Segmentation of User Task Behavior by using Artificial Neural Network

Ruchika Tripathi
BITS
Bhopal

Pankaj Richhariya
BITS
Bhopal

ABSTRACT

As Segmentation of User's Task to understand the user search behavior is the new field of research for various researchers. Massive volumes of search log data have been collected in several search engines. Currently, a commercial search engine collects billions of queries and gathers terabytes of log data on each single day. At times user moves from one site to another because latency time of the site is more, so the researchers found this as an essential subject for research. Proposed work classifies the user query by combining query clustering boundary spread method with the neural network. For training of neural network proposed work evolve binary feature vector from the clustered query obtained from QCBSP method. The experiment was done on user search behavior of different time intervals. Results show that proposed work has achieved a high precision, accuracy for classification of the user query. Proposed scheme reduces execution time as well because of using trained neural network.

Keywords

Information Extraction, weblog, web query ranking, web mining

1. INTRODUCTION

As the internet users are increasing on daily basis, the requirement of the web world is pretty high. In order to increase the transparency and rapidity in the work large amount of work is depend on this digital network. This attracts many researchers 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 focuses 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 a number of people to provide various services. But targeting the correct customer is the basic requirement of the service or business. Exploration and study in the current area goes with the objectives of helping e-commerce industries to maintain effective internal search engines. This study also helps in efficiently designing of good websites and also helps the user while navigating through the websites.

2. LITERATURE SURVEY

All Hai Dong, Farookh Hussain, and Elizabeth Chang in [1] proposed Web Query Classification technique which depends on web distance normalization. In this architecture, middle categorized queries are sent 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 this system, to record the target category Taxonomy-Bridging Algorithm is used. The ODP (Open Directory Project) is used to develop an ODP-based classifier. Then, this classification is mapped to the target categories using Taxonomy-Bridging Algorithm. Thus, the post-retrieval query into the ODP taxonomy is first classified and the taxonomies are then recorded into the target categories for web query.

Classification of web query to the user intendant query is a major task for any information retrieval system. Myomyo Thannaing [2] proposed Query Classification Algorithm which is used for the classification of web query fired by the user into the user intended categories. Domain ontology was used by Myomyo Thannaing. Ontology is useful to match the retrieved category to target category. The user query is extracted in Domain terms and used as input to the query classification algorithm. Terms which are coordinating with each domain term are taken out in further subcategories. Then system computes the probability for matched categories,then all queries are classified by their probability and shows to the user's desk.

Ernesto William De Luca and Andreas Nürnberger [3] proposed the 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 a semantic category. Then Search results are indexed by using sense folder.

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

One more study suggested an algorithm named (QQSSA) Query-Query Semantic Based Similarity Algorithm. The approach of this algorithm is new as first it filters and breaks the long Query into small words. Then it filters all probable article, prepositions, conjunction, special characters and other sentence delimiters from the query. Afterwards, expand the query into logical similar words to form the collection of similar words. By this user can construct the Hyponym Tree for first and second query etc. and then classification of query is done based upon some distance measures.

Another approach is Classification methodology by S. loelyn Rose, K R Chandran and M Nithya [5]. The classification methodology can be fragmented into the following phases. lllFeature 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.

3. PROPOSED MODEL

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

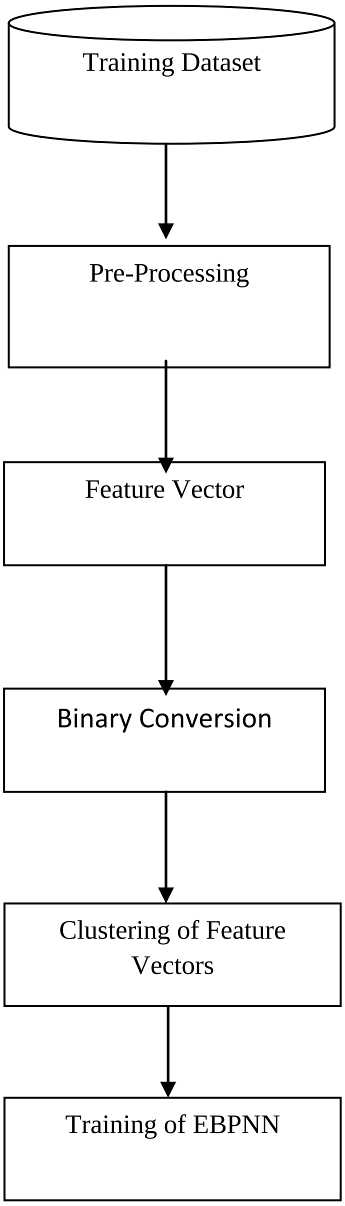

Fig.1 Proposed work training module.

