International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 10138-10144
© Research India Publications. http://www.ripublication.com

Adaptive Feature Selection Based Improved Support Vector Machine Classifier Using Adaboost and Genetic Algorithm for Web Interaction Mining

B. Kaviyarasu
Research Scholar, PG and Research Department of Computer Applications,
Hindusthan College of Arts and Science, Coimbatore – 38, Tamil Nadu, India.

Dr. A. V. Senthil Kumar
Director, PG and Research Department of Computer Applications,
Hindusthan College of Arts and Science, Coimbatore – 38, Tamil Nadu, India.

Abstract

Foreseeing the purpose of internet users has different applications in the areas such as e-commerce, entertainment in online, and several internet-based applications. The significant part of the classifying internet queries are based on available features namely contextual information, keywords and their semantic relationships. This research paper presents adaptive feature selection based improved support vector machine classifier that makes use of ad boost genetic algorithmic approach towards web interaction mining. Around 31 participants are chosen and given topics to search web contents. Parameters such as precision and F1 score are taken for comparing the proposed classifier with the classical support vector machine and improved support vector machine. Results proved that the proposed classifier achieves better performance than that of the conventional SVM and improved SVM [16].

Keywords: Web interaction mining, algorithm, adaptive feature selection, support vector machine, classifier, ad boost, genetic algorithm.

INTRODUCTION

Web mining is the use of data mining strategies to add learning from web data, together with web archives, hyperlinks between records, use logs of web locales, and numerous others. Web mining is the withdrawal of possibly significant examples and understood comprehension from interest identified with the webpage. This extricated learning will likewise be additional used to improve web use with the end goal that prediction of consequent page liable to got to through purchaser, crime detection and future prediction, individual profiling and to perceive about individual seeking pastimes [Monika Dhandi, Rajesh Kumar Chakrawarti.,2016] [8].

Web Mining can be exhaustively confined into three specific classes, as showed by the sorts of data to be mined. The survey of the three arrangements of web mining [T. Srivastava et al.,2013] [11] examined underneath are (1) Web Content Mining (2) Web Structure Mining (3) Web Interaction Mining.

1.1. Web Content Mining (WCM): WCM is the route toward removing accommodating data from the substance of web chronicles. Portrayed data identifies with the get-together of assurances of a web page were planned to pass on to the customers. It may include content, pictures, sound, video, or sorted out records, for instance, records and tables.

1.2. Web Structure Mining (WSM): The structure of a particular web includes Web pages as hubs, and web connect as edges partner related pages. Web Structure Mining is the path toward discovering structure data from the Web. This can be additionally parceled into two sorts in perspective of the kind of structure data used.

(a) Hyperlinks: A Hyperlink is an essential unit that interfaces a zone in a page to emerge area, either inside the indistinct web page or on a substitute page.

(b) Document Structure: Besides, the substance inside a page will in like manner be created in a tree-composed structure, headquartered on the more than a few HTML and XML marks inside the website page. Mining attempts right have fascinated without a doubt by isolating record question display (DOM) structures out of archives.

1.3. Web Interaction Mining (WIM): WIM is the utilization of data mining systems to discover captivating use plans from Web data, with a particular ultimate objective to grasp and better serve the prerequisites of Web-based applications. Utilization of data gets the character or

wellspring of web customers close by their examining conduct at a webpage. WUM itself can be gathered further dependent upon the kind of utilization data considered:

(a) Web Server Data: The customer logs are accumulated by Web server. Little scope of the data fuses IP address, page reference and get the opportunity to time.

(b) Application Server Data: Business application servers, for instance, Web-rationale, Story-Server have essential parts to engage E-trade applications to be founded over them with little effort. A key part is the ability to track diverse sorts of business events and log them in application server logs.

(c) Application Level Data: New sorts of events can be portrayed in an application, and logging can be turned on for them - creating histories of these remarkably described events.

In earlier works [17] and [18] improved extreme learning machine classifier and penta-layered artificial neural networks are developed for web interaction mining. In this phase of research an adaptive feature selection is employed which aims to improve the performance of classifier in terms of precision and F-1 score.

This paper is organized as follows. This section gives a brief introduction about the research. Section 2 portrays the related works carried out. Section 3 emphasizes the proposed work. Section 4 discusses on results. Section 5 presents concluding remarks.

