International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

# A Survey on Domain Name Categorization Using Artificial Neural Networks

**Akshay S. Dhomble[1], Disha Deotale[2]**

[1]ME Computer (Networks), Savitribai Phule Pune University, G.H Raisoni College of Engg. & Technology, Wagholi, Pune, India

[2] Assistant Professor, Department of Computer Engineering, Savitribai Phule Pune University, G. H .Raisoni Collage of Engg. & Techonology, Wagholi, Pune, India

**Abstract:** ***This paper describes automatic domain name categorization using Artificial Neural Networks (ANN). The categorization of Web page is one of the challenging tasks in the world of ever expanding web innovations. There are numerous methods for categorization of Web pages using distinct approach and features. In this paper using an Artificial Neural Networks (ANN) proposes another dimension in the way of Web pages categorization through extracting the features automatically. The entire procedure of the proposed system is done in three progressive stages. In the first stage, the features are consequently extricated through breaking down the wellspring of the web pages. The second stage incorporates altering the information estimations of the neural system; every one of the qualities stay somewhere around 0 and 1. The variations in those qualities affect the output. At long last the third stage decides the class of a certain web page out of seven predefined classes. This stage is done utilizing back propagation algorithm of artificial neural network. The proposed idea will encourage web mining, recoveries of data from the web furthermore the search engines.***

**Keywords:** Artificial Neural Networks, Classification, Web Mining

## 1. Introduction

A Web page categorization (or classification) also known as Domain Name Categorization (DNC) is carried out using various ways. The distinct authors was categorized the web pages using distinct ways using different techniques. In this paper we describe automatic domain name categorization using Artificial Neural Networks (ANN). In machine learning, Artificial Neural Networks are a group of models propelled by biological neural networks and are utilized to gauge or surmised functions that can rely on upon a large number of inputs and are for the most part obscure. Artificial neural networks are by and large displayed as systems of interconnected "neurons" exchange messages between each other. The associations have numeric weights that can be tuned in view of experience, making neural nets versatile to inputs and equipped for learning.

For instance, a neural system for handwriting recognition is characterized by an arrangement of input neurons which may be actuated by the pixels of an input image. In the wake of being weighted and changed by a function (dictated by the system's architect), the actuations of these neurons are then gone on to different neurons. This procedure is rehashed until at last, a yield neuron is enacted. This figures out which character was read.

Like other machine learning routines/ways - systems that gain from data - neural systems have been utilized to explain a wide variety of assignments/task that are difficult to illuminate utilizing ordinary rule - based programming, including personal computer (PC) vision and speech recognition.

Classification assumes an essential part in numerous data administration and recovery errands. On the Web, it is the essential to focused crawling when classification of page content, to the helped improvement of web directories, to point particular Web join examination, to logical promoting, and to investigation of the topical structure of the Web. Web page classification can likewise enhance the nature of Web search.

Classification is a supervised learning problem in which a classifier is prepared on a set of information labeled with predefined categories and after that connected to label future examples. It assumes a key part in various key errands on data recovery and administration On the Web, classification of page content is crucial to focused crawling, to help advancement of web directories such as those provided by Yahoo and the Open Directory Project (ODP), to topic-specific web link analysis, to analysis of the topical structure of the Web, and to contextual advertising.

A classifier is usually evaluated with regard to how accurately it can label unseen instances. The accuracy of the classifier often directly affects the performance of the system built on top of it. Inaccurate classification results will lead to an overall performance degradation, which in turn adversely affects user experience, and sometimes causes direct monetary loss. For example, in a ranking system, if an important page is incorrectly classified into a category that has no connection with the query, it will be viewed as less important by the positioning calculation, and in this manner not be positioned as high as it should be.

In contextual advertising, if a page around an auto dashing amusement is inaccurately classified as an automobile merchant, the notice coordinating system will show insignificant advertisements on the page, losing the clicks it could have attracted using correct classification. Since numerous data recovery undertakings rely on upon accurate classification, research in advanced classification approaches

will advantage systems that inquiry or oversee information on the Web, as well as other types of information in general.

## 2. Concept of Artificial Neural Networks

A machine learning approach that is artificial neural network (ANN) that models human brain and it comprises of various artificial neurons. Neurons in ANNs have a tendency to have less connection than biological neurons. Every neuron in ANN gets various inputs. An actuation function is connected to these inputs which bring about initiation level of neuron (yield estimation of the neuron). Information about the learning undertaking is given as samples called training examples.

