Privacy measures for free text documents: bridging the gap between theory and practice

Geng, Liqiang; You, Yonghua; Wang, Yunli; Liu, Hongyu

https://doi.org/10.1007/978-3-642-22890-2_14

Trust, Privacy and Security in Digital Business: 8th International Conference, TrustBus 2011, Toulouse, France, August 29 - September 2, 2011. Proceedings, Lecture Notes in Computer Science; no. 6863, pp. 161-173, 2011-10-01

Privacy Measures for Free Text Documents: Bridging the Gap between Theory and Practice

Liqiang Geng, Yonghua You, Yunli Wang and Hongyu Liu

National Research Council of Canada, Institute for Information Technology,
46 Dineen Drive, Fredericton, Canada E3B 9W4
{Liqiang.Geng, Yonghua.You, Yunli.Wang, Hongyu.Liu}@nrc,gc.ca

Abstract. Privacy compliance for free text documents is a challenge facing many organizations. Named entity recognition techniques and machine learning methods can be used to detect private information, such as personally identifiable information (PII) and personal health information (PHI) in free text documents. However, these methods cannot measure the level of privacy embodied in the documents. In this paper, we propose a framework to measure the privacy content in free text documents. The measure consists of two factors: the probability that the text can be used to uniquely identify a person and the degree of sensitivity of the private entities associated with the person. We then instantiate the framework in the scenario of detection and protection of PHI in medical records, which is a challenge for many hospitals, clinics, and other medical institutions. We did experiments on a real dataset to show the effectiveness of the proposed measure.

Keywords: Privacy compliance, ontology, privacy measure, personal health information.

1 Introduction

Privacy compliance has been an important issue that faces most organizations, as more and more privacy legislation and organizational privacy policies became mandatory. It is especially difficult, yet important, for organizations to reinforce privacy compliance on free text documents due to the following reasons. First, approximately 80% of corporate data is in free text format. Secondly, the free text documents are more easily accessed and transmitted than structured data stored in databases. Thirdly, technically it is more challenging to deal with privacy in free text documents where no data schema is available.

Natural language processing and machine learning techniques can be used to identify private entities, such as persons' names, email addresses, telephone numbers. heath records, and credit card numbers. Korba et al. proposed to use named entity recognition to identify private entities and use machine learning method to extract relations between the private entities [1]. That solution is based on the assumption that if one or more of the private entities and their proprietor's name are found in a document, the document is considered as containing private information. A drawback

of this method is that it may retrieve huge number of documents as containing private information as long as these documents contain a person's name and his/her private entities. However, among the retrieved documents, only a small proportion may be practically considered as real private information. Furthermore, this method only classifies the documents into two categories: containing private information and not containing private information. Sokolova and Emam did a similar work that proposed a two-phase approach to identify personal health information (PHI). In the first phase, personally identifiable information (PII) is detected. In the second phase, the PHI is detected. They also proposed two measures to evaluate their approach. [2]. However, the method does not measure the degree of private information contained in the medical documents either.

Defining privacy measures to evaluate privacy levels in documents has several advantages. First, privacy measures can be used as a standard for de-identification and de-sensitivity. De-identification and de-sensitivity require that privacy should be protected while as much information can be released for data analysis as possible, i.e., to balance the private information protection and the quality of information released for data analysis. For example, the user may set a privacy degree threshold to determine if a document may be released. If the value of the privacy measure for a document is above the threshold, the system should remove some private entities until the privacy measure of the document is below the threshold. Second, when huge numbers of documents containing privacy are detected, privacy measures can rank them so that the privacy experts may focus on documents with more serious private information.

In this paper we present a method to measure the private information in free text documents and to address the above-mentioned difficulties in practice. Section 2 reviews the related work. Section 3 presents the theoretic framework for measuring privacy in free text. Sections 4 uses PHI as a case study to show how this framework can be implemented in the case of personal health records. Section 5 presents preliminary experimental results. Section 6 concludes the paper.

