ISSN 2348 – 7968

An Efficient Clustering Technique for Weblogs

V.Vidyapriya[1], S.Kalaivani[2].

[1] ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE,
QUAID-E-MILLATH GOVERNMENT COLLEGE FOR WOMEN,
CHENNAI - 600 002.

[2] RESEARCH SCHOLAR, PG AND RESEARCH DEPARTMENT OF COMPUTER SCIENCE,
QUAID-E-MILLATH GOVERNMENT COLLEGE FOR WOMEN,
CHENNAI - 600 002.

ABSTRACT- Web mining research includes several research communities such as database, information recovery and artificial intelligence. Web mining divided into three categories they are web usage mining, web content mining and web structure mining. Web content mining is used to extract useful information from the webpages. Structure mining deals with in link and out link and web usage mining is to discover interesting usage patterns from the web data. The pre-processing step can improve text quality by eliminating irrelevant data. Clustering is to group the data based on their similarities. This paper mainly focuses on clustering web log data to identify user access pattern. In first phase going to pre-process the web logs data. Pre-processing step is done to make the sample raw web logs more efficient. If the web logs are large in size with unwanted data it will not provide better result. So, Pre-processing step is done. In second phase clustering techniques is used. Sample Web logs are clustered using k-means clustering and farthest first clustering technique. By clustering the webpages we can easily identify the user interest and the access pattern. In third phase visualize the clustered same weblogs.

Keywords: Web mining, Pre-processing, k-means, farthest first clustering algorithm, cluster.

1. INTRODUCTION

World Wide Web consists of enormous information and still growing. The reporting of the information is also very extensive and different. One can find any information on the web. All types of data exist on the web. Information we find in the web is unrelated. Due to varied authorship of web pages, many pages may present the same or similar information using completely different words and formats. This makes grouping of information from multiple pages a problem.

Web mining is used to discover useful information from web hyperlink structure, page content and usage data. Web mining uses many data mining techniques which includes supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining. Web mining is related to data mining process. We can find difference in data collection. In traditional data mining the data is already collected and stored in the data warehouse. For web mining, data collection is an important task especially for web structure and content mining which involves crawling a large number of target web pages

Purpose of Web Mine:

There are enormous information are available on the web but we need only certain information which are related to our query. It is possible to extract interesting pieces of information. Web has become an authoritative platform for recovering information and discovering knowledge from web data [1].

Web Applications:

Web applications can be lists such as to aim potential customers for e-commerce, to improve the quality and delivery of web information service to end user, to enhance the web server program's performance, to personalize the delivery of web content, to improve web designs, to improve customer satisfaction, to recognize the prospective main advertisement locations, to facilitate adaptive sites, to expand site design, to ensure fraud detection and to predict user behavior.

Web mining issues:

Now a day's internet has become a part of our life and more popular. Internet plays a vital role in our daily life. Many people are using internet which includes school students, business people and in medical field. There are number of web users increasing at exponential speed [3]. Many different types of data are shared on web which includes text, images, audio, video, XML, and HTML. Web datasets can be very vast ranges from ten to hundreds of tera bytes. So it couldn't be extracted from a single server. To store large datasets we need large number of servers.

2. WEB USAGE MINING

Web mining is also same as data mining but the difference is that it extracts data from the web. Some of the data mining techniques which include clustering, classification, association rule and sequential rule mining are used to identify user access patterns. Web usage mining is a type of web mining technique to discover interesting usage patterns from the web log files.

Mining techniques for web usage mining

Clustering is an important technique which is used to group the data which are similar to each other. It is a suitable technique used to analyse large dataset and then group the data based on the similarity.

Classification is a technique which is used in web mining. It is mainly used to group the users based on their browsing history. ***Association rule mining*** is an important class of regularities in data. Mining of association rules is a fundamental data mining task. Its objective is to find all co-occurrence relationships among data items. ***Sequential rule mining*** is useful to predict user navigation behaviour which is used to improve server's response time [5]. These data mining techniques cannot be directly applied to log files due to redundancy, irrelevant data, and unstructured data. To avoid these problems data pre-processing is needed [4].

Data Sources for web usage mining

There are three data sources are used to gather log data for web usage mining. Data sources include server log, client log, and proxy server log.

