iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 4 **Issue:** V **Month of publication:** May 2016

**DOI:**

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# Domain Based Categorization Using Adaptive Preprocessing

Anam Nikhil[1], Supriye Tiwari[2], Ms. Arti Deshpande[3], Deepak Kaul[4], Saurabh Gaikwad[5]

***Abstract:*** ***As the number users accessing network for various purposes increases and simultaneously size of the Network and Internet traffic increase so, there is need for categorization web pages according to domain for easy access and also to improve the performance of system. As ANN provides Massive Parallelism, Distributed representation, Learning ability, Generalization ability, Fault tolerance. In real world applications, preprocessing plays a vital role of data mining process, as real data often comes from Different complex resources which may often noisy and redundant. So we are using an Artificial Neural Network (ANN) with adaptive pre-processing technique.***

***Keywords: Web classification, artificial neural network, training, adaptive pre-processing, unsupervised.***

## I. INTRODUCTION

In this section we analyze the two most important concepts involved for developing an firewall and adaptive classifier.

### *A. Firewall*

In computing, a firewall is a network security system that monitors and controls the incoming and outgoing network traffic based on predetermined security rules. A firewall typically establishes a barrier between a trusted, secure internal network and another outside network, such as the Internet, that is assumed to not be secure or trusted. Firewalls are often categorized as either *network firewalls* or *host-based firewalls*. Network firewalls are a software appliance running on general purpose hardware or hardware-based firewall computer appliances that filter traffic between two or more networks. Host-based firewalls provide a layer of software on one host that controls network traffic in and out of that single machine. Firewall appliances may also offer other functionality to the internal network they protect such as acting as a DHCP or VPN server for that network. A firewall has a set of rules which are applied to each packet. The rules decide if a packet can pass, or whether it is discarded. Usually a firewall is placed between a network that is trusted, and one that is less trusted. When a large network needs to be protected, the firewall software often runs on a dedicated hardware, which does nothing else.

A firewall protects one part of the network against unauthorized access.

*1) Different kinds of Firewalls*

*a) Packet filtering:* Data travels on the internet in small pieces; these are called packets. Each packet has certain metadata attached, like where it is coming from, and where it should be sent to. The easiest thing to do is to look at the metadata. Based on rules, certain packets are then dropped or rejected. All firewalls can do this.it is known as network layer.

*b) Stateful packet inspection:* In addition to the simple packet filtering (above) this kind of firewall also keeps track of connections. A packet can be the start of a new connection, or it can be part of an existing connection. If it is neither of the two, it is probably useless and can be dropped.

*c) Application-layer firewalls:* Application-layer firewalls do not just look at the metadata; they also look at the actual data transported. They know how certain protocols work, for example FTP or HTTP. They can then look if the data that is in the packet is valid (for that protocol). If it is not, it can be dropped.


Screenshot of Gufw. The firewall shows its settings for incoming and outgoing traffic.

2) *Firewall characteristics:* The following design goals for a firewall:

*a)* All traffic from inside to outside, and viceversa, must pass through the firewall.This is achieved by physically blocking all access to the local network exceptvia the firewall. Various configurations are possible, as explained later in this chapter.

*b)* Only authorized traffic, as defined by the local security policy, will be allowed to pass. Various types of firewalls are used, which implement various types of security policies, as explained later in this chapter.

*c)* The firewall itself is immune to penetration. This implies the use of a hardened system with a secured operating system. Trusted computer systems are suitable for hosting a firewall and often required in government applications.

## II. EXISTING SYSTEMS

There is an exponential increase in the amount of data available on the web recently. According to , the number of pages available on the web is around 1 billion with almost another 1.5 million are being added daily. This enormous amount of data in addition to the interactive and content-rich nature of the web has made it very popular. However, these pages vary to a great extent in both the information content and quality. Moreover, the organization of these pages does not allow for easy search. So an efficient and accurate method for classifying this huge amount of data is very essential if the web is to be exploited to its full potential. This has been felt for a long time and many approaches have been tried to solve this problem.

To start with, classification was done manually by domain experts. But very soon, the classification began to be done semi automatically or automatically. Some of the approaches used include text-categorization

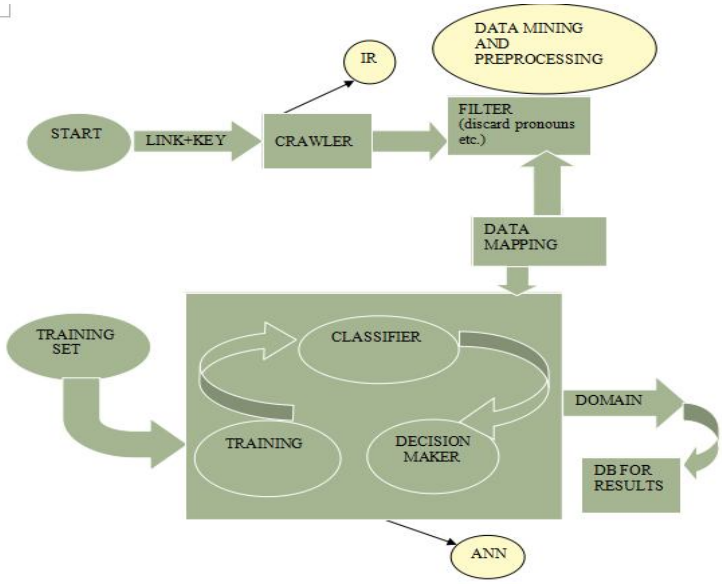
based on statistical and machine-learning algorithms , K-Nearest Neighbor approach , Bayesian probabilistic models , inductive rule learning, decision trees , neural networks and support vector machines. An effort was made to classify web content based on hierarchical structure. However, besides the text content of the web page, the images, video and other multimedia content together with the structure of the document also provide a lot of information aiding in the classification of a page. Existing classification algorithms, which rely solely on text content for classification, are not exploiting these features.

Several attempts have been made to categorize the web pages with varying degree of success. The major classifications can be classified into the following broad categories

a) Manual categorization: The traditional manual approach to classification involved the analysis of the contents of the web page by a number of domain experts and the classification was based on the textual content as is done to some extent by Yahoo. The sheer volume of data on the web rules out this approach. Moreover, such a classification would be subjective and hence open to question.
b) Clustering Approaches: Clustering algorithms have been used widely as the clusters can be formed directly without any background information. However, most of the clustering algorithms like K-Means etc. require the number of clusters to be specified in advance. Moreover, clustering approaches are computationally more expensive.
c) META tags based categorization: These classification techniques rely solely on attributes of the meta tags (<META name=''keywords''> and <META name=''description''>). However, there is a possibility of the web page author to include keywords, which do not reflect the content of the page, just to increase the hit-rate of his page in search engine results.

## III. PROPOSED SYSTEM

In this section, we present our proposal : first we introduce some concepts that we use throught the article;then we present our crawler and our pattern builder. In text-based approaches, first a database of keywords in a category is prepared as follows - the frequency of the occurrence of words, phrases etc in a category is computed from an existing corpus (a large collection of text). The commonly occurring words (called stop words) are removed from this list. The remaining words are the keywords for that particular category and can be used for classification. To classify a document, all the stop words are removed and the remaining keywords/phrases are represented as a feature vector. This document is then classified into an appropriate category using the K-Nearest Neighbor classification algorithm. These approaches rely on a number of high quality training documents for accurate classification. However, the contents of web pages vary greatly in quality as well as quantity. It has been observed that 94.65% of the web pages contain less than 500 distinct words. Also the average word frequency of almost all documents is less than 2.0, which means that most of the words in a web document will rarely appear more than 2 times. Hence the traditional method based on keyword frequency analysis cannot be used for web documents.




## IV. SYSTEM COMPONENTS

The system will be consisting of the following components: The System comprises of the following components that work in coordination in order to categorise the domain.

Crawler, Filter, Data Mapping, ANN(Artificial Neural Network), Database

*A. Crawlers*

With the explosion in size of world wide web, web search engines are becoming increasingly important as the primary means of locating relevant information.Such search engines rely on massive collections of web pages that are acquired with the help of web crawlers,which traverse the web by following hyper-links and storing downloaded pages in a large database that is later indexed for efficient execution of user queries.

A crawler for a large search engine has to address two issues:

1) *A good crawling strategy* : a strategy for deciding which pages to download next

2) *Highly optimized system* architecture :that can download a large number of pages per second while being robust against crashes, manageable, and considerate of resources and web servers. Crawler Structure For simplicity, we separate our crawler design into two main components, referred to as:

a) *Crawling Application* :The crawling application decides what page to request next given the current state and the previously crawled pages, and issues a stream of requests (URLs) to the crawling system.

b).*Crawling System*: The crawling system (eventually) downloads the requested pages and supplies them to the crawling application for analysis and storage. The crawling system is in charge of tasks such as robot exclusion, speed control, and DNS resolution that are common to most scenarios,while the application implements crawling strategies. The crawling system itself consists of several specialized components, in particular

*c) Crawl Manager*: The crawl manager is responsible for receiving the URL input stream from the applications and forwarding it to the available downloaders and DNS resolvers while enforcing rules about robot exclusion and crawl speed.

*3) .Downloader* : A downloader is a high-performance asynchronous HTTP client capable of downloading hun-dreds of web pages in parallel.

4) *DNS Resolvers* : while a DNS resolver is an optimized stub DNS resolver that forwards queries to local DNS servers. All of these components, plus the crawling application, can run on different machines (and operating systems) and can be replicated to increase the system performance. downloaded Data is then marshaled into less located in a directory determined by the application and accessible via NFS. Since a downloader often receives more than a hundred pages per second, a large number of pages have to be written out in one disk operation. We note that the way pages are assigned to these data les is unrelated to the structure of the request les sent by the application to the manager.

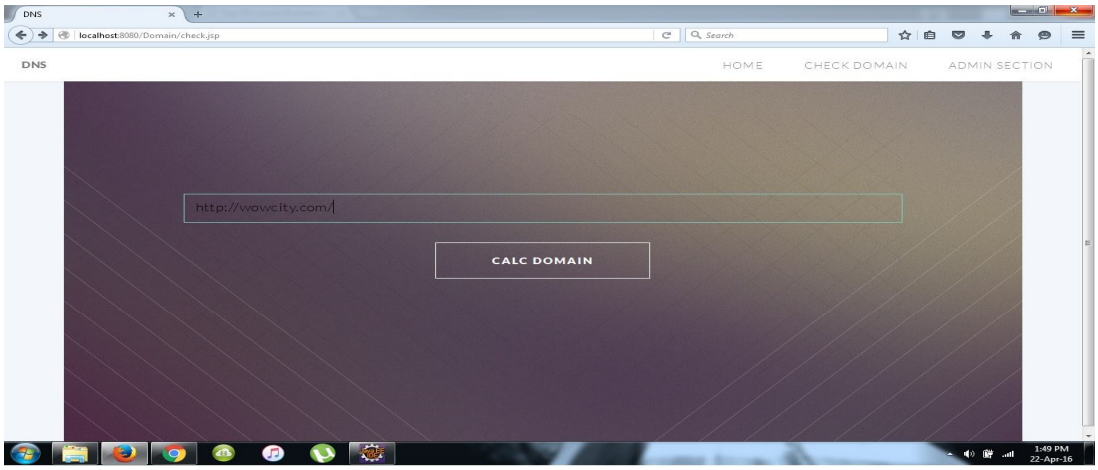
## V. ANN (ARTIFICIAL NEURAL NETWORK)

ANN's provide the system a capability to learn and expand its knowledge .There is no need to explicitly program a neural network. For instance, it can learn from training samples.Here we have used the back propogation algorithm of ANN. Artificial neural network (ANN) is a machine learning approach that models human brain and consists of a number of artificial neurons. Neuron in ANNs tend to have fewer connections than biological neurons. Each neuron in ANN receives a number of inputs. An activation function is applied to these inputs which results in activation level of neuron (output value of the neuron). Knowledge about the learning task is given in the form of examples called training examples.

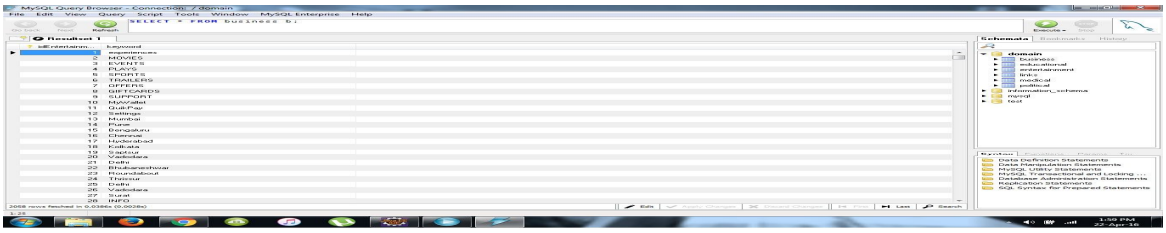
Working of the system:

The proposed system comprises of the following steps

Step 1: The user enters the URL into the GUI at the initial stage.