3.1 Preprocessing

Preprocessing is a process used for conversion of web content query into a feature vector. Just like text categorizations the preprocessing also has controversy about its division [1, 7]. Web content preprocessing comprise of words which are responsible for the falling of the performance of learning models. Data preprocessin significantly reduces the extent of the input text documents . Activities like Elimination of stop words, sentence boundary determination are involved to get the database. Stop words are functional words which occur frequently in the language of the text (for example a, the, to, of etc. in the English language) as they are not useful for classification.

3.2 Feature Vector

The vector which contains the pre-processed keywords is used 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 increased by one. So the refined vector will act as the feature vector for that document.The keywords or feature of that document are measured by the List of words which are crossing the threshold.

[feature] = mini_threshold ([processed_text])

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

3.3 Clustering of Feature Vector

In order to cluster the feature vector as per their required field, QC-BSP algorithm was used from [13]. Here QC-BSP 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.

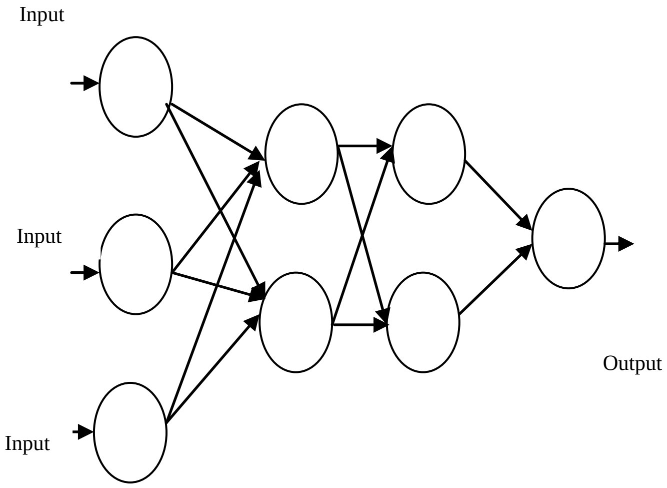

Fig. 2 Error Back propagation step And Network activation Forward Step

3.4 Binary Conversion

In this step keywords obtained from the features of the document are need to be inserted into the neural network for classification but as we know that text words cannot be inserted into the neural network. So the representative of those words is required. As each keyword is a set of ASCII value for example keyword "ABCD" ASCII set is [65 66 67 68]. Now each ASCII number is replace by its binary number as 65={ 1000001}, 66={ 1000010}, 67={ 1000011}, 68={ 1000100}. So in this, we will replace ABCD by its binary number that is {1000001100001010000111000100}.

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

3.5 Training of Error Back Propagation Neural Network (EBPNN)

- Consider a network of three layers as shown in fig 3.
- Now let us consider i to denote the nodes in the input layer, and take j to denote the nodes in the hidden layer and k to denote the nodes in the output layer.

The weight of the connection between a node of input layer and hidden layer is denoted as Wij.

- The following equation is used to derive the output value Yj of node j

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

where $X_j = \sum x_i \,.\, w_{ij} - \theta_j$, $1 \leq\ i \leq n$;

n shows the number of inputs to node j,

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:

$$\mathrm{E(n)} = \sum \mathrm{e}_{\mathrm{k}}^{2}(\mathrm{n})$$

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

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

- The Back propagation 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.
- 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 then Iteration of the Back propagation Algorithm is terminated

$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad \Delta \mathrm{w}_{\mathrm{ij}} = \text{-}\eta \frac{\partial E}{\partial \mathrm{w}_{\mathrm{ij}}}$$

