A Novel Web Page Filtering System by Combining Texts and Images

Zhouyao Chen , Ou Wu, Mingliang Zhu and Weiming Hu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
No.95 East of Zhong Guan Cun Road, Haidian District, Beijing, 100080, Mailbox 2728
Email: {zychen82, wmhu}@nlpr.ia.ac.cn

Abstract

With the rapid development of the Internet, people benefit much from the sharing of information. Meanwhile, the WWW era is a double-edged sword which spreads harmful and erotic content widely. In this paper, a new statistical approach has been exploited by combining the results of two or more different classification methods using our filtering system. We first briefly introduce the classification of discrete texts, continuous texts and images separately, and then describe the specific way we have been exploring to merge the text and image classification result. Also there is a section illustrating our system framework. Finally we assess our method by demonstrating the experimental results and comparing it to some common-used filtering methods.

1. Introduction

The Internet, as a global data center, facilitates people all over the world with various services to share and exchange their information. Being efficient, versatile and powerful, it has created new possibilities for communication, publication, and instruction. However, some people have taken advantage of these benefits to spread harmful contents. It has been a long time since researchers began to deal with filtering pornography web pages. The overall trend of solutions to the filtering is from manual to automatic in purpose of adapting itself to the more dynamic web environment. In this literature the filtering methods can be categorized into three aspects as below:

Mechanical filtering: The process of this method is building two lists of either URLs or IP addresses. One is called Blacklist which consists of web sites that are forbidden to access, whereas Whitelist is made up of those which are allowed to. When new request of accessing new web page comes, the judgment is made by the matching result. [1].

Text-based filtering: It works mainly on sensitive texts which consist of specific words or phrases such as "sex", "adult" which are the most remarkable features to identify erotic web pages from normal ones. After extracting key words from texts, many classification methods can be applied.

Image-based filtering: with the new development in the image processing and pattern recognition fields (Automatic detection of human nudes by Forsyth [2]). More researchers focused on using the results of pornography images detection as a novel way to filter sensitive web pages [3].

The shortcomings of these methods are obvious: all of them are supposed to work with specific web pages well, but can not ensure the robustness in other kinds of web pages. A dynamic internet environment contains different forms of web elements, so the solution is not to abandon these traditional filtering methods, in other words, we can make use of them; combine them together according to both textual and visual clues. Hammami et al [4] have developed "WebGuard" system, which extends adult content detection accuracy by usage of textual, structural and visual based features. The deficiency is that it simply serializes the procedure of text-based filtering and image-based filtering.

In this paper, our core invention is applying divide and conquer strategy and proposing a fusion algorithm to combine the results obtained from text as well as image classification. The remainder of this paper is organized as follows. Section 2 briefly introduces the categorization of three typical web page forms and three independent classifiers. Section 3 summarizes the major steps of the proposed merging algorithm with the statistical reasoning. The architecture of the whole filtering system is described in Section 4. Experimental results are given and discussed in Section 5, prior to the conclusion in Section 6.

2. Divide and Conquer

Web page classification is much more complex than pure text classification due to a large variety of multimedia embedded in web pages. To handle these complicated problems, we have observed that there are three typical forms of web pages: Form 1 only contains

Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE

a long article, for example, novels, biography and news. Form 2 consists of mages or graphics but almost no texts, known as galleries; Form 3 is the most regular pages which provide us both with images and texts, such as homepages. Our first task is to "divide" various kinds of web pages into three forms via some classification technique and then "conquer" them by corresponding filtering methods.

With the consideration of efficiency, it leads us to a well-accepted method of classification, Decision Trees [7]. In our approach, C4.5 takes charge of the task to divide incoming web pages into the three forms. Since this task is an accessorial step, so the time cost must be restricted within an acceptable amount which depends much on the number of attributes. As illustrated below, we pick up some very first and most obvious discriminating attributes distinguishing the structure of a web page from others (see Table 1).

