AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

Hybridization of Linear Support Vector Machine and Spectral Clustering Technique for User Web Log Files

[1]R. Deepa and [2]Dr. R Manicka Chezian

[1]Research Scholar, Department of Computer Science, NGM College, Pollachi, Coimbatore, India
[2]Associate Professor, Department of Computer Science, NGM College, Pollachi, Coimbatore India

Address For Correspondence:
R.Deepa, Research Scholar, Department Of Computer Science, NGM College, Pollachi, Coimbatore, India
E-mail: deepaharini2015@gmail.com, Chezian_r@yahoo.co.in

ARTICLE INFO

Article history:
Received 19 September 2016
Accepted 10 December 2016
Published 31 December 2016

Keywords:
SVM based Clustering; Web Usage Mining; Hyper Plane; Web Server;

ABSTRACT

Web Usage mining (WUM) has recently become an important tool for the electronic marketing and it is widely useful for knowing about the behavior of the users. The queries can be gathered through sources like web server and proxy server. The navigation pattern of user's are very hard for the website owners for improving the way information is presented and increase the number of users to the website. This paper determines the navigation pattern of the web user for their query. The information inferred for the mostly clicked web page link and the most time spent web page. This task can be achieved using the enhanced SVM (Support Vector Machine) based clustering technique that is chosen for managing the profiles on the basis of the user's query data. This paper hybrid the Support Vector Machine binary classification and spectral clustering techniques which first classifies the user profile and then cluster the data to get the required result.

INTRODUCTION

The web users find information on the WWW (WorldWideWeb) that satisfies their need for information. The resources on the WWW are large and increases by day and minute. Under these web search engine plays an important role in surfing information from web. Till today the search engines are not trained for the query. The biggest problem facing users of Web search engines today is the quality of the results they get back. While the results are often amusing and expand users' horizons, they are often frustrating and consume precious time [S. Brin and L. Page]. It finds the information by matching the keyword to keyword. In general, such user has different information needs for their query. For example, for the query keyword "windows "some users may be interested in documents dealing with the computer operating system windows while others may want documents related to the "house windows" models. Thus the web search results should adapt to users with different information needs, there are several approaches applying data mining techniques to extract usage patterns from Weblogs [M. Spiliopoulou and L. Faulstich]. Thus the machine learning algorithms are used to train the patterns of web user navigation from the data. The SVM based clustering algorithm introduces cluster data with no prior knowledge of input classes. The initialization step of classifying data using binary classification the SVM confidence parameters necessary for classification on each of the training instances can be obtained. The lowest confidence data (e.g., the worst of the mislabelled data) then has its' labels moved to the other class label. The SVM is then run again on the data set (with partly re-labelled data) and is assured for convergence in this same condition as it joined previously, and now there are fewer data points to carry with mislabelling penalties. This technique seems to limit being vulnerable to the local minima traps common with other techniques. Thus, there is an improvement in the algorithm on its weakly convergent result with the help of SVM re-training after each

To Cite This Article: R. Deepa and Dr. R Manicka Chezian., Hybridization of Linear Support Vector Machine and Spectral Clustering Technique for User Web Log Files. ***Aust. J. Basic & Appl. Sci.,*** *10(18): 24-33, 2016*

re-labelling on the worst of the misclassified vectors i.e., those feature vectors which possess confidence factor values beyond a set threshold. The repetition of the above process helps in the improvement of the accuracy, which is a measure of reparability, till there are no errors in classifications.

The Support vector Machine (SVM) is a classification algorithm used for classifying the datasets using support vectors. The clustering categories in SVM are the state of art in supervised learning. The Web page recommender systems need the support of supervised learning algorithms for clustering data. The web link is visited by the users and the browsing history is cached as snippets and the data of user profile management is maintained for learning the behavior of the web users. Many Business intelligent systems are highly needed of this type of recommender system to analyze their commercial value and needs (Finley, Thomas, and Thorsten Joachims 2005).The supervised clustering algorithm results better than the unsupervised hence combines the clustering algorithms with classification technique (Wang, Xiang-Yang, *et al*. 2016). The SVM based spectral clustering techniques overcome the k-means clustering based SVM. The spectral clustering technique aligns the data in adjacency matrix and the binary classifiers instead of maximum margin are used for the web recommender system based on the browsing history of the web users (Sugiyama *et al*, 2004). Thus the proposed system hybrids the spectral clustering algorithm with the SVM Classification technique.

