Combining Classifiers for Web Violent Content Detection and Filtering

Radhouane Guermazi¹, Mohamed Hammami², and Abdelmajid Ben Hamadou¹

¹ Miracl-Isims, Route Mharza Km 1 BP 1030 Sfax Tunisie

² Miracl-Fss, Route Sokra Km 3 BP 802, 3018 Sfax Tunisie

`rguermazi@laposte.net`,

`mohamed.hammami@fss.rnu.tn`,

`abdelmajid.benhamadou@isimsf.rnu.tn`

`http://www.isimsf.rnu.tn/`

Abstract. Keeping people away from litigious information becomes one of the most important research area in network information security. Indeed, Web filtering is used to prevent access to undesirable Web pages. In this paper we review some existing solutions, then we propose a violent Web content detection and filtering system called “WebAngels filter” which uses textual and structural analysis. “WebAngels filter” has the advantage of combining several data-mining algorithms for Web site classification. We discuss how the combination learning based methods can improve filtering performances. Our preliminary results show that it can detect and filter violent content effectively.

Keywords: Web classification and categorization, data-mining, Web textual and structural content, violent website filtering.

1 Introduction

The growth of the Web and the increasing number of documents electronically available has been paralleled by the emergence of harmful Web pages content such as pornography, violence, racism, etc. This emergence involved the necessity of providing filtering systems designed to secure the internet access. Most of them process mainly the adult content and focus on blocking pornography. Whereas, the other litigious characters, in particular the violent one, are marginalized. In this paper, we propose a textual and structural content-based analysis using several major-data mining algorithms for automatic violent website classification and filtering. We focus our attention on the combination of classifiers, and we demonstrate that it can be applied to improve the filtering efficiency of violent web pages. The remainder of this paper is organized as follows. We overview related work according to web filtering in section 2. In the next section, we present our approach. The extraction of features vector is described, and the experimentation of each used algorithm is studied. In the fourth section, we show the efficiency of combining data-mining techniques to improve the filtering accuracy rate. In section 5 we describe the architecture and the functionalities of the

prototype “WebAngels filter” as well as its results compared to the most known software on the market. Finally section 6 presents some concluding remarks and future work directions.

2 Web Filtering: Related Works

Several litigious website filtering approaches were proposed. Among these approaches, we can quote (1) the Platform for Internet Content Selection (PICS)¹, (2) the exclusion filtering approach, which allows all access except to sites belonging to a manually constructed black-list, (3) the inclusion filtering approach, which only allows access to sites belonging to a manually constructed white-list and (4) the automated content filtering approach, designed to classify and filter Web sites and URLs automatically in order to reflect the highly dynamic evolution of the Web. In this approach, we can enumerate two types: (a) Keyword blocking where a list of prohibited words is used to identify undesirable Web pages. (b) Intelligent content Web filtering which falls in the general problem of automatic website categorization and uses machine learning. At least, three categories of intelligent content Web filtering can be distinguished : (i) textual content Web filtering [1], (ii) structural content Web filtering [2], [3] and (iii) Visual content Web filtering [4]. Other Web filtering solutions are based on an analysis of textual, structural and visual contents, of a Web page [5].

Litigious Web pages classification and filtering are diversified. However, the majority of these works treat particularity adult character. We propose in the following section our approach for the violent sites classification.

3 Our Approach to Violent Web Filtering

Black-lists and white-lists are hard to generate and maintain. Also filtering based on naive keyword-matching can be easily circumvented by deliberate mis-spelling of keywords. We propose to build an automatic violent content detection solution based on a machine learning approach using a set of manually classified sites, in order to produce a prediction model, which makes it possible to know which URLs are suspect and which are not. To classify the sites into two classes, we are based on the KDD process for extracting useful knowledge from volumes data [6]. The general principle of the approach of classification is the following: Let S be the population of samples to be classified. To each sample s of S one can associate a particular attribute, namely, its class label C . C takes its value in the class of labels(0 for violent, 1 for nonviolent)

$$C : S \rightarrow \Gamma = \{Violent, nonViolent\}$$

$$s \in S \mapsto C(s) \in \Gamma$$

Our study consists in building a means to predict the attribute class of each website. To do it three major steps are necessary: (a) selecting and pre-processing

¹ <http://www.w3.org/PICS>

step which consists on selecting the features which best discriminate classes and extracting the features from the training data set; (b) data-mining step which looks for a synthetic and generalizable model by the use of various algorithms; (c) evaluation and validation step which consists on assessing the quality of the learned model on the training data set and on the test data set.

3.1 Data Preparation

In this stage we identify exploitable information and check their quality and their effectiveness in order to build a two-dimensional table from our training corpus. Each table row represents a web page and each column represent a feature. in the last column, we save the web page class (0 or 1).

Construction of Training Set. The data-mining process for violent website classification requires a representative training data set consisting on a significant set of manually classified websites. We choused diversified websites according to their content, treated languages and structure. Our training data set is composed of 700 sites from which 350 are violent.

