National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Private data discovery for privacy compliance in collaborative environments **

Korba, L., Wang, Y., Geng, L., Song, R., Yee, G., Patrick, A.S., Buffett, S., Liu, H., You, Y.
September 2008

* published in the Proceedings of the Fifth International Conference on Cooperative Design, Visualization and Engineering (CDVE 2008). Palma de Mallorca, Mallorca. September 21-25, 2008. NRC 50386.

Canada

Private Data Discovery for Privacy Compliance in Collaborative Environments[1]

Larry Korba, Yunli Wang, Liqiang Geng, Ronggong Song, George Yee,
Andrew S. Patrick, Scott Buffett, Hongyu Liu, Yonghua You

Institute for Information Technology, National Research Council of Canada
Building M-50, Montreal Road, Ottawa, Ontario K1A 0R6
{larry.korba, yunli.wang, liqiang.geng, ronggong.song, george.yee,
andrew.patrick, scott.buffett, hongyu.liu, yonghua.you}@nrc.ca
http://iit-iti.nrc-cnrc.gc.ca

Abstract. With the growing use of computers and the Internet, it has become difficult for organizations to locate and effectively manage sensitive personally identifiable information (PII). This problem becomes even more evident in collaborative computing environments. PII may be hidden anywhere within the file system of a computer. As well, in the course of different activities, via collaboration or not, personally identifiable information may migrate from computer to computer. This makes meeting the organizational privacy requirements all the more complex. Our particular interest is to develop technology that would automatically discover workflow across organizational collaborators that would include private data. Since in this context, it is important to understand where and when the private data is discovered, in this paper, we focus on PII discovery, i.e. automatically identifying private data existant in semi-structured and unstructured (free text) documents. The first part of the process involves identifying PII via named entity recognition. The second part determines relationships between those entities based upon a supervised machine learning method. We present test results of our methods using publicly-available data generated from different collaborative activities to provide an assessment of scalability in cooperative computing environment. .

Keywords: collaborative computing, privacy, compliance, text mining, machine learning, privacy management, personally identifiable information.

1 Introduction

As the cost of computers and networks have decreased, and with innovations in computing environments, there has been a dramatic increase in the use of networks and collaboration tools within all organizations today. Collaborative environments are facilitated by a myriad of software including: messaging tools such as email and chat, audio and video conferencing, file sharing systems, electronic whiteboards, desktop sharing, among other innovations. During the course of their work, employees may handle many different pieces of data. Some of this data may include different types of

[1] National Research Council Paper Number 50386.

sensitive personal information belonging to themselves, other employees, or customers. Within many organizations maintaining compliance with privacy legislation and/or organizational privacy policies is mandatory. Unfettered access to personal data allows easy collaboration, but increases the likelihood ofpersonal data leaks. In the face of the collaborative and distributed nature of work within organizations today, important challenges arise in finding and identifying the personal data subject to compliance. Our research involves the development of innovations in several technological domains to produce automated solutions that would allow organizations to locate the data they must manage, understand how it moves throughout the organization, and determine when it is being manipulated inappropriately, within a framework that secures against further data leakage [1].

Finding the private data is the important starting point towards building an automated privacy compliance solution. Ideally a solution would find private data whether it is at rest (on hard disc drives on the different computers across an organization), in motion (while it is being transmitted across the organization), and in use (for instance, when users type or copy sensitive private data). In this paper we focus our attention on the discovery of private data. Our objective for private data discovery is to develop ways to extract private data efficiently and effectively from unstructured and semi-structured content soas not to interfere with work activities. The private data may emerge from any type of computer-based activity, whether it is collaborative or not,

2 Our Approach

Private data discovery involves two steps: named entity recognition (NER) and relationship extraction (RE). In NER, privacy-related named entities are extracted. The semantic relationships between these entities associated with individual identity are extracted in RE. In effect, NER is a preprocessing step for RE. The left side of Figure 1 shows the original text in a document. As the first step of extracting private data, NER identifies each entity: Person, Address, Phone and Email. RE extracts the related entity pairs: Person-Address, Person-Phone, and Person-Email as PII. The result of both steps is shown on the right side of figure 1.

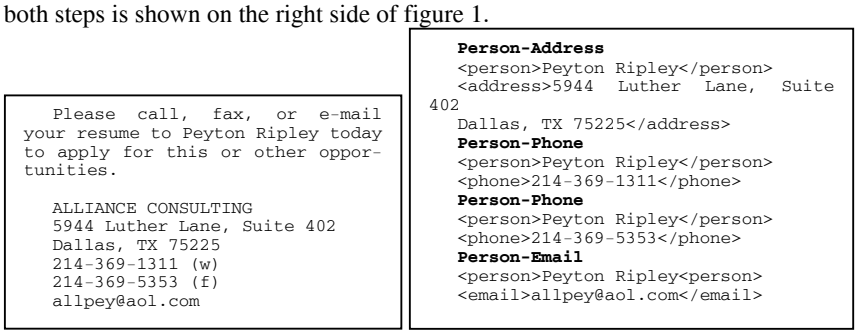

Figure 1. On the left is a section of a document, on the right is the result of private data discovery.

2.1 Named entity recognition

The named entities for private data discovery we extract are: Person, Organization, Email, Address, Phone number, Money, Date, Credit card number and Social insurance number (SIN). We extract the Address entity, not Location, since Address is more informative than a series of Locations. For example, "5944 Luther Lane, Suite 402 Dallas, TX 75225" in Figure 1 expresses three Location entities: "5944 Luther Lane Suite 402", "Dallas" and "TX"; however the Address entity is "5944 Luther Lane, Suite 402 Dallas, TX 75225".

Named entities are usually extracted by matching with patterns. We use a gazetteer and regular expressions for named entity recognition. A gazetteer is a list of names of people, organizations, locations and other named entities. A regular expression is a string used to describe or match a set of strings, according to certain syntax rules. For instance, a regular expression for a North American Phone number is [0-9]{3}-[0-9]{3}-[0-9]{4} (i.e. three digits-three digits- four digits), can be matched with two phone numbers 214-369-1311 and 214-369-5353 in **Figure 1**.

Regular expression matching is one of the common techniques for named entity detection. Using this technique, we are able to detect the entities: Email, Phone, Money, and Date. For detecting Credit Card Numbers (CCN) and Social Insurance Numbers (SIN), we used the Luhn's algorithm to verify strings of numbers that appear to be in CCN or SIN format [6]. We also include some keywords as part of our regular expression rules. We formulated some keywords for each entity. For instance, "Phone" and "Fax" are often used as keywords when introducing phone numbers. One regular expression rule is "Phone: [0-9]{3}-[0-9]{3}-[0-9]{4}", which means the keyword "Phone" precedes the phone number pattern. A priority level is defined for each rule in terms of its reliability of appearance in free-form text.

For complex named entities, such as Person, Organization and Address, we perform multi-level extraction.

- Gazetteers are collected. First name, last name dictionaries are used for extracting Person, company endings for Organization and City, Province (State), and Country names for Address.
- Regular expressions for each entity are defined. We use patterns of names [7] for Person, morphological regular expressions for Organization, and regular expressions or Street and PO BOX.
- Sequential patterns for each data type are defined and used in analysis, e.g. usually street name appears before city name and province name before country name.
- Some rules for distinguishing ambiguous strings are applied, e.g. "John Smith Ltd." is considered as a candidate for Person and Organization. A rule "person name + company ending = organization" applies resulting in a classification as Organization.

2.2 Relation extraction

The relation extraction for private data discovery is targeted for any relations that may identify a particular person, such as person-phone number, person-email, person-birth date, person-income, etc. as pair-wise relations. In this work we focus on ex-

tracting person-email, person-address, and person-phone number because of the frequency of appearance of such data in data sets as described in Section 3.

As the core part of information extraction (IE), pattern discovery is the task of identifying the extraction patterns. Patterns can be discovered automatically, semi-automatically or manually. Choosing the pattern for private data discovery depends on the characteristics of private data. Private data could be in semi-structured or within unstructured (free) text. It may be hidden anywhere in a file. There is no common template among these files, i.e. no metadata to provide hints of the presence of private data. Regular expression is one of the more common manual pattern discovery methods. Usually regular expression rules are sufficiently powerful for semi-structured text, especially template-based since we usually find common tokens surrounding the data to be extracted. For instance, most Web pages on the same Web site follow similar templates. Machine learning approaches: supervised, semi-supervised or unsupervised, are often used to automate relation extraction. These approaches are most successfully applied to template-based inputs. Techniques used in unsupervised IE systems are hard to extend to free text and even non-template inputs since many heuristics are applicable only to template-based inputs [8]. For the reasons described above, we used a supervised machine learning approach for the RE task in private data discovery.

We use a statistical machine learning model since the effectiveness of this method for relation extraction has been proven [9]. Usually the IE problem is translated into a classification problem in statistical learning methods. Decision tree, naïve Bayes and support vector machine (SVM) can be applied as statistical learning methods. We chose decision tree as the classifier for its good performance (execution speed) in various domains.

Feature selection is key for the performance of machine learning algorithms. Our task is to choose the feature sets that may work for extracting privacy related relations from semi-structured and free text. We use semantic, structural, and lexical features for relation extraction. For each pair of entities, various semantic, structural and lexical features are extracted. Semantic relations between two entities are determined using the decision tree algorithm.

To better assess the performance of these private data discovery features, we compare the following parameters and combinations of different feature sets in Section 3:

- Semantic features: entity type (e.g. Person, Email, Phone, and Address), entity sequence (i.e. the sequence of entities).
- Structural features: entity and word distance between two targeted entities.
- Lexical features: unigram, bigram, and trigram.

3 Results

For the purpose of testing private data discovery, we used different data sets containing privacy-related entities. The document header data set, available from Carnegie-Mellon University, is semi-structured and is considered highly-structured formal documents (i.e. research papers) [10]. The job posting data set [11] is also semi-structured but considered as informal documents from Usenet posts in jobs-related

discussion newsgroups. The Enron email data set [12] comprises a subset of email exchanges amongst employees of Enron. It is free text and is considered as informal documents, in this case, shared during work activities.

For our tests we used a randomized subset of these data sets. Some characteristics of these three data sets used in our tests are summarized in **Table 1**.

Table 1. Data sets used to test PII discovery

Data Set	Size	Input	Relation Extraction
Header	347K (246 headers)	Semi-structured	Person-Email Person-Address Person-Phone
Job posting	644K (85 messages)	Semi-structured	
Enron	1.2M (571 emails)	Free text	

A semi-automatic annotation method was used to generate training data for relation extraction. Entity recognition was used as a pre-processing step for relation extraction. Named entities were detected by scanning documents, followed by relation candidate extraction using the algorithm in **Figure 2**.

```
For each document d in data set D
    For each targeted relation R= (e1, e2)
        For each pair of adjacent entity (e1, e2) or (e2, e1) present in d,
            Extract content from d between these two entities, and tag them as relation candidates.
```

Figure 2. Algorithm for extracting candidate relations between entities.

Next, these extracted relations were manually classified into positive and negative sample cases (see **Table 2**).

Table 2. The number of positive and negative samples in three annotated data sets

	Person-Email		Person-Address		Person-Phone	
	Pos	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
Header	148	242	210	180	21	369
Job posting	87	130	75	123	30	84
Enron	376	766	39	48	176	92

Recall and precision are often used to measure the effectiveness of information extraction systems. In our approach, recall measures the ratio of correctly classified relations to all the positive relations. Precision measures the ratio of correctly classified relations to all classified positive relations. We used the F-score to determine the performance of the three data sets for relation extraction. F is the geometric mean between recall (R) and precision (P).

$$F = \frac{2PR}{P+R}$$

We used Weka [13] and the decision tree algorithm C4.5 in Weka to test the performance of these three data sets. As well, we used a 10-fold cross validation method. Testing was conducted in two steps: 1) test the performance of extracting combinations of features for each data set, and choose the best parameters; 2) test each rela-

tion in the three data sets using the best parameter and compare the effectiveness for different data sets. The result of the first step was the Person-Email discovery parameter.

3.1 Experiment 1

In the original Header data set, each header is composed of tagged entities: title, author, affiliation, email, phone, etc. Header data set represents semi-structured formal documents. We obtained tagged entities: Person, Email, Phone and Address from NER, then generated training data in a semi-automatic annotation process as presented in figure 2. Each header in the Header data set is only composed of entities, and there are no words between entities. Therefore, only semantic and structural features were extracted. In the Header data set, semantic features are entity type and entity sequence. Structural features are entity distance of a NE pair.

Using the entity type as a baseline, we tested the performance of combining entity sequence with entity type (**Table 3**). The results show that this approach provides a statistically-significant performance improvement as compared with entity type alone. The entity sequence is probably an informative feature set for semi-structured inputs with certain implicit template. Adding entity distance on entity type and entity sequence also significantly improves the performance.

3.2 Experiment 2

The Job posting data set represents semi-structured informal documents. We used it for extracting relations between private data and followed the same procedure as with the Header data set to get the training and testing data. Although the Job posting data set is also considered as semi-structured input, it differs from the header data set in that each relation candidate is composed of entities and words. Figure 1 is one section of a job posting message.

Still the entity type was used as a baseline. We tested the performance of combined entity sequence and entity type. We found no statistically-significant difference between combined features and entity type alone (the data is not shown here). This may indicate there is no sequential pattern or template among job posting messages. Therefore, for the Job posting data set, only entity type was used as semantic features, distance of entity and word between a NE pair was used as the structural feature, and unigrams were used as the lexical features (Table 3). Our emphasis was on the performance of lexical features in the job posting data set. For extracting lexical features, stemming was conducted. We tested the performance of combined semantic, structural and lexical features. The results show that the combined semantic and structural features and the combined semantic, structural and lexical features offer a statistically-significant, improvement in performance over using semantic features alone. The lexical features are very useful and improve the F-measure by 5.63.

3.3 Experiment 3

The Enron email data set represents informal documents comprised of free text. We used it to test the performance of semantic, structural and lexical features for free text

inputs and compare with semi-structured inputs. A randomized subset of the Enron data set was used and training data was generated in a semi-automated annotation process. Using the same approach as with the Job Posting data set, we used the entity types as the semantic features, distance as structural features, and word unigrams as lexical features for the Enron email data set (**Table 3**).

Unlike the other two data sets, the performance of combined entity type and distance is worse than entity type alone. This may be due to some noise in the free text inputs. Still, combined entity type, distance and word unigram features reach the best performance for the Enron data set. Lexical features contribute the largest to improved performance.

Table 3. Comparison of feature sets in three data sets (F-measure)

Feature set	Header	Job posting	Enron
(1)Entity Type	88.15	89.32	73.21
(1) + (2)Entity Sequence	90.92	-	-
(1) + (2) + (3)Distance	**91.82**	-	-
(1) + (3)	-	93.78	72.39
(1) +(3) + (4)Word	-	**99.41**	**88.68**

Table 4. Performance of relation extraction for three data sets

	Person-Email	Person-Address	Person-Phone
Header	91.82	97.21	83.78
Job Posting	99.41	98.66	100
Enron	88.68	83.59	94.05

The common trend of these three data sets is the combination of entity, word and distance feature sets to reach the best performance. We tested the performance of three relations using the combined feature set (Table 4). We observed that the effectiveness is influenced by both the input type and the training data. In general, the performance of free text input is worse than semi-structured inputs. However, there are some exceptions. Some relations such as person-phone in the Header data set and person-address in the Enron data set are significantly worse than others. It may be due to their quite small sample sizes (Table 2). The Header and Job Posting data sets are both semi-structured inputs, but the performance of the Job Posting data set is better than that of the Header data set using entity type as the feature set. This indicates that the performance obtained in one data set may not be generalized to other data sets. This is one of reasons why we tested the system on three quite different data sets. Within one data set, the performances of the three relations are different since the challenges of the tasks are different. The success of extracting private data can vary for different domains, task, format, and types of document collections. Nevertheless, we can still conclude that the combined semantic, structural and lexical feature sets reached the best performance in all three data sets.

4 Related Work

Our team has found no research performed exploring the detection of private data within documents in the context of collaborative work. Aura et al. proposed a method for detecting certain predefined PII for the purpose of retaining anonymity in scholarly manuscript review [2]. They only addressed situations where the author is the person who requires anonymity. Another difference between their work and ours is that Aura et al. extracted individual PII, but we extracted privacy related relations.

Other related work with private data discovery is in the area of IE. More specifically, two crucial and related IE techniques: NER and RE are used in private data discovery. NER involves the task of identifying entities such as Person, Organization and Location in text [3]. RE is the task of identifying semantic relationships between entities in the text, such as a person's birth date, which relates a person's name in the text to a date that is the person's birth date [3]. While many information extraction systems have been developed, to our knowledge, this is the first attempt to extract privacy related entities and identify relationships between these entities for the purpose of automating privacy management. RE systems used various features: syntactic, semantic, and lexical features [3-5]. In this study, we identify and compare the effectiveness of these features in extracting private data from semi-structured and free text.

5 Discussion and Conclusions

It is impossible to maintain privacy compliance in collaborative environments without the ability to determine when and where private data appears. Towards the goal of attaining this ability, this paper addresses the issue of extracting private data from semi-structured and unstructured documents. There are two steps involved in extracting private data: NER and RE. We used a supervised machine learning approach for RE, and tested its effectiveness for various feature sets using three different data sets. The results show the combined semantic, structural and lexical features are most effective for extracting relations within and across sentences. Personally identifiable information discovery based on this method is effective for both semi-structured and unstructured (free-text) inputs. We have applied these techniques in our prototypes to discover private data within files of different types (Word, PDF, Excel, and text), and within editing operations performed by computer users collaborating across an organization. The prototype software also correlates work activities across different users to discover and map collaborative work. The idea of "spying" on individuals and analyzing their work patterns compared to those of others may seem counterproductive to maintaining privacy compliance. Discovering private data on an individual by individual basis allows us to restrict our work pattern searches to activities that only touch private data, ignoring the rest and alleviating the amount of data collected and shared across the organization for the automated compliance process.

PII discovery is an important step in the process of automating privacy policy compliance verification. PII discovery indexes all personally identifiable information in a computer workstation, pinpointing the location, type of PII across all computers and computer-based activities within an organization and limiting compliance analy-

sis to situations involving private data. Clearly, automating PII discovery must include the means for protecting the discovered PII information. To this end, we are developing technologies for efficient key distribution, authorization, and access control to control access to the discovered PII. The techniques for discovering the workflow associated with private data across working groups must also be done in such a way as to prevent disclosure of either the PII being managed, or the nature of the PII workflow knowledge discovered. For this purpose, we are developing privacy-aware workflow discovery techniques [14].

References

1. L. Korba, R. Song, G. Yee, A.S. Patrick, S. Buffett, Y. Wang, L. Geng, "Private data management in collaborative environments", Proc. of the Fourth International Conference on Cooperative Design, Visualization and Engineering (CDVE 2007), Shanghai, China, September 16-20, 2007.
2. T. Aura, T.A. Kuhn, M. Roe, "Scanning electronic documents for personally identifiable information", Proc. of the Workshop on Privacy in the Electronic Society (WPES'06), Washington, DC, Oct 2006, pp. 41-49.
3. E Agichtein, S Cucerzan, "Predicting accuracy of extracting information from unstructured text collections", CIKM'05, 2005, Bremen, Germany, pp. 413-420.
4. N. Kambhatla, "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations", Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain, July 21-26, 2004.
5. S. Miller, H. Fox, L. Ramshaw, et al, "Description of the SIFT system used for MUC-7", Proc. of the 7th Message Understanding Conference (MUC-7), 1998.
6. Luhn's Algorithm on Wikipedia, last accessed: March 20, 2007, http://en.wikipedia.org/wiki/Luhn_algorithm
7. H. Han, CL. Giles, E. Manavoglu, H. Zha, Z. Zhang, EA. Fox, "Automatic document metadata extraction using support vector machines", Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03), Houston, Texas, May 27-31, 2003, pp. 37-48.
8. C.H. Chang, M. Kayed, M.R. Girgis, K.F. Shaalan, "A Survey of Web Information Extraction Systems", IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10), pp. 1411-1428.
9. J. Turmo, A. Ageno, N. Catala, "Adaptive information extraction", ACM Computing Surveys, 38(2): 4, July, 2006.
10. Headers data, available at: http://www.cs.cmu.edu/~kseymore/ie.html
11. Job posting data, available at: http://www.cs.utexas.edu/users/ml/index.cgi?page=resourcesrepo
12. Enron random subset, available at: http://www.cs.cmu.edu/~wcohen/
13. Weka, available at: http://www.cs.waikato.ac.nz/ml/weka/
14. Ronggong Song, Larry Korba, George Yee, An Efficient Privacy-Preserving Data Mining Platform, The 4th Int. Conf. on Data Mining (DMIN'08), Las Vegas, Nevada, July 14-17, 2008.