LEXICAL ENTAILMENT FOR PRIVACY PROTECTION IN MEDICAL RECORDS

Lila Ghemri, Raji Kannah
Department of Computer Science
Texas Southern University
Houston, Texas
ghemri_lx@tsu.edu
rajikannah@gmail.com

Abstract— Most privacy preserving data mining methods apply transformations to the data that result in the loss of original data and reduces the effectiveness of the underlying mining results. Our goal in this work is to define privacy preserving methods that would reduce the difference between the mining results obtained with the original data and the "anonymized" data. We propose to use lexical entailment in order to replace specific features in the text with a semantically close category, this way the general meaning of the document is not overtly modified and the mining results will still be mostly valid.

Keywords: Text Mining; privacy; lexical entailment, clustering, medical records.

1. INTRODUCTION

Medical records are usually in a text format. Although, directly identifying information, such as the patient's name, Social security number, address and so on are usually blacked out or replaced with dummy strings, other information, particularly related to treatment and medication is usually left untouched in publicly available medical datasets. However, these attributes, considered as pseudo-identifiers can be used in conjunction with other records to uniquely identify a patient as shown in [1], [2] and [3].

In this work, we propose to investigate the use of lexical entailment as the basis for developing privacy de-identification methods in medical text, especially in patients' records. The main motivation behind this work is the need to comply with HIPAA (Health Insurance Portability and Accountability Act) requirements on preserving patient's privacy before making the data public [4]. A second motivation is the need to maintain some kind of semantic integrity to the document being anonymized so that the mining results still have some validity

2. PRIVACY-PRESERVING DATA TRANSFORMATION METHODS

Data transformation or perturbation techniques refer to modifications on the data. The goal of performing such data transformation is to disguise the sensitive data contained in the published datasets. On the other hand, these transformations are expected to preserve the data that is critical so as to build meaningful data mining models. Consequently, any data transformation technique has to be take into account these two competing goals: Hiding sensitive data and preserving meaningful mining results[5].

3. CHALLENGES WITH TEXT MINING

While most of the work on privacy-preserving data mining has focused on numerical and categorical data stored in databases and ways to preserve the underlying sensitive information coded in these data; mining text presents its specific challenges. Text is usually too high dimensionally to work effectively with standard anonymization techniques. Furthermore, unlike databases, in which each column is labeled with its meaning, it is difficult, in a piece of text, to select the words (attributes or features) that should be anonymized or hidden [6].

4. LEXICAL ENTAILMENT

Lexical entailment is a semantic relation that holds between lexical elements when the meaning of one element can be inferred from the meaning of the other [7].

This work was supported by a grant from the US Homeland Security Department for DHS Center for Dynamic Data Analysis

4.1 Linguistic Semantic Relations:

Linguistic semantic relations, such as synonyms, hypernyms, hyponyms, have been widely used in natural language processing applications, such as automatic ontology building [8], information extractions, question-answering systems[9], etc. In our work, we use the linguistic semantic relations to modify the data by generalizing it; and therefore focus on the *hypernym* relation.

4.2 Definition

A term **w** is a hypernym of another term **u** if **w** has an extensive meaning that forms a category under which **u** and other more specific terms fall. For instance, a *dog* is a hypernym of *terrier*.
Converse: If **u** is a hyponym of **w** (and **w** is a hypernym of **u**) then **u** is a *kind of* **w**. This relation is also called the *IS-A* relation. In this relation the hypernym **w** can often replace its hyponyms while preserving and generalizing the meaning.
For instance, replacing *terrier* with its hypernym *dog* in the sentence "The boy played with the terrier" still preserves the meaning of the sentence, even though some details are lost.

5. DATA DESCRIPTION

Our dataset consists of 2048 Intensive Care Unit (ICU) records which were collected by the Intensive Care Unit Safety Reporting System (ICUSRS) and stored in a publicly available database. The dataset describes adverse events in ICU. An adverse event is an injury or harm related to (or from) the delivery of care as noted in [10].

Although each record consisted of several fields, we collected a single field labeled Description. This is a textual field that describes, in the medical personnel own words, the ICU adverse event. They are usually filed either by the registered nurse or the doctor on shift. An example of such field content is:
"Dobutamine and heparin infusing via PIC line. Y site occluded with white precipitate. PT. noted to be diaphoretic and weak. Drugs had been running for several days. Dobutamine mixed with NS since 1/14/04."

6. DATA MODIFICATION RULES

As noted above, it is complex in such texts to know what information should be anonymized and what words should be generalized. Obviously, generalizing each word is not a workable solution. To solve this problem, we referred to what HIPAA considers as private information. According to [4], HIPAA mandates the protection of patients' medical records, especially patient identifiers, medical conditions, treatments, and procedures. While most directly identifying data is hidden in document, information related to medical conditions, or treatment and procedure is usually left intact. We, therefore, focused on anonymizing text that relates to treatment and especially expressions for medications administered and their dosages. The first step in this work was to detect the expressions that indicate dosages and/or drug names.

