CLASSIFICATION OF ILLICIT WEB PAGES USING NEURAL NETWORKS

Lee Zhi Sam[1] , Mohd Aizaini bin Maarof[2], Ali Selamat[3], Siti Mariyam Shamsuddin[4]

Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia,
81300 Skudai, Johor, Malaysia.
Email: samleecomp@gmail.com[1], aizaini@utm.my[2], aselamat@utm.my[3], mariyam@utm.my[4]

Abstract: The illicit web contents such as pornography, violence, gambling, etc, have greatly polluted the mind of web users especially children and teenagers. Due to some popular web filtering techniques like Uniform Resource Locator (URL) blocking and Platform for Internet Content Selection (PICS) checking are limited against today dynamic web content, hence content based analysis techniques with effective model are highly desired. In this paper we propose textual content analysis model using entropy term weighting scheme to classify pornography and sex education web pages. We examine the entropy scheme with two other common term weighting schemes which are TFIDF and Glasgow. Those techniques are examined extensively with artificial neural network using small class dataset. In this study, we found that our proposed model archive better performance from the aspects of accuracy, convergence speed and stability.

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

1. INTRODUCTION

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

Internet is an information superhighway but also the most danger place [6]. Web users always need to pay the risk for theft of information, spamming, virus threat and mental pollution of harmful resource. The objectionable web content such as pornography, violence, gambling, etc. are greatly pollute the mind of immature web users. Pornography perhaps is the one of the biggest threat related to current children's and teenagers' healthy mental life. There is thousands of pornography sites on the internet can be easily found and detected. In order to analysis the impact of pornography web pages to the social, Jerry Ropelato[1] had conduct a web survey. Table 1 and Figure 1 show the statistics for top adult search request during year 2006 based on the web survey. Figure 2 illustrates the average proportion for different ages of people that involve in adult search request using search engine. Based upon figure 1 and 2, we found that 23% of figure 1 result is contributed by teenager and children who their ages are below 18. This will certainly become a detrimental factor to letting children and teenagers access internet without proper guiding.

Table 1. Top ten request term for adult search request 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

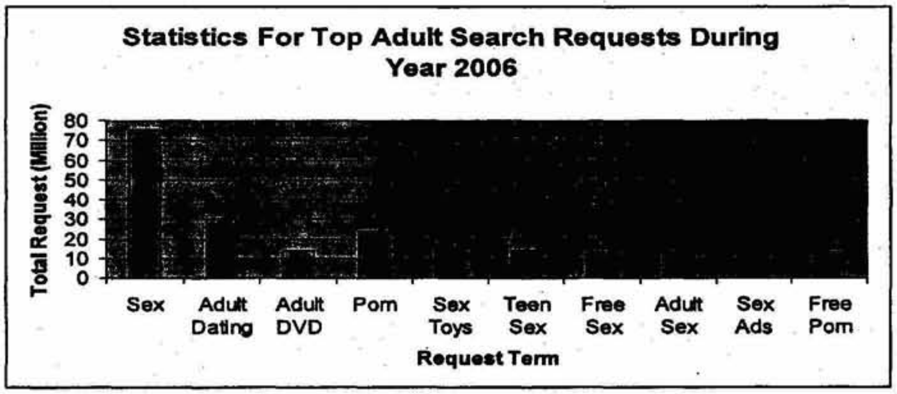


Figure 1. Statistics for top adult search requests during year 2006 by Jerry Ropelato [1]

Teenagers browsing the adult web pages without acknowledgement from parents make web filtering and monitoring system is highly required in family and education environment.

In fact there are various commercial web filtering products available in the market. Some popular products are CyberPatrol, NetNanny, Websense, CyberSitter, etc (see [1]). The simplest and most popular solution to filter the harmful resources is simply block the Uniform Resource Locator (URL) or Internet Protocol (IP) address of particular link. Some other techniques are using the Platform for Internet Content Selection (PICS) checking or keyword matching. The PICS checking is a technique that checking the labels content of PICS (which also call metadata) for web pages and further block the web page if it is harmful resource. Keyword matching is a technique that will block 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 above technique especially URL and IP address blocking.

