IJETST- Vol.||04||Issue||09||Pages 5879-5886||September||ISSN 2348-9480 2017

Open access Journal International Journal of Emerging Trends in Science and Technology

IC Value: 76.89 (Index Copernicus) Impact Factor: 4.219 DOI: https://dx.doi.org/10.18535/ijetst/v4i9.04

Web Interaction Mining Using Adaptive Feature Selection Basedpenta Layered Artificial Neural Network (AFS-Plan) Classifier

Authors

B. Kaviyarasu, Dr. A. V. Senthil Kumar

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

Abstract

Predicting the objective of internet users contains different applications in the areas such as e-commerce, entertainment in online, and several internet-based applications. The integral section of classifying the internet queries based on accessible features such as contextual information, keywords and their semantic relationships. This research article aims in proposing Adaptive Feature Selection based Penta Layered Artificial Neural Network Classifier 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 AFS-PLANN classifier with the ANN and PLANN. Results proved that the proposed classifier attains better performance than that of the conventional ANN and PLANN.

Keywords: Web interaction mining, algorithm, neural network, adaptive feature selection, classifier, precision, recall, F1-Score

1. Introduction

Web mining is that the use of data mining procedures to remove learning from web data, in conjunction with web archives, hyperlinks between records, use logs of web destinations, and a lot of others. web mining is that the withdrawal of without a doubt significant examples and certain comprehension from intrigue related with the area. This furtherected data will be additional wont to upgrade web usage indicated prediction of resulting page conceivable to got to through customer, crime detection and future prediction, individual distinguishing proof and to recognize concerning individual watching out leisure activities [Monika Dhandi, Rajesh Kumar Chakrawarti.,2016] [8].

Web Mining can be comprehensively remoted into 3 unique instructions, as indicated by way of the forms of records to be mined. The evaluate of the three classifications of internet mining [T. Srivastava et al.,2013] [11] discussed beneath are (1) Web Content Mining (2) Web Structure Mining (3) Web Interaction Mining.

Web Content Mining (WCM): WCM is the manner in the direction of extricating beneficial facts from the substance of net records. Depicted facts pertains to the collection of certainties of an internet web page were intended to pass directly to the clients. It might include of content, pictures, sound, video, or organized statistics, as an example, information and tables.

Web Structure Mining (WSM): The structure of a distinctive internet contains of Web pages as nodes, and web link as edges associating associated pages. Web Structure Mining is the way toward locating structure information from the Web. This can be further partitioned into sorts in view of the type of structure data utilized.

(a) Hyperlinks: A Hyperlink is a simple unit that interfaces a place in a web page to face-out vicinity, both in the indistinguishable internet page or on an change web page.

(b) Document Structure: Moreover, the substance inner a web page will likewise be composed in a tree-organized structure, situated on the greater than multiple HTML and XML labels inside the website web page. Mining endeavors proper have intrigued certainly by means of isolating record item version (DOM) systems out of documents.

Web Interaction Mining (WIM): WIM is the usage of statistics mining processes to find exciting usage designs from Web statistics, with a particular cease intention to realise and better serve the necessities of Web-primarily based packages. Use of data catches the person or supply of net customers alongside their perusing behavior at a web site. WUM itself may be grouped further contingent upon the type of use records taken into consideration:

(a)**Web Server Data:** The customer logs are amassed by Web server. Small range of the data includes IP cope with, page reference and get to time.

(b)**Application Server Data:** Commercial utility servers, for instance, Web-common sense, Story-Server have noteworthy components to empower E-trade applications to be primarily based on pinnacle of them with little exertion. A key thing is the ability to tune one of a kind varieties of business activities and log them in utility server logs.

(c)**Application Level Data:** New varieties of activities may be characterised in an utility, and logging can be became on for them - producing histories of these uniquely characterised events.

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 supplied three facts services: entity synonym records service, query-to-entity facts service and entity tagging know-how company. The entity synonym provider used to be an in-advent information service that used to be currently available even as the alternative are information services currently in development at Microsoft. Their experiments on product datasets showcase (i) these understanding offerings have immoderate pleasant and (ii) they have large have an impact on on customer studies on e-tailer internet web sites.

M. Nayrolles and A. Hamou-Lhadj.,2016 [7] proposed BUMPER (BUg Metarepository for dEvelopers and Researchers), a normal infrastructure for developers and researchers interested in mining information from many (heterogeneous) repositories. BUMPER was an open supply web-based environment that extracts facts from a diffusion of BR repositories and variation manage structures. It become as soon as equipped with a sturdy search engine to useful resource clients speedy question the repositories utilizing a unmarried factor of access. X.