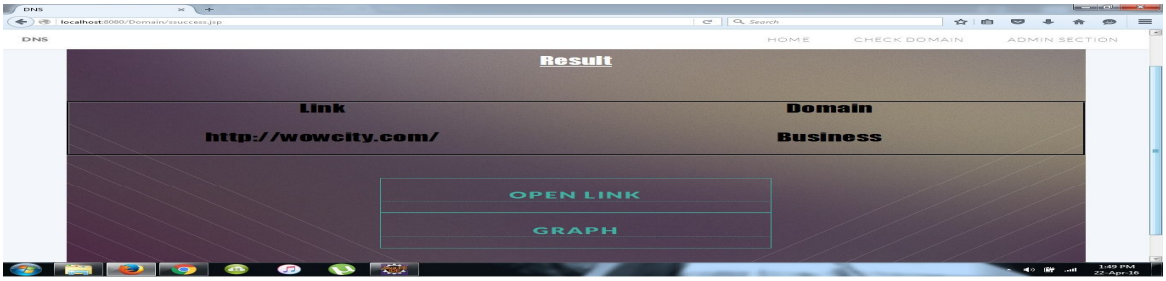

Step 2:After the link is entered the crawler retrives the information from the link and passes on to filter.Filter here discards the irrelevant data( Nouns and Pronouns).The relevant data is mapped and passed on to the classifier



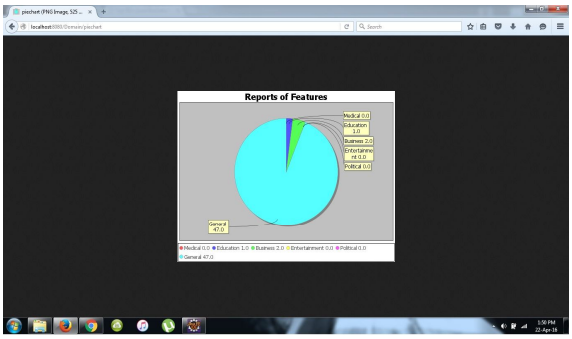
Step 3: Depending upon the data fetched the classifier categorizes the domain into one of the five categories viz:

1) Entertainment
2) Medical
3) Politics
4) Business
5) Education

Volume 4 Issue V, May 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)




Step 4: If the categorized domain is unaccessible i.e. the category has been blocked then the system displays “ACCESS IS DENIED”. Else it asks the user if he/she wants to open the link or not and simultaneously asks for the generation of the graph which specifies the number of words in a particular category.




System Configuration and Tools:

| Client Side(Recommended) | | | |
|---|---|---|---|
| | Processor | RAM | Disk Space |
| Internet Explorer -6 | Intel Pentium III or AMD-800 MHz | 256 MB | 100MB |
| Speaker | - | - | - |

| Server Side | | | |
|---|---|---|---|
| RAD | Intel Pentium III or AMD-800 MHz | 2 GB | 3.5 GB |
| DB2-9.5 | | 512 MB | 500 MB |

Software Interfaces:

Client on Internet :Web Browser, Operating System (any) .

Client on Intranet :Web Browser, Operating System (any) .

Web Server :APACHE TOMCAT, Operating System (any) .

Data Base Server :Mysql, Operating System (any).

Development End :Eclipse (J2EE, Java, Java Bean, Servlets, HTML, XML, AJAX), DB2, OS (Windows), Web Sphere(Web Server).