An Artificial Neural Network is indicated by:

- **Neuron model:** Is the data preparing unit of the Neural Networks (NN) in Artificial Neural Networks (ANN)
- **Architecture:** A set of neurons, inputs and its associating connections towards neurons. Every connection has a weight.
- **A learning algorithm:** It's utilized for modifying so as to prepare the Neural Networks the weights keeping in mind the end goal to show a specific learning assignment effectively on the training examples.

The aim of Artificial Neural Networks is to acquire a Neural Networks that is prepared and sums up well. And also it should behave effectively on new occurrences of the learning task. The neuron is the fundamental data processing unit of a Neural Networks.

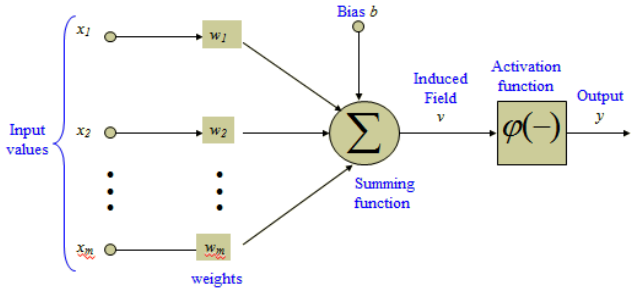


**Figure 1:** A Neuron Diagram

It consists of following three main points.

1. A set of links that are describing the neuron inputs with weights $W_1, W_2, \ldots, W_m$
2. An adder function (straight combiner) for processing the weighted entirety of the inputs: (Real numbers)

$$u = \sum_{j=0}^{m} wjxj \quad (1)$$

3. Activation functions $\varphi$ for constraining the sufficiency of the neuron yield. Here "b" signifies bias

$$Y = \varphi(u + b) \quad (2)$$

In figure 1 bias ***b*** has of applying a change to the weighted sum ***u***

$$v = u + b \quad (3)$$

The bias is an outside parameter of the neuron. It can be modeled by including an additional input, ***v*** is called **induced field** of the neuron.

$$v = \sum_{j=0}^{m} wjxj \quad (4)$$

$$w0 = b$$

## 3. Network Architecture of ANN

A neural network is a capable information displaying instrument that has the capacity catch and speaks to complex data/yield connections. The inspiration for the advancement of neural network innovation originated from the yearning to add to an artificial system that could perform "brilliant" endeavors like those performed by the human mind. Neural systems take after the human mind in the accompanying two ways:

- A neural network procures knowledge through learning.
- A neural network's knowledge is put away inside between neuron association qualities known as synaptic weights.

The genuine power and point of preference of neural networks lies in their capacity to speak to both straight and non-direct connections and in their capacity to gain these connections straightforwardly from the information being displayed. Conventional linear models are essentially deficient with regards to demonstrating information that contains non- linear characteristics.

In the model of neural networks there are three distinct classes of network architectures is as follows:

1) Single-layer feed-forward
2) Multi-layer feed-forward
3) Recurrent

Neural network architecture is linked with the learning algorithm used to train. The following three figures are differentiates architecture of Neural Networks (NN).

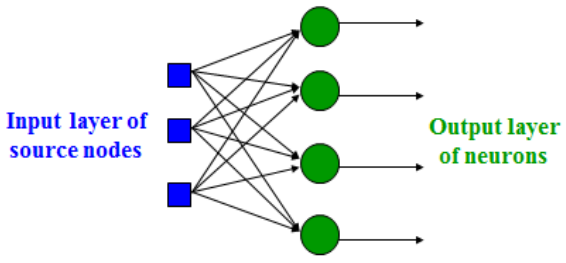


**Figure 2:** Single-layer feed-forward Architecture

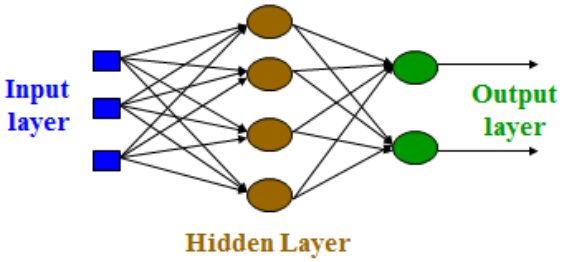


**Figure 3:** Multi-layer feed-forward Architecture

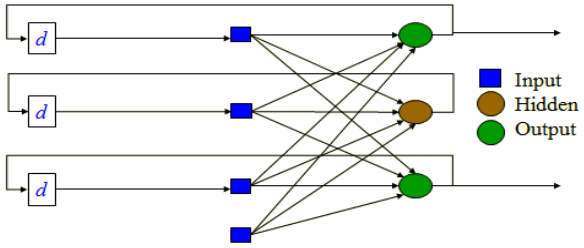


**Figure 4:** Recurrent Architecture

The most widely recognized neural networks model is the multilayer perceptron (MLP). This kind of neural networks is known as a supervised system on the grounds that it requires a wanted yield with a specific end goal to learn. The objective of this kind of system is to make a model that effectively maps the info to the yield utilizing historical data so that the model can then be utilized to create the yield when the craved yield is obscure.

