IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 1 (Sep-Oct. 2012), PP 20-25
www.iosrjournals.org

On The Automated Classification of Web Pages Using Artificial Neural Network

[1]Pikakshi Manchanda, [2]Sonali Gupta, [3]Komal Kumar Bhatia
[1,2,3] Department of Computer Science, YMCA University of Science & Technology, Faridabad, India

Abstract: *The World Wide Web is growing at an uncontrollable rate. Hundreds of thousands of web sites 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. As the number of Internet users is growing, so is the need for classification of web pages with greater precision in order to present the users with web pages of their desired class. However, web page classification has been accomplished mostly by using textual categorization methods. Herein, we propose a novel approach for web page classification that uses the HTML information present in a web page for its classification. There are many ways of achieving classification of web pages into various domains. This paper proposes an entirely new dimension towards web page classification using Artificial Neural Networks (ANN).*

Index Terms: *World Wide Web, Web page classification, textual categorization, HTML, Artificial Neural Networks, ANN.*

I. INTRODUCTION

The World Wide Web is a huge repository of information that has been growing exponentially over the years. This rapid growing nature of the web has led to the invention of various techniques for managing the vast amount of content available online in order to realize its potential as a useful information resource [13, 22]. The various techniques that have been invented for managing the information content available online include a number of automatic categorization techniques of web pages into different classes or categories [7, 10]. Automatic categorization of web pages has been studied extensively, and most of these categorization techniques are usually based on similarity between documents contents or their structures [1] [2] [3].

Web page classification may be considered as an important process for managing various web directories such as Yahoo! [8], Looksmart [9], and the Open Directory Project [4]. These web directories may be considered as a way to reach the Web documents [20]. Typically these directories have been manually created over the years. In other words, the decisions regarding the category to which a web page belongs have been done by human editors. As a result, creation, management and maintenance of these web directories have been a time-consuming and cumbersome process. It is simply not feasible to keep up with the pace of growth and change on the web through manual classification without expending immense time and effort. It has, therefore, desirable to be able to learn an automatic classifier that tests membership to a given category [32].

Web page classification, also known as web page categorization, is the process of assigning a web page to one or more predefined categories. Classification is often posed as a supervised learning problem (Mitchell 1997) in which a set of labelled data is used to train a classifier, which is then applied to label future samples [26].

The general problem of web page classification can be further divided into multiple sub-problems such as subject classification, functional classification, sentiment classification, and other types of classification [26]. *Subject classification* is concerned about the subject or topic of a web page. For example, judging whether a page is about “arts”, “business” or “sports” is an instance of subject classification. *Functional classification* cares about the role that the web page plays. For example, deciding a page to be a “personal homepage”, “course page” or “admission page” is an instance of functional classification [26]. *Sentiment classification* focuses on the opinion that is presented in a web page, i.e., the author’s attitude about some particular topic. Other types of classification include genre classification [26] (e.g., (zu Eissen and Stein 2004)), search engine spam classification (e.g., (Gy¨ongyi and Garcia-Molina 2005b; Castillo, Donato, Gionis, Murdock, and Silvestri 2007)) and so on.

Based on the number of classes in the problem, classification can also be divided into binary classification and multi-class classification [26], wherein binary classification categorizes instances into exactly one of two classes (as in Fig. 1(a)); multi-class classification deals with more than two classes. If a problem is multi-class, say four-class classification, it means four classes are involved, say Arts, Business, Computers, and Sports (as in Fig. 1(b)). Classes may refer to categories here, such as 2 categories in binary classification or multiple categories in multi-class classification.

Thus, it is pretty evident that popularity and need of web page classification is very significant, not only from the point of view of academic needs for continuous knowledge growth, but also for the needs of industry for quick, efficient solutions to information gathering and analysis in maintaining up-to-date information that is critical to the business success [30]. Pioneering work has been done in the fields of classification of web pages based on their textual content, visual layout of information on a web site (i.e., placement of images, graphics, tables, forms etc. on the web page), classification into high-level domains such as informational, research, transactional etc.