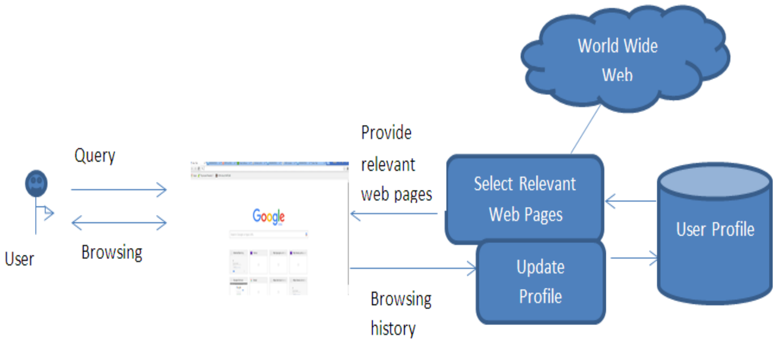


Fig. 2: System Overview

The Browsing History:

A Web log is a file to which the Web server writes information each time a user requests a Web Site from that particular server. A log file can be located in three different places:

- Web Servers
- Web proxy Servers
- Client browsers

The log file that resides in the web server notes the activity of the client who accesses the Web Server for a web site through the browser. A Proxy server is said to be an intermediate server that exist between the client and the Webserver. Therefore if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers maintain a separate log file for gathering the information of the user. This kind of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window the entries to the log file is done only by the Web server. Hundred users are invited to use the search engine application and the application is run on the system. The browsing histories of the users are catched in the snippet and the data is clustered using the SVM.

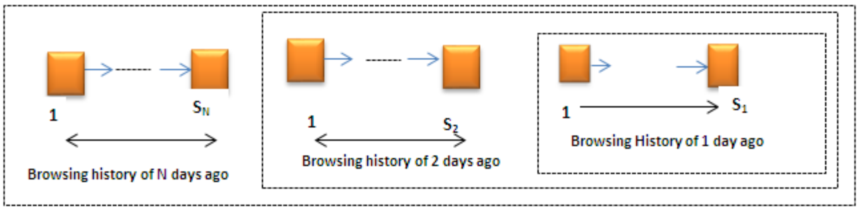

1
S_N
Browsing history of N days ago
1
S_2
Browsing history of 2 days ago
1
S_1
Browsing History of 1 day ago

Fig. 2: Users Browsing History for n days

2. Related Work:

A navigation pattern is a directed graph that summarizes the traversal movements of a group of visitors and satisfies certain human-centric criteria that make it "interesting" (M. Spiliopoulou and L. Faulstich, 1998). (S. Brin and L. Page, 2012) suggested that Search engines index tens to hundreds of millions of webpages involving a comparable number of different items. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the Web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and Web proliferation, creating a Web search engine today is very different from three years ago. Sugiyama *et al*, (2004) Says that the search engine is personalized under the approaches the relevancy, browsing history and because of the feedback of the people. Wang, Xiang-Yang, *et al*. (2016) proposed KMSVM algorithm that is to build classifiers by combining the K-means clustering technique with SVM. Experiments on the real-world databases have shown that compared with SVM, the KMSVM algorithm can build classifiers with both a higher response speed and a similar testing accuracy. Finley, Thomas, and Thorsten Joachims (2005), formulated a supervised clustering method SVMcluster based on an SVM framework for learning structured outputs. The algorithm accepts a series of "training clusters," a series of sets of items and clusterings over that set. Filippone, Maurizio, *et al*. Clustering is a classical problem in pattern recognition. Recently spectral and kernel methods for clustering have provided new ideas and interpretations to the solution of this problem (Zheleva, *et al* 2009). The key novel idea is that in addition to friendship links, groups can be carriers of significant information. A novel recommender system was formulated by Kumar, A., and P. Thambidurai, (2010). Their approach named, "FARM" is a recommendation system designed to help users sift through the mammoth amount of information obtainable in the World Wide Web. Their proposed approach is the combination of fuzzy and association rule mining algorithm. The combination exploits the advantages of both the methods thereby avoiding the shortcomings. FARM hybrid structure can be used for automatic recognition of emergent issues relevant to various groups of users. It also enables two scaling problems, pertaining to the rising number of users and documents, to be addressed. [Khabia, Apeksha, *et al*. (2014)] describes a review on most of document clustering technique and cluster based classification techniques used so far. Clustering based classification of text data is of great importance. Its goal is to automatically classify the text documents into different clusters and then exploit them to train the classifier. Among the large amount of text information available in electronic format, only 2% to 5% words of text corpus are used for text analysis and other words such as stop words, white spaces, header, footer etc. are not used for frequent pattern analysis and clustering the documents. Thus, lot of text pre-processing which is task of text mining, is required before text analysis and extracting knowledge from these text data. Alsmadi, *et al*. (2015) gives a brief review on clustering based classification techniques. The central theme in many of these is providing dimensionality reduction to improve text document classification. Clustering helps in reduction of the number of redundant features, which subsequently help in reducing the dimensions. Lee, Gyemin (2015), Proposed algorithm successfully identified the cluster components in the datasets. In this paper, presented a hierarchical clustering algorithm employing the OC-SVM. Rather than using distance to indirectly find the high-density regions, we use the OC-SVM to estimate level sets and to directly locate the high-density regions.

