Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 11 (2017), pp. 7695-7707
© Research India Publications
http://www.ripublication.com

Web Interaction Mining Using Adaptive Feature Selection Based Improved Extreme Learning Machine (AFS-IELM) Classifier

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

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

Abstract

Assessing the objective of internet users holds diverse applications in the areas such as e-commerce, entertainment in online, and several internet-based applications. The central section of classifying the internet queries based on available features such as contextual information, keywords and their semantic relationships carries significant research scope. This research article aims in proposing adaptive feature selection based improved extreme learning machine classifier shortly coined as AFS-IELM for web interaction mining. Around 31 participants are chosen and given topics to search web contents. Parameters such as precision, recall and F1 score are taken for comparing the proposed classifier with the ELM [16]. Results proved that the proposed classifier attains better performance than that of the conventional ELM.

Keywords: Web interaction mining, algorithm, extreme learning machine, classifier, precision, recall, F1-Score.

1. INTRODUCTION

Web mining is that the application data mining strategies to extract knowledge from web information, in conjunction with web documents, hyperlinks between records, usage logs of internet sites, and plenty of others. internet mining is that the withdrawal of doubtless valuable patterns and implicit understanding from interest associated with the location. This furthercted data will be extra wont to enhance internet utilization specified prediction of consequent page possible to accessed through shopper, crime detection and future prediction, person identification and to acknowledge concerning person looking out hobbies [Monika Dhandi, Rajesh Kumar Chakrawarti.,2016] [8].

Web Mining may be comprehensively isolated into three explicit categories, as indicated by the types of knowledge to be mined . The review of the three classifications of internet mining [T. Srivastava et al.,2013] [11] mentioned below square measure (1) web content Mining (2) web Structure Mining (3) web Interaction Mining.

WCM is that the approach toward extricating useful information from the substance of net archives. delineated data relates to the gathering of certainties of an online page were supposed to expire to the purchasers. it'd comprise of content, pictures, sound, video, or organized records, as an example, records and tables.

Web Structure Mining (WSM): The structure of a particular web includes of sites as nodes, and internet link as edges associating connected pages. internet Structure Mining is that the means toward finding structure knowledge from the net. This may be additional partitioned off into 2 kinds visible of the type of structure knowledge used.

Hyperlinks: A hyperlink could be a basic unit that interfaces a vicinity during a page to stand-out region, either within the indistinguishable web content or on an alternate page.

Document Structure: Moreover, the substance within a page can likewise be composed during a tree-organized structure, headquartered on the quite a few of hypertext markup language and XML labels within the web site page. Mining endeavors right have intrigued beyond any doubt by separating document object model (DOM) structures out of documents.

Web Interaction Mining (WIM): WIM is that the use data mining procedures to search out intriguing utilization styles from net information, with a selected finish goal to grasp and higher serve the necessities of Web-based applications. Use of knowledge catches the character or supply of net shoppers aboard their poring over conduct at a webpage. WUM itself are often classified any contingent upon the type of use info considered:

Web Server Data: The client logs square measure gathered by internet server. little vary of the data incorporates information processing address, page reference and acquire to time.

Application Server Data: Commercial application servers, for instance, Web-logic, Story-Server have noteworthy parts to empower E-trade applications to be supported high of them with very little labour. A key part is that the capability to trace different types of business occasions and log them in application server logs.

Application Level Data: New styles of occasions is characterised in associate application, and work is turned on for them - manufacturing histories of those unambiguously characterised occasions.

In earlier works [17] – [19] improved support vector machine, 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.

2. RELATED WORKS

T. Cheng et al.,2013 [9] have provided three info offerings: entity equivalent word info carrier, query-to-entity info service and entity tagging data supplier. The entity equivalent word service accustomed be AN in-creation data carrier that accustomed be presently accessible while the opposite 2 square measure info services presently current at Microsoft. Their experiments on product datasets exhibit (i) these data offerings have excessive best and (ii) they've large influence on client experiences on e-tailer websites.

M. Nayrolles and A. Hamou-Lhadj.,2016 [7] planned BUMPER (BUg Metarepository for dEvelopers and Researchers), a customary infrastructure for developers and researchers interested by mining data from several (heterogeneous) repositories. BUMPER accustomed be associate open provide web-founded atmosphere that extracts data from a spread of BR repositories and variant manipulate systems. it absolutely was once equipped with a robust computer programme to assist customers quickly question the repositories utilizing one purpose of access. X.

Ye et al.,2015 [12] authors planned a replacement learning technique by means that of a generalized loss perform to capture the delicate connectedness variations of coaching samples once a further granular label constitution was once existing. Authors have utilised it to the Xbox One's pic military mission the place session-headquartered person conduct understanding was once to be had and therefore the granular connectedness variations of work samples square measure derived from the session logs. once place next with the prevailing technique, their new generalized loss perform has tested subtle experiment potency measured by means that of many consumer-engagement metrics.

