Jurnal Teknologi, 50(D) Jun 2009: 73–91

TEXT CONTENT ANALYSIS FOR ILLICIT WEB PAGES BY USING NEURAL NETWORKS

LEE ZHI SAM[1], MOHD AIZAINI MAAROF[2], ALI SELAMAT[3] & SITI MARIYAM SHAMSUDDIN[4]

Abstract. Illicit web contents such as pornography, violence, and gambling have greatly polluted the mind of web users especially children and teenagers. Due to the ineffectiveness of some popular web filtering techniques like Uniform Resource Locator (URL) blocking and Platform for Internet Content Selection (PICS) checking against today's dynamic web contents, content based analysis techniques with effective model are highly desired. In this paper, we have proposed a textual content analysis model using entropy term weighting scheme to classify pornography and sex education web pages. We have examined the entropy scheme with two other common term weighting schemes that are TFIDF and Glasgow. Those techniques have been tested with artificial neural network using small class dataset. In this study, we found that our proposed model has achieved better performance in terms accuracy, convergence speed, and stability compared to the other techniques.

Keywords: Artificial neural network; term weighting scheme; textual content analysis; web pages classification

Abstrak. Kandungan laman web haram seperti pornografi, keganasan dan perjudian telah dengan meluasnya mencemarkan pemikiran pengguna internet terutamanya golongan muda seperti kanak-kanak dan muda-mudi. Oleh kerana kurang berkesannya beberapa teknik penapisan saringan laman sesawang yang popular seperti penyekatan *Uniform Resource Locator* (URL) dan penyemakan *Platform for Internet Content Selection* (PICS) terhadap kandungan sesawang yang dinamik pada masa kini, maka teknik penapisan yang berasaskan analisis kandungan sesawang secara berkesan amat diperlukan. Demi mengatasi masalah ini, kami telah mencadangkan suatu model penganalisis kandungan web berasaskan teks dengan menggunakan skema *entropy term weighting* untuk mengelaskan laman pornografi dan laman pendidikan seks dalam penulisan ini. Kajian terhadap keberkesanan skema *entropy* dijalankan dengan membandingkan skema *entropy* dengan dua skema pemberat perkataan yang umum, iaitu TFIDF dan Glasgow. Teknik-teknik ini telah diuji dengan rangkaian neural menggunakan dataset berkelas kecil. Dalam kajian ini, kami mendapati model yang dicadangkan telah mencapai prestasi yang lebih baik dari segi kejituan, kecepatan penumpuan dan kestabilan.

Kata kunci: Rangkaian neural buatan; skema pemberat perkataan; penganalisis kandungan berasaskan teks; pengelasan saringan laman sesawang

[1,2,3&4] Faculty of Computer Science and Information Systems (FSKSM), Universiti Teknologi Malaysia (UTM), Skudai, 81310, Johor, Malaysia
Tel: +607-55 32222 (ext.), Fax: +607-5532210. Email: [1]samleecomp@gmail.com, [2]aselamat@utm.my, [3]aizaini@utm.my, [4]mariyam@utm.my

1.0 INTRODUCTION

The impressive growth of the Internet has made a new evolution in human life. Twenty years ago, the term Internet was practically anonymous to most of the people [2]. Today, it has become a very powerful tool for human throughout the world and has even become a part of human life [3]. The Internet is constructed of a huge amount of information which consists of almost any subjects of our life such as environment, law, health, etc [4]. Nowadays, many e-services provided through the Internet are much more effective and cost saving than the traditional services. With such a collection of various resources and services, we almost can do anything at our finger tips through the Internet [5].

Despite the benefits of the Internet as information superhighway, it is also the most dangerous place [6]. Web users always need to pay the risk for theft of information, spamming, and virus threat as well as mental pollution because of harmful resources. The objectionable web contents such as pornography, violence, and gambling have polluted the mind of immature web users. Pornography perhaps is the biggest threat to our children's and teenagers' healthy mental life. There are thousands of pornography sites on the Internet that can easily be found and detected. In order to analyze the impact of pornography web pages to the public at large, Jerry Ropelato[1] has conducted a web survey. Table 1 and Figure 1 show the statistics for top adult search requests during year 2006 based on the web survey. Figure 2 illustrates the average proportion for different age groups of people that involve in adult search request using search engine. Based on Figure 1 and 2, we found that 23% of Figure 1 result was contributed by teenager and children age below 18. This has certainly become a detrimental factor to let children and teenagers access the Internet without proper guiding.

The issue of children and teenagers browsing adult web pages without acknowledgement from parents makes web filtering and monitoring system highly required in the family and education environments. In fact, there are various commercial web filtering products available in the market. Some popular products are CyberPatrol,

Table 1 Top ten requested terms from adult search requests during year 2006 [1]

Index	Request Term	Total Request
1	Sex	75,608,612
2	Adult Dating	30,288,325
3	Adult DVD	13,684,718
4	Porn	23,629,211
5	Sex Toys	15,955,566
6	Teen Sex	13,982,729
7	Free Sex	13,484,769
8	Adult Sex	13,362,995
9	Sex Ads	13,230,137
10	Free Porn	12,964,651

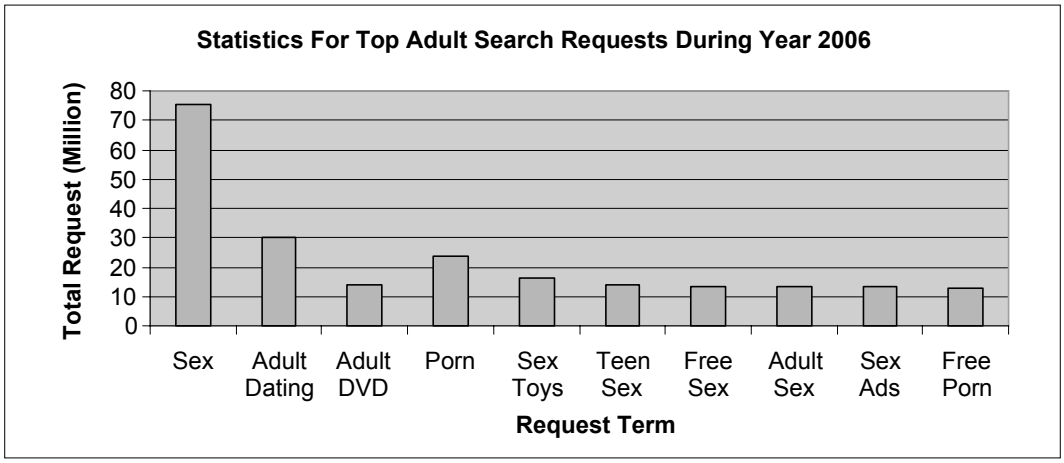


Figure 1 Statistics for top adult search requests during year 2006 by Jerry Ropelato [1]

