ARES project
CONSOLIDER-INGENIO 2010 CSD2007-00004
Workpackage 4 - Task 2 (WP4.T2)
New privacy-preserving data mining techniques
Deliverable Report

G. Navarro-Arribas, V. Torra

Institut d'Investigació en Intel·ligència Artificial,
Consejo Superior de Investigaciones científicas,
Campus UAB,E-08193 Bellaterra, Catalonia (SPAIN
e-mail {guille,vtorra}@iiia.csic.es

November 12, 2010

Contents

Abstract

We present some new techniques for privacy preserving data mining (PPDM). On the one hand, we introduce the use of semantic or general ontology tools to protect non-numerical data and provide PPDM in the context of confidential documents. On the other hand, we analyze the use of previously developed protection techniques for web access and web query logs in the context of PPDM with some encouraging results.

Chapter 1

Introduction

The term Privacy Preserving Data Mining (PPDM) [3] denotes techniques which for example enable e-commerce companies having collected transaction data to conduct joint exploitation with other companies without revealing the individual data they own. In general, individuals can benefit from PPDM when, for example, their data is collected by any Web site and then used for profiling [7].

Although PPDM has been a relatively active research field, it still needs a lot of improvements and advances to become widespread in e-commerce and broadly used by companies or administrations.

In this report, we introduce two different research lines for PPDM resulting from WP4.T4:

1. An attempt to open novel research lines in the field of *semantic* PPDM. PPDM has been mainly applied to statistical data, mostly numeric. Little attention has been paid to categorical or general non-numeric data. In this report we introduce the protection of textual data based on its semantics to provide semantic based PPDM. This is in fact a relatively new research field, where few works are available.

2. Extending the use of PPDM techniques to Web usage mining and query logs mining. This work complements the outcome of WP2 on secure e-commerce. There, statistical disclosure techniques (SDC) were introduced to protect both access and query logs [28]. We show that these techniques are specially suited for PPDM.

These two research lines are presented in Sections 2 and 3 respectively. Finally, Section 4 concludes the report and provides a list of contributions. We have also included in Appendix A a brief description of microaggregation, an statistical disclosure technique which is used to some extend in all the proposals.

Chapter 2

Towards semantic PPDM

We have introduced relatively novel approaches to implement PPDM techniques for non-numeric data relaying in the semantics of the data in what can be seen as an ontology or semantic based PPDM. Although some of the techniques could be seen as generic protection methods for non-numeric data in PPDM, we have focused our work to a concrete case, which is the analysis of generic documents. The techniques and results presented here were primarily focused towards information retrieval, but they can be easily accommodated for other data mining approaches. This work is presented in the following sections.

2.1 Problem description

It is not uncommon to find situations where it is needed to provide information regarding the contents of a set of confidential documents. Documents such as research project proposals, research papers submitted for publication, confidential law suites, medical records, confidential reports (by companies of law enforcement agencies), etc. cannot be publicly revealed, but being able to provide some information about them might be very useful. The information disclosed should provide enough accuracy to allow generic tasks such as the classification of documents into categories of topics, but preserving their confidentiality.

A simple approach in these situations is to provide, for each document, a vector of keywords or terms. These terms can be manually specified, or most commonly, automatically generated. For instance, by providing the N most frequent terms for each document. The terms, are then used by classification, clustering or generic information retrieval algorithms. The problem with this approach is to ensure that the vector of terms from the document does not reveal confidential information. A relatively large vector provides more accuracy but at the same time might reveal too much information regarding the contents of the document.

To deal with this problem, we provide a vector of terms for each confidential document, which ensures that a certain degree of confidentiality is preserved. The degree of confidentiality or privacy is measured in terms of k-anonymity with respect to the whole set of documents. That is, in the resulting set of document vectors, there will be k indistinguishable vectors. To do so, we rely on semantic generalizations through the use of a semantic microaggregation

approach for categorical data, which makes use of WordNet [26].

2.2 Document vectors

We have a set of m confidential documents D, and each document is represented by a document vector, which contains the most relevant terms of the document. The relevance of the selected terms is determined by their frequency. The documents are automatically parsed and tokenized following [21], then, we eliminate common English stop-words, words with less than three letters and words which are not in WordNet. The resulting set of terms are used to calculate the document vectors.

By considering only the words included in WordNet we are eliminating some words that can result in a loss of information. These words are normally common names or very specific terms used in specific research fields. It is important to remark that in this work we are using WordNet as a generic ontology for the English language. When the application domain is known, other domain-specific ontologies can be used, such as the UMLS (Unified Medical Language System) [40] for biomedical data.

Each document is represented by a vector d, which contains the N most frequent terms t with their associated frequency weight w. The weight $\omega_{i,j}$ is computed as:

$$\omega_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.1}$$

where $n_{i,j}$ is the number of occurrences of the term t_i in document d_j, and $\sum_k n_{k,j}$ is the number of occurrences of all the terms in document d_j.

The document vector d_j for the jth document is sorted by frequencies. Formally,

$$d_j = ((t_{\sigma(1),j}, \omega_{\sigma(1),j}), (t_{\sigma(2),j}, \omega_{\sigma(2),j}), \ldots, (t_{\sigma(N),j}, \omega_{\sigma(N),j}) \tag{2.2}$$

where σ is a permutation such that $\omega_{\sigma(i),j} \geq \omega_{\sigma(i+1),j}$ for all $i = 1, \ldots, N-1$.

The vector of relevant terms based on their frequency provides a good approximation to the contents of the document. Moreover, it can be easily used for classification of documents in categories or topics by automated algorithms.

