Journal of Korea Multimedia Society Vol. 13, No. 6, JUNE 2010(pp. 841-848)

Personalized Anti-spam Filter Considering Users' Different Preferences

Jongwan Kim†

ABSTRACT

Conventional filters using email header and body information equally judge whether an incoming email is spam or not. However this is unrealistic in everyday life because each person has different criteria to judge what is spam or not. To resolve this problem, we consider user preference information as well as email category information derived from the email content. In this paper, we have developed a personalized anti-spam system using ontologies constructed from rules derived in a data mining process. The reason why traditional content-based filters are not applicable to the proposed experimental situation is described. In also, several experiments constructing classifiers to decide email category and comparing classification rule learners are performed. Especially, an ID3 decision tree algorithm improved the overall accuracy around 17% compared to a conventional SVM text miner on the decision of email category. Some discussions about the axioms generated from the experimental dataset are given too.

Key words: personalized anti-spam filter, user preference, email categorization, classification rule learner.

1. INTRODUCTION

Anti-spam filters can be classified into content-based and user-based ones. Conventional filters are usually working based on machine learning algorithms such as naive Bayesian classifier (NBC) and support vector machine (SVM) [1]. While the content-based filters achieve statistically impressive accuracies according to summary digests and selective tagging in order to manage their spam mails, they ignore that some email is spam to someone but ham (=legitimate or non-spam) to others in many real situations. According to this user-centered view, several personalized anti-spam systems are currently operating. Most of them require users' selection about what they accept or not based on the recommendation of the anti-spam filter. They are mainly based on other user's advice in the same network group [2] or other email account information of the same user [3] or user's judgment to accept or not based on the recommendation of the system [4].

However, we aim to develop a personalized anti-spam filtering system which is entirely dependent on user preferences and user responses in this research. We will present that our previous user preference based anti-spam system [5] is a reasonable solution by showing that traditional content-based email classifiers are not applicable in the situation where users' preferences are considered (see section 3.1). Also we will validate the proposed method by constructing classifiers deciding email categories and showing usefulness of axioms in user preference ontology.

A dynamic personalized approach combining global and personal anti-spam filtering was presented [6]. While a global or conventional anti-spam filter is trained on spam and ham email from

※ Corresponding Author : Jongwan Kim, Address : 15 Nairi Jillyang Gyeongsan Gyeongbuk 712-714, TEL : +82-53-850-6575, FAX : +82-53-850-6589, E-mail : jwkim@daegu.ac.kr.

Receipt date : Jan. 15, 2010, Revision date : Mar 3, 2010
Approval date : Mar. 16, 2010

† Member, Professor, Division of Computer and Information Technology, Daegu University

※ This research was supported by the Daegu University Research Grant, 2009.

a large collection of users, personally trained filters have the advantage of allowing each person to provide their own personal definition of spam. So both approaches have their own advantages and limits. The basic idea of dynamic personalization is to apply personal data as the amount of personalized data grows but to use global data when the personal data set is small. Since there are no user-specific labels available for any of the test corpus, Segal makes the unrealistic assumption that each user assigns the same label to each message and assigns each user the same judgment-the message's correct label as indicated by the test corpus. As a result of this assumption, the experiments tend to underestimate the value of personalized classification. The point is absolutely different from this research.

In this paper, we collected email corpus without considering email recipients and performed a data mining process on the email corpus with user's preferences for every email. Then we judge that an email is either spam or ham based on both of user's preference and email category derived from email content.

The paper is organized as the following. Section 2 describes the proposed system. Section 3 gives experiments and discussions. Conclusions are given in section 4.

2. PROPOSED SYSTEM

2.1 System Architecture

Each user can respond differently to a mail with even identical mail header and content. This situation is mainly caused by personal preferences and potential modes of behaviors. However, we do not consider users' changable behaviors in this work. We started from this assumption and decided to show that it was valid in real situations. Thus, we collected preferences for a group of users. To analyze potential responses of users to various emails, we provided sample emails to a user group and asked them to respond of the predefined (Reply, Hold, Delete, Spam) actions. In this work, "Reply", "Hold", and "Delete" responses are considered ham emails but only "Spam" response is considered as spam. Thus, this research is different from conventional spam filtering works because we classify user's specific responses into 4 categories instead of spam and ham.

The proposed architecture is given in Figure 1. User profile was collected from several participant users, user logfile was also built with their responses to sample emails, and email category (henceafter, ECat) values were given to individual mails by an email categorizer which SVM has been used in this work. A detailed procedure which SVM will be applied to the system, will be given in section 3.2.2.

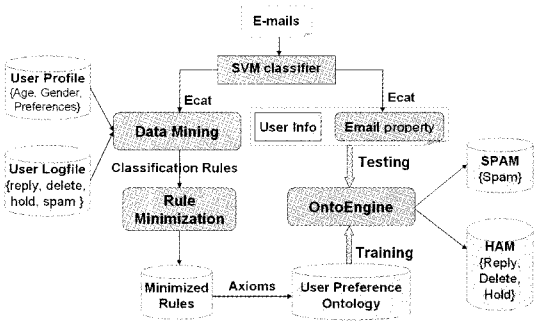

Fig. 1. The Proposed System Architecture.

We used the WEKA tool [7] to find some association and classification rules between preferences and responses. User preference ontology was constructed after data mining and rule minimization [8]. Using the ontology, especially those axioms which we constructed, our inference engine OntoEngine [9], which is a first order logic reasoner using generalized modus ponens, can classify the emails into four categories-Spam, Reply, Delete, and Hold-based on user preferences, email category, and personal information by forward chaining.

Conventional anti-spam software provides content-oriented filtering service. However, the proposed system can give user oriented anti-spam

service, because our system not only gives information about spam mail but also estimates user's response to an incoming mail. This approach can be an essential service for email clients suffering from lots of spam messages.

2.2 Ontology Construction

Ontologies, which can be defined as a formal specification of vocabulary of concepts and their relationships, play a key role to define the semantics of information for intelligent systems [10]. To achieve personalized anti-spam system, it will be helpful to construct a domain ontology which can formally define user behaviors based on their preferences. To do that, we mainly perform three steps [11]. The first step is to use association and classification mining to find relationships, i.e., rules between several users' preferences and their email responses. In the next step, we apply a rule pruning procedure eliminating redundant rules and preserving highly comprehension. Translation from optimized rules to axioms in domain ontology is performed during the final step. The details are described as follows.

To find unknown correlations between user profiles and user logfiles are required, which include user responses to sample emails. Thus, we chose the typical decision tree algorithm, ID3 [7], to train sample email preference data. After ID3 mining was performed, a decision tree is generated. We convert the decision tree into decision rules by describing each path of the tree with a rule.

A rule minimization procedure is applied to exclude redundant rules and select highly comprehensible ones inspired from logic synthesis [8]. There are several variables in a rule set derived from ID3 mining. Most variables are Boolean or binary but some of them are multi-valued. In our HLSRP (Hybrid Logic Synthesis based Rule Pruning) algorithm [11], if two or more corresponding logics of a specific variable are merged into one and then the corresponding antecedent condition of the variable will be omitted in a merged rule. Resulting from this pruning operation, the number of rules and the number of terms in a rule are reduced. Even though only one logic value among three or more attributes in a specific categorical variable was different, the multi-valued categorical variable was merged in the rule minimization algorithm in [8]. However, whenever a multi-valued categorical variable has totally different logic values in the involved rules, the proposed HLSRP method omits the multi-valued variable and merges into one rule. As a result, the classification rule accuracy is improved too.

In the final step, we interpret the derived classification rules to an ontology and axioms using a formal language, Web-PDDL [9], a strongly typed first order language especially for representing ontologies and rules. The detail procedure describing concepts as classes and properties and translating axiom rules into an anti-spam domain ontology in Web-PDDL are given in [11]. For your information, we present an example that the rules generated from classification rule learning can be put into the user preference ontology as axioms. For example, Rule: if Age = JS and ECat = Adults and FStrength = Neutral and Adults = True then Response = Hold can be represented in Web-PDDL as axioms:

```
(axioms:
(forall (c-Client e-Email)
  (if  (and (age c "JS") (FStrength c "Neutral")
       (prefer c Adults) (category e "Adults")
       (respond c e Hold)))
```

3. EXPERIMENTS AND DISCUSSION

3.1 Conventional Filters Are Not Good for Personalized Spam Filtering

We point out that traditional content-based email classifiers are not applicable in the proposed

situation where users' preferences are considered. Users may have distinct response to even the same email, yielding inconsistent dataset which solely based on email content. Table 1 illustrates the distribution of users' response for samples used in our experiment, which signifies the fact that opinions diverge on emails due to users' different preferences. In Table 1, each M_i represents the email sequence. Therefore there is no straightforward way to compare the performance of our system with traditional approaches.

Table 1. Distribution of users' responses for samples

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	···
Reply	5	2	1	17	1	2	2	
Spam	31	29	24	6	25	40	8	···
Hold	10	3	13	22	9	1	30	
Delete	21	30	28	18	31	23	26	

3.2 Experimental Results

3.2.1 Dataset

To evaluate the proposed approach based on the user preference ontology, we collected emails from mostly 6 predefined categories including adults, entertainment, finance, jobs, shopping, and travel. Since we chose 150 samples for each category from email correspondences of anonymous users, a total of 900 emails in Korean and English were collected. But 249 emails in English are utilized in this experiment. We also collected responses to those 249 emails from 74 college students together with their respective preference over several options. But since a few users have a confusion to decide their own responses for some emails, very few 91 instances among original 18,426 (= 74 * 249) records were left blank. And thus we got a dataset with 18,335 records; each record consists of the email category, user preferences, and the corresponding response. We used 12,223 instances, almost 2/3 instances of the dataset in the training process, and the rest of the total instances for testing.

3.2.2 Construction of Classifiers Deciding the Email Category

Since user response to a given email normally depends on both of user preference and email category based on the content, the email category information should be included into the system. In general, two kinds of methods such as NBC and SVM are used to classify an email into several categories such as adults, finance, entertainment, and so on. There is no big difference in classification performance between NBC and SVM. So, we used the standard SVM [12] as an email category classifier in this work. There are several choices when designing a multi-class text classifier based on SVM. Typically in text categorization, a feature is a word. There is one feature vector per message and there are various alternatives in assigning value to the component of the vector. We consider the followings in our experiment:

TF (Term frequency): The i-th component of the feature vector is the number of times that a specific word appears in that document.

TF-IDF: uses the above TF multiplied by the IDF (inverse document frequency).

Binary representation: indicates whether a particular word occurs in a particular document.

We also consider whether to apply a stop word filter and word stemming [13]. The argument against using a stop list is that it is not obvious which words, beyond the trivial, should be on the stop list. And by applying word stemming, the size of the feature vector is lowered but it is also tricky since certain forms of a word may be important in classification. We experimented with various configuration settings and the classification results are illustrated in Table 2. From the results we observed that TF method with stop list and word stemming works best in this application of multi-category email classification, achieving the accuracy of 81.27%.

Table 2. Performance of email classifier with different configuration settings

	Stop list and stemming	No stop list or stemming
TF	81.27%	80.13%
TF-IDF	78.33%	78.89%
Binary	76.08%	74.41%

3.2.3 Comparison of Classification Rule Learners: SVM vs ID3

We implemented two methods, namely the SVM and decision tree learner, to take into consideration the users' preference information in the classification rule generating process of the data mining module in Figure 1. In both methods we integrated preferences as binary-value attributes together with the email category to predict on the users' response. The email categories are manually assigned by users in the training phase but decided by SVM in the test mode. But in practice, the accuracy of the whole system depends heavily on the decision of email category in the first place. Results of the two classifiers on our test dataset showed that SVM achieved 63.02% overall accuracy with 4 responses while ID3 79.85%. Other performance measures including precision and recall are presented in Table 3.

Table 3. Precision and Recall results for the two classification learners: SVM and ID3

Classifier / Response	SVM		ID3	
	Precision	Recall	Precision	Recall
spam	0.56	0.53	0.77	0.75
delete	0.62	0.86	0.78	0.90
hold	0.74	0.41	0.82	0.69
reply	0	0	0.51	0.24

Thus, we chose ID3 decision tree classifier as a classification rule mining algorithm for the implementation of our system. Not only in favour of its relatively high accuracy, but we also appreciate the ability of composing comprehensible rules from decision trees. We construct the preference based ontology by adding in axioms translated from the rules "read-off" from the decision tree. We present and discuss the axioms generated from the experimental dataset in the following section.

3.2.4 Experiments Showing Usefulness of Axioms

We obtained 304 rules from ID3 learning process. To evaluate the performance of derived rules, we applied them to the 6,112 test instances and carried out the HLSRP rule minimization approach, which then reduced the rules to 238. We interpreted the rules in logic axiom form and some representative ones are shown in Table 4. Each rule in the table can be explained as the way that a user responds to a certain email with respect to his or her specific preference. For example, ECat has "Travel" attribute in the four rules 7, 8, 10, and 15, but their Response values are not the same. Because users' preferences for Travel are True in the rules 7,10 and 15, they tend to hold for the kind of emails. In the case of the rule 8, they almost delete them. This point convinces us that the proposed personalized anti-spam approach can be a more adaptable solution.

3.3 Discussions

Through the data gathering process and implementation of the proposed architecture, we believe that an adaptable and readily customizable anti-spam filter is increasingly in demand. We proposed a personalized anti-spam system using ontologies constructed from rules derived in decision tree mining. From the above experiments, it was shown that the proposed approach can be a new solution for users suffering from lots of disgusting spam. The performance of the system relies on the following factors.

Preference modelling: in this paper we define users' preference in terms of a collection of Boolean

Table 4. Axioms derived by the proposed approach and their accuracies (T: True and F: False)

No	Axiom Rule	accuracy
1	Age=JS & Gender=Male & ECat=Adults & Adults=F & Entertainment=F & Sports=T & Kids=F → Response=Spam	1
2	Age=JS & Gender=Female & ECat=Adults & Adults=F → Response=Spam	1
3	Age=FS & FStrength=Neutral & ECat=Adults & Adults=F & Shopping=F & Travel=T → Response=Spam	1
4	Age=FS & ECat=Adults & Adults=F & Travel=F → Response=Spam	1
5	FStrength=Neutral & ECat=Entertainment & Entertainment=F → Response=Delete	1
6	Age=JS & FStrength=Strong & Gender=Male & ECat=Entertainment & Adults=F & Entertainment=F & Shopping=F & News=F & Sports=F → Response=Delete	1
7	FStrength=Strong & ECat=Travel & Adults=F & Entertainment=F & Finance=F & Shopping=F & Travel=T & Kids=F → Response=Hold	0.978
8	FStrength=Neutral & ECat=Travel & IT=T & Travel=F & Sports=T → Response=Delete	0.95
9	Age=FS & ECat=Adults & Adults=F & Shopping=T & Travel=T → Response=Spam	0.947
10	FStrength=Neutral & ECat=Travel & Finance=F & Travel=T & Sports=T → Response=Hold	0.939
11	FStrength=Strong & Gender=Male & ECat=Shopping & Adults=F & Finance=F & Shopping=F & Travel=T & Sports=T → Response=Delete	0.938
12	FStrength=Neutral & ECat=Etc & IT=F & Finance=F → Response=Delete	0.853
13	ECat=Shopping & Finance=F & Travel=F & News=F & Sports=F → Response=Delete	0.815
14	FStrength=Neutral & ECat=Entertainment & Entertainment=T & Shopping=T & Sports=T → Response=Hold	0.810
15	Age=FS & FStrength=Strong & ECat=Travel & Finance=T & Shopping=F & Travel=T → Response=Hold	0.767

attributes. These attributes are selected based on some preliminary empirical categorization. We argue that a more sophisticated modelling from a thorough study can improve the system's performance but it is definitely a non-trivial job. The modelling itself could be a learning task that determines a set of attributes contributing most to the user's response to a specific email.

Email categorization: in order to perform personalized anti-spam filter in a fine granularity, we need to obtain the knowledge of the email as much in detail as possible. Therefore as oppose to the traditional filters that typically only deal with a binary class classification problem, we need to develop a multi-class email classifier. Experiments showed that our SVM classifier with specific configuration settings (using TF, stop list and stemming in text vectorization) achieved 81.27% accuracy. State of the art multi-class text mining techniques can be plugged in and improve the performance.

In the previous research [11], the importance of interpreting the classification rules to an axiom ontology was focused. However, the following three contributions are conducted in this work: the two experiments that construct classifiers deciding the email category and compare classification rule learners in a data mining module are given, and the aforementioned performance modelling and email category classifier are the key factors affecting the performance of the proposed system.

4. CONCLUSION

We presented the demand of a personalized anti-spam mail system by showing that the traditional content-based filters are not applicable to the real world, which each user's response can be different to exactly identical mail. The important feature of the proposed approach is to allow users to give different responses to the same email based on their preferences. It is different from conventional systems that normally judge which mail is spam based on the email content and expect every user to equally respond to the same email. Especially this assumption is not true by illustrating individuals respond differently to the same email is shown. It is a step forward for building personalized anti-spam mail service considering user preference and previous response history as well as email content. Another contribution of this work is that a user preference ontology can explain reasonably why a certain mail is decided to be spam or ham. For the future work, we need to extend the proposed system in order to accomodate every client demand up to the level for launching in the market.

ACKNOWLEDGEMENTS

The author appreciates Prof. Dou and Haishan Liu at the Computer and Information Science Department at University of Oregon to support this research. I also thank to all participants who read sample emails and provided their responses.

REFERENCES

[1] G. Cormack and T. Lynam, "On-line Supervised Spam Filter Evaluation," *ACM Trans. on Information Systems*, Vol.25, No.3, article 11, 2007.

[2] A. Gray and M. Haahr, "Personalized, Collaborative Spam Filtering," Proc. of the First Conference on Email and Anti-Spam, Mountain View, CA, 2004.

[3] J. Ravi, W. Shi, and C. Xu, "Personalized Email Management at Network Edges," *IEEE Internet Computing*, Vol.9, No.2, pp. 54-60, 2005.

[4] Anti-Spam Firewall, http://www.barracuda-networks.com/ns/products/anti_spam_tech.php.

[5] J. Kim, D. Dou, H. Liu, and D. Kwak, "Constructing A User Preference Ontology for Anti-spam Mail Systems," *Lecture Notes in Artificial Intelligence*, Vol.4509, pp. 272-283, 2007.

[6] R. Segal, "Combining Global and Personal Anti-Spam Filtering," Proc. of the 4th Conf. on Email and Anti-Spam, http://www.ceas.cc/papers-2007/, 2007.

[7] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and Techniques with java implementations*, 2nd Ed., Morgan Kaufmann, San Francisco, CA, 2005.

[8] J. Kim, "A Method to Minimize Classification Rules Based on Data Mining and Logic Synthesis," *Journal of Korea Multimedia Society*, Vol.11, No.12, pp. 1739-1748, 2008.

[9] D. Dou, V. McDermott, and P. Qi, "Ontology translation on the semantic web," *Journal of Data Semantics*, Vol.2, pp. 35-57, 2004.

[10] T. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *Int. Journal of Human-Computer Studies*, Vol.43, pp. 907-928, 1995.

[11] J. Kim, "From Computing Distribution of Email Responses for Each User To Construct User Preference based Anti-spam Mail System," *Journal of Korean Institute of Intelligent Systems*, Vol.19, No.3, pp. 343-349, 2009. (in Korean)

[12] 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.

[13] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

Jongwan 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 2006-2007, he was a visiting professor at Computer and Information Science Department of University of Oregon, working on the user preference ontology based anti-spam systems with the partial support of KRF. He has written several papers in the areas of artificial intelligence, fuzzy systems, and anti-spam systems. His current research areas include artificial intelligence, data mining, internet malfunction protection, and IT convergence service.