6.1 Dosage Recognizer

A sample of 300 records was examined in order to find expressions that denote dosages. Most dosages were found to contain a series of digits followed by units, such as 5 mg,
100 units/day, 45 mg/hour, 100 units/100cc, .2units, 0.3mL, 5kg, 100meq and so on. Since, there was a strong regularity in these patterns, we developed a grammar that recognizes dosage expressions.

<u>The Dosage Expressions Grammar:</u>

S → NA|B|C|N/B|N-NA
N → D|DN|F|N|ε
B -→ NA/H|NH
C-→NA/F|NA/A/F
D→0|1|2|3|4|5|6|7|8|9
F → day|hour|year|hr|min
I →.DA|.DNAA
A→mg|kg|ug|ml|cc|u|units|mcg|meq|hgb

A program that implements this grammar was designed in Perl. The program runs on the dataset, identifies the dosages expressions and replaces each with the word "dosage".

6.2 Drug Name Recognizer

Recognizing and extracting drug names from documents is an instance of named entities recognition. Named entities extraction, including organizations, people, and locations, along with date/time expressions and monetary and percentage expressions has been an active research topic. Several approaches have been proposed to capture these types of terms, corpus-based methods using machine learning or rule-based methods are used to extract the named entity of interest. In our work, the rule-based approach has been adopted.

We examined our dataset for semantic patterns that indicate the presence of a medication and discovered a number of patterns that are defined in the table below. These patterns form a lexical grammar that is used to "parse" the documents to detect and extract the drug name.

Table 1. Semantic Patterns for Drug Name Extraction

Pattern	Example
concentration of <medication> <medication> concentration	concentration of Norcuron Fentanyl concentration
<medication> drip	Ativan drip
ordered (dosage)<medication>/ <medication> ordered	ordered ethacrynic acid Bacitracin ordered
dosage of <medication>/ <medication> dosage	dosage of tacrolimus NPH insulin 15U
<medication> at dosage	Ativan at 1/2mg
infusion of <medication>	infusion of Vecuronium

The Drug Name recognizer has been implemented in Java. The patterns defined above are run and all occurrences of drug names are found and replaced with the word "medication".

After the original dataset is processed by the de-identification rules, we obtained two "anonymized" datasets: one with the dosage removed, one with the drug name replaced.

Table 2. Example of text modification

Original Text	**Dosage**	**Drug**
Vasopressin ordered at **.2 units** per hour. Pharmacy mixed drip of **100units/500cc** instead of standard concentration on **100units/100cc**	Vasopressin ordered at **dosage**. Pharmacy mixed drip of **dosage** instead of standard concentration on **dosage**	Medication ordered at **dosage.** Pharmacy mixed drip of **dosage** instead of standard concentration on **dosage**

7. EVALUATION

A natural consequence of applying privacy preservation methods, such as perturbation and anonymization, is the information loss. This loss of information may not only affect the data quality, but also may change the mining results, rendering them useless in extreme cases.

In order to assess the validity of our approach and the impact of the data transformation procedures on the mining process, we run our datasets using the K-means algorithm[11]. First, with the original dataset. Then we run it on the anonymized versions, dosage removed, then with drug name replaced. We then compared the clusters. Our intent is to determine the number of documents from the modified sample that stayed in their original cluster, we call this persistence. If the persistence is high, then the de-identification rules are not drastically compromising the mining process. Conversely, if the persistence is low, this means that the anonymization rules have caused records to cluster in different groups and changed the outcome of the original clustering process.

We used RapidMiner™ as a tool to perform the text mining tasks. RapidMiner™ is an environment for machine learning, data mining, text mining and predictive analytics. It is distributed under the

AGPL open source license and has been hosted by SourceForge since 2004 [12]. We selected the K-means clustering algorithm and set the number of clusters to 5. We run the algorithm on the three datasets and obtain the following results for the 5 clusters.

Table 3. Distribution of Records in Clusters across Datasets

	Original Text	-Dosage	-Medication
Cluster 0	537	528	547
Cluster 1	316	301	66
Cluster 2	1043	1074	1238
Cluster 3	66	66	83
Cluster 4	87	80	192

7.1 Persistence Definition:

Our notion of Persistence amounts to comparing clusters by measuring a distance between clusters. We define the distance between two clusters C1 and C2, by the number of changes that C2 has to undergo in order to be the same as C1. For this, we considered the Earth Mover Distance (EMS) [13]. EMS is the measure of distance between two probability distribution. Informally if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region *D*, the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved. The EMD is widely used in content-based image retrieval to compute distances between the color histograms of two digital images [14]

7.2 Results

We loosely borrowed the EMD approach and adapted it to our problem: we considered the clusters as the distributions and each record is a parcel of “dirt”, the region is the total set of features. The cost of turning cluster C2 into C1 is the sum of the number of record deletions from C2 and records additions to C2.

