Màster en Seguretat Informàtica i Sistemes Intel·ligents

UNIVERSITAT ROVIRA I VIRGILI

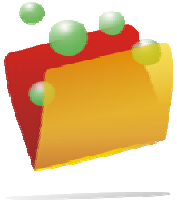
MASTER THESIS

PRIVACY PRESERVING IN CATEGORICAL MICRODATA USING SEMANTIC KNOWLEDGE

Sergio Martínez Lluís

Advisors: Dra. Aïda Valls and Dr. David Sánchez

June, 2010



iTAKA

Intelligent Technologies for Advanced Knowledge Acquisition

Acknowledgements

This work is part of a CONSOLIDER- INGENIO research project funded by the Spanish ministry, called ARES (Advanced Research in Information Security and Privacy). ARES gathers around 60 people from six of the most dynamic Spanish research groups in the area of information security and virtually all existing Spanish groups in the area of information privacy. This work belongs to the workpackage 4, devoted to Data Privacy Technologies.

In particular, the author has worked, as a member of the ITAKA research group, in the team called IF-PAD (Information Fusion for Privacy and Decision) that also includes people from Institut d' Investigació en Intel·ligència Artificial (IIIA-CSIC). The author has been supported by the Universitat Rovira i Virgili predoctoral research grant.

Thanks are given to "Observatori de la Funcació d'Estudis Turístics Costa Daurada" and "Parc Nacional del Delta de l'Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)" for providing us the data collected from the visitors of the park.

I would like to thank ITAKA group members, especially Aïda Valls and David Sánchez.

To my family.

To my wife and my son.

Summary

Exploitation of microdata provided by statistical agencies is very important for many organizations in order to have better knowledge of their customers. The exploration of microdata is usually done with data mining methods that permit to extract useful knowledge from large databases. However, from the Artificial Intelligence (AI) community, very few attention has been paid to the fact that this data often refers to sensible information which can be directly or indirectly associated to individuals. A proper anonymization process is required to minimize the disclosure risk. Several masking methods have been developed for dealing with numerical data or bounded categorical values, but approaches tackling the anonymization of textual values are scarce and shallow. Artificial Intelligence has been a field that has traditionally focused on symbolic data rather than numerical, developing also different techniques to deal with linguistic or textual information.

In this work we present a new masking method aimed to anonymize unbounded textual values using techniques from the field of AI. This method is based on the substitution of sensible record values with other semantically similar ones, creating groups of k-indistinguishable individuals. Taking special attention to the utility of textual information from the data exploitation point of view (which is closely related to the preservation of its meaning), our method relies on the structured knowledge representation given by ontologies to introduce a background context into the masking process. In particular, we exploit the semantic relations offered by WordNet. Ontologies and the theory of semantic similarity are used to guide the masking process towards value substitutions that best preserve the semantics of the original data.

Since textual data typically consist on large and heterogeneous value sets, our method focuses on providing a computationally efficient algorithm by relying on several heuristics instead of exhaustive searches. The method has been evaluated with real data both from theoretical and practical points of view, comparing our results against those provided by more classical approaches, which omit or shallowly consider background knowledge. Evaluation results show that a semantically-grounded anonymization method preserves better the utility of data, offering a low the probability of record linkage.

Index

List of figures

List of tables

1 Introduction

Nowadays the protection of the individuals' privacy is a very important issue in our society because it is a fundamental right. To guarantee and protect the civil liberties and the person rights to the people that participate in surveys and provide their data (usually by means of questionnaires), it is necessary to develop new tools that ensure the privacy of these persons. Usually these data is collected for the National Statistical Offices and some of them are made public in order to enable to third parties (companies, research institutions) to perform studies on them. Statistical Offices never publish directly data that could reveal the person identity (such as D.N.I. or full name), however, sometimes it can be deduced the person identity from a combination of other published data values. For example, in small towns, by publishing the birthplace, birth year and occupation of an individual, one could re-identify the person because, due to the limited data size, this value combination unequivocally identifies him/her. If other sensible data (for example the income, investment actions or others) are also published and associated to values which enable the re-identification, the privacy of confidential data will be compromised.

With the enormous growth of the Information Society and the necessity to enable the access and exploitation of large amounts of data referred to individuals, the preservation of their confidentiality has become a crucial issue. Any survey's respondent (i.e. a person, business or other organization) must be guaranteed that the individual information provided will be kept confidential. *Statistical Disclosure Control* discipline aims at protecting statistical data in a way that it can be released and exploited without publishing any private information that could be linked with or identify a concrete individual. This is achieved by means of a masking algorithm that creates a new anonymized version of the original dataset.

Statistical agencies provide numerical and non-numerical data. In the past, many masking methods have been designed to deal with numerical data [1]. Numbers are easy to manage and compare; so, the quality of the resulting dataset from the utility point of view can be optimized by retaining a set of statistical characteristics [1]. However, the extension of these methods to non-numerical attributes is not straightforward, because of the limitations on defining appropriate aggregation operators on symbols, which have a restricted set of possible

operations. Non-numerical attributes have been treated as categorical variables, defining methods based on a comparison of the words at a string level, or considering some kind of ordering between the words. In those methods, the quality of masked data obtained is typically considered by preserving the distribution of input data.

In this work, we extend previous methods that consider textual data in a categorical fashion by dealing with unbounded variable values which can take labels from a free list of linguistic terms (i.e. potentially the complete language vocabulary). That is, the user is allowed to write the answer to a specific question of the survey using any noun phrase. Some examples of this type of attributes can be “Main hobby” or “Most preferred type of food”.

Unbounded textual variables provide a new way of obtaining information from individuals, which has not been exploited due to the lack of proper anonymization tools. Allowing a free answer, we are able to obtain more precise knowledge of the individual characteristics, which may be interesting for the study that is being conducted. However, at the same time, the privacy of the individuals is more critical, as the disclosure risk increases due to the uniqueness of the answers.

Moreover, this kind of attributes may have a potentially large and rich set of modalities if the individuals are allowed to give responses in textual form. Due to the nature of this kind of values and the ambiguity of human languages, the definition of appropriate aggregation operators is even more difficult. Word semantics play a crucial role in the proper interpretation of this data, a dimension which is commonly ignored in the literature that does not taking into account the semantics of the values. In fact, retaining the semantics of the dataset plays an important role when one aims to extract conclusions by means of intelligence data analysis techniques [2].

This work studies how to integrate Artificial Intelligence techniques to deal with domain knowledge with anonymization methods, traditionally studied in the context of cryptography and information hiding. Its originality consists on treating textual data from a semantic point of view rather than from a categorical (i.e. symbolic) way. The main objective of this work is anonymize textual attributes from a semantic point of view, aiming to get an anonymized dataset as

semantically similar as possible with respect to original data, i.e., retaining the utility of data as a function of their semantics.

Semantic interpretation of textual attribute values for masking purposes requires the exploitation of some sort of structured knowledge sources which allow a mapping between words and semantically interrelated concepts. The use of well-defined general purpose semantic structures, as ontologies (a rigorous and exhaustive organization of some knowledge domain [3]) will allow a better interpretation of data.

1.1 Objectives

To achieve the objective of developing a masking method for textual data using ontologies, we have made an study of current state of the art of privacy preserving methodologies, especially those dealing with non-numerical data (typically in a categorical fashion). Then, a new method will be designed to anonymize unbounded textual attributes semantically. The method has been evaluated using a dataset collected from visitors to the *National Park Delta de l'Ebre*, in Catalonia, Spain using different quality measures.

The main objectives of the work can be summarized as follows:

- Study of works dealing with Statistical Disclosure Control.
- Study of masking methods dealing with categorical data.
- Study of quality metrics aimed to optimize the utility of the categorical data anonymization.
- Study of the possibility of using ontologies as knowledge bases to assist the anonymization of textual data.
- Study of semantic similarity functions which may aid to guide the anonymization in a semantic fashion.
- Design of an algorithm to anonymize potentially unbounded categorical data with a masking method based on semantic similarity and ontologies.
- Implement the proposed method and test it in a case of study created from a real dataset.
- Evaluate the information loss and the quality of the masked dataset, checking up to which point, with the masked dataset, it is possible to

classify and interpret the data in the same way that with non-masked dataset.

- Evaluate the re-identification risk. Assuming that an attacker has the original information of the quasi-identifiers masked, checking up to which point one can associate their registers with the masked file.
- Compare the results (risk and quality) of the new method with respect to related works dealing with textual data in a categorical fashion.

1.2 Document structure

The documentation is divided into:

- Related work: study of privacy methodologies, study of semantic measures and study of quality metrics (section 2). Analyzes the state of the art of anonymizing methodologies, specifically, studies the recoding masking methods for categorical data, also is studied in the section 3 the different semantic similarity measures and quality metrics for the treatment of datasets containing categorical attributes.
- New proposal to anonymize categorical attributes (section 4): In this section is presented our proposal of a methodology to anonymize categorical data taking into account the words semantic, is proposed an algorithm to anonymize categorical data using our methodology.
- In the section 5 we present the tests and evaluations made with statistical text data extracted from a survey. The conclusions and future work (section 6 and 7 respectively) explain the results, the main contributions of this master thesis and the possible extensions in the area, which will be the starting point of my Ph.D. Thesis.
- Appendix A consists on a summary of four research papers that are the result of the work done in the Master Thesis. The summary includes: title, authors, abstract, conference or journal, dates and state of the papers. Three of them have been submitted and we are waiting for the answer, another one has been accepted in an international conference. In Appendix B one can find the full paper of these publications.

1.3 Framework of this master thesis

This work is part of a CONSOLIDER- INGENIO research project funded by the Spanish ministry, called ARES (Advanced Research in Information Security and Privacy). ARES gathers around 60 people from six of the most dynamic Spanish research groups in the area of information security and virtually all existing Spanish groups in the area of information privacy. In particular, I have worked, as a member of the ITAKA research group, in the team called IF-PAD (Information Fusion for Privacy and Decision) that also includes people from *Institut d' Investigació en Intel·ligència Artificial* (IIIA-CSIC).

This work belongs to the workpackage 4 devoted to Data Privacy Technologies.

2 Related work

2.1 Statistical Disclosure Control

Inference control in statistical databases or Statistical Disclosure Control (SDC) aims to disseminate statistical data while preserving confidentiality. Statistical Disclosure Control techniques transform the original database into a new database, taking into account that the protected data satisfies simultaneously utility and security conditions. The dataset will be useful if it is representative of the original dataset and it will be secure if it doesn't allows the re-identification of the original data. There are several areas of application of SDC techniques, which include but are not limited to the following:

Official statistics. Most countries have legislation to guarantee statistical confidentiality when they release data collected from citizens or companies. This justifies the research on SDC (e.g. ESSnet project [4] on the European Union).

Health information. This is an important area regarding privacy. E.g., the Privacy Rule of the Health Insurance Portability and Accountability Act in the U. S., (HIPAA [5]) requires the strict regulation of protected health information for use in medical research. In most others countries, the situation is similar.

E-commerce. The extensive use of electronic commerce generates automatically collection of large amounts of consumer data. This wealth of information is very useful to companies, which are often interested in sharing it with their subsidiaries or partners. Such consumer information transfer should not result in public profiling of individuals and is subject to strict regulation; e.g. in [6] we can consult regulations in the European Union.

The information confidentiality is guaranteed when it is minimized its disclosure risk. The concepts of confidentiality and disclosure are defined in [7] as follow:

Confidentiality: it assures that the dissemination of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited, so that the data are immune from legal processes. Confidentiality differs from privacy because it applies to business as well as

individuals. Privacy is an individual right whereas confidentiality often applies to data on organizations and firms. *Disclosure*: relates to an inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure).

The protection provided by SDC techniques normally entails some degree of data modification, which is an intermediate option between no modification (maximum utility, but no disclosure protection) and data encryption (maximum protection but no utility for the user without clearance).

The challenge for SDC is to modify data in such a way that sufficient protection is provided while keeping at a minimum the information loss. The protection provided by SDC techniques normally entails some degree of data modification, which is an intermediate option between no modification (maximum utility, but no disclosure protection) and data encryption (maximum protection but no utility for the user without clearance).

Statistic databases are those that contain statistic information and can be divided into the following formats:

- *Tabular data*: have been the traditional outputs of national statistical offices. The goal here is to publish static aggregate information, i.e. tables, in such a way that no confidential information on specific individuals among those to which the table refers can be inferred.

- *Dynamic databases*: The scenario here is a database to which the user can submit statistical queries (sums, averages, etc.). The aggregate information obtained by a user as a result of successive queries should not allow him to infer information on specific individuals.