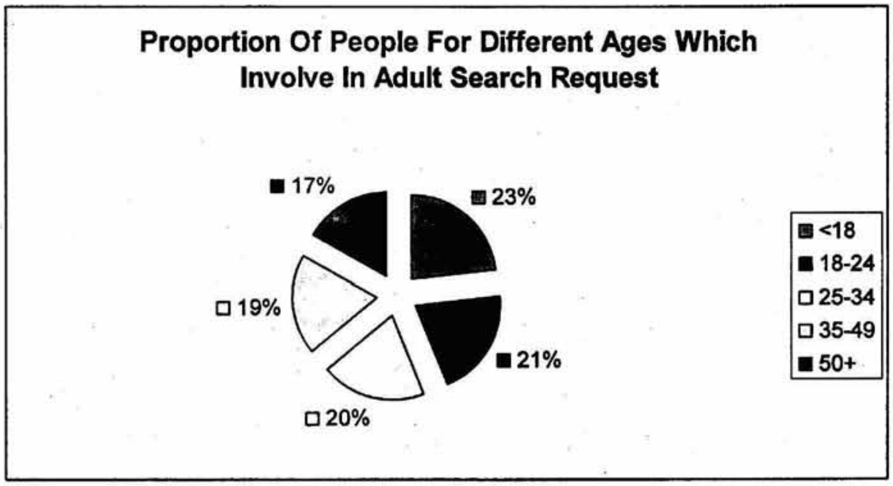


Figure 2. Statistic based on the proportion of people for different category of ages which involves in adult search request [1].

The above techniques are fast and require only short processing time, however they are always fail to block those unknown web pages which is not in their list. The technique such as URL or IP blocking has even become insufficient against current dynamic web content [8]. It is due to URL blocking will only block particular web pages regarding to the database where store those black listed URL or IP address [9]. Unfortunately with current limited technology, it is hardly obtaining the complete URL list of whole World Wide Web (WWW) since there are million of web pages being added daily [10]. The trust issue is always an argument for PICS checking technique since the web publishers having the right to label whatever content to the metadata [9]. Hence PICS only suggested as supplementary filtering technique due to it is weak against ever-changing web content. The keyword matching technique is designed to overcome the dynamic content issues; however it is not efficient during different subjects but having similar terminologies web pages [12]. For instances this technique will block both pornography and sex education web pages since intentionally the users only need to block pornography web pages. Thus, heuristic content analysis technique with effective model is highly desire in order to archive a better performance for pornography web page classification.

We use 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 able to analysis the content body of web pages and provide efficient classification result against unknown web pages [8]. This technique is adapted in our pornography web page classification with textual content analysis model. In order represent the textual content of web pages to be understood by computer machine, efficient term weighting scheme is crucial. This paper we propose textual content analysis model using entropy term weighting scheme to classify pornography and sex education web pages. We examine the entropy scheme with Term Frequent Inverse Document Frequency (TFIDF) and Glasgow term weighting schemes. The output of the schemes will be the input for artificial neural network in order to test the effectiveness of each scheme.

2. 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 hardly to understand the meaning of natural language without some proper ways. Term weighting scheme is statistical measure used to evaluate how important a word is to a document in a collection [13]. In other word, sets of numeric number would obtain through term weighting scheme. These sets of number will be understood by computer machine and use for further analysis and document classification. Normally there are three main factors term weighting which are term frequency factor, collection frequency factor and length normalization factor [14]. These three factor are multiplied together to make the resulting the term weight. We compare the effectiveness of entropy weighting scheme with two other weighting scheme which often used by search engines to score and rank a document's relevance given a user query. Those weighting scheme are Term Frequent Inverse Document Frequent (tfidf) and Glasgow weighting scheme.

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 will be represented as the document-term frequency matrix (Doc$_j$ x TF$_{jk)}$. The TFIDF function weights each vector component of each document 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 illustrate 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) measure how infrequent a term is in the collection. The value is estimate using the whole collection documents. The TFIDF is based on believe that if a word is infrequent in the text collection, it is estimated to be very relevant for 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]

$$x_{jk} = TF_{jk} \text{ x } idf_k \qquad (1)$$

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 scheme with main advantages that too long documents and queries can be penalized [17].

