Online Web Page Classification

Talar M. Muhammed[1*] Laith A.Flaih[2]

1. Computer Science Department University of Salahaddin, Erbil, Iraq

2.Computer Science Department , Cihan University

*E-mail of the corresponding author: talar2007hiwa@yahoo.com

*E-mail of corresponding author: d.leth1974@yahoo.com

Abstract

The World Wide Web (www) is growing at an uncontrollable rate, hundreds of thousands of web site appear every day, with the added challenge of keeping the web directories up-to-date , further the uncontrolled nature of web presents difficulties for w

eb page classification , the proposed system about using neural network technique for automatically online web pages classification according to their domain , the system provide ability to online web page classification ,which make the system sensible to any change happen to the website .

Keywords: Classifier, Neural Network classifier, web crawling, data cleaning.

1. Introduction

One of the most common theme in analyzing complex data is the classification, or categorization, of elements, the task is to classify a given data instance into a prespecified set of categories, the classification work by given a set of categories (subjects, topics) and a collection of text documents, the process of finding the correct topic (or topics) for each document[1],.

Classification problem occurs when an object needs to be assigned into a predefined group or classified based number of observed attribute related to that object, many ideas have emerged over the years on how to achieve quality results from web Classification systems, thus there are different approaches that can be used to a degree such as Clustering, Na ïve Bays (NB) and Bayesian Networks, Neural Networks (NNs), Decision Trees (DTs), Support Vector Machines (SVM) etc…,[2,3].

NN are powerful techniques for representing complex relationships between inputs and outputs. based on the neural structure of the brain [4], NNs are complicated and it can be used enormous for certain domains, containing a large number of nodes and synapses [5,6].

In this research, a new approach based on rule based NN classification has been proposed ,the approach starting by analyzing the URL which entered by the user ,takes the advantage of web mining, text mining, and data mining tasks, the proposed method work by on line assign the web documents to there categorizes ,this work belongs to the category of multi-class classification, with multiple classes or categories mean the documents belongs two more than one categories.

2. Classifier

Pattern recognition has a wide variety of application in many different filed, such that it is not possible to come up with a single classifier that give good results in all the cases, classifier operates well on different aspects of training or test feature vector[7].

Classifiers are functions which partition a set into two classes (for example, the set of rainy days and the set of sunny days), classifiers appear to be the most simple nontrivial decision making element so their study often has implications for other learning system classifiers are sufficiently complex that many phenomena observed in machine learning (theoretically or experimentally) can be observed in the classification setting, classifiers are simple enough to make their analysis easy to understand, this combination of sufficient yet minimal complexity for capturing phenomena makes the study of classifiers especially fruitful ,the classifiers are built which learn from a pre labeled set of data the characteristics of the categories, the performance of commonly used classifiers varies depending on the data and the nature of the task[9,10].

3.Neural Network Classifiers

NN are powerful techniques for representing complex relationships between inputs and outputs, based on the neural structure of the brain [8], conventional search procedure for training neural classifier equally use all the examples to minimize a sample estimate of the selected cost function however the real problem is to define appropriate classification borders, which is not exactly equivalent to any of these procedure but the key is to obtain good generalization[11].

Advantages of NN, however, include their high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained. In addition, several algorithms have recently been developed for the extraction of rules from trained neural networks. These factors contribute towards the usefulness of neural networks for classification in data mining, NN can be built to perform text categorization, usually the input nodes of the network receive the feature values, the output nodes produce the categorization status values, and the link weights represent dependence relations [1, 13].

