Journal of Theoretical and Applied Information Technology
10th January 2014. Vol. 59 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVEMENTS IN NEURAL NETWORK FOR CLASSIFICATION OF WEB PAGES

[1]J. B. LEELA DEVI, [2]Dr. A. SANKAR

[1]University Research Scholar, Anna University, Tamil Nadu, India

[2.] Associate Professor, PSG College of Technology, Coimbatore, India

E-mail: leeladevi_2008@rediffmail.com

ABSTRACT

Web page classification differs from traditional text classification due to additional information by Hyper Text Markup Language (HTML) structure and the presence of hyperlinks. While effort was taken to exploit hyperlinks for classification, web pages structured nature is rarely considered. A noticeable HTML documents feature is HTML tags and respective attributes that ensure that HTML documents are viewed in browsers and other user agents. This paper proposes a semantic-based feature selection to improve web pages search and retrieval over large document repositories. Web page classification using HTML tags is evaluated using the 4 Universities Dataset. The features are classified using Proposed Neural Network. The experimental results show improved precision and recall with the presented method.
Keywords: *Hyper Text Markup Language (HTML), Web page classification, HTML tags, Neural Network*

1. INTRODUCTION

World Wide Web is the globe's biggest information system. Not only the sheer size of the web but its rapid growth making web analysis difficult. Web page classification, also called web page categorization, is a process of assigning a web page to one or more category labels. Compared to standard text classification, web content classification is different in the following aspects. First, conventional text classification is usually performed on "structured corpora with well-controlled authoring styles" [1], while web collections lack this facility. Second, web pages are semi-structured HTML documents making them user friendly visuals. Though other document collections could include embedded information to make them a semi-structured format, such markup is usually stripped for classification purposes. Finally, web documents exist in a hypertext, linked to and from other documents. While not unique, this feature is central to the web's definition and is not seen in text classification problems. Hence, web classification is both important and distinguished from traditional text classification,

Semantic technologies aim to overcome IR limitation through use of explicit descriptions, internal structure, content and services overall structure [2]. The Semantic Web is an information mesh linked in such a way to enable easy process by machines, globally. This can be thought of an efficient way to represent data on the World Wide Web, or as a globally linked database. All meanings/information conveyed by content in unstructured form (like text or audio-visual content) cannot be fully translated to a clear/formal semantic representation, for practical reasons. However, it is possible to describe parts of conveyed information, albeit to an incomplete extent, as metadata.

Metadata is data about other data (e.g., the ISBN number and the author's name are metadata about a book) [3]. For the same reasons it is useful to keep both information (data and metadata) parts in the system. It is also relevant to have a link connecting both, commonly called annotation. Different syntactic support standards were proposed for metadata and annotation representations. Markup languages like HTML (Hyper Text Markup Language) and XML and their features are effectively used for web page classification. Current studies on semantic query interfaces are in 4 categories, i.e.; keyword-based, form based, view-based and natural language-based systems reviewed in [4].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Current webpage classification techniques use information variety to classify a target page: the text of the page, its hyperlink structure, the link structure and anchor text from pages pointing to target page and its location (provided by its URL). A web page's uniform resource locator (URL) is the least expensive to obtain, of this information, and one of the more informative sources as regards classification [5, 6]. Web pages contain extra information like HTML tags, hyperlinks and anchor text, mainly because web content is written HTML. This study uses these features - located on the page - for classification.

An obvious HTML documents feature but not in plain text documents are HTML tags and their attributes which ensure that HTML documents are viewed in browsers and other user agents. It was proved that using information from tags boosted classifier performance. Golub and Ardo [7] derived significance indicators for different tags textual content. Tag use is advantageous for structural information embedded in HTML files but usually ignored by plain text approaches.

An HTML element is a HTML document's individual component which in turn is made up of a tree of HTML elements and other nodes like text nodes with every element having specified attributes. Elements have content, including other elements and text. HTML represents semantics/ meaning. For example, title element is representative of the document title. In HTML syntax, elements are usually written with a start and end tag the content coming in between. Tags include the element's name and are surrounded by angle brackets. An end tag includes a slash after opening angle bracket to differentiate it from a start tag.

