A Hierarchical Text Rating System for Objectionable Documents

Chi Yoon Jeong*, Seung Wan Han*, and Taek Yong Nam*

Abstract: In this paper, we classified the objectionable texts into four rates according to their harmfulness and proposed the hierarchical text rating system for objectionable documents. Since the documents in the same category have similarities in used words, expressions and structure of the document, the text rating system, which uses a single classification model, has low accuracy. To solve this problem, we separate objectionable documents into several subsets by using their properties, and then classify the subsets hierarchically. The proposed system consists of three layers. In each layer, we select features using the chi-square statistics, and then the weight of the features, which is calculated by using the TF-IDF weighting scheme, is used as an input of the non-linear SVM classifier. By means of a hierarchical scheme using the different features and the different number of features in each layer, we can characterize the objectionability of documents more effectively and expect to improve the performance of the rating system. We compared the performance of the proposed system and performance of several text rating systems and experimental results show that the proposed system can archive an excellent classification performance.

Keywords: Objectionable documents, document analysis, text classification, hierarchical system, SVM

1. Introduction

As the Internet has recently been rapidly expanded, we can find information easily and quickly. A lot of useful information exists on the Internet, but there is also harmful information involving pornography, drug abuse, violence, etc. Therefore, a filtering system is necessary to protect the children from this harmful information.

Many filtering systems, for examples Honorguard, SurfControl and S4F, have been developed and used to block harmful information in homes, schools, libraries, and so on. There are two kinds of filtering approach to blocking harmful information: the metadata-based approach and the content-based approach.

The metadata-based approach depends on the results of URL and IP address blocking. This approach was mainly used in early filtering systems. Because the Internet is very dynamic and the URL/IP addresses of web sites are always changing, this approach has the problem of having to periodically update blocking lists. To solve such problem, several studies of content-based filtering have been conducted [1, 2, 3, 4, 5, 6, 7, 8].

Although there are many applications of content-based filtering according to the types of the media and the methods of analyzing contents, we limit the discussion to objectionable document filtering, a serious problem on the Internet. In this paper, the described objectionable document is a pornographic document written in Korean.

Automatic objectionable document filtering methods can be classified into two main approaches: keyword matching and inductive learning model. Keyword matching is the most commonly used method of classifying the objectionable documents and has an advantage in the processing time. But, this method has low accuracy and produces many false positives. There have been few studies that try to classify the objectionable texts using the inductive learning model [8, 9]. Gui-yang et al. [9] proposed the hybrid method which combines keyword matching and the inductive learning model in order to increase the precision of objectionable text filtering. But they focused on two-class problems, whether the documents were objectionable or not. Since documents about sexual medicine or consultation on sexuality can be classified as objectionable texts in two-class problems, the filtering system comes up with many false positives. In addition, to classify the texts into objectionable and unobjectionable texts limits the user's option to access diverse information. For these reasons, a rating system of objectionable texts is needed. Lee et al. [8] proposed a text rating system which consists of a non-harmful documents screen and inductive learning model. But they used a single classification model which assigns a document into one of three rates. Since the text rating system based on a single inductive learning model selects features from the training samples of all rates, it leads to confusion in rating documents and decreases accuracy of the text rating system. So a novel approach to classifying the rates of documents is necessary.

In this paper, we classify the objectionable texts into four rates and propose a novel hierarchical text rating system for objectionable documents. The proposed system consists of three layers and each of them uses its own classification model which has different features and a different number of features.

Manuscript received October 4, 2005; accepted November 7, 2005.

* Privacy Protection Research Team, ETRI, Daejeon, Korea ({iamready, hansw, tynam}@etri.re.kr)

The remainder of this paper is organized as follows. In section 2, the rate of objectionable texts and design concept are described. The hierarchical text rating system for objectionable documents is described in section 3 and experimental results are given in section 4. In section 5, we conclude this work.

2. The Proposed System

2.1 The rate of objectionable texts

Up to now, the filtering approaches of the objectionable text have been focused on distinguishing between objectionable texts and unobjectionable texts. Consequently, there have been many false positives, which degrade the performance of filtering systems, and restrict the user's access right. To solve these problems, it is necessary to classify the documents according to their objectionability. In this paper, we use the four rates for the classification of objectionable texts, as described in Table 1.

Table 1. The rate of objectionable texts

	Description
Rate 0	Texts which do not have any objectionable expression.
Rate 1	Texts which have the objectionable words but represent u seful information such as sexual medicine, sexual consult ation, etc.
Rate 2	Texts which describe the human body or sexuality
Rate 3	Texts which describe sexual perversion.

2.2 Design Concept

The automatic text categorization system assigns the predefined categories to free text documents and generally consists of feature selection, representation, and the classification process, as illustrated in Figure 1.

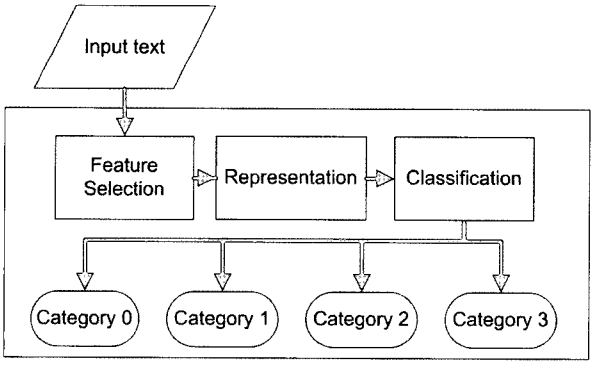


Fig. 1. Automatic text categorization system

In the automatic text categorization system, there exist sharp distinctions between documents in other categories. However, in the text rating system for objectionable documents, which classifies the documents in the same category according to their objectionability, there is no clear distinction between documents in the other rates. It is due to the similarities in used words, expressions, and structure of the document in the same category. Therefore, if we extract and select features from documents of all rates at the same time, the performance of the rating system will be decreased owing to confusion between rates.

It may be helpful to consider some important characteristics of rating the documents here. Since the documents in rate 0 do not contain objectionable words or expressions, they are definitely distinguished from the documents in the other rates. In addition the documents in rate 1 not only contain the objectionable words describing sexuality or sexual organs, but also contain the words which are used in medicine or consultation. So it also differentiates between the documents in rate 1 and those in rate 2 or 3. The documents in rate 2 and rate 3 contain the same words expressing sexual acts, but the contents of them are very different. The documents in rate 2 generally contain implicit expressions on sexual acts and the other documents in rate 3 usually contain explicit expressions on sexual perversion. From the results of the above observation, we can classify objectionable documents into several subsets by using the properties of them. Also we can reduce the confusion of rated documents and improve the performance of the text rating system by classifying the subsets hierarchically.

In this paper, we divide objectionable documents into five subsets: subset A (documents in rate 0), subset B (documents in rate 1, 2 and 3), subset C (documents in rate 1), subset D (documents in rate 2 and 3), subset E (documents in rate 2) and subset F (documents in rate 3). The subsets are classified by the proposed hierarchical text rating system. The proposed system consists of three layers. In the first layer, the text is classified as subset A or B and the text classified as subset B passes to the next layer. The text is classified as subset C or D in the second layer and the text classified as subset D in the second layer is classified as subset E or F in the last layer. At each layer, the different features and the different number of features are used to classify the subsets.