The MLP and numerous other neural networks learn utilizing a calculation called back-propagation. With back-propagation, the input data is over and again introduced to the neural network. With every presentation the output of the neural network is compared to the desired output and an error is processed. This error is then sustained back (back-propagated) to the neural network and used to adjust the weights such that the mistake diminishes with every cycle and the neural model gets closer and closer to delivering the fancied output. This procedure is known as "training".

A decent approach to acquaint the topic is to take a look at a typical application of neural networks. A number of today's document scanners for the personal computer accomplished programming that performs a undertaking known as optical character recognition (OCR). OCR programming permits you to examine in a printed record afterward convert the checked image into to an electronic text format for example, a Word document, empowering you to control the text. With a specific end goal to perform this conversion the software must analyze every group of pixels (0's and 1's) that shape a letter and create a value that relates to that letter. Some of the OCR software on the business sector utilizes a neural network as the classification engine.
Obviously character recognition is not by any means the only issue that neural networks can solve. Neural networks have been effectively applied to broad spectrum of data-intensive applications, for example,

- **Process Modeling and Control** - Developing a neural network model for a physical plant then utilizing that model to decide the best control settings for the plant.
- **Machine Diagnostics** - Detect when a machine has fizzled so that the system can naturally close down the machine when this happens.
- **Portfolio Management** - Allocate the benefits in a portfolio in a way that augments return and minimizes hazard.
- **Target Recognition** - Military application which utilizes video and/or infrared image data to figure out whether an adversary target is available.
- **Medical Diagnosis** - Analyzing so as to assist specialists with their finding the reported indications and/or image data such as MRIs or X-rays.
- **Targeted Marketing** - Finding the set of demographics which have the most elevated reaction rate for a specific marketing effort.
- **Credit Rating** - Automatically assigning an organization's or individual's credit rating based on their money related condition.
- **Voice Recognition** - Transcribing talked words into ASCII text.
- **Financial Forecasting** - Using the historical data for the security foresee the future development of that security.
- **Quality Control** - Attaching a camera or sensor to the end of a generation procedure to naturally examine for deformities.
- **Intelligent Searching** - An internet search engine that gives the most applicable substance and pennant promotions in light of the clients' past conduct.
- **Fraud Detection** - Detect fake Visa exchanges and nature.

## 4. Literature Survey

In [1], he stated that web page categorization is one of the challenging tasks in the world of ever increasing web technologies. For web page categorization there are many ways based on different approach and features. He's proposes a new dimension in the way of categorization of web pages using artificial neural network (ANN) through extracting the features automatically.

In [2], these authors mainly focused on traditional textual classification. He said that, the automated classification of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of reclassified documents, the characteristics of the categories.

In [3], these authors reviewed Web mining research in general as opposed to concentrating on classification. He said that, classification of Web page content is essential to many tasks in Web information retrieval such as maintaining Web directories and focused crawling. The uncontrolled nature of Web content presents additional challenges to Web page classification as compared to traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process.

In [4], this author reviewed a number of text-learning intelligent agents, some of which are Web-specific. However, her focus was on document representation and feature selection. This author focuses on three key criteria: what representation the Particular application uses for documents, how it selects features, and what learning algorithm it uses. She then describes personal web watcher, a content based intelligent agent that uses text-learning for User-customized web browsing.

In [5], numerous datasets of interest today are best portrayed as a linked collection of interrelated objects. These may

speak to homogeneous networks, in which there is a single-item sort and link sort, or richer, heterogeneous networks, in which there may be numerous item and link sorts (and potentially other semantic data). These authors reviewed a data mining techniques which explicitly consider links among objects, with Web classification being one of such areas.

In [6], this author stated that classification of web page content is essential to many tasks in web information retrieval such as maintaining web directories and focused crawling. The uncontrolled nature of web content presents additional challenges to web page classification as compared to traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process. He's also reviewed a various aspects of Web mining, including a brief discussion on the use of link structure to improve Web classification.

In [7], which described the state-of-the art techniques and subsystems used to build automatic Web page classification systems. These authors propose a new bidirectional hierarchical clustering system for addressing challenges of web mining. The key feature of our approach is that it aims to maximize the intra cluster similarity in the bottom-up cluster-merging phase and it ensures to minimize the inter-cluster similarity in the top-down refinement phase. This two-pass approach achieves better clustering than existing one-pass approaches.

In [8], these authors stated that, Virtual integration systems retrieve information according to the user's interest. This information is retrieved from several web applications, but it is presented to the user uniformly, in an online process.

**Table 1:** Comparison of Web Page Classification Approaches

| Approach | Task | Reported Improv. | Evaluation Dataset |
|---|---|---|---|
| Mladenic [1999] | Topical | N/A | Yahoo! directory |
| Chakrabarti et al. [1998] | Topical | 32% - 75% (accuracy) | Yahoo! directory |
| Furnkranz [1999] | Functional | 70% - 86.6% (accuracy) | WebKB |
| Know and Lee [2000] | Topical | 18% - 19.2% (accuracy) | Hanmir |
| Furnkranz [2001] | Functional | 70% - 86.6% (accuracy) | WebKB |
| Calado et al. [2003] | Topical | 39% - 81.6% (accuracy) | Cade directory |
| Qi and Davison [2006] | Topical | 73% - 91.3% (accuracy) | ODP directory |

## 5. Conclusion

Robotized classification of web pages can prompt better web retrieval tools with the included comfort of selecting among legitimately sorted out registries. In this paper a topic based web page categorization is proposed which extract the features consequently through breaking down the html source, and classify the pages into few major classes utilizing back propagation algorithm. The web site pages are categorized in light of five major characteristics and likenesses of distinctive pages of same types.

## References

[1] S. M. Kamruzzaman, Web Page Categorization Using Artificial Neural Networks, Proceedings of the 4th ICEE and 2nd APM, January 2006.
[2] Joachims, Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 147.
[3] Chakrabarti, Kosala, Blockeel, Web Page Classification: Features and Algorithms, ACM Computing Surveys, Vol. 41, No. 2, Article 12, Publication date: February 2009.
[4] Mladenic, Text-Learning and Related Intelligent Agents: A Survey, IEEE intellegent system, July/August 1999.
[5] Getoor, Diehl, Link Mining: A Survey, SIGKDD Explorations, Vol. 7, Issue 2.
[6] Furnkranz, Web Page Classification: Features and Algorithms, Department of Computer Science and Engineering Lehigh University June 2007.
[7] Choi, Yao, Web Page Classification, Computer Science, College of Engineering and Science Louisiana Tech University, Ruston, LA 71272, USA.
[8] Vinod Anupam, Juliana Freire, Bharat Kumar, Daniel F. Lieuwen, Automating web navigation with the WebVCR, Comput. Netw. 3 (1-6) (2000) 503517, http://dx.doi.org/10.1016/S1389- 286(00)00073-6.
[9] Arvind Arasu, Hector Garcia-Molina, Extracting structured data from web pages, in: SIGMOD Conference, 2003, pp. 337348, http://dx.doi.org/10.1145/872757.872799.
[10] Ziv Bar-Yossef, Idit Keidar, Uri Schonfeld, Do not crawl in the DUST: Different URLs with similar text, TWEB 3 (1) (2009) 3, http://dx.doi.org/10.1145/1462148.1462151.
[11] Ziv Bar-Yossef, Sridhar Rajagopalan, Template detection via data mining and its applications, in: WWW, 2002, pp. 580591, http://dx.doi.org/10.1145/511446.511522.
[12] Eda Baykan, Monika Henzinger, Ludmila Marian, Ingmar Weber, A comprehensive study of features and algorithms for URL-based topic classification, TWEB 5 (3) (2011) 15, http://dx.doi.org/10.1145/ 1993053.1993057.
[13] Eda Baykan, Monika Rauch Henzinger, Ludmila Marian, Ingmar Weber, Purely URL-based topic classification, in: WWW, 2009, pp. 11091110, http://dx.doi.org/10.1145/1526709.1526880.
[14] Florian Beil, Martin Ester, Xiaowei Xu, Frequent term-based text clustering, in:KDD, 2002, pp. 436442, http://dx.doi.org/10.1145/775047.775110.
[15] Asa Ben-Hur, David Horn, Hava T. Siegelmann, Vladimir Vapnik, Support vector clustering, J. Mach. Learn. Res. 2 (2001) 125137. doi: 10.1.1.22.6663.
[16] Adam L. Berger, Stephen Della Pietra, Vincent J. Della Pietra, A maximum entropy approach to natural language processing, Comput. Linguist. 22 (1) (1996) 3971. doi: 10.1.1.103.7637.