4.The Proposed System

In this section we will try to convey the component of the proposed system , the below diagram show the main component of the proposed system ,as shown below in Figure 4

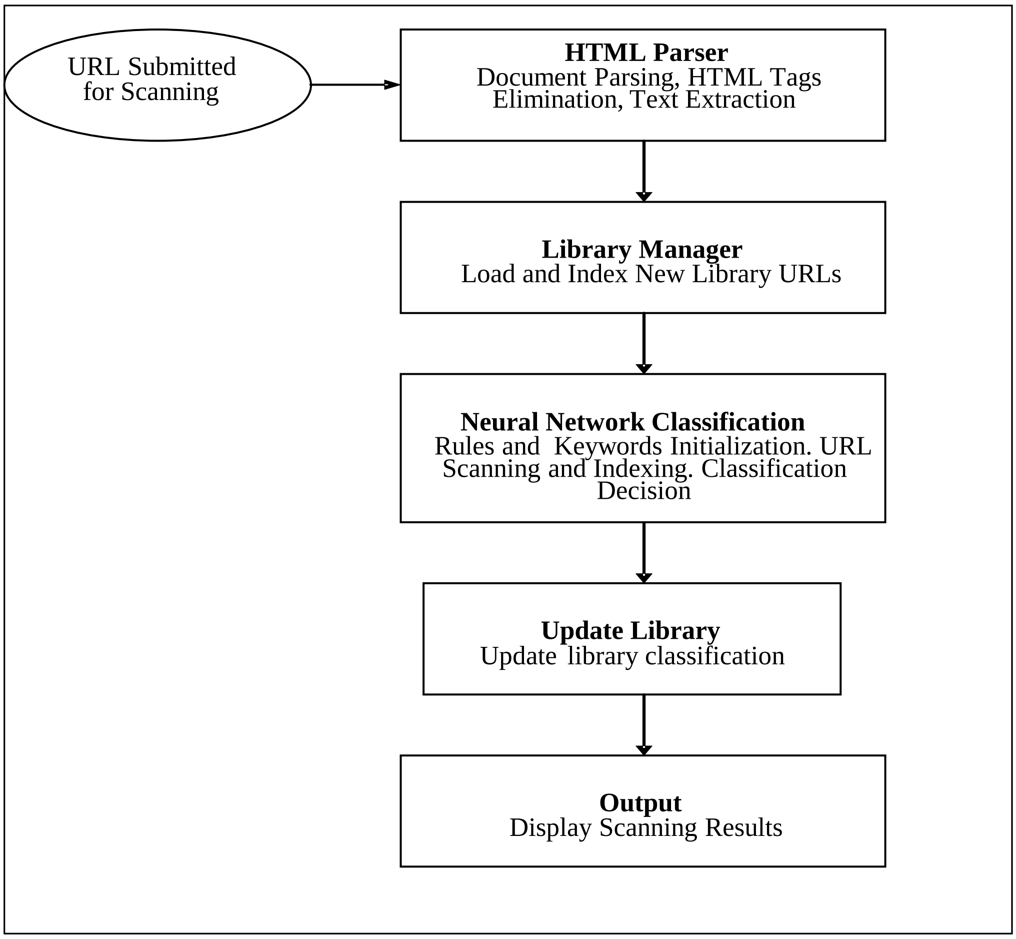

URL Submitted for Scanning
HTML Parser
Document Parsing, HTML Tags Elimination, Text Extraction
Library Manager
Load and Index New Library URLs
Neural Network Classification
Rules and Keywords Initialization. URL Scanning and Indexing. Classification Decision
Update Library
Update library classification
Output
Display Scanning Results

Figure 1.Main block diagram of proposed system

4.1.Web Crawler

Web crawler is a relatively automated program, or script that methodically scans or crawls through Internet pages to create an index of the data it's looking for, alternative names for a web crawler include web spider, web robot, bot , crawler, and automatic indexer, there are many different uses for a web crawler. Probably the most common use associated with the term is related to search engines, web crawler is an automated program searches the HTML document for given web Site in Internet and it stores a copy of page that has been visited by search engines and indexing is performed indexing means the data regarding HTML document is collected and parsed this is used to make information retrieval fast and accurate.

