nternational Journal of Fuzzy Logic and Intelligent Systems, vol. 6, no. 2, June 2006 pp. 105-109

Analyzing the Effect of Lexical and Conceptual Information in Spam-mail Filtering System

Sin-Jae Kang, Jong-Wan Kim

School of Computer and Information Technology, Daegu University

Abstract

In this paper, we constructed a two-phase spam-mail filtering system based on the lexical and conceptual information. There are two kinds of information that can distinguish the spam mail from the ham (non-spam) mail. The definite information is the mail sender's information, URL, a certain spam keyword list, and the less definite information is the word list and concept codes extracted from the mail body. We first classified the spam mail by using the definite information, and then used the less definite information. We used the lexical information and concept codes contained in the email body for SVM learning in the 2nd phase. According to our results the ham misclassification rate was reduced if more lexical information was used as features, and the spam misclassification rate was reduced when the concept codes were included in features as well.

Key Words : Information filtering, spam-mail filtering, thesaurus, concept information

1. Introduction

With the popularization of the Internet, low cost, and fast delivery of message, email has become an indispensable method for people to communicate each other. Though email brought us such huge convenience, it also caused us trouble of managing the large quantities of spam mails received everyday. Spam mails, which are unsolicited commercial emails or junk mails, flood mailboxes, exposing young people to unsuitable content, and wasting network bandwidth [1]. Most software for email clients provides some automatic spam mail filtering mechanism, typically in the form of blacklists or keyword-based filters. Unfortunately constructing these lists and filters is manual time-consuming process, and is not perfect for a variety of cases in real situation.

The spam filtering problem can be seen as a particular case of the text categorization problem. Several information retrieval (IR) techniques are well suited for addressing this problem, and in addition it is a two-class problem: spam or non-spam. A variety of machine learning algorithms have been used for email categorization task on different metadata [2, 4, 5, 6]. Sahami et al. [2] focuses on the more specific problem of filtering spam mails using a Naïve Bayesian classifier and incorporating domain knowledge using manually constructed domain-specific attributes such as phrasal features and various non-textual features. In most cases, support vector machines (SVM), developed by Vapnik [3], outperforms conventional classifiers and therefore has been used for automatic filtering of spam mails as well as for classifying email text [4, 5]. Yang et al. [6] demonstrate that Naïve Bayesian and SVM classifier is by far superior to TFIDF. In particular, the best result was obtained when SVM was applied to the header with feature subset selection. Accordingly, we can conclude that SVM classifier is slightly better in distinguishing the two-class problem.

In this paper, a two-phase filtering system for filtering spam mails based on lexical and conceptual information is given. In the first phase, definite information such as sender's URL, email addresses, and spam keyword lists is applied. In the second phase, remaining, that is, unclassified emails are classified using less definite information, extracted not only from email header and body but also by fetching web pages. Through these phases, we will analyze the effect of lexical and conceptual Information in spam-mail filtering system.

Manuscript received Nov. 11, 2005; revised Mar. 6, 2006.

2. Training Phase

To extract features or attributes about spam mail filtering and use efficiently, we divide them into two kinds of information: definite information and less definite information. The processing flow of training phase is shown in Fig. 1 and Fig. 2.

2.1 Definite Information

Definite information for filtering spam mails is sender's information, such as email and URL addresses, and definite spam keyword (including key phrase) lists, such as "porno," "big money" and "advertisement" (see Table 1). If an incoming email contains one of sender's information, it has a very high probability of being a spam mail. Therefore, we can regard the email as a spam mail. If an incoming email contains one of definite spam keyword lists in the subject line, or if definite spam keywords are appeared in the body of incoming emails over three times, it is also regarded as a spam mail. We extract sender's information automatically from spam mail corpus, and select definite spam keyword lists

manually.

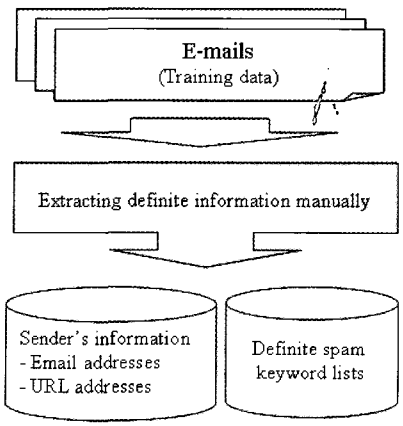

Fig. 1. Training process for 1st phase

Table 1. Examples of definite information

Definite Information	Examples
Sender's email address	minjizbfzzzzg@hotmail.com, kiss@korea.com, robot@helloluck.com, ssycapital_loan1500@empal.com, ...
Sender's URL	sexybra.net, www.09king.com, www.1004movie.com, www.asloan.co.kr, wkwkwk.wo.to, ...
Definite Spam Keywords (or Key Phrases)	advertisement, porno, swapping, big money, credit card loan, no-guarantee loan, ...

2.2 Less Definite Information

There are many particular features of email, which provide evidence as to whether an email is spam or not. For example, the individual words in the text of an email (not included in definite spam keyword lists), domain type of the sender (e.g. co, com), receiving time of an email, or the percentage of non-alphanumeric characters in the subject of an email are indicative of a spam mail [2]. This phase is based on feature vector space representation, in which each feature in the email document mainly corresponds to a single word or thesaurus concept and then uses the SVM learning algorithm to classify the emails. Since the body of a spam mail has little text information (recently, it often has only image data), our system follows hyperlinks contained in the email, fetches contents of a remote webpage, and regards this webpage as extended email body. For extracting textual information (words or phrases in the email), that is hints, each email document is preprocessed to remove symbols, performed morphological analysis, and removed stopwords. Among the extracted features, those with low differential power are not helpful in spam filtering, and thus should be omitted from feature vectors. Feature selection involves searching through all possible combination of features in the candidate feature set to find which subset of features works best for prediction. To select features having high discriminating power, we used the fuzzy inference method [7].

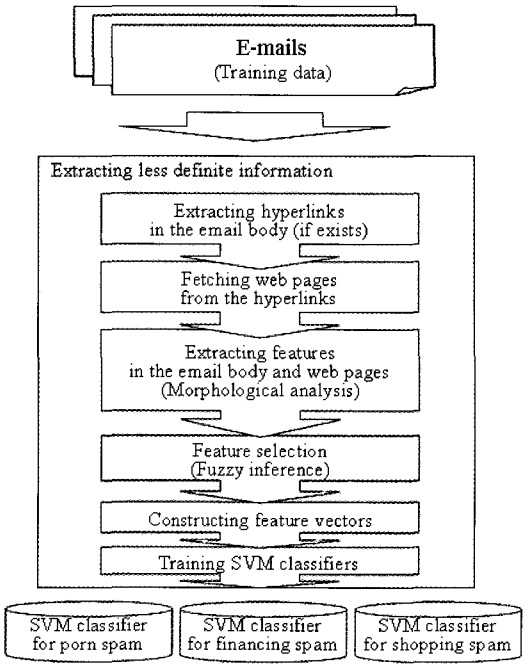

Fig. 2. Training process for 2nd phase

2.3 Kadokawa Thesaurus

The Kadokawa thesaurus [8] has 1,110 semantic categories and a 4-level hierarchy as a taxonomic relation. Semantic categories in level L1, L10, and L100 are further divided into 10 subclasses. The root node is merely a dummy node. Noun and verb categories coexist in the same taxonomic hierarchy of the Kadokawa thesaurus. Verb categories mainly correspond to the code 2xx, 3xx, and 4xx in level L1000. Some resources are readily available, such as bilingual dictionaries of COBALT-J/K (Collocation- Based Language Translator from Japanese to Korean) [9] and COBALT-K/J (Collocation-Based Language Translator from Korean to Japanese) [10] developed by POSTECH (Pohang University of Science and Technology). The Kadokawa thesaurus has proven useful for providing a fundamental foundation to build lexical disambiguation knowledge in COBALT-J/K and COBALT-K/J machine translation systems [11]. All words in these bilingual dictionaries are already annotated with the three-digit concept code at level L1000 in the Kadokawa thesaurus. So we can easily find the relevant concept codes for each lexical word.

2.4 Constructing Feature Vectors

The feature vector is constituted with the selected lexical features above, concept codes and other particular features such as domain type of sender, receiving time of an email, etc. Namely, each email is represented by a vector , where x1, x2,··· ··· Xn. Since SVM performed best when using binary features [4], we used binary representation. In case of the selected lexical features above, the feature values are defined by binary representation that indicates whether a particular word occurs in an email. In case of concept codes, which are constituted by the 1,000 concept codes at level L1000 in the Kadokawa thesaurus, the

feature values are set to 1, if a corresponding concept occurs in an email. In other cases, the attribute values are moderately defined by scaling original data according to their own properties. For example, if an email is arrived between 12 pm and 5 am, the feature value of receiving time is 1, otherwise 0. We will finally classify incoming emails into two categories: non-spam and spam mail. However, since each spam category has its own properties and SVM can only compute two-way categorization, we construct three SVM classifiers separately according to the kinds of spam mails. Three SVM classifiers are generated using the feature vectors, where three is the number of spam categories: porn spam, financing spam, and shopping spam. It is more effective than constructing only one SVM classifier for filtering all spam mails.

3. Applying Phase

Incoming emails are processed by using the several information and SVM classifiers constructed in the training phase (Fig. 3).

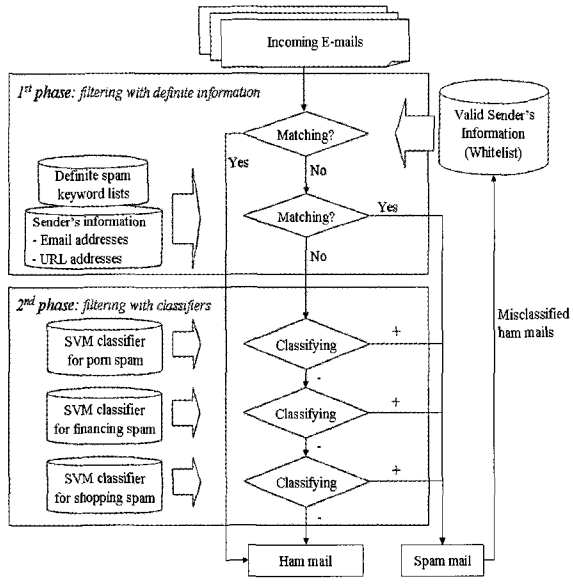

Fig. 3. Applying process for filtering spam mails

We already divided hints into two kinds of information: definite information and less definite information. In case that an email contains one of the definite information, there is no need to perform machine learning algorithms, since it has a very high probability of being spam mails. In other case that the email has no definite information, it is evaluated using the SVM classifiers. That is to say, if an email contains one of the definite information, it is regarded as a spam mail. Otherwise, it is passed to the next SVM applying phase. SVM classifier for porn spam mails is applied first. If an email is classified as a spam mail, the second applying phase is over. If not, it is passed to the next SVM classifier for financing spam. When the email is classified as a financing spam mail, the second applying phase is over too. Like the above two SVM classifiers, the last SVM classifier for shopping spam is performed in sequence if needed.

If a ham (non-spam) mail was classified as a spam mail, a loss of import information may occur. In order to avoid such a problem, a white list consisted of ham mail addresses can be constructed from misclassified ham mails. The spam mail lists classified by the system are also delivered to the users so that the users can double check the spam mail list.

4. Experiments

The email corpus used in the experimental evaluation contained a total of 5,018 emails and 4 categories: 1,737 for ham mail, 1,214 for porn spam, 1,506 for financing spam, and 561 for shopping spam. To apply the 1st phase, the definite information is extracted manually from the email corpus such as 1,461 for sender's email address, 1,023 for URL, and 603 for spam keyword list. To select important features, which can be applied in the 2nd phase, we used the weka.attributeSelection package provided by WEKA [12]. WEKA is a workbench designed to aid in the application of machine learning techniques to real world data sets. WEKA contains a number of classification models. The SVM classifier used in this experiment was also provided by WEKA. It is a John Platt's sequential minimal optimization algorithm for training a support vector classifier. SVM is tested with its default parameters settings within the WEKA.

To evaluate the filtering performance on the email corpus, we use the ham misclassification rate (hm%) and spam misclassification rate (sm%) used in Spam TREC [13], and Accuracy (A), and Error rate (E) commonly employed in the information retrieval. Given the two-way contingency table (Table 2), we can define hm% and sm% as follows.

Ham misclassification rate (hm%) = c / (a + c)
Spam misclassification rate (sm%) = b / (b + d)

Table 2. Contingency table

Filter(System) Classification \ Gold Standard Judgement	Ham (non-spam)	Spam
Ham (non-spam)	a	b
Spam	c	d

In email filtering, it is extremely important that ham emails are not filtered out. In comparison, a user may be satisfied if some spam-mail was not filtered, in order not to miss any good email. So we used accuracy and misclassification error rate [14] as evaluation measures. Accuracy represents the ratio of the correct predictions over total emails received and error rate represents the ratio of the incorrect predictions over total mails. Thus, good email filtering should show high accuracy and low error rate. In general, it is important to reduce the

average error rate than to increase the average accuracy in spam mail filtering software. Accuracy and error rates are defined as:

$$\text{Accuracy}(A) = \frac{\textit{number of correct predictons}}{\textit{total number of emails}} = \frac{a+d}{a+b+c+d}$$

$$\text{Error}(E) = \frac{\textit{number of incorrect predictons}}{\textit{total number of emails}} = \frac{b+c}{a+b+c+d}$$

In our experiments, we used ten-fold cross validation to reduce random variation. E-mail corpus was randomly partitioned into ten parts, and each experiment was repeated ten times, each time reserving a different part for testing, and using the remaining nine parts for training. Results were then averaged over the ten runs. Before using concept codes of the Kadokawa thesaurus as features in the 2nd phase, we experimented on various ways to select desirable parameter settings. Generally, it is a difficult problem to select a correct concept code for each lexical word in a certain context. For example, the Korean word "nwun" has several meanings such as eye, snow, and bud. In natural language processing (NLP), this problem is called as word sense disambiguation (WSD), which is a significant issue up to the present. To solve the WSD problem, we need many language resources such as sense-tagged corpus, collocation patterns, and NLP programs. Because of these reasons we chose a simple method of collecting all possible concept codes from emails. We collected every concept codes for each lexical word from the bilingual dictionaries, already annotated with the concept code at level L1000 in the Kadokawa thesaurus. The collected codes may contain some irrelevant codes in a certain context. The first test was performed to check the effect of irrelevant concept codes. We can think about filtering irrelevant codes by setting the value of features according to the frequency of concept codes. Table 3 compared the performance according to the frequency of concept codes. The "over 1" means that if a concept is appearing more than one in an email body, the feature value is set to 1, otherwise 0. As the Table 3 indicates, the case of "over 1" shows the best performance in most measures.

Table 3. Experimental results according to the frequency of concept codes in the 2nd phase

Measures / Threshold	hm	sm	Accuracy	Error rate
Over 1	3.0	22.4	91.2	8.8
Over 2	3.9	33.8	87.1	12.9
Over 3	3.4	62.0	79.0	21.0
Over 4	1.7	80.4	74.6	25.4

Table 4 shows that the effect of lexical and conceptual information in spam-mail filtering. If we construct a feature vector with same feature number, in case of using lexical information only, the ham misclassification rate shows the best performance, and in other case of using lexical and conceptual information together, the spam misclassification rate shows the best performance, since lexical words which have similar meanings can be categorized into a same concept code. Therefore we can reconfigure the constituent of the feature vector according to the user's needs.

Table 4. Experimental results according to the feature composition in the 2nd phase

Measures / Feature composition	hm	sm	Accuracy	Error rate
2000 lexical features only	1.0	37.1	95.0	5.0
3000 lexical features only	1.1	33.9	95.3	4.7
2000 lexical features + 1000 concept features	1.8	29.4	95.1	4.9

The 4,514 (90%) emails among 5,018 ones are used for training SVM classifiers and the remaining 504 (10%) are used for testing the two-phase system's performance with conceptual information. Testing emails are used to determine whether the mails are spam or not using the information and classifiers constructed during the training phase.

We found from Table 5 that the proposed two-phase method was more effective than the method applying each phase separately, since the 1st phase undertook some portion of the 2nd phase's workload with very high precision. The two-phase method reduced the error rate by 4.9% and 38.8% over the 1st phase or the 2nd phase only works, respectively. We can recognize from these results that conceptual information plays an important role in improving system performance.

Table 5. Performance of the proposed system (with 1000 concept features) (%)

Measures / Phase	hm	sm	Accuracy	Error rate
1st phase only	0.0	17.0	88.9	11.1
2ndphase only	11.5	62.7	55.0	45.0
1st + 2nd phase	7.5	5.5	93.8	6.2

5. Conclusion

In this paper, we constructed a spam-mail filtering system based on the lexical and conceptual information, and then analyzed the effect of the conceptual information. We see from Table 4 that the ham misclassification rate was reduced if more lexical information was used as features, and the spam misclassification rate was reduced when the conceptual information was included in feature vectors. Therefore we can reconfigure the constituent of the feature vector according to the user's needs.

In also, we proposed a two-phase method for filtering spam

mails based on lexical and conceptual information, and hyperlinks. Since the body of a spam mail has little text information recently, it provides insufficient hints to distinguish spam mails from ham mails. To resolve this problem, we utilized hyperlinks contained in the email body and extracted all possible hints from original email body and the fetched webpage. These hints are used to construct SVM classifiers. We divided hints into two kinds of information: definite information and less definite information. In the two-phase approach, definite information is used first and then less definite textual and conceptual information is applied.

We discovered that fetching hyperlinks is very useful in filtering spam mails, and the two-phase method is more effective than the method using machine learning algorithm only, blacklists, or keyword-based filters.

This research is very important in that our system can prevent young people from accessing pornography materials on spam mails by chance, and save valuable time by lightening the email checking work. We will do further research on how to find more features by considering images in email messages and constructing ontology on spam domain.

References

[1] L. F. Cranor, and B. A. LaMacchia, "Spam!," *Communications of ACM*, vol.41, no.8, pp. 74-83, 1998.

[2] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," In *AAAI-98 Workshop on Learning for Text Categorization*, pp. 55-62, 1998.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[4] H. Drucker, D. Wu, and V. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE Trans. on Neural Networks*, vol.10, no.5, pp. 1048-1054, 1999.

[5] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *ECML*, Claire Nédellec and Céline Rouveirol (ed.), 1998.

[6] J. Yang, V. Chalasani, and S. Park, "Intelligent email categorization based on textual information and metadata," *IEICE Transactions on Information and System*, vol.E86-D, no.7, pp. 1280-1288, 2003.

[7] Kim, J. W., Kim, H. J., Kang, S. J., and Kim, B. M., "Determination of Usenet News Groups by Fuzzy Inference and Kohonen Network," *Lecture Notes in Artificial Intelligence*, vol.3157, Springer-Verlag, pp. 654-663, 2004.

[8] S. Ohno, and M. Hamanishi, *New Synonyms Dictionary*, Kadokawa Shoten, Tokyo, 1981.

[9] C. J. Park, J. H. Lee, G. B. Lee, and K. Kakechi, "Collocation-Based Transfer Method in Japanese-Korean Machine Translation," *Transaction of Information Processing Society of Japan*, vol.38, no.4, pp. 707-718, 1997.

[10] K. H. Moon, and J. H. Lee, "Representation and Recognition Method for Multi-Word Translation Units in Korean-to-Japanese MT System," In *the 18th International Conference on Computational Linguistics (COLING 2000)*, Germany, pp. 544-550, 2000.

[11] H. F. Li, N. W. Heo, K. H. Moon, J. H. Lee, and G. B. Lee, "Lexical Transfer Ambiguity Resolution Using Automatically-Extracted Concept Co-occurrence Information," *International Journal of Computer Processing of Oriental Languages*, World Scientific Pub., vol.13, no.1, pp. 53-68, 2000.

[12] I. H. Witten, and E. Frank, *Data Mining: Practical machine learning tools and Techniques with java implementations*, Morgan Kaufmann, 2000.

[13] Gordon V. Cormack, *Overview of the TREC 2005 Spam Track*, http://plg.uwaterloo.ca/~gvcormac/ trecspamtrack05, 2005.

[14] P. J. Resnick, D. L. Hansen, and C. R. Richardson, "Calculating Error Rates for Filtering Software," *Communications of ACM*, vol.47, no.9, pp. 67-71, 2004.

Sin-Jae Kang

received his B.S. degree in Computer Engineering from Kyungpook National University in 1995, and the M.S. and Ph.D. degrees in Computer Science and Engineering from Pohang University of Science and Technology (POSTECH) in 1997 and 2002, respectively. He was an assistant research engineer from 1997 to 1998 at SK Telecom, Korea. Since 2002 he is working at Daegu university. His research interests include semantic web, natural language processing, and information retrieval.

Phone : 053-850-6584
Fax : 053-850-6589
E-mail : sjkang@daegu.ac.kr

Jong-Wan Kim

received the BS, the MS, and the PhD degree in Dept. of Computer Engineering from Seoul National University, Korea, in 1987, 1989, and 1994, respectively. He has been with Daegu University since 1995 and is currently a professor. From 1999 to 2000, he was a visiting scholar at Computer Science Department of University of Massachusetts, Amherst. His research areas include artificial intelligence and data mining. Currently, he has been working on the construction of anti-spam system using fuzzy inference and ontology techniques as a visiting professor at the CIS Department in U of Oregon.

Phone : 053-850-6575
Fax : 053-850-6589
E-mail : jwkim@daegu.ac.kr