Ye et al.,2015 [12] authors proposed a new studying method through a generalized loss characteristic to seize the subtle relevance variations of schooling samples while a extra granular label charter turned into once accessible. Authors have utilized it to the Xbox One's film seek challenge the location consultation-centered individual behavior information changed into as soon as available and the granular relevance differences of coaching samples are derived from the consultation logs. When placed subsequent with the prevailing method, their new generalized loss characteristic has examined state-of-the-art experiment performance measured via a few patron-engagement metrics.

The purpose of T. F. Lin and Y. P. Chi.,2014 [10] became to utilize the implemented sciences of TF-IDF, adequate-technique clustering and indexing wonderful exam to set up the combo of key terms a good way to advantage seo. The study tested that it'd in all likelihood effectively enhance the internet website's development of ranking on search engine, growth internet website's exposure level and click on thru expense.

G. Dhivya et al.,2015 [3] analyzed character behavior by means of the usage of mining enriched web entry log data. The few net interaction mining tactics for extracting treasured factors used to be discussed and rent a majority of these strategies to cluster the users of the domain to take a look at their behaviors comprehensively. The contributions of this thesis are an statistics enrichment that turned into content material and starting vicinity located and a treelike visualization of generic navigational sequences. This visualization makes it feasible for a easily interpretable tree-like view of patterns with highlighted primary understanding.

Z. Liao et al.,2014 [15] delivered "project path" to recognize person search behaviors. Authors define a mission to be an atomic man or woman know-how need, whereas a undertaking path represents all individual hobbies interior that precise venture, equal to question reformulations, URL clicks. Previously, internet search logs have been studied by using and massive at session or question degree the region clients can also positioned up several queries inside one task and manage several responsibilities inner one consultation.

A. Yang et al.,2014 [2] have offered a solution that first identifies the clients whose kNN's likely plagued through the newly arrived content, after which replace their kNN's respectively. Authors proposed a brand new index charter named HDR-tree with a purpose to support the effective search of affected clients. HDR-tree keeps dimensionality discount through clustering and precept element assessment (PCA) for you to make more potent the hunt effectiveness. To greater reduce reaction time, authors proposed a version of HDR-tree, referred to as HDR-tree, that enables extra effective however approximate answers.

A. U. R. Khan et al.,2015 [5] have offered a cloud carrier to explain how the reputation of the mass media information may be assessed making use of users online usage behavior. Authors used understanding from Google and Wikipedia for this assessment undertaking. Google statistics turned into useful in expertise the have an effect on of testimonies on web searches whereas information from Wikipedia enabled us to take into account that articles associated with rising records content material moreover locate lot of interest.

J. Jojo and N. Sugana.,2013 [4] proposed a hybrid approach which makes use of the ant-based clustering and LCS classification methods to are seeking for out and expect consumer's navigation conduct. As a end result person profile will also be tracked in dynamic pages. Personalized seek may be used to cope with challenge inside the internet search community, founded on the idea that a consumer's ordinary choice may additionally simply useful resource the quest engine disambiguate the actual goal of a query.

M. A. Potey et al.,2013 [6] reviewed and compared the available tactics to offer an insight into the area of query log processing for expertise retrieval.

A. Vinupriya and S. Gomathi.,2016 [1] proposed a modern-day scheme named as WPP (net web page Personalization) for effective net web page tips. WPP encompass web page hit rely, complete time spent in each hyperlink, number of downloads and hyperlink separation. Founded on those parameters the personalization has been proposed. The system proposes a latest implicit consumer remarks and event link get entry to schemes for splendid net internet web page customization collectively with area ontology.

Y. C. Fan et al.,2016 [14] proposed an statistics cleaning and expertise enrichment framework for permitting customer opportunity operating out by using manner of Wi-Fi logs, and introduces a sequence of filters for cleaning, correcting, and refining Wi-Fi logs.

Y. Kiyota et al.,2015 described discover ways to assemble a property seek conduct corpus derived from micro blogging timelines, wherein tweets concerning property search are annotated. Authors applied micro mission-hooked up crowd sourcing to tweet knowledge, and construct a corpus which contains timelines of special clients which are annotated with property seek stages.

3. PROPOSED WORK

The proposed work has two main phases. The first phase emphasizes on the adaptive feature selection method. The second phase concentrates on PLANN which is employed for performing the classification task.

