International Journal of Research in Advent Technology, Vol.5, No.2, February 2017
E-ISSN: 2321-9637
Available online at www.ijrat.org

Classification of User Task Behavior by Using Error Back Propagation Neural Network

Arti Dwivedi[1], Prof. Umesh Lilhore[2]
Computer Science and Engineering, NRI Institute of Information Science &Technology
Email: arti.mcpl@gmail.com[1]

Abstract- Web miners are elaborating the various field of research out of those user task based behavior learning is new area of research. So reducing the load on the servers is prior requirement of the work by analyzing the user task behaviour. This paper focus on developing a model that can classify the user task by through its searching or navigation patterns. Here for classification error bach propogation neural network was used. Clustering into multi-class is obtained which increase the dimension of the work as well. Experiment was done on web user navigation and searching behavior at different time duration. Results shows that proposed has achieve a high precision, accuracy for clustering of user query. Proposed scheme reduce searching time as classification is done by trained artificial model.

Index Terms- Information Extraction, web log, web query ranking, web mining.

1 Introduction

As the internet users are increasing on daily basis. So the requirement of the web world is quit high. In order to increase the transparency and quickness in the work large amount of work is depend on this digital network. This attracts many researcher for improving the performance of the network and reduce the latency time of the internet, so that things get easy and fast for the daily users. Here hardware part is the way of optimizing the network but in parallel software also need to update. This paper focus on optimizing the web power by learning the user behavior for reducing the latency time of searching the required matter of interest. As websites are very important source of information for almost all kind of things, so this gathering of people attract number of people to provide various services. But targeting the correct customer is basic requirement of the service or business. Research in this area has the objectives of helping e-commerce businesses in their decision making, assisting in the design of good Web sites and assisting the user when navigating the Web.

Web Structure Mining: This feature develops the website page structure where links between the pages shows relation between pages. Web structure mining helps in finding the similar pages where relation between the websites are also look closer. Here importance of this feature in web mining is quit low as compare to other features.

User's behavior in web browsing can be categorized into two states ,that is, the search state and the browse state. In the search state, a user fires a query to a search engine and get on the search results returned by the search engine selectively. Afterwards, the user may further improve the query and come to an end in the interaction with the search engine. When the user visits a web page other than a search result page, user is considered to be in the browse state. The major difference between the search state and the browse state lies in the type of server which the user interacts with: if a user is interacting with a search engine, user is in the search state else the user is in the browsing state. Users frequently make changeovers between the search state and the browse state. When a user browses a web page, user may want to go to a search engine and search contents related to the page. In this case, the user transfers from the browse state to the search state. On the other way, a user may also transfer from the search state to the browse state. To be definite, after the user clicks on a web page in the search result, user may further trail the hyper-links of the web page and consent the interaction with the search engine.

Query Trail: It represents a sequence of user behaviors of one of the user starting from a query followed by sequence of browsing behaviors that are triggered by this query.

Session Trail: It represents a sequence of user behaviors of one of the user where user behaviors are consecutive and any two consecutive occurred within the time threshold.

Drawbacks of current system: In case of query trails the semantic suggestion amongst immediate query trails are lost. In case of session trails it strictly follows the linear order of user behaviors in search logs. Time threshold settings for session clustering are not able to satisfy our predefined goals. Sessions contain several nuclear information need which are unrelated semantically. ODP (Observer Design Pattern) category data is essential to foresee user search interest. Co-occurrence based query suggestion methods based on task trail and are connected with similar methods based on session trails and click through bipartite graph or Shared graph.

2. Related work

Dong, Farookh Hussain and Elizabeth Chang in [1] proposed Web Query Clustering technique which was depend on web distance normalization. In this architecture middle categorized queries are send to the target class by normalizing and mapping the web queries. By defining the frequency, position and position frequency categories are ranked into three class. In the system Taxonomy-Bridging Algorithm is used to map target category. The Open Directory Project (ODP) is used to build an ODP-based classifier. This taxonomy is then mapped to the target categories using Taxonomy-Bridging Algorithm. Thus, the post-retrieval query query is first classified into the ODP taxonomy, and the classifications are then mapped into the target categories for web query.

Classification of web query to the user intendant query is major task for any information retrieval system. MyoMyo ThanNaing [2] proposed Query Classification Algorithm. To classify the web query inputted by the user into the user intended categories, MyoMyo ThanNaing use the domain ontology. Ontology is useful to matching of retrieve category to target category. User query are extracted in Domain terms are used as input to the query classification algorithm. Matched terms of each domain term are extracted in further sub category. Compute the probability for matched categories. Then all querys are ranked by their probability and displays to the user's desk.