RELATED WORKS

T. Cheng et al.,2013 [9] have given three data offerings: element equivalent word data transporter, question to-substance data administration and element labeling learning supplier. The element equivalent word benefit used to be an in-creation learning bearer that used to be by and by accessible while the other two are data benefits by and by in advance at Microsoft. Their investigations on item datasets show (i) these information offerings have unnecessary best and (ii) they've massive impact on shopper encounters on e-rear web destinations.

M. Nayrolles and A. Hamou-Lhadj.,2016 [7] proposed BUMPER (BUg Metarepository for dEvelopers and Researchers), a standard framework for engineers and specialists curious about mining data from numerous (heterogeneous) vaults. Guard used to be an open supply web-established condition that concentrates data from an assortment of BR stores and variation control frameworks. It was once outfitted with a solid web index to help clients rapidly inquiry the vaults using a solitary purpose of get to X.

Ye et al.,2015 [12] creators proposed another considering technique by methods for a summed up misfortune capacity to catch the unobtrusive significance varieties of preparing tests when an additional granular name constitution was once close by. Creators have used it to the Xbox One's film look mission the place session-headquartered individual direct comprehension was once to be had and the granular importance contrasts of instructing tests are gotten from the session logs. At the point when put next with the predominant technique, their new summed up misfortune work has tried complex trial effectiveness measured by methods for a couple of buyer engagement measurements.

The reason for T. F. Lin and Y. P. Chi.,2014 [10] was to make utilization of the connected sciences of TF-IDF, alright approach bunching and ordering superb examination to set up the combo of key expressions to have the capacity to advantage website design enhancement. The learn showed that it may likely easily upgrade the web website's advancement of positioning on web index, increment web webpage's attention level and tap on through cost.

G. Dhivya et al.,2015 [3] dissected individual lead by utilizing mining advanced web section log data. The few net interaction mining approaches for extricating profitable components used to be talked about and utilize every one of these systems to bunch the clients of the area to consider their practices extensively. The commitments of this proposal are a data enhancement that was substance and beginning spot arranged and a treelike perception of bland navigational groupings. This representation makes it feasible for an advantageously interpretable tree-like perspective of examples with featured essential know-how.

Z. Liao et al.,2014 [15] presented "task trail" to comprehend client look practices. Creators layout a mission to be a nuclear individual know-how need, while a test trail speaks to all individual interests inside that exact venture, comparable to address reformulations, URL clicks. Beforehand, net inquiry logs have been considered all things considered at session or question arrange the place clients may set up a few inquiries inside one wander and deal with a few assignments inside one session.

A. Yang et al.,2014 [2] have granted an answer that initially distinguishes the clients whose kNN's conceivably tormented by the recently arrived content, after which supplant their kNN's individually. Creators proposed another file constitution named HDR-tree keeping in mind the end goal to support the compelling hunt of influenced clients. HDR-tree proceeds with dimensionality decrease through grouping and guideline component assessment (PCA) in order to make

more grounded the pursuit adequacy. To additional scale back reaction time, creators proposed a variation of HDR-tree, known as HDR-tree, that helps additional powerful however surmised arrangements.

A. U. R. Khan et al.,2015 [5] have exhibited a cloud transporter to clarify how the status of the broad communications news can be evaluated using clients online use propensities. Creators utilized information from Google and Wikipedia for this correlation challenge. Google data was useful in comprehension the affect of stories on web looks while data from Wikipedia empowered us to comprehend that articles identified with rising data content additionally discover parcel of consideration.

J. Jojo and N. Sugana.,2013 [4] proposed a half breed approach which utilizes the insect established grouping and LCS order techniques to search out and foresee client's route conduct. Subsequently client profile may likewise be followed in powerful pages. Customized inquiry can be utilized to address extend in the web look group, established on the preface that a purchaser's ordinary decision may simply help the mission motor disambiguate the genuine aim of an inquiry.

M. A. Potey et al.,2013 [6] inspected and contrasted the with be had ways to deal with display an understanding into the train of question log handling for ability recovery.

A. Vinupriya and S. Gomathi.,2016 [1] proposed a fresh out of the box new plan named as WPP (web page Personalization) for effective net page recommendations. WPP comprise of page hit depend, finish time spent in every hyperlink, number of downloads and connection detachment. Established on these parameters the personalization has been proposed. The system proposes a fresh out of the plastic new verifiable client input and occasion hyperlink get to plans for astonishing web page customization together with area philosophy.

