Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 6, No. 1, February 2015), Pages: 101-114
www.jacr.iausari.ac.ir

Using an Automatic Weighted Keywords Dictionary for Intelligent Web Content Filtering

Najibeh Farzi Veijouyeh[1]✉, Jamshid Bagherzadeh[2]

1) Islamic Azad University of Shabestar Branch, Shabestar, Iran
2) Assistant professor, Computer Science and Eng. Deptt, Urmia University, Urmia, Iran
Najibeh.Farzi@yahoo.com; J.Bagherzadeh@urmia.ac.ir

Received: 2013/07/12; Accepted: 2014/05/26

Abstract

Filtering of web pages with inappropriate contents is one of the major issues in the field of intelligent network's security. Having a good intelligent filtering method with high accuracy and speed is needed for any country in order to control users' access to the web. So, it has been considered by many researchers. Presenting web pages in an understandable way by machines is one of the most important preprocessing steps. Thus, offering a way to describe web pages with lower dimensions would be very effective, especially in determining the nature of web pages with respect to whether they should be filtered out or not. In this paper, we propose an automatic method to detect forbidden keywords from web pages. Next, we define a new representation of web pages in vector form which consists of weighted sum and frequency of forbidden keywords in different parts of web pages named RWSF. For this, a ranking dictionary of keywords including forbidden keywords is used. To evaluate the proposed method, 2643 pages consisting of 1311 normal pages and 1332 forbidden pages were used. Among these, 1851 pages were used to train the system and 792 pages were used for system evaluation. The system has been assessed using various classifiers such as: k-Nearest Neighbor, Support Vector Machines, Decision Tree and Artificial Neural Networks. Evaluation results indicate the high efficiency and accuracy of the proposed method in all classifiers.

Keywords: *Content based filtering, Forbidden keywords extraction, Ranking keywords, Web page representation*

1. Introduction

The number of web pages has expanded greatly because of the fast growth of the World Wide Web. The indexed Web contains at least 8.33 billion pages until July 8, 2012.

Web page filtering has various purposes. For instance, protection against improper content is one of the major web page filtering purposes. Web provides advantageous space for users to gain all kinds of information. But this space has been filled with a number of harmful web pages , like pornography, violence, racism, and so on. In 2001, the Online Computer Library Center's annual review found 74,000 adult websites accounting for 2% of sites on the net, and they brought in profits of more than $1 billion together; many were small scale, with half making $20,000 a year. Consequently, web filtering can be used to block access to pages that are against the established policy.

Another purpose of filtering is to avoid misusing of the network. A survey on International Data Corporation (IDC) proved that people spend one third of their on-line time, on tasks other than their job- related tasks . It is obvious that Internet accelerates the communication process and makes research activities more effective, however it also has some problems obviously. Employees mostly use the internet for personal activities such as on-line shopping, chatting with friends or downloading material during work hours, which decrease productivity and responsibility of the company they work for.

Hence, in recent years, plenty of researchers have obtained noticeable interest in studying and offering a solution to manage and filter improper information on the web. There have been plenty of filtering methods in the system, which can be approximately divided into four major categories as follows [1, 2, 3, 4]:

- Blacklist and white-list: Blacklist contains banned web sites, which cannot be accessed, and white-list contains the pages, which are allowed. Regarding a new web page, it is available or forbidden depending on the requested URL, matching either blacklist or white-list. There is an obstacle here. Keeping the URL lists complete and up to date is a very tough task.
- PICS: PICS (Platform for Internet Content Selection) can develop ranking for web sites. There are usually two measures to rank the web pages. the first one is self-ranking and the second one is other ranking. The difference between two originates from the case that if the ranking results are given by web publishers or not. Filtering systems can operate by means of ranking information of web sites. The PICS is not an obligatory labeling system, so the ranking information is not always reliable.
- Keywords filtering: This method is an easy approach to block access to web sites which function according to the occurrence of forbidden words . In this method, a list of forbidden words or phrases is often required. Hence, the web page is blocked when the number of forbidden words in the web page is more than predefined limits. The problem with the keywords analysis based filtering systems is that they rely on the keyword lists for a great deal, which need great effort. Besides, finding enough particular keywords in some fields is hard. The meaning of the word depends on the context. For example, if it is supposed to filter contents by matching keywords for instance a word like "sex",it may mistakenly block web sites about genders. For this reason, this method will unavoidably cause over-blocking. In addition, this method can easily be defected due to misspelled words.
- Intelligent approach to web content filtering: A web filtering system can use intelligent approach to analyze the content. For instance, training models or data mining techniques are efficient ways to classify web contents automatically. Content analysis is a worldwide method for web page filtering task because it is well-known that illegal web sites include particular text, image and other information that can assist us to filter them. Supervised learning methods are used broadly in web page filtering systems. The problem with supervised learning methods is that a great set of high-quality labeled samples are needed, and they