TABLE 1

THE ATTRIBUTES SELECTED FOR C4.5 DECISION TREE

Attributes	Value type	Description
Is_Index	bool	Whether URL contains terms like "main" or "index"
Length of normal text	continuous	Character numbers of normal-text
Length of hyper- link text	continuous	Character numbers of hyperlink-text
Number of images(large)	continuous	Number of images with more than 50,000 pixels
Number of images(middle)	continuous	Number of images with 10,000 to 50,000 pixels
Number of images(small)	continuous	Number of images with less than 10,000 pixels

It is not difficult to notice that there are obvious differences between texts appearing in Form 1 and Form 3. In Form1, strong semantic relations can be found among words and phrases all through the article, and these texts are defined as continuous texts. However, words occurred in Form 3 do not have such connections and are organized sporadically in web pages. They always serve as descriptions of the hyperlinks, images or other multimedia. So we define them as discrete texts.

In previous work of our laboratory, three classifiers which make the footstone of the combined filtering system are shown as below:

Continuous Text Classifier: In a long article the words exert influences on each other, so that the statistic features and semantic relations of sensitive texts are analyzed and represented by a CNN-like word net. CNN (Cellular Neural Network) is a massive parallel computing network in discrete N-dimension spaces. Connections enable global disposing by exchanging and processing information in a local manner. Detailed information such as feature selection and training schedule is described in our previous paper work [5].

Discrete Text Classifier: considering that words are relatively independent in discrete text, we use Naive-Bayes classifier [8] to handle them. Usually, the occurrences of key words in text are treated as the basic semantic feature and taken as the input of classifier. Notice that the output are not "Yes" or "No" but a probability which stands for how much the incoming page would be sensitive.

Image Classifier: besides using traditional method like skin model and skin detection, This classifier obtains some reliable features for image classification based on the ROIs (region of interest). For more detail, it is recommended to see Yang's work in [6]. Also there must be some strategy which can apply single image classifier to recognition of a set of images appearing in web pages.

It is clear to see that these classifiers react to some specific web page form greatly, but not all of three forms. For instance, image classifier does well in Form 2 and Form 3 web pages which contains large sum of images, in another aspect, gets an enormously reduced efficiency in Form 1. Therefore we need to assign these separate classifiers to recognize miscellaneous web pages (see Figure 1). Specially, we need to fuse image classifier and discrete text classifier together to deal with Form 3 web pages. The detailed procedure of the fusion algorithm will be discussed in the section 3.

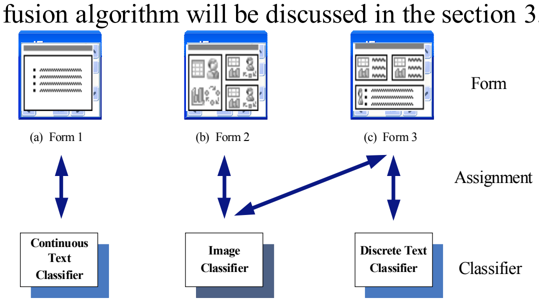

Fig. 1 Assignment between web page Form and Classifier

3. Fusion Algorithm on Form 3

Generally, the majority of images present the main topic of a web page and some pornographic images distinguish the web page from the normal ones. Based on this observation, we induct fairly strong apriori knowledge: Most of identifiable images (large in size, central in position) appearing in the target web page are either sensitive or normal. In light of this statement, the hypothesis is such that the overall web page is considered sensitive same as that all images in it are sensitive.

The probability that a normal image is mistakenly classified sensitive by the image classifier is defined as p_1; the probability that a sensitive image is mistakenly

Proceedings of the 2006 IEEE/WIC/ACM International Conference
on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE

classified normal is defined as p_2. Terms p_1 and p_2 can be acquired by a large number of observations on image classifier. Suppose that the result given by the image classifier corresponding to N-dimensions image set is $\boldsymbol{r}$. In $\boldsymbol{r}$ the notation N_1 represents the number of images which are classified as sensitive by our image classifier, whereas N_2 is the number of normal ones. Let the term *N_sensitive* denote that the web page and all images in it are sensitive, and vice versa as the term *N_normal*. Then we can acquire equations below easily:

$$p(r \mid N\_sensitive) = (1-p_2)^{N_1} p_2^{N_2}$$
$$p(r \mid N\_normal) = p_1^{N_1} (1-p_1)^{N_2} \tag{1}$$