Table 4 computes the cost of turning clusters from the dosage deleted clusters into the original data clusters. Clusters contents are compared pair wise

Table 4. Cluster Distance Original-Dosage

-Dosage					
Original	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 0	*13.4*	96.27	90.5	100	99.6
Cluster 1	92.08	*24.36*	87.07	100	98.73
Cluster 2	96.35	97.41	*7.28*	99	99.23
Cluster 3	100	100	98.48	*1.51*	100
Cluster 4	93.1	95.4	86.20	100	*24.13*

Table 5. Cluster Distance Original/Medication

-Medication					
Original	Cluster 0	Cluster 1	*Cluster 2*	Cluster 3	Cluster 4
Cluster 0	73	97.2	*42.8*	95.5	91.4
Cluster 1	75.6	98	*41*	95.5	89.5
Cluster 2	75.5	97.5	*39.8*	96.3	90.8
Cluster 3	71.2	92.4	*45.5*	97	94
Cluster 4	70	94.2	*50.6*	95.4	89.6

Taking the minimal distance between the clusters, it appears that each cluster C_i in the original set has the strongest similarity with cluster C_i of the modified dataset. We normalize the results and define persistance as: 100-minimal distance we obtain the following results from Table 4

Table 6. Record Persistence(D)

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
86.6%	75.6%	92.7%	98.5%	75.8%

Overall Record Persistence: 86%

Also taking the minimal distance between clusters in Table 5, it surprisingly seems that cluster 2 from the modified dataset is closest to all clusters on the original set.

Table 7. Record Persistence(M)

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
57.17%	58.87%	60.22%	54.85%	49.43%

Overall Record Persistence: 56%

As can be seen, the Record Persistence in the -Dosage dataset is 86%, meaning that its clusters are not very different from the original dataset clusters. However the -Medication dataset clusters seem to have lost about ½ of their members, which indicate that the clustering results have degraded after the de-identification rules for the drug names have been applied.

8. CONCLUSION AND FUTURE WORK

In this work, we presented a de-identification technique for privacy preservation based on lexical entailment. This technique was used to anonymize data related to drug names and their dosages in ICU patients' records. The basic assumption is that if a word is entailed by another in a given context, then replacing the first word by the second, should maintain the semantic integrity of the text and not greatly undermine mining results. This technique was tested on a dataset and clustering results were compared. The approach showed some encouraging results in a restricted domain.

In the future, we plan to exploit the same principle to other datasets to evaluate its portability to other domains.

Additionally, we are interested in finding a more systematic method to recognize features in texts that need to be anonymized and more efficient learning strategies to apply our method to larger corpora.

9. REFERENCES

1. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. In: International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5) pp. 571--588, (2002)
2. Sweeney, L.: Guaranteeing anonymity when sharing medical data, the datafly system. In: Journal of the American Medical Informatics Association. Washington, DC, Hanley & Belfus, Inc. (1997).
3. Sweeney, L.: Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Journal of the American Medical Informatics Association. Washington, DC: Hanley & Belfus, Inc., pp. 333--337 (1996)
4. HIPAA Health Information Privacy, http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary_index.html
5. Aggarwal, C., and Yu, P.S., (eds): Privacy Preserving Data Mining: Models and Algorithms. Springer (2008)
6. Aggarwal, C.: A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In : Aggarwal, C., and Yu, P.S.,(eds) Privacy Preserving Data Mining: Models and Algorithms,. Springer pp. 11—52 (2002)
7. Glickman, O., Dagan, I., and Koppel, M.: A lexical alignment model for probabilistic textual entailment. In: J. Quinonero-Candela, I. , Dagan, B., Magnini, F. (eds). Machine Learning Challenges, 2006. LNCS Vol. 3944 Springer, pp. 287--298 (2006)
8. Giuliano, C., and Gliozzo, A.: Instance Based Lexical Entailment for Ontology Population. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 248--256.(2007)
9. Hovy, E., Hermjakob, U., Lin, C.: The Use of External Knowledge in Factoid QA. In : Proceedings of the Ninth Text Retrieval Conference (TREC-10). (2001).
10. Intensive Care Unit Safety Reporting System (ICUSRS**)**, http://cms.johnshopkinsinternational.com:8001/QSR/Research/ Projects/project_ICUSRS.asp
11. Feldman, R., Sanger, J. : The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data. University Press (2007)
12. RapidMiner™, http://rapid-i.com/content/view/181/190/
13. EMD, http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/RUBNER/emd.htm
14. Rubner, Y., Tomasi, C., and Guibas, L.J. : The Earth Mover's Distance as a Metric for Image Retrieval. In : International Journal of Computer Vision 40(2) pp. 99–121, Kluwer Academic Publishers.(2000)