However Glasgow would only effective with full fill two conditions. First, the value for Glasgow must using normalized frequencies. Second, during defining the length of documents and queries as number of terms must excluding stop words. Glasgow term weighting scheme can be express as follow, as found in M.Sanderson[18]

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

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 term *j* occur

3. Pornography Web Page Classification with Entropy Term Weighting Scheme Model

The main purpose for Pornography Web Page Classification with Entropy Term Weighting Scheme (PWCE) model is to classify the objectionable and healthy web pages which are pornography and sex education web pages. PWCE model is mainly based on textual content analysis which employs entropy method as its term weighting scheme. This model is constructed by several parts which 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 crawl the web pages from internet. Those web pages will go through the stemming and stop world filtering process to reduce those noise features in 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 web page will be classified at ANN classifier. Finally the classification results are categorized to two groups which are pornography and non-pornography (sex education).

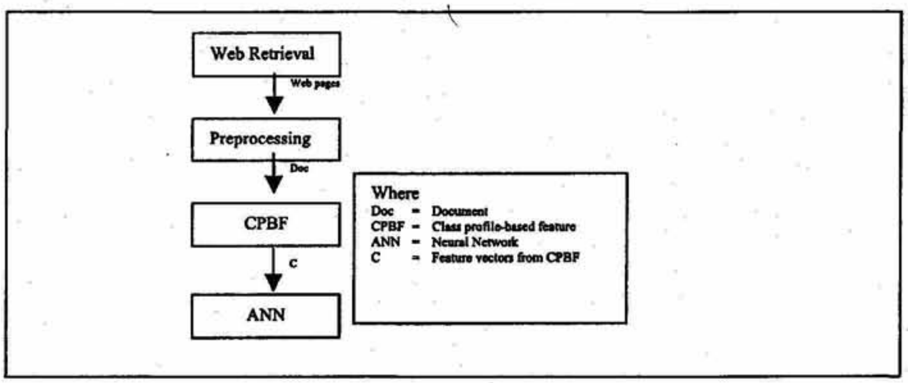


Figure 3. Overview of PWCE Model

3.1 Web Page Retrieval

Web page retrieval is a part that using web robot retrieves the desire web page to database. The operation process for web page retrieval is illustrated in figure 4.

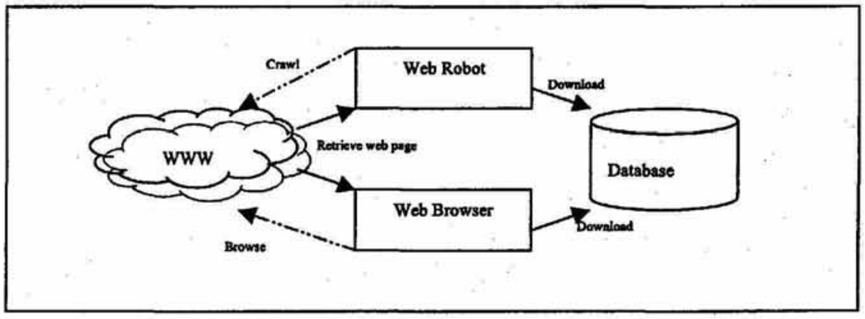


Figure 4. The process of web page retrieval

However web page retrieval part also could be done manually with using web browser. This part can either be a web crawler or web browser as long as the web pages are retrieved as require and store in the database. Shortly, the main task of web page retrieval is to obtain the web resource from WWW and duplicate it to database which will store for further analysis purpose.

3.2 Preprocessing

Web pages in database will go through HTML parsing in order to transform web page become a text document. The documents will perform stopping and stemming before passing to CPBF section. Stop-list is a dictionary that contains the most common and frequent words such as 'I', 'You', 'and' 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 is similar meaning in the same document. It is a process of extract 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'. The workflow of pre-processing is shown in figure 5. The output of the pre-processing would be as input for class profile based features (CPBF) for further process.