4.2 HTML Parser

After the URL has been entered by the user and submitted to the system ,HTML parser will handle the task of URL Validator , it check if URL entered by the user is valid or not ,If the URL not valid then the system abort scanning process and stop, otherwise the system continue with the task of HTML parsing ,in this step the system parsing the valid URL ,this step involve HTML documents parsing , by parsing HTML documents to tag and text and attribute ,then the system continue by checking if the exiting memory exceed for storing parsed documents or not if the memory exceed then reallocated memory step execute otherwise the system execute the text extractor step, text extractor handle the task of extract the text from documents and , cleaning (removing non-HTML files from the document) because our proposed system only work with HTML files , the system save the documents' term in array classification model,figure1.1 show how bock diagram of HTML parser:

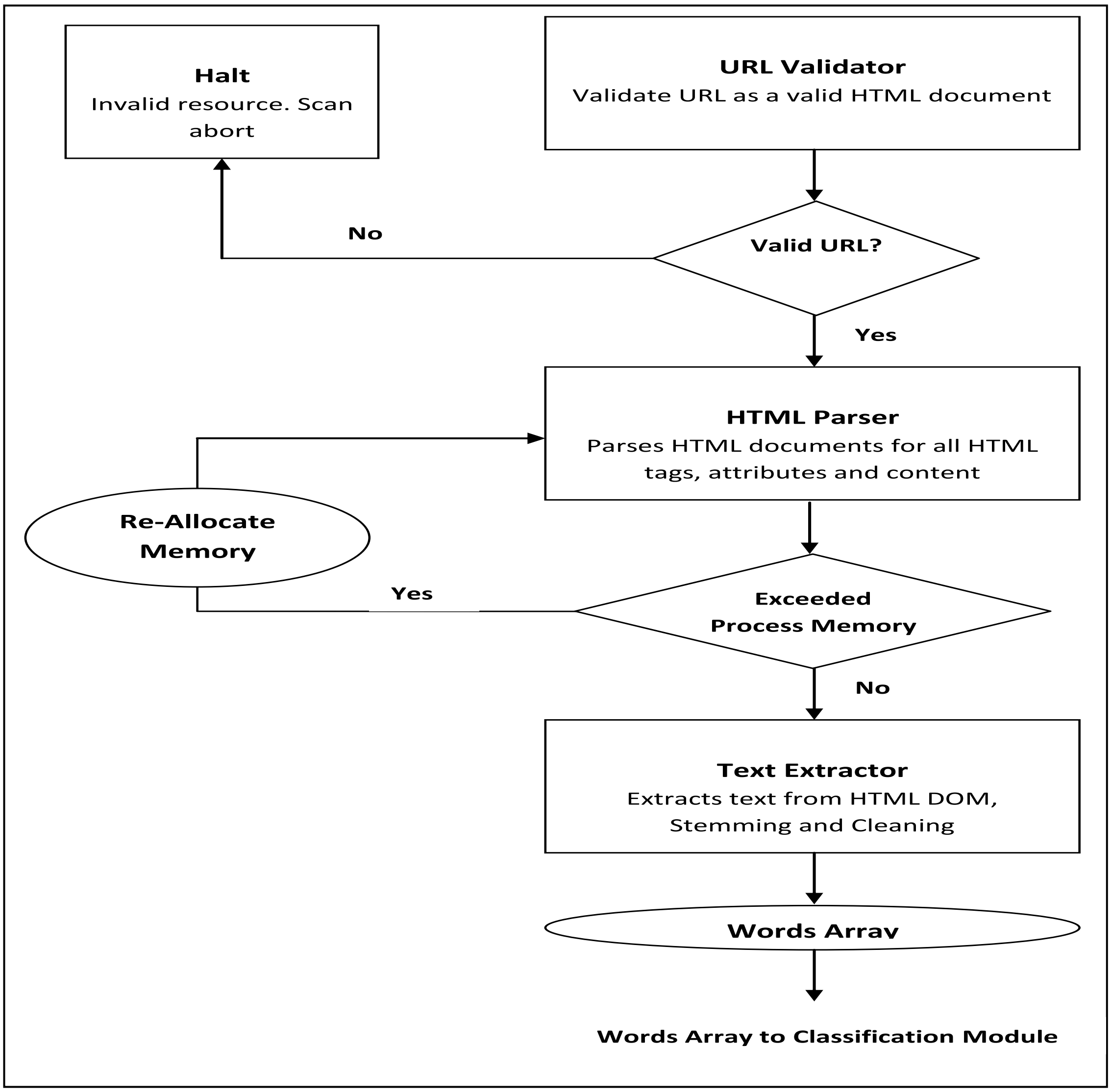

Halt
Invalid resource. Scan abort
URL Validator
Validate URL as a valid HTML document
No
Valid URL?
Yes
HTML Parser
Parses HTML documents for all HTML tags, attributes and content
Re-Allocate Memory
Yes
Exceeded Process Memory
No
Text Extractor
Extracts text from HTML DOM, Stemming and Cleaning
Words Array
Words Array to Classification Module

Figure2. Show Block Diagram of HTML parser

4.2.1 Stemming

Stemming refers to the process of reducing words to their stems or roots ,a stem is the portion of a word that is left after removing its prefixes and suffixes. In English, most variants of a word are generated by the introduction of suffixes (rather than prefixes). Thus, stemming in English usually means suffix removal, or stripping. For example, “computer”, “computing”, and “compute” are reduced to “comput”. “walks”, “walking” and “walker” are reduced to “walk”. Stemming enables different variations of the word to be considered in retrieval, which improves the recall. There are several stemming algorithms, also known as stemmers[14].

4.3 Library manger

After the HTML document parsed ,the system handle the task of library loader ,library loader work by creating an counter for saving all URL , the system continue by checking the counter, if the counter value <0 then load the exiting URL since last scan otherwise the system continue by execute Rules initialization step which composed of loading the(keywords, categorizes ,URL and the ,rules) as array to be ready for entering to classifier model, Figure 3 show block diagram of the library manger