Y. C. Fan et al.,2016 [14] proposed an information cleansing and understanding enrichment framework for enabling consumer alternative working out by way of Wi-Fi logs, and introduces a sequence of filters for cleansing, correcting, and refining Wi-Fi logs.

Y. Kiyota et al.,2015 described learn how to construct a property search habits corpus derived from micro blogging timelines, in which tweets concerning property search are annotated. Authors applied micro task-established crowd sourcing to tweet knowledge, and construct a corpus which contains timelines of special customers that are annotated with property search phases**.**

PROPOSED WORK

Adaptive Feature Selection :

In this adaptive feature selection method, features are ranked and then sorted in descending order by feature selection methods in each feature vector respectively. Once feature ranking is carried out, collection-based features vector (CFV) is obtained. The process of obtaining the CFV and feature subset is given below.

Step 1: Create feature vectors. Let $F=\{f_1,f_2,...,f_N\}$ presents a set of features. Where, N is total number of features and f_i is a feature that can ranks by different feature selection methods, namely $M_1,M_2,...,M_L$. For creating a feature vectors (FV), first, feature are weighted and then features are sorted descending order according to their weight. In feature vector of $FV_j=\left[f_{i1}{}^j,f_{i2}{}^j,...,f_{iN}{}^j\right]$ that created by j th feature selection method, $f_{i1}{}^j$ is a permutation of $\{f_1,f_2,...,f_N\}$.

$$F=\{f_1,f_2,...,f_N\}\rightarrow FV=[x_1,x_2,...,x_N]\ldots (1)$$

Step 2: Integration of FVs. In this step, feature vectors are integrated in order to new feature ranking based on the Equation 1. A new feature ranking is defined as follows:

$$New\ ranking\ of\ (f'_1,f'_2,...,f'_N)=\begin{cases}Rank(f_i^{'}=\sum_{j=1}^{M}indexFV_j(x_i)\\ indexFV_j(x_i)=Place\ of\ x_i\ in\ FV_j\end{cases} ..(2)$$

Where N is number of features. After feature ranking, features are sorted descending order according to their weight in order to create CFV.

Step 3: Generation and evaluation feature subsets. After feature ranking based on collection-based integration, different feature subsets are generated as follows:

$$OFV=[x_1,x_2,...,x_N],\forall_{i,j}\ i<j\rightarrow rank(x_i)\geq rank(x_j)$$

$$Feature\ subsets=\{\{x_1\},\{x_1,x_2\},\{x_1,x_2,x_3\},...,\{x_1,x_2,...,x_N\}\}\quad (3)$$

Where x_i is a feature and N is total number of features.

In this representation, x_1 has the highest rank (or weight) and x_2 has the second highest rank among the feature vectors.

The algorithm is presented below.

Algorithm 1. Adaptive Feature Selection

Input: Web review dataset

Output: Confident features

Create and weight web searches

For pass = 1 : numRepetitions

Initialize first-fold on samples with a start random

For fold = 1 : numKfold

Find training and testing features sets from samples

Rank training-feature set and then create different feature vectors as fellow:

For i = 1 : numFeatureRankingMethods

Apply i^{th} Feature ranking method on training set

Create i^{th} Feature vector by sorting in descending order

End i

Collection-based integration of different feature vectors (called CFV)

Generate feature subsets incrementally based on Equation 3 on CFV

Evaluate different feature subsets:

For wrap = 1 : numFeatureSubsets

Partition web searchers based on number of features

Classification()

End wrap

Save feature subset with highest accuracy value

Adjust next fold

End fold

End pass

In the above algorithm, the number of repetition and folds are constant. The CFV is a vector scored by integrating the ranked feature vectors obtained using adaptive feature selection methods. The main advantages of this method is the reduction in the dependency of the feature vectors. It is to be noted that if the distances between the value of features in CFV vector are low, then this vector will be the best because it means all the feature selection methods selected the feature with a sequence.

