International Journal of Computer and Communication Technology

Volume 1 | Issue 4 Article 4

October 2010

Child Internet Protection System

deepshikha verma
Fairleigh Dickinson University, deepshikha@gmail.com

Mayank Pathak
Department of Computer Science & Engineering Technocrats Institute of Technology Bhopal, India, pathak.mayank@rediffmail.com

Rajesh Nigam
Department of Computer Science & Engineering Technocrats Institute of Technology Bhopal, India, rajesh_rewa@hotmail.com

Bhupendra Verma
Department of Computer Science & Engineering Technocrats Institute of Technology Bhopal, India, bk_verma3@rediffmail.com

Child Internet Protection System

Deepshikha Patel [1], Mayank Pathak[2], Rajesh Nigam[3], Bhupendra Verma[4].
Department of Computer Science & Engineering
Technocrats Institute of Technology
Bhopal, India
23.deepshikha@gmail.com [1], pathak.mayank@rediffmail.com[2], rajesh_rewa@hotmail.com[3], bk_verma3@rediffmail.com[4]

***Abstract*—Nowadays internets is becoming an amazing resource and provide hours of fun for kids. Clearly there are many benefits that result from Internet usage, but there is a side to the internet that can be worrying for any parent. Internet is explored, in particular: child sexual exploitation; children's exposure to sexually explicit or offensive material, so there is need of a system which prevents child abuse from internet. Currently many systems are developed to prevent Internet-related child abuse. In this paper, we have developed a web page classification model, which will protect children from harmful and offensive material available on internet. Our approach uses entropy term weighting scheme, Principal Component Analysis for feature reduction and Back propagation neural network as classifier. Our experiment result shows that our approach performed well as comparison to other approaches.**

Keywords— Feature Selection, Back propagation Neural Network, Principal Component Analysis, Web Page Classification

I. INTRODUCTION

As the internet has evolved, it has become entertainment, communication source that almost every person use as a matter of routine. Internet is now the biggest source to get any type of information, but this is particularly problematic when children are able to access the material with ease. This is becoming serious social issue and parents are much concerned about the access to objectionable and offensive content by children. The need to protect children from objectionable material has led to the development of techniques to facilitate filtering of web content. Now, we can see many parents installing web content filtering software in PC to block illegal and harmful contents. Currently there are various information blocking or filtering application designed for filtering of such contents. Some of them are CyberPatrol, CyberSnoop, NetNanny, and SafeSurf [2]. Controlling access to the objectionable web content typically employs different approaches including self or third party rating, URL blocking, keyword matching and filtering, and intelligent analysis. URL blocking and intelligent content analysis are effective than the others. URL blocking is the simplest one and must be customized for individual uses or highly standardized for specific groups of users. A blacklist/white list database of URLs is manually maintained for the judgment, but for many web sites that might contain multiple URLs and dynamic IP addresses, his method becomes ineffective. The manual effort is required to collect URL in blacklist. This is very time consuming and difficult to maintain without an automatic learning mechanism. Currently content based techniques are used for web filtering.

Although there are many applications of content based filtering according to the types of media and the methods of analyzing contents, we focus on classification of objectionable documents. In this paper we have used a content based approach for filtering of illegal and objectionable documents to protect children from the negative side of internet .

II. THE PROPOSED SYSTEM

A. Basic

Many filtering approaches have been developed for classification of objectionable documents. False positive rate degrade the performance of filtering. To solve this problem, we have considered the length factor of document for its weighting phase, because normally objectionable documents are short in length. We have used the entropy term weighting scheme for feature selection, PCA for feature reduction and neural network as a classifier.

B. Design Concept

Automatic Text categorization assigns the incoming documents into predetermined categories. A text Categorization system consist of preprocessing phase, document representation phase, and classification phase. A text categorization system is shown in figure 1.