- *Microdata*: files where each register corresponds to information of a subject (person or company). It is only recently that data collectors (statistical agencies and the like) have been persuaded to publish microdata. Therefore, microdata protection is the youngest subdiscipline of Statistical Disclosure Control.

In this work it will refer exclusively to databases of microdata and their concrete protection and masking methods, because their disclosure risk is higher than first two. Tabular data publish aggregated information and its aim is not to contain confidential information that can be inferred. Dynamic databases should also ensure that the successive queries do not allow inferring specific information. But midrodata implies a higher risk of disclosure because as it refers to individual information. Due to this reason, microdata is also the most common data used for data mining, implying that the published information must be also analytically useful.

2.1.1 Statistical disclosure control in microdata

There are two main sources of disclosure risk in a microdata file:

- Existence of attributes with high risk:
    - Some registers of a file can represent subjects with unique features that identify them definitely, for example, uncommon works (actor, judge) very high incomes and others.
    - Many registers of a file can be known to belong to the same cluster, for example, family or college.
    - A data dimension is published by a detail level too fine, for example, the publication of the zipcode.
- Possibility of agreement of a microdata file with extern files: there are some persons or firms that have a unique combination of their attributes. Intruders could use extern files with the same attributes and identifiers to link the unique subjects with their file registers of original microdata.

There are various circumstances that positively affect the disclosure prevention:

- Age of data of the microdata file: The individual and firms features may change significantly over time. The age of the extern files with which one tries link the original file may not match with the original.
- Noise in the information of the microdata file and extern files.

- Different definition of variables of microdata file and extern files.
- Other factors: time, effort and economic resources.

Since the purpose of SDC is to prevent confidential information from being linked to specific respondents, we will assume in what follows that original microdata sets to be protected have been pre-processed to remove from them all identifiers.

The purpose of microdata SDC can be stated more formally by saying that, given an original microdata set D with m records (corresponding m individuals) and n values in each record (corresponding to n attributes that are not identifiers), the goal is to release a protected microdata set D_A (with also m records an n attributes) in such a way that:

1. Disclosure risk (i.e. the risk that a user or an intruder can use D_A to determine confidential attributes on a specific individual among those in D) is low.
2. User analysis (regressions, means, data mining, etc.) on D_A and on D yield the same or at least similar results.

Notice, that the use of the data plays an important role in the anonymization process because the masked version must permit to extract the same knowledge than the original one. With respect to Artificial intelligence techniques, this is important specially if data mining analysis must be done in this data, such as clustering, rules induction, profiling, or prediction, among others. In fact, privacy preserving data mining is a new research field that attempts to develop tools to study in an integrated way how to deal with privacy issues while performing data analysis [8].

Microdata protection methods can generate the protected microdata set D_A:

- Either by masking original data, i.e. generating D_A a modified version of the original microdata set D;
- Or by generating synthetic data D_A that preserve some statistical properties of the original data D.

Masking methods try to ensure that statistics computed on the anonymized dataset do not differ significantly from the statistics that would be obtained on the original dataset. I can be divided in two categories depending on their effect on the original data [9]:

- *Perturbative*: data is distorted before publication. The microdata set is distorted before publication. Thus, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality.

- *Non-perturbative*: data values are not altered but generalized or eliminated [9], [10]. The goal is to reduce the detail given by the original data. This can be achieved with the local suppression of certain values or with the publication of a sample of the original data which preserves the anonymity. Recoding by generalization is also another approach, where several categories are combined to form a new and less specific value.

If we consider the type of data on which they can be used, the classification can be divided in:

- *Numerical*. An attribute is considered numerical if arithmetic operations can be performed with it. Examples are income and age. Note that a numerical attribute does not necessarily have an infinite range, as is the case for age. When designing methods to protect continuous data, one has the advantage that arithmetic operations are possible, and the drawback that every combination of numerical values in the original dataset is likely to be unique, which leads to disclosure if no action is taken.

- *Categorical*. An attribute is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Ordinal and nominal scales can be distinguished among categorical attributes. In ordinal scales the order between values is relevant, whereas in nominal scales it is not. In the former case, max and min operations are meaningful while in the latter case only pair wise comparison is possible. The instruction level is an example of

ordinal attribute, whereas eye colour is an example of nominal attribute. In fact, all sensible values in a microdata set are normally categorical nominal. When designing methods to protect categorical data, the inability to perform arithmetic operations is certainly inconvenient. Allowing unbounded categorical attributes (e.g. a free answer), the same drawback of the continuous attributes is present i.e. the privacy of the individuals is critical, as the disclosure risk increases due to the uniqueness of the answers. This work is focused on the privacy protection of this type of attributes, specifically on the categorical unbounded attributes.

2.2 *K*-Anonymity

One important type of privacy attacks is re-identifying individuals by joining multiple public data sources, e.g. according to [11], around the 87% of the population of the United States can be uniquely identified using their zipcode, gender and date of birth. Anonymization methods must mask data in a way that disclosure risk is ensured at an enough level while minimising the loss of accuracy of the data, i.e. the information loss. A common way to achieve a certain level of privacy is to fulfil the *k-anonymity* property, *k*-anonymity was proposed by Samariti and Sweeney [11] and [12].

To define the k-anonymity concept, previously it is necessary to know the classification of types of attributes that can appear in a dataset, in [13] authors enumerate the various (non-disjoint) possible types of attributes:

- *Identifiers*: the attributes that unambiguously identify the individual, such as the social security number, full name or passport number. To preserve the confidential information, we assume that those attributes must be previously removed or encrypted.

- *Quasi-identifiers*: the attributes that may identify some of the respondents, especially if they are combined with the information provided by other attributes. Unlike identifiers, quasi-identifiers cannot be removed from the dataset because any attribute can potentially be a quasi-identifier.

- *Confidential outcome attributes*: the attributes that contain sensitive information. For example: salary, religion, political affiliation, etc.

- *Non-confidential outcome attributes*: the rest of attributes.

The k-anonymity property tries to keep the balance between the information loss and disclosure risk. Once identified the different types of attributes that can appear in a dataset, we can define the *k*-anonymity concept as [13]:

A dataset is said to satisfy k-anonymity for k > 1 if, for each combination of values of key attributes (e.g. name, address, age, gender, etc.), at least k records exist in the dataset sharing that combination.

An evolution of *k*-anonymity property called p-sensitive k-anonymity is defined in [14]:

A dataset is said to satisfy p-sensitive k-anonymity for k > 1 and p ≤ k if it satisfies k-anonymity and, for each group of tuples with the same combination of key attribute values that exists in the dataset, the number of distinct values for each confidential attribute is at least p within the same group.

Once a value for *k* is fixed (considering a value that keeps the re-identification risk low enough), the goal of the masking method is only to make an anonymization with the less information loss as possible.

2.3 Perturbative masking methods

The microdata set are distorted before publication. May include new data, delete and/or modify the existing data, benefiting the statistic confidentiality.

The main perturbative masking methods are:

- *Additive noise*: add noise with the same correlation structure as the original data. Appropriated method for numerical data. The main noise additions algorithms in the literature are:

    - *Masking with uncorrelated noise addition*: A Register r_i of the original dataset is replaced by a vector $z_i = r_i + \epsilon_i$ where ϵ_i is a vector of normally distributed errors drawn from a random

variable $\epsilon_i \sim N(0, \sigma^2_{\epsilon_i})$, such that $Cov(\epsilon_t, \epsilon_l) = 0 \ \forall \ t \neq l$. This does not preserve variances nor correlations.

- *Masking by correlated noise addition*: preserves means and additionally allows preservation of correlation coefficients. The covariance matrix of the errors is now proportional to the covariance matrix of the original data.

- *Masking by noise addition and linear transformation*: This method ensures by additional transformations that the sample covariance matrix of the masked attributes is an unbiased estimator for the covariance matrix of the original attributes.

- *Masking by noise addition and nonlinear transformation*: An algorithm combining simple additive noise and nonlinear transformation. The advantages of this proposal are that it can be applied to discrete attributes and that univariate distributions are preserved. By contrast, the application of this method is very time-consuming and requires expert knowledge on the data set and the algorithm.

Additive noise is not suitable to protect categorical data. On the other hand, it is well suited for continuous data.

- *Data distortion by probability distribution*: distortion the data with estimated series in function of density of the variables.

- *Microaggregation*: Creates small microclusters, these groups are formed using a criterion of maximal similarity. The size of groups (clusters) must be equal o higher than a variable *k* to guarantee the confidentiality. For each attribute, the average value over each group is computed and is used to replace each of the original averaged values. Once the procedure has been completed, the resulting (modified) dataset can be published.

- *Re-sampling*: Originally proposed for protecting tabular data, re-sampling can be used for microdata. Take *t* independent samples $X_1 \ldots X_t$ of the values of an original attribute V_i. Sort the data of each sample. Calculate the average of the first values of each sample.

Replace those values by the calculate average. Repeat the process with the $n - 1$ values of the next positions.

- *Lossy compression*: consider the dataset as a image and apply compression algorithms (e.g. JPEG)

- *Multiple imputation*: generates a new version of the simulated data created from multiples techniques of imputation from the original data. For example, an imputation method consists on making regressions with a random distribution of the error, to impute "unknown" values to a continuous variable.

- *Camouflage*: camouflage the original information in a range (finite set).it is an appropriate method for numerical data, but causes a high information loss.

- *PRAM (Post-Randomization Method)* [15]: is a perturbative method for privacy protection of categorical attributes in microdata files. In the masked files the original values have been replaced by another different information according to a probabilistic mechanism named Markov matrix. The Markov approach makes PRAM very general, because it merges noise addition, data suppression and data recoding. PRAM information loss and disclosure risk depend on the choice of the Markov matrix. The PRAM matrix contains a row for each possible value of each attribute to be protected. This rules excludes this method from being applicable on continuous data.

- *MASSC (Micro Agglomeration, Substitution, Subsampling and Calibration)* [16] is a masking method that has four steps:

    1. Micro agglomeration is applied to divide the original dataset into groups of records which are at a similar risk of disclosure. These groups are formed using the key attributes, i.e. the quasi-identifiers in the records. The idea is that those records with rarer combinations of key attributes are at a higher risk.

    2. Optimal probabilistic substitution is then used to perturb the original data.

3. Optimal probabilistic subsampling is used to suppress some attributes or even entire records.

4. Optimal sampling weight calibration is used to preserve estimates for outcome attributes in the treated database whose accuracy is critical for the intended data use.

The method is interesting because it is the first attempt to design a perturbative masking method where disclosure risk can be quantified. In practice MASSC is a method only suited when continuous attributes are not present.

- *Data swapping and rang swapping*: The basic idea is to transform a database by exchanging values of confidential attributes among individual records. Data swapping is originally presented as a SDC method for datasets that contains only categorical data, in [17] data swapping was introduced to protect continuous and categorical microdata. Another variant of data swapping is rang swapping, although originally described for ordinal attributes, can also be used for numerical attributes.
  In the rang swapping method, values of an attribute V_i are ranked in ascending order, then each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range (e.g. the rank of two swapped values cannot differ by more than $p\%$ of the total number of records, where p is an input parameter).
  It is reasonable to expect that multivariate statistics computed from data swapped with this algorithm will be less distorted than those computed after an unconstrained swap. In an empirical study [18], rank swapping was identified as a particularly well-performing method in terms of the trade off between disclosure risk and information loss.
- *Rounding*: replace original values of attributes with rounded values, choosing values that belong to a predefined rounding set, often the multiples of a base value. The rounding method is suitable for numerical data. In a multivariate original dataset, usually, rounding is performed one attribute at time, however, multivariate rounding is also possible.

2.4 Non-perturbative masking methods

These techniques do not alter the data of the original set but produce partial suppressions or reductions of detail in the original dataset. Some of the methods are suitable for both continuous and categorical data, but others are only usable for categorical data.

The main non-perturbative methods are:

- *Sampling*: [9] publish a sample of the original set of records. This methodology is suitable for categorical microdata, for continuous microdata would be necessary combine with others masking methods, otherwise, the disclosure risk is high.
- *Global recoding*: also know as generalization [12]. The methodology combines several categories to form new (less specific) categories. For continuous attributes, global recoding means replacing an attribute by its discretized version, but the discretization leads very often to an unaffordable loss of information. This technique is more suitable for categorical attributes, some of these techniques rely on hierarchies of terms covering the categorical values observed in the sample, in order to replace a value by another more general one.
- *Top and bottom coding*: methodology that is a special case of global recoding which can be used if the attribute can be ranked, thus, continuous or categorical. The method determines a threshold for top and bottom values and form new categories with these extreme values. It is a concrete case of global recoding method.
- *Local suppression*: removes certain values with the aim of increase the set of records agreeing on a combination of key values. In [19] proposes ways to combine local suppression and global recoding. Local suppression is rather oriented to categorical attributes. Local suppression is not always allowed as anonymization methodology because sometimes the anonymized dataset must have the same number of records as the original dataset.