2 Related Work

Much work has been done on privacy compliance for structured data, i.e., databases. In the scenario of databases, each record in a data table corresponds to the personal information for an individual person. The attributes of the table are classified into quasi-identifying attributes (QIA) and sensitive attributes (SA). QIAs are those that can be used to identify a person, for example, a person's name, address, and so on. SAs are those that contain sensitive information for a person, the disclosure of which may result in harm to the person. SAs include diseases a person has, credit rating a person receives, and so on. The concepts of *K*-anonymity and *L*-diversity were proposed as the standards for privacy information release. *K*-anonymity requires generalization of each record such that it is not distinguishable with at least *K* other records in terms of QIAs [3]. *L*-diversity requires that each equivalence class in terms of the QIAs contains at least *L* "well represented" SA values, so as to reduce the probability of a person's sensitive information disclosure [4]. An alternative standard

to deal with sensitive values is called *t*-closeness, which requires that the distance between the distribution of sensitive attribute values in each equivalence class and that of the attribute values in the entire table is no more than a threshold *t* [5]. Although these studies are focused on the procedures of de-identification and de-sensitivity, and did not explicitly mention the privacy measures, the number of the records in each equivalence class and the diversity of sensitive values in each equivalence class can be considered as factors to measure privacy content for each record in the databases.

In the case of free text, Al-Fedaghi proposed a theoretical definition for measuring private information [6]. In that framework, every assertion involving a person is considered as a unit of private information. The privacy index regarding each person then is defined as $\frac{priv\_u + 1}{priv\_k * pers}$, where *priv_u* denotes the number of units of the person's private information unknown to others, *priv_k* denotes the number of units of his/her private information known to others, and *pers* denotes the number of the persons that know his/her information. This measure is not suitable for practical implementation, since it is impossible to estimate what portion of a person's private information is known or unknown to other people, and how many people know his/her private information.

Fule P. and Roddick proposed a practical method to evaluate the sensitivity of the privacy in the rules obtained from data mining [7]. They specify a sensitivity value for each attribute or attribute-value pair in the rules and proposed various combination functions to calculate the sensitivity values for the rules. However, their approach requires that the sensitivity value for each attribute or attribute-value pair must be specified by users. This is not practical for domains with huge number of attribute-value pairs. Also it does not consider the semantic relationship between the attribute-value pairs.

Literature survey shows that far less research has been done for measuring privacy in free text than in databases. This may be due to the difficulties for measuring privacy in free text documents. First, in databases, the probability of identifying a person in a table is solely based on the information within the table itself. In the case of free text, we have to resort to external sources to determine this probability. Secondly, in the databases, the data is structured and the QIAs and the SAs are already known. In the free text, the sensitive information has to be identified first using some technologies, such as information extraction. Thirdly, sensitive information in free text may involve different entity types, and more sensitive values may be derived based on the information occurred in the documents. For example, with some basic medical knowledge, adversaries can infer that a person is infected with AIDS if cocktail treatment is mentioned in his/her medical record, even if AIDS is not explicitly mentioned.

3 Framework for Privacy Measures

As in the database scenario, we identified two factors for determining privacy degree in free text: quasi-identifying entities (QIE) and sensitive entities (SE). The QIEs refer

to the entities that can be used to identify a specific person. They include persons' names, genders, ages, races, weight, height, addresses, and so on. The more likely a set of QIEs can uniquely identify a person, the more privacy these QIEs contain. For example, the statement "Tom is HIV positive" has lower privacy degree than "Tom, who lives in Yonge Street, Toronto, is HIV positive", because the second statement has one more QIE "Yonge Street, Toronto" which may reduce the scope of the candidates and further help identify the person "Tom". Similarly, the statement "David is HIV positive" has lower privacy degree than "Burt is HIV positive" because fewer people are named Burt than those named David.

Formally, let n denote the number of the persons in the universe matching the QIEs in a document. The probability of identifying a particular person satisfying the QIEs is $1/n$. The difficulty in calculating this probability is that there is not a table available that contains all personal information for all the individuals in the world. In the next section, we propose to use web search engines to obtain an estimate for this probability.