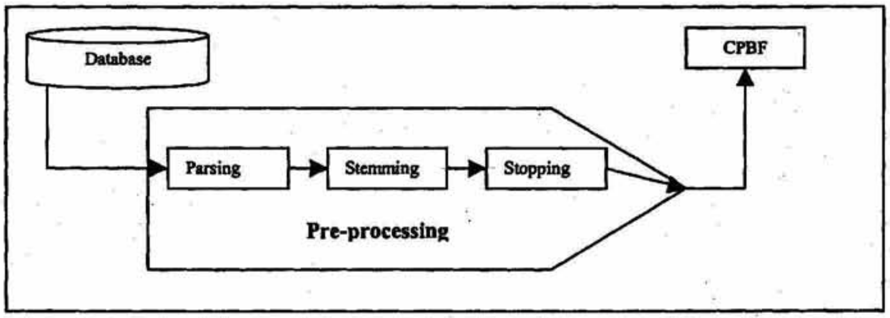


Figure 5. The workflow of pre-processing in PWCE model.

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 the weights of them by implement term weighting scheme. In CPBF, we identify those most regular words and weight them using the entropy term weighting scheme before feed them to 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 providing a more accurate weights especially compare to TFIDF. It is due to it concern the term weighting from two aspects which are local term weighting and global term weighting. This mean once every term receives a weight, it will compose to local and global weights. It is calculate on the range of [0,1], hence the value are normalized. The entropy term weighting scheme which implemented in our experiment can be express as follow, as found in Lee. et.al [19] and Selamat et.al. [20]

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

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

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

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 (or call nodes), each computing a nonlinear function of a weighted sum of its inputs. A typical ANN would constructed by three main layers which 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 would represent the number of category that the network could classify. The more nodes in output layer, the network could classify the data to more categories. 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 are very based on algorithms such as back propagation, KSOM, etc (see [21]) . Basically the learning algorithms are concern to set the weight connections by training the net with a given data set until it archive a certain goal. It is always a very challenging task to design a proper network that would solve a complex issue with more simple 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. ANN is advantage from its learning behavior where it could learn the trait of "train" data during establishing their input-output relationship. However ANN is only learn strictly based on "train" data. In other saying, the trait that not existing in the "train" data are impossible to learn by ANN. Hence the input features for ANN training should be select carefully and efficiently so that it is as representative of the complete data set as possible.

We using CPBF as features selection before the selected features feed to ANN is to ensure only the most representative features are as input for the network.. In this model the artificial feed forward-back propagation neural network (ANN) is adopted as the classifiers. For classifying a test document, its term features are feed to input node. Later, term weight are load into the input units. The activation of these units will propagate 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 due to if a misclassification occurs the error is "backpropagated" so as 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 7. The number of input layers (L) is equal to the number of term features after CPBF which are 30. The number of hidden layers (M) is half of the input layers which are 15. The number of output layers (N) is one since the network only classify the web pages to two classes. We interpret the notation as follow: 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:

$$W_{LM}(i+1) = W_{LM}(i) + \Delta W_{LM}(i+1) \qquad (6)$$

where

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

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

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

$$\tan sig(L) = \frac{2}{(1+e^{-2L})-1} \qquad (9)$$

$$A_L = \tan sig(L) \tag{10}$$

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

$$net_M = \sum_M W_{LM} A_L + \theta_M \tag{11}$$

$$A_M = \tan sig(net_M) = \frac{2}{(1 + e^{-2net_M}) - 1}. \tag{12}$$

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

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

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

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

$$net_N = \sum_N W_{MN} A_M + \theta_N \tag{16}$$

$$A_N = \tan sig(net_N) = \frac{2}{(1 + e^{-2net_N}) - 1}. \tag{17}$$

The detail implementation of BP-ANN for this experiment would be explained in section 4.3.Moreover, the parameters setting for error back-propagation neural network are also indicated in table 7 and 8.

3.5 Result Examination

The classification result will be examine as follow

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

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 that used for examination

The classification result will be examined with (6) in order to evaluate the performance of each term weighting scheme. The higher value of *Accurate,* the better it is.