The following table summarizes and compare all the presented masking methods with respect to the different data type on can be applied:

Table 1. Masking method vs. data types

Method	Type	Continuous data	Categorical data
Additive noise	P	X	
Data distortion by probability distribution	P	X	X
Microaggregation	P	X	X
Re-sampling	P	X	
Lossy compression	P	X	
Multiple imputation	P	X	
Camouflage	P	X	
PRAM	P		X
MASSC	P		X
Data swapping	P	X	X
Rounding	P	X	
Sampling	NP		X
Global recoding	NP	X	X
Top and bottom coding	NP	X	X
Local suppression	NP		X

N: Perturbative NP: Non-Perturbative X: denotes applicable

As shown in Table 1, on categorical data some of the techniques cannot be applied due to the lack of properly interprets the values that permit to quantify those values in some sense.

In order to fulfil the k-anonymity property, making methods have been designed aiming to build groups of k indistinguishable registers by substituting the original values with a prototype. Obviously, this process results in a loss of information which may compromise the utility of the anonymized data for a further exploitation with data mining techniques. Ideally, the masking method should minimize this loss and maximize data utility according to a certain metric. We can distinguish between global anonymization methods in which all identifier or quasi identifier attributes are considered and anonymized at the same time (i.e. records will fulfil k-anonymity) and local ones in which each attribute is anonymized independently (i.e. each attribute will fulfil k-anonymity individually). In the latter case, the information loss of the whole dataset is not optimized because the transformations only have a local view of the problem.

Notice that the methods for categorical data mainly consider the values a enumerated set of terms, for which only Boolean word matching operations can be performed. On one hand, we can find methods based on data swapping (which exchange values of two different records) and methods that add of some kind of

noise (such as the replacement of values according to some probability distribution done in PRAM [15], [20]). On the other hand, other authors [12], [21] perform local suppressions of certain values or select a sample of the original data aimed to fulfil *k*-anonymity property (see section 2.2) while maintaining the information distribution of input data. We can see that the approaches do not make use of intelligent techniques for dealing with linguistic or textual information, making neither a use of background knowledge to support the anonymization task.

Even though those methods are effective in achieving a certain degree of privacy in an easy and efficient manner, they fail to preserve the meaning of the original dataset, due to their complete lack of semantic analysis. Some exceptions exist in the set of recoding methods, as it is explained in the next section.

2.5 Categorical data anonymization

In this section we analyze in more detail the methods that incorporate some semantics in the anonymization process. Masking of categorical data is not straightforward due to the textual nature of attribute values. Due to this reason, in recent years, some authors have incorporated some kind of knowledge background to the masking process.In particular, some global recoding methods have included a way of performing a semantic interpretation of the categorical attributes.

Most recoding methods (also known as generalization) rely on hierarchies of terms covering the categorical values observed in the sample, in order to replace a value by a more general one. This replacing mechanism uses the semantics given by those hierarchies of terms to determine which value will be used to make the masking. Therefore, these recoding methods are the most similar ones to our proposal using ontologies.

In some recoding methods, the set of values of each categorical attribute of the input records in the dataset are structured by means of Value Generalization Hierarchies (VGHs). Those are ad-hoc and manually constructed tree-like structures defined according to a given input dataset, where categorical labels of an attribute represent leafs of the hierarchy and they are recursively subsumed by common generalizations. The recoding masking process consists on, for each attribute, substituting several original values by a more general one, obtained from

the hierarchical structure associated to that attribute. This generalization process decreases the number of distinct tuples in the dataset and, in consequence, increases the level of k-anonymity. In general, for each value, different generalizations are possible according to the depth of the tree. The concrete substitution is selected according to a metric that measures the information loss of each substitution with regards to the original data.

The following summarizes the related work that uses recoding masking methods for categorical data.

- Samariti & Sweeney in [12], Bayardo & Agrawal in [22] and Lefreve, DeWitt and Ramakrishnan in [23] propose a global hierarchical scheme in which all values of each attribute are generalized to the same level of the VGH. The number of valid generalizations for each attribute is the height of the VGH for that attribute. For each attribute, the method picks the minimal generalization which is common to all the record values for that attribute. In this case, the level of generalization is used as a measure of information loss.
- Iyengar [24] presented a more flexible scheme that also uses a VGH, where a value of each attribute can be generalized to a different level of the hierarchy in different steps. This scheme allows a much larger space of possible generalizations. Again, for all values and attributes, all the possible generalizations fulfilling the k-anonymity are generated. Then, a genetic algorithm finds the optimum one according to a set of information loss metrics measuring the distributional differences with regards to the original dataset.
- T. Li and N. Li [25] propose three global generalization schemes:
    - The Set Partitioning Scheme (SPS) represents an unsupervised approach in which each possible partition of the attribute values represents a generalization. This supposes the most flexible generalization scheme but the size of the solution space grows enormously, meanwhile the benefits of a semantically coherent VGH are not exploited.
    - The Guided Set Partitioning Scheme (GSPS) uses a VGH per attribute to restrict the partitions of the corresponding attribute and uses the height of the lowest common ancestor of two values as a metric of semantic distance.

- The Guided Oriented Partition Scheme (GOPS) adds ordering restrictions to the generalized groups of values to restrict even more the set of possible generalizations.

Notice that in the three cases, all the possible generalizations allowed by the proposed scheme for all attributes are constructed, selecting the one that minimizes the information loss (evaluated by means of the discernability metric [22]).

- He and Naughton [26] propose a local partitioning algorithm in which generalizations are created for an attribute individually in a Top-Down fashion (recursively). The best combination, according to quality metric (Normalized Certainty Penalty [27]), is recursively refined.
- Xu et al. [10] also proposes a local generalization algorithm based on individual attribute utilities. In this case, the method defines different "utility" functions for each attribute, according to their importance. Being local methods, each attribute is anonymized independently, resulting in a more constrained space of generalizations (i.e. it is not necessary to evaluate generalization combinations of all attributes at the same time). However, the optimization of information loss for each attribute independently does not imply that the result obtained is optimum when the whole record is considered. As stated in the introduction, non necessary generalizations would be typically done in a local method as each attribute should fulfil *k*-anonymity independently.
- Bayardo and Agrawal [22] propose an alternative to the use of VGHs. Their scheme is based on the definition of a total order over all the values of each attribute. According to this order, partitions are created to define different levels of generalization. As a result, the solution space is exponentially large. The problem here is that the definition of a semantically coherent total order for categorical attributes is very difficult and nearly impossible for unbounded textual data. Moreover, the definition of a total order unnecessarily imposes constraints on the space of valid generalizations.

In order to compare the related work, Table 2 summarizes their main characteristics, regarding to the type of anonymization, the use of different knowledge structures to guide the process of masking, the global vs local

approach, the metric used to measure the quality of the result (more details about these measures are given in the next section) and, finally, the algorithmic search scheme. We have also included the features of the method that we will explain in section 4, in order to facilitate the comparison of our proposal with the previous work.

Table 2. Related work comparison

Work	Anonymization method	Background knowledge	Global/ Local	Quality metric	Algorithm type
Samariti & Sweeney [12]	generalization & suppression	small ad-hoc VGH	global	no	Exhaustive
Bayardo & Agrawal [22]	generalization & suppression	small ad-hoc VGH	global	D.M.	Heuristic
Lefreve, DeWitt & Ramakrishnan [23]	generalization	small ad-hoc VGH	global	D.M.	Exhaustive
Iyengar [24]	generalization	small ad-hoc VGH	global	L.M.	Genetic algorithm
Li & Li [25] SPS	generalization	Partition	global	D.M.	Heuristics
Li & Li [25] GSPS	generalization	small ad-hoc VGH	global	D.M.	Heuristics
Li & Li [25] GOPS	generalization	small ad-hoc VGH	global	D.M.	Heuristics
He and Naughton [26]	generalization	small ad-hoc VGH	local	N.C.P.	Recusrvive
Xu et al. [10]	generalization	small ad-hoc VGH	local	N.C.P. & D.M.	Heuristics
Our method (section 4)	substitution	Ontology (WordNet)	global	Semantic similarity	Heuristics

All the approaches relying on VGHs present some drawbacks. On one hand, VGHs are manually constructed from each attribute value set of the input data (i.e. categorical values directly correspond to leafs in the hierarchy). So, human intervention is needed in order to provide the adequate semantic background in which those algorithms rely. If input data values change, VGHs should be modified accordingly. Even though this fact may be assumable when dealing with reduced sets of categories (e.g. in [25] a dozen of different values per attribute are considered in average), this hampers the scalability and applicability of the approaches, especially when dealing with unbounded textual data (with potentially hundreds or thousands of individual answers). On the other hand, the fact that VGHs are constructed from input data (which represents a limited sample of the underlying domain of knowledge), produces ad-hoc and small hierarchies with a much reduced taxonomical detail. It is common to observe VGHs with

three or four levels of hierarchical depth whereas a detailed taxonomy (such as WordNet) models up to 16 levels [28] (see section 3.1.1). From a semantic point of view, VGHs offer a rough and biased knowledge model compared to fine grained and widely accepted ontologies. As a result, the space for valid generalizations that a VGH offers would be much smaller than when exploiting an ontology. Due to the coarse granularity of VGHs, it is likely to suffer from high information loss due to generalizations. As stated above, some authors try to overcome this problem by trying all the possible generalizations exhaustively, but this introduces a considerable computational burden and lacks of a proper semantic background. Therefore, the quality of the results heavily depends on the structure of VGHs that, due to their limited scope, offer a partial and biased view of each attribute domain.

2.6 Quality metrics for anonymized categorical data

A common way to anonymize a dataset and achieving a certain level of privacy is to fulfill the k-anonymity property, once a value *k* that keeps the re-identification risk low enough is selected, the main objective is to k-anonymize with the least information loss as possible. When anonymizing categorical data, particularly using a recoding masking method, anonymization incurs information loss when a detailed item is generalized to its more generic super-category. The goal of anonymization in general is to find a transformation of the original data that satisfies a privacy model while minimizing the information loss and maximizing the utility of the anonymized data. Thus a metric is necessary to measure the quality of the resulting data. The difference between the original dataset and the anonymized dataset measures the quality of the anonymization. In the literature we can find several ways to measure the quality of the dataset anonymized by a recoding masking method.

The main quality metrics can be grouped as:

- Distributional models: the quality measure only evaluates the distribution of the results groups. These models don't use VGH in the measurement and, therefore, don't incorporate knowledge. The main distributional quality metrics are:
    - *Discernability model (DM)* [22] (1): are used to evaluate the distribution of m records (corresponding to m individuals) into

g groups of identical values, generated after the anonymization process. Concretely, DM assigns to each record a penalty based on the size of the group g_i to which it belongs after the generalization. A uniform distribution of values in groups of similar size would optimize this metric.:

$$C_{DM} = \sum_{i=1}^{m} |g_i|^2 \quad (1)$$

- *Normalized Averaged Equivalence class size metric* (CAVG) [23]: the intuition of the metric is to measure how well the partition in g groups approaches the best case, where each record is generalized in a group of k indistinguishable individuals:

$$C_{AVG} = \frac{\text{number of tuples in the table}}{\text{numbers of group\_bys on quasi\_identifier} \cdot k} \quad (2)$$

- VGH-based models: the quality measure evaluates the information loss as a function of distance calculated on the VGH. These models incorporate knowledge in the measure. The main VGH-based models are:
  - *General loss metric* (LM) [24]: more accurate metric, computed by summing up a normalized information loss for each of these columns. This information loss for a column will be computed as the average loss for each entry in the column. The information loss of each entry is calculated as: Let the total number of leaf nodes in T be denoted by M. Let the number of leaf nodes in the subtree rooted at node P be M_P . Using this simplified model and normalizing using the worst case situation when the generalized node is the root of the taxonomy tree leads to $(M_P - 1)/(M - 1)$ as the loss for this entry. The information loss for a suppressed entry is the same as the loss when the generalized value corresponds to the root of the tree.
  - *Normalized Certainty Penalty* (NCP) [10] (3): is similar to Loss Metric. In the case of categorical attributes NCP is

defined for items in a generalization hierarchy. Let p be an item or its generalization. Then:

$$NCP(p) = \begin{cases} 0, & |u_p| = 1 \\ |u_p|/|I|, & otherwise \end{cases} \quad (3)$$

Where u_p is the node in the tree corresponding to p, and $|u_p|$ is the total number of leaf nodes under that node. The first equation states that when the item is not generalized and published as is, there is no information loss. The second equation states that the information loss for a generalized item is the number of leaves it covers divided by total number of leaves in the hierarchy. The maximum information loss is when an item is generalized to the root of the hierarchy. Then the total information loss according NCP of a generalized database D is defined as (4):

$$NCP(D) = \frac{\sum_{t \in D} \sum_{p \in t} NCP(p)}{\sum_{t \in D} C_t} \quad (4)$$

Thus the overall information loss of an anonymized set is the weighted average of the information loss of all instances of items.