Improved Support Vector Machine Classifier with Adaboost Genetic Algorithmic Approach :

The AdaBoost (adaptive boosting) algorithm is one of the most popular ensemble methods. It creates a collection of moderate classifiers by maintaining a set of weights over training data and adjusting these weights after each learning cycle adaptively. The weights of the training samples which are correctly classified by current classifier will decrease while the weights of the samples which are misclassified will increase. Since the proposed work has the scope to incorporate AdaBoost algorithm, the standard SVM needs to be extended to the SVM for which each training sample has different weights. Now the SVM model is transformed to

$$\min \frac{1}{2}\| \omega \|^2 + C\sum_{i=1}^{t} \mu_i \xi_i$$
$$s.t \quad y_i ((\omega.\phi(x_i)) + b) \geq 1 - \xi_i \qquad \ldots (4)$$
$$\xi_i \geq 0, i = 1,2,....,l$$

where (1,2,..., l) indicates the weight of the sample x_i . And the dual problem is as

follow:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{i}\sum_{j=1}^{i} y_i y_j \alpha_i \alpha_j k(x_i.x_j) - \sum_{j=1}^{i} \alpha_j$$

$$s.t \sum_{i=1}^{i} y_i \alpha_i = 0 \qquad \ldots (5)$$

$$0 \le \alpha_i \le C, i = 1,2,....,l$$

Except for the weights of samples, the classification performance of the proposed Adaboost-SVM is equally affected by its model parameters. During the AdaBoost iterations, if parameters make the classification accuracy of ISVM less than 50%, the requirement on a component classifier in AdaBoost cannot be satisfied. In contrast, if the accuracy is too high, boosting classifiers may become inefficient because the errors of these component classifiers are highly correlated. Hence, how to select appropriate model parameters is important. There are many evolutionary algorithms for searching the suitable solution in real-valued spaces. With the advantages consisting of parallel search, solving complex problems, and large search space, the genetic algorithm (GA) is applied to perform the model parameters selection in the k-fold cross-validation set.

However, the process is time consuming and may cause the overfitting. Therefore, we adjust the model selection procedure so that the GA could be stopped when the cross-validation accuracy is over 0.5. In this result, the component classifiers conform to the condition of AdaBoost and the computational time could be saved. Moreover, the randomness of result produced by GA would decrease significantly after many times of Adaboost iterations. As a result, the outcomes corresponding to several independent runs of the hybrid method are similar to each other. So, the process of the proposed ISVM mechanism is relatively stable.

AdaBoost-GA-ISVM algorithm

Step1: Input - Training samples $\{(x_1, y_1),(x_2, y_2),...,(x_n, y_n)\}$, $x_i \ R^u$ and $y_i \ \{1,...,K\}$; Moderate classifier ISVM; The number of classes K; The total number of the iterations T.

Step 2: Initialize: The weights of training samples: w_i^l $1/n, i\ 1,...,n$; The GA parameters include size of population N ; Maximum number of generations $MaxI$; Length of chromosome of C and kernel parameters l; Crossover rate p_c and mutation rate p_m .

Step 3: For $t = 1,2,...,T.$

a) Select appropriate parameters

i) Encode the parameter C and kernel parameters as an l-bit string which consists of l_1 bits standing for C and l_2 bits standing for kernel parameters, here $l = l_1 + l_2$; Generate an initial population consisted of N strings of binary bit. To avoid trapping into same local optimum in GA process, the parameters are estimated starting from a completely new initial population for each t.

ii) For j = 1,2,..., N, obtain the jth group of parameters by decoding the string j and train a component multi-classifier ISVM g_i^j using these parameters on the k-fold cross-validation data set.

iii) Calculate the average cross-validation error: $E_t = \frac{1}{N}\frac{1}{n}\sum_{j=1}^{N}(\sum_{j=1}^{n} I(Y_i, g_t^j(x_i)))$, where the indicator function I produces 0 if the arguments are equal and 1 if they are different.

iv) If E_t <0.5 , do step v) else the parameters satisfy the requirement of AdaBoost

v) Perform reproduction: selection, crossover and mutation. Then,

a) Generate new offspring population. If the number of generation exceeds MaxI, it means that the t-th moderate classifier is invalid, then stop the GA iteration

b) Train a multi-classifier ISVM G_t using the suitable parameters from a) and obtain a probabilistic output vector probabilistic output vector

c) Compute the training error of G_t.

d) Set weight for the current classifier $G_t : \alpha_t = 0.5 In\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

e) Update the weights: $w_i^{t+1} = w_l^t \exp\{-\alpha_t\ y_i G_i(x_i)\}$

Step 4: Output: When the number of valid classifier reaches T the proposed algorithm is completed.

EXPERIMENTAL RESULTS

31 participants are taken in order to build the dataset for evaluating the proposed model. The people that are chosen belong to heterogeneous age groups and web experience; similar considerations apply for education, even though the majority of them have a computer science or technical background. All participants were requested to perform ten search sessions organized as follows:

- Four guided search sessions;
- Three search sessions in which the participants know the possible destination web sites;

- Three free search sessions in which the participants do not know the destination web sites.