3 Methodologies:

Fuzzy Ontological Keyword-based user profiles are applied in the clustering process for accomplishing personalization effect. The data is linearly separable; hence linearly enhanced SVM computes the binary classifier (Wang, Xiang-Yang, *et al*. 2016). The SVM clustering case is an extension which permits effective clustering of similar text from article. The improved SVM approach deals with multi labeled of concept with 'm' classes and it decomposes the difficulty into 'm' binary problems. Also recent decomposition methods that dominate the field exist. Still, for convenience and for contrast with associated results simple decomposition for SVM clustering is chosen. Support Vector Machine-Clustering is a new approach that clusters data with no prior knowledge of input classes. The support vector Machine clustering is an effective clustering method that runs a binary classifier that finds a hyper plane. It splits optimally the training set. The hyper plane is characterized as by the decision function like f(g)=sgn(<w. Φ(g)> +x), where w is the weight vector orthogonal to the hyperplane, "x" is a scalar that represents the margin of the hyper plane, "g' is the current sample tested, "Φ(g)" is a function that transforms the input data into a higher dimensional feature space and "."is a dot product. The Sgn is the signum function that returns 1 if w has the value greater than to 0 and -1 otherwise. If w has unit length, then <w. Φ(g)> is the length Φ(g) along the direction w. Generally **w** will be scaled by ||**w**||. In the training part the algorithm need to find the normal vector "**w**" that leads to the largest "*x*" of the hyperplane. The problem seems very easy to solve but we have to keep in mind that the optimal classification line should classify correctly all the elements generated by the same given distribution. There are a lot of hyper planes that meet the classification requirements but the algorithm tries to determine the optimum. Feature vectors are denoted by g_k, where index i labels the M feature vectors $(1 \leq g \leq M)$ and index k labels the N feature vector components $(1 \leq g \leq N)$. For the binary SVM, labeling of training data is done using label variable $y_i = \pm 1$ (with sign according to whether the training instance was from the positive or negative class). For hyperplane separability, elements of the training set must satisfy the following conditions: $w_\beta g_{i\beta}-x \geq +1$ for i such that $y_i = +1$, and $w_\beta g_{i\beta}-x \leq -1$ for $y_i = -1$, for some values of the coefficients $w_1,..., w_N$, and x (using the convention of implied sum on repeated Greek indices). This can be written more concisely as: $y_i(w_\beta g_{i\beta}-x) -1 \geq 0$. Data points that satisfy the equality in the above are known as "support vectors" (or "active constraints").First, a query-

Keyword hyperplane is built by the clustering algorithm with one set of points related to the set of user's click, and the other related to the sets of time stamp. The distance between the boundary hyperplanes on the two classes of data are separated by a value 2/w, called as the "margin" where $w^2 = w_\beta w_\beta$. The increase in the margin between the separated data as far as possible leads to the SVM's optimal separating hyperplane. According to the usual SVM formulation, the aim to maximize w^{-1} is resaid as the goal to minimize w^2. The Augmented Lagrangian (AL) function of the dual problem expression then selects an optimum defined at a saddle point.

$$L\_A\ (x, y, w, b, \alpha) = (w\_\beta/2) - \alpha\_\gamma\ y\_(\gamma\beta{}^\wedge-\ )\ (w\_\beta\ x\_(\gamma\beta{}^\wedge-\ )\times b) - \alpha \qquad (1)$$

In principle, the positive parameter w_β in the augmented Lagrangian function, known as the penalty parameter, can also be changed from iteration to iteration, where $\alpha = \sum_\gamma \alpha_\gamma\ \alpha_\gamma > 0\ (1 \le \gamma \le M)$. The saddle point is obtained by reduction according to $\{w_1, \ldots, w_N, b\}$ and maximizing with respect to $\{a_1, \ldots, a_M\}$. Let $\{x_i\}$ be an extracted concepts of N points in a space.

Initially define the surrogate function $f_\eta(w)$as follows:

$$f_\eta(w) = {}^{max}_{\alpha \in R^n, v \in R^n} f_\eta(\alpha, v; w) \qquad (2)$$

Note that from the strong duality

$$f_0(\ w\ ) = \max_{\alpha, v} L_0(\alpha, v; w) \qquad (3)$$

In addition, $since\ L_0(\ \alpha\ , v\ ;\ w) \ge L_\eta(\alpha\ , v\ ; w)$ the inequality $f(w) \ge f_\eta\ (\ w)$ holds. Moreover, since $f_\eta(w) \ge\ L_\eta(\ \alpha^*\ , v^*\ ;\ w) = d\ (\ \alpha^*\ , v^*) =\ f\ (\ w^*)$ (here use $A^\top \alpha^* =\ v^*$ to obtain the first equality), have min $w \in R^n\ f_\eta(w) = f(w^*)$ for any nonnegative η. Moreover, f η (w) is differentiable if η > 0.