On the distributional model, the quality of masked non-numerical data is only considered by preserving the distribution of the input data. On the other hand, the VGH-based model incorporated a poor semantic knowledge (see section 2.5)

Even though data distribution is a dimension of data utility, we argue, as it has been stated by other authors [10] that retaining the semantics of the dataset plays a more important role when one aims to extract conclusions by means of intelligent data analysis. For this reason in section 4 we will propose a new way of measuring the quality of the anonymization, using a knowledge-based approach.

3 Semantic interpretation of categorical data

Semantic interpretation of textual attribute values for masking purposes requires the exploitation of some sort of structured knowledge sources which allow a mapping between words and semantically interrelated concepts. As it explained in Section 2.5, some privacy approaches have incorporated some sort of background knowledge during the masking process. However, the lightweight and ad-hoc nature of that knowledge and the shallow semantic processing of data hamper their applicability as a general-purpose solution. On the contrary, we argue that the use of well-defined general purpose semantic structures, as ontologies, will allow a better interpretation of data [29], [30]. Ontologies are formal and machine readable structures of shared conceptualisations of knowledge domains, expressed by means of semantic relationships. Thanks to initiatives such as the Semantic Web [32], many ontologies have been created in the last years, such as general purpose ones or specific domain or task ontologies. In this section we present a review of the most important concepts related to ontologies, as well as semantic similarity measures.

3.1 Ontologies

Ontology, in information science, can be defined as a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations. It is used to reason about the properties of that domain, and may be used to describe the domain. In this section, the ontological paradigm is formalized, and the knowledge representation possibilities of modern ontological languages are analyzed.

In [3] an ontology (O) has been defined as:

$$O = (C, \leq_C, R, \sigma_R, \leq_R, A, \sigma_A, T)$$

,where

- C, R, A and T represent disjoint sets of concepts, relations, attributes and data types. Concepts (or classes) are sets of real world entities with common features (such as different types of diseases, treatments, actors, etc.). Relations are binary associations between concepts. There exist

inter-concept relations, which are common to any domain (such as hyponymy, meronymy, etc.) and domain-dependant associations (e.g., an Actor performs an Action). Attributes represent quantitative and qualitative features of particular concepts (e.g., the medical code of a Disease), which take values in a given scale defined by the data type (e.g., string, integer, etc.).

- $\leq C$ represents a concept hierarchy or taxonomy for the set C. In this taxonomy, a concept *c1* is a subclass, specialization or subsumed concept of another concept *c2* if and only if every instance of *c1* is also an instance of *c2* (which represent its superclass, generalization or subsumer). Concepts are linked by means of transitive is-a relationships (e.g., if respiratory disease is-a disorder and bronchitis is-a respiratory disease, then it can be inferred that bronchitis is-a disorder). Multiple inheritances (i.e., the fact that a concept may have several hierarchical subsumers) are also supported (for example, dog may be both a subclass of canine and pet).
- $\leq R$ which represents a hierarchy of relations (e.g., has primary cause may be a specialization of the relation has cause, which indicates the origination of a Disorder).
- σR: $R \rightarrow C+$ refers to the signatures of the relations, defining which concepts are involved in one specific relation of the set R. It is worth noting that some of the concepts in $C+$ correspond to the domain (the origin of the relation) and the rest to the range (the destination of the relation). Those relationships may fulfill axioms such as functionality, symmetry, transitivity or being the inverse to another one. Relations between concepts are also called object properties.
- σA: $A \rightarrow C \times T$ represents the signature describing an attribute of a certain concept C, which takes values of a certain data type T (e.g., the number of leukocytes attribute of the concept Blood Analysis, which must be an integer value). Attributes are also called data type properties.

Additionally, an ontology can be populated by instantiating concepts with real world entities (e.g., St. Eligius is an instance of the concept Hospital). Those are called instances.

By default, concepts may represent overlapping sets of real entities (i.e., an individual may be an instance of several concepts, for example a concrete disease

may be both a Disorder and a Cause of another pathology). If necessary, ontology languages permit specifying that two or more concepts are disjoint (i.e., individuals cannot be instances of more than one of those concepts).

3.1.1 WordNet

Nowadays, there exist massive and general purpose ontologies like WordNet [28]. WordNet is a general purpose semantic electronic repository for the English language. It is the most commonly used online lexical and semantic database. In more detail it offers a lexicon, a thesaurus and semantic linkage between the major part of English terms. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. It seeks to classify words into many categories and to interrelate the meanings of those words. It groups English words into sets of synonyms called synsets, provides short, general definitions. A synset is a set of words that are interchangeable in some context, because they share a commonly-agreed upon meaning with little or without variation. Each word in WordNet has a pointer to at least one synset. Each synset, in turn, must point to at least one word. It is useful to think of synsets as nodes in a graph. A semantic pointer is simply a directed edge in the graph whose nodes are synsets.

These relations between synsets vary based on the type of word, In the case of nouns, the main relation types includes:

- Hyponym: *X* is a hyponym of *Y* if *X* is a (kind of) *Y* (*dog* is a hyponym of *canine*).
- Hypernym: *X* is a hypernym of *Y* if *Y* is a (kind of) *X* (*canine* is a hypernym of *dog*).
- Holonym: *X* is a holonym of *Y* if *Y* is a part of *X* (*building* is a holonym of *window*).
- Meronym: *X* is a meronym of *Y* if *X* is a part of *Y* (*window* is a meronym of *building*).
- Coordinate terms: *Y* is a coordinate term of *X* if *X* and *Y* share a hypernym (*wolf* is a coordinate term of *dog*, and *dog* is a coordinate term of *wolf*)

Each synset also contains a description of its meaning that is expressed in natural language as a gloss. Some example sentences of typical usage of that synset are also given.

The Table 3 summarizes the WordNet 2.1 database statistics (number of words, synsets and senses

Table 3. WordNet 2.1 database statistics

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117.097	81.426	145.104
Verb	11.488	13.650	24.890
Adjective	22.141	18.877	31.302
Adverb	4.601	3.644	5.720
Totals	155.327	117.597	207.016

The result is a network of meaningfully related words, where the graph model can be exploited to interpret concept's semantics. Hypernymy is, by far, the most common relation, representing more than an 80% of all the modelled semantic links. The maximum depth of the noun hierarchy is 16. Polysemous words present an average of 2.77 synsets (i.e. they belong to almost three different hierarchies).

Considering those dimensions, the use of WordNet instead of VGHs as semantic background for data anonymization would result in a generalization space which size would be several orders of magnitude bigger. In fact, as most of the related works make generalizations in an exhaustive fashion, the generalization space is exponentially large according to the depth of the hierarchy, the branching factor, the values and the number of attributes to consider. So, those approaches are computationally too expensive and hardly applicable in such a big ontology like WordNet. A solution will be provided in this Master Thesis.

3.2 Ontology-based semantic similarity

In general, the assessment of concept's similarity is based on the estimation of semantic evidence observed in a knowledge resource. So, background knowledge is needed in order to measure the degree of similarity between concepts. From the similarity point of view, taxonomies and, more generally, ontologies, provide a graph model in which semantic interrelations are modeled as links between concepts. Many approaches have been developed to exploit this geometrical model, computing concept similarity as inter-link distance.

In order to guide the anonymization process towards the transformation that would result in the minimum information loss, a similarity measure that evaluates

the semantic difference between the original data and the data resulting from each transformation is needed. To determine the most appropriate measure to guide the masking process, it is necessary the study of the different semantic similarity measures.

In the literature, we can distinguish several different approaches to compute semantic similarity according to the techniques employed and the knowledge exploited to perform the assessment. The most classical approaches exploit structured representations of knowledge as the base to compute similarities.

3.2.1 Edge counting-based measures

By mapping input terms to ontological concepts by means of their textual labels, a straightforward method to calculate the similarity between terms is to evaluate the Path Length connecting their corresponding ontological nodes via is-a links [33]. As the longest the path, the more semantically far the terms appear to be, this defines a semantic distance measure. The most basic edge counting-based measures are:

- *Path Length* [33]: in an is-a hierarchy, is the simplest way to estimate the distance between two concepts c_1 and c_2. Consist of calculating the shortest Path Length (i.e. the minimum number of links) connecting c_1 and c_2 concepts.

$$dis_{pL}(c_1, c_2) = \min \#\ of\ is-a\ edges\ connecting\ c_1\ and\ c_2 \quad (5)$$

- *Leacock and Chodorow* [34] also proposed a measure in order to normalize this distance dividing the path length between two concepts (N_p) by the maximum depth of the taxonomy (D) in a non-linear fashion (6). The function is inverted to measure similarity.

$$sim_{L\&C}(c_1, c_2) = -\log(N_p/2D) \quad (6)$$

- Wu and Palmer [35]: However, those measures omit the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered as less similar than those belonging to a lower level, as they present different degrees of generality. Based on this premise Wu and Palmer's measure also takes into account the depth of the concepts in the hierarchy (6).

$$sim_{W\&P}(c_1, c_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (7)$$

where N_1 and N_2 are the number of is-a links from c_1 and c_2 respectively to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

The main advantage of the presented measures is their simplicity. They only rely on the geometrical model of an input ontology whose evaluation requires a low computational cost. However, several limitations hamper their performance.

In general, any ontology-based measure would depend on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology. So, they require rich and consistent ontologies like WordNet to work properly [36].

A problem of path-based measures typically acknowledged [37] is that they rely on the notion that all links in the taxonomy represent a uniform distance. Wide ontologies with a relatively homogenous distribution of semantic links and good domain coverage minimize these problems [38].

3.2.2 Feature-based measures

On the contrary to edge-counting measures which, as stated above, are based on the notion of path distance (considered in a uniform manner), feature-based approaches assess similarity between concepts as a function of their properties.

By features, authors exploit the information provided by the input ontology. For WordNet, concept synonyms (i.e. synsets, which are sets of linguistically equivalent words), definitions (i.e. glosses, containing textual descriptions of word senses) and different kinds of semantic relationships can be exploited.

- Similarity, in Tversky [39] concepts and their neighbours (according to semantic pointers) are represented by synsets. The similarity is computed as:

$$sim_{Tve}(c_1, c_2) = \frac{|A \cap B|}{|A \cap B| + \gamma(c_1, c_2)|A \backslash B| + \big(1 - \gamma(c_1, c_2)\big)|B \backslash A|} \quad (8)$$

Where *A, B* are the synsets for concepts corresponding to c1 and *c2, A\B* is the set of terms in *A* but not in *B* and *B\A* the set of terms in *B* but not in *A*. *γ(c1, c2)* is computed a function of the depth of *c1* and *c2* in the taxonomy:

$$\gamma(c_1, c_2) = \begin{cases} \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, depth(c_1) \leq depth(c_2) \\ 1 - \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, depth(c_1) > depth(c_2) \end{cases} \quad (9)$$

- Rodriguez [40], the similarity is computed as the weighted sum of similarities between synsets, features and neighbour concepts of evaluated terms:

$$sim_{rod}(c_1, c_2) = w \cdot S_{synsets}(c_1, c_2) + u \cdot S_{features}(c_1, c_2) + v \cdot S_{neighborhoods}(c_1, c_2) \quad (10)$$

- Petrakis [41] a feature-based function called *X-similarity* relies on matching between synsets and concept's glosses extracted from WordNet (i.e. words extracted by parsing term definitions). They consider that two terms are similar if the synsets of their concepts and the synsets of concepts in their neighbourhood (following is-a and part-of links) and their glosses are lexically similar. The similarity function is expressed as follows:

$$sim_{X-Similarity}(c_1, c_2) = \begin{cases} 1, if\ S_{synsets}(c_1, c_2) > 0 \\ max\left(S_{neighborhods}(c_1, c_2), S_{descriptions}(c_1, c_2)\right), if\ S_{synsets}(c_1, c_2) = 0 \end{cases} \quad (11)$$

where $S_{neighborhods}$ is computed as:

$$S_{neighborhods}(c_1, c_2) = \max \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (12)$$

where *A* and *B* denote synsets or description sets for term *a* and *b*.

