Harmful Contents Classification Using the Harmful Word Filtering and SVM

Wonhee Lee[1], Samuel Sangkon Lee[2], Seungjong Chung[1], and Dongun An[1]

[1] Dept. of Computer Engineering, Chonbuk National University, South Korea
{wony0603, sjchung, duan}@chonbuk.ac.kr
[2] Dept. of Computer Engineering, Jeonju University, South Korea
samuel@jj.ac.kr

Abstract. As World Wide Web is more popularized nowadays, it is also creating many problems due to uncontrolled flood of information. The pornographic, violent and other harmful information freely available to the youth, who must be protected by the society, or other users who lack the power of judgment or self-control is creating serious social problems. To resolve those harmful words, various methods proposed and studied. This paper proposes and implements the protecting system that protects internet youth user from harmful contents. To effectively classify harmful/harmless contents, this system uses two steps of classification: harmful word filtering and SVM learning based filtering. We achieved result that the average precision of 92.1%.

Keywords: Harmful Word, Content Classification, Filtering, SVM.

1 Introduction

As World Wide Web is more popularized nowadays, the environment is flooded with the information through the web pages. However, despite such convenience of web, it is also creating many problems due to uncontrolled flood of information. The biggest problems are that the users are facing excessive information, making it difficult to search for the right one, and that there is uncontrolled harmful information. Especially, the pornographic, violent and other harmful information freely available to the youth, who must be protected by the society, or other users who lack the power of judgment or self-control is creating serious social problems. There have been many rules and studies of various types to resolve the problem for example PICS, Keyword filtering, research based on image data, intelligent analysis system. However, these methods all have limitation, and their performance levels are very low in actual application [2].

This paper describes the method of filtering using the text data of the web contents. To increase the accuracy of classification, filtering is performed in two steps. First, the system classifies the contents as harmful or harmless using harmful word filtering. In the second step, the harmful contents are rated using SVM based filtering. In Chapter 2, the preceding studies dealing with web contents classification are reviewed and summarized. In Chapter 3, the proposed algorithm and its execution are described. In

Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 18–25, 2007.

Chapter 4, the experiment using the proposed algorithm is described and evaluated. The conclusion is presented in Chapter 5.

2 Related Work

Study of web content filtering can be mainly divided into platform for Internet content selection, URL interception, keyword filtering, artificially intelligent contents analysis and image based filtering [2], [11].

2.1 PICS (Platform for Internet Content Selection)

PICS is a technical standard that allows detecting and classifying the meta data, which describes the web page, using the computer software. RSACi and SafeSurf are generally used for classification of PIC contents. RSACi (Recreational Software Advisory Council) uses four categories of harsh language, nudity, sex and violence. Each category is then further classified into 5 ratings from 0 (harmless) to 4. The classification system of SafeSurf is more detailed. In order to describe harmfulness of web contents to each age group, it uses eleven categories [2], [8], [5], [10].

2.2 Keyword Filtering

This method intercepts the web contents based on the harmful words or phrases contained in the content. The words and phrases in the web page are compared with predefined keyword dictionary, and the content is intercepted if the used of those words and phrases exceed a critical number. This method can quickly determine if the web page potentially contains the harmful content. However, it has the problem of over-blocking the harmless contents due to double meaning of the words or phrases [2], [11], [15].

2.3 Intelligent Content Analysis

A web filtering system can use intelligent content analysis to automatically classify the web contents. One of them is artificial neural network that can learn according to the applied training case. Such learning and adaptation can distinguish the syntax depended words like 'sex' that appears widely in both pornographic and other web pages [10], [11].

3 Implementation

3.1 System Architecture

The system proposed in this paper executes two steps of harmful word filtering and SVM learning to increase the accuracy of classification and shorten the time for evaluation. The system is structured as shown in Fig. 1.

