Web Page Categorization using Multilayer Perceptron with Reduced Features

Kavitha S
PSGR Krishnammal College
for Women,
Coimbatore, India

Vijaya M S
GR Govindarajulu School of
Applied Computer Technology
Coimbatore, India

ABSTRACT
The web is a huge repository of knowledge and numerous hyperlinks. Web also serves a broad diversity of user communities and global information service centers. Every day the knowledge in web page upwards rapidly. Web pages can be used to convey the knowledge to web users. Such voluminous size of the web makes an intricacy of web information retrieval, web content filtering and web structure mining. Hence, it is essential for proper categorization of web pages. This paper demonstrates the web page categorization problem as the multi classification task and provides a suitable solution using a supervised learning technique namely multilayer perceptron. The classification model is generated by learning the features that have been extracted from HTML structure and URL of the web page. Feature reduction techniques have been applied to select optimum features and a model is learned. The experimental results of the multilayer perceptron models before and after feature reduction has been evaluated and observed that the multilayer perceptron model with reduced features performs well.

Keywords
Categorization, Multilayer perceptron, Training, Web page.

1. INTRODUCTION AND BACKGROUND
With the rapid growth of World Wide Web, the knowledge in web page grows explosively. Due to its mass, the information overload and information unavailability are the problems in every web search engine. In addition the web pages are not uniformly structured. Hence web page classification is a crucial task in every web search engine.

Web page categorization is an important in many information retrieval tasks like retrieval of scientific papers, e-books and digital library from the web. In web usage mining the web page categorization consumes to build customized web services to individual web users. Web structure mining is concerned with discovering the model underlying the link structure on the web page, for example to predict the links between terrorists in social network analysis. In web page filters like an e-mail filter, content filter, web content filtering determines the content that is to be blocked in a web page. This web page categorization helps to achieve efficient web information retrieval, web content filtering, web structure mining and web usage mining.

Various rules based and machine-learning techniques are currently in use for web page classification. In [1], the different supervised learning techniques namely, decision tree, k-nearest neighbor, one r, multilayer perceptron and rbf kernel are adopted for web page categorization. Web page categorization has been implemented using three feature selection techniques namely, filter model, wrapper model and hybrid model along with the page rank algorithm in order to reduce the redundant features in a web page [2].

In [3], the authors have used various features that are extracted from HTML structure and URL with a compound of HTML and URL along with its information on sibling pages, for web page classification. Naive Bayes algorithm is used as a classifier. The naive Bayes algorithm is compared with semi-supervised algorithm such as co-training and expectation maximization and inductive logic programming have been applied to boost the performance in weak learner for web page categorization in [4].

The research work presented in this paper syndicates the features of web pages stated in [1] [2] [3] [4] and identifies few novel features which can contribute more in accurate categorization of web pages. The features such as strings between slashes and dots in the href attribute of all anchor tags, strings between underscores and minus symbols in the href attribute of all anchor tags, defined in HTML structure of a web page are additionally used. These features have been used to incorporate the referral mechanisms available in web pages like tables, footnotes, bibliographies. They also provide the interconnection between the linked web pages which may contain text, images, video and other multimedia contents.

The proposed model also employs new URL features namely, substring between underscores and minus symbols of the URL, substring between two different symbols of the URL, apart from those used in the existing work. These features have been used to provide additional resources to the URL. Hence these features are very much imperative and assured to contribute more in web page categorization.

In most of the existing work, web page categorization was carried out to classify the web pages of the same domain. Here web pages on different domains such as arts, business, culture, education, entertainment, health and wellness have been considered for categorization.

This paper elucidates the implementation of multilayer perceptron for categorizing the web pages of six different domains. The features are extracted from HTML structure and URLs of a set of web pages in different categories. Feature reduction techniques have been employed to select the optimum features. Feature extraction process and the experiments carried out are described in the rest of this paper.

2. PROPOSED WEB PAGE CATEGORIZATION MODEL
The proposed web page categorization model reduces the convolutions in web mining. The different categories of web pages are collected randomly from the search engines. The

acquired web pages are preprocessed and features are extracted from HTML structure and URL using feature extraction methods. The training data set with instances related to six domains, arts, business, culture, education, entertainment, health and wellness is developed. Feature reduction techniques such as information gain attribute evaluation and symmetrical uncertainty attributes evaluation are applied to identify and select the top quality features. The training set with reduced features is also taken for learning. The proposed web page categorization model employs the multilayer perceptron to learn the classification models. Finally the trained models are evaluated and used for predicting the unknown category of web pages. The proposed web page categorization model is shown in Fig. 1.