.

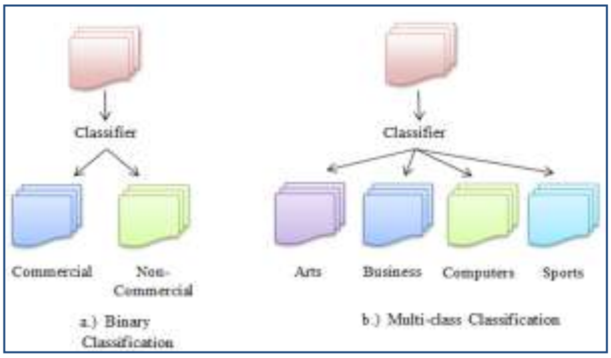

Fig 1. Types of Classification

Existing algorithms have been relying mainly over the usage of text content of a web page in order to identify its domain and classify them accordingly. However web pages include a lot more information other than text content based on which their domains may be identified. One such source of information, apart from the afore-mentioned sources of information, is the information provided by certain HTML elements, such as meta-tags, body tag, title tag etc. that may help in classification of web pages. The presence of additional information, provided by the HTML tags and the hyperlinks gives the researchers idea of exploring new techniques for representing Web sites for automatic classification.

Further, for the purpose of classification of web pages, methods such as Support Vector Machines, Naïve's Bayesian trees, Decision Trees etc. have been used.

In this paper, a novel approach based on Artificial Neural Network (ANN) has been proposed that utilizes the information provided by HTML elements of a web page in order to identify its domain. Our work involves classifying web pages into a number of domains as specified by the user. Also, our work belongs to the category of multi-class classification, with multiple classes or categories such as entertainment, food, medicine, sports, education and the like.

II. RELATED WORK

Classification plays a vital role in many information management and retrieval tasks. On the Web, classification of page content is essential to:

a) assisted development of web directories,
b) help improve the quality of web search,
c) topic-specific web link analysis,
d) analysis of the topical structure of the Web,
e) satisfy the information needs of a large number of Internet users,
f) helping question answering systems, and
g) focused crawling [26].

A. Web Page Classification

This involves classifying the web pages based on various parameters such as text, image, structure of the document etc. The classification mechanisms that have been used so far for web page classification are: [2, 31]

1) Manual classification by domain specific experts,
2) Clustering approaches (manual or automated),
3) Link and Content Analysis.

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 Bayes (NB) and Bayesian Networks, Neural Networks (NNs), Decision Trees (DTs), Support Vector Machines (SVMs) etc. as mentioned above [30].

B. Naïve Bayes Models

NB models are popular in machine learning applications, due to their simplicity in allowing each attribute to contribute towards the final decision equally and independently from the other attributes. This simplicity equates to computational efficiency, which makes NB techniques attractive and suitable for many domains. However, the very same thing that makes them popular is also the reason given by some researchers, who consider this approach to be weak. The conditional independence assumption is strong, and makes NB-based systems incapable of using two or more pieces of evidence together, however, used in appropriate domains; they offer quick training, fast data analysis and decision making, as well as straightforward interpretation of test results [30]. However, a thorough analysis of a large number of training web pages has shown us that the features used in these pages can be independently examined to compute the category for each page. Further, enhancing the standard NB rule or using it in collaboration with other techniques has also been attempted by other researchers. Addin et al in [27] coupled a NB classifier with K-Means clustering to simulate damage detection in engineering materials. NB Tree in [23] induced a hybrid of NB and DTs by using the Bayes rule to construct the decision tree. Other research works ([18], [24]) have modified their NB classifiers to learn from positive and unlabeled examples.

