Towards a private vector space model for confidential documents

Daniel Abril
Institut d'Investigació en Intel·ligència Artificial (IIIA), Consejo Superior de Investigaciones Científicas (CSIC)
dabril@iiia.csic.es

Guillermo Navarro-Arribas
Dep. Enginyeria de la Informació i de les Comunicacions (DEIC), Universitat Autònoma de Barcelona (UAB)
gnavarro@deic.uab.cat

Vicenç Torra
Institut d'Investigació en Intel·ligència Artificial (IIIA), Consejo Superior de Investigaciones Científicas (CSIC)
vtorra@iiia.csic.es

ABSTRACT

We introduce in this paper a method to anonymize document vector spaces. These vector spaces can be used to analyze confidential documents without disclosing private information. The method is inspired in microaggregation, a popular technique used in statistical disclosure control.

Keywords

Privacy, document vector space, indexes, anonymization

1. INTRODUCTION

Management of confidential documents has become an important issue in governments, administrations, public organizations and private corporations owing to the debate between the freedom and the withhold information. That is why many researchers are focusing their efforts in this field. We have focused on the anonymization of indexes which have been built from a set of confidential documents. For instance, consider 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 research 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

The previous example can be extended to several application areas where an organization holds a set of private documents, but wants to give information about them to third parties. In order to do it properly the organization has to ensure that the third parties will not be able to obtain specific information about the documents normally considered private. In the previous example, such information will be detailed information about the project, and most important who proposes the project (researcher, university, ...).

We propose in this paper to rely in the vector space or term vector model [3] normally used in information retrieval systems to provide such information. In a vector space model, it is common to represent a document as a vector of terms with an associated frequency-based weight. Furthermore, in this paper we introduce the anonymization of a vector space for a set of documents to ensure a given degree of privacy. The protected vector space can be used to build indexes for querying the set of documents, to carry classification or categorization tasks, latent semantic analysis, and so on, preserving some degree of privacy.

2. DOCUMENT REPRESENTATION

In order to represent each document as a vector of terms, the documents are read and tokenized. However, not all the words included in a document are useful when using text classification or information retrieval techniques. These useless words, called stop-words, are removed. Moreover, we consider two additional steps in the cleaning process, the first one consists on removing all the words with two or less letters and the latter removes all the words which are not in the WordNet ontology [4]. Note that by considering only the words included in WordNet we are eliminating some words, which can result in a loss of information, because they might be relevant for the document analyzer. At the same time, they can lead to disclosure. Usually, these words are proper names or very specific terms of a particular research field.

Once the documents are cleaned we apply the Porter stemming algorithm [6], which is another common preprocessing step in NLP. Words with the same stem are considered the same word, which also reduces the size of the feature set.

Finally, we also consider a feature selection step so as to decrease the size of the vocabulary and also to avoid noise features. A common feature selection method is to compute term frequencies, but other methods could be used, such as term frequency - inverse document frequency, information gain, mutual information, etc.

The set of all document vectors can be seen as a document-term matrix, where the rows represent each document and

SAC'12 March 26-30, 2012, Riva del Garda, Italy.

the columns are the corresponding term weight, expressed by the feature selection method, in our case we have used the normalized term frequency.

3. ANONYMIZATION OF A VECTOR SPACE

The anonymization process is based on microaggregation [1], which is a popular method used in Statistical Disclosure Control (SDC) [9]. Microaggregation's privacy is ensured by the satisfaction of the k-anonymity principle [7, 8]. Broadly speaking, microaggregation consists of clustering the data into small clusters and then replacing the original data by the centroids of the corresponding clusters.

Privacy is achieved ensuring that all clusters have at least a predefined number of elements, say k to the number of values, and, therefore, there are at least k records with the same value. Note that all the records in the cluster replace a value by the value in the centroid of the cluster. The constant k is a parameter of the method that controls the level of privacy. The larger the k, the more privacy we have in the protected data.

From the operational point of view, microaggregation is defined in terms of partition and aggregation:

- **Partition.** Records are partitioned into several clusters, each of them consisting of at least k records.
- **Aggregation.** For each of the clusters a representative (the centroid) is computed, and then original records are replaced by the representative of the cluster to which they belong to.

It is known that the solution of this problem is NP-HARD [5], that's why heuristic methods have been developed. On example is MDAV [2] (Maximum Distance to Average Vector), which is the one we have used in this work.

3.1 Microaggregation of vector spaces

In order to microaggregate a vector space, we consider vectors as records or attribute values. We need thus, to define the partition and aggregation steps for the vectors.

The partition step is determined by the distance function between vectors, in our case we have considered the cosine distance, which is a common distance used in vector spaces and frequently used in the text mining area in order to compare the dissimilarity of two documents. The cosine distance between two vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$ is defined as:

$$d_{cos}(\vec{V}(d_1), \vec{V}(d_2)) = 1 - \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}$$

where $\cdot$ is the dot product of the vectors.

Finally we need to define the aggregation step, which is determined by an aggregation operator $\mathbb{C}$. In our case we use a component-wise mean to aggregate vectors. That is, for a set of n vectors $\vec{V}(d_1), \ldots, \vec{V}(d_n)$ in a M-dimensional space,

$$\mathbb{C}(\vec{V}(d_1), \ldots, \vec{V}(d_n)) = (\sum_{i=1}^{n} w_{i,1}, \ldots, \sum_{i=1}^{n} w_{i,M}) \quad (1)$$

where, $w_{i,j}$ is the weight associated to term j in document i.

We use these distance and aggregation functions within the MDAV algorithm [2] in order to microaggregate a vector space.

4. CONCLUSIONS

We have introduced in this paper, the anonymization of document vector spaces. The motivation is to provide an anonymized vector space, which could be used to analyze a set of confidential documents. For this purpose we have developed an anonymization method based on microaggregation, which is a popular technique in SDC and PPDM.

A common use of the vector space is to use it as an index to be queried. In order to measure the loss of information we compared the results of querying the original vector space with the results obtained in the protected versions. Using 5 different queries we have obtained relatively good results for lower values of k when comparing the list of documents returned by querying the original index and the protected one.

As future work we plan to further develop the idea introduced in this paper with more accurate anonymization techniques. For instance we are studying different distances and aggregation operators to improve the microaggregation method we have presented, as well as other anonymization techniques.

Acknowledgments

This work is partially funded by projects TSI2007-65406-C03-02, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, TIN2010-15764 and TIN2011-27076-C03-03 of the Spanish Government, and by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement 262608. The work contributed by the first author was carried out as part of the Computer Science Ph.D. program of the Universitat Autònoma de Barcelona (UAB).

5. REFERENCES

[1] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204, 1993.

[2] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowl. and Data Eng.*, 14:189–201, January 2002.

[3] C.D. Manning, P. Raghavan, H. Schütze (2009) *An Introduction to Information Retrieval*. Cambridge University Press.

[4] Miller, G., 2010. WordNet - About Us, *WordNet*, Princeton University. http://wordnet.princeton.edu.

[5] A. Oganian and J. Domingo-ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Commission for Europe*, 18:345–354, 2001.

[6] M.F. Porter. *An algorithm for suffix stripping.* Program Vol. 14, no. 3, pp 130–137, 1980.

[7] P. Samarati. Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, 13:1010–1027, 2001.

[8] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, October 2002.

[9] d. W. T. Willenborg L. *Elements of Statistical Disclosure Control*. Springer Verlag, New York, 2001.