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 $F=\{f_1, f_2,\ldots, f_N\}$ 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 $M_1, M_2,\ldots, M_L$. 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 $FV_j=\left[f_{i1}{}^{j}, f_{i2}{}^{j},\ldots, f_{iN}{}^{j}\right]$ that created by j th feature selection method, $f_{i1}{}^{j}$ is a permutation of $\{f_1, f_2,\ldots, f_N\}$.

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

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:

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

… (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:

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

$$Feature\ subsets=\left\{\{x_1\},\{x_1, x_2\},\{x_1, x_2, x_3\},\ldots,\{x_1, x_2,\ldots, x_N\}\right\}$$

… (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. Penta Layered - ANN (PL – ANN) Classifier

Once when the features are selected using AFS mechanism PL-ANN takes care for performing the classification task. PL - ANN is a five layered RBF based classifier neural network that makes use of gradient descent approach and regression based classification. It optimizes flattening parameter of RBF kernel through gradient descent approach. It consists of five layers named as input, pattern, summation, normalization and output and is portrayed the Fig. 1.

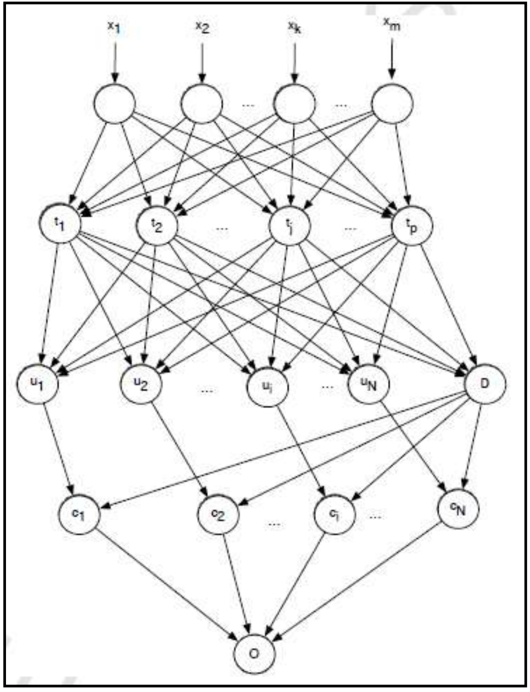


Fig. 1. The Proposed Penta Layered Artificial Neural Network

Applied input vector x is transmitted to pattern layer through input layer. Pattern layer includes one neuron for each training datum with RBF kernel. Squared Euclidean distance between input vector x and training data vector t is calculated as in (4) where p denotes total number of training data at pattern layer.

$$dist(j)=\left\|x-t_j\right\|^2,1\le j\le p \text{ ... (4)}$$

Calculated squared Euclidean distances are used in RBF kernel function as in (5) where r (j) denotes output of jth training data and σ represents flattening parameter. Outputs of RBF kernel function are the output values of pattern layer neurons. Moreover, this layer includes N target values of each training datum determined by corresponding class.

$$r(j)=e^{\left(-1*\frac{dist(j)}{2\sigma^2}\right)},1\le j\le p \text{ ... (5)}$$

When a training datum belongs to ith class then its ith value will be 0.9 and others will be 0.1, as given in (6).

$$y(j,i)=\begin{cases}0.9t_j\ belongs\ to\ i^{th}\ class\, 1\le i\le N\\ 0.1else \qquad\qquad\qquad 1\le j\le p\end{cases} \text{ ... (6)}$$

N+1 neurons are placed at summation layer where N is the total number of classes and additional one term to N is for one neuron to obtain denominator. PL-ANN uses diverge effect term at summation layer to increase the distances among classes. Diverge effect term value is calculated as in (7) where d (j, i) denotes diverge effect term of jth training data and ith class. y_{max} is initialized to 0.9 which denotes the maximum value of y(j, i). y_{max} value is updated with the maximum value of output layer after each iteration of optimization. Diverge effect term is calculated by N neurons of summation layer. This calculation includes exponential form of y (j, i) − y_{max} to increase the effect of y (j, i).

$$d(j,i)=e^{(y(j,i)-y_{\max})}*y(j,i) \text{ ... (7)}$$

Diverge effect term is used in calculating nominator values at summation layer as in (8). Moreover, denominator value is also calculated at this layer as in (9).

$$u_i=\sum_{j=1}^{p}d(j,i)*r(j),1\le i\le N \text{ ... (8)}$$

When N neurons, represented with u_i, calculate nominator values by summing dot product of diverge effect terms and pattern layer outputs, other neuron calculates denominator value the same as PL-ANN represented by D.