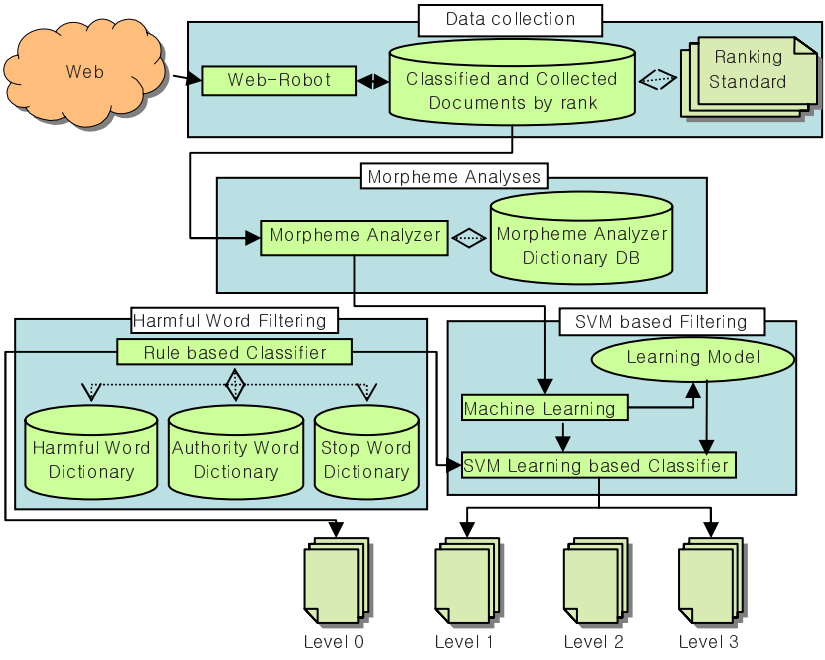

Fig. 1. System Architecture

3.2 Harmful Word Filtering

The harmful word filtering step basically executes keyword filtering. For the filtering work, the harmful word dictionary, authority word dictionary and stop word dictionary are developed and used. If the harmful word dictionary is structured just as a simple list of harmful words, it has the risk of causing the same over-blocking problem as the existing keyword filtering. Therefore, the dictionary adds the information of the words that are used with the harmful words. The added words can lower the fault classification rate due to dual meaning of the words by considering the syntax information of the word. Authority Word dictionary is standard word list for abbreviated or metamorphosed form intentionally or mistakenly by the producer of the content.

3.2.1 Adjacent Word and Non-adjacent Word

A harmful word dictionary consists of the list of the harmful words and the adjacent and non-adjacent words of the harmful word. An adjacent word is defined as the word that can determine the harmfulness level of the harmful word when appearing in the same sentence. A non-adjacent word is the one that appears in the same content as the harmful word, although not in the same sentence, and can still determine the harmful nature of the content. For example, a word 'breast' can be harmful or harmless. In addition, when the word is used in harmful meaning, its level of harmfulness can be very high or little depending on how it is used. If the word 'cancer' appears in the

same sentence as 'breast', then it is very likely that the content is of harmless nature. However, if the word 'tongue' appears, it is likely that the content is of harmful nature. Furthermore, if the word 'rope' appears in the same sentence, the content is likely to be very harmful, containing pervertible nature.

3.2.2 Authority Word Dictionary

When a standard word is presented in abbreviated or metamorphosed form intentionally or mistakenly by the producer of the content, it must be transformed to its standard form. Otherwise, the abbreviated word or metamorphosed word can change the appearance frequency and affect the evaluation of being harmful or harmless.

3.2.3 Harmful Word Filtering

Harmful word filtering is executed in the following procedure:

- Step 1: Tags are removed from the document, and the morpheme analysis is performed.
- Step 2: The list of the morpheme analyzed words is transformed to their standard words.
- Step 3: The harmful word candidates are selected from the word list. The identified word candidates are evaluated of their harmfulness using the adjacent and non-adjacent words. The frequency of the harmful word usage is then calculated. If it is larger than a certain value (α), the document is considered harmful.

The determining function $f(d_i)$ of a document can be expressed as follows:

$$f(d_i) = \frac{FT}{N}, \tag{1}$$

where *FT* is the count of the harmful words in a document and *N* is the total number of harmful words.

3.3 SVM Learning Based Filtering

SVM based filtering consists of the following four steps: (1) Feature Extraction, (2) Indexing, (3) Generation of learning mode, and (4) Classification using the learning model.

3.3.1 Feature Extraction

Feature extraction deals with generating a list of feature to be used for learning based text classification. For the algorithm to extract the feature, DF (Document Frequency), IG (Information Gain), MI (Mutual Information), and CHI-square are tried to determine the best method. DF means the number of documents that a certain word occurs in a set of documents. IG is an algorithm that selects only the feature with high contribution in order to calculate how a word's appearing in the document

contributes to classification of the document. CHI seeks a quantity of term importance by measure the dependence relationship between term *t* and category *c* [7].

3.3.2 Indexing

The extracted qualities needed to be weighted according to their importance in the document. For indexing and weighting process, the TF, TF-IDF(Inverse Document Frequency) and TF-ICF(Inverse Category Frequency) algorithms are used [7].