C. Decision Trees

Unlike NB classifiers, DT classifiers can cope with combinations of terms and can produce impressive results for some domains. Decision trees may be computationally expensive for certain domains, however, they make up for it by offering a genuine simplicity of interpreting models, and helping to consider the most important factors in a dataset first by placing them at the top of the tree [30].

The researchers in [25], [12], [21] all used DTs to allow for both the structure and the content of each web page to determine the category in which they belong. An accuracy of around 85% was achieved by all.

D. Comparison of various approaches

TABLE I
Comparison Of Various Classifiers Used For Web Page Classification

CLASSIFIER	Accuracy	Precision	Recall	F-Measure
NB Classifier	95.20%	99.37%	95.23%	97.26%
DT Classifier	94.85%	98.31%	95.90%	97.09%

The NB classifier was tested against 8725 sampling units of web pages after being trained with only 711 units by [30]. This exact same sample was also analyzed by a DT classifier and the results from all systems were compared to one-another, as shown above in table I.

Table I shows the Accuracy, Precision, Recall and F-Measure results achieved by the NB and DT classifiers. These results show that both classifiers achieve impressive results in the classification of attribute data in the training courses domain. The DT classifier outperforms the NB classifier in execution speed and Recall value (by 0.67%) [30]. However, the NB classifier achieves higher Accuracy, Precision and most importantly, overall F-Measure value, which is a very promising result.

As depicted by table I above, NB classifier outperformed DT classifier during web page classification; it, however, suffered through various drawbacks, as will be discussed in our proposed work, which encouraged us to use Artificial Neural Networks for the purpose of web page classification.

E. Neural Networks

NNs are powerful techniques for representing complex relationships between inputs and outputs. Based on the neural structure of the brain ([19]), NNs are complicated and they can be enormous for certain domains, containing a large number of nodes and synapses. There is a research that has managed to convert NNs into sets of rules in order to discover what the NN has learnt ([6], [5]), however, many other works still refer to NNs as a ‘black box’ approach ([28], [20]), due to the difficulty in understanding the decision making process of the NN, which can lead to not knowing if testing has succeeded. Researchers in [15] and [16] proposed a term frequency method to select the feature vectors for the classification of documents using NNs. A much later research [29] used NNs together with an SVM for better classification performance. The content of each web page was analyzed together with the content of its neighboring pages. The resulting feature scores were also used by the SVM. Using two powerful techniques may radically improve classification; however, this research did not combine the techniques to create a more sophisticated one. They simply used them one after the other on the same data set, which meant that the system took much longer to come up with results.

Further, Chekuri et al. (Chekuri et al. 1997) studied automatic web page classification in order to increase the precision of web search. A statistical classifier, trained on existing web directories, is applied to new web pages and produces an ordered list of categories in which the web page could be placed. At query time the user is asked to specify one or more desired categories so that only the results in those categories are returned, or the search engine returns a list of categories under which the pages would fall.

Furthermore, Chakrabarti et al. (Chakrabarti et al. 1999) proposed an approach called focused crawling, in which only documents relevant to a predefined set of topics are of interest. In this approach, a classifier is used to evaluate the relevance of a web page to the given topics so as to provide evidence for the crawl boundary.

Moreover, an approach proposed by Chen and Dumais (2000) classified search results into a predefined hierarchical structure and presents the categorized view of the results to the user. Their user study demonstrated that the category interface is liked by the users better than the result list interface, and is more effective for users to find the desired information. Compared to the approach suggested by Chekuri et al., this approach is less efficient at query time because it categorizes web pages on-the-fly. However, it does not require the user to specify desired categories; therefore, it is more helpful when the user does not know the query terms well. Similarly, K¨aki (2005) also proposed to present a categorized view of search results to users. Experiments showed that the categorized view is beneficial for the users, especially when the ranking of results is not satisfying.

III. PROPOSED WORK

A critical look at the available literature indicates that in the current scenario, web page classification plays an important role in efficient result retrieval. Most of the work done in the field of web page classification has been carried out using various approaches, such as clustering, Naïve Bayes Model, Bayesian Networks, Decision Trees, Support Vector Machines and so on [30].

However, the approaches discussed above suffer many drawbacks, as listed below:

1) The traditional approach of manual categorization of web pages is a very time consuming and subjective task, and is, thus, open to question.
2) Categorization of web pages based on clustering algorithms requires the number of clusters to be specified in advance by a user.
3) Text-based content categorization is again a subjective task.
4) Web page classification using Naïve Bayes Model also suffers certain drawbacks. For example, the conditional independence assumption used in Naïve Bayes Model is strong, and makes NB-based systems incapable of using two or more pieces of evidence together [30].
5) Further, use of decision trees (DT) for classification suffers with the drawback of a complex process of training the DT classifier and they can get out of hand with the number of nodes created in some cases. According to [19], with six Boolean attributes there would be need for 18,446,744,073,709,551,616 distinct nodes [30].

Moreover, most existing algorithms have used text content of a web page for its classification, while less importance has been given to the information provided by various HTML elements of the web page.

Our work proposes a novel approach to predictively classify a web page into its respective domain using the HTML elements of the corresponding web page by means of usage of an Artificial Neural Network (ANN). Further, various sources of information (like HTML tags) have been tested for web page classification by [17] namely:

- BODY, the content of the BODY tag;
- META, the meta-description of the META tag;
- TITLE, the page's title;
- MT, the union of META and TITLE content;
- BMT, the union of BODY, META and TITLE content.

The experimental results shown in table II that have been obtained by testing the afore-mentioned HTML tags signify the fact that by using a combination of one or more of the afore-mentioned HTML tags, classification performance for web pages may be improved. Here, F1 measure combines precision and recall with equal importance into a single parameter for optimization and is defined as:

$$F1 = 2PR/(P + R) \qquad (1)$$

where P is precision and R is recall [13].

TABLE II
Classification Performance (F1) For Various Representations Of Web Pages

CLASSIFIER	BODY	META	TITLE	MT	BMT
NAIVE BAYES	0.4455	0.5374	0.4015	**0.5587**	0.5086
PERCEPTRON	0.4075	0.4727	0.3707	**0.4996**	0.4691

HTML elements are considered for web page classification. This is so because an HTML document is much more than a simple text file [17]. It is structured and connected with other HTML documents. While a great effort has been made to exploit hyperlinks for classification, the structured nature of web pages is rarely taken into account [17]. Our experiments show that in addition to the content of the web site, using further the HTML elements used in representing the Web sites enhances the performance of the classification.

Various web pages will be tested by our system in order to classify them into their respective domains. Artificial Neural Networks (ANN) will be used by our system for the process of identification of the domains of various web pages. This is so because in case of supervised learning, ANN works by learning from a sample of known examples, and learn from their mistakes.

Generally, Neural Networks are trained initially to perform various complex functions, viz, pattern recognition, identification, classification, prediction, forecasting and the like. They are basically trained so that a particular input leads to a specific output. The network is adjusted, based on a comparison of the output and the target, until the network output matches the target, as shown in Fig. 2 below [33]. Typically, many such input/output pairs are used in this process of supervised learning in order to train a network. In this process, weights or bias parameters are adjusted in order to meet the desired output [33]. The network learns from its mistakes and the known samples of input/output patterns. Thereafter, the trained network is used for testing or predicting the outcomes of unknown samples based on the patterns observed in known samples.

The same process will be deployed by our system in order to predict the category (or domain) to which a web page will belong. The number of neurons in the input layer will vary based on the number of categories as specified by the user, say M. The number of neurons in the hidden layer may also vary, say N, while the number of neurons in the output layer will depend on whether a page belongs to a single domain or multiple domains.