Similar to the nonlinear SVM conceptualization, using a non-linear transformationϕ, transform *x* to a high-dimensional space – *Kernel space* – and the smallest enclosing sphere of radius *R* is looked for. Hence the Hellinger Distance formula [Harsha, *et al.*] for similarity matching is given in eqn (3). A new metric of distance between probability distributions is referred to as Hellinger distance. Using some of the nice properties of this distance, generalize the fooling set argument for deterministic protocols to the randomized setting. For probability distributions $X = \{x_i\}_{i \in [n]}$,$Y = \{y_i\}_{i \in [n]}$supported on [n],the Hellinger distance between them is defined as in eqn(3)

$$h(X, Y) = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{X} - \sqrt{Y} \right\|_2 \qquad (4)$$

By definition, the Hellinger distance is a metric that satisfies triangle inequality. The $\sqrt{2}$ in the definition is to assure that $h(X, Y) \le 1$ for all probability distributions. Thus the cluster assignment is determined as follows. Let a segment of points y, the clustering rule can be represented as the adjacency matrix which is given in eqn(4)

$$A_{ij} = \begin{cases} \forall y & on\ the\ line\ segment\ connecting\ x_i\ and\ x_j \\ 0 & otherwise \end{cases} \qquad (5)$$

Checks are performed on all data points for assigning a specific cluster. In addition, outliers are not classified since their feature space dwell outside the enclosing sphere.

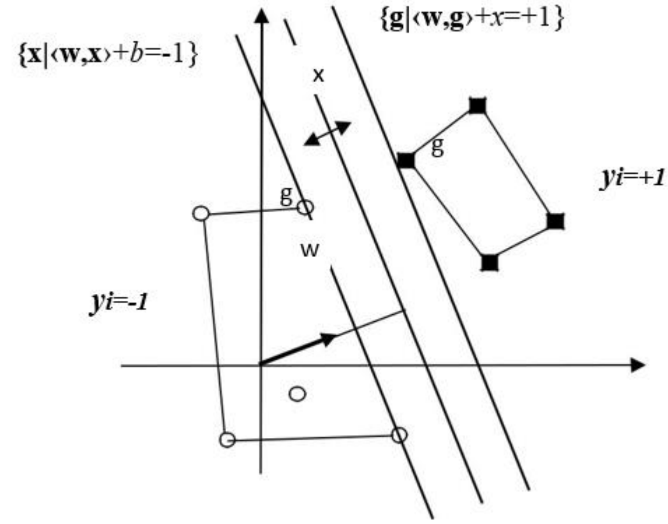


Fig. 3: The Optimal hyper-plane with normal Vector w and Offset x

Existing SVM based Clustering Algorithm:

```
Input: Extracted Web Log Files
Output: Clusters according to the query
Step1. SVM Initialized by running a Maximum Margin Linear SVM classifier against the dataset .
Step2. Each vector is randomly labelled
Step 3. Repeat upto initial convergence Occur
Step 4. Obtained SVM confidence parameter
Step 5. Labelling if the training data is done using $y_i=\pm1$
Step 6 . Build Hyperplane if $(w_{\beta.}\ g_{i\beta})$-x≥+1 for i such that $y_i$=+1 and $(w_{\beta.}\ g_{i\beta})$-x≤-1 for some values of co-efficientsw$_1$…….w$_n$ and x
Step 7. Then Datapoints (Support Vectors) that satisfies the equality is Obtained
Step 8. . Lagrangian formulation applied and saddle point fixed and maximized.
Step 9. Transform “g” to a high dimensional  space
Step10. Use Euclidian distance and triangle inequality satisfied
Step 11. Check performed on all data points
Step 12. Construct a similarity Using K Means Clustering
Step13. Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.
Step 14. Generate a new partition by assigning each pattern to its closest cluster center.
Step 15. Compute new cluster centers.
Step 16. Cluster the points $(y_i)_i$=1…n in $R^k$
Step 17.  Clusters are obtained
```

An Hybridiztion of SVM based Spectral Clustering algorithm:

```
Input: Extracted Web log Files
Output: Clusters according to the query
Step 1: ESVM  Initialized by running a binary SVM classifier against the dataset
Step 2: Each vector is randomly labelled
Step 3.Repeat upto initial convergence Occur
Step 4. Obtained ESVM confidence parameter
Step 5. Labelling of the training data is done using $y_i=\pm1$
Step 6. Build Hyperplane if $(w_{\beta.}\ g_{i\beta})$-x≥+1 for i such that $y_i$=+1 and $(w_{\beta.}\ g_{i\beta)}$-x≤-1for some values of co-efficientsw$_1$….w$_n$ and x
Step 7. Then Datapoints (Support Vectors) that satisfies the equality is obtained
Step 8. Lagrangian formulation applied and saddle point fixed and maximized.
Step 9. Transform “g” to a high dimensional  space
Step 10. Use Helinger distance and triangle inequality satisfied
Step 11. Check performed on all data points
Step 12. Construct a similarity Using Spectral Clustering Technique.
Step 13. Let ‘w be its weighted adjacency matrix.
Step 14. Compute the unnormalized Laplacian L= $D^{1/2}\ LD^{1/2} = I - D^{-1/2}AD^{-1/2}$.
 Step 15. Compute the first k generalized eigenvectors u1,...,uk of the generalized Eigen problem Lu=λDu
Step 16.  Let U ∈ R $^{n\times k}$ be the matrix containing the vectors u1,...,uk as columns.
Step 17. For i = 1,...,n, let $y_i$ ∈R $^k$ ,be the vector corresponding to the i-th row of U.
Step18. Cluster the points $(y_i)_i$=1,...,n in $R^k$
Output: Clusters A1,...,Ak with Ai ={j|yj ∈Ci}.
Clusters are obtained
```