In this formula, $\boldsymbol{r}$ which is given by image classifier is the observed data on the image set in web pages. According to the conditional probabilities and Bayes theorem, the following equations are obtained:

$$p(N\_sensitive \mid r) = \frac{p(r \mid N\_sensitive) * p(N\_sensitive)}{p(r)}$$
$$p(N\_normal \mid r) = \frac{p(r \mid N\_normal) * p(N\_normal)}{p(r)} \tag{2}$$

The two equations calculate the posterior probability on the classification of image set in web pages. From (1) and (2), instead of finding the maximally probable hypothesis on whether the page is sensitive, we can introduce a decision factor $\boldsymbol{f}$, which is the ratio of the two posterior probabilities in (2).

$$f = \frac{p(N\_sensitive \mid r)}{p(N\_normal \mid r)} \tag{3}$$
$$= \frac{(1-p_2)^{N_1} p_2^{N_2}}{p_1^{N_1} (1-p_1)^{N_2}} * \frac{p(N\_sensitive)}{p(N\_normal)}$$

From a large amount of observations, the discrete texts around the image would be a remarkable feature which describes the theme of the web page. So we replace the priori probability $p(N\_sensitive)$ with the result of discrete text classifier P_t. And accordingly, the probability $p(N\_normal)$ is replaced with $1-P_t$.

$$f = \frac{(1-p_2)^{N_1} p_2^{N_2}}{p_1^{N_1} (1-p_1)^{N_2}} * \frac{P_t}{1-P_t} \tag{4}$$

Finally, we reach the goal by defining the decision factor $\boldsymbol{f}$ in a statistical sense. If $\boldsymbol{f} \geqslant 1$, the conclusion is that the target page has sensitive content and vice versa. In addition, besides getting terms p_1 and p_2 from a large number of observations on the image classifier, we can also adjust these two factors p_1 and p_2 by ourselves as a measure to control the degree of tightness of the fusion algorithm.

4. Filtering Framework

The whole filtering procedure over sensitive web pages is shown in Figure 2. It is defined as three core phases including: (1) Obtaining web data flows and pre-processing;(2) Applying corresponding classifiers to text and image data in web pages;(3) Combining classification results to get an overall judgment of Form 3 web pages.

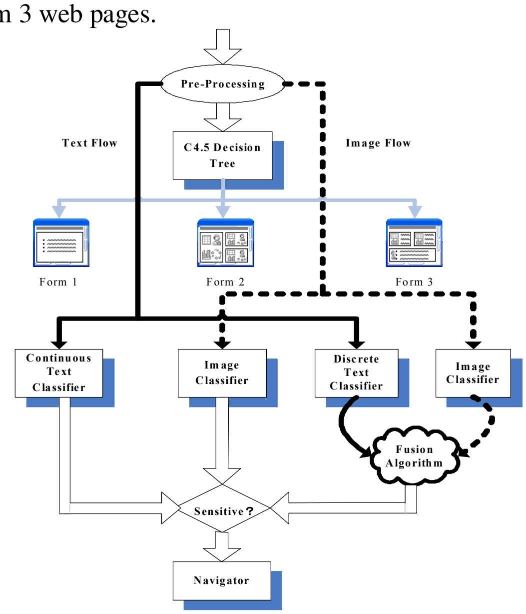

Fig.2 Outline of filtering procedure

5. Experiment

In order to evaluate the performance of the proposed method and our system, some intensive experiments are designed. Considering that there are no public sensitive web pages data sets, 1500 web pages have been manually collected from the Internet, including 500 erotic web pages and 1000 normal pages. Of all these erotic pages, there are 123 plain-text pages, 89 gallery-like pages and the rest 288 pages consist of both texts and images. The normal pages are categorized into 10 sub-folders, named Arts, Business, Science, Computer, News, Education, Entertainment, Society, Health and Sports. In business and health, note that there are some confusable pages such as ads for condoms, sex-related medical material and topics about sex education.