are hard to obtain. Semi-supervised training methods are efficient when the available labeled sample set is not large.

In this paper, we have proposed a brand new method for web page representation. In the proposed method, we have used weighted forbidden keywords dictionary to represent web pages. We have compared it with *TFIDF* method in accuracy, training time and memory usage. We also evaluated the effect of weighted forbidden keywords dictionary in accuracy of the proposed method by using different classifiers.

The remainder of the paper is organized as follows. In the section 2 we start out reviewing the related works on web filtering. The architecture of our filtering system is described in the section 3. Web page classification in our system is explained in the section 4. In the section 5 we describe the proposed method for document representation and weighted keywords dictionary. Experimental results are given and discussed in the section 6, prior to the conclusion in the section 7.

2. Related Work

Machine learning methods such as k-nearest neighbors (kNN), Neural Network, Decision Trees, Support Vector Machines (SVM), Neural Networks (NN) are broadly used in web page filtering problems [5, 6, 7, 8, 9,17].

Du et al. [1] proposed a web filtering system that uses text classification approach to classify web pages into desirable and undesirable ones. Similarities between the input web page and all training web page samples are averaged and compared with a threshold to determine the label of the input page. The system was trained with a training dataset of 487 adult URLs, without any non-adult URLs and we used a database that included 329 adult URLs and 587 non-adult URLs to test the system performance. Their method achieved a high accuracy on the data set containing adult texts from the adult category of Yahoo. Because the styles of pornography texts and stories are not the same, so this approach cannot work well in the real world [10].

In [10], Wu et al. introduced a system like a Cellular Neural Network word net to extract and reflect semantic and statistic aspects of texts. They analyzed different types of keywords alongside obvious keywords, hidden keywords and logical keywords. SVM was applied as a classifier. In order to evaluate the performance of their system, they used a dataset containing 3162 Chinese texts among them 577 were tricky texts, 585 texts were related to sex but normal at the same time and 2000 normal texts. 300 tricky texts, 300 sex-related normal texts and 1000 normal texts were used as training data, and the rest acted as test data. Also they gathered list of 109 expressive terms containing 29 apparent keywords, 33 hidden keywords and 47 logical keywords.

Their experimental results showed that three kinds of keywords can improve the recognition rates noticeably. They obtained the best classification rate using the CNN-

like word net to extract aspects of texts too. It affirms that CNN-like word net can accurately represent the semantic features of tricky texts.

Chen et al. [11] first used a C4.5 decision tree to classify input pages into three classes of continuous texts, discrete texts, and image pages. A CNN net is applied to recognize the semantic relations within continuous texts and a naïve Bayesian algorithm is adopted to identify discrete texts. After that, a fusion classifier based on Bayes Theorem approach integrated texts and images and 91.8% classification rate was gained over 1500 sample pages. Using only URLs and keywords instead of a content based analysis, as well as a small set of test data, and relatively low accuracy rate are some shortcomings of their work [12].

He et al. [2] used a semi-supervised framework for web page filtering. The Adaboost algorithm was used as a classifier. The experimental results show that semi-supervised learning approach outperforms supervised method when available labeled sample set is small.

Feature reduction should be employed to decrease the number of feature terms to an acceptable level before filtering. In [13] authors proposed to use a rough set to reduce original feature terms. After selecting features, all web pages were represented by the feature vector with the weighting function. They also presented a brand new coefficient weighted method based on rough set to Bayesian formula. The method improves filtering performance but it is not very efficient to increase filtering correctness.