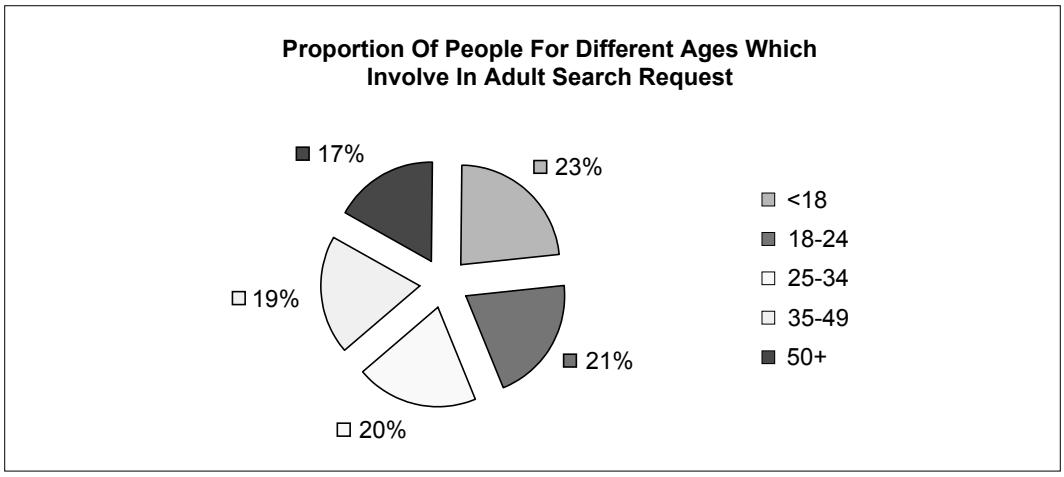


Figure 2 Statistics based on the proportion of people from different age groups who involve in adult search request [1]

NetNanny, Websense, and CyberSitter [1]. The simplest and most popular solution to filter the harmful resources is simply to block the Uniform Resource Locator (URL) or Internet Protocol (IP) address of those particular links. Some other techniques are using the Platform for Internet Content Selection (PICS) checking or keyword matching. The PICS checking is a technique that identifies the PICS content labels (also called metadata) for web pages and further blocks the web page if it contains harmful resources. Keyword matching is a technique that blocks a web page if the number of certain keywords in the web page reaches a pre-set threshold. According to the in depth evaluation of these products (e.g. the one performed in the European Project

NetProtect [7]), their filtering effectiveness is limited by the use of the above techniques especially the URL and IP address blocking.

The above techniques are fast and require only short processing time, however they always fail to block those unknown web pages that are not in their list. The URL or IP blocking technique has even become ineffective against current dynamic web contents [8]. The reason is that the URL blocking only blocks particular web pages that match those black listed URL and IP address stored in the database [9]. Thus, using this limited technology, it is difficult to obtain the complete list the URLs of the whole World Wide Web (WWW) since there are millions of web pages being added daily [10]. On the other hand, trust issue has always been the argument for PICS checking technique since the web publishers have the right to label the content of the metadata [9]. Hence, PICS has been only suggested as the supplementary filtering technique due to its limitation against the ever-changing web contents. The keyword matching technique is designed to overcome the dynamic content issues; however it is not efficient to handle web pages with different subjects but having similar terminologies [12]. For instance, this technique blocks both pornography and sex education web pages although intentionally the users need to block only pornography web pages. Thus, a heuristic content analysis technique with effective model is highly desired in order to effectively perform pornography web page classification.

We have used a textual content analysis technique to solve the issues of dynamic web content and similarity terminologies with different subjects [11]. The advantages of this technique are it is able to analyze the textual content on the body of web pages and provide effective classification result against unknown web pages [8]. This technique has been adapted in our pornography web page classification with textual content analysis model. In order to represent the textual content of web pages to be understood by computer machine, efficient term weighting scheme is crucial. This paper has proposed a textual content analysis model using entropy term weighting scheme to classify pornography and sex education web pages. We examined the entropy scheme with Term Frequent Inverse Document Frequency (TFIDF) and Glasgow term weighting schemes. The output of the schemes is the input for artificial neural network in order to test the effectiveness of each scheme.

2.0 TERM WEIGHTING SCHEME

From the textual content analysis point of view, natural language is very redundant in the sense that many different words share a common or similar meaning. As for computer machine, it is hard to understand the meaning of natural language without proper ways. Term weighting scheme is a statistical measure used to evaluate how important a word is to a document in a collection [13]. In other words, sets of numeric number can be obtained through this term weighting scheme. These sets of number can be understood by computer machine and used for further analysis and document

classification. Normally, there are three main factors of term weighting that are term frequency factor, collection frequency factor, and length normalization factor [14]. These three factors are multiplied together to produce the resulting term weight. We have compared the effectiveness of entropy weighting scheme with the two other weighting schemes often used by search engines to score and rank relevancy of a document given a user query. Those weighting schemes are Term Frequent Inverse Document Frequent (TFIDF) and Glasgow weighting schemes.

2.1 Term Frequent Inverse Document Frequent

TFIDF is one of the most common methods used in information retrieval (IR) field to represent and describe documents in the Vector Space Model [15]. The data is represented as the document-term frequency matrix ($Doc_j \times TF_{jk}$). The TFIDF function weighs each vector component of each document by following several steps. Each vector component is relating to a term or a word (w_k) of vocabulary. The first step is calculating the term frequency in the document. Term frequent illustrates how frequent a word appears in a document. The higher term frequency for a term means, it is estimated that the more significant of the particular term in that document.

On the other hand, Inverse Document Frequency (IDF) measures how infrequent a term is in the collection. The value is estimated using the whole collection of documents. The TFIDF is based on a belief that if a word is infrequent in the text collection, it is estimated to be very relevant to the document. In contrast, if a word is very frequent in the text collection, it is not considered to be representative in the text collection.