Server log: when a person request a particular page on web whenever an entry is logged into a special file is called server log file [6].

Client log: it is also known as browsers log. Web log data can also be gathered from user machine by integrating java applets to the websites, writing java scripts or even modified browsers [7][8].

Proxy server log: it is a server which acts as an intercessor between the user's requests to other web servers [8].

Weblog information can be integrated with web content and web structure mining. When user access a particular page entry is entered in web log server. The interaction details of users with website are recorded automatically in web servers as the form of weblogs [2]. Weblogs are kept as in form of line of text in web server, proxy server and browser.

Transfer/Access Log

The information regarding users request from their web browsers is deposited in access log. Information in transfer log which includes Time, Date, Host name, File requested Amount of data transferred and status of report. Example of access log format given below:

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
```

Referrer Log

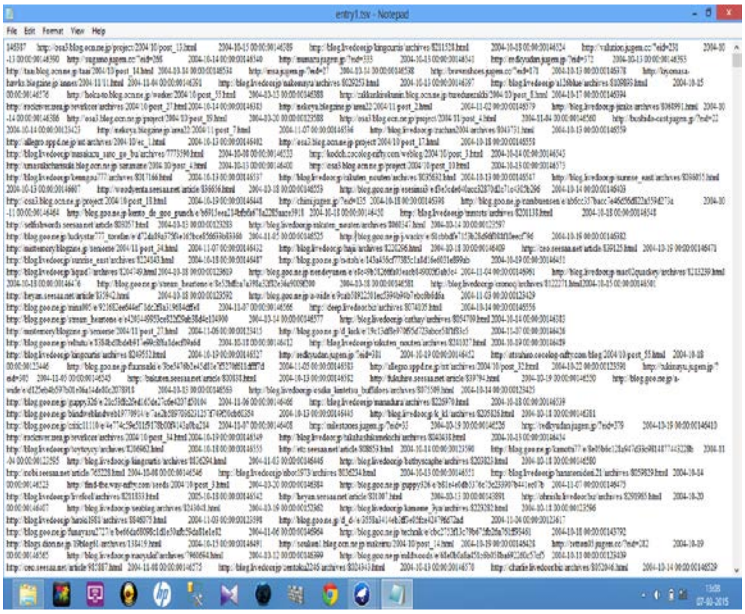
Referrer log includes two fields they are URL and Referrer URL

Error Log

The errors and requests which have failed to access are collected in error log. Not only for the page which grasps links to a file that does not exist, but also for the user who is not permitted to access a particular page, the user request may fail. The information that is contained in most error log entries is the message given below:

```
[Wed Oct 11 14:32:52 2000] [error] [client 127.0.0.1] client denied by server
Configuration: /export/home/live/ap/htdocs/test
```

Sample raw web log data shown here.



Example from collected raw dataset.