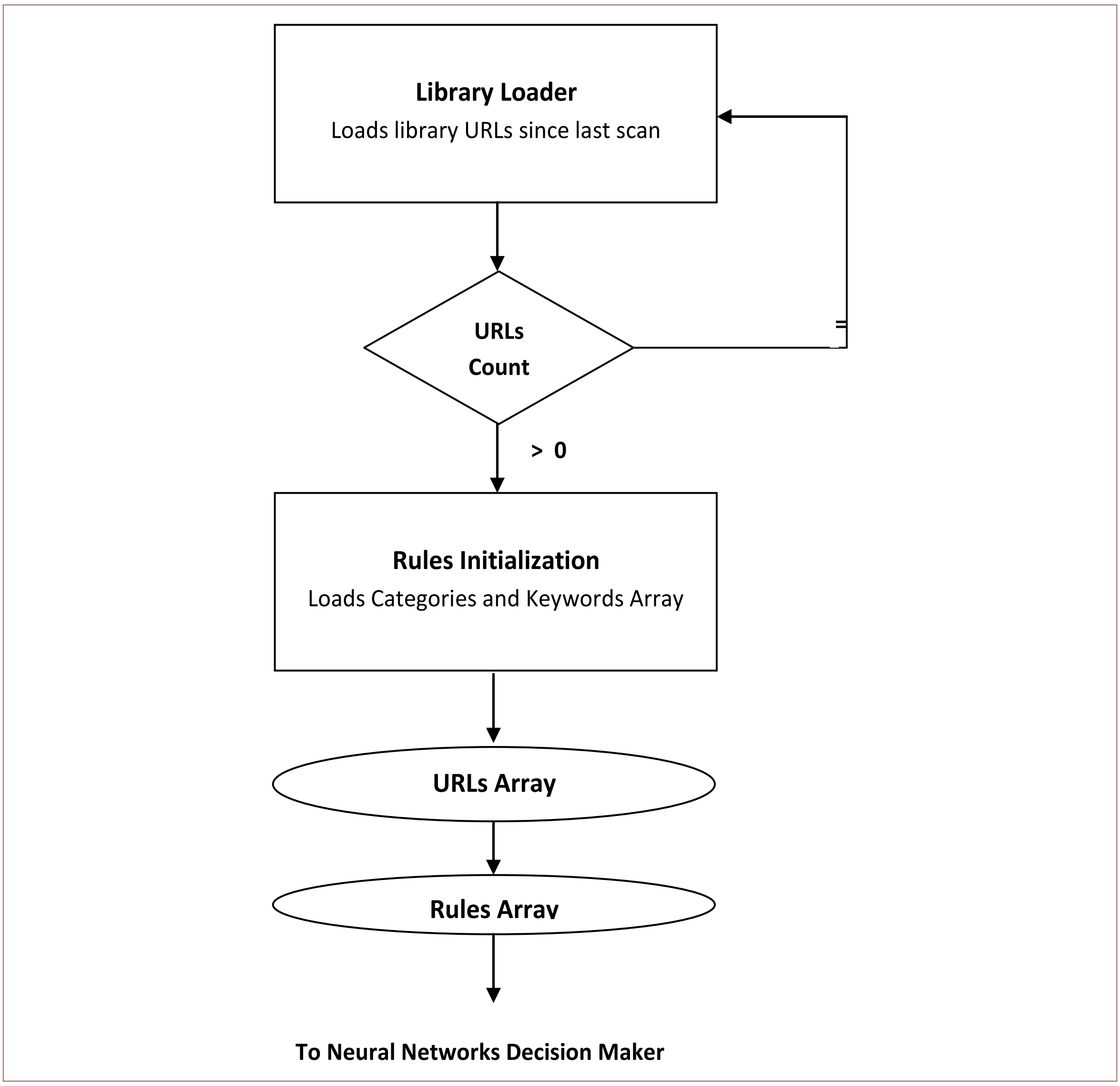

Library Loader
Loads library URLs since last scan
URLs
Count
=
> 0
Rules Initialization
Loads Categories and Keywords Array
URLs Array
Rules Arrav
To Neural Networks Decision Maker

Figure 3.Show Block Diagram of Library Manger

4.4 Neural Network Classifier:

In this step the(URLs , word, rule) arrays are submitted to neural classifier model, the system handle the task of rule scanner, in this step the classifier scan for occurrence of the rules in word array ,and check whether the rule satisfy the current word array or not if the rule not occurs in word array then the system scan other word array, but if the rule accrue in word array then the system create the new rule pairs ,and handle task of classification register in this step the system save rule pair in database and check the rule pair score, if it have greater score than other exiting rule the update rule score and create log operation otherwise, show the output(classification result) Figure 4 explain how the neural classifier work in detail:

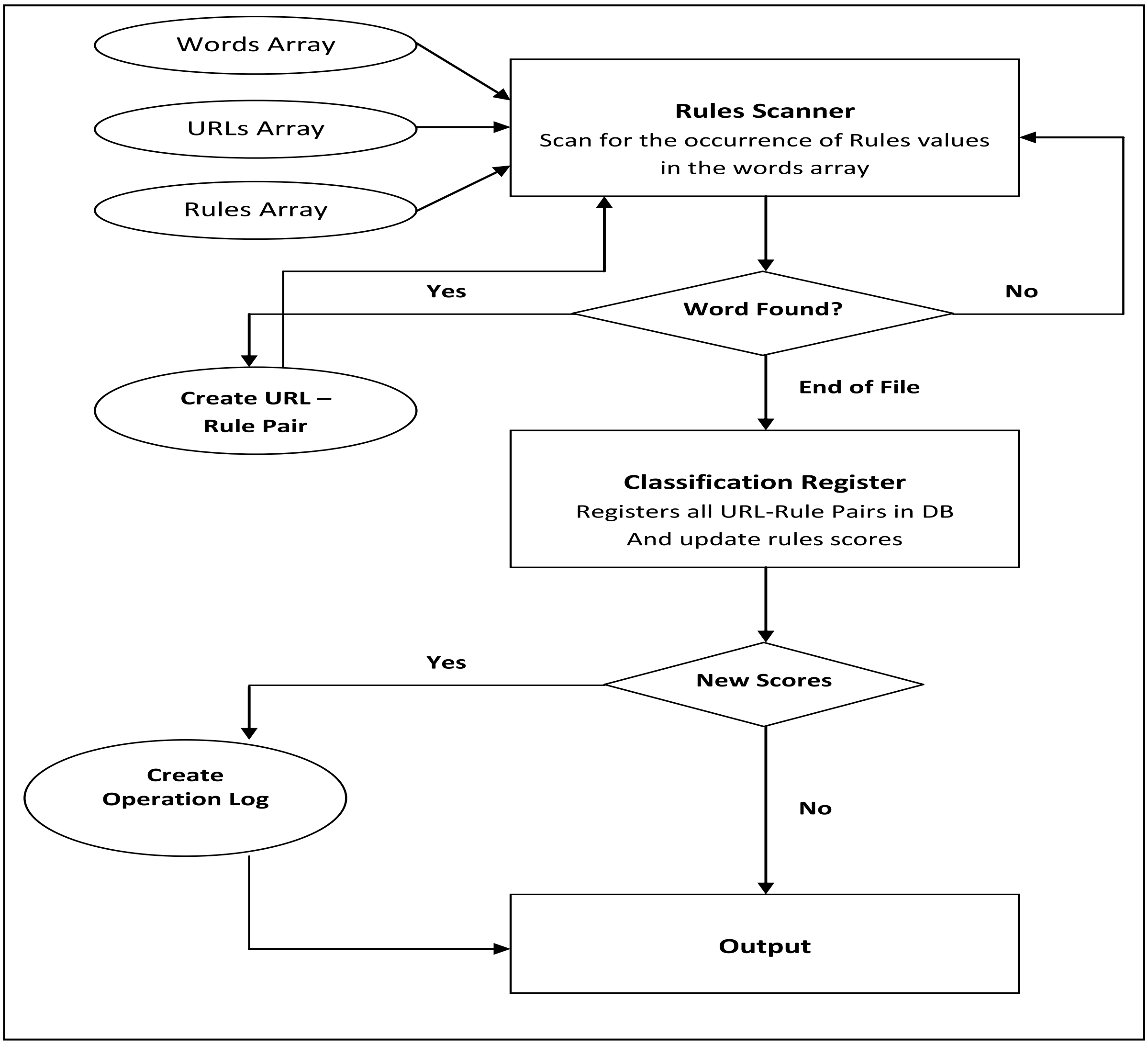