Ernesto William De Luca and Andreas Nürnberger [3] proposed method of web query classification using sense folder. In this method the user query is separated in small terms. These small terms are matched with target categories using ontology. Ontology is set of rules. Word vectors (prototypes) are used to create semantic category. Then Search results are indexed by using sense folder.

At last retrieved querys are displays to the user desk. Suha S. Oleiwi, Azman Yasin [4] proposed method of web query classification using Ontology and classification. All are retrieve querys are indexed according to their probability. Probability depends on how often the querys are search on web by user.

Another study proposed an algorithm named Query-Query Semantic Based Similarity Algorithm (QQSSA). This algorithm works on a new approach it filters and breaks the long Query into small words and filters all possible prepositions, conjunction, article, special characters and other sentence delimiters from the query. And then expand the query into logically similar word to form the collection of similar words. Construct the Hyponym Tree for query1 and query2 etc. And based upon some distance measure he classifies the query.

Another approach is Clustering methodology by S. loelyn Rose, K R Chandran and M Nithya [5]. The clustering methodology can be fragmented into the following phases. Feature Extraction, and Mapping intermediate categories to target categories The features extracted in the first phase are mapped onto various target categories in this second phase by Direct Mapping, Glossary Mapping, Wordnet Mapping, Semantic Similarity Measure.

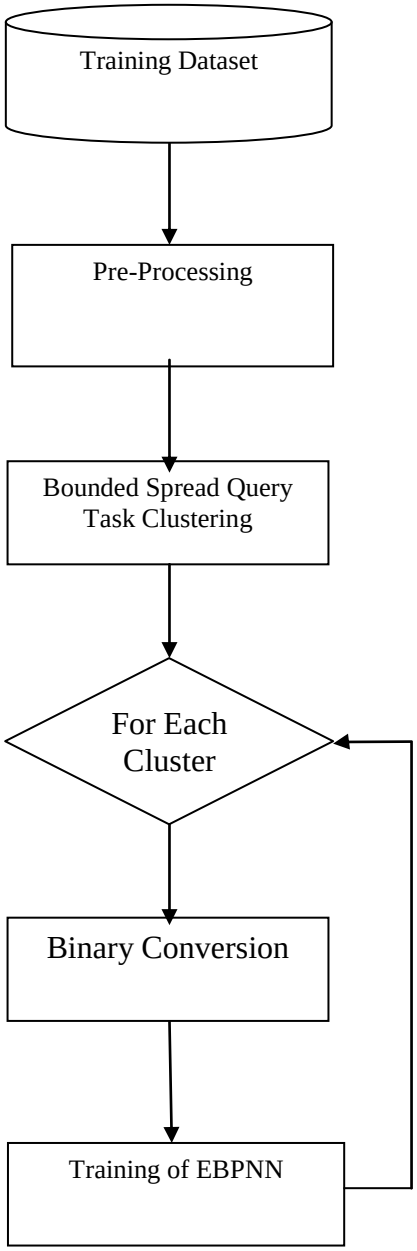

Fig.1 Proposed work training module.

3. Proposed Work

As the mining is utilized in different types of data analysis. So all need to increases the different technique in the required area. So proposed work contribute the web mining by clustering the user query in the group without having any prior knowledge of the user behaviour. In the proposed work no need of any format for the input data such as speaker's identification symbol or special character, here all process is perform by utilizing the different combination of terms features.

3.1 Preprocessing

Preprocessing is a process used for conversion of web content query into feature vector. Just like text categorizations the preprocessing also has controversy about its division [1, 7]. Web content preprocessing is consisting of words which are responsible for lowering the performance of learning models. Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stopword elimination. Stopwords are functional words which occur frequently in the language of the text (for example a, the, an, of etc. in English language), so that they are not useful for clustering.

3.2 Bounded Spread Query task Clustering

The vector which contain the pre-processed keywords is use for collecting feature of that query. This is done by comparing the vector with vector KEY (collection of keywords) of the ontology of different area at each comparison word count is increase by one. So the refined vector will act as the feature vector for that document.

So the list of words which are crossing the threshold are consider as the keywords or feature of that document.

[feature] = mini_threshold ([processed_text])

In this way term feature vector is created from the document.

In order to cluster the feature vector as per there required field algorithm was used from [13]. Here this algorithm is named stands for Query Clustering Bounded SPread method. In this approach user queries of limited time interval is consider for study, so this approach required less number of comparison for clustering.

3.3 Binary Conversion

In this step keywords obtained from the features of the document are need to be insert into neural network for clustering but as text words cannot be insert in the neural network. So a representative of those words are required. As each keyword is a set of ASCI value for example keyword "ABCD" ASCI set is [65 66 67 68]. Now each ASCI number is replace by its binary number as 65={ 1000001}, 66={ 1000010}, 67={ 1000011}, 68={ 1000100}. So in this work ABCD binary is {10000011000010100000111000100}.