The SEs may include diseases, medication, bank account numbers, bank account passwords, religions and so on. The degree of sensitivity for SEs entities can be both objective and subjective. For example, the statement "John is diagnosed with heart disease" is more sensitive than "John suffers from a cold" in the sense that the former statement may incur more personal loss for the individual, such as the increased life insurance premiums and reduced employment opportunities. In this sense, the sensitive degree can be evaluated with objective measures. On the other hand, whether the statement "John was put into the prison" is more sensitive than "John suffers from AIDS" depends on social and cultural factors that may not be objectively measured. In our framework, we consider the degree of sensitivity to be subjective and determined by the privacy experts, since it is difficult to define comprehensive objective measures on different types of SEs. To overcome the bias of the subjective sensitivity values from an expert, a practical way is to let a few privacy experts assign the values to the SEs independently and resolve the inconsistency through discussion.

When the number of SEs in a domain is huge, it is not practical to ask the experts to manually assign the sensitivity values for all the SEs, therefore an ontology is desirable to provide the degree of sensitivity for each SE and to conduct inferences among these entities.

Although the sensitivity values for the SEs are to be determined by the users, the assignment of these values should not be arbitrary due to the semantic relationship between these entities. We propose some principles for ensuring consistency between the SEs.

Let A, B, and $C \in SE$ denote the SEs, which are organized in an ontology. Let S denote the degree of sensitivity, which is a function S: $2^{SE} \rightarrow [0, 1]$. We define five principles for assigning sensitivity values as follows.

1. $0 \le S(A) \le 1$
2. $A \le B \Rightarrow S(A) \ge S(B)$
3. $\max(S(A), S(B)) \le S(A, B)$
4. $A \le B \Rightarrow S(A, B) = S(A)$
5. $A \le B \Rightarrow S(A, C) \ge S(B, C)$

Principle 1 specifies that a sensitivity value should be a normalized positive real value between 0 and 1, with 0 representing no privacy and 1 representing highest degree of privacy. Principle 2 states that if entity *A* is more specific than *B* in the ontology, *A* is considered to be more sensitive than *B*. For example, date of birth is more sensitive than the year of birth. Principle 3 says that the combination of two sensitive entities is more sensitive than or equally sensitive with the maximum sensitivity of each of them considered alone. For example, the combination of the bank account number and password are far more sensitive than each of the two entities alone. Principle 4 states that if between two entities, one is more general than the other, the former does not contribute to the overall sensitivity. Principle 5 states that a more general entity introduces less sensitivity than a more specific one when combined with other entities.

Principles 1 and 2 specify the consistency among the entities defined in ontology, which may be represented in a tree or a graph. The other three principles are useful in deriving sensitivity values for compound entities. A composition function *f* is needed to calculate the sensitivity values for compound entities. For example, suppose we set *S*(*diabetes*) = 0.6, *S*(*heart attack*) = 0.9. Then *S*(*diabetes*, *heart attack*) = *f*(*S*(*diabetes*), *S*(*heart attack*)) may yield a sensitivity value of 0.95.

Adversaries usually have background knowledge and could conduct inferences on the SEs in the documents. For example, if *cocktail treatment* is mentioned as a medical procedure for a person, it is highly likely that this person was infected with *AIDS*.

Let *D* denote a document and *Ent*(*D*) denote the SEs contained in the document. By applying inference rules, we can get the closure of *Ent*(*D*), denoted as *Closure*(*D*). Then we can calculate the sensitivity value *S*(*Closure*(*D*)) for the document *D*.

The procedure for calculating the privacy measure for a free text document is as follows.

1. Preprocess: Extract QIEs and the SEs in the document, identify relations between entities.
2. Calculate the probability *p* that a person can be identified with the QIEs.
3. Use inference rules and ontology to obtain the closure of SEs for the document.
4. Remove the entities that are a more general entity of another entity in the closure.
5. Calculate the sensitivity value *s* for the closure
6. The privacy measure is calculated with *privacy* = *p* * *s*.

It should be noted that theoretically the QIEs and SEs are not necessarily exclusive to each other. For example, *date of birth* may be used as a QIE to identify a person and it also can be considered as a SE that may be used for fraud.

4 Calculating Privacy Measures for PHI

In this section, we use PHI as an example to illustrate the implementation of the proposed framework for calculating privacy measures.