TFIDF is commonly implemented in IR to compare a query vector with a document vector using a similarity such as the cosine similarity function. However, there are still many variants of TFIDF. The following common variant was used in our comparison experiments, as found in Salton [16]

$$x_{jk} = TF_{jk} \times idf_k \tag{1}$$

where the variables are explained in Table 2.

Table 2 Explanation of index for calculation based on term frequency inverse document frequency

Index	Explanation
j	Variable, $j = 1, 2,..., n$
k	Variable, $k = 1, 2,..., m$
x_{jk}	Terms weight with term j in document k
Doc_j	Each web page document that exists in local database
TF_{jk}	Number of how many times the distinct word (term) w_n occurs in document Doc_j
df_k	Total number of documents in the database that contains the word (term) w_k
idf_k	Equal to log (n/df_k) where n is the total number of documents in database

2.2 Glasgow Term Weighting Scheme

Glasgow term weighting scheme (also called Glasgow model) is one of the interesting schemes in which its main advantage is that too long documents and queries can be penalized [17]. However, Glasgow would only be effective if these two conditions are fulfilled. First, the value for Glasgow must use normalized frequencies. Second, when defining the length of documents and queries as number of terms, stop words must be excluded. Glasgow term weighting scheme can be expressed as follows (as found in M. Sanderson [18])

$$x_{jk} = \frac{\log\left(TF_{jk} + 1\right)}{\log\left(len_k\right)} \times \log\left(\frac{N}{Z_j}\right) \tag{2}$$

where the variables are explained in Table 3.

Table 3 Explanation of index for calculation based on Glasgow Term Weighting Scheme

Index	Explanation
x_{jk}	Terms weight with term j in document k
TF_{jk}	Number of how many times the distinct word term j occurs in document k
len_k	Number of unique term in document k
N	Total number of documents in the database
z_j	Number of documents in which term j occur

3.0 PORNOGRAPHY WEB PAGE CLASSIFICATION WITH ENTROPY TERM WEIGHTING SCHEME MODEL

The main purpose of Pornography Web Page Classification with Entropy Term Weighting Scheme (PWCE) model is to classify the objectionable web pages such as pornography and the healthy web pages such as sex education. PWCE model is mainly based on textual content analysis which employs entropy method as its term weighting scheme. This model is constructed of several parts that are web page retrieval, preprocessing (consists of HTML parsing, text stemming and stopping), class profile based feature (CPBF), and artificial neural network (ANN) classifier. Figure 3 illustrates the overview of PWCE model.

Firstly, during the web retrieval part, a web robot crawls the web pages from the Internet. Those web pages will go through the stemming and stop world filtering process to reduce noise features in the preprocessing part. The relationship within features and web pages will be built and calculated using entropy term weighting scheme in CPBF. Meanwhile, each category of features will be saved in different class profile as reference input for ANN classifier. The features will be trained and the web

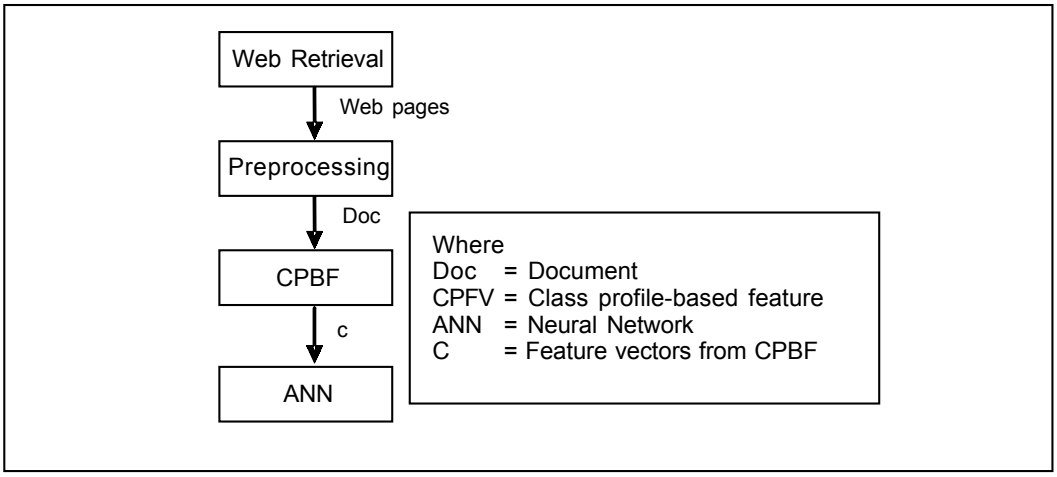


Figure 3 Overview of PWCE model

pages will be classified at ANN classifier. Finally the classification results are categorized into two groups that are pornography and non-pornography (sex education).

3.1 Web Page Retrieval

Web page retrieval is a part that uses web robot to retrieve the desired web page to database. The operation process of this web page retrieval is illustrated in Figure 4. However, web page retrieval part could also be done manually using web browser. This part can be done using either a web crawler or a web browser as long as the web pages are retrieved as required and are stored in the database. Shortly, the main task of the web page retrieval is to obtain the web resources from WWW and duplicate it for the database that will store them for further analysis purpose.