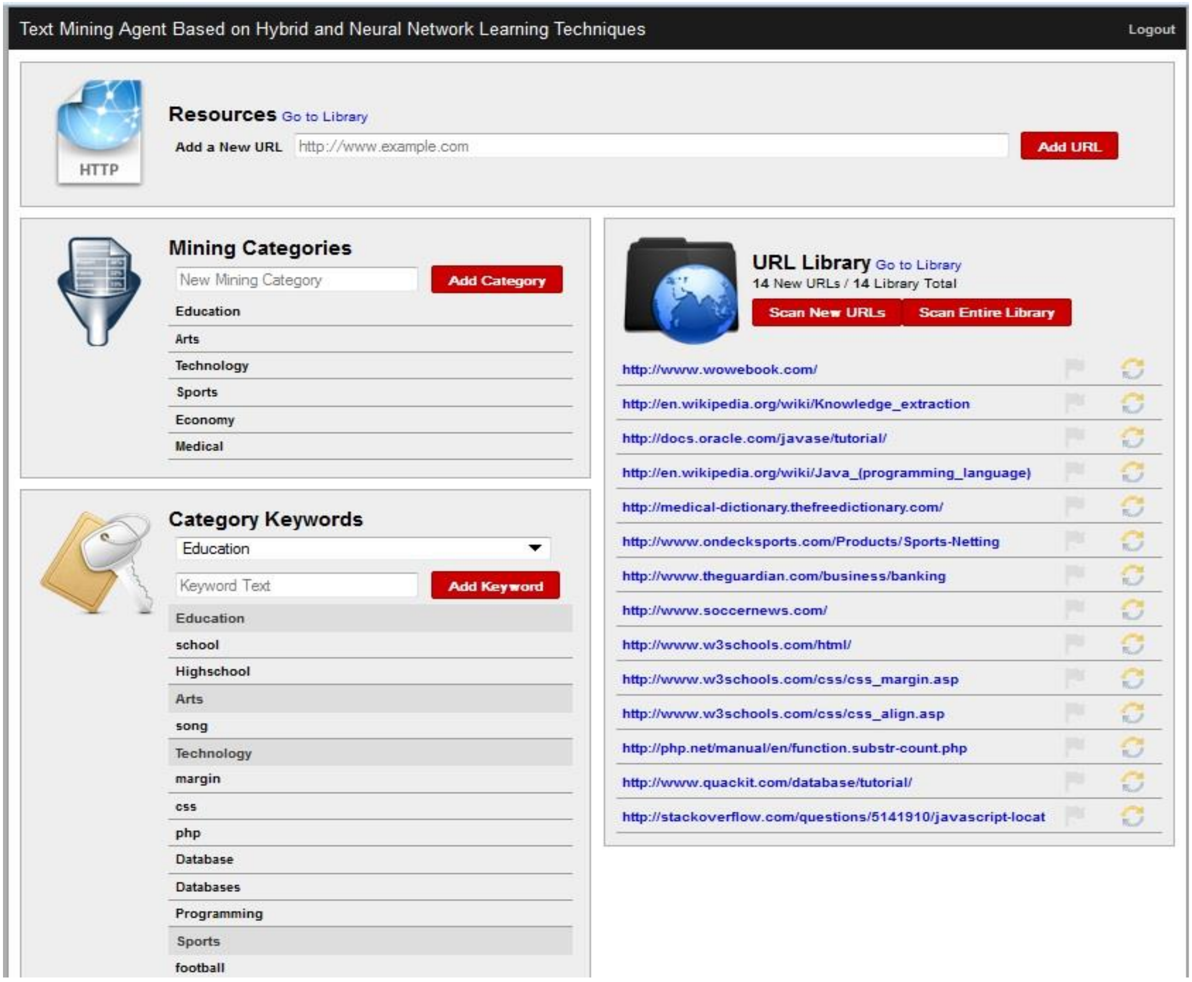