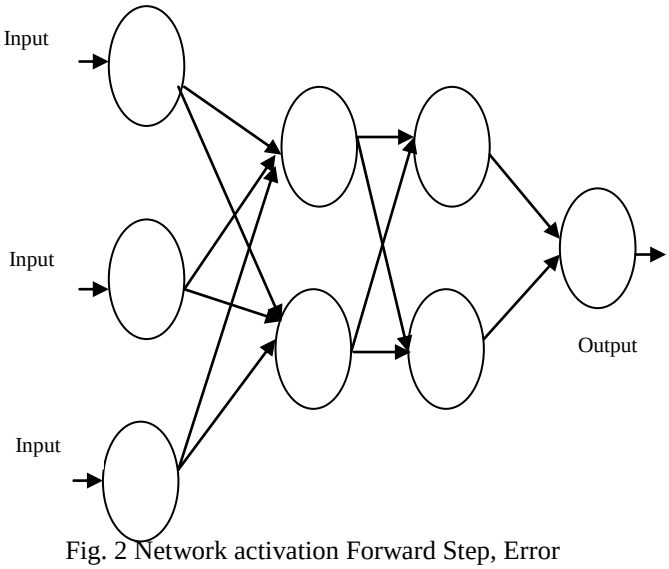

Fig. 2 Network activation Forward Step, Error propagation Backward Step

As each word contain different number of characters so a set of 100 bit is taken as input in the neural network. Where default value is zero in the vector.

3.4 Training of Error Back Propagation Neural Network (EBPNN):

- Consider a network of three layers as shown in fig 3.
- Let us use i to represent nodes in input layer, j to represent nodes in hidden layer and k represent nodes in output layer.
- w_{ij} refers to weight of connection between a node in input layer and node in hidden layer.
- The following equation is used to derive the output value Yj of node j

$$Yj = \frac{1}{1+e^{-X_j}}$$

where, $X_j = \sum x_i \, . \, w_{ij} - \theta_j$, $1 \leq i \leq n$; n is the number of inputs to node j, and θ_j is threshold for node j

- The error of output neuron *k* after the activation of the network on the *n-th* training example (x(n), d(n)) is:

$$e_k(n) = d_k(n) - y_k(n)$$

- The network error is the sum of the squared errors of the output neurons:

$$E(n) = \sum e_k^2(n)$$

- The total mean squared error is the average of the network errors of the training examples.

$$E_{AV} = \frac{1}{N} \sum_{n=1}^{N} E(n)$$

- The Backprop weight update rule is based on the gradient descent method:
  - It takes a step in the direction yielding the maximum decrease of the network error E.
  - This direction is the opposite of the gradient of E.
- Iteration of the Backprop algorithm is usually terminated when the sum of squares of errors of the output values for all training data in an epoch is less than some threshold such as 0.01

$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad \Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

Testing of EBPNN

In this step input query is preprocess as done in the training module, similarly feature vector is create by assigning identification numbers to those keywords. Finally feature vector is input in the EBPNN which give output. Now analysis of that output is done that whether specified class is desired one or not.

3.5 Proposed Algorithm

Input: DD //Training Document Dataset
Output: TNN // Trained Neural Network

1. Loop 1:n // n : number of document in the dataset
2. S←Pre_processing(A) // S: Sentence Matrix.
3. BOW←Feature_selection(S) //Collect terms from processed document
4. EndLoop
5. B ←Binary_Conversion(P)
6. [F C]←Neural_Input(B, C) // C: represent class of the document
7. Loop 1:itr // itr: Iterations
8. TNN←EBPNN(F, C)
9. EndLoop

4 Experiment And Result

In order to implement above algorithm for intrusion detection system MATLAB is use, where dataset is use of different size. Neural Network Toolbox includes command-line functions and apps for creating, training, and simulating neural networks. This make it easy to develop neural networks for tasks such as data-fitting, pattern recognition, and clustering. After creating networks in these tools, it can automatically generate MATLAB code to capture work and automate tasks.

4.1 Evaluation Parameters

As various techniques evolve different steps of working for classifying user query into appropriate category. So it is highly required that proposed techniques or existing work need to be compare on same dataset. But query cluster which are obtained as output is need to be evaluate on the function or formula. So following are some of the evaluation formula which help to judge the clustering techniques ranking.

Precision=True positive/ (True positive + False positives)

Recall=True positives/ (True positive + False negative)

F-Measure= 2*Precision*Recall/ (Precision + Recall)

Accuracy = (True Positive + True Negative) / (True Positive + True Negative+ False Positive + False Negative)

4.2 Results

Table 1. Precision value comparison from trained Neural Network keyword class.

