International Journal of u- and e- Service, Science and Technology
Vol.8, No. 10 (2015), pp.11-20
http://dx.doi.org/10.14257/ijunesst.2015.8.10.02

Classification Technique for Filtering Sentiment Vocabularies for the Enhancement of Accuracy of Opinion Mining

Ji-Hoon Seo, Ho-Sun Lee and Jin-Tak Choi

*Incheon University, 119 Academy-ro, Yeonsu-gu, Incheon, Republic of Korea,
sserz@inu.ac.kr, hilhs21@hanmail.net,choi@inu.ac.kr*

Abstract

This thesis, as part of the creation of a text-mining-based sentiment dictionary to be applied in the Korean grammar structure, solves the problem of the enhancement of accuracy of opinion mining data by applying the filtering model of candidate sentiment vocabularies. The fact that the reliability of sensitive vocabularies shows huge variances according to the filtering modeling method applied has become a decreasing factor for the accuracy of the vocabularies in the opinion mining process, which is attributable to the fact there isn't a success factor in the filtering modeling standard for precise selection of vocabularies. In this thesis, a filtering model of positive and negative vocabularies on candidate Korean sentiment vocabularies and a reliability scale for accuracy were suggested to solve such problems by applying the semi-structured data filtering model for the selection of candidate sentiment vocabularies of the Korean grammar. The study has confirmed through relevant performance assessment when filtering were applied in relation to 30%, 50% and 60% respectively with regard to candidate sentiment vocabularies upon collecting vocabularies obtained via sentence segmentation and classification into positive and negative vocabularies that exceptional accuracy of the opinion sentiment dictionary was shown via the 60% filtering.

Keywords: *Filtering, Text Mining, Opinion Mining, Prediction, Classification*

1. Introduction

F Aside from the gradual development being made in data mining technology today, the technology is increasingly being used as a tool for psychological strategies and marketing analyses on the basis of analyses on semi-structured texts and text mining, and studies on opinion mining are also being conducted through new discoveries made [1, 2]. Opinion mining, as a study that makes judgments on the assessments made on emotional conditions expressed in the language form by extracting the units of positivity and negativity with the use of the sentiment dictionary assorted from text documentations and sentiment vocabularies can be expressed by humans and conducting comparisons with relevant grammar, is similar with text mining, but is clearly a different technology, for analyses are made with sentiment analyses of the vocabularies of humans [3]. The creation of a reliable sentiment dictionary is considered to be a significant factor for increasing the prediction accuracy of opinion mining, and differences may be shown in the data performance based on the filtering method used for the candidate sentiment vocabularies. In the case of Korean grammar, this is a criterion which can increase the prediction accuracy of opinions for the selection of regular sentiment vocabularies in the sentiment candidate group. An optimum filtering method on performance models of sentiment vocabularies in the Korean grammar for the enhancement of accuracy were suggested in this thesis to solve such problems.

ISSN: 2005-4246 IJUNESST

2. Related Work

Opinion mining technology, having advanced the data mining a step further, is developing to be the infrastructure of the NLP (Natural Language Processing) integrated convergence technology and for the most part, is being utilized in managerial consumer marketing and reputation analyses [4, 5]. In spite of the fact that we are approaching nearer to the next generation and future promising technologies capable of collecting information of greater value by analyzing users' of various clients with developments made in internet technologies in recent years, various factors, such as accumulated text data and objective individual sentiment, are entwined together until the creation of a sentiment dictionary is reached, and it continues to show its limitations in the applications of logics and rules of grammar for all languages [6, 7].

2.1 Filtering of the Candidate Sentiment Vocabularies Set

The method in which only the top 20% are extracted based on the importance of the vocabularies, number of documentations and how frequently the vocabularies have appeared in semi-structured texts is generally used in the studies on filtering for the construction of a sentiment dictionary of opinion mining, and cases where the top 15% of the entries and lower 15% of the entries of important vocabularies are used were also suggested. In brief, a clear standard of the opinion process was not set and critical success factors of filtering were not present in these cases [8]. Therefore, the selection range of filtering is determined according to the environmental factors regarding the number of vocabularies and amount of information of the documentation in the selection of vocabularies.

2.2 Methodology on Data Filtering

Up to the present, various filtering methods have been introduced through studies that were based on the morpheme analyses of the semi-structured data. The universal scales presently being used are, one, the method in which the top 20% of the higher significance and weighted value are used, and two, the method in which the top and bottom 20% are used and the remainder are excluded. However, candidates of sentiment vocabularies for applying the opinion by the characteristic of the text documentation, context of the documentation and type of language have sensitive effects on the filtering, and thus variable result values are gained. Consequently, a methodology for seeking an optimum filtering method of the semi-structured data was suggested in this thesis.

3. Proposed Method

In the selection of candidates of sentiment vocabularies to be included in the entry groups of the sentimental dictionary of the Korean grammar being suggested in this thesis, the level of importance of vocabularies are calculated by using the analyses of morphemes in accordance with the properties of the vocabularies, and the level importance of tokenized vocabularies are calculated by using TF-IDF. 40% of the vocabularies of higher importance among the 100% are included in the candidate group of sentiment vocabularies.

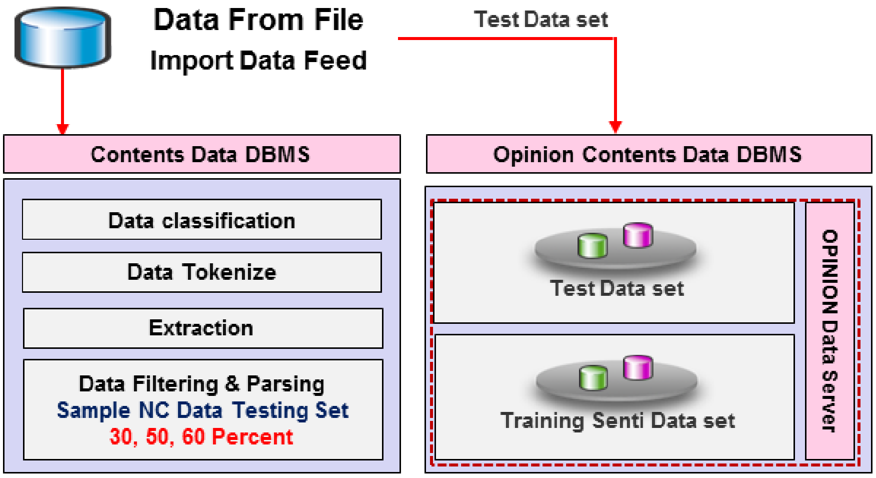

Figure 1. Training Data Derivation

The thesis had employed a multi-server type system, in which a server for the texts and servers required for the work process was constructed for the assortment of semi-structured texts.

3.1 Source Data Type

Collection and classification of the source data are critical parts for the creation of an opinion sentiment dictionary. Candidate sentiment vocabularies are determined by the weighted value of vocabularies, and this, with vocabularies of special characteristics, serves as one of the factors that have effect on the sentiment dictionary. For the initial data, contents such as online news articles, general information on finance and et cetera were collected and assigned. Following the assignment of the target website, themes of the text page items were selected for the extraction of the sample data from the target website, and assignment was made in a way that only the desired data can be collected by sorting out the paths of the target website as a way to remove the nonspecific information in the collection process.

3.2 Morpheme Analysis of the Sentence

The semi-structured data contents used in this thesis were collected on the basis of news data extracted from online finance websites within a two month time frame in which national holidays were excluded, and a total of 3,154 documents were collected for morpheme analysis. Morpheme analysis is a procedure in which meaningful grammatical or relational meanings are tokenized in the smallest units and refers to the minimum semantic element which cannot be analyzed to any further extent. Moreover, lexical analysis is indicative of the analysis of input character strings with a word dictionary as default and use of words in the dictionary and requires an accuracy of a more valid analysis in bigger units than typical morphemes.

In spite of the fact that applications of lexical analyses are commonly used in the text mining in reality, the intent of natural language analyses are met with morpheme analyses. Therefore, this thesis has carried out morpheme analyses prior to the sentiment analysis process of the data with the purpose to increase the degree of accuracy on filtering of vocabularies, and the tokenization process was carried out to tokenize the vocabularies into the smallest forms of strings. Syntax analyses were not carried out in this thesis, for greater emphasis was made on the discovery of the number of appearances of vocabularies which have come about in each document and data on to what degree the tokenized vocabularies were showing their appearance ratio in one day via tokenization were required for the creation of a sentiment dictionary. In the case of plain texts, seeking positivity and negativity appropriate to the document is more efficient than syntactic analyses, for, despite clear realization of sentiment language, the scopes in which vocabularies can be used as sentiment vocabularies are different.

3.3 Filtering of Tokenized Vocabularies in the Document

Vocabularies generated on each document fall under the category of preliminary vocabularies in the candidate sentiment vocabularies. In order that accuracy of opinion is increased by creating sophisticated sentiment vocabularies via filtering, the following rules apply.

First, in order to increase the accuracy of vocabularies in the document extracted from online news articles, 42,358 words which have appeared the most have been extracted by applying the association rule, and nouns amongst words extracted which maintain close link with positive and negative units were treated as important words for partial inclusion into candidate sentiment vocabularies.

Second, all vocabularies being recognized as one word were eliminated during the filtering process, and verbs and antonyms which have a high level of association between two words were extracted.

Third, for the reason that online news articles were data written by professional journalists which are above the average standard of ordinary replies and comments, slangs and unspecified terms did not appear. However, numbers, special symbols, neologism and abbreviations with attributes of internet language were excluded for the accuracy of the data. For instance, although the word was an internet neologism or abbreviation officially registered on Wikipedia or encyclopedia, the word was included in the target for elimination, for classifications were required to be made in the Korean grammar structure with the use of parts of speech, that is, noun (N) and noun (N) + noun (N) and thus will have a low reliability when tokenization of the word is carried out.

3.4 Data Feature Extraction

In this thesis, TF-IDF was used for the extraction of features of semi-structured text documents. Definite articles and conjunctive particles appear frequently in the English grammar and such words intrinsically show a high appearance frequency in the sentences. However, they cannot be interpreted as words with unique characteristics that link the context of the relevant document. In order to accurately find words that clearly express the characteristics of the document, instead of extracting words typically generated, for instance, articles, conjunctive particles and et cetera, unique words that do not appear in most cases while regularly appearing for representing the characteristics of the relevant document are required to be extracted. Such issue is no exception in the natural language process of the Korean language. Due to the principle of conjunctive particles in the Korean unit, which is used regularly and frequently, data becomes less accurate and the problem of reliability in considering the features of the sentence is established. In this thesis, the following feature extraction procedure is carried out to solve such issues.

First, the original data of online news contents on stocks and securities are classified into information data with an interval of one day. This process is a sort procedure for grouping documents via regular classifications and creating document sets for the application of IDF.

Second, the frequency of words in sentences of each document is calculated on the basis of documents with groupings formed, and the derived data are classified into the respective document.

Third, saved vocabularies in respective document are processed with the use of TF-IDF, and the processed data of completed document sets are integrated into one. tf value is calculated to suggest the result of TF-IDF of each vocabulary with respect to n number of document sets.

$$tf_{Docu(N),word} = \frac{Appearance\ frequency\ of\ each\ word}{Apperance\ frequency\ of\ total\ vocabularies\ of\ the\ document\ set}$$

The following equation is established for the idf value relative to the document set by using the natural logarithm.

$$idf_{word} = log \frac{Total\ document\ set + 1}{Number\ of\ corresponding\ vocabularies\ included\ in\ total\ document\ set}$$

The value gained with the addition of integer 1 to the total document set can prevent the idf value from becoming 0 when a certain vocabulary is included in all areas of the document set and thus the total document set and number of corresponding vocabularies included in the entire document set are made equal, and TF-IDF is calculated by using such a method.

3.5 Execution Process of the Vocabulary Classification Model

The two classification models of Korean vocabularies suggested in this thesis are as follows;

First, on the basis of the entire sentence, the grammar of the conjunctive particles by antonyms was considered. Calculation is made with conjunctive particles as a reference point on whether or not the grammatical structure is a positive or negative conjunctive particle, which is determined by the sentiment vocabulary positioned before and after the conjunctive particle.

Second, the relationship between homonyms and multiple meaning words was distinguished to clearly set apart vocabularies falling under the category of verb construction and noun structure.

3.6 Data Filtering

For data filtering, vocabularies of higher importance among the 100% consisted of positivity and negativity were selected as candidate sentiment vocabularies. The construct of filtering can be subdivided into three-level divergence points. First, vocabularies in the top 60% in the selected group are selected as sentiment vocabularies and the remaining 40% of candidate of sentiment vocabularies are eliminated. Second, vocabularies in the top 50% in the selected group are selected as sentiment vocabularies and the bottom 50% are removed. Third, vocabularies in the top 30% in the selected group are selected as sentiment vocabularies and the bottom 70% are eliminated.

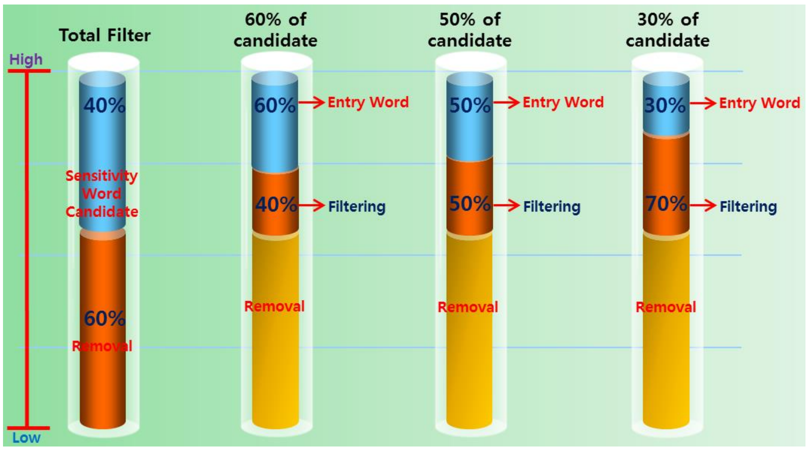

Figure 2. Data Filtering Rule

3.7 Opinion System

The opinion system architecture to be applied on the filtering modeling suggested in this thesis is classified into six phases. Firstly, collection of data is carried out through crawling, and input is made into the database's storage. Secondly, text documents are segmented via the data filtering process to apply the filtering method being suggested in this thesis. Thirdly, tagging is executed for positive and negative vocabularies by the construct of API via opinion mining. Fourthly, rules are applied by assorting the adversative conjunctions and antonyms to be applied in the Korean grammar structure. Fifthly, classify sentiment dictionary's vocabularies applied in the filter and construct an integrated sentiment dictionary accordingly. Sixthly, determine reliability and accuracy of unstructured data by using the sentiment dictionary built using filtering. The architecture flow chart of the opinion analysis system is as follows.

Table 1. Flow Chart of Opinion API Architecture

phase	procedure
STEP 1. Data collection and input	· Collection of data contents. · Import NC Data and save collected data at DBMS.
STEP 2. Data contents DBMS (Data filtering process)	· Classification process of the collected data and tokenization of vocabularies. · Extraction of filtered data and execution of data filtering process. · Sample data filtering with 30%, 50% and 60% settings.
STEP 3. Opinion contents API	- Positive data · Distinction and importation of positive data. · Selection of candidate positive data and tagging of positive vocabularies. - Negative data · Distinction and importation of negative data. · Selection of candidate negative data and tagging of negative vocabularies.
STEP 4. Algorithm	· Application of antonym algorithm of positive and negative data. · Resolution algorithm of conjunctive particles and adversative relations of the Korean grammar. · Enhancement of the Korean grammar and derivation of emotional quotient.
STEP 5. Data union all setting	· Creation of an integrated negative and positive opinion sentiment dictionary & settings for the training data.
STEP 6. Data contents DBMS	· Settings for the source data(test data) and training data. · Join data & opinion analysis. · Data contents analysis, month data analysis, filtering data analysis. · Analysis of reputation data & performance.

4. Performance Evaluation

Following are the domains of vocabularies generated using the 30%, 50% and 60% filtering technique, respectively.

a. Total count(TC) of the total number of entire vocabularies generated per hour frame by applying the opinion of the source data

b. Document count(DC) of number of vocabularies generated per document

c. Word Count(WC) which tallies all the words written within the text per hour frame and accounts any overlapping word as one

Test results have shown 1,655 negative total counts (TC) and 3,366 total counts (TC), which is the number of positive appearance of the entire data, when opinion sentiment dictionary was created with the use of the top 30% filtering technique of the collected data. Consequently, a total of 239 negative vocabularies and 293 positive vocabularies were generated.

When data generated from the top 50% filtering technique were applied, positive total count (TC) was 4,337 and negative total count (TC) was 1,907, and a total of 334 negative vocabularies were used and a total of 485 positive vocabularies were used.

Lastly, twice more vocabularies were recognized than that of the data with the 50% application and a total of 881 positive vocabularies and 173 negative vocabularies were used when the filtering of the top 60% was applied. The study did not find huge differences in the extraction of opinion sentiment vocabularies when 30% and 50% data filtering were applied, and higher opinions were able to be extracted when data was applied with a 10% increase from 50% to 60%.

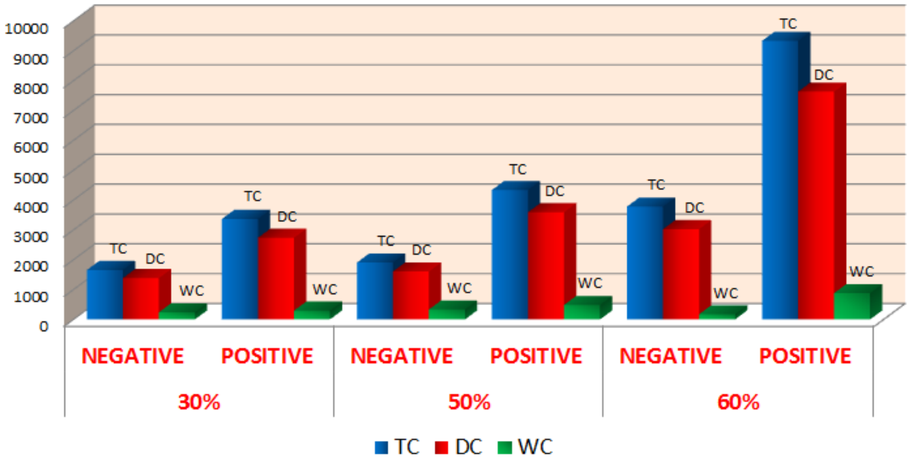

Figure 3. Test for Drawing the Opinion Accuracy with Applications of Filtering Phases

This thesis executes tests on precision, recall and accuracy as a way to verify the performance reliability on various filtering.

In the field of information retrieval and fields related to pattern recognition that make use of binary classification, binary classification is referring to the ratio of the output classified as relevant in the search result, and recall signifies the ratio of items that have actually been searched amongst items classified as relevant. In general, this is a verification method commonly used in studies that create a sentiment dictionary on opinions with statistical information as its target and make comparisons of their prediction accuracies.

The result of the performance evaluation carried out has shown that similar accuracy prediction values were produced when 30% and 50% filtering were applied and thus does not affect the performance model to a significant extent, and there was a 4% increase in accuracy and precision respectively when 60% filtering method was applied which in turn produced a calculation of 98, 97% for their prediction accuracy. For recall value, the case of the 30% filter has produced 86% and a similar value of 85% for the case of 50% filter. However, the 60% filter has produced a performance model of 97%.

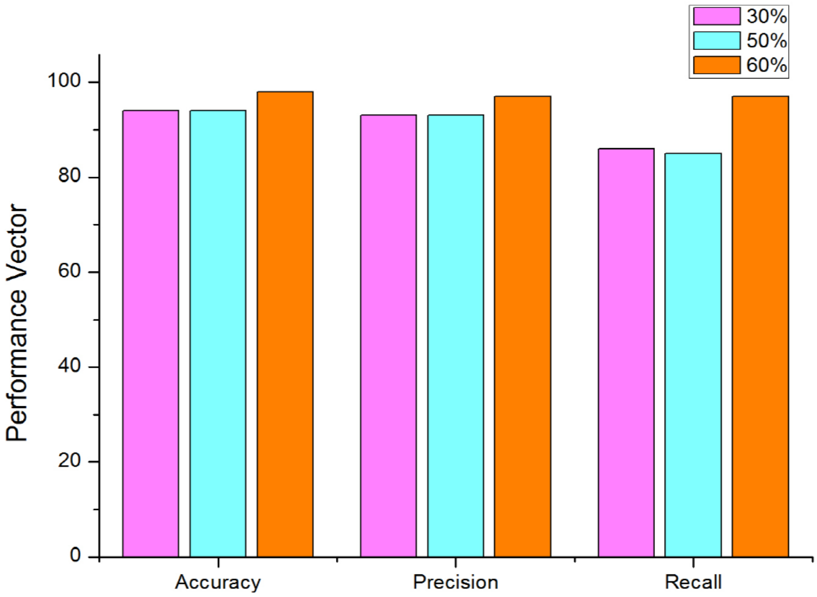

Figure 4. Performance Evaluation

The overall results of the tests have shown there was no difference in the degree of sense of beauty when 30 ~ 50% filtering were applied in the semi-structured data filtering technique, and accuracy of the words and reliability of opinions were largely influenced in the 50~60% filtering.

5. Conclusion

This thesis had suggested an optimum filtering method suited for candidate sentiment vocabularies groups for the enhancement of accuracy prior to generating a sentiment dictionary of opinion mining in the Korean grammar structure, and while the existing filtering process of the opinion text document tended to show differences in their data on the performance model used due to the fact that criteria for calculation and suggestions based on evidence were imprecise, the study has confirmed that performance model with respect to the 60% filter of the entire candidate sentiment vocabularies acquire exceptional performance.

However, customized studies on the natural language process of sentiment dictionaries appropriate to the Korean grammar are required, and the fact that an integrated system, such as SentiWordNet, cannot be applied and can only be applied in themes of specialized areas still remain as restrains for the case of a Korean sentiment dictionary. On that account, there is the need to create an integrated sentiment dictionary in which sentiment data will be utilized for the enhancement of prediction accuracy and reliability of opinions.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (NRF-2014R1A2A1A11050678)

References

[1] P. G. Ipeirotis and A. Sundararajan, “Opinion Mining Using Econometrics : A Case Study on Reputation System” Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, vol. 2, **(2007)**, pp. 416-423.

[2] P. Russom, Ibm Tdwi Research, "Big Data Analytics", Fourth Quarter, **(2011)**.

[3] Chen and D. Zimbra, “AI and Opinion Mining”, IEEE Intelligent Systems, vol. 25, no. 3, **(2010)**, pp. 74-80.

[4] A. Mittermayer and G. F Knolmayer, “Text Mining Systems for Market Response to News: A Survey”, The Institute of Information Systems, University of Bern, Switzerland, **(2006)**.

[5] B. Liu, M. Hu and J. Cheng, “Opinion observer: analyzing and comparing opinions on the Web”, Proceedings of the 14th International Conference on World Wide Web, New York, USA, **(2005)**, pp. 342-351.5.

[6] R. Narayanan, B. Liu and A. Choudhary, “Sentiment analysis of conditional sentences”, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, **(2009)**, pp. 180-189.

[7] S. J. Hoon, “Design of Opinion Sensitivity Dictionary Model for Big Data Management”, **(2015)**.

[8] H. Ming, N. Wenying and L. Xu, “An improved decision tree classification algorithm based on ID3 and the application in score analysis”, Chinese Control and Decision Conference (CCDC), **(2009)**, pp. 1876-1879.

[9] E. Courses and T, Surveys, “Using Sentiment SentiWordNet for multilingual sentiment analysis”, IEEE 24th International Conference on Data Engineering Workshop 2008, Cancun, Mexico, **(2008)**, pp. 507-512.

Authors

Seo-Ji Hoon, Received his Bachelor degree in 2008 at Seoul National University of Science and Technology from department of Safety Engineering. He finished his MS and Ph.D. at Incheon National University from Department of Computer Science and Engineering in 2010 and 2015 respectively. His research interest includes Data Mining, Database Management and Sensor Networking

Ho-Sun Lee, He received the M. S. and Ph. D. degrees in civil engineering from University of Incheon, Incheon, Korea in 2004 and 2010, respectively. He is currently a Director of the Smart Water Grid Research Group at University of Incheon, Incheon, Korea, in 2012. His research interests are in hydraulic and water quality modeling in stormwater and smart water grid.

Choi-Jin Tak, Received his B.S. degree in Mathematics and his M.S. degree in Computer Science from Dongkuk University, Seoul, Korea, in 1977 and 1982, respectively. He received Ph.D. degree in Electronics from Kyunghee University, Seoul, Korea, in 1991.Since 1987, he has been a Faculty Member at the Department of Computer Science of Incheon National University. His research interests include cryptography, database systems, and mobile and distributed computing