Feature-based measures exploit more semantic evidences than edge-counting approaches, evaluating both commonalties and differences of compared concepts. However, by relying on features like glosses or synsets (in addition to taxonomic and non-taxonomic relationships), those measures limit their applicability to ontologies in which this

information is available. Another problem is their dependency on weighting parameters that balance the contribution of each feature

3.2.3 Information Content-based measures

Resnik [42] proposed measure the quality of an anonymized set calculating and comparing the information content of both original and result sets. Information content (IC) of a concept c is the inverse to its probability of occurrence. IC computation is based on the probability $p(c)$ of encountering a concept c in a given corpus. In this way, infrequent words obtain a higher IC.

$$IC(c) = -\log\, p(c) \tag{13}$$

The main IC-based similarity measures are:

- *Resnik* [42] introduced the idea of computing the similarity between a pair of concepts (c_1, c_2) as the IC of their Least Common Subsumer (LCS), which is the most concrete taxonomical ancestor common c_1 and c_2 in a given ontology. This gives an indication of the amount of information that the two concepts share in common. The more specific he subsumer is (higher IC), the more similar the terms are.

$$sim_{res}(c_1, c_2) = IC\big(LCS(c_1, c_2)\big) \tag{14}$$

- Lin similarity [43] is an extension of Resnik's measure. This measure depends on the relation between the information content of the LCS of two concepts and the sum of the information content of the individual concepts (c_1, c_2).

$$sim_{lin}(c_1, c_2) = \frac{2 \times sim_{res}(c_1, c_2)}{\big(IC(c_1) + IC(c_2)\big)} \tag{15}$$

- Jiang and Conrath presented in [38] another extension of Resnik's measure that subtract the information content of the LCS from the sum of the information content of the individual content.

$$dis_{jcn}(c_1, c_2) = \big(IC(c_1) + IC(c_2)\big) - 2 \times sim_{res}(c_1, c_2) \tag{16}$$

Note that this function is a dissimilarity measure because the more different the terms are, the higher the difference from their IC to the IC of their LCS will be.

Finally, other approaches aiming to compute semantic alikeness exploit the notion of concept's Context Vector. They are based on the premise that words are similar if their contexts are similar. In this case, vectors are constructed from the context of words extracted of the text. Then, the semantic relatedness of two concepts c_1 and c_2 is computed as the cosine of the angle between their context vectors [44].

$$rel_{vector}(c_1, c_2) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|v_1| \cdot |v_2|} \quad (17)$$

Where $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ are the context vectors corresponding to c_1 and c_2 respectively.

Using the information offered by WordNet and the measures seen above, it is possible to compute the similarity between concepts. There have been some initiatives for computing some standard measures that have been widely used by several authors, such as the software WordNet::Similarity [45].

WordNet is particularly well suited for similarity measures, since it organizes nouns into is-a hierarchies and, therefore, it can adequate to evaluate taxonomics relationships.

3.3 Evaluation of semantic similarity measures

After the study of the different semantic similarity measures, it is necessary to evaluate and compare between them, in order to determine and select the most appropriate one to guide the masking process of our method.

An objective evaluation of the accuracy of a semantic similarity function is difficult because the notion of similarity is subjective [37]. In order to enable fair comparisons, several authors created evaluation benchmarks consisting on word pairs whose similarity were assessed by a set of humans. Rubenstein and Goodenough [46] defined the first experiment in 1965 in which a group of 51 students, all native English speakers, assessed the similarity of 65 word pairs selected from ordinary English nouns on a scale from 0 (semantically unrelated) to 4 (highly synonymous). Miller and Charles [47] re-created the experiment in 1991 by taking a subset of 30 noun pairs which similarity was reassessed by 38 undergraduate students. The correlation obtained with respect to Rubenstein and Goodenough experiment was 0.97. Resnik [42] replicated again the same

experiment in 1995, in this case, requesting 10 computer science graduate students and post-doc researchers to assess similarity. The correlation with respect to Miller and Charles results was 0.96. Finally, Pierro [36] replicated and compared the three above experiments in 2008, involving 101 human subjects, both English and non-English native speakers. He obtained an average correlation of 0.97. It is interesting to see the high correlation obtained between the experiments even though being performed in a period of more than 40 years and a heterogeneous set of human subjects. This means that similarity between the selected words is stable over the years, making them a reliable source for comparing similarity measures.

In fact, Rubenstein and Goodenough and Miller and Charles benchmarks have become de facto standard tests to evaluate and compare the accuracy of similarity measures. As a result, correlation values obtained against those benchmarks can be used to numerically quantify the closeness of two ratings sets (i.e. the human judgments and the results of the computerized assessment). If the two rating sets are exactly the same, correlations coefficient is 1 whereas 0 means that there is no relation. Correlations coefficients have been commonly used in the literature; both are equivalent is ratings sets are ordered (which is the case). They are also invariant to linear transformations which may be performed over results such as change between distance and similarity (for the corresponding functions) or normalizing values in a range. This enables a fair and objective comparison against different approaches.

So, we have taken the correlation values originally reported by related works for Rubenstein and Goodenough and Miller and Charles benchmarks (when available) and summarized in Table 4. In case in which a concrete measure depends on certain parameters (such as weights or corpora selection/processing) the best correlation value reported by the authors was compiled. It is important to note that, even though some of them rely on different knowledge sources (such as tagged corpora or the Web), all ontology-based ones use WordNet. WordNet 2 is the most common version used in related works. In cases in which original authors used an older version (WordNet 2 was released in July 2003), we took a replication of the measure evaluation performed by another author in order to enable a fair comparison. As a result, we picked up results reported by authors in papers published from 2004 to 2009.

Table 4. Correlation values for each measure. From left to right: measure authors, family type, correlation reported for Miller and Charles benchmark, correlation reported for Rubenstein and Goodenough benchmark.

Measure	Type	M&C	R&G	Evaluated in
Path (Rada)	Edge	0.59	N/A	X-sim06 [41]
Wu & Palmer	Edge	0.74	N/A	X-sim06 [41]
Leacock & Chodorow	Edge	0.74	0.77	EACL06 [44]
Rodriguez	Feature	0.71	N/A	X-sim06 [41]
Tversky	Feature	0.73	N/A	X-sim06 [41]
Petrakis	Feature	0.74	N/A	X-sim06 [41]
Resnik	IC	0.72	0.72	EACL06 [44]
Lin	IC	0.7	0.72	EACL06 [44]
Jiang & Conrath	IC	0.73	0.75	EACL06 [44]

Correlation values indicate that measure accuracies are very similar through the different families. However, the applicability and generality of each measure type depend on the principle they exploit.

On one hand, with respect to feature-based type measures, the main problem is their dependency on weighting parameters that balance the contribution of each feature. In all cases, those parameters should be tuned according the nature of the ontology and even to the evaluated terms. This hampers their applicability as a general purpose solution. Only Petrakis [41] does not depend on weighting parameters, as the maximum similarity provided by each feature alone is taken. Even though this adapts the behavior of the measure to the characteristics of the ontology and the knowledge modeling, by taking only the maximum value at each time the contribution of other features is omitted.

On the other hand, the I.C. type measures need an accurate computation of concept probabilities that requires a proper disambiguation and annotation of each noun found in the corpus. If either the taxonomy or the corpus changes, re-computations are needed to be recursively executed for the affected concepts. So, it is necessary to perform a manual and time-consuming analysis of corpora and resulting probabilities would depend on the size and nature of input corpora. Moreover, the background taxonomy must be as complete as possible (i.e. it should include most of the specializations of a specific concept) in order to provide reliable results. Partial taxonomies with a limited scope may not be suitable for this purpose. All those aspects limit the scalability and applicability of those approaches.

In this work, we have chosen the edge counting-based type family as the similarity measure for testing purposes (see evaluation section 5), because, as they neither depend on corpora nor tuning parameters they present a low computational cost and lack of constraints. This ensures their applicability and generality especially when dealing with large sets of data, which is common when anonymizing data.

4 New proposal to anonymize categorical attributes

Our method addresses the problem of masking a subset of the unbounded categorical attributes with a global masking approach. As it has been said in the section 2.2, four different types of attributes are distinguished: identifiers, quasi-identifiers confidential and non-confidential. Only the first two may lead to the re-identification of individuals. Identifiers are directly removed from the dataset (or encrypted) because they refer to values that are unique for each individual (e.g. personal identification number or social security number). As a consequence, the masking process would be applied over tuples of textual quasi-identifier attributes.

In order to overcome the limitations identified in related works, in this section we propose a global masking method for unbounded textual values, based on the merging of quasi-identifier values of the input records. This method permits to build groups of indistinguishable registers with multiple textual attributes in a way in which k-anonymity is fulfilled. The method relies on the well-defined semantics provided by big and widely used ontologies like WordNet. This permits to properly interpret words' meaning and maximize the quality of the anonymized data from the semantic point of view. The aim is that the conclusions that may be inferred from the masked dataset by means of data analysis methods would be the most similar to those obtained from the original data. Due to potentially large size of ontologies (with respect to ad-hoc knowledge structured exploited in previous approaches [11], [12], [22], [25], [26]) and the fact of dealing with potentially unbounded textual attributes with a large set of distinct values, we propose a non-exhaustive heuristic approach which provides better scalability (with respect to the size of the ontology and the input data) than related works. To perform the anonymization using in a knowledge-based approach, the Wordnet ontology will be used as background knowledge because, as semantic electronic repository for the English language, it is the most commonly used lexical and semantic database.

As explained above (section 3.1.1), exhaustive generalization methods are computationally too expensive to be applicable with unbounded textual attributes and large ontologies like WordNet. Moreover, the fact that values to anonymize correspond to leafs of the VGH implies that values are only substituted by more

general ones (which unnecessarily imposes constraints on the space of valid transformations).

4.1 Ontology-based method to mask textual attributes

Our approach deals with the global masking process in a new way. Thanks to the wide coverage of WordNet, one would be able to map textual attribute values into ontological nodes which do not necessary represent leafs of a hierarchy. As a result, semantically related concepts can be retrieved going through the ontological hierarchy/ies to which the value belongs. Those ontological hierarchies are designed in a much general and fine grained fashion than ad-hoc VGHs and, according to the agreement of domain knowledge experts, not in function on the input data. Those facts open the possibility of substituting values by a much wider and knowledge-coherent set of semantically similar elements. In order to ensure the scalability with regards on the ontology size and the input data, we bound the space of valid value changes to the set of the value combinations that can be found in the input dataset. When changing a value of a record for another, one may represent a taxonomical subsumer to the other (which is the only case covered by generalization method) but also a hierarchical siblings (with the same taxonomical depth) or a specialization (located in a lower level). In fact, in many situations, a specialization may be more similar than a subsumer because, as stated in section 3.2.1, concepts belonging to lower levels of a hierarchy have less differentiated meanings due to their higher concreteness. As a result, the value change would result in less information loss and a higher preservation of data utility from a semantic point of view. This is an interesting characteristic and an improvement over the more restricted data transformations supported by VGH-based generalization methods.

In a nutshell, the method proposed is based on the fusion of quasi-identifier values of each record with the values of another record. In order to select the value that minimizes the information loss resulting from the data substitution, a semantic metric (studied in section 3.2.1 and evaluated in section 3.3) is used to select the most similar one. As a result of the fusion, quasi-identifier values for both records (the one to anonymize and the most semantically similar one) will take the same values and will become indistinguishable; so, the k-anonymity level for both records will increase. By repeating the process iteratively for each non

anonymous record according to a certain value of k-anonymity, the input dataset will be anonymized.

In order to formally present the method, we introduce some definitions.

Let us take an $m \times n$ data matrix, D, where each of the m rows corresponds to the record of a different respondent and each of the n columns is a textual quasi-identifier attribute. Let us name D^A the anonymized version of D. And let us define the records belonging to the original data matrix as $r_i = \{r_{i1}, \dots, r_{in}\}$ and the records of the anonymized version as $r_i^A = \{r_{i1}^A, \dots, r_{in}^A\}$, where r_{ij} and $r^A{}_{ij}$ are attribute values for each record.