The personalized hybrid SVM clustering algorithm iteratively performs the merging of the most similar pair of query points, and then the most similar pair of concept points, and then the merge of the most similar pair of query points, and so on. In the Spectral clustering algorithm, which denotes similar queries submitted by different users by one query point, there is again a necessity to take into consideration about the similar queries submitted by various users separately for the achievement of personalization effect. Otherwise stated, if two given queries, even if they are identical or not, give diverse meanings to two dissimilar users, and their merging should not be done because they are conceptually two distinct sets for the two users. Therefore, each individual query submitted by each user should be treated as an individual vertex in the bipartite graph by labeling each query with a user identifier. After the personalized bipartite graph is created, the initial experiments showed that if the algorithm is directly applied on the bipartite graph, the query clusters which are returned will quickly merge the queries from different users, hence leading to a loss in the personalization effect. Then it was found that identical queries, even though are issued by different users and have various meanings, try to have some general concept points such as ‘information’ in common.

Experiments And Results:

The real-time dataset is collected for the institution domain and the svm based clustering technique is used to cluster the data which is to infer that 1. The most visited websites for the user’s query interest and also 2. The most time spent in the particular website for a particular query. Thus the improved svm based clustering algorithm is applied for the dataset of 1, 00,000 data. The Algorithm is implemented in Pyspark in the spark environment. The Log files in different web servers maintain different types of information. The basic information present in the log file are

• User name:

This identifies who had visited the web site. The identification of the user mostly would be the IP address that is assigned by the Internet Service provider (ISP). This may be a temporary address that has been assigned. Therefore here the unique identification of the user is lagging. In some web sites the user identification is made by getting the user profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified.

Visiting Path:

The path taken by the user while visiting the web site. This may be by using the URL directly or by clicking on a link or through a search engine.

• Path Traversed:

This identifies the path taken by the user with in the web site using the various links.

• Time stamp:

The time spent by the user in each web page while surfing through the web site. This is identified as the session.

• Page last visited:

The page that was visited by the user before he or she leaves the web site.

• Success rate:

The success rate of the web site can be determined by the number of downloads made and the number copying activity under gone by the user. If any purchase of things or software made, this would also add up the success rate.

• User Agent:

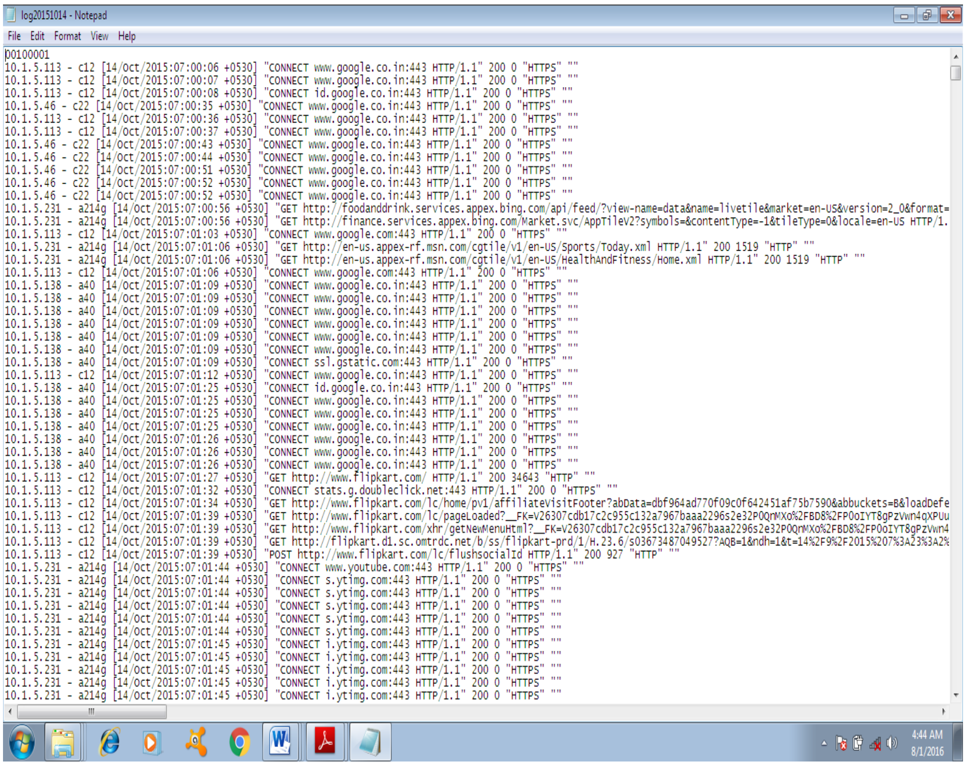
This is nothing but the browser from where the user sends the request to the web server. It’s just a string describing the type and version of browser software being used.