4.1 Using WEB Search Engine to Estimate the Probability of Identifying a Person

When we calculate privacy measures, a difference between the database scenario and the free text scenario is that the former (for example, *l*-diversity for database) assumes that the adversary has the background knowledge about QI information of the target person, and also knows that the person's information is stored in the database table, while the latter assumes that the adversary only knows some QI information of the target person and does not know whether the file matching the QI information refers to the target person. Therefore, we need to model the adversary's background knowledge about the demographic statistics for the free text scenario. This is one of the challenging tasks for defining the privacy measures for free text documents. Using published demographic database for the modeling may be a solution. However, there are two problems. First, the published demographic databases are usually generalized. If we use them, we have to calculate the estimates of the real distributions at the more detailed level. For example, in [8], distribution over date of birth is estimated based on the real distribution over year of birth. However, this calculated distribution may be distorted from the real one. Secondly, no tables contain all kinds of QI information that can be identified from a free text document, such as color of hair.

Inspired by [9], [10], we adopt the Web as a knowledge base to estimate the probability of uniquely identifying a person given QIEs. First, the QIEs are identified. Then all the QIEs are concatenated as a string delimited by spaces. Finally it is submitted as keywords to a search engine, such as Google, to retrieve the number of the web pages containing these keywords. We use the inverse of the number as the estimate of the probability. This probability is not accurate for any inference, but would be enough to represent the relative strength to rank the QI information.

At the current stage, we do not take into account the information in the related documents when we calculate the privacy measure for a document.

In our study, we consider the following QIEs which are directly associated with a person: *name*, *age*, *date of birth*, *telephone number*, *email address*, *address* and *gender*. Other entities that may help identify the person, but are not directly associated with the person, including person's parents' names, spouse's name, time of admission to a hospital, travel date, and so on, are not considered in our work.

4.2 Calculating Sensitivity of Diseases

Some studies use information gain obtained by information disclosure to measure the sensitivity of privacy. For example, Lonpre and Kreinovich [11] used the financial loss to measure the sensitivity of diseases. However, in order to calculate the information gains and utility losses, we must know the related probability distribution for all diseases and financial losses due to disclosure of the diseases. This is practically impossible. Kobsa argues that the further a value is from the normal value, the more privacy the value contains [12]. Also he argues that entities with lower probabilities are more sensitive than the entities with higher probabilities because they can be considered as anomalies [12]. However, this model also needs all probability

distributions among each kind of the entities, and hence it is not practical for implementation.

In this study, we consider sensitivity levels of SEs as subjective because it is determined by social, economic, and cultural factors. For example, a person's age is considered as privacy in North America, but maybe not in some Asian countries. Therefore, we ask the user to specify the sensitivity values for each disease. Our solution consists of three steps. First, the user specifies the sensitivity values to medical terms in an ontology. Then the system extracts medical terms from a document and maps them to the ontology to get the sensitivity values for the terms. Next, the system uses inference and aggregation to calculate the sensitivity value for the document. We used MeSH in our case study. MeSH is a medical ontology that records terms for diseases, medications, procedures, etc. and shows the relationship between them [13]. The terms in MeSH are organized in a tree with root node representing the most general concept and the leaf nodes representing the most specific ones. Our goal is to associate a sensitivity value for each disease in MeSH. Since currently there are more than 10,000 concepts representing diseases in MeSH, it is tedious to assign the values for all diseases. We first specify the default values for all the disease to 0. Then the user can change the default settings for the diseases that are more important from privacy perspective. For example, the user may change the sensitivity value for AIDS to 1.0 and that for lung cancer to 0.9. After the new values are specified, they will be propagated to other concepts. The propagation of sensitivity values should observe the principles proposed in Section 2. We propose an algorithm for sensitivity value propagation, which is shown in Figure 1.

The algorithm first checks the consistency of the initially assigned sensitivity values, i.e., the sensitivity value of a parent node should not be greater than that of a child node. Then the algorithm propagates the sensitivity values upward. Finally, it propagates the sensitivity values downward to populate the entire tree.

We can prove that if the initial assignment is consistent (satisfying principle 1 and 2), the sensitivity values obtained from our propagation algorithm will also satisfy principles 1 and 2. It is straightforward that downward propagation observes principles 1 and 2. We only need to prove that two principles also hold for upward propagation.

Proof using induction:

It is straightforward that the first propagation observes principles 1 and 2.

