Web Sensitive Text Filtering by Combing Semantics and Statistics

Ou WU and Weiming HU
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Science
No.95 East of Zhong Gun Cun Road, Haidian District, Beijing, 100080, Mailbox 2728
Email: {owu, wmhu}@nlpr.ia.ac.cn

***Abstract*-Web sensitive information is defined as texts, pictures and other forms of information which contain erotic content on web. How to filter this harmful information attracts researchers' interests. In order to keep web content safe, governments have also given great support on the research on this problem. This paper first briefly review recent developments in web sensitive information filtering then the statistic and semantic features of sensitive texts are analyzed and represented by a CNN-like word net. Finally a novel method which combines semantics and statistics is proposed to filter sensitive text on web. Experimental results have demonstrated the proposed method's promising performance.**

I. INTRODUCTION

Internet has facilitated ones to obtain and exchange information. However, it also brings us harmful contents such as pornography, violence and other illegal messages. These harmful contents naturally have serious influence on the whole society, especially young people. So sensitive information filtering is of great importance, and has been one of most active research topics recently.

There have been a large number of filtering methods in the literature, which can be roughly classified into three major classes as follows [1][2].

PICS (Platform for Internet Content Selection) is in essence a set of specifications for content-rating systems that can rate web sites. There are usually two measures to rate the web pages. One is Self-rating and the other is the Third-part rating, which are distinguished by whether the rating results are given by web publishers or not. The filtering systems can operate through checking this rating information of web sites.

Blacklists and Whitelists. Blacklists and Whitelists are lists of web sites which compiled manually or automatically beforehand. Blacklists record URLs of web sites which are forbidden to access. Whitelists record URLs of web sites which are allowed to access. For a given a new web page, whether it is allowed to access or not depends on the matching result of the requested URL with the Blacklists or Whitelists.

Keyword-based Filtering. The idea of this approach is that sensitive texts always contain some specific words or phrases while they do not usually appear in normal texts. A word list that is composed of these specific words or phrases is usually needed to be constructed for keyword-based filtering methods, which count he number of words contained in the wordlist by matching the word list and a web page and do not allowed to browse when the number exceeds a predefined threshold.

Each kind of method mentioned above does have its own advantages in sensitive information filtering, but their drawbacks are also obvious. The PICS is not a compulsory labeling system, so the rating information is not always available. It is very difficult to keep the URL lists complete and up to date; the approach of Blacklists and Whitelists thus cannot deal with the sensitive pages effectively. As to the Keyword-based Filtering, many normal texts also contain some specific words in the word list. Therefore, this approach will lead to overblocking inevitably.

There are many commercial Web-filtering systems available currently. In 2001, The NetProtect project [3] launched by European Commission selected fifty commercial web-filtering systems and evaluated their performance. Because most of these systems use one or more traditional approaches above, it is clear that they cannot provide satisfactory results in real applications.

In order to filter the sensitive information on web more accurately, researchers have recently focused on research on the intelligent content recognition. Various algorithms are proposed to detect adult images [4]. However, they can only recognize certain kinds of adult images to some extent. Some other researchers have paid more attention to sensitive text filtering [1][2]. Based on the traditional approach of Keyword-based Filtering, Lee et al. [1] counted the number of key words appearing in the text to obtain a feature vector, and then used the vector as the input into a KSOM neural network for text classification. Although the results in the paper have shown that this method is efficient, it usually gives wrong results when the input text is about sexual healthy and other related topics. Du et al. [2] used text classification to filter sensitive texts on web. On a test data set in which adult texts were collected only from the adult category of Yahoo, their method achieved a high accuracy. In fact, the styles of erotic stories and texts are not in common, so this approach cannot work well in the real world.

In summary, there are three major problems which are not well solved in this area, i.e.

Overblocking problem: How to distinguish sensitive texts from related topic texts such as sex-related health and culture

0-7803-9361-9/05/$20.00 ©2005 IEEE

is a challenging problem which many methods can not solve efficiently.

Mis-spelled problem: Many approaches probably cannot work normally if the specific words are misspelled intentionally or unintentionally.