2.3 Microaggregation of term vectors

In order to provide a semantic microaggregation of the document vectors, we use a semantic approach both for the partition of the data into clusters, and for the aggregation of vectors to compute the centroids of the clusters. The partition is based on a semantic distance on vectors, which relies in the Wu-Palmer similarity [42] to compute distances between terms using WordNet. On the other hand, the aggregation is computed by generalizing terms in a WordNet taxonomy.

WordNet structures nouns, verbs, adjectives, and verbs, into sets of cognitive synonyms called *synsets* which express concrete concepts. These synsets are interlinked by several conceptual-semantic and lexical relations.

The distance between vectors is computed by considering the possible combination of synsets from the terms in each vector. For a detailed description of the distance and aggregation functions used, please refer to [2, 1].

In order to understand better the semantic microaggregation process, we give a toy example using an original small dataset integrated by four documents as input of the process.

Table 2.1 (top) shows firstly the original file, integrated by four documents with three terms and their respective term frequency for each of them. The table also shows the protected output file obtained after the microaggregation process with $k = 2$. As you can see the output file has four documents as the original file, but it only has two different records. The first document centroid represents the set of documents that discuss about computer parts, and the second one joins the two original documents that talk about different animals. Therefore, we can say that with the protected file we can deduce the general topics of the documents, but we cannot know the specific topics of the original dataset.

Original Data File
(('keyboard', 0.3), ('laptop', 0.4), ('software', 0.3))
(('horse', 0.7), ('dog', 0.2), ('cat', 0.1))
(('hardware', 0.3), ('screen', 0.3), ('computer', 0.4)))
(('lion', 0.5), ('monkey', 0.3), ('tiger', 0.2))
Protected Data File
(('abstraction', 0.3), ('computer', 0.4), ('instrumentality', 0.3))
(('big_cat', 0.3), ('carnivore', 0.2), ('placental', 0.5))
(('abstraction', 0.3), ('computer', 0.4), ('instrumentality', 0.3))
(('big_cat', 0.3), ('carnivore', 0.2), ('placental', 0.5))

Table 2.1: Example of semantic microaggregation. Original and its respective protected dataset

2.4 Evaluation

In order to evaluate the semantic microaggregation we have used 50 published papers in the last three years in the *Modeling Decisions for Artificial Intelligence*(MDAI) conference. We have created two different data sets from these 50 documents. One with a set of document vectors with the 50 more frequent terms, and another with the 100 most frequent terms. To simplify, we call them respectively *f50x50* and *f100x50*.

We only consider the words included in WordNet, which result in some minor loss of information. In this concrete case, we lose some common names (for example from the bibliography of each paper), and some very specific terms. More precisely, if we consider the set of words from *f50x50 f100x50* with words included in WordNet and without them, the average similarity measured by the Jaccard similarity between both sets is 0.769557, and 0.771153 respectively[1]. This work is just an illustrative experiment that could be improved by considering domain-specific ontologies.

Both files were protected with different values of the parameter k in the range from 2 to 10, and then, they were compared with different evaluation

[1]The Jaccard similarity coefficient measures the similarity between two sets A and B as $\frac{|A \cap B|}{|A \cup B|}$

measures. We have not computed values of k greater than 10 due to the limited size or the test dataset, and to the fact that, as we will see with $k = 10$, we already have a high degree of information loss.

Information loss is measured in terms of common error estimation measures from data clustering. The first measure, SSE, is the sum of squares to measure homogeneity in clustering. The lower SSE, the higher the within-group homogeneity. The SSA measure evaluates homogeneity between-groups. The higher SSA, the lower the between-groups homogeneity. Then, the SST measure is the total sum of squares ($SST = SSA + SSE$).

The final measure is the normalized information loss defined as

$$L = \frac{SSE}{SST} \times 100 \tag{2.3}$$

The optimal k-partition is defined by the one that minimizes the SSE measure (i.e., maximizes the within-group homogeneity) and maximizes the SSA measure (i.e, minimizes the between-group homogeneity). Note that the higher within-group homogeneity, the lower the information loss.

Table 2.2 shows the evaluation values defining how optimal is the k-partition for each one of these protected files. As expected, the SSE values increases as k increases. It means that within-group homogeneity decreases when the number of documents per cluster increases.

k	**Data Set**	**SSE**	**SSA**	**SST**	**L**
2	f50x50	4.938	30.929	35.867	13.766
	f100x50	4.936	37.119	42.055	11.736
3	f50x50	11.407	21.390	32.797	34.780
	f100x50	12.049	29.733	41.782	28.838
4	f50x50	15.693	21.556	37.249	42.131
	f100x50	16.647	22.759	39.406	42.245
5	f50x50	20.404	11.890	32.294	63.181
	f100x50	21.070	19.157	40.227	52.377
6	f50x50	23.072	17.372	40.444	57.046
	f100x50	24.516	18.336	42.852	57.212
7	f50x50	25.109	11.332	36.441	68.903
	f100x50	26.712	18.981	45.693	58.560
8	f50x50	27.034	8.986	36.0194	75.053
	f100x50	27.662	16.101	43.763	63.209
9	f50x50	28.529	10.085	38.614	73.883
	f100x50	30.107	11.657	41.764	72.088
10	f50x50	31.670	5.680	37.350	84.793
	f100x50	31.455	10.857	42.312	74.341

Table 2.2: Evaluation values of both data sets according to k

On the contrary, SSA values decrease when k decreases. This is reasonable because when k grows, there are less centroids and homogeneity between clusters decreases.