HTML tags provide visual web page representation and are a parameter to highlight content importance. Keyword-based retrieval returns inaccurate/incomplete results when differing keywords describe the same document and queries concept. Concept-based retrieval attempted to overcome this by using manual thesauri with term co-occurrence data, or extracting latent word relationships and concepts from a corpus. Semantic search motivates Semantic Web from inception for classification and retrieval processes. This paper proposes a semantic-based feature selection. Feature extraction is by stemming, stop words, and locating IDF. Word importance based on html tag and ontology mapping are used by the proposed feature extraction to assign features extra weights. Features thus obtained are classified using proposed Neural Network.

2. LITERATURE REVIEW

Trillo et al [8] proposed semantics techniques to group results from a traditional search engine into categories defined through different meanings of input keywords. It discovers keywords' possible meanings to create categories dynamically at run-time by considering web available heterogeneous sources. This method considers knowledge provided by ontologies available on the Web to dynamically define categories. Hence, it is independent of the sources providing results to be grouped. Experimental results prove the proposed approach's effectiveness, specially when users search for information on the Web.

Eissen et al [9] presented results from a user study on Web genre usefulness and also results from a genre classifier construction using discriminant analysis, neural network learning, and support vector machines. A classifier's underlying feature set is emphasised: Aside from the standard feature types, new features based on word frequency classes and which can be computed with minimum computational effort are introduced. Compact feature sets with few elements are constructed to achieve a satisfactory genre diversification. About 70% of Web-documents are assigned their true genre.

Riboni [10] analysed HTML structure and hyperlinks based web page classification peculiarities trying to use them to represent web pages and thereby improve categorization accuracy. Experiments on a corpus of 8000 documents of 10 Yahoo! categories, using Kernel Perceptron and Naive Bayes classifiers revealed dimensionality reduction's usefulness and of a new, structure-oriented weighting technique. A new method representing linked pages with local information ensured hypertext categorization feasibility for real-time applications. Combining usual web pages representation (in form of local words) with a hypertextual one improves classification.

A recent approach to semantic web search combines standard Web search with ontological background knowledge with regular Web search engines as inference motor for Semantic Web search. Amato et al [11] proposed enhancing this to Semantic Web search by using inductive reasoning techniques ensuring abilities to handle inconsistencies, noise, and incompleteness,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

likely in web's distributed and heterogeneous environments. A prototype was reported and the new approach's implementation with extensive experimental results was discussed.

Selamat et al [12] proposed a news web page classification method (WPCM), using a neural network with inputs from both principal components and class profile-based features. Each news web page is represented by a term-weighting scheme. As the unique words number in the collection set is big, principal component analysis (PCA) selects most relevant features for classification. Then the final PCA output is combined with feature vectors from a class-profile having the most regular words in each class. Most regular words existing in each class were manually selected and weighted using an entropy weighting scheme. Regular words fixed number from each class is used as feature vectors with reduced principal components from PCA. Feature vectors are input to neural networks for classification. Experimental evaluation proved that the WPCM method provided acceptable classification accuracy with sports news datasets.

Anagnostopoulos, et al [13] suggested a system to identify and categorise web pages, based on information filtering. The system is a three layer Probabilistic Neural Network (PNN) having biases and radial basis neurons in the middle layer and competitive neurons in the output layer. This is a ecommerce area study domain. Thus, PNN hopes to identify e-commerce web pages to classify them to respective type based on a framework describing commercial transactions fundamental transactions on the web. The system was tested with many web page types demonstrating the method's robustness as no restrictions were imposed except for language content being in English. The probabilistic classifier estimated specific e-commerce web pages population. Potential applications include surveying web activity in commercial servers and web page classification in largely expanding information areas like e-government or news media.

3. METHODOLOGY

The 4 Universities Dataset

The 4 Universities Dataset includes WWW-pages from major university computer science departments collected in January 1997 by CMU text learning group's World Wide Knowledge Base project [14]. The dataset has a total of 8,282 manually classified pages. The classes included Student, Faculty, Staff, Department, Course, Project and Others. The class 'other' includes pages not considered ``main page" representing an instance of earlier six classes. Data set included pages from Cornell, Texas, Washington, Wisconsin and 4,120 mixed pages from other universities. Each class is assigned a directory with each having 5 subdirectories, one for each of the 4 universities and 1 for mixed pages. The directories contain Web-pages.