Wordlist problem: How to construct a sufficient and practical wordlist is a key problem for many keyword-based filtering approaches. However, so far, nobody has focused on this problem.

In this paper, we will specifically divide words which are useful for sensitive information detection into three classes. By combining semantics and statistics of texts, a more efficient text feature is obtained for the purpose of sensitive information filtering. The remainder of this paper is organized as follows. Section 2 briefly introduces the Cellular Neural Network (CNN), and semantic features of sensitive texts are analyzed and a CNN-like word net is designed for feature representation in Section 3. Section 4 simply summaries major steps of the proposed algorithm. Experimental results are given and discussed in Section 5, prior to conclusions in Section 6.

II. CELLULAR NEURAL NETWORK

CNN is a massive parallel computing paradigms defined in discrete N-dimensional spaces [5], in which each cell is a multiple input-single output processor. Cells and connections among them form the network. The main difference between the CNN and other neural networks is that connections are only allowed between adjacent cells, which allows obtaining global processing by exchanging and processing information in a local manner. Figure 1 shows an example of CNN.

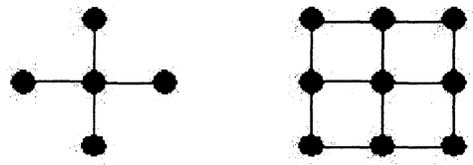

Fig.1. (a) A cell and its adjacent cells and (b) a two-dimensional CNN of 3*3 size.

A CNN dynamical system can operate in continuous (CT-CNN) or discrete time (DT-CNN) [5]. Each cell in a CNN is characterized by an internal state variable. Three parts, namely its internal state, its output and outputs from its adjacent cells, decide its final output. A mathematical description in a discrete time case is as follows:

$$\begin{aligned} x(t+1) &= g(x(t)) + I(t) + f_1(y(t)) + f_2(u(t)) \\ y(t) &= f(x(t)) \end{aligned} \tag{1}$$

where $x(t)$ is the internal state of a cell in time t. $y(t)$ is the output. $u(t)$ is external input from adjacent cells and $I(t)$ is a local value called *bias*. In addition, f_1 and f_2 are two parametric functions respectively.

The theory of the CNN has been widely applied in many areas such as signal and image processing [6][7]. In this study, we will use the main idea of the CNN to construct a CNN-like word net to illustrate semantic features of the input texts.

III. FEATURE ANALYSIS AND REPRESENTATION FOR SENSITIVE TEXT

A. *Statistic feature analysis*

Text categorization is to assign a new text into the predefined categories. The first and also predominate step is to transform texts into a suitable feature representation. There are many methods to define text feature. The common method is using statistical data of words appearing in text. Vector Space Model (VSM) may be the most notable model in text categorization, in which [8], documents are generally represented by vector of words. Let A denote the feature of a text,

$$A = (a_1, a_2, \cdots a_i, \cdots a_N) \tag{2}$$

where a_i is the weight of word i and N is the number of words we will count. The key step here is how to define the words' weights. Kerstin et al. [8] described 6 different weighting strategies in their paper. Let f_i be the frequency of word i in the text. A simple approach is to use the frequency of word as its associated weight, i.e.

$$a_i = f_i \tag{3}$$

The task of sensitive text filtering is to determine an input text sensitive or normal, it may be considered as a text categorization problem. Most of the existing texts filtering approaches are based on this idea, in which a wordlist which contains some specific words is firstly compiled. Then a vector like (2) is created to the text. Obviously, such vector is the statistical feature of a text. Although it is useful to classify the text, almost all of the semantic information about the text is not yet explored if only using this statistical method.

B. *Semantic feature analysis*

Generally, some specific words such as sex and breast are considered as the semantic features of a sensitive text. In fact, many sex-related but normal texts also contain these words. So they may provide error clues to predict the category of a text. In addition, if the Miss-spelled problem occurs, any clues from the text will impossibly be collected correctly. So how to extract right clues from a text will be critical.