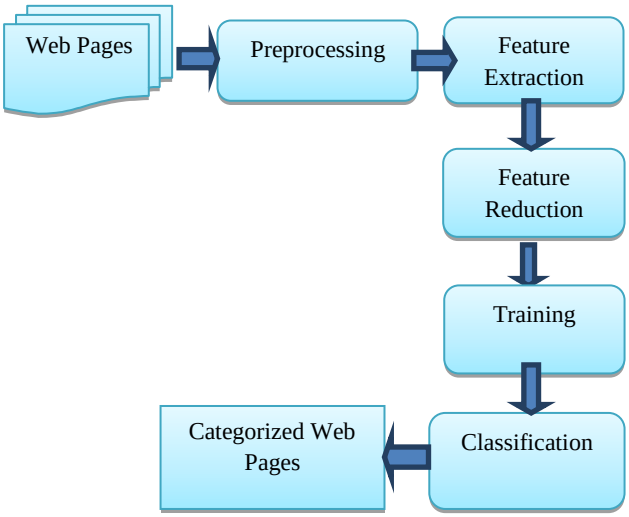

Fig. 1 Proposed web page categorization model

2.1 Preprocessing

Preprocessing can improve computational efficiency and quality of the data. It is used to remove stop words, inadequate html tags and redundant symbols from the HTML structure of web pages.

2.2 Feature Extraction

Feature extraction and selection process plays a vital role in data mining. It can be used to improve the classification effectiveness and computational efficiency. Two kinds of features i.e. HTML features and URL features are extracted. HTML features are obtained from the HTML structure of web pages using term frequency, stemming and structure oriented weighting technique. Term frequency is used for computing the weight of a term in a web page, which is nothing but the number of times the term occurs in a web page. Stemming have been used for morphological scrutiny of words, it reduces the term frequency, which has analogous meaning in the identical web page. The structure oriented weighting technique is used to assign most significant terms that are more appropriate for representing other elements in HTML structure of a web page. Features relevant to the URL are taken from the URL of the corresponding webpage. Categorical values from 1 to 6 are assigned for features with respect to arts, business, culture, education, entertainment, health and wellness respectively.

2.3 HTML Features

The HTML elements defined in HTML structure such as title tag, meta attribute tags, paragraph tag, heading and links describe the content of a web page. So the features pertaining to these elements are extracted to form the training dataset for web page categorization. The HTML features are enumerated below.

2.3.1 Title Tag

The title tag is required in all HTML structure, it defines the title of the document to the web page. This tag plays a vital role in search engine optimization. The syntax of a Title tag is <Title> Title of the web page </Title>. For example this tag <Title>All Indian Sports Websites</Title> returns the value of a feature as 'Sports' and the categorical value 5 is assigned to this feature.

2.3.2 Meta Description Tag

The Meta Description tag is an HTML tag and it describes the contents of a web page. It is used in the head area of the web page. This tag is used after the title tag and before the Meta keyword tag. The Meta description tag provides the next significance to the search engine optimization and it contains the snippet knowledge of the HTML structure. The syntax of this tag is <Meta name="description" content= "description of web page">. For example this tag <Meta name= "description" content="health web page may contain general health, men's health and women's health"> returns 'health' and the categorical value 6 is assigned to this feature.

2.3.3 META Keyword Tag

The Meta Keywords are the list of terms and it can be used to highlight information on the web page. These keywords are separated by comma. The syntax of this tag is <Meta name="keywords" content="list of terms">.This Meta Keyword tag <Meta name="keywords" content="Best sports websites, Best sports teams"> returns 'sports' and the categorical value of this feature is 5.

2.3.4 Paragraph Tag

The intention of Paragraph tag on any web page is to elucidate different concepts related to one topic. Paragraph element is one consideration of an HTML element in which to separate one paragraph with another paragraph in a web page. Paragraph element is simply represented with English alphabet p. The content between the start tag and end tag forms a Paragraph. The format of Paragraph tag is Content of the paragraph . The value of this feature is determined as before.