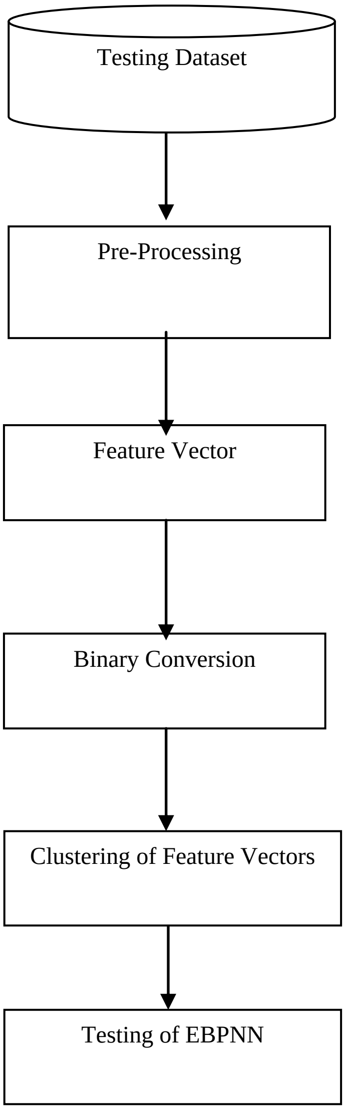

Fig. 3 Proposed work testing module

3.6 Testing of EBPNN

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

4. EXPERIMENT AND RESULTS

Place To implement above algorithm for intrusion detection system, MATLAB is used .The dataset which is used is of different size in the system. Neural Network Toolbox includes applications for creating, training and simulating neural networks and also have command-line functions. This makes 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 Parameter

As various techniques evolve different steps of working for classifying the document into an appropriate category. So it is highly required that proposed techniques or existing work should be compared to the same dataset. But document cluster which is obtained as output needs to be evaluate on the function or formula. So following are some of the evaluation formulae which helps to judge the classification 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 shows that proposed work has achieved a high precision value as the testing files are increasing. It has shown in the table that trained neural network generated value is acceptable for the true positive case.

Table 1. Precision and Recall testing result from trained Neural Network keyword class.

Dataset Percent	Keyword Classification values on Different Testing dataset	
	Precision	Recall
50	0.9545	0.8571
60	0.9565	0.8627
70	0.9583	0.8679
80	0.9615	0.8772

Table 2 shows that proposed work has achieved a high F-Measure and accuracy value as the testing files are increasing. It has shown in the table that trained neural network generated value is acceptable for the keyword classification. Accuracy can further be increased by passing high-quality training dataset.

Table 2. F-Measure and Accuracy testing result from trained Neural Network keyword class

Dataset	Query Classification values on Different Testing dataset	
	F-Measure	Accuracy
50	0.9032	87.7551
60	0.9072	88.2353
70	0.9109	88.6792
80	0.9174	89.4737

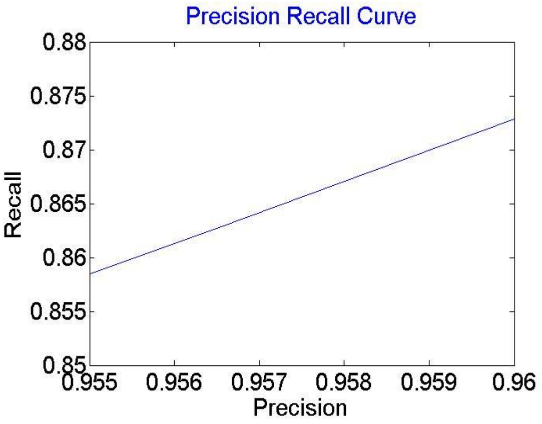

Fig. 4. Precision-recall graph for different dataset size

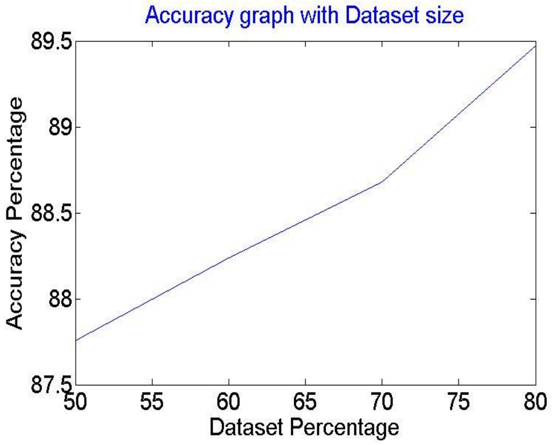

Fig. 5. Accuracy-graph for different dataset size

Above graphs 4 and 5 shows that with the increase in dataset size for testing results gets improve, as a number of missed get reduce. This shows that use of the neural network is highly efficient for query set classification.

5. CONCLUSIONS

The most important role of the information retrieval is to satisfy user. Query recommendation is one of the best methods for assisting users to fulfill the user's information need by suggesting queries related to current users need by maintaining query log processing files. With the proper information of the ontology and the web usage of the web, feature vectors are developed for training the Error back propagation neural network. By the use of Error Back Propagation Neural Network classification technique the queries are handled in effective manner and consumes less time. This work improves the accuracy of the classification so the web server response time will become lesser. Precision and recall values are also good from segmentation point of view. In future, the different genetic approach for segmentation of user query can also be adopted like Firefly, PSO etc. Research work can also be done on different languages as this work is totally concentrated on text mining of English language.

6. 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.