Dataset Percent	**Precision Value Comparison**	
	Previous	**Proposed**
500	**0.1629**	**0.9681**
1000	**0.1792**	**0.9419**

1500	**0.1881**	**0.9397**

Table 1 shows that proposed work has achieved a high precision value as the testing files are increasing. It has shown in table that trained neural network generated value is acceptable for the true positive case.

Table 2. F-measure value comparison from trained Neural Network keyword class.

Dataset Percent	**F-measure Value Comparison**	
	Previous	**Proposed**
500	**0.243**	**0.659**
1000	**0.2639**	**0.6532**
1500	**0.2734**	**0.6527**

Table 2 shows that proposed work has achieved a high F-measure value as the testing files are increasing. It has shown in table that trained neural network generated value is acceptable for the true positive case.

Table 3. Execution time value comparison from trained Neural Network keyword class.

Dataset Percent	**Execution Time (second) Value Comparison**	
	Previous	**Proposed**
500	**9.9469**	**5.326**
1000	**36.678**	**29.218**
1500	**53.632**	**46.6455**

Table 3 shows that proposed work has achieved a low execution time value as compare to previous work. It has shown in table that trained neural network generated value is acceptable for the true positive case

5 Conclusion

As the user satisfaction plays important role in information retrieval. Query recommendation is one of the best method for helping users to satisfy the users information need by suggesting queries related to current users need by maintaining query log processing files. With the proper knowledge from the ontology and the web usage of the web feature vector are create for training the Error back propagation neural network.. By the use of EBPNN clustering of the query get efficient will be less time consuming. This work has increase the accuracy of the clustering so the web server response will be small. Here overall precision and recall values are also good from clustering view. In future one can adopt different genetic approach for clustering of user query as well.

REFERENCES

[1] An Ontology-based Webpage Classification Approach for the Knowledge Grid Environment by Hai Dong, Farookh Hussain and Elizabeth Chang, 2009 Fifth International Conference on Semantics, Knowledge and Grid (IEEE-2009).

[2]Ontology-Based Web Query Classification For Research Paper Searching , By Myomyo Thannaing, International Journal Of Innovations In Engineering And Technology (Ijiet) , Vol. 2 Issue 1 February 2013.

[3] Ontology-Based Semantic Online Classification Of Querys: Supporting Users In Searching The Web By Ernesto William De Luca And Andreas Nürnberger, Ijcst, 2012.

[4] Web Query Classification To Multi Categories Based On Ontology By Suha S. Oleiwi, Azman Yasin, International Journal Of Digital Content Technology And Its Applications(Jdcta) Volume7, Number13, Sep 2013.

[5] S.Lovelyn Rose, K.R.Chandran, M.Nithya An Efficient Approach To Web Query Classification Using State Space Trees., Issn :2229-4333, International Journal Of Computer Science And Technology (Ijcst), June-2011.

[6] Zhao, Y., Karypis, G. 2001. Criterion Functions For Query Clustering:Experiments And Analysis. Technical Report #01-40. University Of Minnesota, Computer Science Department. Minneapolis, Mn (Http://Wwwusers. Cs.Umn.Edu/~Karypis/Publications/Ir.Html)

[7] Zhao, Y., Karypis, G. 2002. Evaluation Of Hierarchical Clustering Algorithms For Query Datasets, Acm Press, 16:515-524.

[8] San San Tint1 And May Yi Aung. "Web Graph Clustering Using Hyperlink Structure ".Advanced Computational Intelligence: An International Journal (Acii), Vol.1, No.2, October 2014

[9] Khan, M. S., & Khor, S. W. (2004). Web Query Clustering Using A Hybrid Neural Network. Applied Soft Computing, 4(4), 423-432. 17

[10] Kleinberg, J. 1997. " Web Usage Mining For Enhancing Search Result Delivery And Helping Users To Find Interesting Web Content",‖ Acm Sigir Conf. Research And Development In Information Retrival (Sigir '13), Pp. 765-769,2013.

[11] Mamoun A. Awad And Issa Khalil "Prediction Of User's Web-Browsing Behavior: Application Of Markov Model". Ieee Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.

[12] Thi Thanh Sang Nguyen, Hai Yan Lu, Jie Lu " Web-Page Recommendation Based On Web Usage And Domain Knowledge" 1041-4347/13/$31.00 © 2013 Ieee.

[13] Zhen Liao, Yang Song, Yalou Huang, Li-Wei He, And Qi He. "Task Trail: An Effective Segmentation Of User Search Behavior" . Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 12, December 2014.