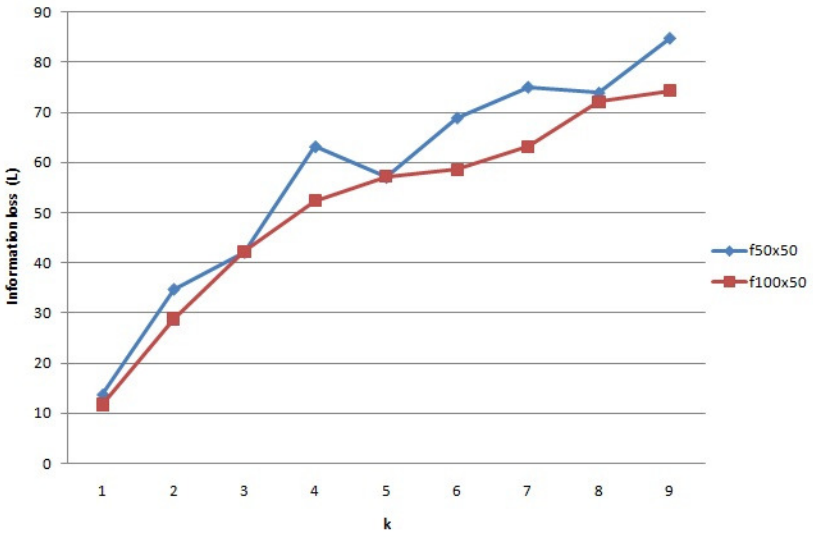


Figure 2.1: Plot of information loss (L) vs. privacy level k.

Finally, we focus on the information loss. As expected, when k increases, the information loss also increases. Moreover, we can appreciate that the dataset with 50 terms, f50x50, results into a higher information loss than the dataset with 100 terms. You can clearly see it in Figure 2.1.

After the analysis, we can say that the best parameter is a k with values between 3 and 5, because they are the ones with a lower information loss value. At this point, we do not consider 2 as an acceptable value for k because, in this case, the protection level is too weak to ensure data confidentiality.

Chapter 3

PPDM for Web usage and query logs

In this cahpter, we present two examples of PPDM for web related logs. The first one (cf. Section 3.1) discusses web mining, which uses common web access logs as primary data. We show that PPDM can be performed to some extent in these logs if they are properly protected. Then, a similar approach is presented in the case of query logs, that is, logs generated by a web search engine (cf. Section 3.2).

3.1 PPDM for Web usage mining

Profiling the use of web site by its customers allows the developers of the service to redesign the service or e-commerce site according to the real needs of the users and even provide personalized browsing experiences for each one. Usage profiling in such scenarios is normally done through Web mining [13], by gathering information from the usage of the Web site (or service) to later on analyze it and come up with usage patterns. More precisely, these specific data mining approaches to determine the usage of a web site are normally denoted as Web usage mining, and complemented with what is known as Web structure mining and Web content mining[15, 22].

Web usage mining normally comprises two main techniques: statistical analysis, and more advanced data mining algorithms such as association rules, sequential patterns, and clustering [16]. While the first approach provides common and consolidated statistical estimations of the usage, the second approach attempts to identify usage patterns. In any case, both approaches rely on huge amounts of data gathered mainly from three sources: web servers, proxy servers and web clients [15], although other sources, such as specific application servers or even application data, can also be used [36]. The need to maintain such data in order to later on use them or even share them with other companies by outsourcing the usage profiling makes privacy issues very important.

The data used in Web usage mining contains sensitive information regarding the individual users, such as IP addresses. Moreover, Web usage mining and user profiling have always aroused privacy concerns among the users, governments, and civil rights organizations [35, 11, 14, 41]. Even so, there are not much

proposals to effectively provide a degree of privacy in Web usage mining data.

3.1.1 Session preservation in data masking

A fundamental step previous to web usage mining is data preprocessing [8], which does an initial analysis of the data gathered from the available sources. One of the main concerns in the data preprocessing and Web usage mining is to be able to identify *sessions*. A session is the set of pages that a single user browses in one visit, normally with the associated time spent in each page.

Logs are normally pre-processed where data is cleaned and formated, and sessions are identified. A session is usually identified as all the log entries having the same IP address and being compressed in a given time interval. In some cases, different users can have the same IP address if they are behind a proxy or doing NAT (Network Address Translation). To overcome such cases, the other information from the log is also used. For example, if two entries share the same IP address but have different user-agent, they are considered from different users, thus different sessions. If the structure of the web page is available, it can also be used considering that within a session there has to be a path from the first page to the last one following the web site structure.

We proposed in [31] a protection method based on microaggregation for Web access logs (see also [28]). This method was intended to provide PPDM, and to that end, it was specially designed to preserve sessions in protected data. In [30] we fine-tune the method and show that this is actually achieved with relatively high success. The results are summarized in the following section.

3.1.2 Evaluation of PPDM in access logs

We have evaluated the protection of the access logs generated by two different sites:

- **Site A** (`http://hacks-galore.org`): small web site consisting of three personal web pages with approximately 10 pages, several downloadable content (mainly PDF files), and a web blog which serves RSS. It is a site with a relatively low volume traffic (approximate average of 14 clicks per hour). An important issue of Site A is that the navigation patterns of the users are quite uniform.

- **Site B** (`http://www.iiia.csic.es`): a site based on Drupal with more than 6000 pages and content most of it dynamically generated, and with a relatively high volume of traffic (approximate average of 50 clicks per hour). In this case the navigation patterns of the users are very irregular.