This led to 129 sessions and 353 web searches, which were recorded and successively analyzed in order to manually classify the intent of the user according to the two-level taxonomy. Starting from web searches, 490 web pages and 2136 sub pages were visited. The interaction features were logged by the inbuilt YAR plug-in that is present in Google Chrome web browser.

For performing query classification, the proposed AFS - ISVM presumes that the queries in a user session are independent; Conditional Random Field (CRF) considers the sequential information between queries, whereas Latent Dynamic Conditional Random Fields (LDCRF) models the sub-structure of user sessions by assigning a disjoint set of hidden state variables to each class label.

In order to evaluate the effectiveness of the proposed model, we adopted the classical evaluation metrics of Information Retrieval: precision, recall, and F1-measure. In order to simulate an operating environment, 60% of user queries were used for training the classifiers, whereas the remaining 40% were used for testing them.

Precision*:* It is the fraction of retrieved web searchers that are relevant to the query which is calculated using (6).

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (6)$$

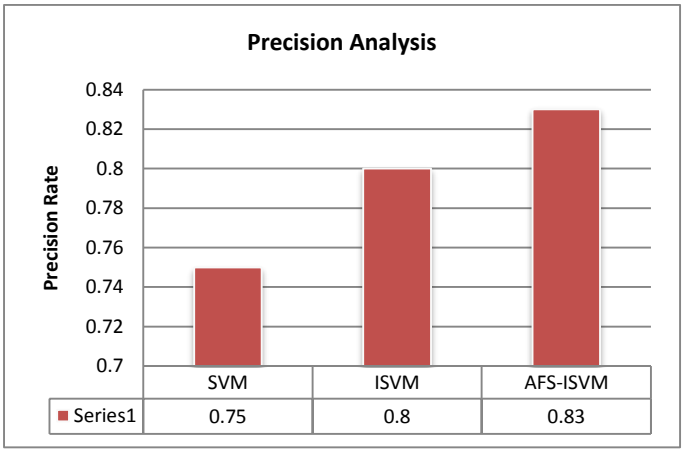

Figure 1: Comparison of Precision

F1 – Measure*:* F1 score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score. The F-1 measure is calculated using (7).

$$F1 = 2.\frac{precision.recall}{precision + recall} \quad \ldots (7)$$

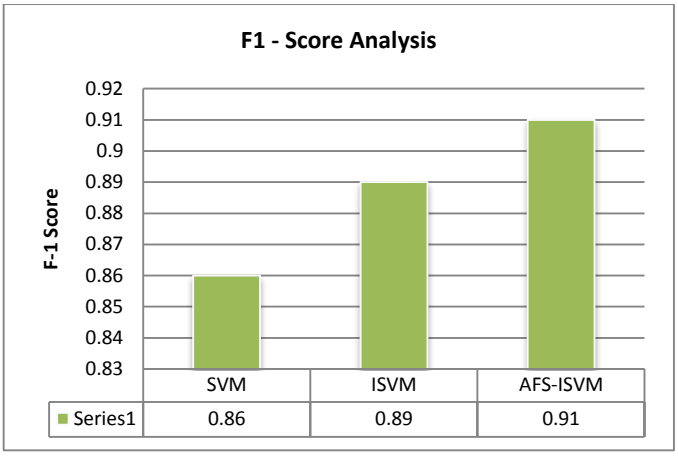

Figure 2: Comparison of F-1 Score

From the results it is evident that the proposed AFS-ISVM outperforms than that of SVM and ISVM classifiers. The precision rate is elevated to 0.83 and the F-1 score measure increases to 0.91. This is due to the incorporation of adaptive feature selection strategy added over ISVM. Due to the selected feature subset the number of false negatives are reduced which results in better performance.

CONCLUSIONS

This research work aims in design and development of adaptive feature selection based improved support vector machine classifier that makes use of adaboost genetic algorithmic approach towards web interaction mining. Feature selection is carried out at first and then with the help of that appropriate parameters are chosen with the help of the genetic algorithm. For improving the performance of the SVM classifier, Adaboost algorithm is employed. Performance metrics such as precision and F-1 score are chosen. From the results it is evident that the proposed AFS - ISVM outperforms SVM and ISVM classifiers.