```
146607        http://woodyenta.seesaa.net/article/836656.html        2004-10-18 00:00:00
```

3. DATA PREPROCESSING

There are three important steps in web usage mining they are data pre-processing, pattern discovery and pattern analysis. In these process data pre-processing is a complex task it takes around 80% of time to do pre-process. Data mining techniques cannot be directly applied on the data sets. So, the data pre-processing is done to remove inconsistent data, redundant data, and noise data. The steps for data pre-processing are data cleaning, User identification, Session Identification, path completion.

Log Data file

The raw data for mining purpose is collected .It contains approximately 1, 00,000 records in Common log file format.

Data Cleaning

Data cleaning is one of the techniques used in data pre-processing. It is used to remove unwanted data from the datasets [12]. Data cleaning make the data more efficient. The data cleaning takes the following steps:

Step1: Removal of entries which contains image files, graphics or multimedia files. After performing this step 50,000 i.e. 50% of records are left.

Step2: Removal of entries with failed status code [16].

Step3: Removal of entries with bytes transferred field zero. It indicates that the page is not opened. So it has to be removed. After removing these records 40,000 i.e. 40% of records are left.

Step4: Removal of records which has been visited less than three times. After removing these records 25,000 i.e. 25% of records are left.

User and Session identification

User identification is used to identify the unique user. Unique user can be identified based on their log name and rfname of the attribute. Session identification [9] is the process of allocating the individual user access logs into sessions.

Path Completion

The incomplete access path of every user session is recognized based on user session identification.

4. K-MEANS CLUSTERING ALGORITHM