• URL:

The resource accessed by the user. It may be an HTML page, a CGI program, or a script.

• Request type:

The method used for information transfer is noted. The methods like GET, POST. These are the contents present in the log file. This log file details are used in case of web usage mining process. According to web usage mining it mines the highly utilized web site. The utilization would be the frequently visited web site or the web site being utilized for longer time duration. Therefore the quantitative usage of the web site can be analyzed if the log file is analyzed.



```
00100001
10.1.5.113 - c12 [14/Oct/2015:07:00:06 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.113 - c12 [14/Oct/2015:07:00:07 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.113 - c12 [14/Oct/2015:07:00:08 +0530] "CONNECT id.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.46 - c22 [14/Oct/2015:07:00:35 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.113 - c12 [14/Oct/2015:07:00:36 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.113 - c12 [14/Oct/2015:07:00:37 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.46 - c22 [14/Oct/2015:07:00:43 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.46 - c22 [14/Oct/2015:07:00:44 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.46 - c22 [14/Oct/2015:07:00:51 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.46 - c22 [14/Oct/2015:07:00:52 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.46 - c22 [14/Oct/2015:07:00:52 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:00:56 +0530] "GET http://foodanddrink.services.appex.bing.com/api/feed/?view-name=data&name=livetile&market=en-US&version=2_0&format=
10.1.5.231 - a214g [14/Oct/2015:07:00:56 +0530] "GET http://finance.services.appex.bing.com/Market.svc/AppTileV2?symbols=&contentType=-1&tileType=0&locale=en-US HTTP/1.
10.1.5.113 - c12 [14/Oct/2015:07:01:03 +0530] "CONNECT www.google.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:06 +0530] "GET http://en-us.appex-rf.msn.com/cgtile/v1/en-US/Sports/Today.xml HTTP/1.1" 200 1519 "HTTP" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:06 +0530] "GET http://en-us.appex-rf.msn.com/cgtile/v1/en-US/HealthAndFitness/Home.xml HTTP/1.1" 200 1519 "HTTP" ""
10.1.5.113 - c12 [14/Oct/2015:07:01:06 +0530] "CONNECT www.google.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:09 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:09 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:09 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:09 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:09 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:09 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:09 +0530] "CONNECT ssl.gstatic.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.113 - c12 [14/Oct/2015:07:01:12 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:25 +0530] "CONNECT id.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:25 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:25 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:25 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:26 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:26 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.138 - a40 [14/Oct/2015:07:01:26 +0530] "CONNECT www.google.co.in:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.113 - c12 [14/Oct/2015:07:01:27 +0530] "GET http://www.flipkart.com/ HTTP/1.1" 200 34643 "HTTP" ""
10.1.5.113 - c12 [14/Oct/2015:07:01:32 +0530] "CONNECT stats.g.doubleclick.net:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.113 - c12 [14/Oct/2015:07:01:34 +0530] "GET http://www.flipkart.com/lc/home/pv1/affiliateVisitFooter?abData=dbf964ad770f09c0f642451af75b7590&abbuckets=B&loadDefe
10.1.5.113 - c12 [14/Oct/2015:07:01:39 +0530] "GET http://www.flipkart.com/lc/pageLoaded?__FK=V26307cdb17c2c955c132a7967baaa2296s2e32P0QnMXo%2FBD8%2FP0oIYT8gPzVwn4qXPUu
10.1.5.113 - c12 [14/Oct/2015:07:01:39 +0530] "GET http://www.flipkart.com/xhr/getNewMenuHtml?__FK=V26307cdb17c2c955c132a7967baaa2296s2e32P0QnMXo%2FBD8%2FP0oIYT8gPzVwn4
10.1.5.113 - c12 [14/Oct/2015:07:01:39 +0530] "GET http://flipkart.d1.sc.omtrdc.net/b/ss/flipkart-prd/1/H.23.6/s03673487049527?AQB=1&ndh=1&t=14%2F9%2F2015%207%3A23%3A2%
10.1.5.113 - c12 [14/Oct/2015:07:01:39 +0530] "POST http://www.flipkart.com/lc/flushsocialId HTTP/1.1" 200 927 "HTTP" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:44 +0530] "CONNECT www.youtube.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:44 +0530] "CONNECT s.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:44 +0530] "CONNECT s.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:44 +0530] "CONNECT s.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:44 +0530] "CONNECT s.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:44 +0530] "CONNECT s.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:45 +0530] "CONNECT i.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:45 +0530] "CONNECT i.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:45 +0530] "CONNECT i.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:45 +0530] "CONNECT i.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
10.1.5.231 - a214g [14/Oct/2015:07:01:45 +0530] "CONNECT i.ytimg.com:443 HTTP/1.1" 200 0 "HTTPS" ""
```

Fig. 4: The Web Log dataset collected from an institution

Thus the most visited webpage and the most time spent webpage is found for the institution domain using the improved SVM based clustering technique and the results are inferred after implementing the hybrid SVM based clustering algorithm in Python.

Table 1: The Web Log Dataset collected from An Institution.

IP ADDRESS	QUERY/KEYWORD	THE MOST CLICK WEBSITE	LOGIN TIME	LOGGED OUT	TIME SPENT
10.1.5.100	Ad exchange	Doubleclickbygoogle.com	[14/Oct/2015:17:10:15 +0530]	[14/Oct/2015:17:26:33 +0530]	16:18
10.1.5.106	Dfp small business	Doubleclickbygoogle.com	[14/Oct/2015:18:14:53 +0530	[14/Oct/2015:18:29:44 +0530	15:11
10.1.5.108	Doubleclick for publishers	Doubleclickbygoogle.com	[14/Oct/2015:17:35:26 +0530]	[14/Oct/2015:18:28:24 +0530]	52:58
10.1.5.109	Doubleclickfor publishers	Doubleclickbygoogle.com	[14/Oct/2015:17:53:07 +0530]	[14/Oct/2015:18:28:10 +0530]	35:3
10.1.5.110	Doubleclickfor publishers	Doubleclickbygoogle.com	[14/Oct/2015:18:12:28 +0530]	[14/Oct/2015:18:24:35 +0530]	12:7
10.1.5.111	How to purchase books?	"CONNECT www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:17:20:52 +0530]	[14/Oct/2015:17:48:48 +0530]	27:56
10.1.5.113	How to purchase books?	www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:17:01:32 +0530]	[14/Oct/2015:18:13:01 +0530]	11:29
10.1.5.113	Handbag	www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:17:03:05 +0530]	[14/Oct/2015:07:24:25 +0530]	21:20
10.1.5.94	How to Get Books?	www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:17:21:05 +0530]	[14/Oct/2015:18:20:36 +0530]	59:31
10.1.5.75	How to get books?	www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:17:27:49 +0530]	[14/Oct/2015:17:37:07 +0530]	10:42

10.1.5.46	Get Books	www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:07:27:57 +0530]	[14/Oct/2015:07:31:19 +0530]	4:26
10.1.5.209	a book	"CONNECT www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:17:45:48 +0530	[14/Oct/2015:18:18:09 +0530	32:27
10.1.5.207	a book	"CONNECT www.amazon.in:443 HTTP/1.1"	14/Oct/2015:07:51:11 +053	[14/Oct/2015:08:03:00 +053	11:49
10.1.5.113	How to purchase a book?	"CONNECT www.amazon.in:443 HTTP/1.1"	[14/Oct/2015:07:03:05 +0530]	[14/Oct/2015:07:24:27 +0530]	21:22

From the Table 1 it is inferred that the most clicked website link that is doubleclick.net for the query doubleclickfor publishers reveals the most stayed in time in different systems in which the user from the system IP 10.1.108 had spent more time in doubleclick.net by Google search engine. And the other keyword related query is amazon and the most clicked link is "GET http://www.amazon.in/ . If the newly developed search engine application is run on the system and any user gets in for search with the query doubleclickfor publishers reveals the rank one.

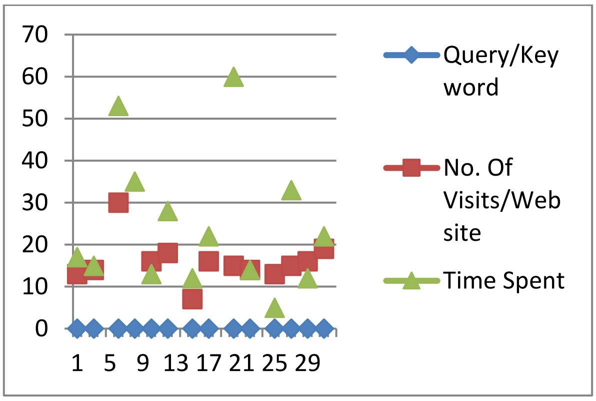


Fig. 5: The Clustered Web Logs

The Fig 5 Shows the most clicked website by the number of visits the user had for a website link and the time spent shows the most spent time as of here in the fig 5 for the query “How to Get Books?” is most the most time spent website after the results are clustered.

Table 2: The number of clusters on the institution web log database using KMSVM.