2.3.5 Heading Tags

Heading tags are indicators and it defines the section headings and subheadings within a web page. This tag can be used to validate document structure and organization. It also represents to generate outlines and tables of contents in a web page. Section headings at different levels namely, <h1> represents the highest level heading, <h2> is the next level down, <h3> for a level below that, and so on to <h6>. H1 tag represents the page title in a web page and H2 tag can be used to provide the all major headings in the web page. These heading tags <h1>Special Education </h1> and <h2>General Education</h2> returns 'Education' and the categorical values of these features are 4.

2.3.6 Base URL in HTML Structure

Base URL specifies the all relative href and other links in a web page. It represents an external resource to the web page. HTML permits only one base element for each web page. The base URL has attributes, but it does not support contents in a web page. This Base URL tag http://dir.yahoo.com/Arts/

Visual_Arts/Mail_Art/ returns 'Arts' and so its categorical value is 1.

2.3.7 Anchor Tag

An anchor element is called an anchor because the web designers can use it to anchor a URL to some text on a web page. When users view the web pages in a system, they can click the text to activate the link and visit the page that the URL has the link. In HTML structure an anchor tag may be either the origin or target of a hyperlink. In the href attribute the anchor becomes a hyperlink to the web page. Alternatively with the name or id attribute set, the element becomes a target. The URL can link to this target via a fragment identifier. In anchor tag the title represents brief information about the link. The syntax of the anchor tag is link text . The anchor tag features are listed below and some of the anchor tag features are shown in Fig. 2.

Fig. 2 Anchor tag Features

2.3.7.1 Strings between slashes

This feature specifies the link between slashes in the href attribute of all anchor tags in HTML source code. For example this href attribute http://dir.yahoo.com /Health/ Dental_Health/ returns 'Health' and the categorical value 6 is assigned to this feature.

2.3.7.2 Strings between dots

This feature stipulates the link between dots in the href attribute of all anchor tags. The value of this feature is determined as before.

2.3.7.3 Strings between underscores

It can be used in the states the link between underscores in the URL of all anchor tags in HTML. This tag http://dir.yahoo. com/Education/Higher_Education returns 'Education' and its categorical value is 4.

2.3.7.4 Strings between minus symbols

It represents the link between minus symbols in the href attribute of all anchor tags in the HTML structure. The value of this feature is determined as before.

2.4 URL Features

Uniform Resource Locator is an exact character string that constitutes a reference to an internet resource and it is a type of uniform resource identifier. In URL the scheme name defines the namespace and domain name gives the destination location to the URL. The port number is optional. The path can be used to find the resource and the query string contains the name and value pairs separated by ampersands of the URL. The fragment identifier specifies the location within a web page. The URL features used in the research work are listed below.

2.4.1 Substring between dots

This feature specifies the substring between dots in the URL. For example, the URL http://rakings.business.com, returns 'business' and so its categorical value is 2.

2.4.2 Substring between slashes

It states the substring between slashes in the URL. For example this URL http://dir.yahoo.com/Arts/ returns 'Arts' and so its categorical value is 1.

2.4.3 Substring between underscores

This feature used to specify the substring between underscores in the URL. For example this URL http://yahoo.com/Health/General_Health returns 'Health' and the categorical value of this feature is 6.

2.4.4 Substring between minus symbols

This feature represents the substring between minus symbols in the URL. This URL http://www.magfact.com/sports/best-sp orts-websites/ returns 'sports' and so its categorical value is 5.

2.5.5 Substring between two different symbols

This feature specifies the substring between two different symbols in the URL. This URL http://dir.yahoo.com/ Society_and_Culture/Cultures_and_Groups returns 'Culture' and the categorical value is 3.

This a list of 16 features is extracted from the HTML source and the URL of a web page. The categorical values for these 16 features are obtained through a PHP code. This list of 16 categorical values can form a feature vector. The feature vectors corresponding to web pages of six different categories are generated and a training dataset is developed.

3. MULTILAYER PERCEPTRON

Neural networks are a different paradigm for computing and it is an inspiration from neuroscience. Neural networks are

particularly effective for predicting events when the networks have a large database. This network imitates the human brain. Artificial neurons or processing elements are highly simplified models of biological neurons. As in biological neurons, artificial neurons have a number of inputs, cell body and output that can be connected to a number of other artificial neurons. This network is densely interconnected together by learning rule which to adjust the strength of the connection between the units in response to externally supplied data.