Figure 4.Show Bock Digram of Neural Network Decision Maker

5 system implementation:

In this section we will specify how the system is implemented, Figure(5.1)the main page of system in which consists of:



Education
Arts
Technology
Sports
Economy
Medical

Education
school
Highschool
Arts
song
Technology
margin
css
php
Database
Databases
Programming
Sports
football

http://www.wowebook.com/
http://en.wikipedia.org/wiki/Knowledge_extraction
http://docs.oracle.com/javase/tutorial/
http://en.wikipedia.org/wiki/Java_(programming_language)
http://medical-dictionary.thefreedictionary.com/
http://www.ondecksports.com/Products/Sports-Netting
http://www.theguardian.com/business/banking
http://www.soccernews.com/
http://www.w3schools.com/html/
http://www.w3schools.com/css/css_margin.asp
http://www.w3schools.com/css/css_align.asp
http://php.net/manual/en/function.substr-count.php
http://www.quackit.com/database/tutorial/
http://stackoverflow.com/questions/5141910/javascript-locat

Figure 5.1:Main page of Proposed system

5.1 The main home page of the system is composed of the following divisions:

5.1.1 Resource

In this division the user enter the specific URL address, that he want to classify it by the system ,then click Add URL tab to submit it by the proposed system as shown in Figure 5.2

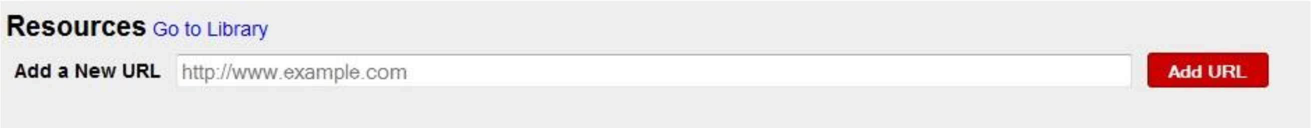


Figure 5.2. Resource section

5.1.2 URL Libraray

In this division the user add new categories to the system ,by writing the specific categories and then click Add Category tab

