Adult Content Filtering through Compression-based Text Classification

Igor Santos, Patxi Galán-García, Aitor Santamaría-Ibirika, Borja Alonso-Isla, Iker Alabau-Sarasola, and Pablo G. Bringas

S^{3}Lab, University of Deusto
Avenida de las Universidades 24, 48007 Bilbao, Spain
{isantos, patxigg, a.santamaria, borjaalonso, ialabau, pablo.garcia.bringas}@deusto.es

Abstract. Internet is a powerful source of information. However, some of the information that is available in the Internet, cannot be shown to every type of public. For instance, pornography is not desirable to be shown to children. To this end, several algorithms for text filtering have been proposed that employ a Vector Space Model representation of the webpages. Nevertheless, these type of filters can be surpassed using different attacks. In this paper, we present the first adult content filtering tool that employs compression algorithms to represent data that is resilient to these attacks. We show that this approach enhances the results of classic VSM models.

Keywords: Content filtering, text-processing, compression-based text classification

1 Introduction

Sometimes, the information available in the Internet, cannot be shown or is not appropriate to show to every type of public. For instance, pornography is not desirable to be shown to children. In fact, sometimes the content is illegal (or barely legal) such as child pornography, violence or racism.

The approach that both the academia and the industry has followed in order to filter these not appropriate contents is web filtering. These filters are broadly used in workplaces, schools or public institutions [1]. Information filtering itself can be viewed as a text categorisation problem (or image categorisation problem if images are used). In particular, in this work, we focus on adult site filtering. An important amount of work have been performed to filter these contents using the image information [2–6].

Regarding the use of textual information for adult website filtering, several works have been developed [7, 6, 8, 9]. These approaches model sites using the *Vector Space Model* (VSM) [10], an algebraic approach for *Information Filtering* (IF), *Information Retrieval* (IR), indexing and ranking. This model represents natural language documents in a mathematical manner through vectors in a multidimensional space.

However, this method has its shortcomings. For instance, in spam filtering, which is a type of text filtering similar to adult website filtering, *Good Word Attack*, a method that modifies the term statistics by appending a set of words that are characteristic of legitimate e-mails, or *tokenisation*, that works against the feature selection of the message by splitting or modifying key message features rendering the term-representation no longer feasible [11], have been applied by spammers.

Against this background, we propose the first compression-based text filtering approach to filter adult websites. Dynamic Markov compression (DMC) [12] has been applied for spam filtering [13], with good results. We have adapted this approach for adult content classification and filtering. In particular, we have used Cross-Entropy and Minimum Description Length, which model a class as an information source, and consider the training data for each class a sample of the type of data generated by the source.

In particular, our main findings are:

- We present the first compression-based adult content filtering method.
- We show how to adopt DMC for adult content filtering tasks.
- We validate our method and show that it can improve the results of the classic VSM model in adult content filtering.

The remainder of this paper is organised as follows. Section 2 describes the DMC approach. Section 3 describes the performed empirical validation. Finally, Section 4 concludes and outline the avenues of future work.

2 Dynamic Markov Chain Compression for Content Filtering

The compression algorithm dynamic Markov compression (DMC) [12] models information with a finite state machine. Associations are built between every possible symbol in the source alphabet and the probability distribution over those symbols. This probability distribution is used to predict the next binary digit. The DMC method starts in a already defined state, changing the state when new bits are read from the entry. The frequency of the transitions to either a 0 or a 1 are summed when a new symbol arrives. The structure can be also be updated using a state cloning method.

DMC has been previously used in spam filtering tasks [13], with good results. We have used a similar approach used in spam filtering for text classification using compression models. In particular, we have used Cross-Entropy and Minimum Description Length, which model a class as an information source, and consider the training data for each class a sample of the type of data generated by the source. In this way, our text analysis system tries to accurately classify web pages into 2 main categories: *adult* or *not adult*, therefore, we are training two different information sources *adult* A or *not adult* $\neg A$.

In order to generate the models, we used the information found within the web page. To represent the web page, we started by parsing the HTML of the

page so only the text remains. Using these parsed websites, we generate the two information sources *adult* A or *not adult* $\neg A$.