In [19], Ma proposed a neural network method for determining the existing status of a requested URL in the large prohibited collection. The simulation results show superior performances in both memory requirement and speed, comparing with a database implementation on the same PC.

3. Filtering Architecture

We use the combination of the three methods including black list, keyword blocking and intelligent content filtering for web page filtering. The formulation of our system architecture is as follows:

1) URL is launched.
2) If the site exists in the blacklist, block the page and stop.
3) Load the page's HTML source code.
4) If the frequency of the forbidden keywords in the page is more than a predefined threshold, classify the page as forbidden page and go to the (6).
5) Analyze the content of the web page and make a further decision on the site regarding whether to allow access or deny it.
6) Block the page if it is judged as a forbidden page and update the blacklist.

Figure 1 shows the general architecture of our filtering system.

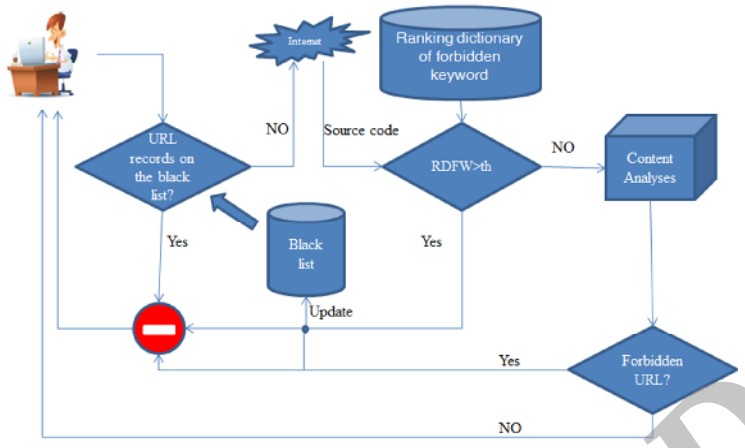

Figure 1. Filtering Architecture

4. Web Page Classification

Machine learning techniques provide us powerful ways to automatically predict forbidden web pages using manually classified web pages. **Figure 2** illustrates the general schema of the proposed approach. It consists of two phases: generating a predication model phase and detection phase.

For web page representation, the web pages have to be transformed from the full text version to web page vectors. The First step consists of tokenization, stop word removal and word stemming to make a vocabulary, where each term occurs at least once in a certain number of web pages. In the second step, we prepare the forbidden keywords using vocabulary and calculate their ranks. After that we represent all the training web pages as vectors of 18 features using the ranking dictionary of forbidden keywords obtained in the previous step. Web page vectors are used as inputs to learn and make a model (classifier) for predication.

In the detection phase a new web page is converted to its corresponding vector using forbidden keywords and their ranks, then the classifier classifies it.

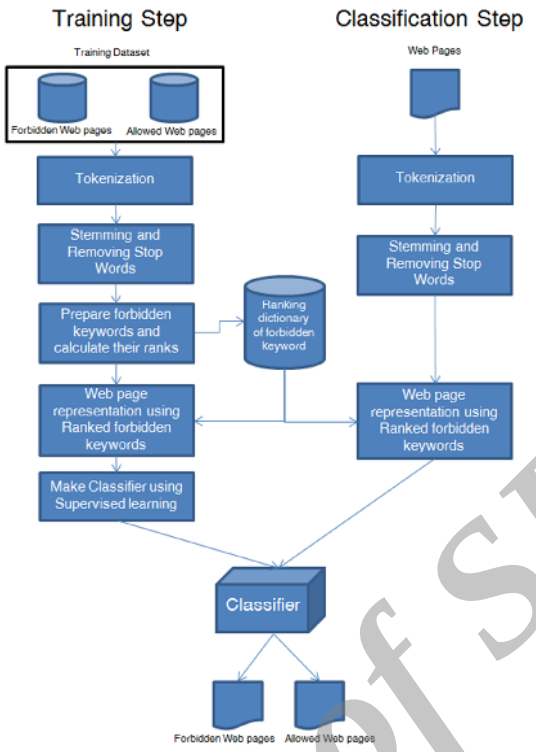

Figure 2. General schema of the proposed approach

5. Web Page Representation

Algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired. Nevertheless, web page representation is one of the preprocessing techniques that is used to reduce the complexity of the documents and make them easier to handle. Web page representation is an important aspect in web page classification, which denotes the mapping of a web page into a compact form of its content.