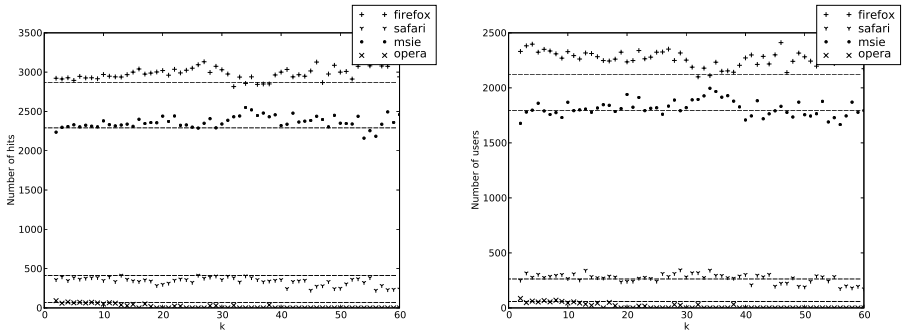
To simplify the results we have evaluated the three different indicators:

1. Number and type of browser per clicks.
2. Number and type of browser per session.
3. Percentage of usage of a given browser per session.

In each case, we attempt to identify sessions both in the original data and the protected data and then compare the results. Sessions are identified with the following procedure:

1. All records with the same IP and user agent (browser, browser version, OS, OS version) are grouped in the same sessions.

2. A timeout of 30 minutes is used to separate sessions. That is, after 30 minutes of inactivity, we consider that a new session starts.



(a) Browser usage *per clicks* (b) Browser usage *per session*

Figure 3.1: Browser usage for *Site A*. Straight lines denote the values for the original file.

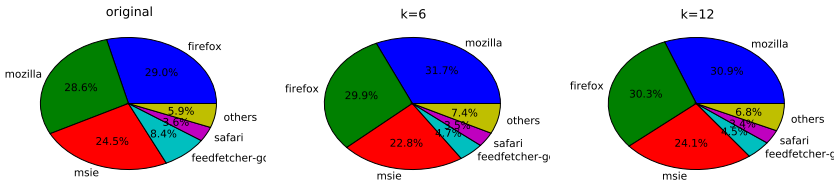


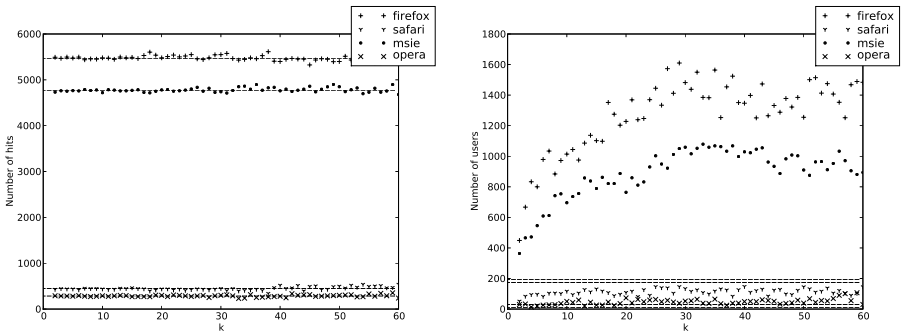
Figure 3.2: Browser usage percentage for $k = 1, 6, 12$ for *Site A*.

For *site A*, Figure 3.1 shows the browser usage per clicks and sessions, and Figure 3.2 the browser percentage for sessions for different protection levels (values of k). Analogously, for *site B*, Figure 3.3 shows the browser usage per clicks and sessions, and Figure 3.4 the browser percentage usage.

Whereas we achieve very good results with *site A* in all tests, the number of browsers per session in *site B* produces quite bad results. This is due to the fact that session identification is altered. More precisely, since the clusters tend to gather records from different sessions based on the request and referrer, and we are using a relatively big dataset (with 15000 records), fields such as the date suffer a relatively high distortion. This makes the identification of sessions by timeout to fail producing more sessions in the microaggregated files than in the original file.

Although the number of sessions in *site B* is altered, it is important to note that the usage percentage is maintained, with relatively low standard deviation and regression degree (cf. Table 3.1).

In order to solve the problem, we can make use of pre-computed session variables (added to the original data and then protected). Figure 3.5 shows the



(a) Browser usage *per clicks*.

(b) Browser usage *per session*.

Figure 3.3: Browser usage *Site B*. Straight lines denote the values for the original file.

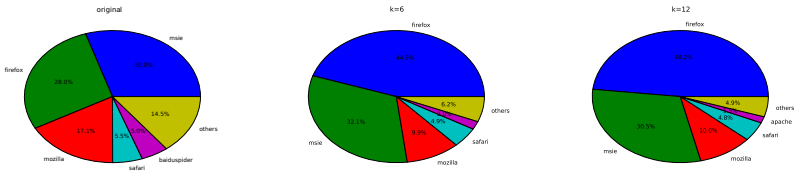


Figure 3.4: Browser usage percentage for $k = 1, 6, 12$ for *Site B*.

browser usage per session making use of a *sessionid* variable to identify sessions. We can see that for values $k < 30$ results are quite good (note that a $k > 20$ is hardly ever used).

3.2 PPDM for Query logs

The search logs generated by a web search engine (from now on WSE) are also very commonly used for data mining. They are a great source of information for researchers or marketing companies, but at the same time their publication may expose the privacy of the users from which the logs were generated. There is at least one well known case of released search logs with poor anonymization, which have been shown to reveal enough information to re-identify some users. The release was done by AOL in an attempt to help the information retrieval research community and ended up with not only important damage to AOL users privacy, but also a major damage to AOL itself with several class actions

	firefox	msie	safari	mozilla	others
σ	2.779324992	3.185317407	1.557858919	3.139720065	3.275318009
b	-0.034676437	0.006376312	0.036435533	0.085361292	-0.093496699

Table 3.1: Browser usage per session: standard deviation (σ), and regression coefficient (b) for *Site B* (all values $\times 100$).