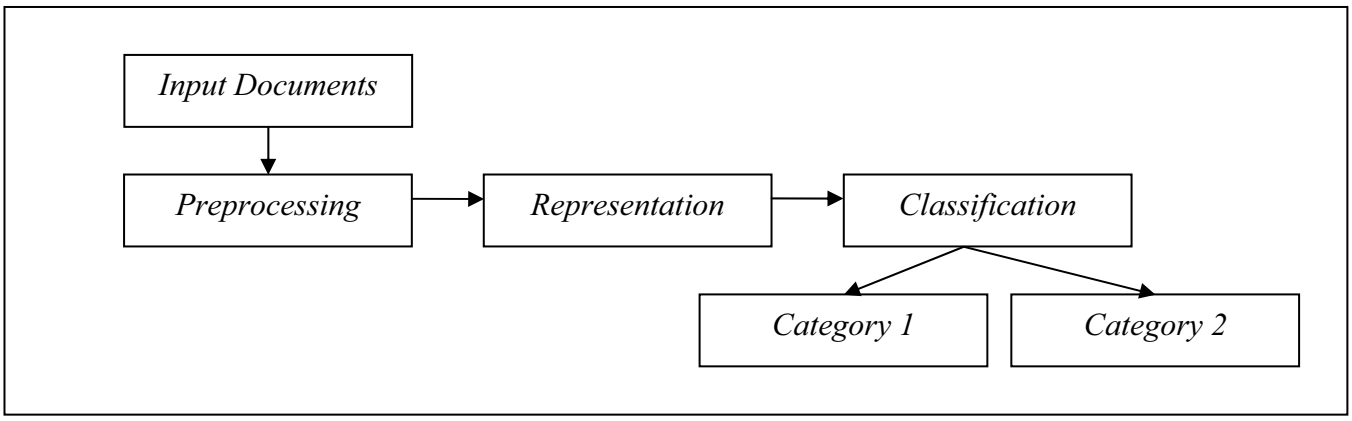

Figure 1: Automatic Text Categorization System

1) Preprocessing

All web documents are HTML pages and contain HTML tags, so there is a need for HTML parsing of web documents so that only texts are extracted excluding those HTML code. Afterward stopping and stemming are performed on HTML parsed documents. Stopping is a process of removing most frequent and common words that exist in a web document by using a stop words dictionary. However stemming reduces the Occurrence of term frequency, which has similar meaning in the same document [1].

2) Representation

To facilitate the process of matching keywords and documents, some preprocessing steps are taken first:

Documents are *tokenized*; that is, all punctuation marks are removed and the character strings without spaces are considered as tokens (words, also called *terms*).

All characters in the documents and in the query are converted to upper or lowercase.

Words are reduced to their canonical form (stem, base, or root). For example, variant forms such as *is* and *are* are replaced with *be*, various endings are removed, or the words are transformed into their root form, such as *programs* and *programming* into *program*. This process, called *stemming*, uses morphological information to allow matching different variants of words.

Articles, prepositions, and other common words that appear frequently in text documents but do not bring any meaning or help distinguish documents are called *stop words*. Examples are *a*, *an*, *the*, *on*, *in*, and *at*. These words are usually removed.

The collection of words that are left in the document after all those steps is different from the original document and may be considered as a formal representation of the document. To emphasize this difference, we call the words in this collection terms. The collection of words (terms) in the entire set of documents is called the text corpus.

3) Classification

In clustering we use the document class labels for evaluation purposes only. In classification they are, however, an essential part of the input to the learning system. The objective of the system is to create a mapping (also called a *model* or *hypothesis*) between a set of documents and a set of class labels. This mapping is then used to determine automatically the class of new (unlabeled) documents. In a narrow sense the latter process is called *classification*, while the general framework for classification includes the model creation phase and other steps. Therefore, the general framework is usually called *supervised learning* (also, *learning from examples*, *concept learning*)

III. CHILD INTERNET PROTECTION SYSTEM (CIPS)

Child Internet Protection System is a system which helps parents to safe internet access for their children. In this paper, we proposed the web page classification model depicted in figure 2. We collected some documents for training and testing from different web sites. First documents of training set are preprocessed, after preprocessing; documents are represented in a term document matrix. And, then, the features are selected using entropy term weighting scheme. After this we apply PCA to reduce the dimensionality of feature space. Finally the output from the PCA is feed to the back propagation neural network. The steps of CIPS are described as follows:

A. Web Documents Collection

In this step, we gathered various web documents from different web sites. Web robots visit internet sites and gather web pages for classification. Those retrieved web pages are stored in the local database for further processing.