Words in sensitive texts may give different semantics from they give in normal texts. But we don't know whether the input text is sensitive or not in the beginning. Other information such as context of words will be needed to decide whether we should extract these words as clues or not. As we know, some words do not contain any sensitive semantics by themselves. But if they are combined with some other words, they can provide sensitive clues. Based on the above consideration, here we specifically divide words which useful

for us to filter into three classes according to semantics as follows:

Obvious Keywords: This class of words approximately only appears in sensitive texts. In a statistical sense, the probabilities of their appearances in normal texts are close to zero. In a semantic sense, they represent erotic meaning.

Hidden Keywords: This class of words does not contain erotic meanings. But for some reasons, there are confused relations between them and sensitive texts. That is to say, the probabilities they appear in sensitive texts are high though they also appear in normal texts.

Logical Keywords: This class of keywords can be further divided into two subclasses. One is multi-semantic word such as breast, which provides normal information in normal texts and erotic information in sensitive texts. The other is that if only some specific words are companied with them, these words are considered as erotic keywords.

Many existing approaches only use obvious keywords and the first subclass of logical keywords. In fact, hidden keywords and the second subclass of logical keywords also can greatly help classify a new text.

There are giant numbers of words in human brains and they are not isolated each other. These words form a huge net by semantic relations among them, which will facilitate to process text information accurately. For example, when we read a word in an article, we may associate with other semantic related words. In addition, when we read a word, the node corresponding to this word accepts it as an input. Then the node's state is changed according to its previous state and the states of its adjacent nodes. This mechanism enlightens us that the three classes of keywords and their semantic relations can represent the semantic features of the sensitive texts reasonably. In the next subsection, we explain the semantic information among them and accordingly construct a CNN-like word net to describe the semantic features. Figure 2 shows the main difference in keyword set or wordlist between the traditional and our approaches.

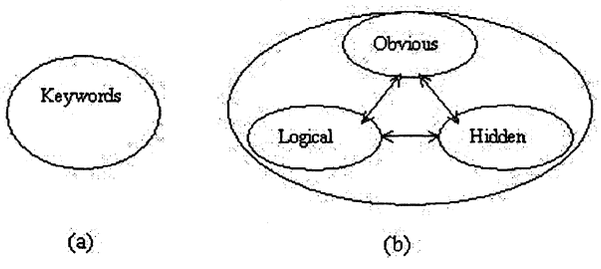

Fig.2. Traditional keyword set and our keyword set.

C. *CNN-like word net*

There exists semantic information among words. When we see the word "teacher", the word "student" may appear in our minds unconsciously. Provided that you only see three words "parents", "study" and "children" in a paragraph, you may say this paragraph is about parents' role in children's education. It means that only several separated words can give us an integrative meaning. When you see a word "students", you may be not sure whether it means middle school students or university students. But, if you see "bachelor" in the following paragraph, you can say it most possibly means university students. Examples like these have shown that semantic information among words can help obtain more informative clues about words. In this study, we will try to explore this information to help us extract right clues from a text.

To construct a CNN-like word net, we define a cell as a word. And four parameters are used to describe a cell. They are state, position set for appearances, number and activated number respectively. A cell has three kinds of states: sleep if it doesn't appear, fallow if it has appeared but not been activated and activity if it is activated. The output of a cell is equal to the internal state and the internal state of a cell is described by the four parameters above. Instead of using strict function to describe the input and output of a cell, we just use some semantic rules. A cell and its adjacent cells are shown in Figure 3.

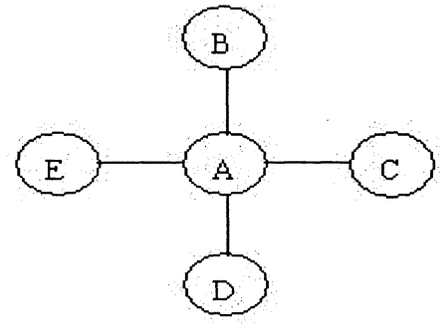

Fig.3. A word cell and its adjacent cells.