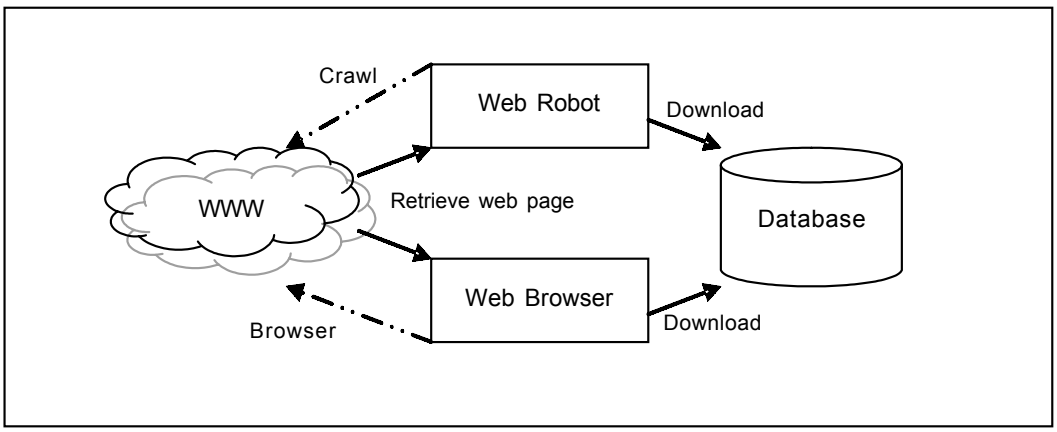


Figure 4 The process of web page retrieval

3.2 Preprocessing

Web pages in the database will go through HTML parsing in order to transform the web pages to become text documents. The documents will perform stopping and stemming before they are passed to the CPBF section. Stop-list is a dictionary that contains the most common and frequent words such as 'I', 'You', 'and', etc. Stopping is a process that filters those common words that exist in web document by using stop-list. Stemming plays an important role to reduce the occurrence of term frequency which has similar meaning in the same document. It is the process of extracting each word from a web document by reducing it to its possible root word. For example, 'beauty' and 'beautiful' have similar meaning. As a result, the stemming algorithm will stem it to its root word 'beauty'. The workflow of pre-processing is shown in Figure 5. The output of this pre-processing will become the input for class profile based features (CPBF) for further process.

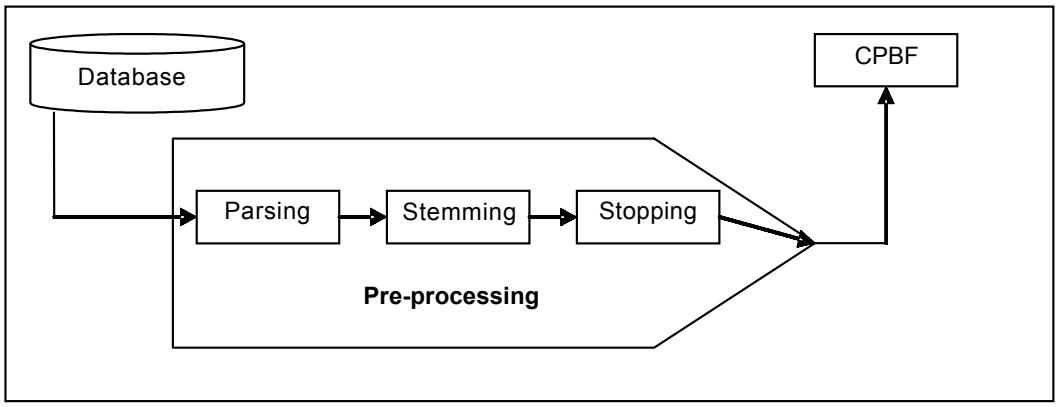


Figure 5 The workflow of pre-processing in PWCE model

3.3 Class Profile Based Features

Class profile-based feature (CPBF) process is a process that identifies those most regular words in each class or category as well as calculates their weights by implementing term weighting scheme. In CPBF, we identifies those most regular words and weighs them using the entropy term weighting scheme before feeding them as input for ANN classifier.

Entropy method is based on a probabilistic analysis of the texts. The main advantage of entropy term weighting scheme is it provides more accurate weights especially when it is compared to TFIDF. It is due to its concern of two aspects that are local term weighting and global term weighting. This means that once every term receives a weight, it will be composed into local and global weights. It is calculated on the range of [0,1], hence the value is normalized. The entropy term weighting scheme

which has been implemented in our experiment can be expressed as follows (Lee. *et al*. [19] and Selamat *et al.* [20]):

$$G_j = \frac{1 + \sum_{j=1}^{n} \frac{TF_{jk}}{F_j} \log\left(\frac{TF_{jk}}{F_j} + 1\right)}{\log n} \tag{3}$$

$$L_{jk} = \begin{cases} 1 + \log TF_{jk} & \left(TF_{jk} > 0\right) \\ 0 & \left(TF_{jk} = 0\right) \end{cases} \tag{4}$$

$$x_{jk} = L_{jk} \times G_k \tag{5}$$

where the variables are explained in Table 4.

Table 4 Explanation of index for calculation based on Entropy Term Weighting Scheme

Index	Explanation
x_{jk}	Terms weight with term j in document k
TF_{jk}	Number of how many times the distinct word term j occurs in document k
F_j	It is a frequency of the term j in the entire document collection
L_{jk}	Local term weighting with term j in document k
G_j	Global term weighting with term j in documents of a collection
n	It is the number of documents in a collection

3.4 Neural Network Classifier

In general, ANN is an interconnected group of artificial neurons (also called nodes), each computing a nonlinear function of a weighted sum of its inputs. A typical ANN is constructed of three main layers that are input, hidden, and output layers. The number of neurons used in the input layer is very depending on the type and amount of input data. Normally, the number of neurons placed in the output layer represents the number of categories that the network could classify. The more nodes in output layer, the more categories into which the data could be classified by the network. The hidden layers are named "hidden" because their output is only available within the network but not available as the global network output. The number of nodes in the hidden layer determines the ability of the network to learn complex relationship. For a simple network, the hidden layer may not exist. The learning behavior of ANNs is very based on algorithms such as back propagation and KSOM [21]. Basically, the learning algorithms are concerned to set the weight connections by training the net with a given data set until it achieves a certain goal. It is always a very challenging task to design a proper network that would solve a complex issue with simpler architecture design.

The main objective to use artificial neural network (ANN) is its learning and generalization characteristics. Learning is the ability to approximate the underlying behavior adaptively from given training data, while generalization is the ability to predict efficiently beyond the trained data. Those characteristics are essential for analyzing complex relationship within data sets which may not be easily perceived by human. The advantage of ANN is its learning behavior in which it could learn the trait of "train" data while establishing their input-output relationship. However, ANN only learns strictly based on the "train" data. In other words, the trait that does not exist in the "train" data is impossible to be learned by ANN. Hence, the input features for ANN training should be selected carefully and efficiently so that the representatives of the data set are as complete as possible.

We uses CPBF as the features selection before the selected features are feed to ANN. It is to ensure only the most representative features are selected as the input for the network. In this model, the artificial feed forward-back propagation neural network (ANN) is adopted as the classifier. For classifying a test document, its term features are feed to input node. Later, term weights are loaded into the input units. The activation of these units will be propagated forward through the network, and finally the value of the output unit determines the categorization decision(s). The backpropagation neural network (BP-ANN) is used to handle a misclassification that occurs so that the error is "backpropagated" in order to change the parameters of the network and minimize or eliminate the error.