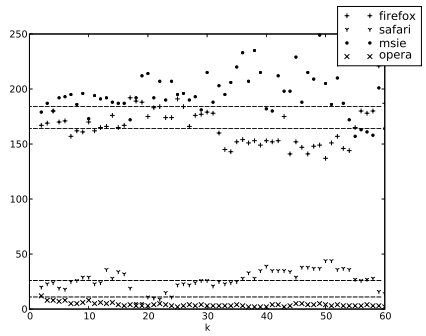


Figure 3.5: Browser usage per session in *Site B* using the *sessionid*. Straight lines denote the values for the original file.

suits and complaints against the company [10, 27]

Ideally, the search logs should be properly anonymized before they become public. The problem is that achieving a desirable degree of privacy in search logs is not easy, and presents an important trade-off between privacy and the usefulness of the data for mining purposes. There are several approaches [9] to anonymize such data, but they are normally reduced to the deletion of specific queries or logs. Moreover, common techniques used in statistical disclosure control (SDC) have not been applied to this specific problem until very recently [17].

Within this context, we proposed a specially aimed method to provide a reliable protection for PPDM. More precisely, the protection method is based on microaggregation and it is fully described in [32]. It was initially proposed in [29] and developed as part of the WP3.T3 [28].

Without getting into details of the protection method, we will show some results of the proposals in the next section.

3.2.1 Evaluation of PPDM in query logs

We have analyzed the frequency of queries in the protected data. Figure 3.6 shows the frequency of the ten most popular queries in the original data, and their evolution in the protected data as the microaggregation k increases. Note that $k = 1$ in the figure corresponds to the original data.

Although there are variations in the frequency, they are very low. Most relevant is that in all the protected data the same 10 queries are the 10 most popular, with some minor exceptions.

If we take a look to the frequency of single words, excluding common stop words, a similar result is obtained. As shown in Figure 3.7, frequency of words is relatively preserved. There are some concrete cases where a given word disappears with big values of k, but the overall result is satisfactory.

Normally, this kind of studies are performed with much bigger datasets (note that the most frequent word appears less than a 2.5%). The fact that our method preserves the frequencies with relatively smaller sets points to an even better preservation in bigger datasets.

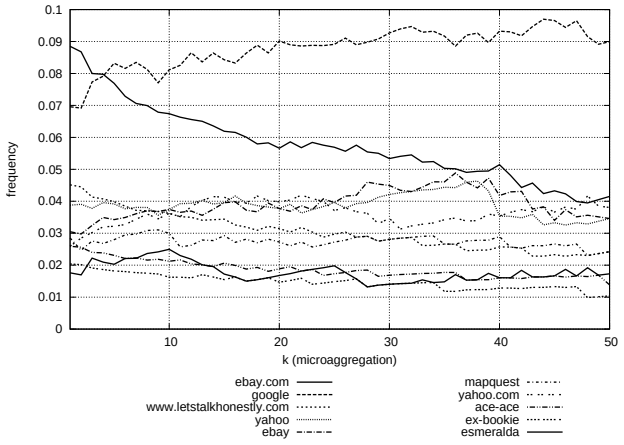


Figure 3.6: Query frequency analysis.

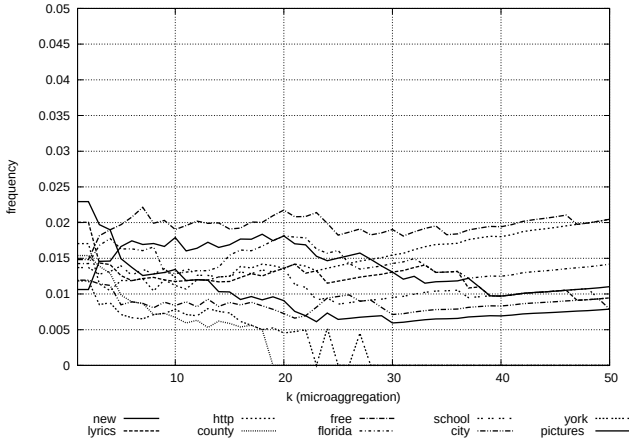


Figure 3.7: Word frequency.

Evaluation for clustering based-PPDM

Query logs are normally used in data mining processes for their analysis. To evaluate the utility of our protection method in data mining, we have considered clustering as a generic data minig process. There are several data mining techniques, from which clustering is one of the most popular [5, 6, 4].

We have compared the clustering of protected data with the clustering of the original data. We have used the k-means algorithm to cluster user query logs relaying in the distance function described in [29].

To compare the clusters obtained in the original data and the protected data, we have used two different well known indexes: the Jaccard, and Rand indexes.

Figure 3.8 shows the Jaccard and Rand indexes comparing the original data with the data protected with $k \in \{3, 5, 10, 20, 30, 50\}$, using the k-means algorithm with $\kappa = 2 \ldots 500$. We use κ to denote the k parameter of the k-means algorithm, which corresponds to the number of clusters. As the number of clus-

ters increases, the indexes are closer to 1, meaning that both partitions are very similar. Regardless of the k used in the microaggregation process, we obtain similar results for both indexes.

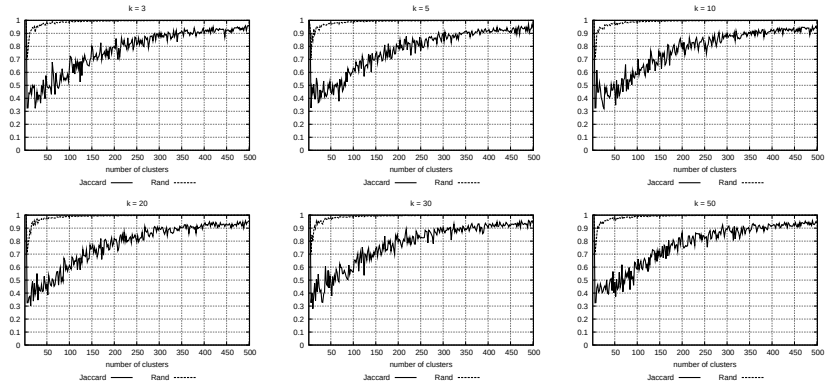


Figure 3.8: Rand and Jaccard indexes for aggregated data with $k \in \{3, 5, 10, 20, 30, 50\}$.