Suppose that the first k propagations observe principles 1 and 2. For the (k+1)th propagation, we only need to prove that in the chosen path, the top node t's sensitivity value is less than its bottom node p's sensitivity value (Figure 2). Suppose t's sensitivity value was obtained by propagation from another node s to node q, which means that t is between s and q. Since path (s q) was chosen over path (p q), according to the algorithm, the increment in the path (s q) is smaller than the increment over the path (p q). We have $\frac{s_p - s_q}{d(p,q)} \geq \frac{s_s - s_q}{d(s,q)}$. We also have $\frac{s_t - s_q}{d(t,q)} = \frac{s_s - s_q}{d(s,q)}$ when it propagates sensitivity values from s to q. Combining these

observations, we have $\frac{s_p - s_q}{d(p,q)} \geq \frac{s_t - s_q}{d(t,q)}$. Since $d(p, q) > d(t, q)$, we have $s_p > s_t$, hence principles 1 and 2 are satisfied.

```
PrivacyValuePropagation(T: a MeSH tree; N: nodes in T. S:
nodes that have obtained sensitivity values){
  if (ConsistencyCheck( )){
    UpwardPropagation( );
    DownwardPropagation( );
  }
}
UpwardPropagation(){
  For each s ∈ S
    Find ancestors sa ∈ S such that there is at least one
    node sp∉S  between s and sa and there is no sp ∈ S
    between s and sa
  Put all the pairs (sa, s) in set R
  while R is not empty
    For each pair (sa, s) ∈ R
      Inc(sa,s) = (s.sensitivity - sa.sensitivy)/length(sa, s)
    Inc = min(Inc(sa,s) )
    (sa1, s1) = argmin(Inc(sa, s))
    For each sp1 between sa1 and s1,
      sp1.sensitivity = sa1.sensitivity + Inc * length(sa1,
      sp1)
    R = R - {(sa1, s1)}
    For each pair (sa1, s) ∈ R
      Find sa2 between sa1 and s1 to replace sa1 in (sa1, s)
      such that sa2 is a newly labeled node and sa2 is the
      closest labeled ancestor of s.
  }
}
DownwardPropagation(){
  Traverse the tree in a breadth first fashion.
  For each non-updated node
    update sensitivity value with the value of its parents
     node
}
CheckConsistency(){
  consistency = true
  Traverse the tree in a breadth first fashion
  for each node p, find its child node c{
    if c is labeled with a sensitivity value
      if p.sensitivy ≥ c.sensitivy
        consistency = false
    Else
        c.sensitivy = p.sensitivity
  }
  return consistency
}
```

Fig. 1. Propagation of sensitivity values in MeSH.

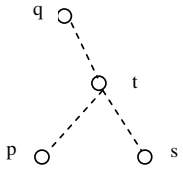
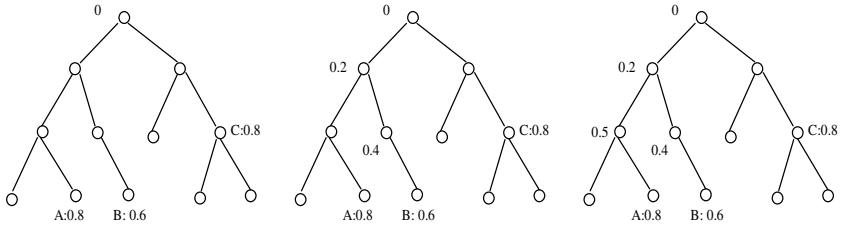


Fig. 2. Proof of conformance to Principles 1 and 2.

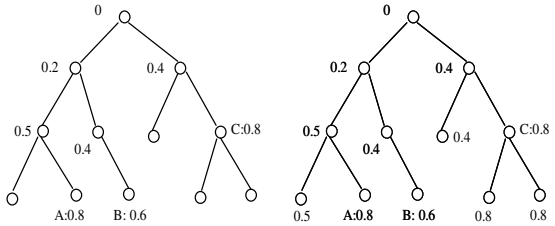
We use an example in Figure 3 to illustrate the propagation steps

Initially we set sensitivity values for nodes *A*, *B*, *C* to 0.8, 0.6, and 0.8 respectively (Figure 3(a)). First, node *B* was chosen for upward propagation (Figure 3(b)), then node *A* and node *C* were chosen in a sequence (Figures 3(c) and 3(d)). Finally, the downward propagation was done (Figure 3(e)).