The architecture of neural network used for this experiment is shown in Figure 6. The number of input layers (L) is equal to the number of term features after CPBF which is 30. The number of hidden layers (M) is half of the input layers which is 15. The number of output layers (N) is one since the network only classifies the web pages to two classes. We interpret the notation as follows: iteration number as i, momentum rate as α, learning rate as η, bias on hidden node as θ_M, bias on output as θ_N, weight between input layer (L) and hidden layer (M) as W_{LM}, weight between hidden layer (M) and output layer (N) as W_{MN}, generalized error during hidden layer (M) as δ_M, generalized error between hidden layer (M) and output layer (N) as δ_{MN}. The adaptation of the weights between input layer (L) and hidden layer (M) is as below:

$$W_{LM}\,(i+1)\;\; W_{LM}\,(i) + \Delta W_{LM}\,(i+1), \tag{6}$$

where

$$\Delta W_{LM}\,(i+1) = \eta \delta_M A_L + \alpha \Delta W_{LM}(i), \tag{7}$$

$$\delta_M = A_M\,(1 - A_M\,)\sum_N \delta_N W_{MN}\,, \tag{8}$$

Note that the transfer function at input layer (L), A_L is given by

$$A_L = \tan\, sig\,(L), \tag{9}$$

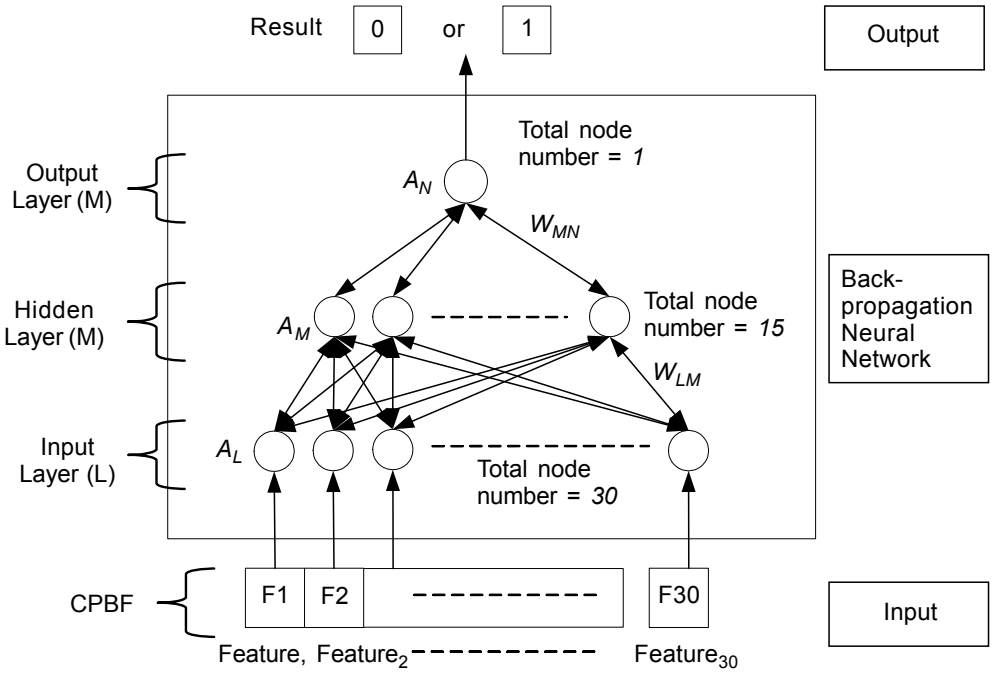


Figure 6 The architecture design of back-propagation neural network in PWCE model

where

$$\tan sig(L) = \frac{2}{\left(1 + e^{-2L}\right) - 1}, \tag{10}$$

and the transfer function at hidden layer (M), A_M is given by

$$A_M = \tan sig(net_M) = \frac{2}{\left(1 + e^{-2net_M}\right) - 1}, \tag{11}$$

where

$$net_M = \sum_M W_{LM} A_L + \theta_M \text{ .,} \tag{12}$$

The adaptation of the weights between hidden layer (M) and output layer (N) is as below:

$$W_{MN}(i + 1) = W_{MN}(i) + \Delta W_{MN}(i + 1), \tag{13}$$

where

$$\Delta W_{MN}(i + 1) = \eta \delta_N A_M + \alpha \Delta W_{MN}(i), \text{ and} \tag{14}$$

$$\delta_N = A_N (1 - A_N)(i_N - A_N)., \tag{15}$$

Finally, the output function at the output layer (N), A_N is given by

$$A_N = \tan sig\left(net_N\right) = \frac{2}{\left(1+e^{-2net_N}\right)-1} \tag{16}$$

$$net_N = \sum_N W_{MN} A_M + \theta_N \tag{17}$$

The detail implementation of BP-ANN for this experiment is explained in Section 4.3. Moreover, the parameters setting for error back-propagation neural network are also indicated in Table 9.

3.5 Result Measurement

The classification result will be examined by the following measurement

$$Accurate = \left(\frac{TotalCorrect}{TotalDocument}\right) \times 100\% \tag{18}$$

where the variables are explained in Table 5

Table 5 Explanation of index for result examination

Index	Explanation
Accurate	The accuracy rate of the classification result
TotalCorrect	The total number of documents fall in the correct category
TotalDocument	The total number of documents used for examination

The classification result will be examined with equation (18) in order to evaluate the performance of each term weighting scheme. The higher the value of *Accurate*, the better it is.

4.0 EXPERIMENTS AND RESULTS

4.1 Data Sets

We collected 700 web pages from the Internet in which the web pages were reviewed to make standard classification. In order to simplify the experiment, this study only classified the web pages to two categories that are pornography and sex education (non-pornography). The pornography web pages are referring to those adult web pages which display sexual activities. On the other hand, the non-pornography web pages in this experiment are referring to those web pages which display useful and informative contents. The sex education web pages here include the subjects related to sex such

as medical sex, sex physiology consultation, health news, and educative information. Table 6 summarizes the data. Due to high similarity between pornography and sex education web pages, this experiment used the mix of medical sex and sex education web pages as the non-pornography category. The purpose to do so is to prove that this model is able to perform extensive classification with textual content analysis.