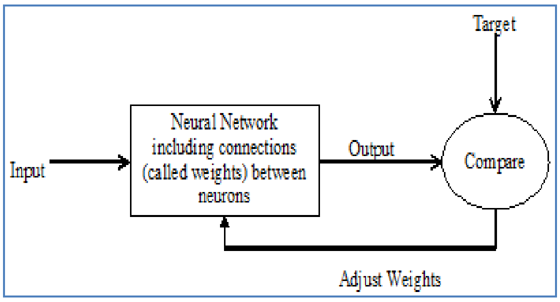
Target
Neural Network including connections (called weights) between neurons
Input
Output
Compare
Adjust Weights

Fig.2. Neural Network Block Diagram

Fig. 3 depicts the architecture of an Artificial Neural Network. The figure shows that the ANN consists of M neurons in the input layer, N neurons in the hidden layer and 2 neurons in the output layer. The neuron in the output layer will define the domain of the web page that needs to be classified. (Herein, value of M is 5, and N is 4.)

IV. CONCLUSION

The World Wide Web is growing widely and within a few years the amount of web content will surely increase tremendously. Hence, there is a great requirement to have algorithms that could classify and list web pages accurately and efficiently. In this paper we have attempted to propose a solution for web page classification using HTML elements of a web page. The proposed model will provide the necessary web page classification technique for fast and efficient working of the search engines. Further, it is also expected to obtain results with high classification accuracy.

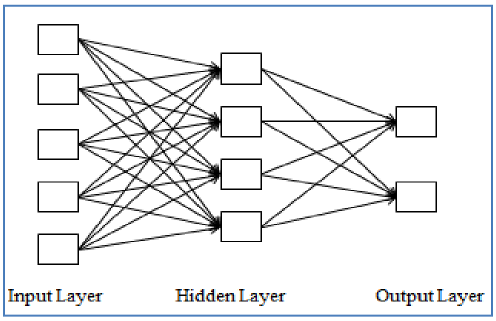

Fig.3. Architecture of the Artificial Neural Netowrk (ANN) [M:N:2]

.

REFERENCES

[1] Aijun An and Xiangji Huang, "Feature selection with rough sets for web page categorization", York University, Toronto, Ontario, Canada.

[2] Arul Prakash Asirvhatam and Kranti Kumar Ravi. "Web Page Categorization based on Document Structure", International Institute of Information Technology, Hyderabad, India 500019.

[3] Chandra Chekuri & Michale, Stamford University; Prabhakar Gold wasser and Eli Uphal, “Web Search Using Automatic Categorization”, IBM alamden Research Center, 650 Harry Road,San Jose CA 95120.

[4] Open Directory Project, http://www.dmoz.org/.

[5] Towell, G. & Shavlik, J. “Extracting Refined Rules from Knowledge-Based Neural Networks”, Machine Learning, Vol. 13, No. 1, 1993, pp. 71-101.

[6] Fletcher, G.P & Hinde, C.J. “Interpretation of Neural Networks as Boolean Transfer Functions”, Knowledge-Based Systems, Vol. 7, No. 3, 1994, 207-214.

[7] Gerry McGovern. "A step to step approach to web page categorization", www.gerrymcgovern.com.

[8] Yahoo!, http://www.yahoo.com/.

[9] Looksmart, http://www.looksmart.com/.

[10] J.Yi and N.Sudershesan, “A classifier for semi structured documents”, In KDD 2000, Boston, MA USA, 2000.

[11] Pierre J. M., Practical issues for automated categorization of Web sites. In Proceedings of ECDL 2000 Workshop on the Semantic Web, Portugal.

[12] Hu, W., Chang, K. & Ritter, G. “WebClass: Web Document Classification Using Modified Decision Trees”, in: 38th Annual Southeast Regional Conference, 2000, pp. 262-263.

[13] On the Automated Classification of web sites, John.M.Pierre, Linkoping Electronic Articles in Computer and Information Science, Vol. 6, 2001.