5.1 Feature's vector with the TFIDF weighting function

A web page is typically represented as a vector of term weights (word features) from a set of terms (vocabulary). Vocabulary is the set of all distinct words and other tokens occurring in any web page from training dataset [18]. A major characteristic of the web page classification problem is the extremely high dimensionality of web page data.

After selecting feature subsets, all documents were represented by the feature vector with the *TFIDF* weighting function. That is, the weight of term t_i in document d_j is calculated by

$$W_{ij} = tfidf\ (t_i, d_j) = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{M} f_{kj}^{2}}} \times log\left(\frac{N}{n_i}\right) \quad (1)$$

Where f_{ij} denotes the number of times, t_i occurs in document d_j, $n(t_i)$ the number of documents in which t_i occurs at least once, N the total number of documents, M is the size of the feature subset.

5.2 Feature's vector with the RW*SF representation's method*

Hammami et al. [4, 14, 15] use another method to represent web pages. They represent web pages as vectors of numbers, which show numbers and frequencies of forbidden keywords in different parts of web pages such as title, body, links, etc. As the speed of filtering is important, this method is a good way for representing web pages. The created vectors would have less dimensions which speed up creating a classifier and consequently web page classification. In all the papers, which use forbidden keywords dictionary to represent web pages, dictionary is made by experts based on forbidden groups, except the method of [15], which creates semi-automatic dictionary based on n-grams that has high accuracy in contrast to manual and automatic methods. In semi-automatic methods there is a need for experts to select keywords which are cost consuming and error prone.

In this paper we propose an automatic method based on Chi-square [9] to select forbidden keywords based on training documents. The term-goodness measure is defined as:

$$X\ (t_i) = \frac{N\ (a_i d_i - b_i c_i)}{\sqrt{(a_i + b_i)(a_i + c_i)(d_i + b_i)(c_i + d_i)}} \quad (2)$$

Where a_i is the number of times t_i occurs in the forbidden web pages, b_i is the number of times t_i occurs in the normal web pages, c_i is the number of forbidden web pages without t_i, d_i is the number of normal web pages without t_i and N is the total number of webpages.

Using this formula we can choose a number k of keywords as forbidden keywords where their goodness is more than the predefined threshold.

In all provided papers and systems which use forbidden keywords dictionary to represent web pages, number and frequency of forbidden keywords have been considered as main features. These methods give equal importance to all forbidden keywords of dictionary. However, when we need k number of keywords, all of them are not equally incorporated in forbidden webpages. We can have high accuracy by ranking forbidden keywords of dictionary and take into account the weighted sum and frequency of forbidden keywords instead of number and frequency of forbidden keywords. We have selected a number of words and have normalized their ranking with respect to their

minimum and maximum values and mapped them into the (1, 40) interval. Then we represent web pages as vectors of 18 features using the ranking dictionary of forbidden keywords. Textual and profile features that we used to represent web pages are shown in Table 1.

Weighted sum and frequency of forbidden keywords in different parts of web pages are calculated by the following formula:

$$\begin{aligned} Weighted\ Sum &= \sum Rank(t_i) \times n(t_i) \\ Weighted\ Frequancy &= \frac{\sum Rank(t_i) \times n(t_i)}{\sum Rank(t_i) \times n(t_i) + m} \end{aligned} \quad (3)$$

Where $n(t_i)$ is the number of times t_i occurs in the target part of the web page and m is the number of non-forbidden words in target part of the web page.