Table 6 The ratio of pornography and non-pornography web pages used as dataset for experiment purpose

Category	Web Pages	Ratio
Pornography	400	57.14
Non pornography (sex education)	300	42.86
Total	700	100

4.2 CPBF as Feature Selection

For the feature selection using the class profile-based approach, we identified the most regular terms that exist in pornography and sex education categories. We weighted the terms using TFIDF, Glasgow, and Entropy term weighting schemes respectively. We selected thirty terms that have the highest value from each term weighting scheme as the input vector for ANN classifier independently as illustrated in Figure 7. The number vector for *t*, *g*, and *e* were fixed as 30. Hence, ANN classifier would act as the base line to examine the three kinds of term weighting scheme. Table 7 indicates the sample of the first ten selected term features after the CPBF used the TFIDF, Glasgow, and Entropy term weighting schemes.

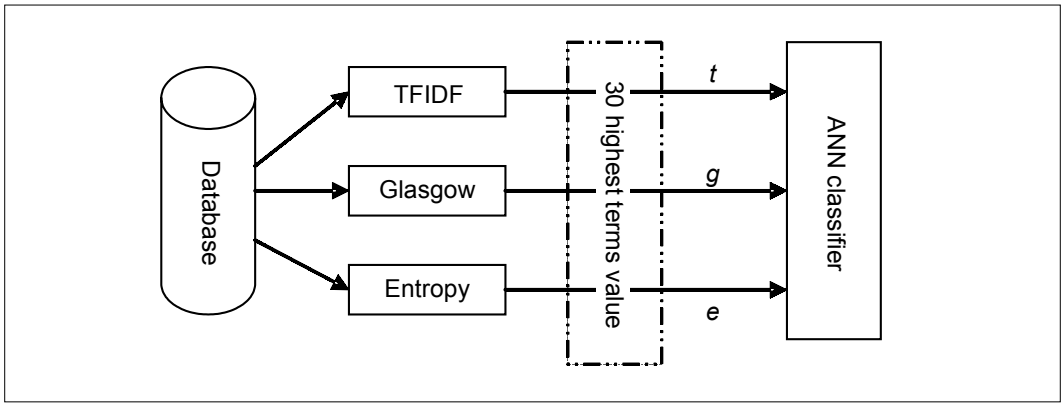


Figure 7 CPBF used TFIDF, Glasgow, and Entropy term weighting schemes as the feature selection methods for ANN classifier

Table 7 Sample of the first ten selected term features for TFIDF, Glasgow, and Entropy weighting schemes

	TFIDF	Glasgow	Entropy
1	Teen	Net	Send
2	Gallery	Bookmark	Print
3	Movie	Girl	Girl
4	Sex	Teen	Legal
5	Picture	Gallery	Image
6	Girl	Medical	Medicine
7	Free	Legal	Breast
8	Baby	Medicine	Picture
9	Education	Sex	Free
10	Young	Penis	18+

4.3 Parameters for ANN Classifier

In order to do the classification, we implemented the back-propagation neural network as our classifier. We have used a set of documents as shown in Table 8 and specification of network summarized in Table 9. We chose 200 documents as training set which consists of 100 pornography and 100 sex education web pages. On the other hand, we selected 500 documents as testing set which consists of 300 pornography and 200 non-pornography web pages.

We designed the architecture of ANN to consist of three layers that are input layer, hidden layer, and output layer. Due to the 30 features (sample of the features are shown in Table 7) from CPBF as input, we designed 30 input nodes at input layer

Table 8 Training and testing set for PWCE with neural network

	Training Documents	Testing Documents
Pornography	100	300
Non Pornography	100	200
Total	200	500

Table 9 Parameters for error back-propagation neural network

Parameters	Value
Learning rate	0.05
Maximum number of interaction	20,000
MSE	0.001
Input layer	30 nodes
Hidden layer	15 nodes
Output layer	1 nodes

where each node feeded one feature. The hidden layer consisted of 15 hidden nodes while the output layer contained one node. The output layer would return two values in which each value represented a category. The return value 0 represented a non-pornography document while the value 1 represented a pornography content. Figure 6 represents the architecture design of BP-ANN in our PWCE model.

During training, the connection weights of the neural network were initialized with some random values. The connection weights were adjusted according to the error back-propagation learning rule. This process was continued until the maximum number of interactions was achieved or the mean squares error (MSE) reached a predefined level.

4.4 Classification Result

The three terms weighting schemes, TFIDF, Glasgow, and Entropy were used to represent term-document in a data collection for CPBF as feature selection purpose. The best performance obtained by these term weighting schemes are reported and compared in Table 11. The classification results using each term weighting scheme are shown in Table 10. The average accuracy rate using TFIDF, Glasgow, and Entropy term weighting schemes are 86.9%, 89.7%, and 91.1% respectively. We compared the accuracy which is calculated based on the formula (6). We examined that the proportions of the documents were classified correctly to their particular categories in a collection of documents. The more documents were classified correctly to their particular categories, the higher the value returned by formula (6) which indicated that it is more accurate.

The fundamental characteristics of intelligent are learning and predicting abilities. As previously mentioned, ANN would only learn the pattern that exists in the training data. Different term weighting schemes would generate their unique pattern of data

Table 10 Accuracy rate for TFIDF, Glasgow, and Entropy

No	**TFIDF**		**Glasgow**		**Entropy**	
	Training Iteration	Accuracy (%)	Training Iteration	Accuracy (%)	Training Iteration	Accuracy (%)
1	12,417	92	20,000	85	6,125	96
2	15,399	84	20,000	92	5,487	94
3	19,898	90	20,000	88	4,283	90
4	14,802	92	20,000	92	3,838	87
5	15,471	92	20,000	92	2,458	88
6	13,285	82	20,000	95	3,576	88
7	20,000	80	20,000	86	1,283	87
8	20,000	86	20,000	94	3,865	93
9	14,232	80	20,000	85	1,579	92
10	20,000	91	20,000	88	2,074	96
Average	**16,550**	**86.9**	**20,000**	**89.7**	**3,457**	**91.1**

and the data was further employed as the training data for ANN. If the data pattern is more representative than the complete data, ANN would be able to learn more traits from the data. The prediction of the ANN would certainly be more accurate if it could learn more completely. Due to the fundamental consideration of feature weighting for Entropy is deeper where it covers the local and global term weighting aspects; hence it generated more representative data than TFIDF and Glasgow. This is the main reason why Entropy has indicated the most accuracy term weighting schemes shown in Table 10.