To classify the new webpages, Cross-Entropy and MDL were used:

– **Cross Entropy.** Following the classic definition, the entropy $H(X)$ of a source X measures the amount of information used by a symbol of the source alphabet:

$$H(X) = \lim_{n \to +\infty} -\frac{1}{n} \sum P(s_1^n) \cdot \log_2 P(s_1^n) \tag{1}$$

The cross-entropy between an information source X and a compression model M is defined as:

$$H(X, M) = \lim_{n \to +\infty} -\frac{1}{n} \sum P(s_1^n) \cdot \log_2 P_M(s_1^n) \tag{2}$$

For a given webpage w, the webpage cross-entropy is the average number of bits per symbol required to encode the document using the model M:

$$H(X, M, w) = -\frac{1}{n} \log_2 P_M(w) = -\frac{1}{n} \sum_{i=1}^{|w|} \log_2(P_M(S_i|S_1^{i-1})) \tag{3}$$

The classification criteria follow the expectation that a model which achieves a low cross-entropy on a given webpage approximates the information source. In this way, to assign the probability of the webpage w to belong to a given class c of a set of classes C is computed as:

$$P(c) = \frac{1}{H(X, M_c, w)^{-}1 \cdot \sum_{c_i \in C} \frac{1}{H(X, M_{c_i}, w)}} \tag{4}$$

– **Minimum Description Length.** Minimum Description Length (MDL) [14] criteria states that the best compression model is the one with the shortest description of the model and the data i.e., the one that compresses best a given document.
The difference with minimum cross- entropy is that the model adapts itself with the test webpage, while the page is being classified.

$$MDL(X, M, w) = -\frac{1}{n} \log_2 P'_M(w) = -\frac{1}{n} \sum_{i=1}^{|w|} \log_2(P'_M(S_i|S_1^{i-1})) \tag{5}$$

where $P'_M(w)$ means that the model is updated with the information found in the webpage w. The classification criteria is the same has the one used with cross-entropy.

$$P(c) = \frac{1}{MDL(X, M_c, w)^{-}1 \cdot \sum_{c_i \in C} \frac{1}{MDL(X, M_{c_i}, w)}} \tag{6}$$

3 Empirical Validation

To validate our approach, we downloaded 4500 web pages of both adult content and other content such as technology, sports and so on. The dataset contained 2000 adult websites and 2500 not adult websites. The collection was conformed by gathering different adult websites and sub-pages within them. A similar approach was used to conform the not adult data.

Once we parse the HTML code from all the web pages, we conducted the following methodology:

- **Cross Validation.** We have performed a *K-fold cross validation* with k=10. In this way, our dataset is 10 times split into 10 different sets of learning (90% of the total dataset) and testing (10% of the total data).
- **Learning the model.** For each fold we have performed the learning phase of the DMC. In this way, we added to the DMC model every website contained in each training dataset, adapting the compression model with each website.
- **Testing the model.** For each fold, we have used different criteria to select the class: Cross-Entropy and MDL. In this way, we measured the True Positive Ratio (TPR), i.e., the number of adult websites correctly detected, divided by the total number of adult webs:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

where TP is the number of adult websites correctly classified (true positives) and FN is the number of adult websites misclassified as not adult sites(false negatives).
We also measured the False Positive Ratio (FPR), i.e., the number of not adult sites misclassified as adult divided by the total number of not adult sites:

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

where FP is the number of not adult websites incorrectly detected as adult and TN is the number of not adult sites correctly classified.
Furthermore, we measured the accuracy, i.e., the total number of the classifier's hits divided by the number of instances in the whole dataset:

$$Accuracy(\%) = \frac{TP + TN}{TP + FP + TP + TN} \cdot 100 \tag{9}$$

Besides, we measured the Area Under the ROC Curve (AUC) that establishes the relation between false negatives and false positives [15]. The ROC curve is obtained by plotting the TPR against the FPR.
- **Comparison other models:** In order to validate our results, we compare the results obtained by the DMC classifiers (with Cross-Entropy and MDL criteria) with the ones obtained with a classic VSM model.
To represent the web page, we start by parsing the HTML of the page so only the text remains. Then, we remove stop-words [16], which are words devoid of content (e.g., 'a','the','is'). These words do not provide any semantic

Table 1. Results of DMC compared with Bayesian classifiers (%).

Classifier	Accuracy	TPR	FPR	AUC
DMC Cross Entropy	99.9778	100.0000	0.0004	1.0000
DMC MDL	99.2889	98.5000	0.0004	1.0000
Naïve Bayes	99.1556	98.9000	0.0060	0.9910
Bayesian Network: K2	99.5111	98.9000	0.0000	0.9970
Bayesian Network: TAN	99.5333	99.0000	0.0000	0.9900

information and add noise to the model [17]. We used the *Term Frequency – Inverse Document Frequency* (TF–IDF) [17] weighting schema, where the weight of the i^{th} term in the j^th document is:

$$weight(i,j) = tf_{i,j} \cdot idf_i \tag{10}$$

where *term frequency* is:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{11}$$

where $n_{i,j}$ is the number of times the term $t_{i,j}$ appears in a document p, and $\sum_k n_{k,j}$ is the total number of terms in the document p.
The inverse term frequency idf_i is defined as:

$$idf_i = \frac{|\mathcal{P}|}{|\mathcal{P} : t_i \in p|} \tag{12}$$

where $|\mathcal{P}|$ is the total number of documents and $|\mathcal{P} : t_i \in p|$ is the number of documents containing the term t_i.
Next, we perform a stemming step [18]. Stemming is the process for reducing inflected words to their stem e.g., 'fishing' to 'fish'. To this end, we used the StringToWord filter in a filtered classifier in the well-known machine-learning tool WEKA [19]. Using this bag of words model, we have trained several Bayesian classifiers: K2 [20] and Tree Augmented Naïve (TAN) [21]. We also performed experiments with a Naïve Bayes classifier [22]. The DMC classifier was implemented by ourselves.

Table 1 shows the obtained results. The DMC classifier using the cross entropy criteria obtained the best results, improving the results of Bayesian classifiers. Indeed, this classifier only failed one of the instances: a false positive. The results with MDL were also high but not as high as when using cross-entropy, indicating that the update of the compression model with the test webpage does not improve the classification phase.

Even though that the classic Bayesian classifiers obtain a high accuracy rate, there may be several limitations due to the representation of webpages. As happens in spam, most of the filtering techniques are based on the frequencies with

which terms appear within messages, and we can modify the webpage to evade such filters.

For example, Good Word Attack[23] is a method that modifies the term statistics by appending a set of words that are characteristic of not adult pages, thereby bypass filters. Another attack, known as tokenisation, works against the feature selection of the message by splitting or modifying key message features, which renders the term-representation as no longer feasible [11]. All of these attacks can be avoid by the use of this compression-based methods, because it.

4 Discussion and Conclusions

Internet is a powerful channel for information distribution. Nevertheless, sometimes not all of the information is desirable to be shown to every type of public. Hence, web filtering is an important research area in order to protect users from not desirable content. One of the possible contents to filter is adult content.

In this research, our main contribution is the first adult content filter that is based in compression techniques for text filtering. In particular, we used DMC as text classifier, and we showed that this approach, enhances the classification results of VSM-based classifiers. Nevertheless, this approach also presents some limitations that should be studied in further work.

There is a problem derived from IR and *Natural Language Processing* (NLP) when dealing with text filtering: *Word Sense Disambiguation* (WSD). An attacker may evade our filter by explicitly exchanging the key words of the mail with other polyseme terms and thus avoid detection. In this way, WSD is considered necessary in order to accomplish most natural language processing tasks [24]. Therefore, we propose the study of different WSD techniques (a survey of different WSD techniques can be found in [25]) capable of providing a more semantics-aware filtering system. However, integrating a disambiguation method with a compression-based text-filtering tool is not feasible. Therefore, in the future, we will adopt a WSD-based method for the classic representation of websites VSM and we keep the compression based method, combining both results into a final categorisation result.

Besides, in our experiments, we used a dataset that is very small in comparison to the real-world size. As the dataset size grows, the issue of scalability becomes a concern. This problem produces excessive storage requirements, increases time complexity and impairs the general accuracy of the models [26]. To reduce disproportionate storage and time costs, it is necessary to reduce the size of the original training set [27]. To solve this issue, data reduction is normally considered an appropriate preprocessing optimisation technique [28, 29]. This type of techniques have many potential advantages such as reducing measurement, storage and transmission; decreasing training and testing times; confronting the problem of dimensionality to improve prediction performance in terms of speed, accuracy and simplicity; and facilitating data visualisation and understanding [30, 31]. Data reduction can be implemented in two ways. Instance selection (IS) seeks to reduce the number of evidences (i.e., number of rows) in

the training set by selecting the most relevant instances or by re-sampling new ones [32]. Feature selection (FS) decreases the number of attributes or features (i.e., columns) in the training set [33].

Future versions of this text filtering tool will be oriented in two main ways. First, we would like to deal with the semantics awareness of adult-content filtering including these capabilities in our filter. Second, we will enhance the requirements of labelling, in order to improve efficiency. Third, we will compare more compression methods.