Fig. 3 illustrates the structure of a node with an activation function. The weighed inputs are combined via a combination function that often consists of a straightforward summation. A relocate function calculates a subsequent value, the invention yielding single output between 0 and 1. Combination function and transfer function makes up the activation of the node. The three common transfer functions used in this network are sigmoid, linear and hyperbolic function.

4. EXPERIMENT AND RESULTS

The web page categorization model is implemented using Multilayer Perceptron in WEKA environment. The WEKA, Open Source, Portable, GUI-based workbench is an assortment of state-of-the-art machine learning algorithms and statistics preprocessing tools. The data set used in the experiment is developed by collecting web pages randomly from search engines. Web pages relevant to six categories namely arts, business, culture, education, entertainment, health and wellness by taking from yahoo, google, etc. For each category 35 web pages have been downloaded and totally 210 web pages are used in experimenting web page categorization.

The features narrating distinguishing characteristics of a web page are extracted and the feature vector of size 16 is generated for all the 210 web pages as described earlier.

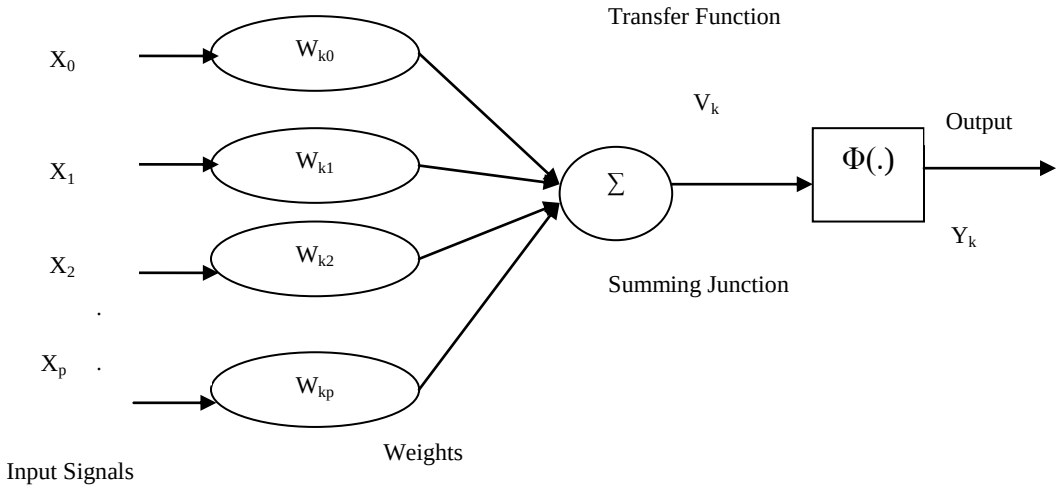

Fig. 3 A typical artificial neuron with an activation function

The multilayer Perceptron network is the most widely used neural network classifier. It is a feed forward artificial neural network model that maps sets of input data into a set of appropriate output. It is a variation of the standard linear perceptron in that it uses three or more layers of nodes with nonlinear activation functions and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable using hyperplane. MLP networks are universal, flexible and nonlinear models consisting of a number of units organized into multiple layers.

In MLP the complexity network can be changed by varying the number of layers and the number of units in each layer. Multilayer Perceptron with back propagation algorithm are the standard algorithm for any supervised learning pattern recognition process and the subject of ongoing research in computational neuroscience and parallel distributed processing. In addition, MLP networks are universal approximators. These tools are valuable in problems when one has little or no knowledge about the form of the relationship between input vectors and their corresponding outputs.

Class labels Arts, Business, Culture, Education, Entertainment and Health are assigned to all the instances pertaining to arts, business, culture, education, entertainment, health and wellness respectively and the training data set is developed. In order to improve the classification accuracy and learning time, the optimum features are selected using feature reduction techniques such as, information gain attribute evaluation and symmetrical uncertainty attribute evaluation. Features namely paragraph tag and h2 tag in HTML structure, substring between minus symbols in URL are ranked low and so removed. Therefore the size of the feature vector is reduced from 16 to 13 and the training data set is developed.

