Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Spherical microaggregation: Anonymizing sparse vector spaces

Daniel Abril [a,c,*], Guillermo Navarro-Arribas [b], Vicenç Torra [a,d]

[a] IIIA, Institut d'Investigació en Intel·ligència Artificial – CSIC, Consejo Superior de Investigaciones Científicas, Campus UAB s/n, 08193, Bellaterra, Spain
[b] DEIC, Dep. Enginyeria de la Informació i de les Comunicacions, UAB, Universitat Autònoma de Barcelona, Campus UAB s/n, 08191, Bellaterra, Spain
[c] UAB, Universitat Autónoma de Barcelona, Campus UAB s/n, 08193, Bellaterra, Spain
[d] School of Informatics, University of Skövde, 54128 Skövde, Sweden

ARTICLE INFO

Article history:
Received 22 February 2014
Received in revised form
5 October 2014
Accepted 18 November 2014
Available online 27 November 2014

Keywords:
Anonymization
Vector space
Privacy preserving
Data mining
Information loss
Sparse data

ABSTRACT

Unstructured texts are a very popular data type and still widely unexplored in the privacy preserving data mining field. We consider the problem of providing public information about a set of confidential documents. To that end we have developed a method to protect a Vector Space Model (VSM), to make it public even if the documents it represents are private. This method is inspired by microaggregation, a popular protection method from statistical disclosure control, and adapted to work with sparse and high dimensional data sets.

1. Introduction

In the last years, the quantity of stored personal information has been increasing exponentially and it has created a problem of privacy due to the inherent tension in mining those sensitive data bases. Since that, research fields such as privacy preserving data mining has got an important role in the community and a number of different methods have been proposed for privacy preserving data mining. Although, sanitization methods explicitly remove identifiers like names, phone numbers, addresses, etc., they are not enough to protect individual's privacy. Therefore, we should delete other additional information so an attacker cannot infer an identity or other sensitive information about an individual based on the remainder information.

In this paper we address the problem of how to release a set of confidential documents without giving away sensitive information that can be linked to specific individuals and also achieved with the minimum loss of information. Given a set of confidential, and thus, private documents we want to provide some public metadata to be used for analysis and mining,

* *Corresponding author*. IIIA, Institut d'Investigació en Intel·ligència Artificial – CSIC, Consejo Superior de Investigaciones Científicas. Campus UAB s/n, 08193, Bellaterra, Spain.
E-mail addresses: dabril@iiia.csic.es (D. Abril), guillermo.navarro@uab.cat (G. Navarro-Arribas), vtorra@his.se (V. Torra).
http://dx.doi.org/10.1016/j.cose.2014.11.005

which preserves the privacy or the anonymity with respect to the original documents. To address the problem we have relied on a well known data representation of a set of documents, the Vector Space Model (VSM) (Salton, 1989), which is widely used in information retrieval and text mining. Following these ideas our proposal can be summarized as providing a secure VSM, which closely represents a set of documents while preserving the anonymity of the documents. The protected VSM can be made public, while the original documents are kept secret. This allows to perform some information retrieval and text mining tasks on the set of documents while preserving the privacy of the documents.

It is important to clarify what do we consider as private information with respect to a set of documents. In this work we have focused to the concrete protection of the document owner, creator, or the entity to which the document is explicitly related. We try to prevent the ability of an attacker to correctly link a given document (or document representation) to a concrete entity (individual, organization, ...). That is, depending on the protection purpose, we want to avoid the possible link an attacker can establish between a document and its writer, or on the other hand, we also want to avoid the possible link an attacker can establish between a document and the entity which it is about. We will discuss several possible scenarios that present this particularity or threat, some clear examples are a set of health patient records, research project proposals, individual posts to a private internet forum, ...

To achieve the protection of the VSM we rely in the k-anonymity property (Samarati, 2001; Sweeney, 2002). Our proposal provides a protection method that yields a k-anonymous VSM. As we will see, this will ensure that at least k vectors (document representations) in the VSM are equal to each other providing an upper bound on the probability to link an entity to the specific document.

1.1. Motivation

To better shape our proposal we present here three motivating scenarios. In short, the presented anonymization technique is suitable for scenarios involving a set of confidential documents in which each document is directly or indirectly related to one or a set of different entities. A direct relation is when the document contains sensitive information of the specific person or institution that must be anonymized, while an indirect relation is when the document does not contain explicit information about the entity to be anonymized, but also there is an implicit relation between the entity and the document that can be inferred through some other document properties. We describe three cases where our proposal has a direct application.

1.1.1. *Private textual datasets for generic research*

A clear application scenario is within the research community, in the information retrieval and text mining fields. Several organizations present their research at scientific journals or conferences. Usually, the experiments presented showing their improvements and validating their research are done over some private datasets. However, when other researchers want to reproduce those experiments it becomes impossible, since the datasets are private and can contain confidential information. Examples are a set of patient health records, user posts to a private Internet forum, a set of user profiles from a social network, or even a set of user queries made to a search engine (recall the infamous AOL search data leak Barbaro et al., 2006).

A possible solution is to publish an anonymized data that represents the original dataset and can be used to reproduce to some extent the research made on the original dataset. This is straightforward in text mining research where the VSM is frequently used to represent a set of documents, but other similar data structures can be envisioned with the same purpose for more specific tasks.

1.1.2. *Private profiling data for advertising*

Personalized online advertisement is another possible area where anonymization should be considered. Lots of web services are offering their services for free in exchange of introducing advertisements on their services. Google, Twitter or Facebook are some examples of companies, which collect and store thousands of users' confidential information in order to analyze and offer targeted advertisements (Rutkin and October 1, 2013; Simonite and November 8, 2013). E-mails, user's posts or even personal documents are some clear examples. In some cases these data could be transferred to specialized companies, which analyze all data in order to define advertisement strategies in a user base.

These user data might be considered confidential, and might not be directly transferable to other parties. Therefore, the solution is to anonymize the data before its transference. The idea behind this approach is that the advertisement company will not be able to distinguish a unique user from a set of k of them. Hence, the advertisements selected for a single user are actually extracted from a mix of several user's profiles.

1.1.3. *Anonymized metadata from public tender*

As a last example, we can consider for instance a government agency managing applications to public research project funding. Such applications should be kept private, but at the same time it can be interesting to be able to give some information about the applications and more precisely of the projects presented by the applicants. This becomes specially difficult if we assume that the projects are written in a free-form text. This information is interesting not only to the community applying for funding but also to the administration and politicians. We are looking for information such as: "this geographic area applies for projects about this topic", or "this methodology is proposed by a given percentage of researchers from these given topics". While this information can be valuable it normally does not reveal specific and private information.

1.2. Contributions and plan of the paper

The major contributions of this paper are as follows. We provide a new approach for the anonymization of very sparse and high-dimensional data sets. We continue and improve the work presented in Abril et al. (2013), in which the authors presented a first microaggregation approach for the