Journal of Korea Multimedia Society Vol. 12, No. 6, JUNE 2009(pp. 867-875)

Context-based classification for harmful web documents and comparison of feature selecting algorithms

Youngsoo Kim[†], Namje Park[††], Dowon Hong[†††], Dongho Won[††††]

ABSTRACT

More and richer information sources and services are available on the web everyday. However, harmful information, such as adult content, is not appropriate for all users, notably children. Since internet is a worldwide open network, it has a limit to regulate users providing harmful contents through each countrie's national laws or systems. Additionally it is not a desirable way of developing a certain system-specific classification technology for harmful contents, because internet users can contact with them in diverse ways, for example, porn sites, harmful spams, or peer-to-peer networks, etc. Therefore, it is being emphasized to research and develop context-based core technologies for classifying harmful contents. In this paper, we propose an efficient text filter for blocking harmful texts of web documents using context-based technologies and examine which algorithms for feature selection, the process that select content terms, as features, can be useful for text categorization in all content term occurs in documents, are suitable for classifying harmful contents through implementation and experiment.

Key words: Text Classification, Harmful Contents, Machine Learning, Feature Selection

1. INTRODUCTION

The World Wide Web is growing ever more rapidly. More and richer information sources and services, such as news, advertisements, consumer information and adult contents are available on the Web everyday. Simultaneously, user communities are becoming increasingly diverse. The openness of the Web allows any user to access almost any type of information. However, some information, such as adult content, is not appropriate for all users, notably children. Additionally for adults, some contents included in abnormal pornographic sites can do ordinary people's mental health harm. However, since the web is an open network linking the world together, it is limited to regulate providers for harmful contents legally or institutionally. The technical approach is the only way to solve the above problems. But it is not desirable to develop system-specific protecting technologies for harmful contents, because people can meet these contents through diverse routes such as adult web sites, harmful spam mails or peer-to-peer networks. There are some products already publicized, but those products have concentrated on IP-based filtering, and their classification of Web sites is mostly manual. However, as we know, the Web is a highly dynamic information source. Not only do many Web sites ap-

※ Corresponding Author : Dongho Won, Address : (440-746) 300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do, Korea, TEL : +82-31-290-7107, FAX : +82-42-290-7686, E-mail : dhwon@security.re.kr

Receipt date : Aug. 27, 2008, Approval date : Oct. 21, 2008

[†] Cryptography Research Team, Electronics and Telecommunications Research Institue
(E-mail : blitzkrieg@etri.re.kr)

[††] RFID/USN Security Research Team, Electronics and Telecommunications Research Institue
(E-mail : parknamja@hotamil.com)

[†††] Cryptography Research Team, Electronics and Telecommunications Research Institue
(E-mail : dwhong@etri.re.kr)

[††††] School of Information and Communication Engineering, Sungkyunkwan University

※ This work was supported by the IT R&D program of MKE/IITA. [2007-S019-02, Development of Digital Forensic System for Information Transparency]

pear everyday while others disappear, but also site content (include linkage information) is updated frequently. Thus, manual classification and filtering systems are largely impractical. The highly dynamic character of the Web calls for new techniques designed to classify and filter Web sites and URLs automatically. Additionally, most of conventional products are only for preventing children from adult contents. In real world, some non-adult sites can contain adult contents like adult education, sex consultation or sex-related gossips. On the other hand, some abnormal adult sites include a lot of objectionable contents being able to do ordinary people's mental health harm. A new criterion is needed like movie world. In that field, all films are released after being graded by their contents and audience's age. In this paper, we propose an efficient text filter for blocking harmful texts of web documents using context-based technologies and examine which algorithms for feature selection, the process that select content terms, as features, can be useful for text categorization in all content term occurs in documents, are suitable for classifying harmful contents through implementation and experiment. It is organized as follows. Chapter 1 is an introduction and we consist of the technology map for text categorization and describe it in chapter 2. In chapter 3, we show the system framework and web documents rating processes, and also we show block composition for implementation and describe each block's role in chapter 4. Additionally, we show experimental results and analyze that which algorithms for feature selection are suitable for classifying harmful contents in chapter 5, and finish it with conclusion in chapter 6.

2. TEXT CATEGORIZATION TECHNOLOGIES

While more and more textual information is available online, efficient managing is not easy. Good indexing and summarization of document content are required and text categorization can be a solution. Text categorization is the problem of automatically assigning predefined categories to free text documents, and can be applied to diverse field such as automatic indexing, document organization, text filtering, word sense disambiguation, or web pages categorization. Text categorization technologies contain category definition, feature selection technologies and text classification technologies. First, category definition is the process that chooses classifying target categories and criteria classifying each chosen categories. This process includes the collecting documents step. Second, feature selection technology is the process that select content terms, as features, can be useful for text categorization in all content term occurs in documents, and is divided into the morphological analysis based feature selection technology and the grammar analysis based feature selection technology. Finally, text classification technology is the step of classifying texts through selected features, and algorithms used in the field of machine learning are used for this technology; rule-based classification, inductive learning-based classification, and similarity-based classification. First, rule-based classification finds rules, learning documents have, which can divide categories and classifies documents using these rules. Second, inductive learning based classification includes Bayesian Probability Model selecting features from documents and using them for probability approach, decision tree model reconstructing a tree structure and deciding categories from the presence of features, and support vector machine representing positive and negative features, generated from learning documents, as vector spaces[1,2]. Finally, similarity-based classification regards a document as query with a view of information retrieval and finds similar documents and includes k-nearest neighbor method and linear classification model. Figure 1 shows the technology map for text categorization.

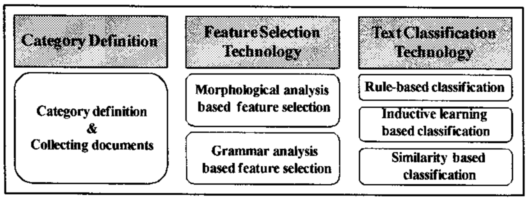
Category Definition
Feature Selection Technology
Text Classification Technology
Category definition & Collecting documents
Morphological analysis based feature selection
Grammar analysis based feature selection
Rule-based classification
Inductive learning based classification
Similarity based classification

Fig. 1. Technology Map for Text Classification

Feature selection is the process that select content terms, as features, can be useful for text categorization in all content term occurs in documents. Since the number of content terms is tens or hundreds of thousands, it takes too much time to learn and classify and has no guarantee of good performances to select all content terms as features. There are studies for measuring amount of information which content terms contain and selecting them have higher amount of information. Feature selection is divided into the morphological analysis based feature selection technology and the grammar analysis based feature selection technology.

2.1 Morphological Analysis based Feature Selection

This method makes lists of features having good quality by extracting candidates of features through morphological analysis, selecting feature terms statistically, and choosing representing features applying word control lists such as thesaurus or authority file. It gives weights to each term on the list of features by judging the degree of importance, for features having the higher weight, so as to have more influence on classification. Generally, feature selecting methods are as follows.

(1) Term Frequency (TF) is the number that a term occurs in all text set. It is the fundamental method and has some variations such as log TF.

(2) Document Frequency (DF) is the number of documents a term occurs in all text set. Usually terms whose DF was less than some predetermined threshold are removed, since they are not considered to be important. The fundamental assumption when we use DF is that rare terms are either non-informative for category prediction, or not influential in global performance.

(3) Mutual Information (MI) uses the amount of information a term has concerning other terms. It represents numerically contributing degree to guess, when an event a term of two occurs, whether the other term occurs or not. If A is the number of times a term t and a category c co-occur, B is the number of time the t occurs without c, C is the number of times c occurs without t, and N is the total number of documents, then the MI between t and c is defined to be

$$MI(t,c) = \log\frac{\Pr(t \wedge c)}{\Pr(t) \times \Pr(c)} \tag{1}$$

and is estimated using

$$MI(t,c) \approx \log\frac{A \times N}{(A+C) \times (A+B)} \tag{2}$$

(4) Information Gain (IG) measures the degree that it has an influence on classification whether a term occurs or not, and selects features having higher degrees. It calculates amount of information gained for all terms and selects terms having higher values than the threshold as features. IG measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. It has better performances than MI in many cases, since it is measured as a sum of MI's average for considering the presence of a term in a document and MI's average for considering the absence of a term in a document. If a set of category is defined as $\{c_1, c_2, \cdots, c_m\}$, then the IG between t and c is defined to be

$$\begin{aligned} IG(t) = & \\ & -\sum_{i=1}^{m}\Pr(c_i)\log\Pr(c_i) + \Pr(t)\sum_{i=1}^{m}\Pr(c_i|t)\log\Pr(c_i|t) \\ & + \Pr(\bar{t})\sum_{i=1}^{m}\Pr(c_i|\bar{t})\log\Pr(c_i|\bar{t}) \end{aligned} \tag{3}$$

(1) χ^2 statistic (CHI) is a method of obtaining a degree of importance by measuring dependency

between a term t and a category c. If the gap between t and c is big, t can be selected as a feature with high possibility. CHI measures differences between each category's occurring distribution and general term's occurring distribution by using document frequency, and chooses terms having higher values than the threshold as features. Usually terms having low frequency are known to be untrustworthy. If A is the number of times a term t and a category c co-occur, B is the number of time that t occurs without c, C is the number of times c occurs without t, D is the number of times neither c nor t occurs, and N is the total number of documents, then the CHI between t and c is defined to be

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{4}$$

To represent collected documents in learning-specific type, weights are given to each feature of documents. That is, each document is represented as a set of values of selected features. Content terms (feature terms) chosen at previous step become features, and term weights become values; <feature: value> expression. General presentation method for documents is the vector space model, which represents a document as a vector using term frequency of features in that document. Even terms with high frequency occur as function terms very often, they are not able to express content of a document, and then a method, gives weights considering both TF and IDF(Inverse Document Frequency) simultaneously, is mostly used. Each feature's weight is represented as multiplication of TF and IDF in a document[3]. If f_{ik} is them frequency of feature k in document i, N is the number of all documents, and n_{ik} is the number of documents feature k occurs, the weight a_{ik} of feature k in document i can be expressed as follows.

$$a_{ik}=f_{ik}\times\log\frac{N}{n_k} \tag{5}$$

TF/ICF(Inverse Category Frequency) is another way of imposing weights, uses category frequency, instead of document frequency. This method gives weights to terms having good ability to classify categories. That is, it imposes high weights to terms occur in minor categories frequently, but gives low weights to terms occur in major categories very often. In case of document classification, since terms can help classifying among categories are very important, TF/ICF is more reasonable method for giving weights than TF/IDF. Meanwhile, TF used for weights has diverse variations such as binary TF, log TF, double-log TF, double-log2 TF and root TF.

2.2 Grammar Analysis based Feature Selection

This method identifies terms or phrases having specific functions through grammar analysis, and uses them as features. Grammar analysis is mainly applied to research for natural language processing, but it is not easy to complete grammar analysis on that field. Comparing to hardness of grammar analysis, the effectiveness of it is not good, so actually it is not possible to implement feature selecting system using grammar analysis including semantic analysis.

3. SYSTEM FRAMEWORK AND OPERATION PROCESSES

Figure 2 shows our system framework and rating processes for web documents. It consists of 5 parts; web-documents collector, preprocessor (morphological analyzer), rule-based text classifier, learn-based text classifier, and harmful URL manger.

(1) Web-Documents Collection is for gathering web documents used at learning or classifying. Web robots visit internet sites and gather all web pages. They are stored at database, after being classified into one of 4 grades according to a new rating criterion proposed in this chapter.

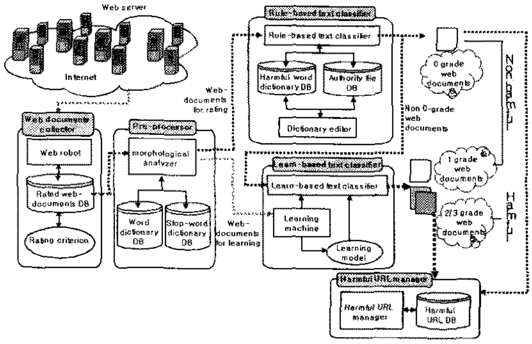

Fig. 2. System Framework and Web Documents Rating Processes

(2) The main part of this function is the morphological analysis. Before doing that, to get some learning samples, we need the pre-processing process. Those web-documents at database contain many HTML tags, so they are HTML-parsed and divided into many morphemes using a morphological analyzer. This process includes deletion process for symbols and stopwords to help the morphological analysis. All web documents at databases need this process for learning and rating

(3) Rule-based Text Classification function extracts non-harmful documents (0-grade web documents) from all collected web documents using a pattern matching algorithm. It decides a document must be non-harmful if it does not have any harmful words. This function has 3 dictionaries; a harmful words dictionary, a stopwords dictionary, and an authority control dictionary[4]. We checked and extracted from 20,000 documents and made words bags to get those dictionaries. The words bags are sorted by frequency and selected by its harmfulness.

(4) Learn-based Text Classification function adapts the svm learning algorithm to classify harmful documents (1-grade, 2-grade, and 3-grade). It is divided into 2 main processes; a learning process and a rating process. A learning process consists of feature selection, indexing, SVM preprocessing[5,6], and generation of learning models. It calculates feature vectors from the result of the morphological analysis. There are some algorithms for selecting features like TF, MI, IG, and c2 statistic[7]. Indexing is for endowing features with weights, because features have different depths of importance at each web documents. We adopted TF/IDF method for giving a weight to each feature. SVM preprocessing part includes normalization for feature vectors and grid search for finding optimal SVM parameters. Finally, a learning model is generated using optimal SVM parameters. A rating part consists of indexing, normalization, and rating. This part is for web documents for rating. They need indexing and normalization for comparing to the learning model generated at learning process. After that, harmful grades are given to those web documents for rating; 1-grade, 2-grade, and 3-grade.

4. IMPLEMENTATION

The proposed text categorization system for harmful web documents was implemented as figure 3 having three blocks; Data Management Block (DMB), Learning Management Block (LMB), and Rating Management Block (RMB). DMB manages harmful dictionaries and sample documents for learning, and has jobs as the preprocessor for functions of LMB and RMB. LMB generates learning models for learn-based text classification and

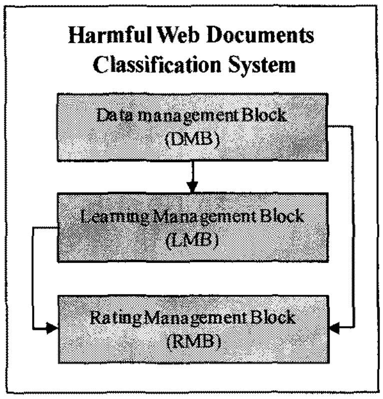
Harmful Web Documents Classification System
Data management Block (DMB)
Learning Management Block (LMB)
Rating Management Block (RMB)

Fig. 3. Block Composition

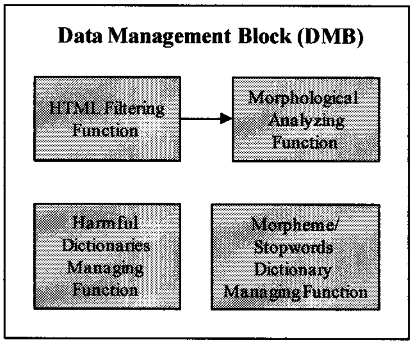
Data Management Block (DMB)
HTML Filtering Function
Morphological Analyzing Function
Harmful Dictionaries Managing Function
Morpheme/ Stopwords Dictionary Managing Function

Fig. 4. Data Management Block (DMB)

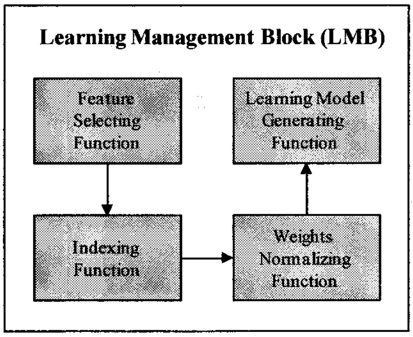
Learning Management Block (LMB)
Feature Selecting Function
Learning Model Generating Function
Indexing Function
Weights Normalizing Function

Fig. 5. Learning Management Block (LMB)

RMB checks harmfulness of texts and decides the final grade for them.

4.1 Data Management Block (DMB)

DMB has the html filtering function for getting ride of html tags showing text structures and the morphological analyzing function for extracting and analyzing Korean/English morphemes from texts. It also has the harmful dictionaries managing function that manages harmful dictionary and authority control dictionary and transforms them into data for rule-based text classification. Additionally, DMB has the morpheme/stopwords dictionary managing function that manages a stopword dictionary and morpheme dictionary and transforms them into data for morphological analysis

4.2 Learning Management Block (LMB)

LMB has the feature selecting function generates a list of features used for learn-based text classification and the indexing function gives a weight to each features to have weights at learning process. It also has the learning model generating function and the weights normalizing function maps all weights to values between -1 and 1 to promote accuracy for classification.

4.3 Rating Management Block (RMB)

RMB has the rule-based text classification function filters non-harmful documents by checking if documents contain harmful words or not. It also has the learn-based text classification function gives grade 1, grade 2 or grade 3 to potential harmful documents using a learning model. Additionally, RMB has the final grade deciding function decides a document's last grade making use of the rule-based text classification function and the learn-based text classification function.

5. EXPERIMENTS AND RESULTS

In this section, we test the proposed harmful documents classification system using huge amount of collected harmful or non-harmful web documents, and analyze the results.

5.1 Experimental Environments

After collecting over 5 thousand Korean or English web documents, we gave grades to them according to the grading criteria of web documents

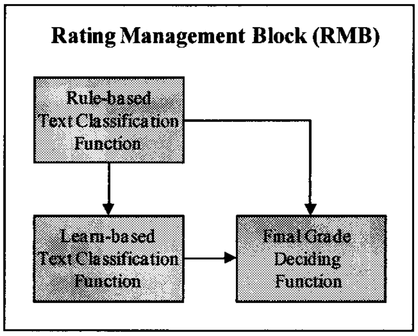
Rating Management Block (RMB)
Rule-based Text Classification Function
Learn-based Text Classification Function
Final Grade Deciding Function

Fig. 6. Rating Management Block (RMB)

Table 1. Grading criteria of web documents

Harmful	Grade	Web site Category	Description
Non-Harmful	0	Non-harmful sites	Not including any harmful contents
	1	Medical sites (Urology, Obstetrics and Gynecology)	Informative for adult or younger generation
		Sex-related consulting sites (webboard)	Consulting for sex-related information for younger generation
		Sex education/Sex clinic sites for the younger generation	
		Sex-related articles at entertainment/sport pages in newspaper sites	Dealt with sex-related contents as gossips
Harmful	2	Sex education/Sex clinic sites for an adult	Consulting for sex-related skills or troubles for an adult
		Sex-related consulting sites (webboard)	
		Normal sex-related stories	Describing normal sex scenes
	3	Checking pages for adult	Front pages for adult sites
		General adult sites	Common adult sites such as playboy.com
		Adult Video Reviews	Usually containing sex-related images
		Abnormal sex-related stories	Describing sex scenes abnormally

Table 2. Numbers of web documents for learning and testing

	For learning		For testing(Classifying)	
	Non-harmful	Harmful	Non-harmful	Harmful
Korean web documents	588	1164	452	509
English web documents	694	936	533	449

Table 3. Description of criteria for evaluating performance

		Proposed System	
		Harmful	Non-Harmful
Real world	Harmful	A	C
	Non-Harmful	B	D

(Table 1). Table 2 shows numbers of web documents which are used for learning and testing.

5.2 Criteria for evaluating performances

We used the following 4 performance criteria to evaluate our experimental results for text classification[8].

(1) Accuracy is the ratio of the number of data which is classified rightly. It can be calculated to (A+D)/(A+B+C+D) at Table 3.

(2) Recall is the ratio of web documents which proposed system decided as harmful in real harmful web documents. It is obtained to calculate A/(A+C) at Table 3, is irrelevant to non-harmful web documents.

(3) Precision is the ratio of the number of real harmful web documents in web documents proposed system decided as harmful and can be calculated to A/(A+B) at Table 3. As a matter of fact, the more the number of web documents proposed system decided as harmful in real non-harmful ones (B at Table 3) are, the lower precision is. Therefore, the system that both precision and recall are high is the best one.

(4) F-measure is the harmonic mean of recall and precision and can be calculated to (2*recall*precision)/(recall+precision).

5.3 Experimental Method

We generated the Korean and English learning models using over 3,000 web documents and measured performances of 4 evaluating criteria inputting test web documents to the proposed system. We made 84 learning models through tuning the following diverse parameters.

(1) Number of Features is 200, 300, 400, 500, 600, 700, 800

(2) Feature Selection Algorithms are TF(Specially we used logTF), IG and c2 statistics

(3) Indexing Algorithms are TF/IDF and TF/ICF

(4) Languages are Korean and English

5.4 Analysis of Experimental Results

In case of Korean web documents, highest accuracy and f-measure are 91.97% and 94.05% each when we use CHI and TF/IDF algorithms with 800 features. Additionally, even though the number of

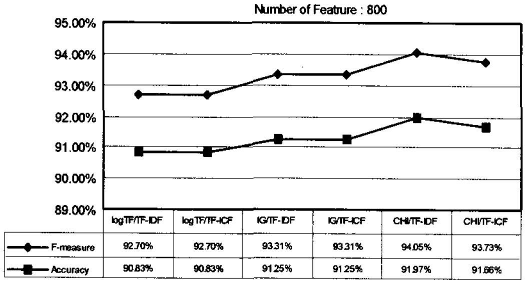

Fig. 7. Comparison of performances (Korean web documents)

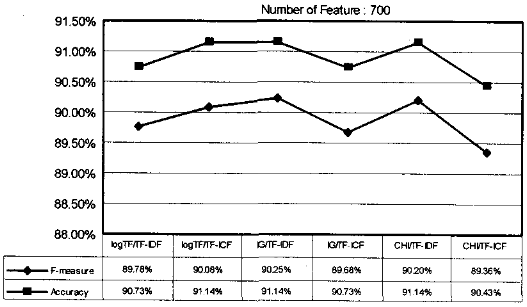

Fig. 8. Comparison of performances (English web documents)

features is tuned, performances are mainly high, specially when we adapt learning models using CHI and TF/IDF algorithms. Figure 7 depicts a performance comparison of each algorithm with 800 features for Korean web documents.

In case of English web documents, highest accuracy and f-measure are 91.34% and 90.46% each when we use CHI and TF/ICF algorithms with 500 features and CHI and TF/IDF algorithms with 600 features. However, the more features are used, the higher performances are when we adapt learning models using CHI and TF/IDF algorithms and IG and TF/IDF algorithms. Figure 8 depicts a performance comparison of each algorithms with 700 features for English web documents.

6. CONCLUDING REMARKS

We proposed an efficient harmful text classification system using pattern matching and machine learning method hierarchically. Composing the technology map for text categorization, we described diverse text categorization technique to classify web documents and showed a method of filtering them step by step. Additionally, we implemented the proposed system and evaluated performances by tuning various parameters after setting evaluating criteria. As we became to know through previous sections, we can get different performances by tuning related parameters such as number of features, method for selecting features, or indexing method. In case of Korean web documents, highest accuracy and f-measure are 91.97% and 94.05% each when we use CHI and TF/IDF algorithms with 800 features, and even though the number of features is tuned, performances are mainly high, specially when we adapt learning models using CHI and TF/IDF algorithms. In case of English web documents, highest accuracy and f-measure are 91.34% and 90.46% each when we use CHI and TF/ICF algorithms with 500 features and CHI and TF/IDF algorithms with 600 features, but, the more features are used, the higher performances are when we adapt learning models using CHI and TF/IDF algorithms and IG and TF/IDF algorithms. However, both cases' gap of performances are small, so it is very difficult to select the best parameters to get the highest performance for specific application, harmful web document classification. Additionally, since testing results can be different by changing number of web documents or contents of them used for learning or testing, continuous collecting web documents as learning samples and generating learning models are only way to get confidence of performances. We are expecting that performance of the proposed system is gradually increasing if intelligence is improving through steady feedback.

REFERENCES

[1] F.Sebastiani, "Machine Learning in Automated

Text Categorization," *ACM Computing Surveys*, Vol.43, No.1, pp. 1-47, 2002.

[2] C.Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, Vol.2, pp. 121-167, 1998.

[3] W.Frakes and R.Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992.

[4] Thesaurus, *Wikipedia: the Free Encyclopedia*, http://en.wikipedia.org/wiki/Thesaurus,2008.

[5] G.Siolas and F.d'Alche-Buc, "Support Vector Machines based on a Semantic Kernel for Text Categorization," *Proceeding of IJCNN 2000*, Vol.5, pp. 205-209, 2000.

[6] Support Vector Machine, *Wikipedia, the free Encyclopedia*, http://en.wikipedia.org/wiki/SVM,2005.

[7] Y.Yang and J.Pederson, "A Comparative Study on Feature Selection in text Categorization," *Proceedings of the 14th International Conference on Machine Learning*, pp.412-420, 1997.

[8] S.Kang, *Korean Morphological Analysis and Information Retrieval*, Hongreung Science Press, 2002.

Youngsoo Kim

Youngsoo Kim received the B.E. and M.E. degrees in Information Engineering and Electronics & Computer Engineering from Sungkyunkwan University, Seoul, Korea, in 1998 and 2000, respectively. He is currently a PhD. Candidate of Computer Engineering in SungkyunkwanUniversity. Since Feb 2000, he has been a senior member of engineering staff in ETRI (Electronics and Telecommunications Research Institute), Korea. His research interests include cryptography and information security.

Namje Park

Namje Park received the B.S. degree in Information Industry from Dongguk University, Korea in 2000 and the M.S. and Ph.D degrees in Information Security and computer engineering from Sungkyunkwan University, Korea, in 2003 and 2008, respectively. Since 2003, he has been a senior member of engineering staff in ETRI (Electronics and Telecommunications Research Institute), Korea. His current research interest is in the area of cryptography and information security.

Dowon Hong

Dowon Hong received the B.S., M.S. and Ph.D. degrees in Mathematics from Korea University, Seoul, Korea, in 1994, 1996, and 2000, respectively. Since 2000, he has been a senior member of engineering staff in the Cryptography Research Team of ETRI (Electronics and Telecommunications Research Institute), Korea. His research interests are in digial forensic analysis, computer security and cryptography.

Dongho Won

Dongho Won received the B.E., M.E. and Ph.D. degrees from Sungkyunkwan University, Seoul, Korea, in 1976, 1978, and 1988, respectively. After working at ETRI (Electronics and Telecommunications Research Institute) from 1978 to 1980, he joined Sungkyunkwan University in 1982, where he is currently Professor of School of Information and Communication Engineering. His interests are on cryptology and information security. Especially, in the year 2002, he was occupied the president of KIISC (Korea Institute of Information Security & Cryptology).