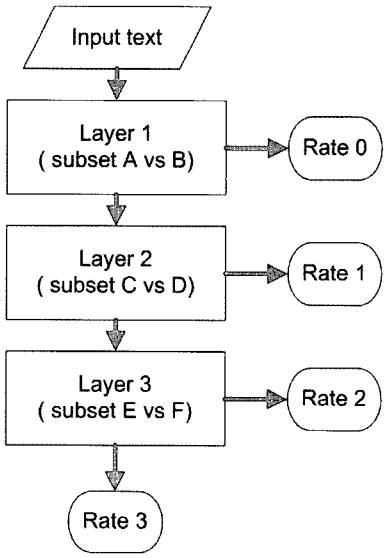


Fig. 2. Proposed system

3. A Hierarchical Text Rating System for Objectionable Documents

In this paper, we propose the hierarchical text rating system for objectionable documents as depicted in Figure 3. The proposed system consists of three layers and each layer uses the different features and classifiers. We gathered sample documents for learning and testing written in Korean. First, we extract nouns from the sample documents using a morphological analyzer. And then, the features are selected using chi-square statistics at each layer and the weight of features are calculated using the TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme. Finally, the weights of features are used as an input to the non-linear SVM (Support Vector Machines) classifier at each layer. We will discuss feature selection, representation and the non-linear SVM classifier in the following sections.

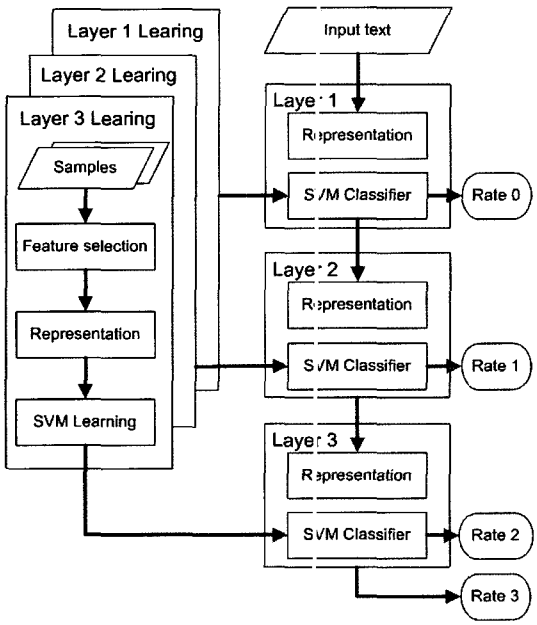


Fig. 3. Hierarchical text rating system for objectionable documents

3.1 Feature selection and representation

Feature selection is to find the characteristic words which can help classify the rates of documents well. The process of feature selection consists of the morphological analyzer and the algorithm by which the amount of information of characteristic words is measured. The morphological analyzer is used to extract the substantives from the documents. But the number of substantives in documents is so high that the method of reducing the number of substantives is needed in order to reduce the processing time of document classification. Therefore the algorithm of measuring the amount of information on substantives is required to find effective words in classifying the rates of documents. Generally, the features are selected by the term-goodness criterion such as DF (Document Frequency), IG (Information Gain), CHI (CHI-square statistics), and MI (Mutual Information). It is known that IG and CHI are the best in automatic document categorization [10].

While the features are selected from the documents of all categories in the automatic text categorization system, the different features are selected from the subsets at each layer in the proposed system. As a result, we characterize the objectionability of texts more effectively and can expect to improve the performance of the rating system. In the proposed system, we use chi-square statistics as the term-goodness criterion of the features at each layer. Chi-square statistics measure the lack of independence between a term t and a rate r and can be compared to the chi-square distribution with one degree of freedom to judge extremeness [10]. It is defined as:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

where A is the number of times t and r co-occur, B is the number of times t occurs without r, C is the number of times r occurs without t, D is the number of times neither r nor t occurs, and N is the total number of documents.

Representation means how to express the documents with features. In this paper, the TF-IDF weighting scheme [11] is used to calculate the weight of features. W_{ik} is the weighting value of the ith feature in the kth document and is defined as:

$$w_{ik} = (\log f_{ik} + 1) \times \log\left(\frac{N}{n_k}\right)$$

where f_{ik} is the frequency of the ith feature in the kth document, N is the total number of documents in the training set and n_k is the number of documents in which the ith feature occurs.

3.2 SVM Classifier

There have been a wide range of statistical learning algorithms such as K-NN, SVM, Bayes probabilistic classification and neural networks applied to the automatic text categorization system [12]. In this paper, we use the SVM classifier.

SVM is based on the Structural Risk Minimization principle from computational learning theory [13]. Non-linear SVM [14] maps the data into a predetermined very high dimensional space via a kernel function and finds the hyper-plane that maximizes the margin between the two classes. Non-linear SVM classification is extremely efficient and robust. Therefore, non-linear SVM is used to classify the rates of documents using the values of features which are calculated by the TF-IDF weighting scheme.

4. Experimental Results

We gathered 3000 HTML documents for each rate from various websites. The texts, which are extracted from

HTML documents using HTML parser, were used for experiment. For each rate, training samples contain 2400 texts and test samples contain 600 texts. We used chi-square statistics to select features from samples, and the TF-IDF weighting scheme was used for feature representation. Finally, Non-linear SVM with the radial basis function as the kernel function was used as a classifier. The best kernel parameters for the SVM classifier were calculated from a grid search and n-fold cross validation.

In layer 1, the harmful words appear on the top of the selected feature list. In layer 2, the words relating to sexual medicine or consultation on sexuality rank highly in the selected feature list. In Layer 3, we can not find a dominant characteristic in the selected feature list.

Table 2. Classification results with the same number of features at each layer

	# of features	Accuracy
Layer 1 (2400)	400	95.58%
Layer 2 (1800)	400	89.72%
Layer 3 (1200)	400	74.00%

Table 3. Classification results with a different number of features at each layer

	# of features	Accuracy
Layer 1 (2400)	400	95.58%
Layer 2 (1800)	600	91.17%
Layer 3 (1200)	800	75.08%

The classification result, which uses the same number of features at each layer, is shown in Table 2. The classification result, which uses a different number of features at each layer, appears in Table 3. The experimental results show that the accuracy of layer 3 is lower than that of the other layers. It is due to the similarities in used words, expressions, and structure of the document in rates 2 and 3. From Table 3, we see that the accuracy of the text rating system is improved by increasing the number of features.

The classification results of the text rating system are shown in Table 4 and Table 5. In Table 4 and 5, the Hierarchical text rating system I represents the proposed system which uses a different number of features and the Hierarchical text rating system II means the proposed system which uses the same number of features. The pair-wise text rating system, which selects features from all rate documents and then classifies the rate of the document using pair-wise classifiers and a voting scheme, is used for performance comparison. The text rating system with a non-harmful documents screen [8] is also used for performance comparison. That system consists of a non-harmful documents screen using the pattern matching algorithm and classifier using the SVM model, which assigns a document into one of three rates.

Experimental results show that the proposed system, which uses a different number of features, has the best rating performance for objectionable document classification.

Table 4. Overall classification results of the text rating system

	Accuracy
Hierarchical rating I	82.96 %
Hierarchical rating II	81.67%
Pair wise rating	81.83%
Text rating system with non-harmful document screen	81.25%

Table 5. The classification results of the text rating system

	Rate 0 (600)	Rate 1 (600)	Rate 2 (600)	Rate 3 (600)
Hierarchical text rating system I	94.00%	87.67%	80.83%	69.33%
Hierarchical text rating system II	94.00%	84.67%	73.17%	74.83%
Pair wise text rating system	94.17%	90.33%	63.17%	79.67%
Text rating system with non-harmful document screen	88.33%	92.00%	72.17%	73.17%

5. Conclusion

In this paper, we classified the objectionable texts into four classes according to their objectionability and proposed the hierarchical text rating system for objectionable documents. We used the three layers for the hierarchical system and selected features independently at each layer. Experimental results showed that the proposed system archived an excellent classification performance. We used only the lexical information in documents and did not use the syntactic information. We expect to improve the performance by using syntactic information. More research on the classification of objectionable texts using syntactic information is required. We expect that this system will help to protect children from objectionable documents.

References

[1] M. Fleck, D. Forsyth, and C. Bregler, "Finding Naked People," In European Conf. on Computer Vision, 1996, vol. II, pp.592–602,

[2] C. Ding, C.H. Chi, J. Deng, and C.L. Dong, "Centralized Content-Based Web Filtering and Blocking: How Far Can It Go?" IEEE SMC'99 Conference Proceedings, 1999, pp.115-119

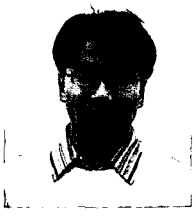
[3] M. J. Jones and J. M. Rehg, "Statistical Color Model with Application to Skin Detection," In Technical Report CRL, 98/11:1--12, 1998.

[4] J. Z. Wang, G Wiederhold and O. Firschein, "System for Screening Objectionable Images," Computer Communications, vol.21:1355--1600, 1998.

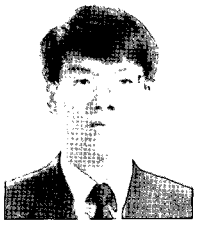
[5] V. Jacob, R. Krishnan, Y.U. Ryu, R. Chandrasekaran and S. Hong. "Filtering Objectionable Internet Content," In Proceedings of the 20th International Conference on Information Systems, pp. 274-278, 1999.

[6] R. Chandrasekaran, , Y. U. Ryu, V. Jacob, and S. Hong, "Isotonic Separation," Working Paper, Depart-

ment of Management Science and Information Systems, School of Management, University of Texas at Dallas, 1998.
[7] M. Hammami, Y. Chahir, and L. Chen, "WebGuard: Web based Adult Content Detection and Filtering System," IEEE/WIC International Conference on Web Intelligence, 2003.
[8] H.G. Lee, Y.S. Kim, C.Y. Jeong, S.W. Han and T.Y. Nam, "Multi level objectionable text classification using SVM and non-harmful document screen," The 4th International Conference on Asian Language Processing and Information Technology, 2005.
[9] G.Y. SU, J.H. LI, Y.H. MA and S.H. LI, "Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model," Journal of Zhejiang University SCIENCE, 2004 Vol. 5 No. 9 pp.1106-1113
[10] Y. Yang and J.O. Pedersen, "A comparative study on feature selection in text categorization," In Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 412-420.
[11] G. Salton and M.J. McGill, "Introduction to Modern Information Retrieval", New York: McGraw-Hill, 1983
[12] K.H. Lee, J. Kay, B.H. Kang, and U. Rosebrock, "A Comparative Study on Statistical Machine Learning Algorithms and Thresholding Strategies for Automatic Text Categorization," In Proceedings of PRICAI, 2002, pp.444-453
[13] VN. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.
[14] T. Joachims, "Making Large-Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning", MIT-Press, 1999.



Chi Yoon Jeong
He received a BS degree in electronic & electrical engineering from POSTECH, Korea, in 2002. He also received an MS degree in electronic & electrical engineering from POSTECH, Korea, in 2004. In 2004 he joined the Electronics & Telecommunications Research Institute, where he is currently a researcher. His current research interests include computer vision and pattern recognition.



Seung Wan Han
He received MS and Ph.D. degrees in computer science from Chonnam National University, Korea, in 1996 and 2001, respectively.
Since 2001, he has been a senior member of the engineering staff in the Information Security Research Division, ETRI (Electronics and Telecommunications Research Institute), Korea. His primary research areas are network security, algorithms, and computation theory. He is currently interested in text and image filtering for blocking objectionable information.

Taek Yong Nam
He received a BS degree in computer science & statistics from Chung Nam National University, Korea, in 1987. He received an MS degree in computer science & statistics from Chung Nam National University, Korea, in 1990. He also received a Ph.D. degree in electronics & information engineering from Hankuk University of Foreign Studies
gn Studies, Korea, in 2005. In 1987 he joined the Electronics & Telecommunications Research Institute, where he is currently a principal researcher. His current research interests include Internet technology, information security, and information classification.