Table 1. Selected features for web page representation

Features	Description
nw-page	Weighted sum of forbidden words that occur in the page
wfw-page	Weighted frequency of forbidden words that occur in the page
nw-body	Weighted sum of forbidden words that occur in the body
wfw-body	Weighted frequency of forbidden words that occur in the body
nw-title	Weighted sum of forbidden words that occur in the title
wfw-title	Weighted frequency of forbidden words that occur in the title
n-URL	Number of URLs in the page
nw-URL	Weighted sum of forbidden words that occur in the URLs
n-link	Number of links in the page
nw-link	Weighted sum of forbidden words that occur in the links
wfw-link	Weighted frequency of forbidden words that occur in the links
n-image	Number of images in the page
nw-image	Weighted sum of forbidden words that occur in the images
Wfw-image	Weighted frequency of forbidden words that occur in the image
nw-src	Weighted sum of forbidden words that occur in the attribute src of the img tag
nw-alt	Weighted sum of forbidden words that occur in the attribute alt of the img tag
nw-meta	Weighted sum of forbidden words that occur in the meta
wfw-meta	Weighted frequency of forbidden words that occur in the meta

For example, the following text is content part of tag Meta of a forbidden page, words of text that are in forbidden words dictionary are specified in underlined form and rank of each words is given in the against table.

"*Brand* New! We have *reviewed* Shemale Sex Dates and it was *awesome*. Horny Shemale *Lovers Take* the *Free Tour* and see for yourself!"

Forbidden keywords	Rank
Shemale	2.98
Sex	40
Horny	10.42

Weighted sum and frequency of this text is calculated as follows.

$$Weighted\ Sum = 2.98\times 2 + 40\times 1 + 10.42\times 1 = 56.38$$

$$Weighted\ Frequancy = \frac{2.98\times 2 + 40\times 1 + 10.42\times 1}{2.98\times 2 + 40\times 1 + 10.42\times 1 + 7} = 0.8896$$

Words extracted from the text regardless of the forbidden words are defined in italics after removing stop words and equals to 7. Weighted sum and frequency for texts related to rest of the Web page were calculated as sample and a vector consisting of 18 attributes is formed for each web page.

6. Experimental Results

6.1 Dataset description

To evaluate the proposed method, 2643 random samples of ODP[1] links have been selected from allowed and illegal categories. Among them 1311 web pages belong to the allowed category and 1332 pages belong to the forbidden category. Among selected samples, 933 legal web pages and 918 forbidden web pages have been randomly chosen as training dataset. Moreover, 792 web pages have been selected in order to assess system accuracy and efficacy that include 393 legal webpages and 399 forbidden web pages.

6.2 Performance measure

Usually blocking and over-blocking rate are used for performance measurement in the filtering systems. Blocking rate measures the percentage of forbidden pages that the filtering system manages to block and over-blocking rate shows the rate of misclassified normal pages as forbidden pages. They are defined by the following equations:

$$\begin{aligned} BlockingRate &= \frac{TP}{TP + FP} \\ Over-BlockingRate &= \frac{FP}{TN + FN} \end{aligned} \quad (4)$$

Where *TP* is the number of test web pages correctly classified under forbidden web pages, *FP* is the number of test web pages incorrectly classified under forbidden web pages, *TN* is the number of test web pages correctly classified under normal web pages,

1. Open Directory Projects http://www.dmoz.org/.

and *FN* is the number of test web pages incorrectly classified under normal web pages. These definitions are shown in **Table 2**.

Table 2. The global contingency table

		Expert Judgment	
		Yes	**No**
Classifier	**Yes**	TP	FP
Judgments	**No**	FN	TN

Another commonly used measure in filtering systems is accuracy that is defined in the equation (5).

$$Accuracy = \frac{TP + TN}{N} \quad (5)$$

Where *N* is the total number of web pages.

7. Comparison Analysis

To evaluate the proposed method, after attaching training web pages to each other, words in the pages are extracted and after removing stop words, the remaining words were stemmed. Porter algorithm [16] is used for word stemming. The number of rooted words in the vocabulary was equal to 58090 after rooting the keywords. Forbidden keywords were selected using the method mentioned in the section 5.

In the next stage, the corresponding vectors of web pages were formed in three ways. In the first way (*TFIDF*), a web page was represented as a vector of words where the words are selected by CHI word selection method [5]. In the second way (RSF), a web page was represented as a vector of numbers and frequencies of forbidden keywords in different parts of web pages. In the third way, a web page was represented as we proposed (*RWSF*), which is introduced in the section 5. Different classifiers including Support Vector Machine, k-Nearest Neighbor, Artificial Neural Network and Decision Tree are used to evaluate all types of representations. In our experiments, all the classifiers were obtained from the framework Weka (Witten and Frank 1999).