Provided that *A* is a hidden keyword and its initial state is sleep. When there is an input for *A*, if the parameters of *B, C, D* and E and previous parameters of *A* meet a certain rule, then *A* will be activated. Otherwise the node *A* turns to the fallow state. This process can help us extract sensitive-prone words efficiently. For example, in a text about sex healthy, although *A* may also appear, but the parameters can't meet a certain rule, then *A* can't be activated. If there is a word similar to *A*, and the parameters meet the rule, we can also activate *A*. Only those activated words are considered as useful clues. If *A* is an obvious or logical keyword, the disposal is similar to this. It is obvious that this process can help solve the Miss-spelled problem. Because when we see a miss-spelled word, we can return to its context or other semantic information to understand it correctly.

The whole CNN-like word net is constructed by a number of paradigms described above. Link between any two cells shows their semantic relations. Rules indicate in what case a cell can be activated. Different paradigms have different rules. Links and rules among the three kinds of words can effectively represent the semantic features of sensitive texts.

IV. THE ALGORITHM

The key step of the proposed approach is constructing the CNN-like word net properly. Keywords are selected according to our keyword classification strategy. Ideally, we had better use machine-learning to automatically attain rules (semantic relations) among cells (words). Considering that only specific texts need to be filtered, here we construct the CNN-like word net manually.

A. Feature extraction

The initial states of all cells are set to sleep. Major steps involved in feature extraction are summarized as follows.

S1. Obtain a word from the input text and match the word with each cell in the CNN-like net gradually. Find a cell which has the highest similarity with the input word. If the similarity score exceeds a predefined threshold, adjust the parameters except the state and activated number of this cell, and then turn to S2; else, turn to S1.

S2. Get the parameters of the cell and its adjacent cells from the CNN-like net. If these parameters meet a certain rule, then this cell is activated. Its activated number is added to one and its state turns to activity. Otherwise, it turns to fallow.

S3. If its state is changed, adjust states of the adjacent cells by the same process as S2. Then adjust the whole net iteratively. If all words of the text have been processed, turn to S4; else turn to S1.

S4. Collect the activated number of each cell and forms a vector.

The vector is used to represent semantic and statistic features of the input text.

B. Training and classification

Support vector machine (SVM) is a very popular classification technique now. It transforms classification to a lineal layout problem. The algorithm finds a hyper plane between different classes of the training data. Once the hyper plane is determined, we can use it to classify a new data [2]. SVM is also applied in text classification [9]. We choose SVM as our classifier for its well performance in text classification.

V. EXPERIMENT

To evaluate the proposed method's performance, 3162 Chinese texts have been collected from Internet, which include 577 sensitive texts, 585 sex-related but normal texts and 2000 normal texts. The normal texts consist of 10 subcategories, namely Arts, Business, Science, Computer, News, Shopping, Game & Recreation, Society, Health and Sports. Each subcategory contains 200 web texts. The Health subcategory was mainly collected from the following web sites: www.xyxy.net, health.21cn.com, www.fm120.com and www.medicch- ina.com. Compared to other text databases for web filtering, only we collected sex-related normal samples such as sexual health, sexual culture and sexual education. 300 sensitive texts, 300 sex-related normal texts and 1000 normal texts are used as training data, and the remaining serves as test data. A list of 109 indicative terms comprising 29 obvious keywords, 33 hidden keywords and 47 logical keywords has been compiled. Some simple rules are constructed to describe the relationships among these three kinds of keywords. All keywords and simple rules form the CNN-like word net.

Three different feature extraction schemes are designed to test the effectiveness of our approach. The first is the traditional scheme, in which only the number of each obvious keyword and a part of logical keywords is counted. The second is to count all of the three kinds of keywords. The third is to count keywords through the CNN-like word net. The free software Libsvm-2.6 [10] is used to train and predict on our database. The experimental results are summarized in Table 1, from which and false recognized texts, we can get several useful conclusions: (1) Comparing scheme 1 with scheme 2, we see that our definition three kinds of keywords and constructing the keyword set according to this definition can improve the recognition rates noticeably. But because the keyword set contains many logical keywords and hidden keywords, it is inevitable to classify some normal texts into sensitive category. (2) Using the CNN-like word net to extract features of texts, we get the best classification rates. It proves that CNN-like word net can represent the semantic feature of sensitive texts properly. (3) The real sensitive texts have protean styles and contents. So in order to get a higher classification rate, we need to enlarge the keyword set. (4) Because we only consider some naïve semantic relations among words, some normal related texts are also predicted as sensitive texts.