More interesting is to notice the differences between the indexes as the microaggregation k increases. As this k increments, we achieve more privacy but less utility. Given a protected dataset with $k = 3$ and another with $k = 50$, Figure 3.9 shows the difference between the clusters of the two datasets, as the k-means κ increases. The straigh line denotes de difference between the Jaccard index of both datasets and the doted line the difference between the Rand index.

Although the values for the indexes in Figure 3.8 seem similar, it can be seen in Figure 3.9 that the differences between the same index but for the two different values of k (3 and 50) decreases as the k-means κ increases. This means that if we are going to cluster data in the data mining process with relatively big κ, we can use a relatively big k in the microaggregation. That is, we can provide high privacy protection to the data while preserving the partitions of the clustering.

Note that we are clustering users (or user profiles) using the same user distance used in the microaggregation process. This anticipates that the relatively good results were expected from how the protected log is generated. We think that this is the main reason that makes microaggregation a good protection technique for privacy-preserving data mining when a distance-based clustering is used in the data mining process. The example described here is just an exemplification of this statement. Some particular data mining applications where our protection method will provide good results are in fact those making use of some clustering techniques: categorization, classification, ...

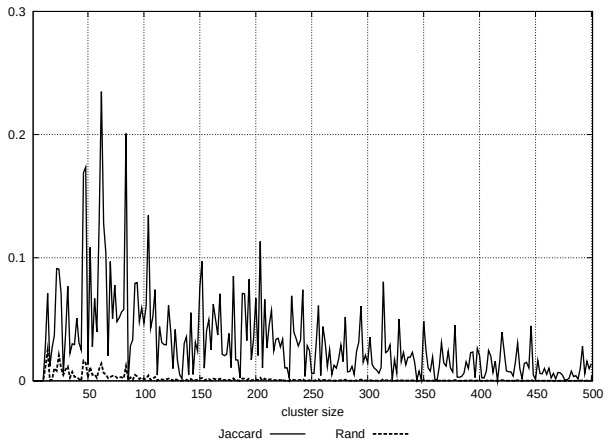


Figure 3.9: Difference of the Jaccard and Rand indexes for data microaggregated with $k = 3$ and $k = 50$.

Chapter 4

Conclusions and contributions

4.1 Conclusions and future research lines

In this report, we have presented some new PPDM techniques. The first ones introduced in Section 2 work towards providing what can be seen as a semantic or ontology-based PPDM. That is, we propose to protect non-numeric data taking into account their semantics to then be used in data mining processes.

We consider the situation where data mining or information retrieval task has to be performed over a set of confidential documents. The documents are summarized with a vector of frequent terms which are then protected so taht they can be safely used for data mining. In Section 2.1, we consider the protection of PDF documents, but the technique could be applied to any set of sensitive documents as in [1], where we apply a similar approach for web pages. Although our proposal does provide good results, there is still a lot of work to be done. The evaluation of specific data mining techniques, the development of more accurate information loss and disclosure risk measures from the semantic perspective, or the overcoming of the limitations of concrete anthologies such as WordNet are just some examples.

In the same line, we are also working on the protection of general unbounded non-numerical data [23, 24, 25]. That is, similarly as most statistical disclosure control (SDC) techniques used for numerical attributes, we apply a semantic or ontology based protection to statistical non-numeric data. Once protected, these data can be used for statistical analysis, but also in data mining processes.

We have also presented an evaluation of the application of PPDM techniques in web access logs and query logs in Section 3. We have seen that these data, once protected, still have utility for data mining processes. It is important to note that we are now working towards introducing some semantic PPDM to query logs [12] since our previous work did only consider protection of the data at a syntactic level.

4.2 List of contributions

A list of published or submitted papers by members of the ARES group and directly related to the work presented in this report follows.

4.2.1 ISI JCR Journals

1. Martínez, S., Sánchez, D., Valls, A., Batet, M. Privacy protection of textual attributes through a semantic-based masking method, *Information Fusion*, (under review).

2. Navarro-Arribas, G., Torra, V., Erola, A., and Castellà-Roca, J. User k-anonymity for privacy preserving data mining of query logs. *Information Processing & Management*, (under review).

3. Navarro-Arribas, G., Torra, V., Privacy-preserving data-mining through microaggregation for web-based e-commerce. *Internet Research*, vol. 20 issue 3, pp. 366–384, 2010.

4.2.2 LNCS

4. Erola, A., Castellà-Roca, J., Navarro-Arribas, G., Torra., V., Semantic microaggregation for the anonymization of query logs. In Privacy in Statistical Databases, volume 6344 of Lecture Notes in Computer Science, pages 127–137. Springer Berlin / Heidelberg, 2011.

5. Abril, D., Navarro-Arribas, G., Torra, V., Towards semantic microaggregation of categorical data for confidential documents. In Modeling Decisions for Artificial In- telligence, MDAI 2010, Lecture Notes in Artificial Intelligence, 2010.

6. Martínez, S., Sánchez, D., Valls, A., Ontology-based anonymization of categorical values, 7^{th} International Conference on Modelling Decisions for Artificial Intelligence, LNAI xx, Perpignan, France, October 2010.