4.1 Classification results before feature reduction

The dataset is trained for multilayer perceptron and the classifier is built. The performance of the learned model has been evaluated using 10-fold cross validation. The categorization accuracy is measured as the ratio of the number of correctly classified instances in the test dataset and the total number of test cases. The statistics of training datasets used in the experiment are given in Table 1.

Table 1. Statistics of training data set

Number of classes	6
Number of features	16
Number of instances in each class	35
Number of instances in training data set	210

The experiment is also carried out using a naïve Bayes algorithm for which the same dataset is employed. The results of web page categorization based on multiplayer perceptron and naïve Bayes classifiers with respect to classification accuracy and learning time are shown in Table 2.

Table 2. Classification performance before feature reduction

Evaluation Criteria	Classifiers	
	MLP	NB
Time taken to build models (Sec)	1.27	0.1
Correctly classified instances	200	192
Incorrectly classified instances	10	18
Classification accuracy (%)	95.2	91.4

4.2 Classification results after feature reduction

The training data set with reduced feature is used here for learning. The statistics of the training dataset after feature reduction used in this experiment is given in Table 3.

Table 3. Statistics of training data set

Number of classes	6
Number of reduced features	3
Number of features after feature reduction	13
Number of instances for each class	35
Number of instances in training data set	210

This dataset is trained using multilayer perceptron and the naïve Bayes algorithm in WEKA environment. The results of this experiment are shown in Table 4.

Table 4. Classification performance after feature reduction

Evaluation Criteria	Classifiers	
	MLP	NB
Time taken to build models (Sec)	1.2	0
Correctly classified instances	203	196
Incorrectly classified instances	7	14
Classification accuracy (%)	96.6	93.3

4.3 Comparative analysis of results

The performance of multilayer perceptron and the naïve Bayes algorithm before and after feature reduction is shown in Fig. 4 and Fig. 5.

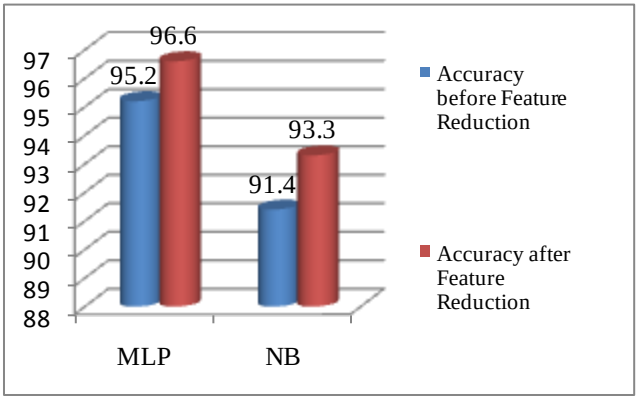

Fig. 4 Classification accuracy

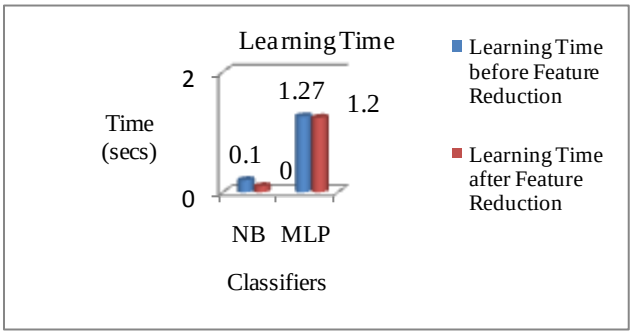

Fig. 5 Learning time of classifiers

From the above comparative analysis it is observed that classification accuracy produced by multilayer perceptron is higher than the naïve Bayes algorithm in both the experiments. The time taken to build the models before and after feature reduction is high in case of multilayer perceptron when compared with naive Bayes. As the web page categorization is adopted in areas like information retrieval, web content filtering and web structure mining built the model can be incorporated into such IR system, so the classification accuracy plays a major role in evaluating the performance of a web page categorization model than learning time. Hence it is concluded that the multilayer perceptron performs well than the naive Bayes algorithm in both the experiments.