FEARURE EXTRACTION

Features extracted from documents used stemming, stop words and finding Inverse Document Frequency (IDF) [15]. Both document and query are represented as vectors in a high dimensional space corresponding to keywords in a vector space model. Similarity measures calculated similarity values between keywords and document with ranking being based on similarity values. The first step is document set keywords identification followed by a list of unrelated/irrelevant words – called a stop list – to avoid indexing them; words like the, a, of, for, with and so are stop words.

In a document set d and a set of terms t, each document is modeled as a vector v in t dimensional space R^t, called a vector space model. Let frequency be denoted by $freq(d,t)$, as it expresses the number of occurrences of term t in document d. The term-frequency matrix $TF(d,t)$ measures term t association regarding the given document d. $TF(d,t)$ has nil value if the document does not contain the term, and a computed number otherwise. The number can be set as $TF(d,t)$ = 1 when term t occurs in document d or uses relative term frequency where frequency versus total occurrences of all document terms. Frequency is generally normalized by:

$$TF(d,t) = \begin{cases} 0 & freq(d,t) = 0 \\ 1+\log\left(1+\log\left(freq(d,t)\right)\right) & otherwise \end{cases}$$

Inverse Document Frequency (IDF), represents scaling factor. If term t occurs frequently in many documents, its IDF value is less as term has lower discriminative power [16]. The $IDF(t)$ is defined as follows:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$IDF(t) = \log \frac{1+|d|}{d_t}$$

d_t is the set of documents containing term t .Similar documents have similar relative term frequencies. Similarity is measured among document sets or between a document and query. Cosine measure locates documents similarity [14]; the cosine measure is got by

$$sim(v_1, v_2) = \frac{v_1 . v_2}{|v_1| \ |v_2|}$$

where v_1 and v_2 are two document vectors, $v_1 . v_2$ defined as $\sum_{i=1}^{t} v_{1i} v_{2i}$ and $|v_1| = \sqrt{v_1 . v_1}$.

PROPOSED NEURAL NETWORK

Multiple layers of computational units make up neural networks, interconnected with each other based on network design. Inputs are fed on input layer and propagated through layers for the output. The latter is computed using weights, bias and activation functions. Propagation rule trains the network through back propagating errors and changing nodes weights. The difference between obtained output and desired output is the error.

Neural networks are an important classification tool. Recent research in neural classification proved that neural networks are a good alternative to various conventional classification methods. Their advantage is in the following theoretical aspects. First, neural networks are data driven self-adaptive methods as they can adjust to data without any specification of functional or distributional form of the underlying model. Second, they are functional approximators as they can approximate any function with arbitrary accuracy. As classification procedures seek functional relationships between group memberships and object attributes, accurate identification of underlying function is important. Third, neural networks are nonlinear, making them flexible in modelling real world complex relationships. Finally, neural networks can estimate posterior probabilities that provide the base to establish classification rule and perform statistical analysis.

A difficulty in using artificial neural networks to solve large-scale real-world problems is in dividing a problem into smaller and simpler sub problems; and also in how to assign a network module to learn each sub-problem; and how to recombine them into solutions to the original problem. In the last few years, researchers studied modular neural network learning approaches to deal with this problem [17-19]. A modular neural network means a neural network made up of several sub-networks, arranged hierarchically.

The proposed neural network is made up of two sub-networks. Each network is made up of two hidden layers, with differing transfer function. In this study, the transfer function used are sigmoid and tan h. The advantage of different functions is that the mutual interference is minimized during simultaneous processing and execution of complex task. Table 1 gives the parameters of the proposed neural network.