The purpose of T. F. Lin and Y. P. Chi.,2014 [10] was to create use of the applied sciences of TF-IDF, ok-approach bunch and compartmentalization high-quality examination to ascertain the jazz group of key phrases to be able to advantage seo. The learn incontestable that it'd most likely well enhance the net site's advancement of grading on program, increase web site's message level and click on on through expense.

G. Dhivya et al.,2015 [3] analyzed person conduct by mistreatment mining enriched internet entry log info. The few internet interaction mining approaches for extracting valuable components accustomed be mentioned and use of these methods to cluster the users of the domain to check their behaviors comprehensively. The contributions of this thesis square measure associate degree info enrichment that was content and origin set and a dendroid image of generic steering sequences. This image makes it attainable for a handily explicable tree-like read of patterns with highlighted primary power.

Z. Liao et al.,2014 [15] introduced “task trail” to grasp user search behaviors. Authors define a mission to be associate atomic person power wish, whereas a challenge path represents all person pursuits within that precise project, comparable to question reformulations, URL clicks. Previously, web search logs are studied by and enormous at session or question stage the place customers might place up many queries at intervals one venture and manage many tasks within one session.

A. Yang et al.,2014 [2] have awarded an answer that initial identifies the shoppers whose kNN's presumably suffering from the freshly arrived content, once that replace their kNN's severally. Authors planned a brand new index constitution named HDR-tree so as to support the effective search of affected customers. HDR-tree continues spatiality reduction through agglomeration and principle part analysis (PCA) thus on support the search effectiveness. To further cut back reaction time, authors planned a variant of HDR-tree, referred to as HDR-tree, that helps further effective however approximate solutions.

A. U. R. Khan et al.,2015 [5] have bestowed a cloud carrier to clarify however the standing of the mass media news are often assessed utilizing users on-line utilization habits. Authors used data from Google and Wikipedia for this comparison challenge. Google knowledge was useful in understanding the have a sway on of stories on internet searches whereas knowledge from Wikipedia enabled United States of America to grasp that articles associated with rising data content to boot realize ton of attention.

J. Jojo and N. Sugana.,2013 [4] planned a hybrid approach that uses the ant-founded clump and LCS classification ways to hunt out and predict user's navigation behavior.

As a result user profile can also be caterpillar-tracked in dynamic pages. customized search are often wont to address project within the net search community, based on the premise that a consumer's traditional alternative could aid the hunt engine elucidate the important intention of a matter.

M. A. Potey et al.,2013 [6] reviewed Associate in Nursingd compared the to be had approaches to gift an insight into the discipline of question log process for experience retrieval.

A. Vinupriya and S. Gomathi.,2016 [1] planned a fresh theme named as WPP (web page Personalization) for powerful internet page suggestions. WPP incorporates page hit swear, complete time spent in every link, variety of downloads and link separation. based on these parameters the personalization has been planned. The procedure proposes a fresh implicit user feedback and event link access schemes for superb net web content customization along side domain metaphysics.

Y. C. Fan et al.,2016 [14] projected associate info cleansing and understanding enrichment framework for enabling client different understanding by method of Wi-Fi logs, and introduces a sequence of filters for cleansing, correcting, and purification Wi-Fi logs.

Y. Kiyota et al.,2015 delineated find out how to construct a property search habits corpus derived from small blogging timelines, within which internet patterns regarding property search square measure annotated. Authors applied small task-established crowd sourcing to tweet information, and construct a corpus that contains timelines of special customers that square measure annotated with property search phases.

3. PROPOSED WORK

3.1. 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:

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

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:

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

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

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.

3.2. AFS based Improved Extreme Learning Machine Classifier

Once when the feature collection task is completed, AFS-IELM is employed for performing classification task. Given a set of N training samples (x_i, t_i) and 2L hidden neurons in total (that is, each of the two hidden layer has L hidden neurons) with the activation function g (x). At first randomly initialize the connection cost matrix between the input layer and the first hidden layer W and the bias matrix of the first hidden layer B, and then calculate the cost matrix β between the second hidden layer and the output layer.

$$g(W_H\, H + B_1) = H_1 \quad \text{... (4)}$$

where W_H denotes the cost matrix between the first hidden layer and the second hidden layer. It is presumed that the first and second hidden layers have the same number of neurons, and thus W_H is a square matrix. The notation H denotes the output between the first hidden layer with respect to all N training samples. The matrices B_1 and H_1 respectively represent the bias and the expected output of the second hidden layer.