References

1. Gómez Hidalgo, J., Sanz, E., García, F., Rodríguez, M.: Web Content Filtering. Advances in Computers **76** (2009) 257–306
2. Duan, L., Cui, G., Gao, W., Zhang, H.: Adult image detection method base-on skin color model and support vector machine. In: Asian Conference on Computer Vision. (2002) 797–800
3. Zheng, H., Daoudi, M., Jedynak, B.: Blocking adult images based on statistical skin detection. Electronic Letters on Computer Vision and Image Analysis **4**(2) (2004) 1–14
4. Lee, J., Kuo, Y., Chung, P., Chen, E., et al.: Naked image detection based on adaptive and extensible skin color model. Pattern Recognition **40**(8) (2007) 2261–2270
5. Choi, B., Chung, B., Ryou, J.: Adult Image Detection Using Bayesian Decision Rule Weighted by SVM Probability. In: 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, IEEE (2009) 659–662
6. : Poesia filter `http://www.poesia-filter.org/`.
7. Du, R., Safavi-Naini, R., Susilo, W.: Web filtering using text classification. In: Networks, 2003. ICON2003. The 11th IEEE International Conference on, IEEE (2003) 325–330
8. Kim, Y., Nam, T.: An efficient text filter for adult web documents. In: Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference. Volume 1., IEEE (2006) 3–pp
9. Ho, W., Watters, P.: Statistical and structural approaches to filtering internet pornography. In: Systems, Man and Cybernetics, 2004 IEEE International Conference on. Volume 5., IEEE (2004) 4792–4798
10. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
11. Wittel, G., Wu, S.: On attacking statistical spam filters. In: Proceedings of the 1^{st} Conference on Email and Anti-Spam (CEAS). (2004)
12. Cormack, G., Horspool, R.: Data compression using dynamic markov modelling. The Computer Journal **30**(6) (1987) 541
13. Bratko, A., Filipič, B., Cormack, G., Lynam, T., Zupan, B.: Spam filtering using statistical data compression models. The Journal of Machine Learning Research **7** (2006) 2673–2698
14. Rissanen, J.: Modeling by shortest data description. Automatica **14**(5) (1978) 465–471

15. Singh, Y., Kaur, A., Malhotra, R.: Comparative analysis of regression and machine learning methods for predicting fault proneness models. Int. J. Comput. Appl. Technol. **35** (June 2009) 183–193
16. Wilbur, W., Sirotkin, K.: The automatic identification of stop words. Journal of information science **18**(1) (1992) 45–55
17. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-Hill New York (1983)
18. Lovins, J.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics **11**(1) (1968) 22–31
19. Garner, S.: Weka: The Waikato environment for knowledge analysis. In: Proceedings of the New Zealand Computer Science Research Students Conference. (1995) 57–64
20. Cooper, G.F., Herskovits, E.: A bayesian method for constructing bayesian belief networks from databases. In: Proceedings of the 7^{th} conference on Uncertainty in artificial intelligence. (1991)
21. Geiger, D., Goldszmidt, M., Provan, G., Langley, P., Smyth, P.: Bayesian network classifiers. In: Machine Learning. (1997) 131–163
22. Lewis, D.: Naive (Bayes) at forty: The independence assumption in information retrieval. Lecture Notes in Computer Science **1398** (1998) 4–18
23. Dietterich, T., Lathrop, R., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence **89**(1-2) (1997) 31–71
24. Ide, N., Véronis, J.: Introduction to the special issue on word sense disambiguation: the state of the art. Computational linguistics **24**(1) (1998) 2–40
25. Navigli, R.: Word sense disambiguation: a survey. ACM Computing Surveys (CSUR) **41**(2) (2009) 10
26. Cano, J., Herrera, F., Lozano, M.: On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining. Applied Soft Computing Journal **6**(3) (2006) 323–332
27. Czarnowski, I., Jedrzejowicz, P.: Instance reduction approach to machine learning and multi-database mining. In: Proceedings of the Scientific Session organized during XXI Fall Meeting of the Polish Information Processing Society, Informatica, ANNALES Universitatis Mariae Curie-Skłodowska, Lublin. (2006) 60–71
28. Pyle, D.: Data preparation for data mining. Morgan Kaufmann (1999)
29. Tsang, E., Yeung, D., Wang, X.: OFFSS: optimal fuzzy-valued feature subset selection. IEEE transactions on fuzzy systems **11**(2) (2003) 202–213
30. Torkkola, K.: Feature extraction by non parametric mutual information maximization. The Journal of Machine Learning Research **3** (2003) 1415–1438
31. Dash, M., Liu, H.: Consistency-based search in feature selection. Artificial Intelligence **151**(1-2) (2003) 155–176
32. Liu, H., Motoda, H.: Instance selection and construction for data mining. Kluwer Academic Pub (2001)
33. Liu, H., Motoda, H.: Computational methods of feature selection. Chapman & Hall/CRC (2008)