3.3.3 Generation of Learning Model

SVM (Support Vector Machines) are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. This learning strategy introduced by Vapnik and co-workers is a principled and very powerful method that in the few years since its introduction has already outperformed most other systems in a wide variety of application [9]. SVM selects the optimum hyperplane from two dimensional classifications and uses it as the decision border surface. The optimum hyperplane identifies two linear separable groups and maximizes the margin. However, since the linear separable case is rare in practical problem, the non-linear space is mapped to the linear space using the kernel function and then classifies using the linear SVM. In this paper, C-SVC and RBF are used as the SVM type and kernel, respectively. The deciding function of C-SVC is

$$\operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x_j) + b\right), \tag{2}$$

and kernel function is

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \quad \gamma > 0 \tag{3}$$

3.3.4 Classification by Using the Learning Model

Classification is the process of rating the document using the generated learning model. The document to be rated creates the feature using the morpheme analysis and goes through indexing and normalization using the created feature. The document data generated through normalization is then rated using the learned model generated through learning.

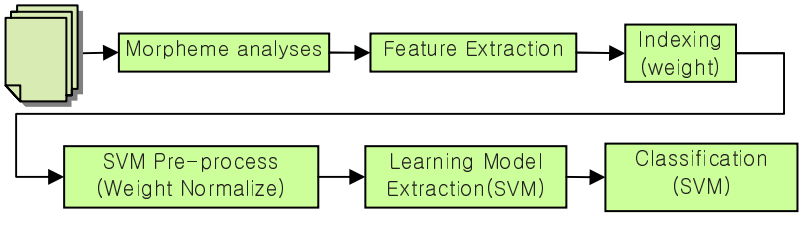

Fig. 2. Processing of Learning and Classification Using SVM

4 Experiment and Evaluation

The environment for the experiment and evaluation is as follows: EHDS-2000 (ETRI Harmful Data Set), which was built as the test based harmful data set, is used. EHDS-2000 is harmful document set built by ETRI(Electronics and Telecommunications Research Institute) in South Korea. EHDS-2000 is composed of Korean Documents and English Documents shown in Table 1. The learning and test data used for the experiment and evaluation are as follows:

Table 1. Data Set

		Collection Documents		Training Documents		Test Documents	
		harmful	harmless	harmful	harmless	harmful	harmless
Languages	Korean	2,572	2,126	1,164	588	509	462
	English	12,250	3,340	936	694	449	533

The experiment was conducted in two steps. First, the optimal algorithm set is determined from the qualifier extraction algorithm (logTF, IG and CHI) and indexing algorithm (TF-IDF and TF-ICF). For the experiment, the qualifier counts between 200 and 800, were adjusted, incrementing by 100. The result of the experiment is shown in Fig. 3. The figure indicates that the combination of CHI and TF-IDF showed the best result at the qualifier count of 800.

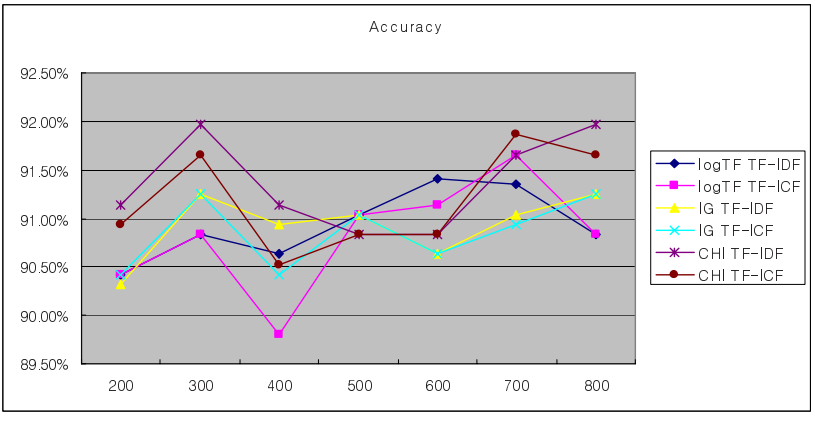

Fig. 3. Comparison of Feature Extracting Algorithm

The second step involved two experiments. One was the classification using harmful word filtering only while the other one was the classification using both harmful word filtering and SVM based filtering. Table 2 shows the result of harmful word filtering only. It indicates that the accuracy level 97.4% for the harmful words. However, the accuracy for the harmless words was only 51.3%. And 48.7% of inaccuracy included the sex advice and medical information document.

Table 2. Result of Classification Using Harmful Word Filtering Only

Non-Harmful(995)		Harmful(958)		Overall Accuracy
correctly	incorrectly	correctly	incorrectly	
511 (51.3%)	484 (48.7%)	933 (97.4%)	25 (2.6%)	74.35%

The inaccuracy was caused by the harmful words in the document. To improve the result, the document sets were separated into three types. First, the harmless documents are divided into type-0 and type-1. Those that contain no harmful words were classified as type-0. And the sex advice or medical information documents that contain the harmful words were classified as type-1. Then the harmful documents were classified as type-2. Table 3 shows the result of the new experiment. It indicates 92.1% accuracy in average. For the harmless documents, the type-0 showed 95.9% accuracy and type-1 90.9%, making the average accuracy of 93.4%, a big improvement. The accuracy of the type-2 (harmful) documents actually was lowered from 97.4%. 7.9% reduction mostly came from the sex advice or medical information contained in the harmful documents. The cases of type-1 or type-2 documents assessed as type-0 were 1.7% and 2.6% for type-1 and type-2, respectively. They were mostly image and the rest of the text was very little.

Table 3. Result of Classification Using both Harmful Word Filtering and SVM based Filtering

Input Data \ Result	Type-0	Type-1	Type-2
type-0(524)	503(95.9%)	14(2.7%)	7(1.4%)
type-1(471)	8(1.7%)	428(90.9%)	35(7.4%)
type-2(958)	25(2.6%)	76(7.9%)	857(89.5%)

5 Conclusion

At present, Internet is flooded with information including the harmful contents of pornographic, violent and suicidal nature. To solve the problem, various methods such as the rating system – PICS, keyword filtering, intelligent analysis system are proposed and studied. But these methods all have certain limitations for intercepting the harmful contents.

This paper proposes the system to protect the young Internet users from harmful sites by accurately identifying the harmful contents widely available on the web pages. For effective classification of harmful and harmless contents, the paper proposes a two step system of harmful word filtering and SVM learning based filtering. For harmful word filtering, the syntax information using the adjacent words and non-adjacent words was added to the harmful word dictionary. That improved the classification accuracy. For the quality of the contents, the various algorithms such as logTF, IG, CHI, TF-IDF and TF-ICF were tried to select the optimal extraction and

normalization algorithm. As the result, the accuracy level of 92.1% was achieved. In the future, the accuracy needs to be further improved by analyzing the various content elements such as the image, sound or document structure.

Acknowledgments. This work was supported by the second stage of Brain Korea 21 Project. And this work was also financially supported by the Jeonju University.

References

1. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. : A Practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
2. Christopher D. Hunter : Internet Filter Effectiveness: Testing Over and Underinclusive Blocking Decisions of Four Popular Filters, Proceedings of the Tenth Conference on Computers, Freedom and Privacy: Challenging the Assumptions, (2000) 287-294
3. Dequan Zheng, Yi Hu, Tiejun Zhao, Hao Yu, and Sheng Li : Research of Machine Learning Method for Specific Information Recognition on the Internet, IEEE International Conference on Multimedia Interfaces (ICMI), (2002)
4. Huicheng Zheng, Hongmei Liu, and Mohamed Daoudi : Blocking Objectionable Image: Adult Images and Harmful Symbols, IEEE International Conference on Multimedia and Expo (ICME), (2004) 1223-1226
5. Jae-Sun Lee, and Young-Hee Jeon : A Study on the Effective Selective Filtering Technology of Harmful Website Using Internet Content Rating Service, Communication of KIPS Review, Vol. 9, No. 2, (2002)
6. KwangHyun Kim, JoungMi Choi, and JoonHo Lee : Detecting Harmful Web Documents Based on Web Document Analyses, Communication of KIPS Review, Vol. 12-D, No. 5, (2005) 683-688
7. JH Jeong, WH Lee, SW Lee, DU An and SJ Chung : Study of Feature Extraction Algorithm for Harmful word Filtering, KCC Summer Conference, Vol. 33. No. 01. (2006) 7-9 (in Korean)
8. Mohamed Hammami, Youssef Chahir, and Liming Chen : WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis, IEEE Transaction on Knowledge and Data Engineering, Vol. 18, No. 2, (2006)
9. Nello Cristianini, and John Shawe-Taylor : An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, (2000)
10. P. Y. Lee, and S. C. Hui : An Intelligent Categorization Engine for Bilingual Web Content Filtering, IEEE Transaction on Multimedia, Vol. 7, No. 6, (2005)
11. P. Y. Lee, S. C. Hui, and A. C. M. Fong : Neural Networks for Web Content Filtering, IEEE Intelligent Systems, (2002) 48-57
12. Yun-Jung Jang, Taehun Lee, Kyu Cheol Jung, and Kihong Park : The Method of Hurtfulness Site Interception Using Poisonous Character Weight, KIPS Spring Conference, Vol. 10, No. 1, (2003) 2185-2188 (in Korean)