[14] Meta tags, Frontware International.

[15] Enhong, C., Shangfei, W., Zhenya, Z. & W. Xufa. “Document classification with CC4 neural network”, in: Proceedings of ICONIP, Shanghai, China, 2001.

[16] Liu, Z. & Zhang, Y. “A competitive neural network approach to web-page categorization”, International Journal of Uncertainty, Fuzziness & Knowledge Systems, Vol. 9, 2001, pp. 731-741.

[17] Riboni D. Feature Selection for Web Page Classification. In Proceedings of EURASIA-ICT 2002 Workshop.

[18] Denis, F., Laurent, A., Gilleron, R., Tommasi, M. “Text classification and co-training from positive and unlabelled examples”, in: ICML Workshop: The Continuum from Labeled to Unlabeled Data, 2003, pp. 80-87.

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

[20] Tal, B. “Neural Network - Based System of Leading Indicators”, CIBC World Markets, 2003.

[21] Orallo, J. “Extending Decision Trees for Web Categorisation”, in: 2nd Annual Conference of the ICT for EUIndia Cross Cultural Dissemination, 2005.

[22] Web Page Categorization Using Artificial Neural Networks, S. M. Kamruzzaman Department of Computer Science and Engineering Manarat International University, Dhaka, Bangladesh, Proceedings of the 4th International Conference on Electrical Engineering & 2nd Annual Paper Meet 26-28 January, 2006.

[23] Wang, L., Li, X., Cao, C. & Yuan, S. “Combining decision tree and Naïve Bayes for classification”, Knowledge Based Systems, Vol. 19, 2006, pp. 511-515.

[24] Wang C., Ding C., Meraz R., Holbrook S. “PSoL: a positive sample only learning algorithm for finding non-coding RNA genes”, Bioinformatics, Vol. 22, No. 21, 2006, pp. 2590-2596.

[25] Estruch, V., Ferri, C., Hernández-Orallo, J., & Ramírez- Quintana, M. J. “Web Categorisation Using Distance-Based Decision Trees”, in: International Workshop on Automated Specification and Verification of Web Site, 2006, pp. 35-40.

[26] Web Page Classification: Features and Algorithms, Xiaoguang Qi and Brian D. Davison, Department of Computer Science & Engineering Lehigh University, June 2007.

[27] Addin, O., Sapuan, S. M., Mahdi, E., & Othman, M. “A Naive-Bayes classifier for damage detection in engineering materials”, Materials and Design, 2007, pp. 2379-2386.

[28] Segaran, T. Programming Collective Intelligence, U.S.A: O’Reilly Media Inc, 2007.

[29] Chau, M., & Chen, H. “A machine learning approach to web page filtering using content and structure analysis”, Decision Support Systems, Vol. 44, No. 2, 2007, pp. 482-494.

[30] Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages Daniela XHEMALI1, Christopher J. HINDE2 and Roger G. STONE3 IJCSI International Journal of Computer Science Issues, Vol. 4, No 1, 2009.

[31] A combination approach for Web Page Classification using Page Rank and Feature Selection Technique, Sini Shibu[1], Aishwarya Vishwakarma[2] and Niket Bhargava[3], International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010.

[32] A Comparative Study on Representation of Web Pages in Automatic Text Categorization, Seyda Ertekin[1], C. Lee Giles[1,2], [1]Dept of Computer Science & Engineering, [2]The School of Information & Technology, The Pennsylvania State University, University Park , PA.

[33] Hand Gesture Recognition Using Neural Networks, Klimis Symeonidis, Centre for Vision, Speech and Signal Processing, August 2000.

[34] TheMathsWork,http://www.mathworks.com/products/neuralnet/.

[35] Adar, Eytan, Teevan, Jaime, Dumais, Susan T., and Elsas, Jonathan L. , “The web changes everything, Understanding the dynamics of web content”, In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 282-289, February 2009.