We evaluate the performance of *TFIDF* method by varying the number of features from 100 to1000. The results of our experiments are shown in the **Figure 3**. As seen in the figure, SMO (a version of SVM implemented in Weka) has a high accuracy of 120 words, Neural Networks has a high accuracy of 160 words, and k-Nearest Neighbor has a high accuracy of 100 words using *TFIDF* method.

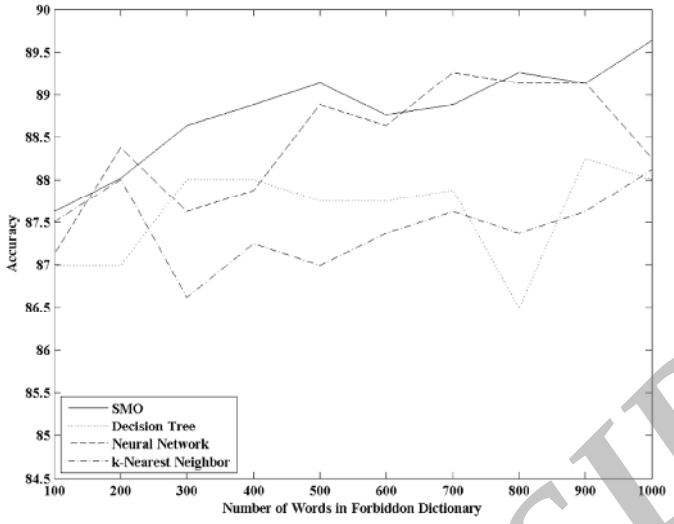

Figure 3. Comparison of the classifiers in the TFIDF approach

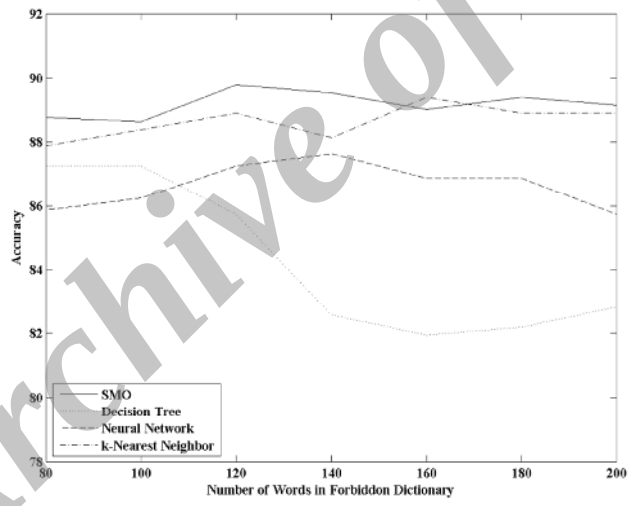

Figure 4. Comparison of the classifiers in the proposed approach

Figure 4 presents the comparison of the classifiers with number of different keywords in the dictionary. According to the results of experiments, the SMO classifier has a high accuracy of 89.6465 percent with 1000 keywords. The *kNN* classifier with k = 10 at best mood has accuracy of 88.1313 percent. Neural Network classifier has a high accuracy equal to 89.2677 percent with dictionary including 700 keywords. The Decision Tree classifier has the high accuracy of 88.257 percent with a dictionary that includes 800 forbidden keywords.

To compare our method with *TFIDF*, we selected the best result of the two methods in each classifier (has shown in *Table 3*) and calculated the percentage of increase or

decrease in the accuracy, training execution time, and memory usage for saving training data after preprocessing step by the following formula:

$$\frac{Result_{new} - Result_{old}}{Result_{old}} \times 100 \quad (6)$$

Table 3. Comparing different classifiers in each method of web page representation in the best way of accuracy

	TFIDF			*RWSF*		
	Accuracy	**Training Time (S)**	**Memory Usage (KB)**	**Accuracy**	**Training Time (S)**	**Memory Usage (KB)**
SMO	89.7727	0.787	579	89.6465	0.3467	169
DT	87.6263	2.102	665	88.257	0.2356	169
ANN	89.3939	811.4578	753	89.2677	14.8811	167
k-NN	87.2475	0	489	88.1313	0	169

The result of our experiments are shown in **Figure 5**. Although our approach does not have much effect on increasing the accuracy of the system comparing to the *tfidf* method, it is very effective in decreasing the training time and memory usage.