Stability performance of a designed network should be one of the considered points during the evaluation of a network performance. Stability performance in this context is also referred to as consistency performance. A network output that always remains in a consistent state or less variation from the standard represents high stability. This would reflect the issues whether the performance of the network should be trusted or not.

Figure 8 indicates the accuracy performance of the network using different kinds of term weighting schemes. Each of the term weighting scheme has been tested ten times and the accuracy rate for each run time has been recorded. The purpose is to observe the stability of the network when different kinds of weighting schemes were implemented.

The variation of the best and poorest accuracy performance for the network is called accuracy gap. If the accuracy gap is smaller, it means the network performance is more stable. The stability of the network could be observed from Figure 8 and Table 10. We noticed that TFIDF has the biggest accuracy gap compared to the Glasgow and Entropy. Table 10 shows the performance of each term weighting scheme for ten

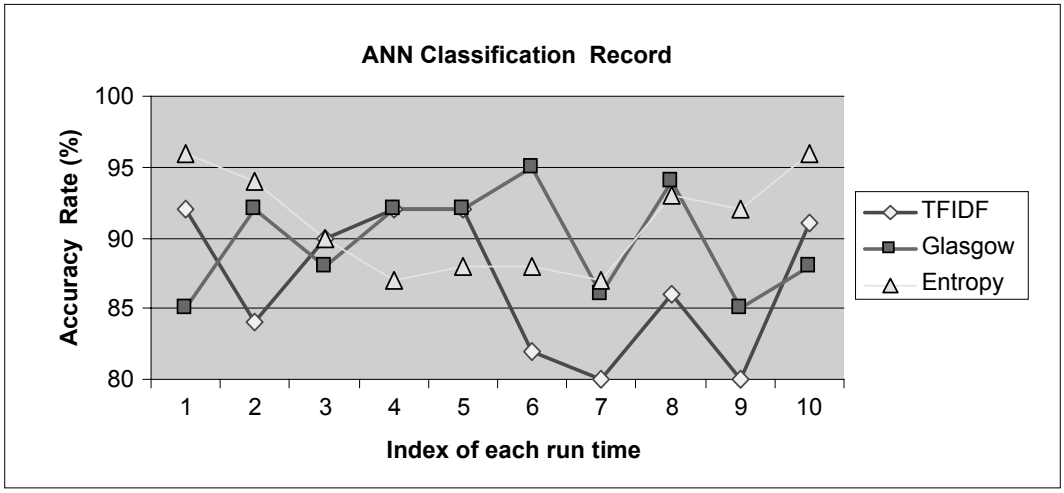


Figure 8 The record for accuracy rates using different kinds of term weighting schemes by ten independent times of neural network classification

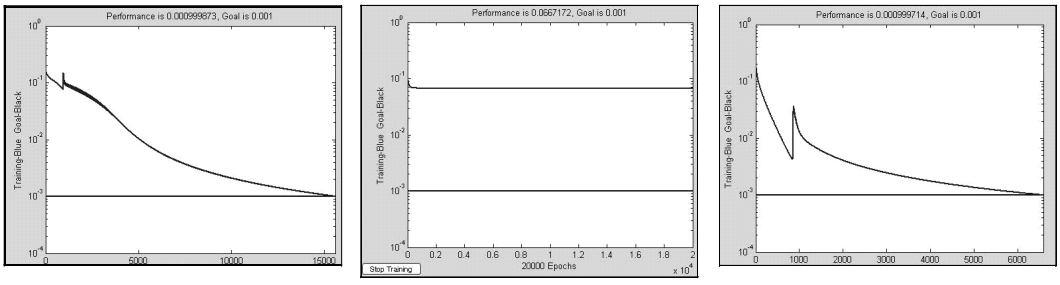
independent run times. From the aspect of stability, the Glasgow and Entropy have provided less performance variation which also means that they are more stable. We believe that TFIDF has represented data with more noise, which lead its accuracy gap of performance to become bigger. In this study, we found that Entropy is the most appropriate term weighting scheme to represent data as input features for neural network because of its lowest accuracy gap as shown in Table 11.

Table 11 Examination of performance gap for term weighting schemes

Scheme	Best Performance	Poorest Performance	Gap
TFIDF	92%	80%	12%
Glasgow	95%	85%	10%
Entropy	96%	87%	9%

On the other hand, we noticed that each of the term weighting schemes have had their unique training patterns as illustrated in Figure 9. In order to identify the learning behavior for each term weighting scheme, we conducted the experiment accordingly and reported the results in Table 8. Among the three term weighting schemes, Entropy achieves the fastest convergence during neural network training which averagely taking 1,283 iterations. However, the average convergences for TFIDF and Glasgow are 16,550 and 20,000 respectively as shown in Table 8.

Figure 9 illustrates the training pattern of BP-ANN during the implementation of TFIDF, Glasgow, and Entropy term weighting schemes as their input features. Meanwhile, the training iterations in this experiment were directly affecting the learning time duration. The more iteration it spends, the longer learning duration it takes. We believe that after Entropy-CPBF, those features were more adaptable to the nature of ANN learning behavior. As a result, Entropy has achieved a faster convergence than the rest two term weighting schemes (it also means consuming less learning duration).



(a) (b) (c)

Figure 9 Training pattern of (a) TFIDF, (b) Galsgow and (c) Entropy term weighting schemes using neural network

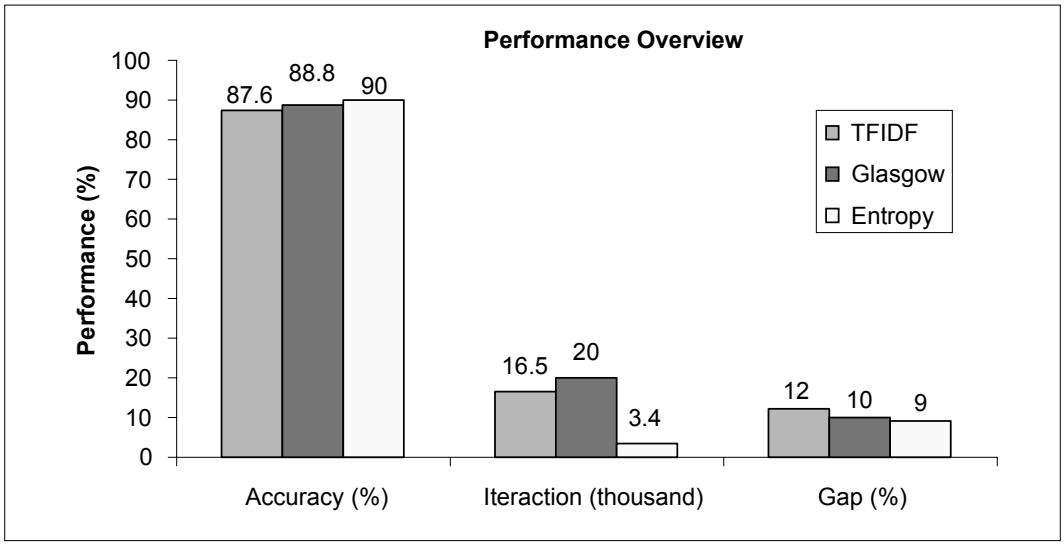


Figure 10 The comparison of overall performance for BP-ANN when it implemented different kinds of term weighting schemes as its input features

