Online Web Page Classification
Talar M. Muhammed1* Laith A.Flaih2
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 web 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 pre-specified 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, Naïve 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 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 non-trivial 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

![Diagram](Image)

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 checks if the URL entered by the user is valid or not. If the URL is not valid, then the system aborts the scanning process and stops. Otherwise, the system continues with the task of HTML parsing. In this step, the system parses the valid URL. This step involves parsing HTML documents to extract tag and text and attributes. Then, the system checks if the existing memory exceeds the memory for storing parsed documents or not. If the memory exceeds, the memory allocation step executes; otherwise, the text extractor step is executed. Text extractor handles the task of extracting text from documents and cleaning (removing non-HTML files from the document) because our proposed system only works with HTML files. The system saves the documents' terms in an array classification model. Figure 1.1 shows how the 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].

---

![Figure 2. Show Block Diagram of HTML parser](image-url)
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.

**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.
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:

![Flowchart of Neural Network Decision Maker](image)

Figure 4. Show Bock Digram of Neural Network Decision Maker
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.

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

---

Figure 5.1: Main page of Proposed system

Figure 5.2. Resource section

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 Category Keyword](image)

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 Category Keyword](image)

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