The expected output of the second hidden layer can be calculated as

$$\mathrm{H}_1 = T\beta^+ \quad \dots (5)$$

where β† is the MP generalized inverse of the matrix β. The calculating method of β† is the same as previously discussed for H†, namely $\beta^\dagger = (\beta^\mathrm{T}\beta)^{-1}\beta^\mathrm{T}\ if\ \beta^\mathrm{T}\beta$ is nonsingular, or alternatively $\beta^\dagger = \beta^T(\beta^\mathrm{T}\beta)^{-1}\ if\ \beta\beta^T$ is nonsingular. Consequently it is defined the augmented matrix $W_{HE} = [B_1 W_H]$, and calculate it as

$$W_{HE} = g^{-1}(H_1)H_E^\dagger \dots (6)$$

where $H_E^\dagger$ is the generalized inverse of $H_E = [1H]^T$, 1 denotes a one-column vector of size N whose elements are the scalar unit 1, where the notation g(x) indicates the inverse of The calculation of $H^\dagger$ proceeds in the fashion described before. The experiments conducted to test the performance of the ELM algorithm. In order to perform the classification task extensively used logistic sigmoid function g (x) = $1/(1+e^{-x})$ is used. The actual output of the second hidden layer is calculated as

$$H_2 = g(W_{HE}\, H_E) \dots (7)$$

and finally, the cost matrix β_new between the second hidden layer and the output layer is calculated as

$$\beta_\mathrm{new} = H_2^\dagger T \quad \dots (8)$$

where $H_2^\dagger$ is the MP generalized inverse of H_2, obtained using the approach discussed before. The ELM output after training can be expressed as

$$\mathrm{f(x)} = H_2\, \beta_{new} \dots (9)$$

Algorithm 1. AFS-IELM Algorithm

Input: N training samples X = $[x_1, x_2, \dots\dots, x_N]^T$, T = $[t_1, t_2, \dots\dots, t_N]^T$ and 2L hidden neurons in total with activation function g (x)

1: Randomly generate the connection cost matrix between the input layer and the first hidden layer W and the bias matrix of the first hidden layer B and for simplicity, W_IE is defined as[B W] and similarly, X_E is defined as $[1\ X]^T$.

2: Calculate $H = g(W_{IE}\ X_E)$;

3: Obtain cost matrix between the second hidden layer and the output layer β = $\mathrm{H}^\dagger\mathrm{T}$

4: Calculate the expected output of the second hidden layer $\mathrm{H}_1 = \mathrm{T}\beta^{\dagger}$

5: Determine the parameters of the second hidden layer (connection cost matrix between the first and second hidden layer and the bias of the second hidden layer) $W_{HE} = g^{-1}(H_1)H_{\mathrm{E}}^{\dagger}$

6: Obtain the actual output of the second hidden layer $H_2 = g(W_{HE}\ H_E)$

7: Recalculate the cost matrix between the second hidden layer and the output layer $\beta_{new} = H_2^{\dagger}T$

Output: The final output of AFS-IELM is $f(x) = \{[W_H g(W\,X + B) + B_1]\}\beta_{new}$

4. 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-IELM 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, 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 documents that are relevant to the query which is calculated using (10).

$$precision = \frac{\left|\{relevant\ documents\} \cap \{retrieved\ documents\}\right|}{\left|\{retrieved\ documents\}\right|} \quad \ldots (10)$$