To make the propagated values accurately reflect the real sensitivity levels for the user, the rule of thumb on which nodes the user should assign initial sensitivity values is that the nodes with significant difference with parent or sibling nodes should be specified.



(a) Initial assignment (b) Upward propagation from node *B* (c) Upward propagation from node *A*



(d) Upward propagation from node *C* (e) Downward propagation

Fig. 3. A propagation example.

PHI may contain other medical terms in addition to diseases. If some medical documents contain medical procedures or medication, it is easy for the adversaries to infer the diseases that are closed related to these procedures and drugs. Using web search engine to find correlation between sensitive keywords [10] provides a solution for inference between medical procedures, medication, and diseases. However, in the

PHI scenario, ontologies are readily available. Using an ontology for inference can provide more accurate results than using web search. In MeSH, the entities are classified into different categories, which include *Disease*, *Medication*, and *Procedure*. In this study, we only take into account these three categories. We only assign sensitivity values to the disease in the ontology. The medication and procedure are used to infer diseases.

Let R_i denote a medication or procedure and D_i denote a disease

From MeSH, we can obtain rules in the following format

$$R_i \rightarrow \{(D_{i1}, p_{i1}),...\{D_{in_i}, p_{in_i})\}$$

This rule states that medication or procedure R_i may infer diseases D_{i1} with probability p_{i1}, and so on. Suppose disease D_{ij} has sensitivity value S_{ij}. The sensitivity value for R_i is determined by

$$S(R_i) = \sum_{j=1}^{n_i} S_{ij} * p_{ij}$$

When multiple medications and procedures are found in a document, we need an aggregation method to calculate the sensitivity value for the set of terms.

The method we used to calculate the aggregation is shown in Figure 4. It is straightforward that this aggregation algorithm satisfies the Principle 3.

Input: n rules $R_i \rightarrow \{(D_{i1}, p_{i1}),...\{D_{in_i}, p_{in_i})\}$, where $0 < i \leq n$

Rank pairs (D_{ij}, p_{ij}) according to $S(D_{ij})$ in a descending order

```
Prob = 0
Num = 0
While prob <1, in the ordered list do{
  prob = prob + Dij
  num = num +1
}
Adopt the top-num pairs for aggregation
Adjust the num-th probability with 1-prob
```

Calculate sensitivity value with $S(R_i) = \sum_{j=1}^{n_i} S_{ij} * p_{ij}$

Fig. 4. Sensitivity value aggregation method.

5 Experimental Results

We conducted our experiments on a dataset downloaded from an online health discussion forum Dipex (http://www.dipex.org.uk/), which consists of 250 messages. The messages posted on this forum do not have real names of patients, but they have nicknames such as "Paul123" and other identifiers such as locations, email addresses, phone numbers, posted dates and times. The advantages of using this real dataset are that the data is anonymous and they are publically available.

We used the PHI detection system described in [14] to pre-process the data, i.e., to identify personally identifiable information and the medical terms. Then we manually checked and rectified the results that became the input for our method.

Then we set the sensitivity values for 5 diseases in MeSH as shown in Table 1.

Table 1. Sensitivity values assigned to five diseases.

Disease	Sensitivity Value
Infection	0.2
Arthritis	0.5
Sarcoma	0.9
Cough	0.3
Anaemia	0.6

The system propagated these values in MeSH and calculated the privacy measures for all the 250 messages. We set the threshold to 5×10^{-5} to classify the messages into 175 PHI and 75 non-PHI. We also manually reviewed and classified the messages into 155 PHI and 95 non-PHI as golden standard. Then we compared the results from the system with the golden standard. Table 2 shows the confusion matrix.

Table 2. Confusion matrix.

	Positive Identified	Negative Identified
Real Positive	131	24
Real Negative	44	51

We have Precision = 0.749, Recall = 0.845, and F = 0.794.

We then took a look at the ranked messages and found that the top ranked messages indeed contain more privacy information than the other messages. For example, the message ranked first contains the disease names such as *ovarian cancer* and *tumors*. It also contains medication terms such as *Doxil* and *Gemcitabine*. This leads to a high aggregated sensitivity value of 0.99. The author of this message had used a very uncommon nickname which generated the probability that the person can be identified to be 6.6×10^{-4}. This probability value was calculated by calling the Google search engine.

The experiment was conducted on a PC with Intel Core 2 duo CPU of 2.20GHz and memory of 3.25GB. The sensitivity value propagation in MeSH took 69 minutes and 8 seconds. The privacy index calculation for 250 files took 185 minutes and 6 seconds.

6. Conclusion

We proposed a general framework for defining privacy measures for free text documents. We also proposed principles for evaluating sensitivity levels of private information. We then proposed a practical solution to estimate the privacy levels in the PHI scenario. Preliminary experimental results show the effectiveness of our approach.

For the database scenario, there is no problem about correlation between QIEs and SEs, because each row in a table refers to one person and the relationship is already embodied in the tables. However, in the case of free text, correlations between QIEs

and SEs are a practical problem. In our experiments, we assumed that the correlations between QIEs and SEs have been perfectly identified. Identifying the correlations between these entities is our future work.

We also assume that each file contains PHI or PII for only one person. This is the case for many health records. However, in a more general scenario, one document may contain privacy for several persons. Defining privacy degree for this situation is also a future work.

Another future work is to detect correlation for different documents that may contain private information for one person. In this scenario, the combination of the private information in different documents may disclose more private information.

References

1. Korba, L., Wang, Y., Geng, L., Song, R., Yee, G., Patrick, A.S., Buffett, S., Liu, H., You, Y.: Private Data Discovery for Privacy Compliance in Collaborative Environments. In: Proceedings of the Fifth International Conference on Cooperative Design, Visualization and Engineering (CDVE 2008), pp 21--25. Palma de Mallorca, Spain (2008)
2. Sokolova, M., Emam, K.: Evaluation of Learning from Screened Positive Examples. In: Proceedings of the 3rd Workshop on Evaluation Methods for Machine Learning, in conjunction with the 25th International Conference on Machine Learning (ICML2008). Helsinki, Finland (2008)
3. Sweeney, L: K-Anonymity: a Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems. 10(5), 557--570 (2002)
4. Machanavajjhala A., Gehrke, J., Kifer. D.: l-Diversity: Privacy Beyond k-Anonymity. In: Proceedings of the 22nd International conference on Data Engineering (ICDE 2006), pp. 24. Atalanta, USA (2006)
5. Li, N., Li, T., Venkatasubramanian, S.: Privacy Beyond k-Anonymity and l-Diversity. In: Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007), pp. 106--115. Istanbul, Turkey (2007)
6. Al-Fedaghi S.S.: How to Calculate the Information Privacy. In: Proceedings of the Third Annual Conference on Privacy, Security and Trust, pp. 12--14. St. Andrews, Canada (2005)
7. Fule P., Roddick, J.F.: Detecting Privacy and Ethical Sensitivity in Data Mining Results. In: Proceedings of the Twenty-Seventh Australasian Computer Science Conference (ACSC2004), pp. 159--166. Dunedin, New Zealand (2004)
8. Golle P.: Revisiting the Uniqueness of Simple Demographics in the US Population. In: Workshop on Privacy in the Electronic Society (WPES'06), pp.77--80. Alexandria, USA (2006)
9. Chow R., Golle, P., Staddon J.: Detecting Privacy Leaks Using Corpus-based Association Rules. In: Proceedings of KDD'08, pp. 893--901. Las Vegas, Nevada (2008)
10. Staddon, J., Golle, P., Zimny, B.: Web-Based Inference Detection. In: Proceedings of the 16th UNENIX Security Symposium, pp. 71--86. Boston, MA (2007)
11. Lonpre, L., Kreinovich, V.: How to Measure Loss of Privacy. http://www.cs.utep.edu/vladik/2006/tr06-24.pdf
12. Kobsa A.: Privacy-Enhanced Web Personalization. In: Brusilovsky et al. (eds.) The Adaptive Web: Methods and Strategies of Web Personalization. Springer Verlag (2007)
13. U.S. National Library of Medicine. http://www.nlm.nih.gov/mesh/
14. Wang, Y., Liu, L., Geng, L., Keays, M.S., You, Y.: Automatic Detecting Documents Containing Personal Health Information. In: Proceedings of AIME 2009, pp335--344. Verona, Italy (2009)