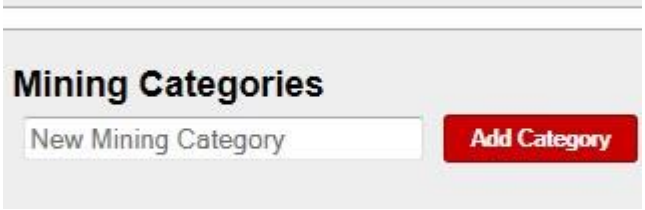


Figure 5.2. Resource section

5.1.3 Category Keyword

In this division the user specify the category name according to it the entered URL, will be classify , by then click Add category tab ,and add the keyword(term) ,which use by the system make decision about URL class, then click on Add Keyword tab to be submit by the system ,as shown in Figure 5.3

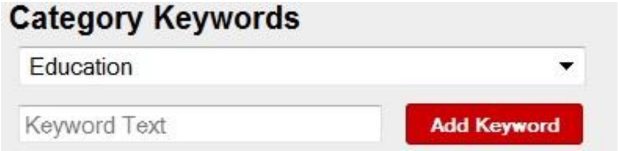


Figure 5.3.Cateogory Keyword

5.1.4 Scan

In this division the user able to run the classification model by click on Scan New URLS, clicking on this tab mean classify only the recent URL which submitted by the system ,While clicking on Scan Entire Library tab mean execute classification process for all URLs found in the system library,as shown in Figure 5.4

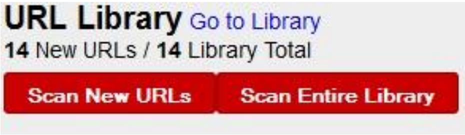


Figure 5.4.Cateogory Keyword

Conclusion and future work

With speed growth of the internet ,and the important roles of web for provide the user with wide range of information ,web page classification become very important challenge ,it make searching web information very effective and speed, in this research on line web page classification based rule based neural network classifier system was proposed ,the system provide effective approach for classified HTML documents, it able to classify the URL link entered by user according to their domain ,it also provide knowledge extraction form entered URLs by ranking them by their interests

Reference

[1] Ronen F., James ,S.(2006)." THE TEXT MINING HANDBOOK Advanced Approaches in Analyzing Unstructured Data". USA:Cambridge unversity

[2] Guoqiang Peter Zhang.,("2000")."Neural network for classification :A Survey, IEEE Transactions on system man ,and cybernetics-part C:Application and review",vol.(30) ,NO.4

[3] Daniela X.,et all.,(2009)" Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages ", International Journal of Computer Science Issues, Vol.(4), No 1.P: 16-23

[4] Russell, S. & Norvig, P. Artificial Intelligence: A Modern Approach, London: Prentice Hall, 2003.

[5] Towell, G., Shavlik, J.,(1993)."Extracting Refined Rules from Knowledge-Based Neural Network Machine Learning". International Journal of Computer Science Issues Vol(13), No. 1,.P.:71-101.

[6] Fletcher, G., Hinde,C.,(1994)."Interpretation of Neural Networks as Boolean Transfer Function Knowledge-Based Systems", International Journal of Computer Science Issues ,Vol.(7), No. 3, P: 207-214.

[7] Eleni M., Audrey T,.(2008)." Real –time Web text classification and analysis of reading difficulty" university of pennsylvaina , philadphia: USA.

[8] Russell S., Norvig P.,(2003)." Artificial Intelligence: A Modern Approach", London: Prentice Hall.

[9] John L.,(2005),"Tutorial on Practical Prediction Theory for Classification", Journal of Machine Learning, P: 273–306.

[10]Karwan H.,(2004)," web page classification using neural network with CPBF". Thesis(phD), Salahaddin Unversity.

[11] Ismal F.,(2004). "examining learning algorithm for text classification in digital libraries" university of groningen department of alfa-informatica.

[12] Jiawei H.,Micheline K.,(2006)."data mining concepts and technique",2nd :USA..