10):867–85.
21. Chitraa V, Davamani AS, A survey on preprocessing methods for Web usage data. International Journal of Computer Science and Information Security. 2010 Mar; 7(3):78–83.
22. Dixit D, Kiruthika M. Preprocessimg of Web Logs. International Journal on Computer Science and Engineering. 2010; 02(07):2447–52 .
23. Sudheer Reddy K, Kantha Reddy M, Sitaramulu V. An effective Data Preprocessing method for Web Usage Mining. International Conference on Information Communication and Embedded Systems. 2013 Feb 21-22; p. 7–10.
24. Kewen L. Analysis of preprocessing methods for Web Usage Data. International Conference on Measurement, Information and Control. 2012 May 18-20; p. 383–6.