- *Definition 1*. A set of indistinguishable records with respect to a given record r_i is defined as $I(r_i) = \{r_k | r_{kj} = r_{ij} \forall j = 1..n\}$. That means that two records are indistinguishable if they have exactly the same value for all of their quasi identifier attributes. Let us call $\Psi = \{I_1, \dots, I_p\}$, the set formed by sets of indistinguishable records.
- *Definition 2*. A set indistinguishable records I_l is considered anonymous (A) iff $|I_l| \geq k$ (i.e, it contains at least k elements, where k is the level of anonymity). Then, $\Lambda = \{\mathrm{A}_1, \dots, \mathrm{A}_\mathrm{q}\}$ is the group of anonymous sets of records built from the dataset D.
- *Definition 3*. The similarity between two records r_i and $r_k \in D$ is defined as the mean of the semantic similarity of each of their attribute values as follows:

$$record_{similarity(r_i, r_k)} = \frac{\sum_{j=1}^{n} sim_{sem}(r_{ij}, r_{kj})}{n} \tag{18}$$

- where for each attribute value pair, the function sim_{sem} can be any of the semantic similarity measures presented in section 3.2.1, 3.2.2 and 3.2.3. As stated before, in this paper, we choose Wu & Palmer similarity (eq. 7) for testing purposes (see section 3.3).
- *Definition 4*. Let us consider a record r_i such that $\forall A_i \in \Lambda, r_i \notin A_i$ (i.e. it is not anonymous). Then, we maximum similarity with regards to any other record available in D will represent the quality of the best data transformation for that record.

$$best_{quality(r_i)} = max\big(record\_similarity(r_i, r_k)\big) \ \forall r_k \in D \qquad (19)$$

- *Definition 5*. The minimum degree of anonymity achievable with the fusion of the values of a record r_i with respect to any other record r_k available in D is given by:

$$\min\_achievable\_anonymity(r_i) = min(|I(r_i) \cup I(r_k)|) \ \ \forall r_k \in D \qquad (20)$$

- *Definition 6*. The quality of D^A with regard to D from a semantic point of view is defined as the inverse of the information loss derived from the transformation of D in its anonymized version D^A. Information loss is usually given by the absolute difference [48], so the quality is measured in terms of semantic similarity (sim_{sem}).

$$semantic_{quality}(D^A) = \sum_{i=1}^{m}\sum_{j=1}^{n} sim_{sem}\big(r_{ij}, r_{ij}^A\big) \qquad (21)$$

This value can be normalized in the range of the sim_{sem} values by dividing it by the total number of records (m) in the set and the total number of attributes (n)

$$norm\_semantic\_quality(D^A) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} sim_{sem}\big(r_{ij}, r_{ij}^A\big)}{m * n} \qquad (22)$$

Based on a semantic similarity measure, which evaluates the quality of the best data transformation, our method aims to find the best value fusion between records that leads to a partition formed by anonymized record sets (i.e. $\forall r_i \in D \exists A_i \in \Lambda, r_i \in A_i$). The optimum anonymous partition is the one that maximizes the utility of the data, by preserving the meaning of the values. In our case, this is a partition that minimizes the information loss from a semantic point of view, which is calculated with eq. 22.

4.2 Heuristics

As noted in section 2.5, finding the optimum anonymous partition requires the generation of all the possible value fusions for all the non-anonymous records, which has an exponential cost. In order to ensure the scalability of our approach, we opted for a greedy algorithm which selects, at each iteration, a set of indistinguishable records (I_l) and finds a feasible value fusion. However, with an

uninformed approach, the quality of the result would depend on the selection of the records at each step. To solve this, an exhaustive method that tests all the combinations can be used, with a factorial cost with respect to the number of non-anonymous records. This approach is again computationally too expensive because, as records are defined by unbounded textual attributes, they usually correspond to a high number of combinations, many of them being unique, leading to a high amount of records not fulfilling *k*-anonymity. In order to ensure the scalability of the method and guide the anonymization towards a minimization of information loss, we have designed several heuristics (*H*) that permit the select, at each iteration, the best set of indistinguishable records (I_l) to transform:

- H_1) From *D*, select the group of sets of indistinguishable records $S_1 \subseteq \Psi$ whose record value tuples have the lowest number of repetitions in the original set. That is the ones with minimum $|I_i|$, which correspond to the least anonymous ones.
- H_2) From S_1, select a subset $S_2 \subseteq S_1$ that contains sets of indistinguishable records for whom the best merging of values leads to the minimum semantic information loss. The aim is to maximize the quality of the anonymized dataset of the result at each iteration. That is the $I(r_i)$ with maximum $best\_quality(r_i)$.
- H_3) From S2, select the subset $S_3 \subseteq S_2$ for which the minimum achievable degree of anonymity of their records (after the transformation) is lower. That is the $I(r_i)$ that minimize min $\_achievable\_anonymity(r_i)$. In this way, the records that are more difficult to anonymize are prioritized, as they will require more value fusions.

Those criteria are applied in the order indicated above. In this way, if the set S_1 obtained with H_1 contains more than one element, we apply H_2 to S_1. In the same way, if the resulting set S_2 obtained with H_2 has not a unique element then H_3 is applied. Through tests performed over real data, those three criteria are

enough to obtain a unique $I(r_i)$ whose values are merged with the ones of the $I(r_k)$ that allows the maximization of $best\_quality(r_i)$, increasing the k-anonymity level of both $I(r_i)$ and $I(r_k)$. However, if using those three criteria it was not possible to find a unique *I*, a random one in S_3 would be selected.

4.3 Algorithm

Algorithmically, the method works as follows:

Algorithm

```
Inputs: D (dataset), k (level of anonymity)
Output: DA (a transformation of D that fulfils the k-anonymity
level).

1    D^A := D
2    min_repetitions := min |I(r_i)| for all r_i ∈ D^A
3    while (min_repetitions < k) do
4      S_1 := set of I(r_i), r_i ∈ D^A with |I(r_i)|=min_repetitions
5      S_2 := set of I(r_i) ∈ S_1 with maximum best_quality(r_i)
6      S_3 := set of I(ri) ∈ S_2 with minimum min_achievable_anonymity(r_i)
7      Take an I(r_i) randomly from S_3
8      Find a I(r_k), r_k ∈ D^A so that r_k = argmax(record_similarity(r_i, r_k))
9      for all (r_i ∈ I(r_i)) do
10        r_ij := r_kj   ∀j=1..n
11     min_repetitions := min |I(r_i)| for all r_i ∈ DA
12   end while
13   output DA
```

As a result of the iterative process, a dataset in which all records are at least k-anonymous is obtained (i.e. $\forall r_i \in D \exists A_i \in \Lambda, r_i \in A_i$).

The algorithm works as follows. First, it is created a new data file that will be the masked version of the original one, which initially contains a copy of the input set (line #1). The sets of indistinguishable records are generated (*Definition 1*) and the minimum number of record repetitions on the set is obtained (line # 2). This number (min_repetitions) is the number of repetitions of the record with the

lowest k-anonymity level. If this value fulfils the k-anonymity level established by the user, the algorithm can stop (line #3) because according to *definition 2* all the sets or indistinguishable records are anonymous.

Otherwise, the set must be anonymized. The algorithm selects all the values the same minimum number of repetitions (line #4) and finds another record in the dataset with results in the best quality according to the *definition 4* (eq. 19). If several substitutions are equally optimum, it is selected the record whose replacement results in the minimum degree of anonymity achievable (definition 5, eq. 20) (line #6). If this record is not unique, a random record that accomplished these criteria is selected (line #7). Once the best candidate to be anonymized has been selected, we find the record that is the most similar from a semantic point of view (eq. 18) (line #9). All the occurrences in the dataset for that value are substituted (lines #9 and #10). Finally, the minimum number of record repetitions on the set is calculated on the new version of the masked dataset (line #11) and the loop is repeated again. The process finishes when no more replacements are needed, because the dataset is k-anonymous.

4.4 Cost analysis

With this method, the cost of the anonymization is $O(p^3)$, being p the number of different records in the dataset ($p \leq m$). In fact, the computationally most expensive step is the calculation of the semantic similarity between all the pairs of different records that is required in step #5 in order to find the subset with maximum *best_quality*(r_i) (eq 19). Since each record has n values, this operation requires to execute $n \cdot p^2$ times the semantic similarity between a pair of single values. In the worst case, we require p iterations to build the valid partition (loop in line #3), so the final cost of the algorithm is $n \cdot p^2 \cdot p = n \cdot p^3$ times, with n being a relative small number when compared with p, because the set of quasi-identifier attributes is usually small.

For big datasets, where p can be large due to the unbound nature of values, the scalability is more critical. For this reason we have optimized the implementation. Notice that the semantic similarity between records is measured in line #5 to calculate *best_quality(R)* and again in line #8 to find the most similar record, and repeated each iteration. As the set of different attribute values and distinct record tuples is known a priori and does not change during the masking process (unlike for generalization methods), it is possible to pre-calculate and store the similarities

between all of them. This avoids repeating the calculus of the similarity for already evaluated value pairs and, more generally, record pairs. In this manner, the calculation of the similarity measure is executed a priori only $n \cdot p^2$ times, leading to an efficiency for the most expensive function of $O(p^2)$. As it will illustrated in the evaluation section, with this modification the execution of the algorithm stays in the range of milliseconds for hundred-sized datasets.

It is important to note that the computational cost of our algorithm uniquely depends on the number of different tuples (p), unlike the related works that depend on the total size of the dataset (m), and on the depth and branching factor of the hierarchy (which represent an exponentially large generalization space of substitutions to evaluate).

5 Evaluation

We have evaluated the proposed method by applying it to a dataset consisting on answers to polls made by the "Observatori de la Fundació d'Estudis Turístics Costa Daurada" at the Catalan National Park "Delta de l'Ebre". Visitants were requested to respond several questions regarding the main reasons and preferences when visiting the park. Each record, which corresponds to an individual, includes a set of textual answers expressed by means of a noun phrase (with one or several words). Due to the variety of answers, the disclosure risk is high and, therefore, individuals are easily identifiable. So, we consider textual answers as quasi identifiers which should be anonymized.

The dataset is composed by 975 individual records, for which we considered two attributes as quasi-identifiers. They refer to the first and second reasons for visiting that natural park. The reasons are quite diverse and the visitors used brief textual expressions to answer these two questions.

Considering those two attributes, a total of 211 different responses were identified, being 118 of them unique (i.e. identifiers). Fig. 1 and Table 5 show the distribution of values for the pair of attributes according to their degree of repetition. Note that this sample represents a much wider and heterogeneous test bed than those reported in related works [12], [25], which are focused on bounded categorical values and usually have been tested with a small datasets.

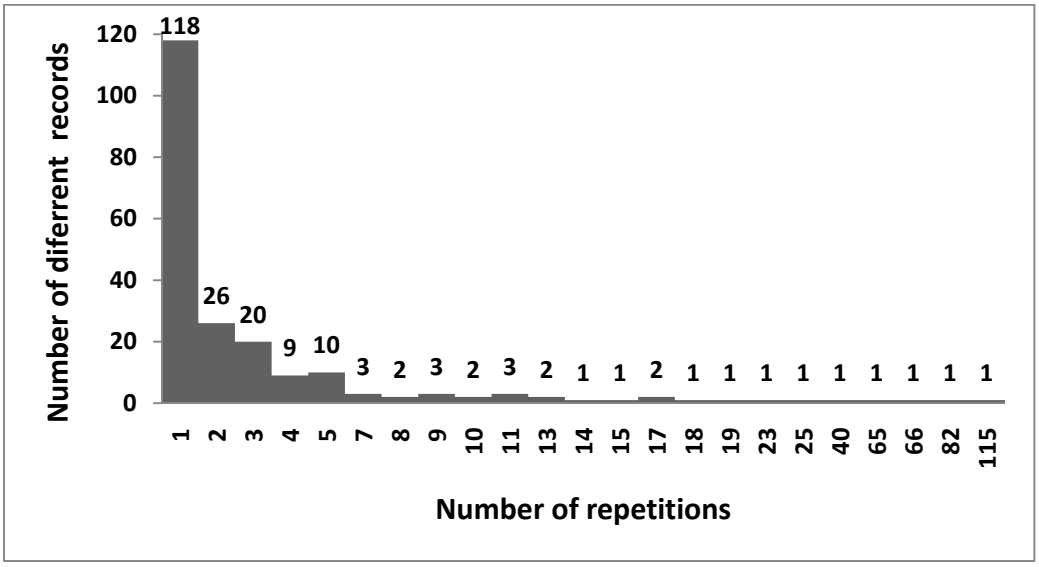


Fig. 1. Attribute distribution according to answer repetitions

Table 5. Distribution of answers in the evaluation dataset (975 registers in total).