Textual and Structural Contents Analysis. The selection of features used in a machine learning process is a key step which directly affects the performance of a classifier. Our study of the state of the art and manual collection of our test data set helped us a lot to gain intuition on violent website characteristics and to understand discriminating features between violent web pages and inoffensive ones. These intuition and understanding suggested us to select both textual and structural content-based features for better discrimination purpose. We used in the calculation of these features a manually collected violent words vocabulary (or dictionary). The features used to classify the Web pages are `n_v_words_page` (total number of violent words in the current Web page), `%v_words_page` (frequency of the violent words of the page), `n_v_words_url` (total number of violent words in the URL), `%v_words_url` (frequency of violent words in the URL), `n_v_words_title` (total number of violent words which appear in the title tag), `%v_words_title` (frequency of violent words which appear in the title tag), `n_v_words_body` (number of violent words which appear in the body tag), `%v_words_body` (frequency of violent words which appear in the body tag), `n_v_words_meta` (number of violent words which appear in the meta tag), `%v_words_meta` (frequency of violent words which appear in the meta tag), `n_links` (total number of links), `n_links_v` (total number of links containing at least one violent word), `%links_v` (frequency of links containing at least a violent word), `n_img` (total number of images), `n_img_v` (total number of images whose name contains at least a violent word), `n_v_img_src` (total number of violent words in the attribute src of the img tag), `n_v_img_alt` (total number of violent words in the attribute alt of the img tag) and `%img_v` (frequency of the images containing a violent word). We took into serious consideration the construction of this dictionary and, unlike a lot of commercial filtering products, we built a multilingual dictionary.

3.2 Supervised Learning

In the literature, there are several techniques of supervised learning, each having its advantages and disadvantages. But the most important criterion for comparing classification techniques remains the classification accuracy rate. We have also considered another criterion which seems to us very important: the comprehensibility of the learned model. In our approach, we used the graphs of decision tree [7]. In a decision tree, we begin with a learning data set and look for the particular attribute which will produce the best partitioning by maximizing the variation of uncertainty $\mathfrak{S}_\lambda$ between the current partition and the previous one. As $I_\lambda(S_i)$ is a measure of entropy for partition S_i and $I_\lambda(S_{i+1})$ is the measure of entropy of the following partition S_{i+1} . The variation of uncertainty is:

$$\mathfrak{S}_\lambda = I_\lambda(S_i) - I_\lambda(S_{i+1}) \quad (1)$$

For $I_\lambda(S_i)$ we can make use the quadratic entropy (2) or Shannon entropy (3) according to the selected method:

$$I_\lambda(S_i) = \sum_{j=1}^K \frac{n_j}{n} \left(- \sum_{i=1}^m \frac{n_{ij} + \lambda}{n_i + m\lambda} \left(1 - \frac{n_{ij} + \lambda}{n_i + m\lambda} \right) \right) \quad (2)$$

$$I_\lambda(S_i) = \sum_{j=1}^K \frac{n_j}{n} \left(- \sum_{i=1}^m \frac{n_{ij} + \lambda}{n_i + m\lambda} \log_2 \frac{n_{ij} + \lambda}{n_i + m\lambda} \right) \quad (3)$$

Where n_{ij} is the number of elements of class i at the node S_j with $i \in \{c_1, c_2\}$; n_i is the total number of elements of the class i , $n_i = \sum_{j=1}^K n_{nj}$; n_j is the number of elements of the node S_j , $n_j = \sum_{i=1}^2 n_{ij}$; n is the total number of elements, $n = \sum_{i=1}^2 n_i$; $m = 2$ is the number of classes (c_1, c_2). λ is a variable controlling effectiveness of graph construction, it penalizes the nodes with insufficient effective.

In our work, four data-mining algorithms including ID3 [8], C4.5 [9], IMPROVED C4.5 [10], SIPINA [11] has been experimented.

3.3 Experiments

We present two series of experiments: in the first, we experimented with four data-mining techniques on the training data set and validate the quality of the learned models using random error rate techniques. Three measures have been used: *global error rate*, *a priori error rate* and *a posteriori error rate*. As *global error rate* is the complement of classification accuracy rate, while *a priori error rate* (respectively, *a posteriori error rate*) is the complement of the classical *recall rate* (respectively, *precision rate*). Figure 1 presents the obtained results by the four algorithms on the training data set. It shows that the best algorithm is SIPINA. These results can be explained by the fact that this algorithm tries to reduce the disadvantages of the arborescent methods (C4.5, Improved C4.5, ID3) on the one hand by the introduction of fusion operation and on the other

hand by the use of a measurement sensitive to effective. A good decision tree obtained by a data mining algorithm from the learning data set should not only produce classification performance on data already seen but also on unseen data as well. In the second experiments, in order to ensure the performance stability of our learned models from the learning data, we thus also tested the learned models on our test data set consisting of 300 Web sites from which 150 are violent. As we can see in the figure 2, all of four data-mining algorithms echoed

Fig. 1. Experimental results by the four algorithms on training data set

very similar performance. The strategy we choose consists on using the model which ensures a better compromise between the *recall* and the *precision*. For instance, when SIPINA displayed the best performance (taking into account both obtained results), we choose it as the best algorithm.