5. CONCLUSIONS AND FUTURE WORK

This paper demonstrates the web page categorization problem as the multi classification task. The proposed model is implemented using multilayer perceptron and naive Bayes algorithm. Features are extracted from HTML structure and URL of different categories of web pages. Feature reduction techniques are used to select the optimum features. The experiment and results of the models were evaluated and observed that the multilayer perceptron performs well with reduced features. Web page categorization can be extended to web community mining as a scope for further work.

6. REFERENCES

[1] Alamelu Mangai, J., and Santhosh Kumar, V. 2011. A Novel Approach for Web Page Classification using

Optimum Features, in Proceedings of International Journal of Computer Science and Network Security, Vol.11, No.5.

[2] SiniShibu, Aishwarya Vishwakarma, and Niket Bhargava. 2010. A Combination Approach for Web Page Classification using Page Rank and Feature Selection Technique, in Proceedings of International Journal of Computer Theory and Engineering, Vol.2, No.6.

[3] Sara Meshkizadeh, and Amir Masound Rahmani. 2010. "Web page Classification based on Compound of Using HTML Features and URL Features and Features of Sibling Pages", in Proceedings of International Journal of Advancements in Computing Technology, Vol.2, No.4.

[4] Nuanwan Soonthornphisaj, and Boonserm Kijsirikul. 2005. Combining ILP with Semi-supervised Learning for Web Page Categorization, in Proceedings of International Journal of Information and Mathematical Sciences, Vol.1, No.4.

[5] Santhana Lakshmi, V., and Vijaya, M. S. 2011. The SVM Based Interactive Tool for Predicting Phishing Websites, in Proceedings of the International Journal of Computer Science and Information Security, Vol.9, No.10.

[6] Rekha Jain, and Purohit G. N. 2011. Page Ranking Algorithms for Web Mining, in Proceedings of International Journal of Computer Application, Vol.13.

[7] Ting, S. L., W.H.IP, Albert, H. C. T. 2011. Is Naïve Bayes a Good Classifier for Document Classification", in Proceedings of International Journal of Software Engineering and its Applications", Vol.5, No.3.

[8] Zhihua Wei, Hongyun Zhang_, Zhifei Zhang, Wen Li, Duoqian Miao, 2011. A Naïve Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Results, in Proceedings of International Journal of Advanced Intelligence", Vol.3, No.2, pp. 173-188.

[9] BinduMadhuri, C. H., AnandChandulal, J., Ramya, K., and Phanidra, M. 2011. Analysis of Users Web Navigation Behavior using GRPA with Variable Length Markov Chains, in Proceedings of International Journal of Data Mining and Knowledge Management Process", Vol.1, No.2.

[10] Pooja Sharma, and Pawan Bhadana. 2010. Weighted Page Content Rank for Ordering Web Search Result, in Proceedings of International Journal of Engineering Science and Technology, Vol.2.

[11] Wongkot Sriurai, Phayung Meesad and Choochart Haruechaiyasak. 2010. Hierarchical Web page Classification based on a Topic Model and Neighboring Pages Integration, in Proceedings of International Journal of Computer Science and Information Security, Vol.7, No.2.

[12] Selvakuberan, K., Indradevi M and Rajaram R. 2008. Combined Feature Selection and Classification-A Noval Approach for the Categorization of Web Pages, in Proceedings of International Journal of Information and Computing Science, Vol.3, No.2, Pp.083-089.

[13] Brown E. N., Kass R. E., and Mitra P. P. 2004. Multiple neural spike train data analysis: state-of-the-art and future challenges, Nature Neuroscience, 7 (5): 456–61.

[14] Arabib and Michael A, The Handbook of Brain Theory and Neural Networks.

[15] Russell and Ingrid. 2012. Neural Networks Module.

[16] Yogendra kumar jain, and Sandeep wadekar .2011. Classification based Retrieval Methods to Enhance Information Discovery on the Web, in Proceedings of International Journal of Managing Information Technology, Vol. 3, No. 1.

[17] Shiqun Yin, Yuhui Qiu, Chengwen, Zhong, Jifu Zhou. 2007. Study of Web Information Extraction and Classification Method, IEEE International Conference on Wireless Communications, Networking and Mobile Computing, Wicom, PP.5548-5552.

[18] Lilac A. E., Al-safadi. 2009. Auto Classification for Search Intelligence, in Proceedings of World Academy of Science, Engineering and Technology.