The evaluation results of comparing *RSF* with *RWSF* methods are shown in **Figure 6**. As shown, the use of ranking dictionary is mentioned by various classifiers to evaluate its effect in achieving higher accuracy.

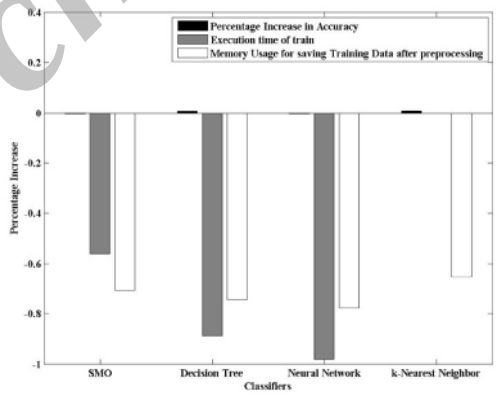

Figure 5. Percentage increase in accuracy, training time and memory usage for saving data in the proposed method compared to TFIDF using different classifiers.

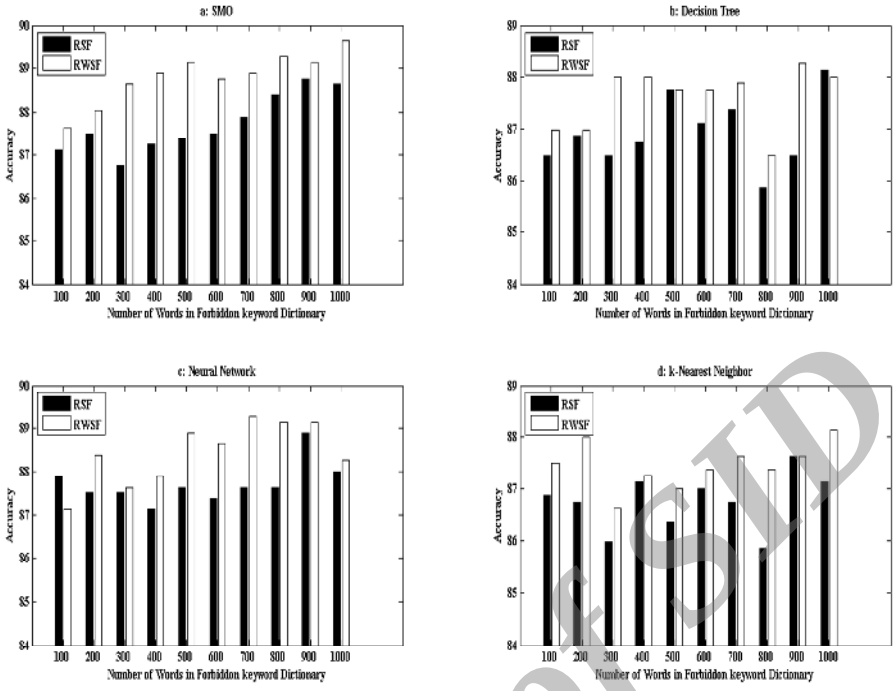

Figure 6. Comparing the accuracy of filtering between use of forbidden keywords and weighted forbidden keyword with different classifiers

8. Conclusion

Invention of Web has made it the main place to publish any kind of information. There are various types of information including a large number of inappropriate web pages, which are useless for some groups of people. Some organizations need to filter access of their community to erotic pages. Recently, some intelligent techniques based on classification methods of texts were proposed to prevent users to access forbidden web pages. In this paper, we have proposed a new intelligent automatic way to forbidden keywords dictionary formation. We presented webpages using various features obtained based on forbidden keywords dictionary. Then we assessed our filtering system using different classification techniques such as Decision Tree, Support Vector Machine, k-Nearest Neighbor and Artificial Neural Network. The results of all classifications show that the proposed method has high efficiency.