Number of repetitions	Number of different responses	Total amount of responses
1	118	118
2	26	52
3	20	60
4	9	36
5	10	50
7	3	21
8	2	16
9	3	27
10	2	20
11	3	33
13	2	26
14	1	14
15	1	15
17	2	34
18	1	18
19	1	19
23	1	23
25	1	25
40	1	40
65	1	65
66	1	66
82	1	82
115	1	115
Total	211	975

The answers given in those two attributes are general and widely used concepts (i.e. sports, beach, nature, wildlife, relax, etc.) all of them have been found in WordNet 2.1, which permits to use this ontology for performing the semantic similarity measurement. However, as we are dealing with values represented by text labels, it was necessary to morphologically process them in order to detect different lexicalizations of the same concept (e.g. singular/plural forms). We apply the Porter Stemming Algorithm [49] to both text labels of attributes and ontological labels in order to extract the morphological root of words and to be able to map values to ontological concepts and to detect conceptually equivalent values in the dataset.

5.1 Comparing edge counting-based semantic measures

In a first study we compare the influence of using different semantic similarity functions in our algorithm, concretely in step #8, when we find similar records to a given one (eq. 21) in order to make the best substitution. We have considered three functions based on edge counting: Wu & Palmer, Leacock & Chodorow and Path Length.

To compare the quality of the masked dataset with regards to these three particular semantic similarity measures, we compared how semantically similar the replaced values are, in average, with respect to the original ones, using eq. 22. This equation has also been applied between the original and the anonymized datasets using the Wu & Palmer's similarity (Fig. 2) and Path Length distance (Fig. 3) measures.

Analyzing the figures with respect to different levels of *k*-anonymity, one can observe a linear tendency with a very smooth growth. This is very convenient and shows that our approach performs well regardless the desired level of anonymization. Regarding the different semantic similarity measures, they provide very similar and highly correlated results. This is coherent, as all of them are based on the same ontological features (i.e. absolute path length and/or the taxonomical depth) and, even though similarity values are different, the relative ranking of words is very similar. In fact, Path length and Leacock & Chorodow measures gave identical results as the later is equivalent to the former but normalized to a constant factor (i.e. the ontology depth).

As Wu & Palmer's measure incorporates more semantic features that the other measures (i.e. absolute path length normalized by relative depth in the taxonomy), it gives higher similarity scores than the other ones. See for example the increase on the similarity between the original and the masked datafiles in the range 8-10 or 12-15 in Fig. 2. The same is observed in Fig. 3 (interval 8-10). According to this results, we have taken Wu & Palmer as the best metric to measure semantic similarity during the anonymization process.

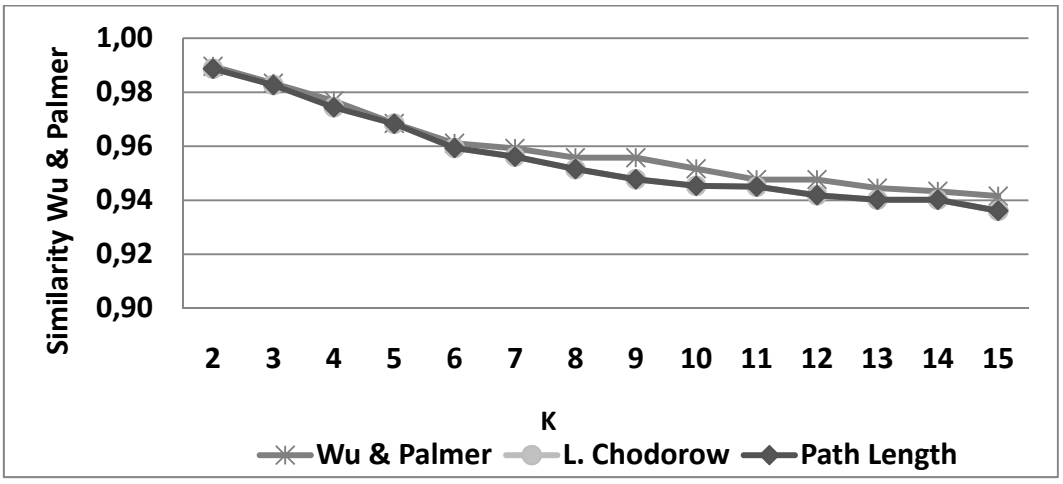


Fig. 2. Semantic similarity of the anonymized dataset

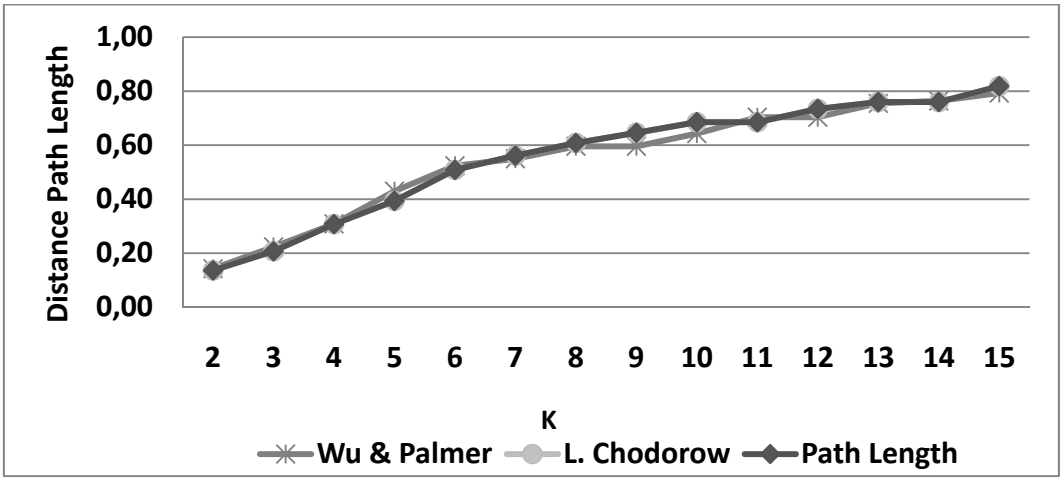


Fig. 3. Distance Path Length of the anonymized dataset

5.2 Evaluation of the heuristics

In the second study, we evaluate the contribution of each of the designed heuristics in guiding the substitution process towards minimizing the information

loss form a semantic point of view (as detailed in section 4.2). The quality of the masked dataset has been evaluated by measuring the information loss according to how semantically similar the masked values are, in average, with respect to the original ones. As before, information loss has been computed and normalized as defined in eq. 22. The test has been done with different levels of k-anonymity.

In order to show the contribution of each heuristic in minimizing the information loss of the results, we replaced the heuristic substitution by a naïve replacement that changes each sensible record by a random one from the same dataset. Following the same basic algorithm presented in section 4.3, each random change would increase the level of k-anonymity until all records are anonymized. For the random substitution, records are ordered alphabetically, in order to avoid depending on the initial order of data. The results obtained for the random substitution are the average of 5 executions. The three heuristics proposed in section 4.2, were gradually introduced instead of the random substitution, in a way that permits to quantify the contribution of each one in the results' quality. The results of this test are shown in Fig. 4: considering no heuristic at all, only the first one, only the first and the second one and all three together.

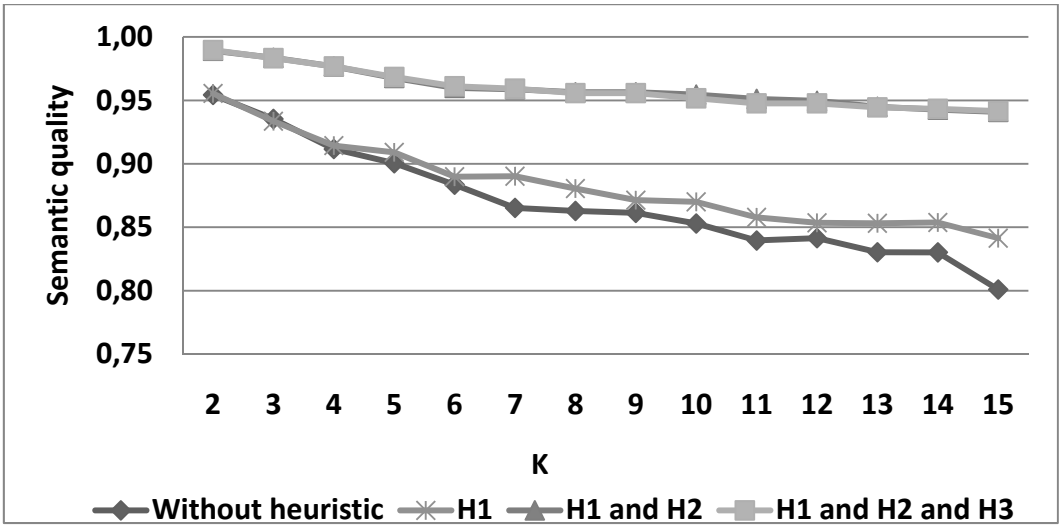


Fig. 4. Contribution of each heuristic to the anonymized dataset quality

Results reflected in Fig. 4 are coherent to what it was expected from the design of each heuristic. The first one, which only re-orders input data according to the degree of record repetition in order to prioritize the less anonymous records, produces a slight improvement over the complete random substitution. The second one, which incorporates the semantic similarity function as a metric to

guide the value fusion process towards the minimization of the semantic loss produces the most significant improvement. The incorporation of the third heuristic produces a very slight improvement in some situations, as it is only executed in case of tie (i.e. when there exists several replacements with an equal value of maximum similarity, which is a quite scarce situation).

As a result of the heuristic fusion process, our approach is able to improve the naïve replacement by a considerable margin. This is even more noticeable for a high k-anonymity level (above 5), when using the three heuristics we clearly outperform the semantic loss of the random version. This is very convenient and shows that our approach performs well regardless the desired level of privacy protection.

5.3 Comparing semantic and distributional approaches

In order to show the importance of a semantically focused anonymization, we compared it with a more traditional schema, focused on the distributional characteristics of the masked dataset (as stated at the beginning of section 2.6). This has been done by including the Discernability metric (eq. 1) in our algorithm instead of the Wu & Palmer's measure as metric, in order to guide the masking process (step #8). Both semantic and distributional approaches have been compared by evaluating the semantic difference between the original and masked dataset as stated in eq. 22. Again two measures have been used, Wu & Palmer results are shown in Fig. 5 and Path Length comparison is displayed in Fig. 6. Moreover, we have compared the masked datafile by computing the Discernability penalty with respect to the original data, as stated in eq. 1, section 2.6 (see Fig. 7).

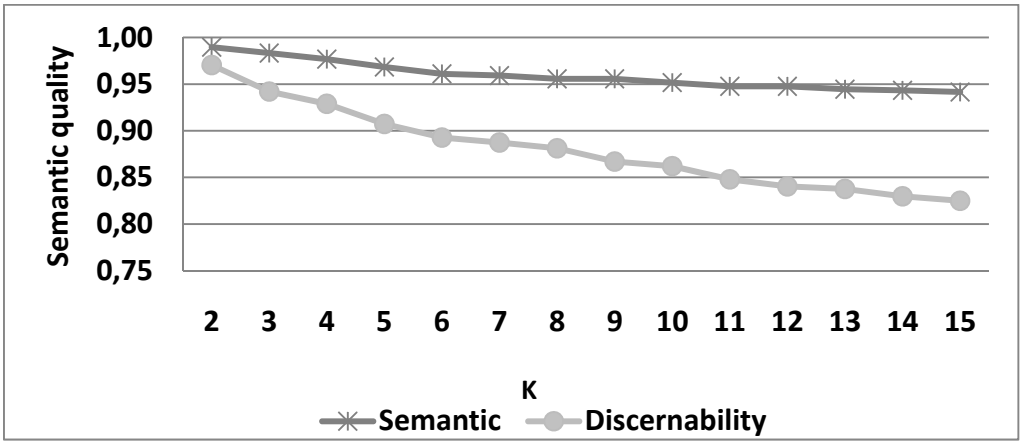


Fig. 5. Similarity against original data for semantic and distributional anonymizations.