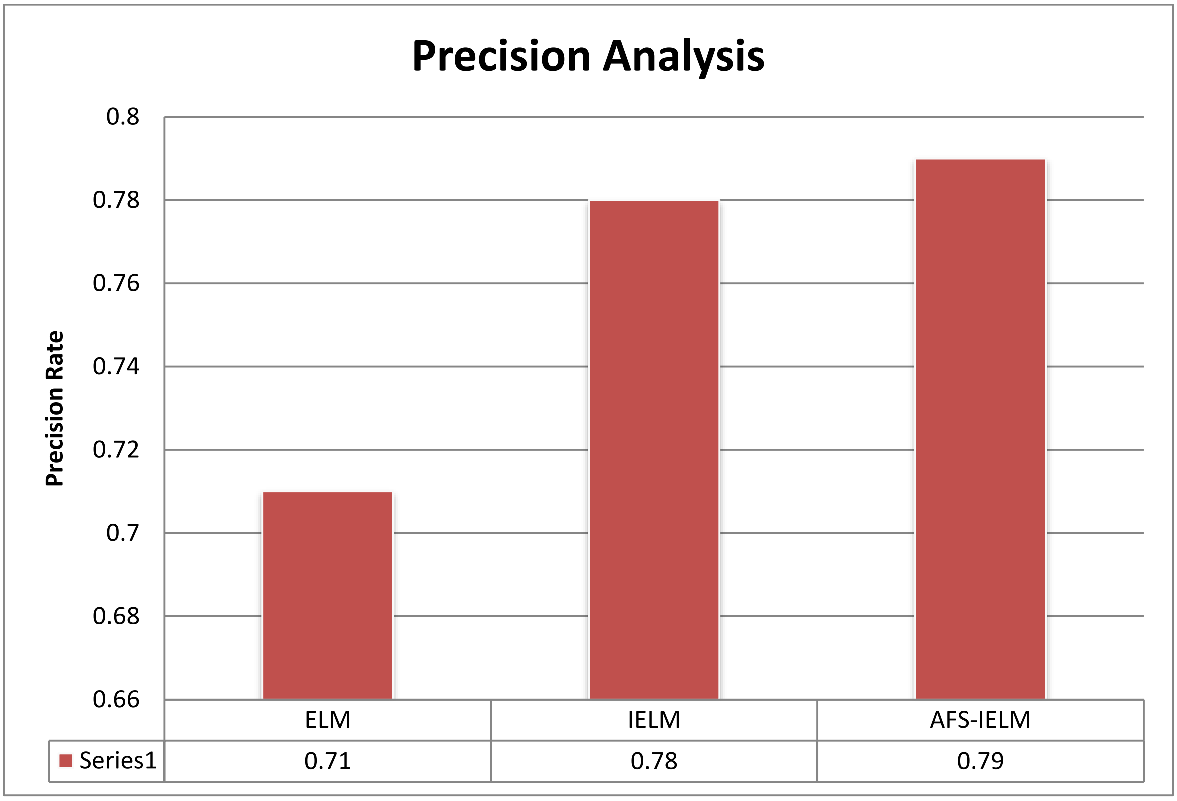

Figure 1. Comparison of Precision

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 (11).

$$F1 = 2.\frac{precision.recall}{precision + recall} \quad \ldots (11)$$

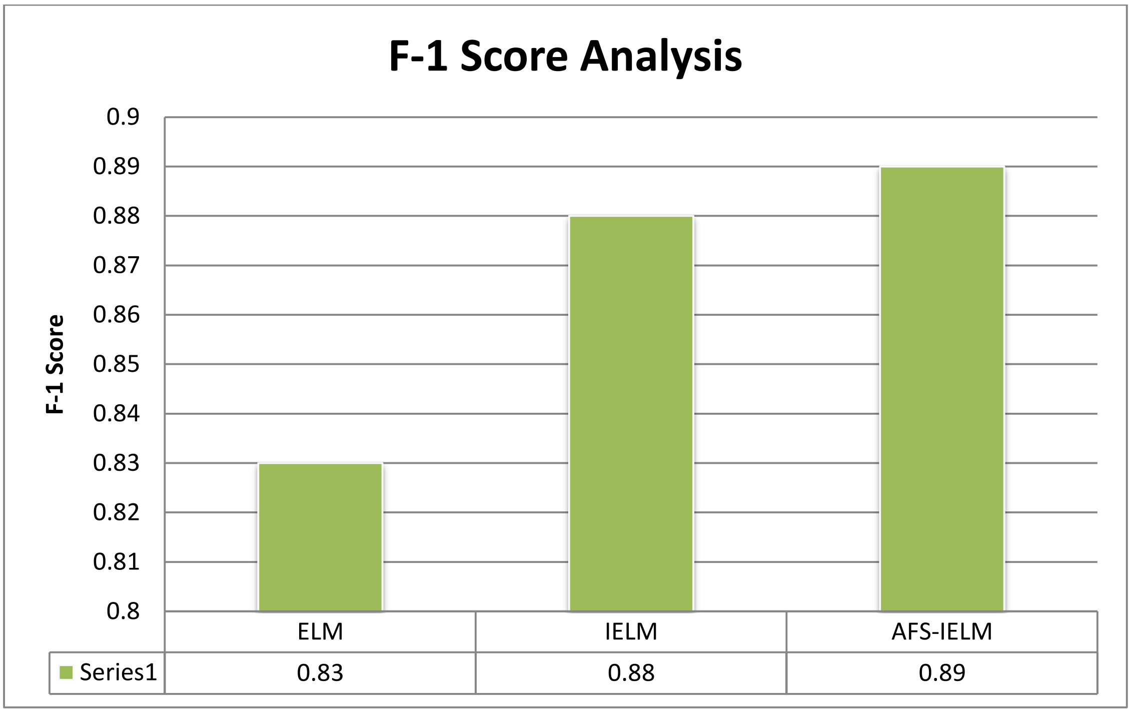

Figure 2. Comparison of F-1 Score

5. Conclusions

This research work aims in design and development of improved extreme learning machine classifier using adaptive feature selection in order to perform web interaction mining. Modification is made in conventional extreme learning machine classifier with the help of improved feature collection using grading method strategy. Performance metrics such as precision, recall and F-1 score are chosen. From the results it is evident that the proposed AFS-IELM algorithm outperforms than ELM and IELM 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] Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, vol. 70, pp. 489-501, 2006.

[17] 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.

[18] 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.

[19] 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.