Fig. 2. Experimental results by the four algorithms on test data set

4 Classifier Combination

As shown figure 1 and figure 2, the four data-mining algorithms (ID3, C4.5, SIPINA and Improved C4.5) displayed different performances on the various error rate. Although SIPINA gives the best results, Combining algorithms may provide better results. In our work, we experiment two combining classifier methods:

1. Majority voting: The page is considered violent if the majority of classifiers (SIPINA, C4.5, Improved C4.5, ID3) considered it as violent and in case of conflict votes, we consider the decision of SIPINA.
2. Scores Combination: we thus affect a weight associated to the score classification decision data-mining algorithm according to the formula (4):

$$Score_{page} = \sum_{i=1}^4 \alpha_i S_i \quad (4)$$

Where S_i presents the score to be a violent web page according to i -th data-mining algorithm ($i = 1 \dots 4$). α_i presents the believe value we accord to the algorithm i . The score is calculated for each algorithm as follows: having the decision tree, we seek for each page the node that corresponds to the values of its characteristics vector. Once found the score is calculated as the ratio of pages judged as violent between S_i and S_{i+1} . The page is judged as violent if its score is greater than 0,5.

Figure 3 presents a comparison between the SIPINA predict model, the majority voting model and scores combination model on the test data set. The obtained

Fig. 3. Comparaison SIPINA model with combination approaches on test data set

results confirm the well interest to use the combining classifiers. Indeed, both majority voting and scores combination approaches provide better results than SIPINA. One clearly sees the reduction in the global error rate of 0.23 for SIPINA predict model to 0.19 by the majority voting and to 0.16 by the scores combination approach.

5 “WebAngels Filter” Tool

5.1 Architecture

The classifiers combination scores is used to propose an efficient violent Web filtering named “WebAngels filter” (figure 4). When a URL is launched, it carries out the following actions: (1) Block the Web page if the site is recorded on

the “black list” else load its HTML code source, (2) Analyze the textual and structural information code, (3) Decide if the access to the Web page is denied or allowed, (4) Update the black list, (5) Update the history of navigation, (6) Block the page if it is judged as violent.

Fig. 4. “WebAngels filter” Architecture

Fig. 5. Classification accuracy rate of “WebAngels filter” compared to some products

5.2 Comparison with Others Products

To evaluate our violent Web filtering tool, we carry out an experiment which compares, on the test data set , “WebAngels filter” with four litigious content detection and filtering systems, knowing, control kids² , content protect³ , k9-webprotection⁴ and Cyber patrol⁵. these tools were been parametric to filter violent Web Content.

² <http://www.controlkids.com/fr>
³ <http://www.contentwatch.com>
⁴ <http://www.k9webprotection.com>
⁵ <http://www.cyberpatrol.com>

Figure 5 further highlights the performance of “WebAngels filter” compared to other violent content detection and filtering systems.

6 Conclusion

In this paper, we studied and highlighted a combination of several violent web classifiers based on a textual and structural content-based analysis for improving Web filtering through our tool “WebAngels filter”. Our experimental evaluation shows the effectiveness of our approach in such systems. However, many future work directions can be considered. Actually, the elaboration of an indicative keywords vocabulary which play a big role in the improvement of the tool performances was manual, and a laborious automatic elaboration of this vocabulary can be one of the directions of our future work. Also, we must think to integrate the treatment of the visual content in the predict models in order to cure the difficulties in classifying violent Web sites which includes only images.

References

1. Caulkins, J.-P., Ding, W., Duncan, G., Krishnan, R., Nyberg, E.: A Method for Managing Access to Web Pages: Filtering by Statistical Classification (FSC) Applied to Text. *Decision Support Systems* (To appear)
2. Lee, P.Y., Hui, S.C., Fong, A.C.M.: Neural Networks for Web Content Filtering. *IEEE Intelligent Systems* (2002) 48–57
3. Ho, W.H., Watters, P.A.: Statistical and structural approaches to filtering Internet pornography. *IEEE Conference on Systems, Man and Cybernetics* (2004) 4792–4798
4. Arentz, W.-A., Olstad, B.: Classifying offensive sites based on image content. *Computer Vision and Image Understanding*. Vol. 94. (2004) 295–310
5. Hammami, M., Chahir, Y., Chen, L.: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis. *IEEE Transactions on Knowledge and Data Engineering* (2006) 272–284
6. Fayyad, U., PIATETSKY-SHAPIRO, G., SMYTH, P.: The KDD process for extracting useful knowledge from volumes data. *Communication of the ACM* (1996) 27–34
7. Zighed, D.A., Rakotomalala, R.: *Graphes d’Induction - Apprentissage et Data Mining*. Hermes (2000)
8. Quinlan, J. R.: *Induction of Decision Trees*. Machine Learning. Induction of Decision Trees (1986) 81–106
9. Quinlan, J. R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann (1993)
10. Rakotomalala, R., Lallich, S.: Handling noise with generalized entropy of type beta in induction graphs algorithm. *International Conference on Computer Science and Informatics* (1998) 25–27
11. Zighed, D.A., Rakotomalala, R.: A method for non arborescent induction graphs. Technical report (1996). Laboratory ERIC, University of Lyon 2