K-means clustering is the heuristic method. In k-means clustering each cluster is represented by the centre of the cluster. Different kinds of measures can be used they are Manhattan distance (L_1 norm), Euclidean distance (L_2 norm), cosine similarity.

Execution of k-means algorithm

Given K, the number of clusters, the k-means clustering algorithm is outlined as follows

1. Select K points as initial centroid
2. Repeat
   (i) Form K clusters by assigning each point to its closest centroid.
   (ii) Recompute the centroids (i.e. mean point) of each cluster

3. Until convergence criterion is satisfied.

Efficiency:

O(tKn) where n: number of objects, K: number of cluster and t: number of iteration. K-means clustering often terminates at local optimal. Initialization can be important to find high-quality clusters. Need to specify K, the number of clusters in advance. There are ways to automatically determine the “best” K. In practice, one often runs a range of values and selected the “best” K value. K-means is applicable only to objects in continuous n-dimensional space.

Top entry pages clustered using k-means clustering

Top visited pages	**Number of hits**	**Clustered instances**
http://osa3.blog.ocn.ne.jp/project/2004/11/post_4.html	**5673**	**22%**
http://blog.livedoor.jp/livefool/archives/8211833.html	**5478**	**21%**
http://blog.livedoor.jp/get_life/archives/6498588.html	**4050**	**16%**
http://rakutenlolita.seesaa.net/article/846539.html	**4158**	**16%**
http://ore-1gpy.cocolog-nifty.com/1/2004/11/post_50.html	**3264**	**13%**
http://blog.livedoor.jp/asudeki/archives/4273346.html	**3045**	**12%**

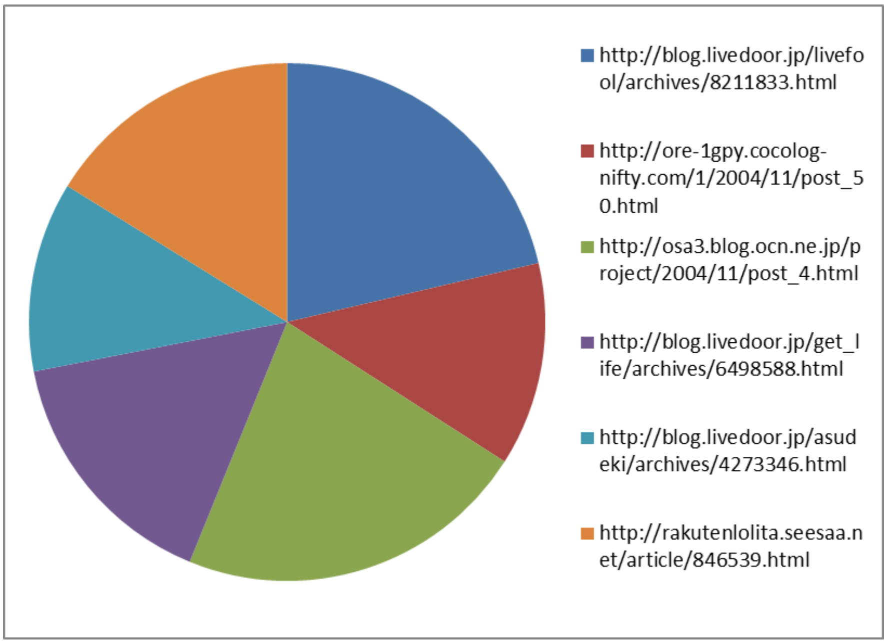
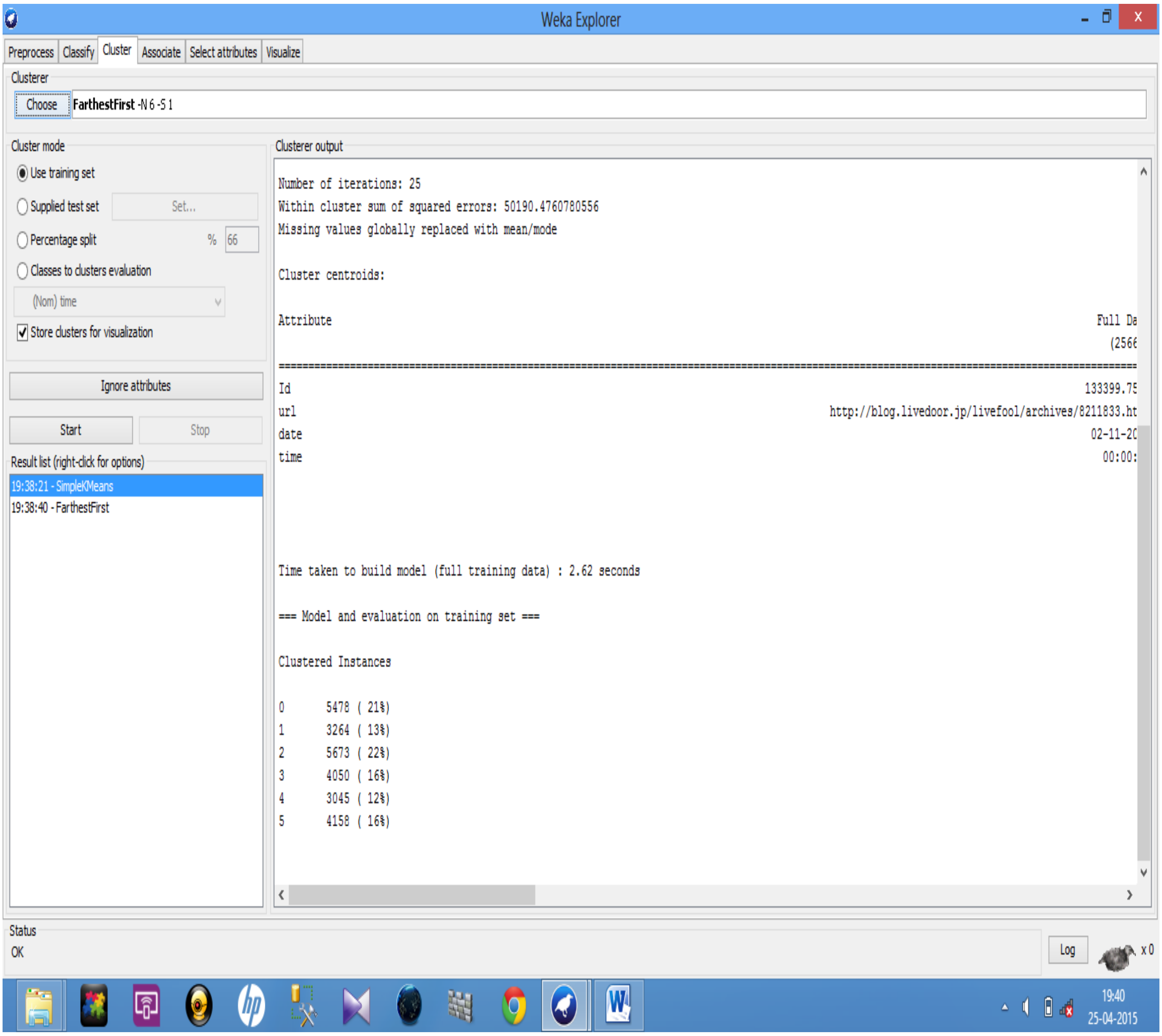


Fig 1 pie chart shorwing frequently accessed webpages using k-means cluster



Clusterer
FarthestFirst -N 6 -S 1

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

Result list (right-click for options)
19:38:21 - SimpleKMeans
19:38:40 - FarthestFirst

Clusterer output