4.2.3 Conferences and non ISI-JCR Journals

7. Martínez, S., Sánchez, D., Valls, A., Batet, M., The role of ontologies in the anonymization of textual variables, 13th International Conference of the Catalan Association for Artificial Intelligence, IOS Press, Espluga de Francolí, Spain, October 2010.

8. Navarro-Arribas, G., Torra, V., Erola, A., Castellà-Roca, J., Microagregación para el k-anonimato en registros de buscadores web. In RECSI 2010. XI Reunión Española sobre Criptología y Seguridad de la Información, pages 135–140. Publicacions URV, 2010.

9. Abril, D., Navarro-Arribas, G., Torra, V., Towards privacy preserving information retrieval through semantic microaggregation. In Web Intelligence / Intelligent Agent Technology, pp. 296–299, 2010.

10. Martínez, S, Valls, A., Sánchez, D., Anonymizing categorical data with a recoding method based on semantic similarity, 13^{th} International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), published by Springer: Communications in Computer and Information Sciece, n. 81, pp. 602-611, Dortmund (Germany), July 2010.

4.2.4 Others

Members of our group have also directed the following research Master thesis related to the topic of this report:

- Sergio Martínez Lluís. *Privacy preserving in categorical midrocata using semantic knowledge* Master Thesis, Universidad Rovira i Virgili. 2010.

Bibliography

[1] Abril, D., Navarro-Arribas, G., Torra, V., (2010) Towards privacy preserving information retrieval through semantic microaggregation. In Web Intelligence / Intelligent Agent Technology., pp. 296–299.

[2] Abril, D., Navarro-Arribas, G., Torra, V., (2010) Towards semantic microaggregation of categorical data for confidential documents. In Modeling Decisions for Artificial Intelligence, MDAI 2010, Lecture Notes in Artificial Intelligence.

[3] Agrawal, R., Srikant, R. (2000). Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 439–450.

[4] Baeza-Yates, R., Hurtado, C., Mendoza, M., (2007). Improving search engines by query clustering. Journal of the American Society for Information Science and Technology 58 (12), 1793–1804.

[5] Beeferman, D., Berger, A., (2000). Agglomerative clustering of a search engine query log. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 407–416.

[6] Beitzel, S., Jensen, E., Chowdhury, A., Frieder, O., Grossman, D., (2007). Temporal analysis of a very large topically categorized web query log. Journal of the American Society for Information Science and Technology 58 (2), 166–178.

[7] Broder, A., (1999). Data Mining, the Internet, and Privacy, in Web Usage Analysis and User Profiling: International WEBKDD'99 Workshop. vol. 1836 of LNCS, pp. 56-73.

[8] Cooley, R., Mobasher, B. and Srivastava, J.: (1997), Web mining: information and pattern discovery on the world wide web, Tools with Artificial Intelligence.

[9] Cooper, A., (2008). A survey of query log privacy-enhancing techniques from a policy perspective. ACM Transactions on the Web 2 (4), 1–27.

[10] EFF, 2009. AOL's massive data leak. Electronic Frontier Foundation, `http://w2.eff.org/Privacy/AOL/`.

[11] Eirinaki, M., Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1), February 2003.

[12] Erola, A., Castellà-Roca, J., Navarro-Arribas, G., Torra., V., (2011) Semantic microaggregation for the anonymization of query logs. In Privacy in Statistical Databases, volume 6344 of Lecture Notes in Computer Science, pages 127–137.

[13] Etzioni, O. (1996). The world-wide web: quagmire or gold mine? *Communications of the ACM*, 39(11), November 1996.

[14] EU. Directive 94/46/ec of the European parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Community (281), 1995.

[15] Facca, F., Lanzi, P. (2005). Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*, 5(3):225–241, January 2005.

[16] Han, J., Kamber, M., (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, January 2006.

[17] Hong, Y., He, X., Vaidya, J., Adam, N., Atluri, V., (2009). Effective anonymization of query logs. In: Proceeding of the 18th ACM conference on Information and knowledge management (CIKM'09). pp. 1465–1468.

[18] Defays, D., Nanopoulos, P.: (1993), Panels of enterprises and confidentiality: the small aggregates method, *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada*, pp. 195–204.

[19] Domingo-Ferrer, J., Mateo-Sanz, J. (2002), Practical data-oriented microaggregation for statistical disclosure control, *Knowledge and Data Engineering, IEEE Transactions on* **14**(1), 189 – 201.

[20] Domingo-Ferrer, J., Torra, V. (2005), Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery* **11**(2), 195–212.

[21] Kiss, T., Strunk, J.(2006) Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[22] Kosala, R., Blockeel, H. (2000). Web mining research: a survey. *ACM SIGKDD Explorations Newsletter*, January 2000.

[23] Martínez, S., Sánchez, D., Valls, A., Batet, M., (2010) The role of ontologies in the anonymization of textual variables, 13th International Conference of the Catalan Association for Artificial Intelligence, IOS Press.

[24] Martínez, S., Sánchez, D., Valls, A., (2010) Ontology-based anonymization of categorical values, 7^{th} International Conference on Modelling Decisions for Artificial Intelligence, Springer: Lecture Notes in Artificial Intelligence.

[25] Martínez, S, Valls, A., Sánchez, D., (2010) Anonymizing categorical data with a recoding method based on semantic similarity, 13^{th} International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), Springer: Communications in Computer and Information Science, n. 81, pp. 602-611.

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

[27] Mills, E., Sep. 2006. AOL sued over web search data release. CNET News, `http://news.cnet.com/8301-10784\_3-6119218-7.html`.