Figure 10 indicates the performance overview for BP-ANN in implementing TFIDF, Glasgow, and Entropy term weighting schemes as its input features. The network that implemented Entropy term weighting scheme has averagely achieved the best accuracy rate, the fastest convergence times, and the smallest performance gap among the three term weighting schemes. Regarding the classification result, it has been proven that our proposed PWCE model with entropy scheme providing a better performance than the other two weighting schemes.

5.0 CONCLUSION AND FUTURE WORK

The current existing web filtering approaches are not efficient enough against today's dynamic web contents. Thus, content based analysis techniques with effective model are highly desired. This paper has proposed PWCE model to classify pornography and sex education web pages. We have examined the model from three aspects that are accuracy, convergence speed, and stability. The model with entropy method has performed better result than the TFIDF and Glasgow term weighting schemes in terms of classification accuracy and neural network training convergence speed. The model provides a satisfying stability performance. We have also proved that PWCE model is also efficient in classifying pornography web pages for small class datasets.

We believe that there are still rooms of improvement for PWCE model. In the future, we plan to further expand the test using larger class of datasets and more extensive analysis will be done. In addition, the architecture of ANN for PWCE model could be further improved so that the future PWCE model will perform even better in terms of accuracy rate, convergence speed, and stability.

ACKNOWLEDGEMENT

The authors have been grateful for the comments received from the reviewers especially during PARS'07. Authors also would like to thank the Ministry of Science & Technology and Innovation, Malaysia and the Universiti Teknologi Malaysia for granting the research fund and providing conducive environments for us to conduct this research under the Vot 79210 and Vot 79200.

REFERENCES

[1] *Internet Filter Review.* 2007. Available at http://www.internet-filter-review.toptenreviews.com, 12 May 2007.

[2] *History of Internet.* 2003. Available at http://www.isoc.org/internet/history/brief.shtml, 6 June 2003.

[3] *Road and Crossroads of Internet History.* 2006. Available at http://www.netvalley.com/cgi-bin/intval/net_history.pl?chapter=1, 3 July 2006.

[4] Clark, D. D., K. Sollins, J. Wroclawski and T. Faber. 2003. *Addressing Reality: An Architectural Response To Real-World Demands On The Evolving Internet*. Proceedings of the ACM SIGCOMM 2003 Workshops. Karlsruhe, Germany. 247–257.

[5] Claire A. Simmers. 2002. Aligning Internet Usage With Business Priorities. *Communication of the ACM*. 45 (1): 71–74.

[6] *Usenet - A Breeding Ground for Viruses*. 2001. Available at http://www.computerweekly.com/Articles/2001/06/26/181052/usenet-a-breeding-ground-for-viruses.htm, 26 June 2001.

[7] *NetProtect.* 2003. Available at http://www.net-protect.org/en/default.htm.

[8] Lee, P. Y., Hui, S. C. Fong, A. C. M. Fong. 2002. *Neural Network for Web Content Filtering. IEEE Intelligent Systems*. 17(5): 48–57.

[9] Churchanroenkrung, N., Y. S. Kim and B. H. Kang. 2005. *Dynamic Web Content Filtering based on User's Knowledge*. Proceeding of ITCC'05. 1: 184–188.

[10] Pierre, J. 2000. *Practical Issues for Automated Categorization of Web Sites.* Lisbon, Portugal. Available at http://www.ics.forth.gr/isl/SemWeb/proceedings/session3-3/html_version/semanticweb.html, 1 September 2000.

[11] Du, R., R. Safavi-Naini and W. Susilo. 2003. *Web Filtering Using Text Classification*. The 11th IEEE International Conference on Network (ICON) 2003. 325–330.

[12] Hammami, M., Y.Chahir and L.Chen. 2006. WebGuard: Web Filtering Engine Combining Textual, Structural and Visual Content Based Analysis. *IEEE Transaction On Knowledge And Data Engineering*. 18(2): 272–284.

[13] Chowdhury, G. G. 1999. *Introduction to Modern Information Retrieval*. London: Library Association Publishing.

[14] Yiming Yang, J. O. P. 1997. *A Comparative Study on Feature Selection in Text Categorization*. Presented at Proceedings of ICML-97, 14th International Conference on Machine Learning. 412–420.

[15] Sebastiani, F. 2002. Machine Learning In Automated Text Categorization. *ACM Computing Surveys*. 34(1): 1–47.

[16] Salton and McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGrawHill. USA.

[17] *The Glasgow Model.* 2007. Available at http://www.miislita.com/term-vector/term-vector-4.html, 12 May 2007.

[18] Sanderson, M. and I. Ruthven. 1996. *Report on the Glasgow IR Group (glair4) Submission.* The proceedings of the 5th TREC conference (TREC-5). 517–520.

[19] Lee, Z. S., M. A. Maarof and A. Selamat. 2006. *Automated Web Pages Classification with Integration of Principal Component Analysis and Independent Component Analysis as Feature Reduction*. International Conference Man Machine System (ICoMMS06), Langkawi Island, Malaysia. Available at http://eprints.utm.my/3129/, 15 September 2006.

[20] Selamat, A. and S. Omatu. 2004. Feature Selection and Categorization of Web Page using Neural Networks. *Int. Journal of Information Sciences.* Elsevier Science Inc. 158: 69–88.

[21] Halpin, S. and R. Burch. 1997. Applicability of Neural Networks to Industrial and Commercial Power System: A Tutorial Overiview. *IEEE Trans. Industry Applications*. 33(5): 1355–1361.