```
Number of iterations: 25
Within cluster sum of squared errors: 50190.4760780556
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute                                                                    Full Da
                                                                               (2566
====================================================================================
Id                                                                         133399.75
url                                  http://blog.livedoor.jp/livefool/archives/8211833.ht
date                                                                           02-11-20
time                                                                            00:00:




Time taken to build model (full training data) : 2.62 seconds

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

Clustered Instances

0       5478 ( 21%)
1       3264 ( 13%)
2       5673 ( 22%)
3       4050 ( 16%)
4       3045 ( 12%)
5       4158 ( 16%)
```

Status
OK

19:40
25-04-2015

Number of iterations: 4
Sum of squared errors: 50192.95
Time taken to build model: 2.62 sec

5. FARTHEST FIRST CLUSTERING ALGORITHM

Farthest first algorithm is first proposed by Hoch Baum and shmoys (1985). Farthest first clustering algorithm is the heuristic method. It has two phases centroid selection and cluster assignment. Centroid selection begins by selecting a random data points as the original clustering centre then chooses the next centre as the data point furthest (according to the distance metric) from the first centre[17].

Farthest first algorithm also differs from k-means in that all centroids are actual data points and not geometric centre of clusters [17]. It achieves good performance in terms of centroid selection.

ISSN 2348 – 7968

Top entry pages clustered using farthest first clustering algorithm

Top visited pages	Number of users	Clustered instances
http://hama.way-nifty.com/hama/2004/09/post_6.html	6436	25%
http://Mint7.blog.ocn.ne.jp/mint7/2004/07/__7.html	4607	18%
http://seilan-sara.jugem.jp/?eid=8	4404	17%
http://dv.blogdns.com/archives/000109.html	4031	16%
http://merrystyle.cocolog-nifty.com/merrystyle_blogs/2004/10/2_1.html	4119	16%
http://blog.livedoor.jp/softcream/archives/8233022.html	2071	8%

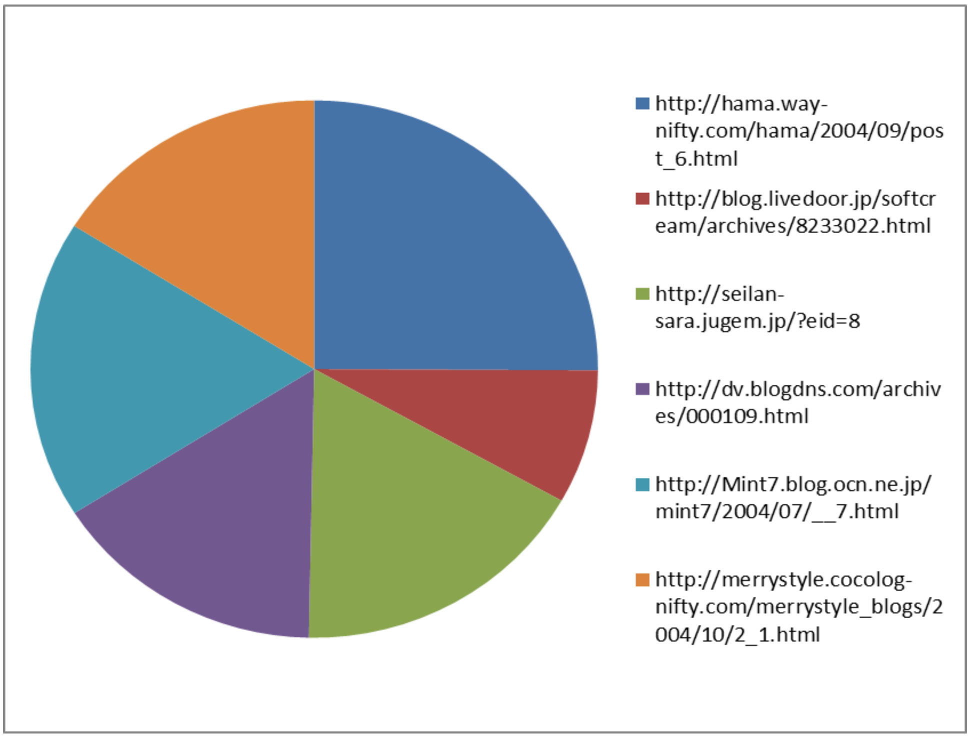
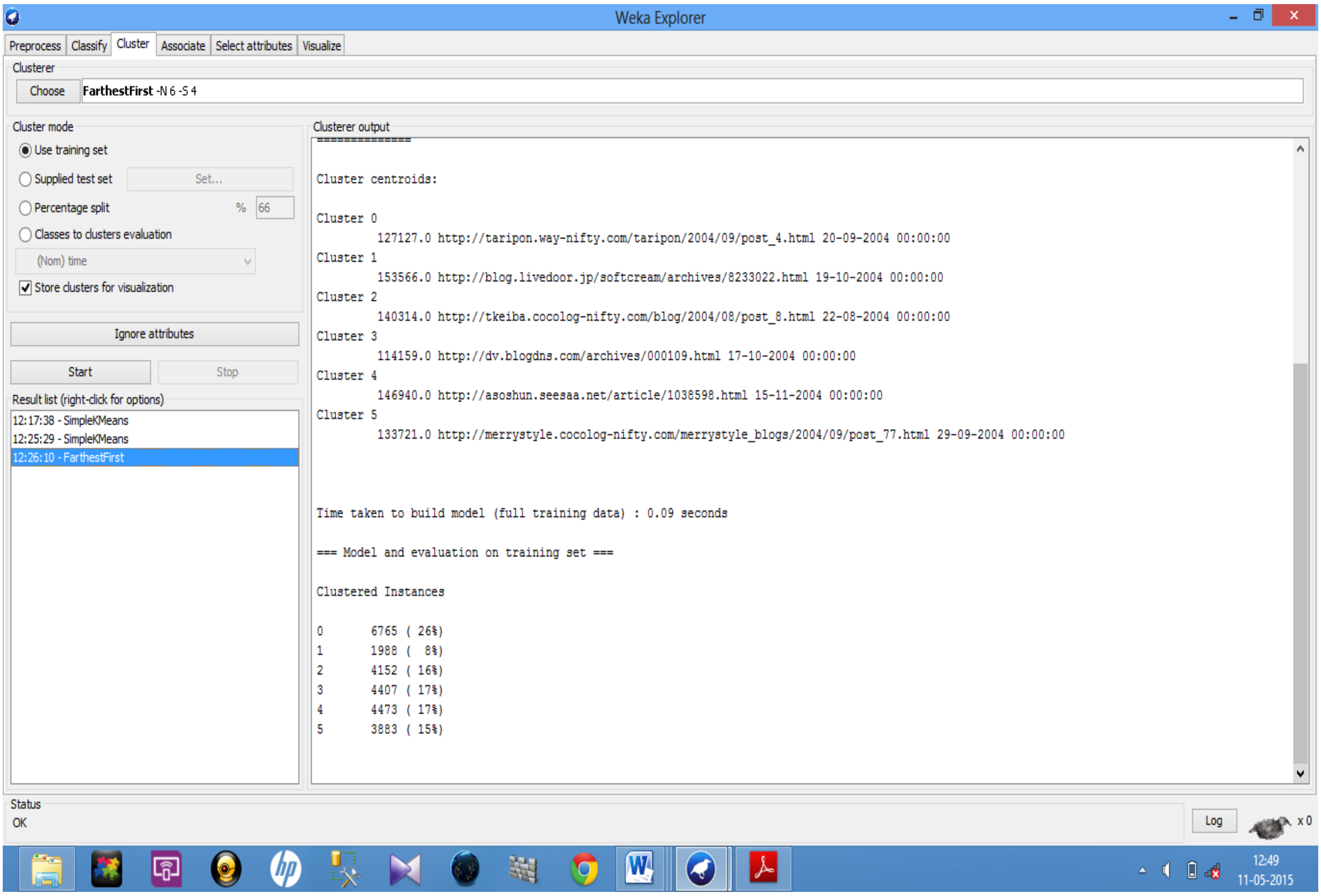


Fig 2 pie chart shorwing frequently accessed webpages using farthest first clustering technique



Clusterer
Choose FarthestFirst -N 6 -S 4

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

Result list (right-click for options)
12:17:38 - SimpleKMeans
12:25:29 - SimpleKMeans
12:26:10 - FarthestFirst

Clusterer output

