Toward Time-Sensitive Structure Analysis for SPAM Filtering: A Data Mining Approach

Atsushi Inoue, Isao Hayashi, Toshiyuki Maeda, Yoshinori Arai and Takashi Kobayashi

Abstract— **The rapid growth of the Internet has caused issues concerning SPAM messages. As a result of this being a widespread social problem, many Mail User Agents (MUAs) are facilitated with SPAM filters. Unfortunately, the most of those filters use probability-based methods that concern only their contents such as word frequencies, thus do not properly perform SPAM filtering based on their intrinsic structure. Moreover, hardly any SPAM filters are sensitive to change in attributes over time. This paper introduces our anticipation toward their time-sensitive structure analysis. Taking a data mining approach on both headers and contents of mail messages, we analyze significance, validity and utility for SPAM filtering. Attributes are selected from their header fields as well as various summarization of their contents. Such a data mining approach is then taken in consecutive time periods in order to study time-sensitivity, i.e. change of significant attributes over time. Some cross validation is conducted in order to show validity. Decision Tree Learning (DTL) is currently deployed for its advantage of identifying significance based on Information Entropy. Finally, we present results throughout experiments using mail messages sent to an actual site.**

I. Introduction

In general, SPAM refers to mail messages distributed anonymously for their various intention including, but not limited to, advertisement, sales presentation, phishing and fraudulence. Furthermore, they usually ignore intentions of recipients and often contain computer virus and warms that turn their machines yet another SPAM distributors [1]. SPAM has been said to occupy up to 97% of the entire mail messages being transmitted over the Internet [2], thus is lead to significant waste of network bandwidth and accessible computational resources. Clearly, there is no need for recipients to receive unwanted messages, that include most certainly SPAM messages.

Conventional methods of SPAM classification significantly rely on probabilistic classification criteria, that do not necessarily reflect on the intrinsic structure, thus often result in misclassification. We consider that such inconsistencies are caused by lack of structure identification intrinsic to SPAM, whereas many methods deploy criteria opt for filtering and optimal combination and selection of attributes.

In this study, we anticipate a method that is capable of the following:

1) extracting attributes both from header information and contents of mail messages;
2) not only extracting attributes but also analyzing their significance, validity and utility for SPAM classification; and,
3) capturing change of structure (time-sensitivity) in attributes over time.

We consider such a method more beneficial as a base of SPAM filtering and unique in comparison with many other conventional methods. By introducing the analysis over time, we are anticipating an extension from classification to prediction. To an extreme extent, we are envisioning a *SPAM forecasting* simply as a matter of scaling up our method.

For our initial stage, we take an off-line data mining approach in order to analyze structure in attributes with respect to the following: significance, validity, utility and time-sensitivity. We use a Decision Tree Learning algorithm (DTL) for data mining so that the significance can be caught as a result of determining the level of tree nodes, i.e. attributes. Datasets are generated for consecutive time periods for the analysis of time-sensitivity. The validity is analyzed as a result of cross validation. Finally, we can determine the utility by observing classification correctness.

II. Related Works and Utility

In general, there are three types of the study on SPAM filtering:

- improvement of classification correctness and scaling,
- identification of attributes for classification, and
- feature extraction and its optimization for SPAM filtering.

Yong Hu, et. al. [5] have proposed a SPAM filtering framework as a result of only analyzing header information. Lee, et. al. [3] have anticipated to improve classification correctness through optimization of combining many extracted features. Almeida, et. al. [4] have improved conventional SPAM classification methods according to their proposed set of SPAM filtering criteria. Soranamagesswari, et. al. [6] have proposed a feature extraction method using neural networks in order to classify SPAM messages with images. All such works are mainly concerned with improvement of conventional methods and that of feature extraction.

On the other hand, we have studied acquisition of rules in order to identify the optimal location of WiFi access points using Fuzzy ID3 method [7]. From this, we leaned the effectiveness of DTLs for structure analysis as well as that of 'fuzzyfication' in order to simplify the tree(s) without losing the significance.

Atsushi Inoue is with Eastern Washington University, USA. Isao Hayashi and Takashi Kobayashi are with Kansai University, Japan. Toshiyuki Maeda is with Hannan University, Japan. Yoshinori Arai is with Tokyo Polytechnic University, Japan.

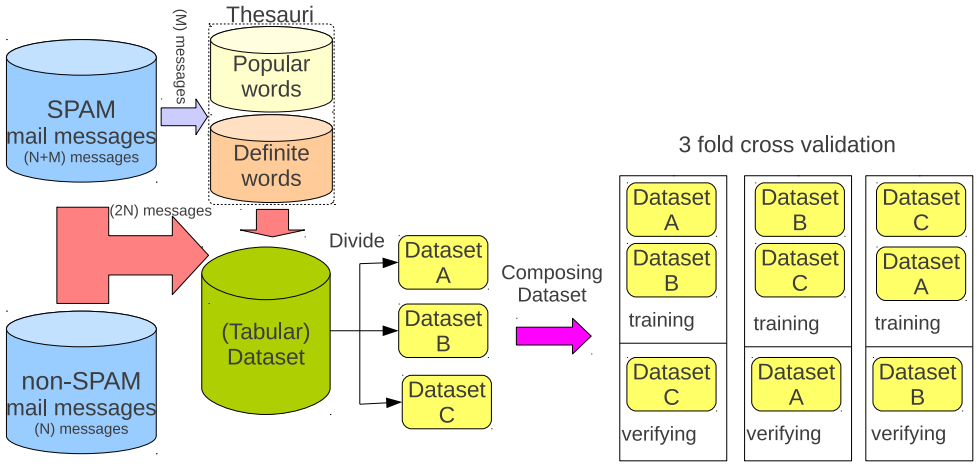

Fig. 1. data generation overview

III. Dataset Generation

First, we simply collected a large number of mail messages (Japanese only at this time) over fifteen months. All those messages are labeled either as SPAM or as non-SPAM by users. We then divide them into five collections such that each collection holds messages within a three-month period. For the each collection, we perform the following steps (fig. 1):

1) Generate two thesauri containing words from sample SPAM messages. The samples are randomly selected from the collection among its SPAM messages. Those thesauri are to be used for feature extraction in the next step.
2) Generate the dataset from the remaining collection. This dataset contains the equal number of data for SPAM and non-SPAM messages respectively.
3) Arrange the generated dataset for the 3-fold cross validation such that the dataset is divided into three portions with equal size. Each portion contains the equal number of data for SPAM and non-SPAM messages respectively. Two out of those three portions are used for learning and the remaining for testing (aka classification). Do so for all (three) possible combinations of those three portions.

In fig. 1, quantity of mail message distribution in order to generate necessary data and thesauri is represented by M and N. Each step is described in the following subsections.

A. Thesauri

Two thesauri are generated for feature extraction that serves as an indication of SPAM. First, a SPAM word list together with its word frequencies is generated from the bag of all SPAM samples using a Japanese/Chinese morphological analyzer, ChaSen [8]. A stop-word list is prepared and applied to remove unnecessary words. That contains words like symbols, one-syllabary words (i.e. Hiragana and Katakana), western alphabets and numbers. The following two thesauri are generated based on that list of words.

1) List of popular SPAM words: P: This list consists of the first 50% of the words in the SPAM word list sorted by their frequencies in its descending order. This is used to determine the degree of a message likely being SPAM $s_1(m)$ such that

$$s_1(m) = \frac{\Sigma_i fr(w_i,\ m)}{fr(m)} \tag{1}$$

where m is a message, $fr(w_i,\ m)$ is a word frequency of the i-th word in the list $w_i \in P$ in the message, and $fr(m)$ is the number (frequency) of words in the message.

2) List of definite SPAM words: D: This consists of unpopular words, especially with very low frequencies, e.g. 1, that SPAM messages definitely contain. In our study, we pick words that frequency is one and that appear only in at least one SPAM message but not in any non-SPAM messages within the sample collection. Then an alternative to the degree of a message likely being SPAM, namely $s_2(m)$,

is defined such that

$$s_2(m) = \begin{cases} 1 & \text{if } w \in D \wedge w \subseteq m \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Message m is definitely SPAM if at least one word in this list of definite SPAM words appears.

B. Dataset

The generation of the dataset transforms a collection of mail messages to a set of data in a tabular form that consists of attributes and the class label (SPAM or non-SPAM determined by users). As a result of a study on feature extraction for SPAM filtering [9], we extracted ten attributes, of which six are extracted from header information and the other four are extracted from message contents. For the extraction of the four attributes from message contents, we use ChaSen, the same tool to generate the thesauri, for parsing words and tagging and write some scripts for other necessary computation in order to extract attributes. For convenience of applying various data mining algorithms, those attributes in this dataset are all numerical. Categorical attributes are properly coded (mostly indexed or hashed) so that they are represented as numbers. More details of those ten attributes and the class label are given as follows:

- **IP address (IP)**. Extracted from Received field in the *h*eader. All IP addresses are extracted in a SPAM message (since any address can be a fake), but only those outside the gateway (i.e. of hosts outside the network) are extracted in a non-SPAM message.
- **Matching degree of domain names between Message-ID and Sender fields (matching)**. Extracted from Message-ID and Sender fields in the *h*eader. Non-SPAM messages tend to have similar domains in Message-ID and Sender fields. On the other hand, SPAM messages tend to have exactly the same domains or even exactly the same server name in Message-ID and Receiver fields. Given such a tendency, we use a matching degree of such domain names.
- **Subject (subj.)**. Since SPAM messages tend to use the same Subject in the *h*eader over time, this is useful for SPAM filtering.
- **Name**. Extracted from From field in the *h*eader. Since SPAM messages tend to use the same name over time, this is useful for SPAM filtering.
- **Content type (cont.)**. Extracted from Content-Type field in the *h*eader and encoded as follows:
  1) HTML-based content
  2) text-based content

  There is a tendency for SPAM messages to hold a certain kind of contents. In such a case, this attribute should contribute significantly.
- **Attachments (attach.)**. Extracted from Content-Type field in the *h*eader and encoded as follows:
  1) no attachments
  2) text attachment
  3) non-text attachment

  SPAM messages may have certain tendency concerning attachments. In such a case, this attribute should contribute significantly.
- **Number of URLs in the message content (URL#)**. Extracted (parsed) from the content. SPAM may consist only of URLs or no URL at all. Therefore, this attribute is likely useful for classification.
- **URL ratio in the message (URL%)**. Extracted (parsed) from the content. This ratio is based on the number of bytes (letters and syllabaries) that URLs takes against the total number of bytes in the message. This attribute is very likely critical for analysis of SPAM (as well as non-SPAM) messages due to variation of URL apperance in mail messages.
- **SPAM word ratio (SPAM%)**. Extracted (parsed) from the content. This ratio looks into the occupation of SPAM words identified in the thesauri. Therefore, this is the best to be the byte-based ratio – the number of bytes taken by SPAM words divided by the total number of bytes taken by the entire message. As this quantity becomes larger, the more message space is occupied by SPAM words (thus more likely a SPAM message).
- **SPAM degree ($s(m)$)**. Extracted (parsed) from the content. This is an alternative to the SPAM word ratio, thus is based on word frequency such that a linear combination of the degrees of a message likely being SPAM, namely $s_1(m)$ and $s_2(m)$, and defined as follows:

  $$s(m) = w1 \cdot s_1(m) + w_2 \cdot s_2(m) \quad (3)$$

  We need to select $w1 >> w2$ due to the characteristic of $s_1(m)$, i.e. a normalized word frequency, that tends to be significantly small in many cases.
- **Class label**. Assigned by users as a result of clicking report SPAM button in their mailers (MUAs) and encoded as follows:
  1) Non-SPAM
  2) SPAM

  Note that this is entirely on users' judgment, thus may not necessarily be consistent with tendency and characteristics of extracted attributes.

C. Arrangements for Cross Validation

The generated data are converted for cross validation using MUSASHI [10], a powerful data processing mining (association rules) tool based on XML technologies. We use 3-fold cross validation, thus the converted data are divided into three equal portions. Two portions constitute a training data set while the remaining serves as a testing (classification) data set. A pair of learning and classification operations take place on altogether three possible combinations of those three equal portions.

IV. Results

A. Analyses

We use a well integrated data mining tool Weka [11] for our experiments. We selected a DTL algorithm J48 within

TABLE I

#SAMPLES IN THE DATASET

Period	SPAM	Non-SPAM
DEC08-FEB09	395	332
MAR09-MAY09	333	289
JUN09-AUG09	298	253
SEP09-NOV09	216	183
DEC09-FEB10	647	527

Weka at this time and conduct analyses of significance, validity, utility and time-sensitivity on the generated datasets.

We use $w1 = \frac{50}{51}$ and $w2 = \frac{1}{51}$ for equation 3.

Mail messages were collected over fifteen months from December, 2008 to February, 2010 and were divided into five portions corresponding to each and every three-month period as indicated in Table I. We have received around more or less 100 SPAM messages in every month in Japanese. They are all manually and keenly labeled by users whether SPAM or non-SPAM. As a result of this, only trivial SPAM messages are claimed as SPAM in this study. In other words, there is very slim chance of finding non-SPAM in selected SPAM message, while the other way around may occur with a (slightly) better change.

B. Results

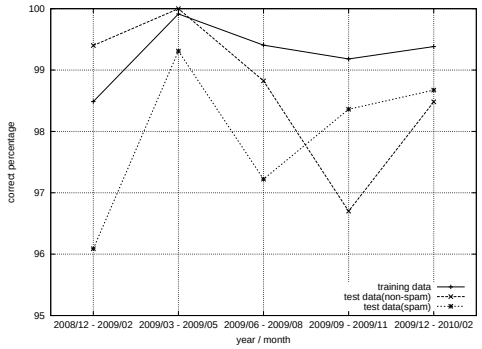

Fig. 2. classification results

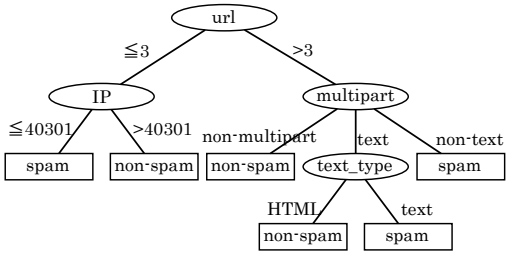

Fig. 3. decision tree: JUN09-AUG09

Classification correctness by DTL J48 over those five periods is shown in Fig. 2. Three induced decision trees in the periods of June 2009-August 2009, September 2009-November 2009 and December 2009-February 2010 are drawn in Fig. 3, Fig. 4 and Fig. 5 respectively for the study on

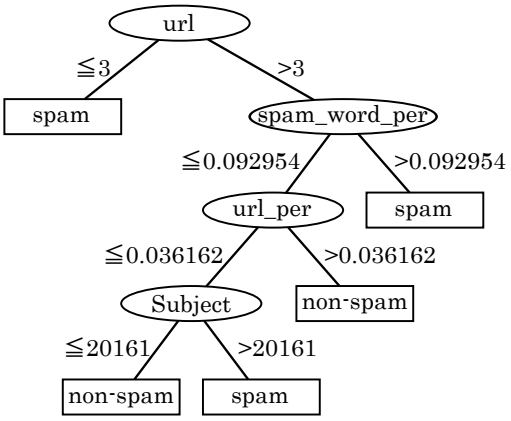

Fig. 4. decision tree: SEP09-NOV09

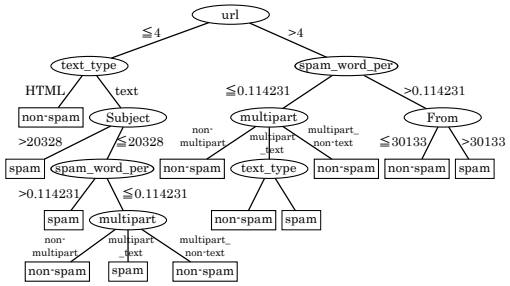

Fig. 5. decision tree: DEC09-FEB10

significance as well as time-sensitivity. The cross validation results should indicate validity of those decision trees.

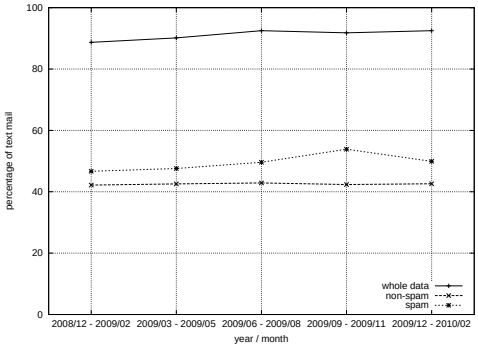

Fig. 6. perceptage of text messages

In addition, we looked into the following that has lead us additional discovery about SPAM: the percentage of text mail messages in Fig. 6, that of HTML mail messages in Fig. 7, that of mail messages without any attachments in Fig. 8, and that of mail messages with at least one text attachment in Fig. 9.

V. STUDY

A. Utility and Validation of Induced Decision Trees

As shown in Fig. 2, classification of both SPAM and non-SPAM always resulted in better than 96%. This suggests a high utility of SPAM filtering based on those results.

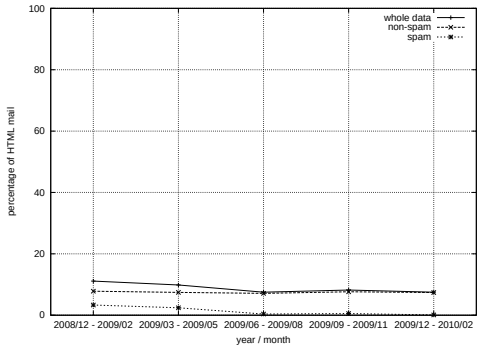

Fig. 7. percentage of HTML messages

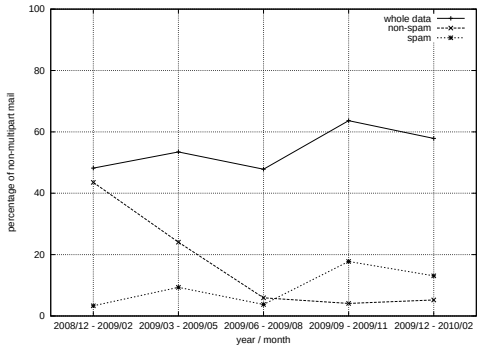

Fig. 8. perceptage of messages without attachments

When applying the 3-fold cross validation, all those three cases resulted similarly. Furthermore, similar decision trees were induced in those cases. Therefore, those good results are said to be validated.

B. Time Sensitivity

From each decision tree in respective period, we obtained attributes that are significant for classification (i.e. in terms of the number of messages correctly classified) as shown in Table II. This result shows that 91% of SPAM messages and 94% of non-SPAM messages were classified with the node of attribute URL ratio (URL%) in the period from June, 2009 to August 2009. In the following period, i.e. from September 2009 to November 2009, around 80% of SPAM messages were classified with this decision node and the remaining 20% were classified with the same node and the other of SPAM word ratio (SPAM%).

The number of attributes significant for classification be-

TABLE II
SIGNIFICANT ATTRIBUTES FOR SPAM CLASSIFICATION

Period	Input attributes
DEC08-FEB09	IP
MAR09-MAY09	attach., name, IP
JUN09-AUG09	URL#, IP
SEP09-NOV09	URL#, SPAM%, URL%
DEC09-FEB10	URL#, SPAM%, attach.

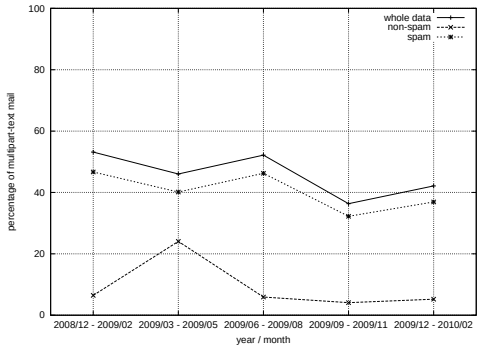

Fig. 9. perceptage of messages with text attachment(s)

comes larger in the next period from December 2009 to February 2010. As can be observed, the complexity of the decision tree (determined based on the depth and the number of nodes) becomes higher as time goes by (i.e. proceeding to the next period). This is a good support of our expectation on time-sensitivity.

Despite our expected significance of message subjects for classification, this only appeared at the lower levels of decision trees (or never appeared). On the other hand, content-based attributes based on URL (e.g. URL% and URL#) and word frequency ($s(m)$) tend to be significant for classification. This suggests that analysis of message contents is more essential than that of header information. However, there are still a few header information that are significant. Overall, both types of attributes are essential for a better SPAM filtering.

C. Additional Findings

Correctness of machine learning algorithms is supported under assumption of so-called the *s*anitary condition. Likewise, their completeness is supported under assumption of a "complete" feature extraction (i.e. identification of attributes). Unfortunately, many problem domains are too complex and dynamic to make such assuptions. SPAM filtering is no exception. As a matter of fact we found what follows.

In Fig. 6 and Fig. 7, we observed that majority of mail messages are text based. The ratio of SPAM and non-SPAM messages among those text-based messages are equally likely (aka 50:50). However, non-SPAM messages are clearly higher in that ratio among HTML-based messages.

Fig 8 and Fig 9 indicated that many text-based SPAM messages have some text attachments while many non-SPAM messages do not have any attachments. Therefore, attributes content type (cont.) and attachments (attach.) are significant. This is somewhat consistent as they appear in decision trees in Fig. 3 and Fig. 5.

As a result, we may say that DTL does not necessarily capture well the relations as described above, i.e.:

- For SPAM classification, text content and text attachments serve as a good attribute.
- For non-SPAM classification, HTML content and no attachments serve as a good attribute.

VI. Concluding Summary

In this paper, we presented our initial anticipation toward a time-sensitive structure analysis for SPAM filtering that very likely extend from classification to prediction; as well as, to an extreme extend, forecasting as a matter of scaling up. We are taking a data mining approach utilizing conventionally well integrated tools due to the scale, complexity and dynamic nature of SPAM filtering.

Overall, we obtained a satisfactory result in analysis of significance in attributes, that of validity supported by the stable result of 3-fold cross validation, that of utility supported mostly by the classification correctness at least and mostly better than 96%. Some critical additional findings are made as well. Furthermore, time-sensitivity was successfully observed on change of significance in attributes, i.e. that of decision trees induced in consecutive periods.

Our future works include, but are not necessarily limited to,

- applications of other machine learning algorithms, e.g. SVM, association rules, neural networks, etc;
- finer granulation of periods, e.g. from three months to one month, in order to analyze SPAM and non-SPAM messages more extensively;
- identical IP address extraction between SPAM and non-SPAM messages that may end up with different results; and most importantly,
- deployment of Soft Computing approaches such as Fuzzy ID3 as an alternative machine learning algorithm and Granular Computing frameworks applied to dataset generations (e.g. granulation management on the time periods).

References

[1] A. Watabe and K. Aiko, *The SPAM Mail Textbook.* Data House, Japan, in Japanese, 2006.

[2] K. Yoshizawa, *CNET Japan.* `http://japan.cnet.com/` (complete URL omitted), extracted on July 16th, in Japanese, 2010.

[3] S. Lee, et. al., "SPAM Detection Using Feature Selection and Parameters Optimization," *CISIS2010.* pp.883-888, 2008.

[4] T. Almeida, A. Yamakami and J. Almeida, "Filtering SPAMs using the Minimum Description Length Principle," *SAC'10.* pp.1854-1858, 2010.

[5] Y. Hu, et. al., "A Scalable Intelligent Non-content-based SPAM-filtering Framework," *Expert Systems with Applications.* Elsevier, 2010.

[6] M. Soranamageswari and C. Meena, "Statistical Feature Extraction for Classification of Image SPAM Using Artificial Neural Networks," *ICMLC.* pp.101-105, 2010.

[7] I. Hayashi, T. Kobayashi, Y. Arai, T. Maeda and A. Inoue, "Optimal Location of Wireless LAN Access Points Using Fuzzy ID3," *Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.13, No.2*, pp.128-134, 2009.

[8] Nara Institute of Science and Technology, *ChaSen Legacy.* `http://chasen-legacy.sourceforge.jp/` extracted on July 16th, 2010.

[9] Rie Sasaki, "Study on Feature Extractions for SPAM Classification," *Kansai University Undergraduate Thesis.* 2009.

[10] *MUSASHI.* `http://musashi.sourceforge.jp/`, extracted on July 16th, 2010.

[11] *Weka 3.* `http://www.cs.waikato.ac.nz/ml/weka/`, extracted on July 16th, 2010.