$$D=\sum_{j=1}^{p}r(j) \text{ ... (9)}$$

Each class is represented with a neuron at normalization layer. These neurons divide corresponding nominator value by denominator value calculated at summation layer, according to (10) where c_i denotes normalized output of ith class.

$$c_i=\frac{u_i}{D},1\le i\le N \text{ ... (10)}$$

Class of input vector is determined at output layer through the champ decision mechanism as given in (11) where c is the output vector of normalization layer, ci_d and id denote champ neuron value and indices of the class, respectively.

$$[c_{id}, id] = \max(c) \dots (11)$$

Gradient descent based interactive learning is utilized in PL-ANN for obtaining optimized flattening parameter value. Each training datum at pattern layer is sequentially applied to neural network and three steps are executed until maximum iteration limit exceeds. Firstly, squared error e is calculated for each input, as in (12) where y(z, id) represents the value of z^{th} training input data for id^{th} class and c_{id} is value of champ class.

$$e = (y(z, id) - c_{id})^2 \dots (12)$$

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.

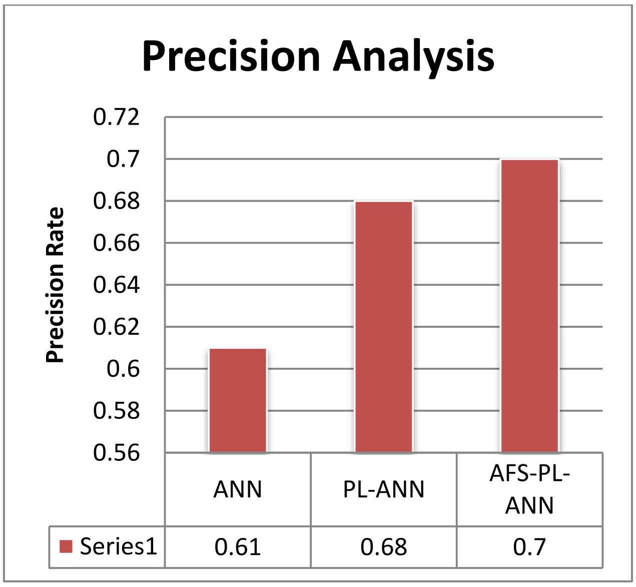
This led to 129 sessions and 353 web searches, which were recorded and successively analyzed in order to manually classify the intent of the user according to the two-level taxonomy. Starting from web searches, 490 web pages and 2136 sub pages were visited. The interaction features were logged by the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

For performing query classification, the proposed PL-ANN 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. The values of the precision, recall and F1-Score of the participants are given in Annexure 1.

Precision: It is the fraction of retrieved documents that are relevant to the query which is calculated using (13).

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \dots (13)$$



	ANN	PL-ANN	AFS-PL-ANN
🟥 Series1	0.61	0.68	0.7

Fig 2. 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 (14).

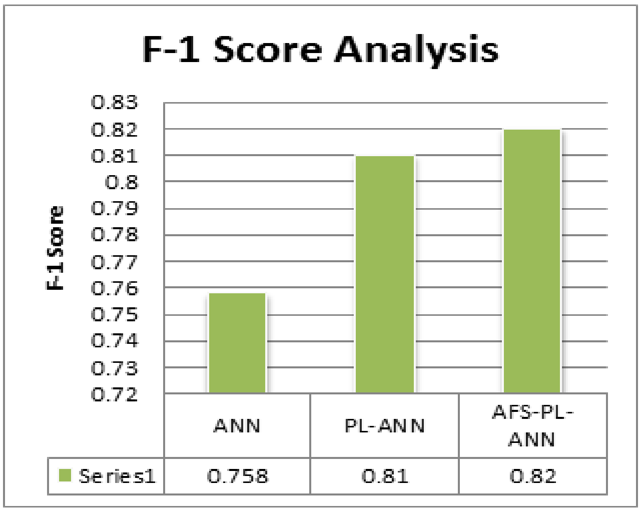


Figure 3. Comparison of F-1 Score

$$F1=2.\frac{precision.recall}{precision+recall} \ldots (14)$$

5. Conclusions

This research work aims in design and development of adaptive feature selection based penta layered artificial neural network (PL-ANN) in order to perform web interaction mining. Adaptive feature selection mechanism is employed in order to select the features. Certain modifications are made in conventional neural network machine learning classifier. 5 layers are designed namely; input layer, pattern layer, summation layer, normalization layer and output layer. Performance metrics such as precision, recall and F-1 score are chosen. From the results it is evident that the proposed AFS-PL-ANN algorithm outperforms than ANN and PL-ANN classifier.

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.