To illustrate our filtering system efficiency, we divide the experiments into two schemes as follows:

Scheme 1 is the test on 500 web pages which are labeled as sensitive in original 1500 training data. In the purpose of assessing the filtering method after the fusion, we compare our filtering system with three single classification methods: Naive-Bayes, kNN, majority vote using Yang image classifiers. In this scheme, samples are segmented into 3 parts apparently representing three typical web page forms described in Section 2. From Table 2 we draw several useful conclusions: (1)single classification methods all get a favorable score on corresponding category of samples, however, none of them can show good performance in other categories. (2)after merging three classifiers, there have been some dropdown in the recognition of certain category, but the fusion system gives us an encouraging result (91.8%) in principle which proves the correctness and efficiency of our system in a multiform web environment.

TABLE 2

THE FILTERING RESULTS OF SCHEME 1

	Texts Mainly	Images Mainly	Texts& Images	Total
Naive-Bayes on texts	82.1%	29.2%	74.3%	68.2%
kNN on texts	91.9%	31.4%	83.0%	76.0%
Yang method on images	4.9%	91.0%	74.7%	60.4%
Fusion System	87.8%	92.1%	93.4%	91.8%

Scheme 2 is the test of the fusion algorithm on all of the 682 test data set. As afore mentioned, some of the normal web pages contain confusable information which may lead to over blocking (69 medical websites and 44 sex education websites). This makes the filtering problem more challenging. After applying our fusion algorithm to this test set, the fusion system is able to overcome incorrect classification on confusable samples. With 113 specific pages which are medical or sex-related, the over-blocking rate is controlled within 5%. Meanwhile the experimental result also that the algorithm also ensures the correctness of filtering, it maintains a high filtering rate as 91.62%.

TABLE 3

THE FILTERING RESULT ON TEST DATA

	Blocking Rate	Over Blocking Rate	Avg. Time Cost
Result on all 682 samples	91.62%	3.06%	1650ms
Result on 113 specific samples	/	4.42%	1740ms

From the experiment results, we draw some critical conclusions: compared to “WebGuard”, our system provide a relatively low over-blocking rates and lesser time cost due to the novel framework, that is quite reasonable for that in “WebGuard”, structural-based and visual-based filters are simply organized as cascade, so the error rate and time complexity are summed up. While in our system, the complexity is the maximum of separate classifiers, and the error rate also does not accumulate.

6. Conclusion

The traditional methods use classification of either texts or images. In these years there have been some simple explorations which try to use them both [4]. This paper is an attempt to make an fusion filtering method by introducing divide and conquer strategy into filtering framework and combining texts and images in a Bayes Theorem approach. Future work will focus on: 1)considering the apriori knowledge we use is too strong, the fusion algorithm can be improved by other new information fusion method; 2)enlarging the training set to get a more accurate classification result.

7. Acknowledgement

This work is partly supported by NSFC (Grant No. 60520120099), Natural Science Foundation of Beijing (Grant No. 4041004).

8. References

[1] Cybersitter 2002 Copyright © 1995-2003, Solid Oak Software, Inc., All Rights Reserved, www.cybersitter.com, 2002.

[2] D. Forsyth and M. Fleck, “Automatic Detection of Human Nudes”, *International Journal of Computer Vision*, 32 (1): pp.63-77, 1999.

[3] Liangliang Cao, Xuelong Li, Menghai Yu, and Zhengkai Liu, “Naked people retrieval based on adaboost learning”, *Proceedings of IEEE International Conference on Machine learning and Cybernetics*, Nov 4-5 2002.

[4] M. Hammami; Y. Chahir; L. Chen; “WebGuard: Web based adult content detection and filtering system”, *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference* on13-17 pp.574 - 578 Oct. 2003.

[5] O. Wu and W.M. Hu, “Web Sensitive Text Filtering by Combining Semantics and Statistics”, *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp.663-667, 2005.

[6] J.F. Yang, Z.Y. Fu, T.N. Tan and W.M. Hu, “A novel approach to detecting adult images.” *the 17th International Conference of Pattern Recognition, Volume: 4* , Aug. 23-26, pp.479-482, 2004.

[7] J.R. Quinlan, “Induction of Decision Trees”, *Machine Learning, vol. 1*, pp. 81-106, 1986.

[8] Lewis, D. “Naive Bayes at forty: The independence assumption in information retrieval”. *Proceedings of the Tenth European Conference on Machine Learning*. 1998.