4. EXPERIMENTS AND RESULTS

4.1 Data Sets

We collect 70 web pages which are reviewed to make standard classification. In order to simplify the experiment, this study will only classify the web pages to two categories which are pornography and sex education (non-pornography). The pornography web pages are referring to those adult web pages which display the sexual activity. On the other hand, the non-pornography web pages in this experiment are referring to those web pages which display useful and informatics contain. The sex education web pages here including the subject such as medical sex, sex physiology consultation, health news and education information related to sex. Table 5 summarizes the data. Due to pornography, and sex education web pages always having high similarity within each other, hence this experiment use the mix of medical sex and sex education web pages as non-pornography category. The purpose to do so is to prove that this model able to perform extensive classification with textual content analysis.

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

Category	Web Pages	Ratio
Pornography	40	57.14
Non pornography (sex education)	30	42.86
Total	70	100

4.2 CPBF AS FEATURE SELECTION

For the feature selection using the class profile-based approach, we identify the most regular terms that exist in pornography and sex education categories. We weight the terms using TFIDF, Glasgow and Entropy term weighting scheme respectively. We select thirty terms that having the highest value from each term weighting scheme as input vector for ANN classifier independently which illustrated at figure 4. The number vector for *t*, *g* and *e* are fixed as 30. Hence ANN classifier will act as the base line to examine the three kind of term weighting scheme.

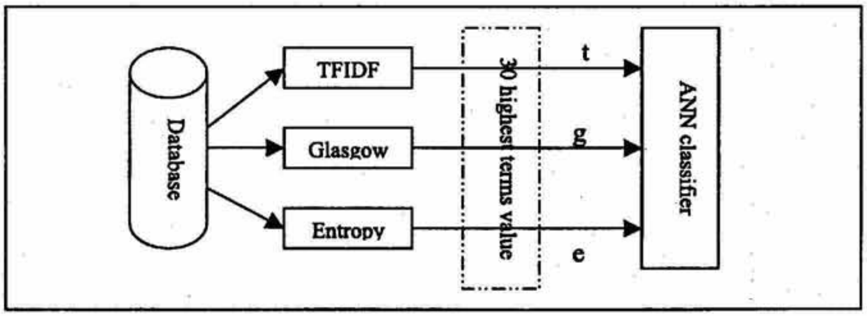


Figure 6. CPBF using TFIDF, Glasgow and Entropy term weighting scheme as each feature selection method for ANN classifier.

4.3 Parameters for ANN Classifier

In order to do the classification, we implement the back-propagation neural network as our classifier. We have used a set of documents as shown in table 4 and specification of network which summarized in table 5. We choose 20 documents as training set which consists of 10 pornography and 10 sex education web pages. On the other hand, we select 50 documents as testing set which consists of 30 pornography and 20 non-pornography web pages.

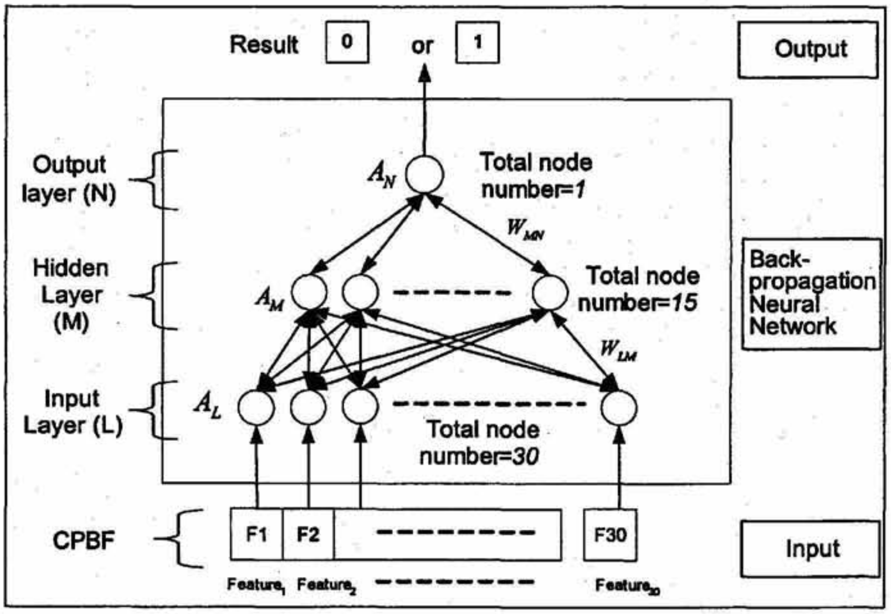


Figure 7. The architecture design of back-propagation neural network in PWCE model.