In this paper, we filter web pages only using textual information of web pages. The accuracy needs to be further improved by analyzing the various multimedia in the web pages, including audios, images and videos.

9. References

[1] Du R, Safavi-Naini R, Susilo W. Web Filtering Using Text Classification. In Networks 2003 ICON2003 The 11th IEEE International Conference on; 28 Sept.-1 Oct. 2003; pp. 325-330.

[2] He Z, Li X, Hu W. A boosted semi-supervised learning framework for web page filtering. In Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics (SMC'09); 11-14 Oct. 2009; IEEE Press, Piscataway, NJ, USA, pp. 2133-2136.

[3] Lee PY, Hui SC, Fong ACM. Neural Networks for Web Content Filtering. IEEE Intelligent Systems 2002; 17: 48-57.

[4] Guermazi R, Hammami M, Hamadou AB. Combinition Classifiers for Web Violent Content Detection and Filtering. ICCS '07 Proceedings of the 7th international conference on Computational; 2007, pp. 773-780.

[5] Baharudin B, Lee LH, Khan K. A Review of Machine Learning Algorithms for Text Documents Classification. Journal of Advances in Information Technology 2010; 1(1): 4-20.

[6] Harish B, Guru D, Manjunath S. Representation and Classification of Text Documents: A Brief Review IJCA,Special Issue on RTIPPR; 2010, 2:110–119.

[7] Mitchell TM. Machine Learning. Annual Review of Computer Science 1997; 4: 417-433.

[8] Mitra V, Wang CJ, Banerjee S. Text classification: A least square support vector machine approach, Applied Soft Computing Journal. 2007, 7 (3), pp. 908-914.

[9] Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys. 2001; 34(1): 1-47.

[10] Wu O, Hu W. Web Sensitive Text Filtering by Combining Semantics and Statistics. IEEE International Conference on Natural Language Processing and Knowledge Engineering. 30 Oct.-1 Nov. 2005, IEEE NLP-KE '05, pp. 663-667.

[11] Chen Z, Wu O, Zhu W, Hu W. A Novel Web Page Filtering System by Combining Texts and Images. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI '06). 18-22 Dec. 2006. IEEE Computer Society, Washington, DC, USA, pp. 732-735.

[12] Ahmadi A, Fotouhi M, Khaleghi M. Intelligent classification of web pages using contextual and visual features. APPL SOFT COMPUT; 2011; 11(2): 1638-647.

[13] Wu Y, She K, Zhu W, Yue X, Luo H. A Web Text Filter Based on Rough Set Weighted Bayesian. In Proceedings of the 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC '09). IEEE Computer Society, Washington, DC, USA, pp. 241-245.

[14] Hammami M, Chahir Y, Chen L. Combining Text and Image Analysis in the Web Filtering System "Webguard". International Association for Development of the Information Society – IADIS. Novembre 2003, pp. 611-618.

[15] Guermazi R, Hammami M, Hamadou AB. Using a Semi-automatic Keyword 9 Dictionary for Improving Violent Web Site Filtering. 2007 Third International IEEE Conference on Signal Image Technologies and Internet Based System, 16-18 Dec. 2007, pp. 337-344.

[16] Porter M. An algorithm for suffix stripping. Automated Library and Information Systems, 1980; 14(3): 130-137.

[17] S. Ramasundaram and S.P. Victor; Algorithms for Text Categorization : A Comparative Study; World Applied Sciences Journal 22 (9): pp. 1232-1240, ISSN 1818-4952, 2013.

[18] Y. Zhao, Chapter 10 - Text Mining, In: Yangchang Zhao, Editor(s), R and Data Mining, Academic Press, 2013, Pages 105-122, R and Data Mining, ISBN 9780123969637, http://dx.doi.org/10.1016/B978-0-12-396963-7.00010-6.

[19] H. Ma, "Fast Blocking of Undesirable Web Pages on Client PC by Discriminating URL Using Neural Networks," Expert Systems With Applications (ESWA), vol. 34, no. 2, pp. 1533-1540, February 2008.