Table 1: Parameters for the proposed MLP NN

Parameter	Values
Input Neuron	50
Output Neuron	4
Number of Hidden Layer	2
Number of processing elements upper	4
Number of processing elements lower	4
Transfer function of hidden layer upper	Tanh
Transfer function of hidden layer lower	Sigmoid
Learning Rule of hidden layer	Momentum
Step size	0.1
Momentum	0.7
Transfer function of output layer	Tanh
Learning Rule of output layer	Momentum
Step size	0.1
Momentum	0.7

Journal of Theoretical and Applied Information Technology
10th January 2014. Vol. 59 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. RESULTS AND DISCUSSION

The proposed semantic based feature selection for web page classification using HTML tags is evaluated using the 4 Universities Dataset and compared with IDF feature extraction method. Four classes are classified (Student, Course, Faculty and Project). The main computing technique is to recall cases where query words occur. Recall and precision are measured for both proposed semantic and keyword techniques allowing absolute and relative performance measures to be calculated using standard measures. The accuracy, precision, recall and f measure are computed as follows:

Accuracy (%) = (TN + TP) / (TN + FN + FP + TP)

$$precision = \frac{TP}{TP + FN}$$

$$recall = \frac{TP}{TP + FP}$$

$$f\ Measure = \frac{2 * recall * precision}{recall + precision}$$

where TN (True Negative) = Number of correct predictions that an instance is invalid

FP (False Positive) = Number of incorrect predictions that an instance is valid

FN (False Negative) = Number of incorrect predictions that an instance is invalid

TP (True Positive) = Number of correct predictions that an instance is valid

SVM with linear kernel, MLP-NN and the proposed Neural Network classify keywords and semantic based features. Experimental results are detailed in the following tables and figures. Table 2 and Figure 1 detail classification accuracy and root mean squared error obtained for IDF and proposed feature extraction.

Table 2: Classification Accuracy and Root Mean Squared Error

Method Used	Classification Accuracy %	RMSE
SVM-linear-IDF	82	0.3
MLP NN with IDF	84	0.255
Proposed NN with IDF	88	0.217
SVM-linear-Proposed feature extraction	87	0.26
MLP NN with proposed feature extraction	89	0.1987
Proposed NN with proposed feature extraction	93	0.1426

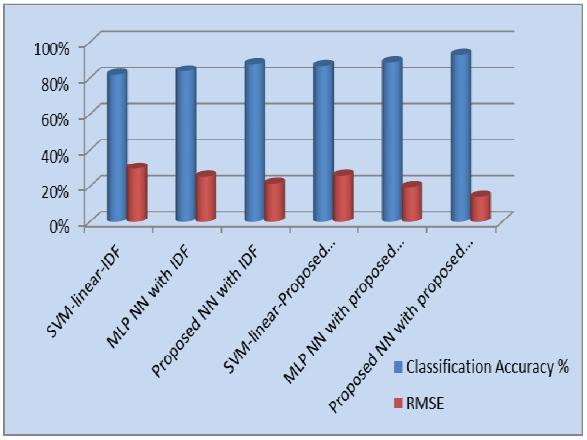

Figure 1: Classification Accuracy and Root Mean Squared Error

Figure 1 shows that the proposed feature extraction performs better than the IDF. The precision, recall for the different methods is shown in Table 3 and Figure 2.

Table 3: Precision, Recall and F Measure

Method Used	Precision	Recall
SVM-linear-IDF	0.826	0.82
MLP NN with IDF	0.905384	0.779384
Proposed NN with IDF	0.910163	0.843169
SVM-linear-Proposed feature extraction	0.887	0.87
MLP NN with proposed feature extraction	0.911797	0.861775
Proposed NN with proposed feature extraction	0.957428	0.904071

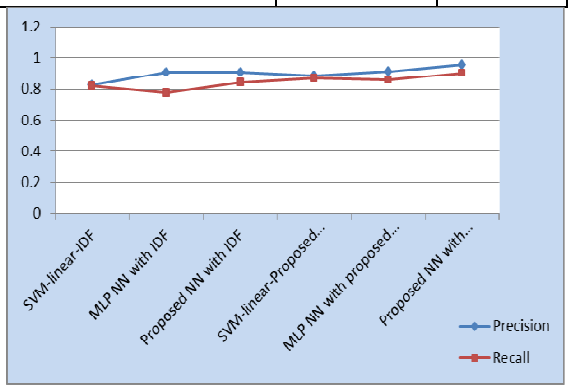

Figure 2: Precision and Recall

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

It is observed from Figure 2 that the precision and recall for the proposed proposed neural network with proposed feature extraction has high precision and recall.