We design the architecture of ANN to three-layer which are one input layer, one hidden layer and one output layer. Due to they are 30 features from CPBF as input, we design 30 input nodes at input layer where each node feeds one feature. The hidden layer consists of 15 hidden nodes and one node for output layer. The output layer will return two values which represent one category for each value. The return value is 0 which represent the document is non-pornography content. In contrast, the value is 1 which the document is pornography content.. Figure 7 represents the architecture design of BP-ANN in our PWCE model.

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

Table 7 Training and Testing Set for PWCE with Neural Network.

	Training Documents	Testing Documents
Pornography	10	30
Non Pornography	10	20
Total	20	50

Table 8 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

4.4 Classification Result

The three terms weighting schemes, TFIDF, Glasgow and Entropy are 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 8. The classification result using each term weighting scheme are shown in table 8. The average accuracy rate using TFIDF, Glasgow, and Entropy term weighting schemes are 87.6%, 86.4% and 90% respectively. We compare the accuracy is calculated based on the formula (6). We examine the proportions of the documents are classify correctly to their particular categories in a collection of documents. The more documents are classified correctly to their particular categories; the value that returns by formula (6) would be higher which indicate more accurate.

The fundamental characteristic of intelligent is learning and prediction ability. As mentioned previously, ANN would only learn the pattern that existed in train data. Different term weighting schemes would generate their unique pattern of data and the data is further employed as the train data for ANN. If the pattern data is more representative as complete data, ANN would 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 a more representative data than TFIDF and Glasgow. This is the main reason that Entropy indicated as most accuracy term weighting scheme which shown in table 8.