```
==============

Cluster centroids:

Cluster 0
         127127.0 http://taripon.way-nifty.com/taripon/2004/09/post_4.html 20-09-2004 00:00:00
Cluster 1
         153566.0 http://blog.livedoor.jp/softcream/archives/8233022.html 19-10-2004 00:00:00
Cluster 2
         140314.0 http://tkeiba.cocolog-nifty.com/blog/2004/08/post_8.html 22-08-2004 00:00:00
Cluster 3
         114159.0 http://dv.blogdns.com/archives/000109.html 17-10-2004 00:00:00
Cluster 4
         146940.0 http://asoshun.seesaa.net/article/1038598.html 15-11-2004 00:00:00
Cluster 5
         133721.0 http://merrystyle.cocolog-nifty.com/merrystyle_blogs/2004/09/post_77.html 29-09-2004 00:00:00



Time taken to build model (full training data) : 0.09 seconds

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

Clustered Instances

0       6765 ( 26%)
1       1988 (  8%)
2       4152 ( 16%)
3       4407 ( 17%)
4       4473 ( 17%)
5       3883 ( 15%)
```

Status
OK

Number of iterations: 4
Time taken to build model: 0.09 sec

When comparing both k-means clustering and farthest first clustering algorithm. The time taken to build model is minimum in farthest first clustering algorithm and maximum time taken to build model in simple k means clustering algorithm.Web log data is Pre-processed and Clustered using weka tool.

6. CONCLUSION AND FUTURE ENCHANCEMENT

Web usage mining has developed as the essential tool for understanding more user friendly and business web services. Web usage mining is used in e-commerce, e-banking sites to organize their sites and increase their profits. K-means clustering and farthest first clustering algorithm are compared. K-means clustering algorithm couldn’t handle overlapping data points because it classifies a point based on its distance from the estimated mean values. Farthest first clustering algorithm took minimum time to cluster in both partitioning and non-partitioning. Cluster is mainly measured as an essential work. It helps to do web usage analysis based on their browsing history. It can be further extended using more than two clustering technique; many pre-processing can be applied effectively on the web logs.

7. REFERENCES

1. Jiawai Han and Micheline Kamber, ”Data mining-Concepts and techniques”, secondedition, Elsevier, Reprint 2010.

2. http://www.galeas.de/webmining.html.

3. Hussain.T, Asghar.S and Masood.N, “Web usage mining: A survey on preprocessing of web log file”, Information and emerging technologies, 2010

4. Pabarskaite Z (2002), Implementing advanced cleaning and end-user interpretability technologies in web log mining in 24th International Conference on Information Technology Interfaces (ITI), Vol. 1 Page(s): 109-113.

5. Zidrina Pabarskaite, Aistis Raudys (2007), A process of knowledge discovery from web usage data: Systemization and critical review in Journal of Intelligent Information System, Springer Vol.28 Issue.1 Page(s)

6. Zidrina Pabarskaite, Aistis Raudys (2007), A process of knowledge discovery from web usage data: Systemization and critical review in Journal of Intelligent Information System, Springer Vol.28 Issue.1 Page(s): 79-104.

7. C. Shahabi, F. Banaei-Kashani (2002), A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking in WEBKDD Third International Workshop on Mining Web Log Data, Page(s): 113-144.

8. D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos (2003), Web usage mining as a tool for personalization: A survey in User Modeling and User Adapted Interaction journal, Vol. 13 Issues. 4 Page(s): 311-372.

9. Sheetal A. Raiyani and, Shailendra Jain, “Efficient Preprocessing technique using Web log mining, International Journal of Advancements in Research & Technology”, 1(6) ISSN 2278-7763, 2012.

10. J.Srivatsava, R.Cooley, M.Deshpande, and P.N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." ACM SIGKDD Explorat. Newsletter, 2000.

11. V.Chitraa, Dr.Antony Selvadoss Devamani, “A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing”, International Journal of Computer Applications, Volume 34– No.9, 2012

12. Vijayashri Losarwar and Dr. Madhuri Joshi, Data Preprocessing in Web Usage Mining, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, Singapore, 2012.

13. R. Suguna et.al,"User interest level based preprocessing algorithms using web usage mining", International Journal on Computer Science and Engineering.

14. Navin Kumar Tyagi1, A.K. Solanki2& Sanjay Tyagi3, An algorithmic approach to data preprocessing in web usage mining, International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 279-283

15. Cooley, R., B. Mobasher and J. Srivatsava, 1997. Web mining: Information and pattern discovery on the World Wide Web. Proceeding of the 9th IEEE International

16. http://en.wikipedia.org/wiki/List_of_HTTP_status_codes.

17. http://cse-wiki.uni.edu/wiki/index.php/clustering techniques