5. CONCLUSION

This paper proposes a model to exploit semantic-based feature selection to improve search and retrieval of web pages in large document repositories. The features are classified using the proposed Neural Network. Keyword-based retrieval returns inaccurate/incomplete results when differing keywords describe same document and queries concept. HTML tags provide visual web page representation and are considered a parameter to highlight content importance. An obvious HTML documents feature is HTML tags and their attributes that create HTML documents to enable being viewed in browsers and other user agents. The proposed semantic based feature selection for web page classification using HTML tags is evaluated using the 4 Universities Dataset and compared with IDF feature extraction method. Experiments show improved precision and recall with the proposed method.

REFRENCES:

[1]. Chekuri, C., M. Goldwasser, P. Raghavan, and E. Upfal (1997, April). Web search using automated classification. In Proceedings of the Sixth International World Wide Web Conference, Santa Clara, CA. Poster POS725.

[2]. M. Fernández, V. López, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells. Semantic Search meets the Web. 2nd IEEE International Conference on Semantic Computing (ICSC 2008). Santa Clara, CA, USA, August 2008.

[3]. V. López, M. Fernández, E. Motta, M. Sabou, V. Uren. Question Answering on the Real Semantic Web. Poster and demo at the 6th International Semantic Web Conference (ISWC 2007). Busan, Korea, November 2007.

[4]. Victoria Uren, Yuangui Lei, Vanessa Lopez, Haiming Liu, Enrico Motta, and Marina Giordanino. The usability of semantic search tools: A review. Knowl. Eng. Rev., 22(4):361–377, 2007.

[5]. M.-Y. Kan. Web page classification without the web page. In Proc. of WWW '04, 2004. Poster paper

[6]. L. K. Shih and D. Karger. Using URLs and table layout for web classification tasks. In Proc. of WWW '04, 2004.

[7]. Golub, K. and A. Ardo (2005, September). Importance of HTML structural elements and metadata in automated subject classification. In Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Volume 3652 of LNCS, Berlin, pp. 368–378. Springer.

[8]. Trillo. R, Po. L, Bergamaschi. S and Mena E., (2011) "Using Semantic Techniques to Access Web Data@ Information Systems, Special Issue 36(2): 117-133.

[9]. Meyer zu Eissen, S., & Stein, B. (2004). Genre classification of web pages. KI 2004: Advances in Artificial Intelligence, 256-269.

[10]. Riboni, D. (2002). Feature selection for web page classification. In EURASIA-ICT 2002 Proceedings of the Workshop (pp. 473-477).

[11]. d'Amato, C., Fanizzi, N., Fazzinga, B., Gottlob, G., & Lukasiewicz, T. (2010). Combining Semantic Web search with the power of inductive reasoning. Scalable Uncertainty Management, 137-150.

[12]. Selamat, A., & Omatu, S. (2004). Web page feature selection and classification using neural networks. Information Sciences, 158, 69-88.

[13]. Anagnostopoulos, I., Anagnostopoulos, C., Loumos, V., & Kayafas, E. (2004, June). Classifying Web pages employing a probabilistic neural network. In Software, IEE Proceedings- (Vol. 151, No. 3, pp. 139-150). IET.

[14]. Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2), 103-134.

[15]. Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation, 60(5), 503-520.

[16]. Papineni, K. (2001, June). Why inverse document frequency?. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1-8). Association for Computational Linguistics.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[17]. J. M. J. Murre, Learning and Categorization in Modular Neural Networks. London, U.K.: Harvester Wheatsheaf, 1992

[18]. Lu, B. L., & Ito, M. (1999). Task decomposition and module combination based on class relations: A modular neural network for pattern classification. Neural Networks, IEEE Transactions on, 10(5), 1244-1256.

[19]. Happel, B. L., & Murre, J. M. (1994). Design and evolution of modular neural network architectures. Neural networks, 7(6), 985-1004.