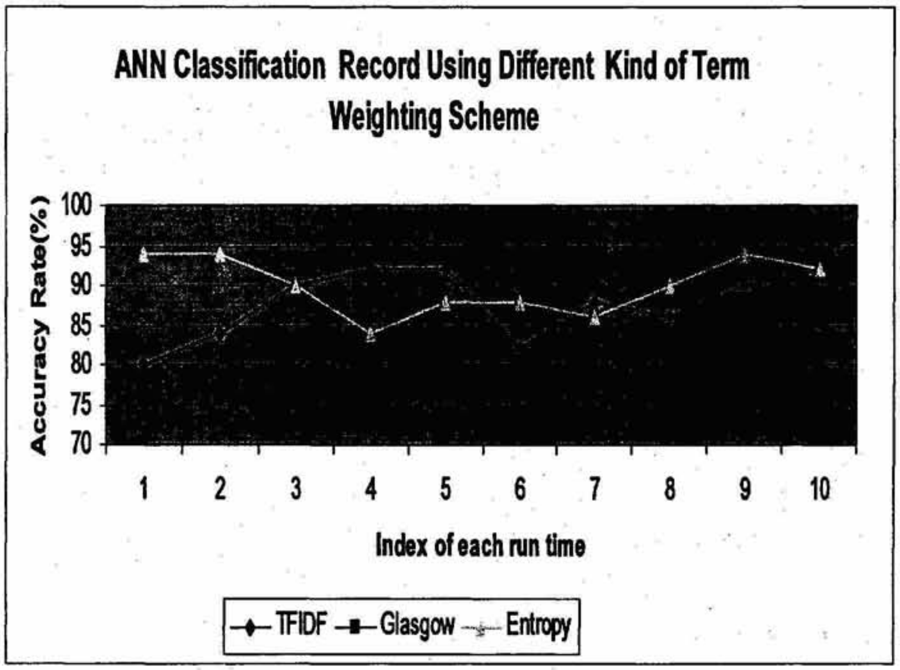


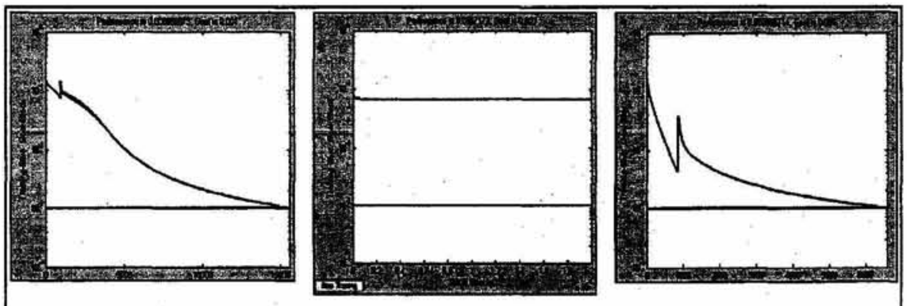
Figure 8. The record for accuracy rate which using different kind of term weighting scheme by ten independent times of Neural Network classification.

Stability performance of a designed network should be one of the consideration point during evaluate a network performance. This would reflect the issues that should be the performance of the network being trusted or not. A network output that always remains in a constant state or less variation from the standard is represent as more stable. Figure 8 indicates the accuracy performance of the network when the network is using different kind of term weighting scheme. Each of the term weighting scheme is being tested ten times and the accuracy rate for each run time is recorded. The purpose to do so is to observe the stability of the network when different kind of weighting scheme is being implemented.

The variation of the best and poorest accuracy performance for the network is called accuracy gap. The smaller gap within the best and poorest accuracy performance of the network, the more stable it is. The stability of the network could be observed from figure 8 and table 9. We notice that TFIDF having the biggest gap within best and poorest accuracy rate when it compare with Glasgow and Entropy. Table 10 shows the gap of performance for each term weighting scheme with 10 independent run times. From the aspect of stability, Glasgow and Entropy provide a less performance variation which also means they are more stable. We believe that TFIDF represent the data with more noise, so the gaps of performance are bigger. In other word, Glasgow and Entropy would be more appropriate to represent data as input features for neural network.

Table 9. Accuracy rate for TFIDF, Glasgow and Entropy

No	TFIDF		Glasgow		Entropy	
	Training Iteration	Accuracy (%)	Training Iteration	Accuracy (%)	Training Iteration	Accuracy (%)
1	13,771	80	20,000	88	7,165	94
2	20,000	84	20,000	88	4,497	94
3	20,000	90	20,000	88	3,293	90
4	14,802	92	20,000	92	4,838	84
5	16,671	92	20,000	88	1,458	88
6	15,585	82	20,000	88	6,576	88
7	20,000	88	20,000	86	5,283	86
8	20,000	86	20,000	94	3,865	90
9	20,000	90	20,000	84	1,179	94
10	20,000	92	20,000	92	1,074	92
Average	**18,087**	**87.6**	**20,000**	**88.8**	**3,923**	**90**



(a) TFIDF

(b) Glasgow

(c) Entropy

Figure 9. Training pattern of (a) TFIDF, (b) Galsgow and (c) Entropy term weighting scheme using Neural Network.

Table 10 Examination of gap performance for term weighting schemes

Scheme	Best Performance	Poorest Performance	Gap
TFIDF	92%	80%	12%
Glasgow	94%	84%	10%
Entropy	94%	84%	10%

We notice that each of the term weighting scheme have their unique training pattern. In order to identify the learning behavior for each term weighting scheme, we done the experiment accordingly and reported in table 8. Among the three term weighting schemes, Entropy archive the fastest convergence during neural network training which averagely taking 3,923 iteration. However the average convergence for TFIDF and Glasgow are 18,087 and 20,000 respectively which shown by table 8. Figure 9 illustrate the training pattern of BP-ANN when implement TFIDF, Glasgow and Entropy term weighting scheme as its input features. Meanwhile, the training iterations in this experiment are directly affects the learning time duration. The more iteration it spends, the longer learning duration it takes. We believe that those features which after Entropy-CPBF are more adaptable to the nature of ANN learning behavior, as result Entropy do archive a faster convergence than the rest two (it also mean consume less learning duration).

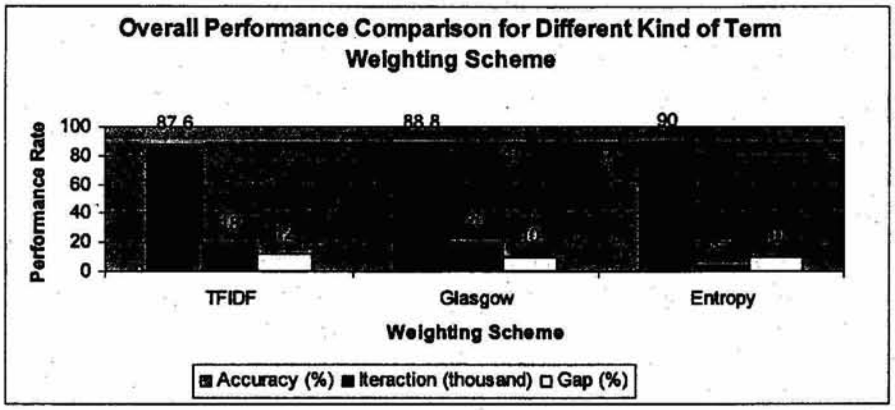


Figure 10. The comparison of overall performance for BP-ANN when it is implements different kind of term weighting scheme as its input features.

Figure 10 indicates the comparison of overall performance for BP-ANN when it is implements TFIDF, Glasgow and Entropy term weighting scheme as its input features. The network that implement Entropy term weighting scheme averagely achieve the best accuracy rate, the fastest convergence times and the smallest performance gap among three term weighting scheme. Regarding to the classification result, it prove that our proposed PWCE model with entropy scheme providing a better performance than other two weighting scheme.

5. Conclusion and Future Work

The current existing web filtering approaches are not efficient enough against today dynamic web content. Thus content based analysis techniques with effective model are highly desired. This paper we proposed PWCE model the classify pornography and sex education web pages. We examine the model from three aspects which are accuracy, convergence speed and stability. The model with entropy method does perform a better result than TFIDF and Glasgow term weighting schemes in term of classification accuracy and neural network training convergence speed. The model provide a satisfy stability performance. We do prove that PWCE model is efficient against pornography web pages classification for small class datasets.

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

Acknowledgement

The authors are graceful for the comments from the PARS'07 and reviewers. Authors also would like to thanks the Ministry of Science & Technology and Innovation, Malaysia and Universiti Teknologi Malaysia for supporting the research fund and conducive environments in conducting the research under the Vot 79210 and Vot 78099.

References

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

[2] "History of Internet" available at http://www.isoc.org/internet/history/brief.shtml, 10 December 2003.

[3] "Road and Crossroads of Internet History" available at http://www.netvalley.com/cgi-bin/intval/net_history.pl?chapter=1, 18 October 2006.

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

[5] Claire A. Simmers, Aligning Internet Usage With Business Priorities, Communication of the ACM, Vol 45 Iss 1, page(s): 71-74, January 2002.

[6] "Usenet - a breeding ground for viruses" available at http://www.computerweekly.com/Articles/2001/06/26/181052/usenet-a-breeding-ground-for-viruses.htm, 26 Jun 2001.

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

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

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

[10] J. Pierre, Practical Issues for Automated Categorization of Web Sites, Lisbon, Portugal 2000

[11] R. Du, R. Safavi-Naini, W. Susilo, Web Filtering Using Text Classification, The 11th IEEE International Conference on Network (ICON) 2003, page(s): 325-330, 28 September- 1 October 2003.

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

[13] G. G. Chowdhury, Introduction to modern information retrieval, London: Library Association Publishing, 1999.

[14] J. O. P. Yiming Yang, A comparative study on feature selection in text categorization, presented at Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997

[15] F.Sebastiani , Machine Learning In Automated Text Categorization, ACM Computing Surveys, vol 34, no. 1, pp. 1-47, 2002.

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

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

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

[19] Z.S. Lee, M.A. Maarof, A. Selamat, 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, 15th-16th September 2006.

[20] A. Selamat and S. Omatu, Feature selection and Categorization of Web Page using Neural Networks, Int. Journal of Information Sciences, Elsevier Science Inc. Vol. 158, pp. 69-88, January 2004.

[21] S.Halpin and R. Burch, Applicability of Neural Networks to Industrial and Commercial Power System: A Tutorial Overiview," IEEE Trans. Industry Applications, vol. 33, no.5 pp. 1355-1361, 1997.