## VI. FUTURESCOPE

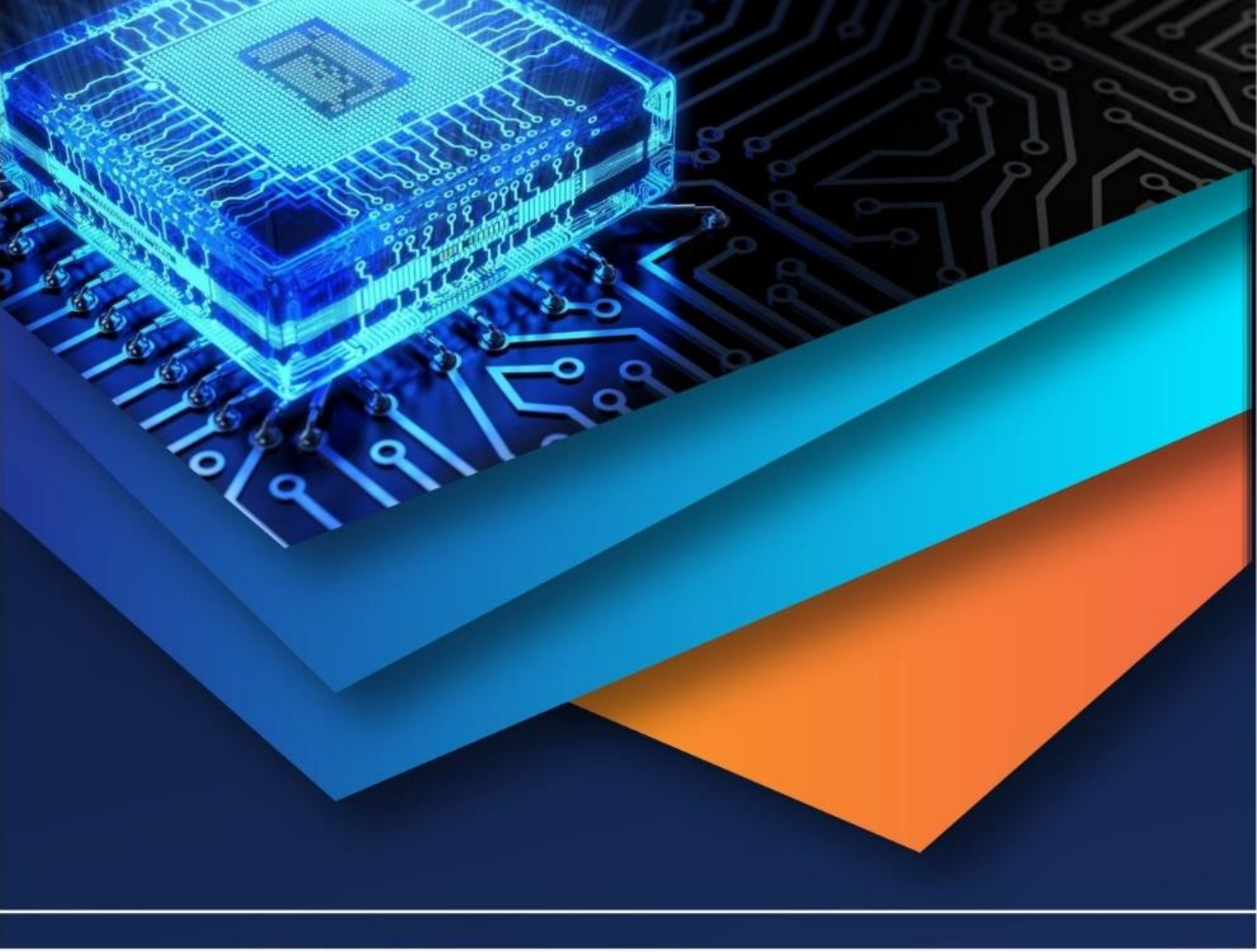
The future updation can be made possible in such way that the system will ask which category has to be blocked and will directly ask at user's GUI.The user's control over the system will increase and hence it will lead to accurate and precise output of the system.

## VII. CONCLUSION

In this article, we have presented a proposal to automatically generate content-based web page classifiers that can be used in the context of enterprise web information integration systems. Our proposal takes the URL of a web page with a keyword-based search form as input, and it outputs as category of the url.Our current implementation uses a method for classification, wherein the weights assigned to each feature are set manually during the initial training phase. A neural network based classification approach could be employed to automate the training process. Adding a few more features based on heuristics, (e.g. the classification of a page as a home page by detecting a face at the top) would increase the classification accuracy.

## REFERENCES

[1] I.Hernandez.."CALA An unsupervised URL-based web page classification system", 2013, http://www.elsevier.com/locate/knosys
[2] Indre Zliobaite Bogdan Gabry"Adaptive Preprocessing for Stream-ing Data",2014,http://dl.acm.org/citation.cfm .
[3] Arul Prakash Asirvatham , Kranthi Kumar. Ravi "Web Page Classification based on Document Structure ", 2012, http://www.cs.utah.edu/ arul- papers / webpage classif ication:pdf .



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY