# Using Data Mining Methods to Predict Personally Identifiable Information in Emails

Liqiang Geng[1], Larry Korba[1], Xin Wang[2], Yunli Wang[1], Hongyu Liu[1], and Yonghua You[1]

[1] Institute of Information Technology, National Research Council of Canada
Fredericton, New Brunswick, Canada
{liqiang.geng,larry.korba,yunli.wang,hongyu.liu,
yonghua.you}nrc-cnrc.gc.ca
[2] Department of Geomatics Engineering, University of Calgary, Calgary, Alberta, Canada
xcwang@ucalgary.ca

**Abstract.** Private information management and compliance are important issues nowadays for most of organizations. As a major communication tool for organizations, email is one of the many potential sources for privacy leaks. Information extraction methods have been applied to detect private information in text files. However, since email messages usually consist of low quality text, information extraction methods for private information detection may not achieve good performance. In this paper, we address the problem of predicting the presence of private information in email using data mining and text mining methods. Two prediction models are proposed. The first model is based on association rules that predict one type of private information based on other types of private information identified in emails. The second model is based on classification models that predict private information according to the content of the emails. Experiments on the Enron email dataset show promising results.

## 1 Introduction

With the information explosion arising from marketing and other business requirements, today's organizations are facing increasing demands to assure privacy compliance with both internal and external policies, regulations, and laws. Externally, corporations are required to comply with a variety of regulations depending on their operational domains and sectors of business. Internally, good corporate practices demand effective management, good decision making, clear accountability, effective risk management, corporate integrity, etc. with respect to the collection, storage and use of personally identifiable information (PII). Because of these increasing demands, corporations are in need of automated tools that can assist in the monitoring of PII data access in workflows, and in ensuring and demonstrating compliance with various privacy requirements.

Email is one of the most important communication and information exchange tools for modern organizations. It has greatly improved work efficiency of organizations. However, with the increasing use of email, leaks and violations of the use of client PII have become a serious issue that organizations are facing.

C. Tang et al. (Eds.): ADMA 2008, LNAI 5139, pp. 272 – 281, 2008.

Two approaches can be used to ensure and to demonstrate privacy compliance for email. First, we can monitor the email content before the emails are sent out. Emails can be blocked if any violations have been detected according to policies and operational context. We call this approach *prevention*. The advantage of this approach is that the privacy violations can be prevented before an email is sent out. The disadvantage is that it may not be flexible enough. In real applications, there may be exceptions and emergencies which require immediate access to private information but which may not be allowed in normal process. The prevention approach may be a hindrance in these situations. In the second approach, which we called *audit*, we need software tools to find traces of the privacy leaks on an as-needed basis, or to report privacy violations on a regular basis. The advantage of this approach is that it provides the user great flexibility. The disadvantage is that it may work after the damage has been done.

The related work on detecting private information in emails is very limited. Armour et al. proposed an email compliance engine to detect privacy violations [3, 4]. This engine consists of two components: The entity extraction module which can identify private information such as names, phone numbers, social insurance numbers, student numbers, and addresses, and the privacy verification module which determines if the email should be blocked according to the type of private information detected, the recipients of the email, and the electronic policies or rules stored in a database. This method is aimed to prevent the private information leaks.

Carvalho and Cohen deal with private information leaks in email from another perspective [5]. They try to prevent the messages from accidentally being addressed to non-desired recipients. The approach does not detect private information in the emails directly. Instead, it predicts if an email is being sent to the wrong person based on the content of the email, the recipient, and the email history of the sender. They used an outlier detection method, which incorporates both content analysis and social network analysis.

It should be noted that private information discovery is different from privacy-preserving data mining [2, 6]. First, privacy-preserving data mining usually works on structured data, like databases, while private information discovery usually works on free text. Secondly, privacy-preserving data mining assumes that the location of the private information is known and uses this information as constraints for data mining process, while the aim of private information discovery is to identify the location and type of the private information. Thirdly, the information in databases for privacy-preserving data mining is more easily manageable than the free text used for private information discovery.

Our work on private information discovery is part of the Social Network Applied to Privacy (SNAP) project [8] underway within National Research Council of Canada. The project addresses the comprehensive issues in managing PII within organizations. The most important issues include identifying the PII in different kinds of documents stored in a computer and in network traffic, tracing user's access and entry of PII, and detecting the inappropriate access to or use of PII in workflow. The first step and the core of these functionalities is to identify the private information, such as email addresses, telephone numbers, addresses, social insurance numbers, etc.

Currently, PII detection algorithms have been implemented in SNAP to detect private information. For example, pattern matching and dictionary lookup [7] are used to

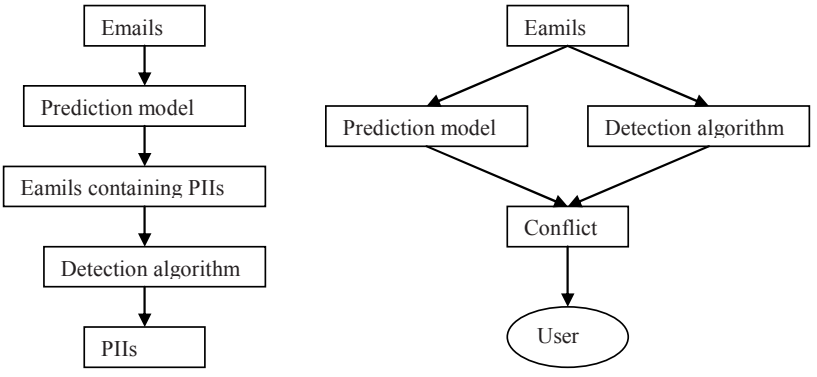
identify the addresses, and Luhn algorithm [13] is used to identify credit card numbers. These algorithms can achieve high precision and recall for high quality documentations, but they may obtain poor performance for text of low quality, like email (i.e. text containing many acronyms, abbreviations, and misspellings). For example, a misspelling "Avenu" cannot be found in the dictionary of street types (we only store "Avenue" and "Ave" in the dictionary as the street type). Therefore, the algorithm will fail to identify the address containing "Avenu". Secondly, the information in the dictionary may evolve. If the dictionary cannot be updated in time, we will fail to detect the PII. For example, if new area codes have been added for the telephone numbers, they have to be updated in the dictionary to make sure that telephone numbers with the newly added area code can be identified. Thirdly, it is very easy for the sender of the email to circumvent the detection system by simply inserting some characters in the PII such that it is easy for human to comprehend, and yet difficult to detect with algorithms (an approach used in spam email messages). For example, the sentence "Tom's credit card number is 5*1*8*1*3*4*5*6*4*5*6*7*5*6*7*8" contains a credit card number, but most algorithms can not detect the pattern with "*" inserted.

## 2 Prediction Methods

In this paper, we address the problem of predicting PII in email using association rule mining and classification model mining. We do not intend to identify the detailed private information. Instead, we only predict if there is a private entity occurring in an email based on the subject and content of the email. This predicting model can be integrated with the existing information extraction-based private information detection algorithms in two ways, as shown in Figure 1. In Figure 1(a), we use the prediction model to improve the efficiency of the private information detection for the audit process. In practice, only a small portion of the emails contain private information. In this case, we can apply prediction model first to identify the emails that contain the private information with a high probability. Then we apply more time consuming detection algorithms to identify the detailed private information.

Figure 1(b) shows the second integration. It combines prediction models and the detection algorithms to improve the recall and precision of the detection system. In the case that the detection algorithm fails to identify private information and the prediction algorithm gives a positive prediction, we present the results to the user and the user must check the email manually.

In this paper, we consider four types of PII elements, *email addresses*, *telephone numbers*, *addresses*, and *money*. We employed two models for prediction, association rules and classification models. The association rules represent correlation between two itemsets [1]. In our case, we want to find the correlation among these four types of the PII elements and use the correlation to predict one type of PII elements based on other types of PII elements already detected or predicted in an email. The second model is based on the classification model. It predicts the private information based on the subject and content of the emails.




(a) Filter Emails for Detection (b) Detect Conflict

**Fig. 1.** Integration of Prediction Model and the Detection Algorithms

### 2.1 Data Set

We worked on the Enron email dataset for our experiments. Enron was a U.S. corporation that conducted business in the energy sector from the 1980s until the early 2000s. With the rapid success of the business, Enron soon expanded their scope to include brokerage of a variety of commodities, including advertising time and network bandwidth. In 2001 the company collapsed due to accounting scandals. During the investigations that followed the collapse of the company, the Federal Energy Regulatory Commission made a large number of corporate email messages public. These emails have since been used as a useful source and benchmark for research in fields like link analysis, social network analysis, fraud detection, and textual analysis. There are different versions of the cleaned dataset. We chose the cleaned version from [12] to work on. The email dataset contains 252,759 messages from 151 employees distributed in around 3000 user defined folders. It is stored in a MySql database.

### 2.2 Using Association Rules to Predict Private Information

Association rule mining is a data mining approach which originated from the market basket problem. It discovers items that co-occur frequently within a data set. More formally, an association rule is defined in the following way [1]: Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items. Let $D$ be a set of transactions, where each transaction $T$ is a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \Phi$. The rule $X \rightarrow Y$ holds for the dataset $D$ with support $s$ and confidence $c$ if $s\%$ of transactions in $D$ contain $X \cup Y$ and $c\%$ of transactions in $D$ that contain $X$ also contain $Y$. The confidence and support are measures that ensure the patterns found are statistically accurate and significant. For example, {milk, eggs} $\rightarrow$ {bread} is an association rule that says that when milk and eggs are purchased, bread is likely to be purchased as well.

In private information detection, we chose the emails sent during the period of January 1, 1999 to December 31, 2000 for our experiments. The number of the emails is 73,817. We first applied the privacy algorithm implemented in SNAP [8] to identify the private information in the emails. Then we constructed a two dimensional table, in which each row corresponds to an email and each column corresponds to a type of private elements. Therefore there are four attributes in the table, each representing *email address*, *phone number*, *address*, and *money*, respectively. If an email contains an *email address*, we put 1 to the cell corresponding to the email and the attribute *email address*. Otherwise, we put 0. We did the same for three other types of private elements. We then applied Apriori algorithm implemented in Weka [11] to mine association rules. We set support to 0.1% and confidence to 60%. The rules mined are shown in the following in the descending order of confidence.

EMAIL=true ADDRESS=true ==> PHONE=true conf:(84%)

PHONE=true ADDRESS=true MONEY=true ==> EMAIL=true conf:(82%)

EMAIL=true ADDRESS=true MONEY=true ==> PHONE=true conf:(78%)

PHONE=true ADDRESS=true ==> EMAIL=true conf:(73%)

PHONE=true MONEY=true ==> EMAIL=true conf:(65%)

ADDRESS=true ==> PHONE=true conf:(62%)

The first rule states that if an email contains an *email address* and an *address*, it will also contain a *telephone number* with the probability of 0.84 according to the dataset. In this case, if an *email address* and an *address* are detected in a new email, but the *phone number* is not detected. It may be worth a manual check into the email, because it is likely that the detection algorithm failed to find a phone number which occurs in the email.

Note that the experimental results may not be applied to other email dataset in different companies or sectors, because different companies may deal with different kinds of private information. However, the approach itself may be applied generally to different situations.

### 2.3 Using Classification Models to Predict Private Information

The classification problem is one of the major tasks in the area of data mining and machine learning. To address a classification problem, a classification model is built according to observed examples, which are represented in a two dimensional table. Each row in the table represents an object and each column represents an attribute. There is one attribute called the decision attribute, which represents the classes of the objects. All the other attributes are conditional attributes which serve as the condition for the classification. When a new example comes, its values for conditional attributes are inputted into the classification model. The model will classify the example into an appropriate class. The most well known classification systems include: Bayesian networks, neural networks, classification rules, and support vector machines [9].

We convert the private information prediction problem into a text classification problem. The conditional attributes are the words occurring in the emails, which we call features. The decision attribute is the attribute that indicates if there is a private element of some kind in the email. We first pre-process the email messages in the following way. We scan subjects and content of emails, remove stop words, and stem words using Porter algorithm [10]. After this processing, 5585 words are identified. Next, we select a subset of these words as features to train the classification model.

We use two methods to select the features for training. The first feature selection method that we use is based on the frequency of the words, and therefore is unsupervised. In this method, we choose the most frequent words as the features. Figure 2 shows the distribution of the words. The X-axis denotes the frequency threshold for pruning. The Y-axis denotes the number of words left after pruning. The distribution coincides with Zipf's law. For our experiments, we choose 100 for the frequency threshold and obtain 578 words as features. We use three representations: binary, frequency, and weighted, to construct the table. In the binary representation, if an email contains a word, we fill in the corresponding cell with 1, otherwise with 0. In the frequency representation, we record the frequency of a word occurring in the email. In the weighted frequency, we set the weight of the words in subject to 3, and weight of words in the content to 1. For example, if "trip" occurs once in subject and twice in content, its total importance value will be 5.

The second feature selection method is based on entropy and information gain, which is a supervised method. Our classification problem is a binary classification problem. If an email contains a piece of private element, we set the decision attribute $C = 1$, otherwise, we set $C = 0$. If a word $w$ occurs in an email, we set condition attribute $w$ to 1, otherwise it is set to 0. The entropy of the data is defined as

$$entropy\ (D) = -P(C = 0)\log_2 P(C = 0) - P(C = 1)\log_2 P(C = 1),$$

where $P(C = i)$ = *the number of the examples with class i / the number of all examples* denotes the probability of occurrences of class $i$. The entropy of data given a word $w$ is defined as

$$entropy(D \mid w) =$$
$$-P(w = 0)(P(C = 0 \mid w = 0)\log_2 P(C = 0 \mid w = 0) + (P(C = 1 \mid w = 0)\log_2 P(C = 1 \mid w = 0))$$
$$-P(w = 1)(P(C = 0 \mid w = 1)\log_2 P(C = 0 \mid w = 1) + (P(C = 1 \mid w = 1)\log_2 P(C = 1 \mid w = 1))$$

where $P(w = 1)$ = *the number of the emails that contain word w / the number of all emails*, $P(w = 0)$ = *the number of the emails that do not contain word w / the number of all emails.* $P(C=0 \mid w= 0)$ denotes the conditional probability that an email does not contain private information given the word $w$ does not occur in the email. With the same convention, it is straightforward to define $P(C= 1 \mid w= 0)$, $P(C= 0 \mid w= 1)$, $P(C= 1 \mid w= 1)$.

The information gain for word $w$ is defined as

$$gain(w) = entropy(S) - entropy(S \mid w).$$

We chose the top 10% of words with the greatest information gain as features for training the classification models.

We chose C4.5 and SVM algorithms for the experiments. The C4.5 algorithm generates classification rules as classification models. It is an efficient algorithm and the

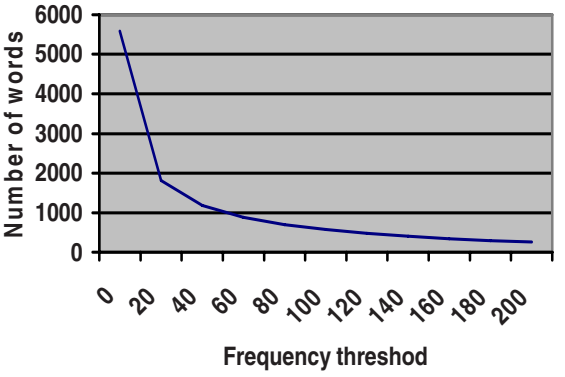



**Fig. 2.** Word distribution in terms of frequency

results offer easy comprehension. SVM is an optimization-based algorithm and is claimed to have good precision and recall on most datasets. We use the false positive rate (FPR) and false negative rate (FNR) to measure the performance of the classification modes. The false positive rate is defined as the ratio of the number of false positives to the number of all negatives. The false negative rate is defined as the ratio of the number of false negatives to the number of all positives. The reason that we use these two measures is due to the purpose of our task. In the case of the pre-detection, we do not want to miss any positive cases. Therefore, low false negative rate is desired. In the case of conflict detection, we do not want to provide too many false positive examples to increase the user's burden. Therefore, low false positive rate is desired.

We first did experiments for predicting *email addresses*. Table 1 shows the false positive rate. Table 2 shows the false negative rate.

**Table 1.** False positive rate for C4.5 and SVM

| FPR | Binary | Frequency | Weighted | Information gain |
|---|---|---|---|---|
| C4.5 | 4.4% | 3.8% | 3.6% | 3.0% |
| SVM | 5.4% | 1.8% | 1.6% | 3.2% |

**Table 2.** False negative rate for C4.5 and SVM

| FNR | Binary | Frequency | Weighted | Information gain |
|---|---|---|---|---|
| C4.5 | 55.3% | 55.9% | 54.9% | 56.3% |
| SVM | 44.4% | 64.7% | 61.0% | 48.2% |

From these two tables, we have several observations. First, the weighted representation always performs better than the frequency representation. Secondly, C4.5 is more insensitive to the representations than SVM is. Thirdly, the false negative rates are very high regardless of feature selection method. This is due to the fact that most of the examples are negative in terms of the private information. This makes our

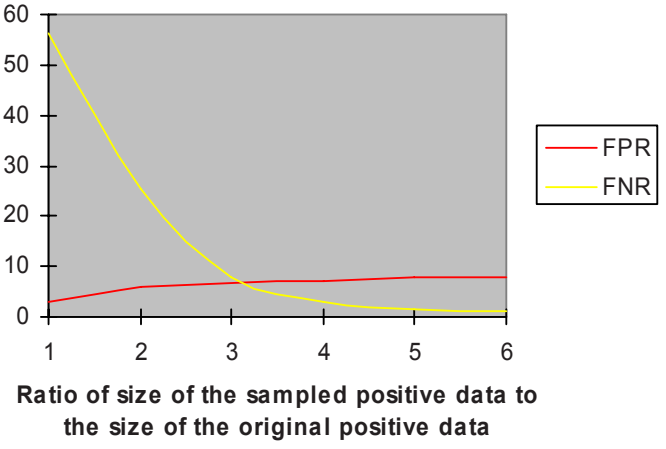



**Fig. 3.** Oversampling positive examples

classification problem an unbalanced one. To deal with this problem, we used an oversampling method to increase the number of the positive training examples. Figure 3 shows the false positive rate and false negative rate of the classification models learned from oversmapling.

The X-axis denotes the ratio of the size of the sampled positive data to the size of the original positive data. We can see that when the number of the positive samples for training increases, the false negative rate increases slightly, while the false positive rate decreases significantly. In this experimental setting, when we sample positive examples more than 5 times, there is no further improvement in either FPR or FNR.

We also did experiments for predicting *telephone number*, *addresses*, and *money*. We chose the C4.5 algorithm and the information gain feature selection method. We set the sampling ratio to 5. The experimental results are shown in Table 3.

**Table 3.** Oversampling for information gain feature selection

| | Email | Telephone no. | Address | Money |
|---|---|---|---|---|
| FPR | 7.9% | 7.8% | 1.9% | 2.7% |
| FNR | 1.5% | 1.6% | 0.4% | 1.2% |

We also experimented on the frequency selection method. The experimental results are shown in Table 4. We can see the information gain selection method produces smaller variances for FPR and FNR than frequency feature selection.

**Table 4.** Oversampling for frequency feature selection

| | Email | Telephone no. | Address | Money |
|---|---|---|---|---|
| FPR | 8.8% | 3.2% | 1.5% | 1.1% |
| FNR | 0.6% | 11.7% | 14.3% | 16.9% |

Finally, we tested the impact of the number of selected features on the performance. We chose C4.5 as classification algorithm and information gain as the feature selection method. We oversampled positive examples 5 times. The experiment was conducted on *email addresses*. The results are shown in Figure 4. The X-axis denotes the number of the features selected. The Y axis denotes the false positive rate and false negative rate in percentage. We can see that when we increase the number of the features from 50 to 250, the FPR decreases. When the number of features exceeds 250, the FPR fluctuates slightly. For the FNR, when we increase the number of features from 50 to 350, the FNR decreases. When the number of features exceeds 300, there is no improvement in FNR.

Based on the experimental results, we conclude that the sampling size and the number of the selected features have great impact on the performance of the classification models, while the data mining methods and feature selection methods do not show significant difference for the performance.

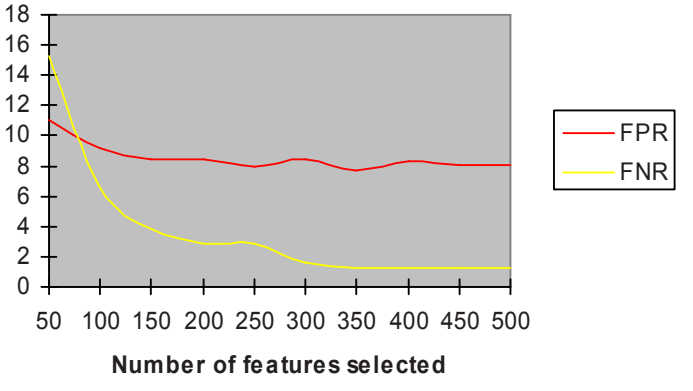



**Fig. 4.** Number of selected features verses FPR and FNR

## 3 Conclusions and Future Work

In this paper we presented a data mining-based method to predict the presence of personally identifiable information in emails. We adopted association rule mining to predict private information according to other PII identified. We used classification models to predict the PII according to the content of the emails. Experimental results on the Enron data set show that our methods can achieve satisfactory false positive and false negative rates.

Currently, we only predict if there is a PII element of a certain type occurring in email. In the future, we may predict how many private elements of a certain type occur. We will also incorporate more information from the email to improve the performance of the system. For example, since different people have different wording habits, incorporating the user information may improve the prediction results. We can also utilize email thread information. For example, an email from A to B at time 0 requests some private information, and another email at time 1, in the same thread, from B to A may provide the private information. In this case, the content of the time

0 email may help predict private information in the time 1 email. Also, in the future, we will try to automatically determine the parameters in training the classification model, such as the number of the features and the size of the samples. Finally, we will do experiments to compare our method with the related methods for private information discovery.

## References

[1] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, pp. 487–499 (1994)
[2] Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, pp. 439–450 (2000)
[3] Armour, Q., Elazmeh, W., El-Kadri, N., Japkowicz, N., Matwin, S.: Privacy Compliance Enforcement in Email. In: Canadian Conference on AI, pp. 194–204 (2005)
[4] Boufaden, N., Elazmeh, W., Ma, Y., Matwin, S., El-Kadri, N., Japkowicz, N.: PEEP - An Information Extraction base approach for Privacy Protection in Email. In: CEAS (2005)
[5] Carvalho, V.R., Cohen, W.W.: Preventing Information Leaks in Email. In: SDM (2007)
[6] Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy Preserving Mining of Association Rules. In: Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2002)
[7] Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic Document Metadata Extraction Using Support Vector Machines. In: Proceedings of the 2003 Joint Conference o Digital Libraries (JDCL 2003), pp. 37–48 (2003)
[8] Korba, L., Song, R., Yee, G., Patrick, A., Buffett, S., Wang, Y., Geng, L.: Private Data Management in Collaborative Environments. In: Luo, Y. (ed.) CDVE 2007. LNCS, vol. 4674, pp. 88–96. Springer, Heidelberg (2007)
[9] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Micheline Kamber Publishers (2006)
[10] Jones, K.S., Willet, P.: Readings in Information Retrieval. Morgan Kaufmann, San Francisco (1997)
[11] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
[12] http://www.isi.edu/~adibi/Enron/Enron.htm
[13] http://en.wikipedia.org/wiki/Luhn_algorithm