[28] Navarro-Arribas, G., Castellà-Roca, J. (2009). W3.T3: Development of privacy preserving transaction logs. Deliverable Project ARES, CONSOLIDER INGENIO 2010 CSD2007-00004.

[29] Navarro-Arribas, G., Torra, V. (2009). Tree-based microaggregation for the anonymization of search logs. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'09). pp. 155–158.

[30] Navarro-Arribas, G., Torra, V., (2010) Privacy-preserving data-mining through microaggregation for web-based e-commerce. *Internet Research*, vol. 20 issue 3, pp. 366–384.

[31] Navarro-Arribas, G., Torra, V. (2009), Towards microaggregation of log files for Web usage mining in B2C e-commerce. The 28th North American Fuzzy Information Processing Society Annual Conference. Special session: Intelligent E-Services and Multi-Agent Systems for Web-Based B2C E-Commerce.

[32] Navarro-Arribas, G., Torra, V., Erola, A., and Castellà-Roca, J. (*under review*). User k-anonymity for privacy preserving data mining of query logs. *Information Processing & Management*.

[33] Oganian, A. and Domingo-Ferrer, J. (2001), On the complexity of optimal microaggregation for statistical disclosure control, *Statistical Journal of the United Nations Economic Commision for Europe* **18**(4), 345–353.

[34] Samarati, P. (2001), Protecting respondents identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering* **13**(6), 1010–1027.

[35] Schafer, J., Konstan, J., Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5(1-2), January 2001.

[36] Srivastava, J., Desikan, P., Kumar, V. (2005). *Foundations and Advances in Data Mining, volume 180/2005 of Studies in Fuzziness and Soft Computing*, chapter Web Mining-Concepts, Applications and Research Directions, pages 275–307. Springer, 2005.

[37] Sweeney, L. (2002), k-anonymity: A model for protecting privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**(5), 557–570.

[38] Torra, V. (2004), Microaggregation for categorical variables: A median based approach, *Proc. Privacy in Statistical Databases (PSD 2004)*, Vol. 3050 of *LNCS*, pp. 162–174.

[39] Torra, V. (2008), Constrained microaggregation: Adding constraints for data editing, *Transactions on Data Privacy* **1**(2), 86–104.

[40] U.S. National Library of Medicine, National Institutes of Health. 2010, Unified Medical Language System (UMLS). `http://www.nlm.nih.gov/research/umls/`

[41] W3C. Platform for privacy preferences (P3P) project, 2008. URL `http://www.w3.org/P3P/`.

[42] Wu, Z., Palmer., M., (1994) Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138. Association for Computational Linguistics.

Appendix A

Overview of Microaggregation

Microaggregation is a statistical disclosure control technique that provides privacy by means of clustering the data into small clusters and then replacing the original data by the centroids of the corresponding clusters.

Privacy is achieved because all clusters have at least a predefined number of elements 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.

Microaggregation was originally [18] defined for numerical attributes, but later extended to other domains. E.g., to categorical data in [38] (see also [20]), and in constrained domains in [39].

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.

From a formal point of view, microaggregation can be defined as an optimization problem with some constraints. We give a formalization below using u_{ij} to describe the partition of the records in the sensitive data set X. That is, $u_{ij} = 1$ if record j is assigned to the ith cluster. Let v_i be the representative of the ith cluster, then a general formulation of microaggregation with g clusters and a given k is as follows:

$$\begin{aligned} \text{Minimize} \quad & SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij} (d(x_j, v_i))^2 \\ \text{Subject to} \quad & \sum_{i=1}^{g} u_{ij} = 1 \text{ for all } j = 1, \ldots, n \\ & 2k \geq \sum_{j=1}^{n} u_{ij} \geq k \text{ for all } i = 1, \ldots, g \\ & u_{ij} \in \{0, 1\} \end{aligned}$$

Algorithm 1: MDAV

Data: X: original data set, k: integer
Result: X': protected data set

```
1  begin
2      while (|X| ≥ 3 * k) do
3          Compute average record x̄ of all records in X;
4          Consider the most distant record x_r to the average record x̄;
5          Form a cluster around x_r. The cluster contains x_r together with
           the k − 1 closest records to x_r;
6          Remove these records from data set X;
7          Find the most distant record x_s from record x_r;
8          Form a cluster around x_s. The cluster contains x_s together with
           the k − 1 closest records to x_s;
9          Remove these records from data set X;
10     if (|X| >= 2 * k) then
11         Compute the average record x̄ of all records in X;
12         Consider the most distant record x_r to the average record x̄;
13         Form a cluster around x_r. The cluster contains x_r together with
           the k − 1 closest records to x_r;
14         Remove these records from data set X;
15     Form a cluster with the remaining records;
16 end
```

For numerical data it is usual to require that $d(x, v)$ is the Euclidean distance. In the general case, when attributes $\mathbf{V} = (V_1, \ldots, V_s)$ are considered, x and v are vectors, and d becomes $d^2(x, v) = \sum_{V_i \in \mathbf{V}} (x_i - v_i)^2$. In addition, it is also common to require for numerical data that v_i is defined as the arithmetic mean of the records in the cluster. I.e., $v_i = \sum_{j=1}^{n} u_{ij} x_i / \sum_{j=1}^{n} u_{ij}$. As the solution of this problem is NP-Hard [33] when we consider more than one variable at a time (multivariate microaggregation), heuristic methods have been developed.

MDAV [19] (Maximum Distance to Average Vector) is one of such existing algorithms. It is explained in detail in Algorithm 1, when applied to a data set X with n records and A attributes. The implementation of MDAV for categorical data is given in [20].

Note that when all variables are considered at once, microaggregation is a way to implement k-anonymity [34, 37].