REFERENCES

[1] A.Vinupriya and S. Gomathi, "Web Page Personalization and link prediction using generalized inverted index and flame clustering," 2016 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2016, pp. 1-8.

[2] A.Yang, X. Yu and Y. Liu, "Continuous KNN Join Processing for Real-Time Recommendation," 2014 IEEE International Conference on Data Mining, Shenzhen, 2014, pp. 640-649.

[3] G. Dhivya, K. Deepika, J. Kavitha and V. N. Kumari, "Enriched content mining for web applications," Innovations in Information, Embedded and

Communication Systems (ICIIECS), 2015 International Conference on, Coimbatore, 2015, pp. 1-5.

[4] J. Jojo and N. Sugana, "User profile creation based on navigation pattern for modeling user behaviour with personalised search," Current Trends in Engineering and Technology (ICCTET), 2013 International Conference on, Coimbatore, 2013, pp. 371-374.

[5] A.U. R. Khan, M. B. Khan and K. Mahmood, "Cloud service for assessment of news' Popularity in internet based on Google and Wikipedia indicators," Information Technology: Towards New Smart World (NSITNSW), 2015 5th National Symposium on, Riyadh, 2015, pp. 1-8.

[6] M. A. Potey, D. A. Patel and P. K. Sinha, "A survey of query log processing techniques and evaluation of web query intent identification," Advance Computing Conference (IACC), 2013 IEEE 3rd International, Ghaziabad, 2013, pp. 1330-1335.

[7] M. Nayrolles and A. Hamou-Lhadj, "BUMPER: A Tool for Coping with Natural Language Searches of Millions of Bugs and Fixes," 2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER), Suita, 2016, pp. 649-652.

[8] Monika Dhandi, Rajesh Kumar Chakrawarti, "A Comprehensive Study of Web Usage Mining", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), INDORE, India, 2016, Pages: 1 - 5.

[9] T. Cheng, K. Chakrabarti, S. Chaudhuri, V. Narasayya and M. Syamala, "Data services for E-tailers leveraging web search engine assets," Data Engineering (ICDE), 2013 IEEE 29th International Conference on, Brisbane, QLD, 2013, pp. 1153-1164.

[10] T. F. Lin and Y. P. Chi, "Application of Webpage Optimization for Clustering System on Search Engine V Google Study," Computer, Consumer and Control (IS3C), 2014 International Symposium on, Taichung, 2014, pp. 698-701.

[11] T. Srivastava, P. Desikan, V. Kumar, "Web Mining – Concepts, Applications and Research Directions", Studies in Fuzziness and Soft Computing Foundations and Advances in Data Mining, Springer Berlin Heidelberg, 2013, pp 275-307.

[12] X. Ye, Z. Qi, X. Song, X. He and D. Massey, "Generalized Learning of Neural Network Based Semantic Similarity Models and Its Application in Movie Search," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 86-93.

[13] Y. C. Fan, Y. C. Chen, K. C. Tung, K. C. Wu and A. L. P. Chen, "A framework for enabling user preference profiling through Wi-Fi logs," 2016 IEEE 32nd International Conference on Data Engineering (ICDE), Helsinki, Finland, 2016, pp. 1550-1551.

[14] Y. Kiyota, Y. Nirei, K. Shinoda, S. Kurihara and H. Suwa, "Mining User Experience through Crowdsourcing: A Property Search Behavior Corpus Derived from Microblogging Timelines," 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Singapore, 2015, pp. 17-21.

[15] Z. Liao, Y. Song, Y. Huang, L. w. He and Q. He, "Task Trail: An Effective Segmentation of User Search Behavior," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 12, pp. 3090-3102, Dec. 1 2014.

[16] B. Kaviyarasu, Dr. A. V. Senthil Kumar, "An Improved Support Vector Machine Classifier Using AdaBoost and Genetic Algorithmic Approach towards Web Interaction Mining", International Journal of Advanced Networking and Applications, vol.8, no.5, pp. 3201 – 3208, 2017.

[17] B. Kaviyarasu, Dr. A. V. Senthil Kumar, "Web Interaction Mining using Improved Extreme Learning Machine Classifier", International Journal of Research in Science Engineering and Technology, vol.3, no.12, pp.45 – 51, 2016.

[18] B. Kaviyarasu, Dr. A. V. Senthil Kumar, "Web Interaction Mining using Penta Layered Artificial Neural Network Classifier", International Journal of Computer Science Engineering and Technology, vol.3, no.1, pp.64 – 70, 2017.