Methods	SVM			KMSVM			
Compression rate	1	10	20	30	40	50	60
Number of SVs	1450	324	190	137	126	99	83
Response time (s)	17.0	5.9	4.6	4.1	3.9	3.6	3.6
Testing accuracy (%)	95.5	93.4	92.53	91.88	91.98	91.38	90.88

The CR = 1 means that the classifier is built by SVM and CR = 10 … 60 means that the classifiers are built by the KMSVM algorithm with different numbers of clusters

Table 3: The number of clusters on the institution web log database using SCSVM.

Methods	SVM			SCSVM			
Compression rate	1	10	20	30	40	50	60
Number of SVs	1450	323	171	132	124	96	76
Response time (s)	17.0	5.4	4.5	4.0	3.6	3.3	3.3
Testing accuracy(%)	95.5	92.4	91.53	90.88	90.09	88.98	86.99

The CR = 1 means that the classifier is built by SVM and CR = 10 … 60 means that the classifiers are built by the SCSVM algorithm with different numbers of clusters

Conclusion:

The SVM based spectral clustering techniques overcome the SVM based k-means clustering. The spectral clustering technique aligns the data in adjacency matrix and the binary classifiers instead of maximum margin are used for the web recommender system based on the browsing history of the web users. Thus the proposed system hybrids the spectral clustering algorithm with the SVM Classification technique gives a better response time and better accuracy than the KMSVM. The user profile is clustered for the number of clicks and the time spent most in the website for a particular query. The SVM based spectral clustering gives the better result than the SVM based k-Means clustering. Further the work can be extended with all other clustering techniques. Thus the SVM based clustering can give better result for minimizing the support vectors and also enhance the response time and accuracy of the algorithm.

REFERENCES

Alsmadi, Izzat and Ikdam Alhami, 2015 Clustering and classification of email contents." Journal of King Saud University-Computer and Information Sciences, 27(1): 46-57.

Brin, S. and L. Page, 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proc. of the 7th International World Wide Web Conference (WWW7), pp: 107-117.

Carrizosa, Emilio, Amaya Nogales-Gómez, and Dolores Romero Morales, 2016 Clustering categories in support vector machines.

Ding, Shifei, *et al*. 2016. Recent advances in Support Vector Machines. *Neurocomputing*.

D'Orangeville, Vincent, *et al*. 2013. Efficient Cluster Labeling for Support Vector Clustering." IEEE Transactions on Knowledge and Data Engineering, 25(11): 2494-2506.

Filippone, Maurizio, *et al*. 2008. A survey of kernel and spectral methods for clustering. *Pattern recognition,* 41(1): 176-190.

Finley, Thomas, and Thorsten Joachims, 2005 Supervised clustering with support vector machines. *Proceedings of the 22nd international conference on Machine learning*. ACM.

Jahanshahi, S.M.A., A. Habibi Rad and V. Fakoor, 2016. A Goodness-of-Fit Test for Rayleigh Distribution Based on Hellinger Distance. Annals of Data Science, pp: 1-11.

Khabia, Apeksha and M.B. Chandak, 2014. A Cluster Based Approach for Classification of Web Results." International Journal of Advanced Computer Research, 4(4): 934.

Kumar, A., and P. Thambidurai, 2010. Collaborative web recommendation systems based on an effective fuzzy association rule mining algorithm (FARM)." Indian Journal of Computer Science and Engineering, 1(3): 184-191.

Lee, Gyemin, 2015. Hierarchical Clustering Using One-Class Support Vector Machines. Symmetry, 7(3): 1164-1175.

Simpson, Douglas G., 1987 Minimum Hellinger distance estimation for the analysis of count data." Journal of the American statistical Association, 82(399): 802-807.

Spiliopoulou, M. and L. Faulstich, 1998. WUM–A Tool for WWW Utilization Analysis. In Proc. of the International Workshop on the World Wide Web and Databases (WebDB'98), pp: 184-203.

Sugiyama, Kazunari, *et al*. 2004. User-Oriented Adaptive Web Information Retrieval Based on Implicit Observations. *Asia-Pacific Web Conference*. Springer Berlin Heidelberg.

Sugiyama, Kazunari, Kenji Hatano, and Masatoshi Yoshikawa. 2004. Adaptive web search based on user profile constructed without any effort from users. *Proceedings of the 13th international conference on World Wide Web*. ACM.

Wang, Jiaqi, Xindong Wu, and Chengqi Zhang, 2005. Support vector machines based on K-means clustering for real-time business intelligence systems."International Journal of Business Intelligence and Data Mining, 1(1): 54-64.

Wang, Xiang-Yang, *et al*. 2016. A new SVM-based relevance feedback image retrieval using probabilistic feature and weighted kernel function. Journal of Visual Communication and Image Representation, 38: 256-275.

Zheleva, Elena, and Lise Getoor, 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles." *Proceedings of the 18th international conference on World wide web*. ACM.