TABLE I

THE CLASSIFICATION RATES OF THREE SCHEMES

	Sensitive Texts	Related Texts	Normal Texts	Total
Scheme1	93.14%	95.08%	100%	97.88%
Scheme2	96.38%	97.54%	99.80%	98.78%
Scheme3	97.83%	98.24%	100%	99.29%

Du et al. [2] applied text category to filter sensitive web pages. Although their training and test sets are English texts and ours are Chinese, we make comparison between their approach and ours. Because both their and our methods are not specific language-oriented. In their approach, it needs to set the threshold t manually. Table 2 shows the results of the two approaches, where Blocking Rate is the fraction of the correct classified texts in the sensitive text set, while Overblocking Rate is the fraction of the false classified texts in the non-sensitive text set. The results of Du's method are taken directly from his paper. It can be seen that our result is better than theirs in an overall manner, even though our text data set is more challenging than theirs. They collected test

data about adult only from the adult category of Yahoo and didn't contain sex-related normal texts.

TABLE II

THE CLASSIFICATION RATES OF THREE SCHEMES

	Blocking Rate	Overblocking Rate
Our approach	97.83%	0.39%
Du's approach (t=0.18)	97.41%	0.48%
Du's approach (t=0.10)	99.35%	4.09%

Experimental results also show that that our approach can solve the Overblocking problem and Wordlist problem efficiently.Our classification of keywords can guide us how to select keywords to construct wordlist (or keyword set) more efficiently. Our CNN-like word net can help extract right clues from text and avoid blocking normal texts. We have not yet done experiments about the Mss-spelled problem. To English words, it is easy to calculate the similarity between two similar words. For example, the words 'University' may be spelled as 'Uinervtisy'. But it is relatively more difficult to handle the Chinese words because they may be similar in pronunciation or shape. But it is believable that, if there is a simple method to calculate the similarity between the Chinese words, the proposed approach can solve the Miss-spelled problem.

VI. CONCLUSIONS

This paper have defined three kinds of keywords and constructed a CNN-like word net to extract and represent semantic and statistic features of texts. This study is an attempt to use semantics to solve the three unsolved problems in this area. Experimental results have shown that the proposed approach is very promising. Future work will focus on (1) Enlarging our keyword set and designing more accurate semantic rules among words so as to construct a better CNN-like word net and (2) Finding a feasible way to calculate the similarity between two Chinese words.

ACKNOWLEDGEMENT

We are very appreciative to Jeffrey L. Wang (Dept.of Computing, Imperial College London) for his revision and valuable suggestions in English. We also thank Mingliang Zhu and Zhouyao Chen (National Laboratory of Pattern Recognition, CAS) for their hard work in web page collection.

REFERENCES

[1] P.Y. Lee, S.C. Hui and A. Fong, "Neural Networks for Web Content Filtering", *Intelligent Systems*, 17(5): 48-57, 2002

[2] R. Du, R. Safavi-Naini and W. Susilo, "Web Filtering Using Text Classification", *Proc. the 11th IEEE Intl. Conf. on Network*, pp. 325 - 330, 2003.

[3] NetProtect Research Group, "Report in Filtering Techniques and Approaches NETPROTECT: WP2: 2.3 V1.0 23", *Technical Report*, Oct 2001

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

[5] http://www.ce.unipr.it/pardis/CNN/#InterPoint

[6] A. Lukianiuk, "Capacity of Cellular Neural Networks as Associative Memories ", *Proc. the Fourth IEEE Int. Workshop on Cellular Neural Networks and Their Applications*, pp. 37-40, 1996.

[7] M. G. Milanova, A. C. Campilho, and M. V. Correia, "Cellular Neural Networks for Segmentation of Image Sequence", *Proc. the 11th Portuguese Conference on Pattern Recognition*, pp. 49-54, 2000.

[8] K. Aas and L. Eikvil, "Text Categorization: a Survey", *Technical Report -941*, Norwegian Computing Center, 1999.

[9] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *Proc. the 10th European Conference on Machine Learning*, pp.137-142, 1998.

[10] http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf