International Journal of Computer Science Engineering & Technology
APPROVED BY
NATIONAL SCIENCE LIBRARY (NSL)
NATIONAL INSTITUTE OF SCIENCE COMMUNICATION AND INFORMATION RESOURCES (NISCAIR)
COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR)– NEW DELHI INDIA
ISSN :2455-9091

WEB INTERACTION MINING USING PENTA LAYERED ARTIFICIAL NEURAL NETWORK CLASSIFIER

[1] B. Kaviyarasu, [2] Dr. A. V. Senthil Kumar
[1] Research Scholar, [2] Director,
[1, 2] PG and Research Department of Computer Computer Applications,
[1, 2] Hindusthan College of Arts and Science,
[1, 2] Coimbatore – 38.

ABSTRACT: Predicting the intention 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 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 classifier with the ANN. Results proved that the proposed classifier attains better performance than that of the conventional ANN.

Keywords: [Web interaction mining, algorithm, neural network, classifier, precision, recall, F1-Score.]

1. INTRODUCTION

Web mining is the application of data mining methods to extract knowledge from internet information, together with internet documents, hyperlinks between records, usage logs of web sites, and many others. Web mining is the withdrawal of potentially valuable patterns and implicit understanding from pastime related to the site. This extracted knowledge will also be extra used to enhance web utilization such that prediction of subsequent page likely to accessed through consumer, crime detection and future prediction, person profiling and to recognize about person searching hobbies [Monika Dhandi, Rajesh Kumar Chakrawarti.,2016] [8].

Web Mining can be comprehensively isolated into three particular classes, as indicated by the sorts of information to be mined. The review of the three classifications of web mining [T. Srivastava et al.,2013] [11] discussed below are (1) Web Content Mining (2) Web Structure Mining (3) Web Interaction Mining.

Web Content Mining (WCM): WCM is the way toward extricating helpful data from the substance of web archives. Depicted information relates to the gathering of certainties of a web page were intended to pass on to the clients. It might comprise of content, pictures, sound, video, or organized records, for example, records and tables.

Web Structure Mining (WSM): The structure of a distinctive web comprises of Web pages as nodes, and web link as edges associating related pages. Web Structure Mining is the way toward finding structure data from the Web. This can be further partitioned into two sorts in view of the sort of structure data utilized.

Hyperlinks: A Hyperlink is a basic unit that interfaces an area in a page to stand-out region, either inside the indistinguishable web page or on an alternate page.

Document Structure: Moreover, the substance inside a page will likewise be composed in a tree-organized structure, headquartered on the more than a couple of HTML and XML labels inside the website page. Mining endeavors right have intrigued undoubtedly by separating document object model (DOM) structures out of documents.

Web Interaction Mining (WIM): WIM is the use of data mining procedures to find intriguing utilization designs from Web information, with a specific end goal to comprehend and better serve the requirements of Web-based applications. Use of information catches the character or source of web clients alongside their perusing conduct at a webpage. WUM itself can be grouped further contingent upon the sort of use information considered:

Web Server Data: The client logs are gathered by Web server. Small range of the information incorporates IP address, page reference and get to time.

Application Server Data: Commercial application servers, for example, Web-logic, Story-Server have noteworthy components to empower E-trade applications to be based on top of them with little exertion. A key component is the capacity to track different sorts of business occasions and log them in application server logs.

Application Level Data: New sorts of occasions can be characterized in an application, and logging can be turned on for them - producing histories of these uniquely characterized occasions.

2. RELATED WORKS

T. Cheng et al.,2013 [9] have provided three information offerings: entity synonym information carrier, query-to-entity information service and entity tagging knowledge provider. The entity synonym service used to be an in-creation knowledge carrier that used to be presently available whilst the other two are information services presently in progress at Microsoft. Their experiments on product datasets exhibit (i) these knowledge offerings have excessive best and (ii) they've gigantic influence on consumer experiences on e-tailer web sites.

M. Nayrolles and A. Hamou-Lhadj.,2016 [7] proposed BUMPER (BUg Metarepository for dEvelopers and Researchers), a customary infrastructure for developers and researchers inquisitive about mining information from many (heterogeneous) repositories. BUMPER used to be an open supply web-founded environment that extracts information from a variety of BR repositories and variant manipulate systems. It was once equipped with a strong search engine to aid customers quickly query the repositories utilizing a single point of access. X.

Ye et al.,2015 [12] authors proposed a new studying method by means of a generalized loss function to capture the subtle relevance variations of training samples when a extra granular label constitution was once on hand. Authors have utilized it to the Xbox One's movie search mission the place session-headquartered person conduct understanding was once to be had and the granular relevance differences of coaching samples are derived from the session logs. When put next with the prevailing method, their new generalized loss function has tested sophisticated experiment efficiency measured by means of a few consumer-engagement metrics.

The purpose of T. F. Lin and Y. P. Chi.,2014 [10] was to make use of the applied sciences of TF-IDF, ok-approach clustering and indexing high-quality examination to establish the combo of key phrases to be able to advantage seo. The learn demonstrated that it might probably comfortably enhance the internet site's advancement of ranking on search engine, increase internet site's publicity level and click on through expense.

G. Dhivya et al.,2015 [3] analyzed person conduct by using mining enriched web entry log information. The few net interaction mining approaches for extracting valuable elements used to be discussed and employ all these strategies to cluster the users of the domain to study their behaviors comprehensively. The contributions of this thesis are an information enrichment that was content and starting place situated and a treelike visualization of generic navigational sequences. This visualization makes it possible for a conveniently interpretable tree-like view of patterns with highlighted primary know-how.

Z. Liao et al.,2014 [15] introduced "task trail" to understand user search behaviors. Authors outline a mission to be an atomic person know-how want, whereas a challenge trail represents all person pursuits inside that precise project, equivalent to question reformulations, URL clicks. Previously, net search logs have been studied by and large at session or question stage the place customers may put up several queries within one venture and manage several tasks inside one session.

A. Yang et al.,2014 [2] have awarded a solution that first identifies the customers whose kNN's possibly plagued by the newly arrived content, after which replace their kNN's respectively. Authors proposed a new index constitution named HDR-tree in order to support the effective search of affected customers. HDR-tree continues dimensionality reduction through clustering and principle element evaluation (PCA) so as to make stronger the search effectiveness. To extra scale back response time, authors proposed a variant of HDR-tree, known as HDR-tree, that helps extra effective but approximate solutions.

A. U. R. Khan et al.,2015 [5] have presented a cloud carrier to explain how the status of the mass media news can be assessed utilizing users online utilization habits. Authors used knowledge from Google and Wikipedia for this comparison challenge. Google data was helpful in understanding the have an effect on of stories on web searches whereas data from Wikipedia enabled us to understand that articles related to rising information content additionally find lot of attention.

J. Jojo and N. Sugana.,2013 [4] proposed a hybrid approach which uses the ant-founded clustering and LCS classification methods to seek out and predict user's navigation behavior. As a result user profile may also be tracked in dynamic pages. Personalized search can be used to address project in the internet search community, founded on the premise that a consumer's normal choice may just aid the quest engine disambiguate the real intention of a question.

M. A. Potey et al.,2013 [6] reviewed and compared the to be had approaches to present an insight into the discipline of query log processing for expertise retrieval.

A. Vinupriya and S. Gomathi.,2016 [1] proposed a brand new scheme named as WPP (web page Personalization) for powerful net page suggestions. WPP consist of page hit rely, complete time spent in each hyperlink, number of downloads and link separation. Founded on these parameters the personalization has been proposed. The procedure proposes a brand new implicit user feedback and event hyperlink access schemes for amazing internet web page customization together with domain ontology.

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**.**

The Proposed Penta Layered - ANN (PL – ANN) Classifier

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. 2.

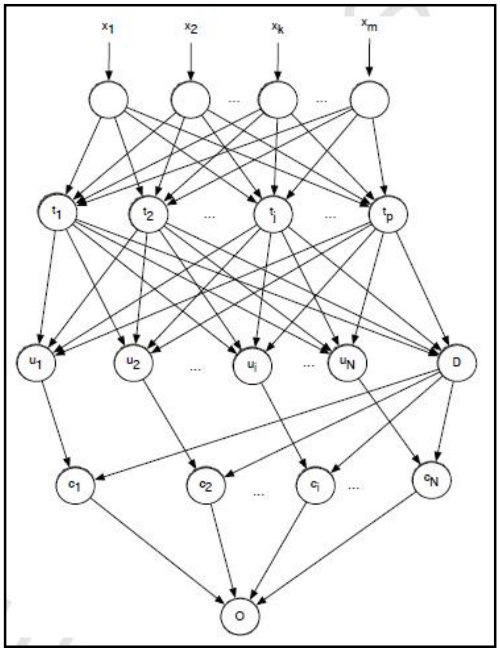

Figure - 2 The Proposed Penta Layered Artificial Neural Network

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 (1) where p denotes total number of training data at pattern layer.

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

Calculated squared Euclidean distances are used in RBF kernel function as in (2) where r (j) denotes output of j^{th} training data and ± 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.

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

When a training datum belongs to i^{th} class then its i^{th} value will be 0.9 and others will be 0.1, as given in (3).

$$y(j,i) = \begin{cases} 0.9 t_j \ belongs\, to\, i^{th}\ class\, 1 \le i \le N \\ 0.1 else \qquad\qquad 1 \le j \le p \end{cases} \text{ ... (3)}$$

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 (4) where d (j, i) denotes diverge effect term of j^{th} training data and i^{th} 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).

$$d(j,i) = e^{(y(j,i) - y_{\max})} * y(j,i) \text{ ... (4)}$$

Diverge effect term is used in calculating nominator values at summation layer as in (5). Moreover, denominator value is also calculated at this layer as in (6).

$$u_i = \sum_{j=1}^{p} d(j,i) * r(j), 1 \le i \le N \text{ ... (5)}$$

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.

$$D = \sum_{j=1}^{p} r(j) \text{ ... (6)}$$

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 (7) where c_i denotes normalized output of i^{th} class.

$$c_i = \frac{u_i}{D}, 1 \le i \le N \text{ ... (7)}$$

Class of input vector is determined at output layer through the champ decision mechanism as given in (8) where c is the output vector of normalization layer, ci_d and id denote champ neuron value and indices of the class, respectively.

$$[c_{id}, id] = \max(c) \text{ ... (8)}$$

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 (9) 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.

$$e = (y(z,id) - c_{id})^2 \ldots (9)$$

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 (6).

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

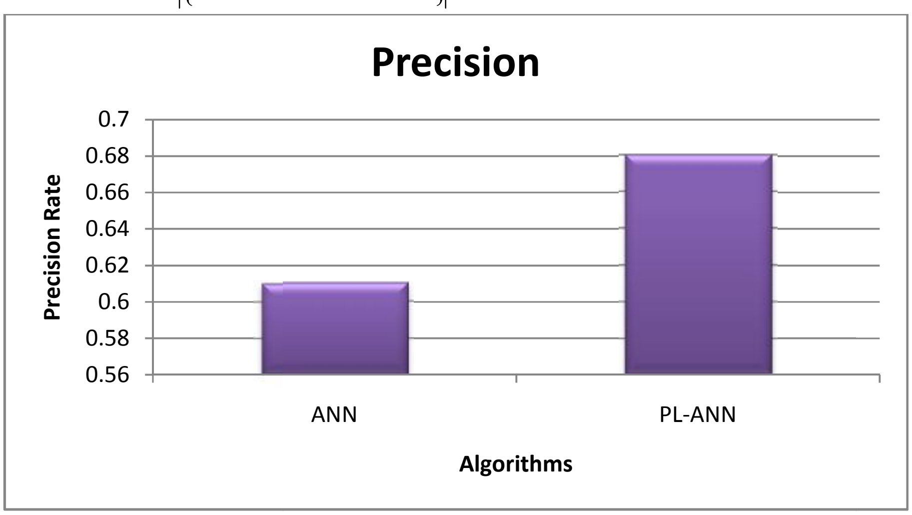

Figure - 1 Comparison of Precision

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).

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

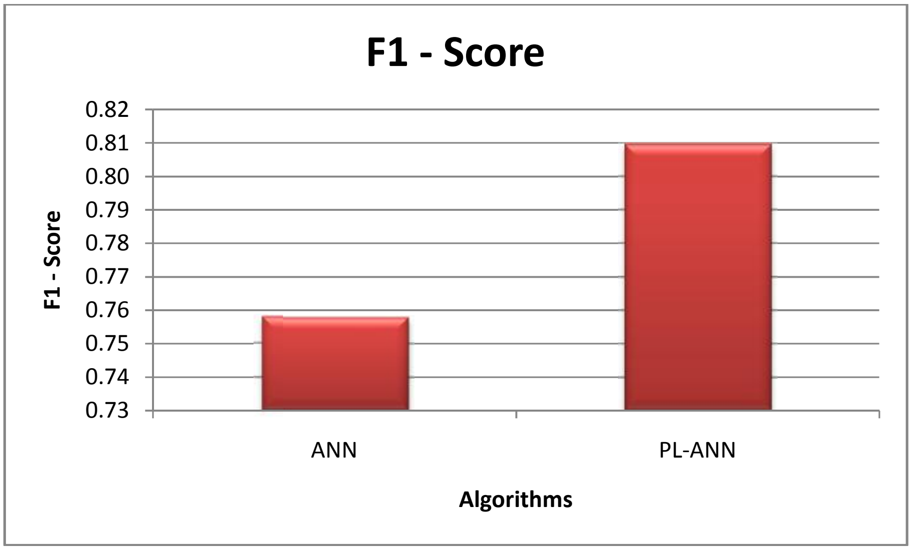

Figure - 2 Comparison of F-1 Score

CONCLUSIONS

This research work aims in design and development of penta layered artificial neural network (PL-ANN) in order to perform web interaction mining. Modification is 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 PL-ANN algorithm outperforms 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.