B. Preprocessing

Web page retrieval is a process that retrieves collections of web documents to database from online internet with the help of web crawler. Those retrieved web pages will be stored in local database for further process. Stop-list is a dictionary that contains the most common and frequent words such as 'I', 'You', 'and' and etc. Stopping is a process, which filters those common words that exist in web document by using stop-list. Stemming plays an important role in reducing the occurrence of term frequency that has similar meaning in the same document. It is a process of extracting each word from a web document by reducing it to a possible root word. For example, 'beauty' and 'beautiful' have the similar meanings. As a result, the stemming algorithm will stem it to its root word 'beauty'

C. Representation of Web Pages

After training pages are selected, we apply Porter's stemming algorithm to transfer each word in a web page into its stem. We then remove all the stop words according to a standard stop words list. For each category, we count the number of occurrences of each remaining word stem in all the pages that belong to the category. The word stems in each category are then sorted according to the number of occurrences. We use word stem counts to represent a web document. A web document may contain a huge number of words and not all the words in the global space appear in every document. If we use all the words in the global space to represent the documents, the dimensionality of the data set is prohibitively high for the learning system. In addition, even though our learning system can handle thousands of features, many of the features are irrelevant to the learning task. The presence of irrelevant features in the training data introduces noise and extra learning time. Therefore, it is necessary to reduce the dimensionality of the feature set by removing words with low frequencies.

D. Feature Selection

Feature selection is the process to find the characteristic features, which can help classifying well. The number of features identified by feature extraction may be extremely large, generally high dimensionality of the term space can made the classifier run slowly. Generally, the features are selected by the term goodness criterion, such as DF (Document Frequency), CHI (CHI Statistics), and MI (Mutual Information), IG (Information Gain etc) .In this paper we are using Modified Entropy Term Weighting Scheme and PCA (Principal Component Analysis) as feature selection methods.

1) Modified Entropy Term Weighting Scheme

The Text categorization problems normally involve an extremely high dimensional feature space. A standard procedure to reduce features dimensionality is feature selection. We will use modified entropy scheme [6,10] to extract features from text documents. The term with highest term weights would select as best features. Later, these selected features would form as feature vector. The modified term weighting scheme is as follows.

$$G_i = \left(\frac{\log_{10} DF_i}{\log_{10} n} + 1 \right) \tag{1}$$

$$L_{ij} = \begin{cases} K_{ij} (TF_{ij} > 0) \\ 0 \quad (TF_{ij} = 0) \end{cases} \tag{2}$$

$$K_{ij} = \left(\frac{\log_{10}(TF_{ij})}{\log_{10}(lenDoc_j)} + 1 \right) * \left(\frac{\log_{10}(TF_{ij})}{\log_{10} T(i)} + 1 \right),$$

$$T(i) = \sum_{j=1}^{n} TF_{ij} \tag{3}$$

$$W_{ij} = L_{ij} \times G_i \tag{4}$$

Where,

DF_i =no. of documents that contain *i*th term in a collection.

$lenDoc_j$ =no. of total words that exists in jth document. This modified term weighting scheme contains three main factors including term frequency, collection frequency and document length factor. More detailed discussion about modified entropy term weighting scheme can be found in the work of Lee et al. [3].

2) Feature Reduction Using PCA

Using PCA, the dimension reduction process will reduce the original data vector into small number of relevant features [4, 5].

Let M to be the matrix of document terms weights as follows.

$$M = \begin{pmatrix} a_{11} a_{12} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} a_{n2} & \dots & a_{nm} \end{pmatrix}$$

Where,

a_{ij} =All terms in the collection of documents.

n= Number of terms.

m= Number of documents.

Then we calculate the mean a and subtract it from each data points $a - \bar{a}$. After variance-covariance matrix M can be calculated, where the new value of a_{ij} = ($a_j - \bar{a}$) ($a_i - \bar{a}$).Then we determine eigenvalues and eigenvectors of the matrix M which *C* is a real symmetric matrix so a positive real number λ and a nonzero vector α can be found such that, $C\alpha = \lambda\alpha$ where λ is called an eigenvalue and α is an eigenvector of *C* .In order to find a nonzero vector α the characteristic equation $| C - \lambda I |$ must be solved. If *C* is an $n \times n$ matrix of full rank, *n* eigenvalues can be found such that ($\lambda 1, \lambda 2, .., \lambda n$). By using $(C - \lambda I)\ \alpha = 0$, all corresponding eigenvectors can be found. The eigenvalues and corresponding eigenvectors will be sorted so that $\lambda 1 \geq \lambda 2 \geq ... \geq \lambda n$. Then we select the first $d = n$ eigenvectors where *d* is the desired value.

E. Input Data to the Neural Networks

After the pre-processing of web pages, a vocabulary that contains all the unique words in the news database has been created. Each of the word in the vocabulary represents one feature vector. Each feature vector contains the document-terms weight. The high dimensionality of feature vectors to be as an input to the neural networks is not practical due to poor scalability and performance. Therefore, the PCA has been used to reduce the original feature vectors into a small number of principal components. The neural network parameters for the NN-PCA and the CIPS model are shown in Table 1. We have

selected 100 features from the PCA method and feed them to the neural network for training.

TABLE I. BACK PROPAGATION NEURAL NETWORK PARAMETERS

NN Parameters	Values
Learning rate	0.005
Momentum rate	0.5
Number of iteration	200
Means square error (MSE)	0.05

The parameters of the error-back propagation neural networks are shown in Table I. The numbers of inputs to the neural networks are 100 with 20 hidden layers and 2 output layers.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2 \times \mathrm{Re}\,call \times \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision} \quad (7)$$

Where, FP (False Positive) refers to the number of legitimate, non-objectionable web pages incorrectly classified as illicit web pages. FN (False Negative), TN (True Negative), TP (True Positive) is defined accordingly (see TABLE II).

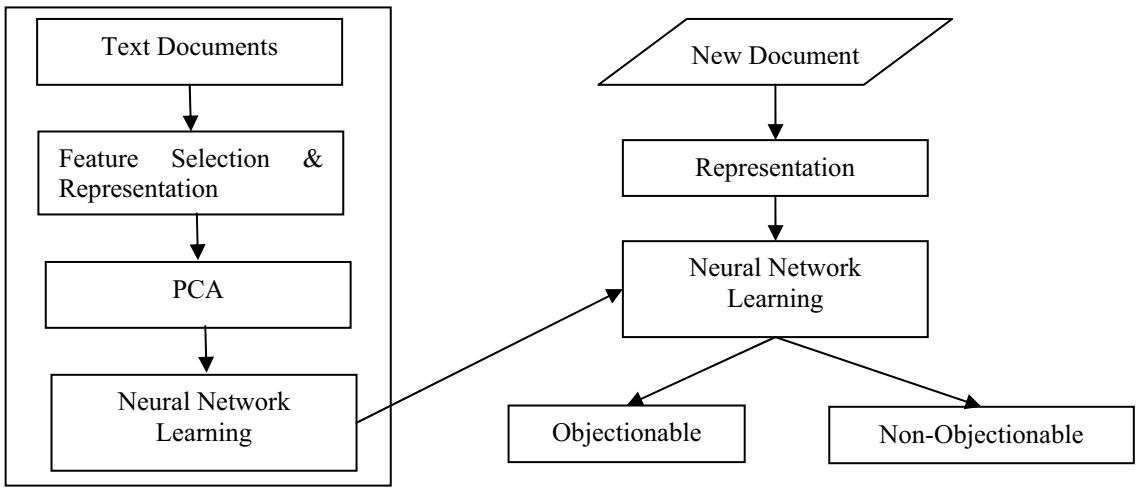

Figure 2: Child Internet Protection System

IV. EXPERIMENTS

A. Data Set

The first task is to build a suitable test corpus of web pages. We collected 200 non-objectionable pages and 50 objectionable web pages from different web sites, from which we have randomly, select 150 web pages for training set and 50 web pages for test set. This study will classify the web pages into two categories, which are objectionable and non-objectionable. Objectionable web pages include adult content, pornographic material etc. On the other hand non-objectionable web pages include informative web pages related to news, business, education etc.

B. Classification Performance

Classification performance is usually measured in terms of the classic Information Retrieval notions of Precision, Recall and F1 measure. [7, 8, 9]. These can be expressed as:

TABLE II. EXPLANATION OF PARAMETERS TP, TN, FP AND FN

Category Set		Expert Judgment	
		Objectionable	*Non-objectionable*
Classifier Judgement	Objectionable	TP	FP
	Non-objectionable	FN	TN

TABLE III. CLASSIFICATION RESULTS USING CHILD INTERNET PROTECTION SYSTEM

Category	Precision	Recall	F1
Objectionable	90.00	94.73	92.30
Non-objectionable	90.00	81.81	85.70
Average	90.00	88.27	89.00

V. RESULTS

We presented new promising results on dimensionality reduction of objectionable web page classification using NN-PCA. In our method, much pre-processing work needs to be done in the selection and calculation of feature vectors of the web documents to be classified. The classification results are shown in Table III. The average of precision, recall, and F1 measures using our approach are 90.00 %, 88.27 %, and 89.00% respectively. We believe that if the feature vectors are selected carefully, the improvement of web page classification using NN-PCA will increase the classification accuracy.

VI. CONCLUSION

The explosive growth of information in internet makes us hard to distinguish between useful and harmful web content. The high similarity within the web pages such as pornography, gynecology and sexology web pages, become a challenging task to content analysis approach in order to classify them correctly. Currently no perfect web classification because each classification design is highly dependable on the content of web sites. In this paper we propose a new model, which uses Principal Component Analysis (PCA) and Modified Entropy Term Weighting Scheme as feature reduction for web pages classification to protect children from offensive material available on the web. We expect the combination of PCA and Modified Entropy Term Weighting Scheme provide better results to control difficult conditions to differentiate between two categories. In addition, by using the combination of PCA and Modified Entropy Term Weighting Scheme as its feature reduction with Neural Network as classifier, we also expect that this model will yield better classification result than the others.

REFERENCES

[1] R. A. Calvo, M. Partridge, and M. A. Jabri, 1998. A Comparative Study of Principal Component Analysis Techniques, presented at In Proc. Ninth Australian Conf. on Neural Networks, Brisbane.

[2] ”Internet Filter Review”, available at http://www.internet-filter-review.toptenreviews.com/, May 2007.

[3] Z. S. Lee, M. A. Maarof,A. Selamat, S.M. Shamsuddin,“Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classofication,”IEEE 8th Int. Conference on Intelligent System Design and Applications, 2008, pp.145-150.

[4] A. Selamat, 2003. Studies on Mobile Agents for Query Retrieval and Web Page Categorization Using Neural Networks, in Division of Computer and Systems Sciences, Gradute School of Engineering, vol. Doctoral. Osaka: Osaka Prefecture University , pp. 94.

[5] R. A. Calvo, M. Partridge, and M. A. Jabri, 1998. A Comparative Study of Principal Component Analysis Techniques, presented at In Proc. Ninth Australian Conf. on Neural Networks, Brisbane.

[6] A. Selamat and S. Omatu ,”Feature Selection and Categorization of Web Pages Using Neural Networks”, Int. Journal of Information Sciences, Elsevier Science Inc. Vol. 158, January 2004.

[7] S. Ali and O. Sigeru, 2004. Web page feature selection and classification using neural networks, Inf. Sci. Inf. Comput. Sci., vol. 158, pp. 69-88.

[8] S. Livingstone and M. Bober. Final report of key project findings. Technical report, UK Children Go Online, 2005.

[9] J. Pierre, 2000. Practical Issues for Automated Categorization of Web Sites.

[10] Z. S. Lee, M. A. Maarof,A. Selamat, S.M. Shamsuddin,“Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classofication,”IEEE 8th Int. Conference on Intelligent System Design and Applications, 2008, pp.145-150.