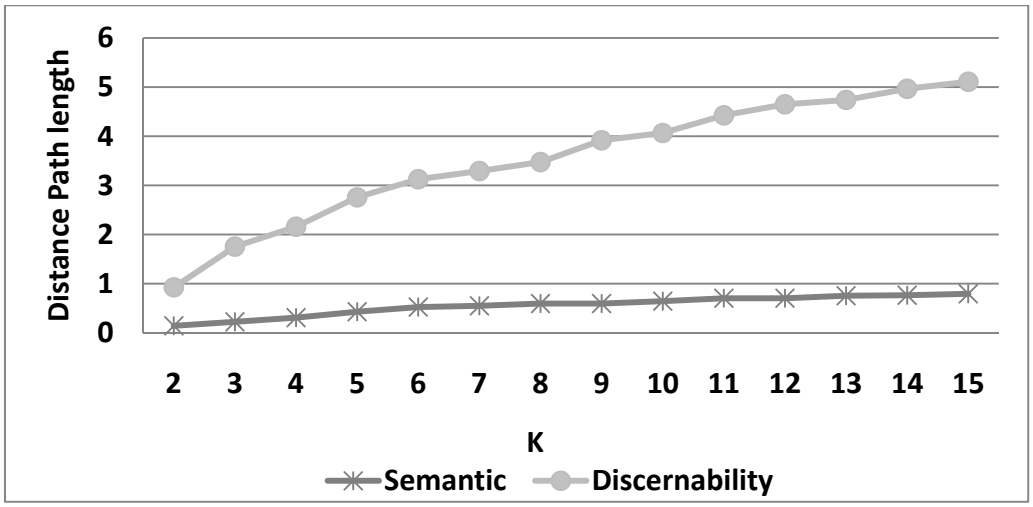


Fig. 6. Distance Path Length against original data for semantic and distributional anonymizations.

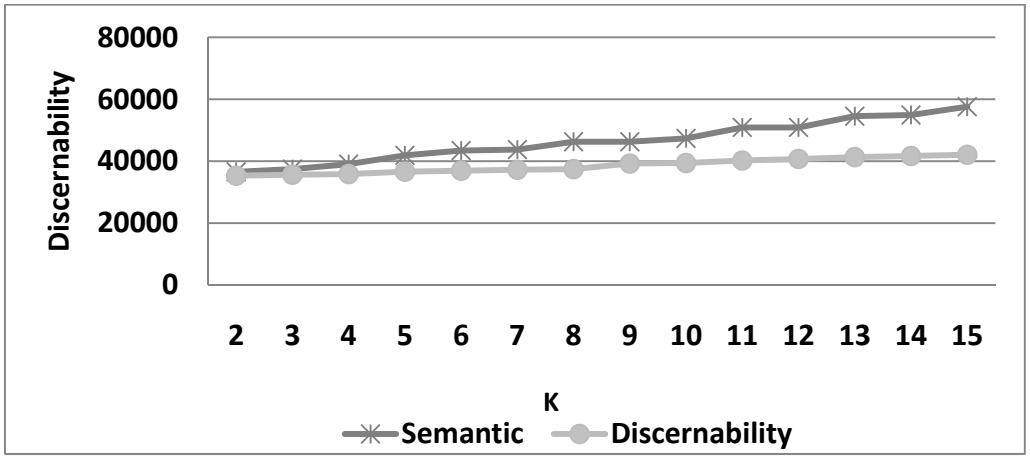


Fig. 7. Discernibilty penalty against original data for semantic and distributional anonymizations.

These figures show that the optimization of the dataset distribution and the preservation of records' semantics are not correlated. In fact, there exists a very noticeable semantic loss in the resulting dataset for k-anonymity values above 5 for the distributional approach. As stated in the introduction, the utility of textual information from the data analysis point of view is highly dependent on its semantics. One can see that classical approaches focused on providing uniform groups of masked values may significantly modify dataset's meaning, hampering their exploitation for knowledge extraction. Otherwise, they give a better discernibility score (lower values in Fig. 7), because the algorithm optimizes the criterion of using substitution values that preserve the distribution of the records as in the original file.

It is also interesting to compare our approximation that optimize the semantic similarity of the anonymized set not only with the distributional approach (Discernability metric) but also with the naïve replacement method that changes each sensible record using a random criterion.

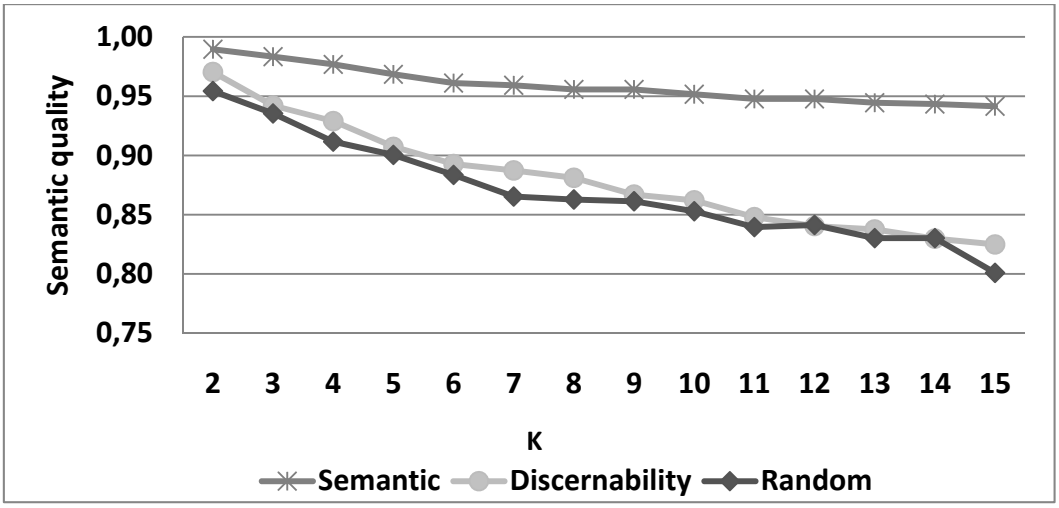


Fig. 8. Comparing the semantic quality by semantic, discernability and random approaches.

Fig. 8 shows that, from a semantic point of view, both anonymized sets with distributional and random approaches, are very similar and worse than the anonymized set with our semantic approach. So, from the point of view of the utility of the data for further analysis, the random and discernability approaches are similar because they do not preserve the semantics of the terms.

5.4 Evaluation of data utility for semantic clustering

In order to evaluate the hypothesis that a semantic-driven anonymization retains better the utility of the original data than distributional approaches from the data exploitation point of view, we next compared the utility of the dataset resulting from both approaches in a concrete data mining setting.

As stated in the introduction, data acquired by statistical agencies are of great interest for data analysis in order to, for example, extract user profiles, detect preferences or perform recommendations [1]. Data mining and, more concretely, clustering algorithms are widely used for organizing and classifying data into a number of homogenous groups. Even though clustering algorithms have been traditionally focused on numerical data or bounded categorical data, the increase in volume and importance of textual data have motivated authors in developing semantically grounded clustering algorithms [50].

In this section, we evaluate the role of ontologies in aiding the anonymization process in comparison to more simple approaches, based on ad-hoc VGHs, and other approaches without any kind of semantic background, based on optimizing data distribution. The quality of the data obtained will be studied in the context of unsupervised clustering. In particular, we will use the method presented in [51], which is hierarchical clustering algorithm that deals with both numerical and textual variables. In this method, ontologies are used as a resource to map textual features in order to semantically interpret the values of semantic features. Then, concepts' alikeness is assessed by means of semantic similarity measures. According to those similarities, an iterative aggregation process of objects is performed based on the Ward's method [52]. As a result, a hierarchical classification of non-overlapping sets of objects is constructed from the evaluation of their individual features. The height of the internal nodes in the resulting dendogram reflects the distance between each pair of aggregated elements.

By means of this algorithm, and using WordNet as the background ontology, we evaluated the utility of data from the semantic clustering point of view. We compare the clusters obtained from the original dataset against those resulting from the execution of the clustering process, both for distributional (i.e. discernibility-based) and semantic (i.e. Wu and Palmer's similarity-based) anonymization procedures.

The dataset has been masked with the method detailed in section 4 in three different configurations:

- Using WordNet 2.1 as ontology and the Wu & Palmer similarity (eq. 7) to guide the anonymization process. This will show the performance of a semantically grounded anonymization process in the preservation of data semantics.

- Using an ad-hoc VGH (see Fig. 9), constructed according to the labels in which textual attributes are expressed in the dataset instead of WordNet. The same similarity metric as above is maintained. This will potentially show the limitations introduced by the use of simple and ad-hoc VGHs (as discussed in section 2.5) with regards to the semantic interpretation of data.

- No semantics are employed. The anonymization process is guided by a metric aimed to optimize the data distribution of the masked data. The discernibility measure (eq. 1) introduced in section 2.6 is used.

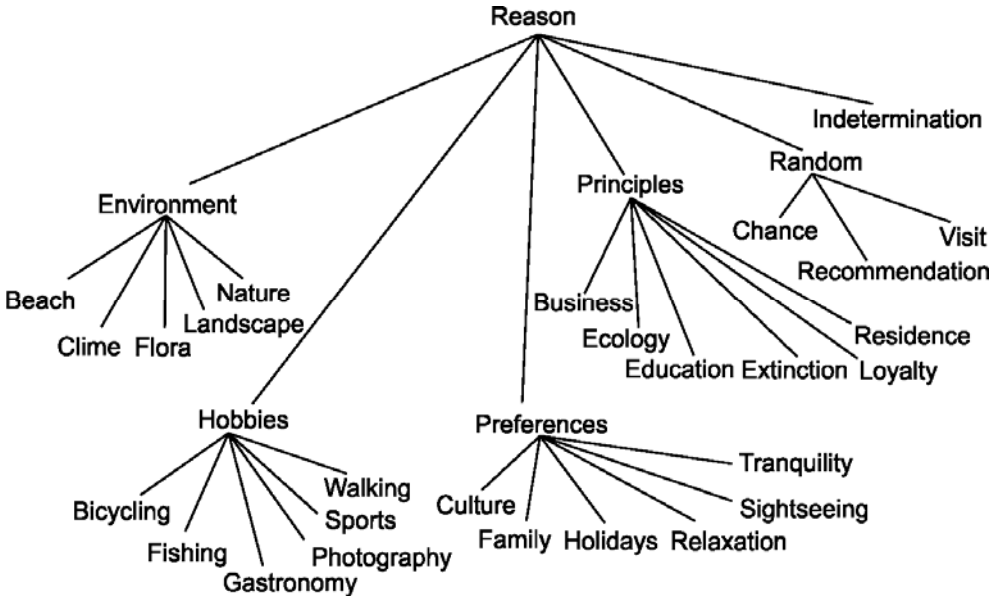


Fig. 9. VGH constructed according to textual labels of sensible attributes.

The dendograms obtained using these three approaches are presented in the following figures: Fig. 10 shows the dendogram of the original dataset; Fig. 11 is the classification obtained with the semantic-based masking method proposed,

using Wordnet 2.1; Fig. 12 has the result obtained with the distributional masking; finally, in Fig. 13 the data set has been anonymized using the VGH.

A first analysis can be done by looking at the dendograms to see the global distribution of the objects in groups. Visually comparing the dendograms one can see that the most similar dendogram to the original is the one generated from the semantic anonymized dataset (Fig. 11). The clusters are formed in a similar way (two large initial clusters of similar size that are similarly divided).

Usually these hierarchical classifications are cut at a certain level to have a partition of the objects. The cut must be done at a level that optimizes the ration between intra and inter cluster variability. That is, we want clusters with low within variability (i.e. *cohesioned*) and with high difference with respect to the rest of the clusters. Looking at an appropriate level for cutting the tree and generating a partition of the individuals, we can see that in all the dendrograms there is a clear cut in 3 clusters. However, this cut is not very useful for the manager or decision maker, because the groups are too large (remember that the dataset has 975 individuals). A finer partition should be done. In the original dataset we can see a partition in 5 clusters. In the dataset anonymized with the ontology-based approach there is also a possible cut in 5 clusters. However, in the distributional approach the cut in 5 clusters does not fulfill the variance ratio indicated, and the cut should be done in 6 clusters. In the VGH-based version, which is also based on some kind of knowledge, a partition in 5 clusters is possible, although it is not as clear as the one generated in the ontology-based clustering. These results seem to indicate that the semantically-guided versions (with Wordnet and VGH) retain better the structure of the data than the distributional one.

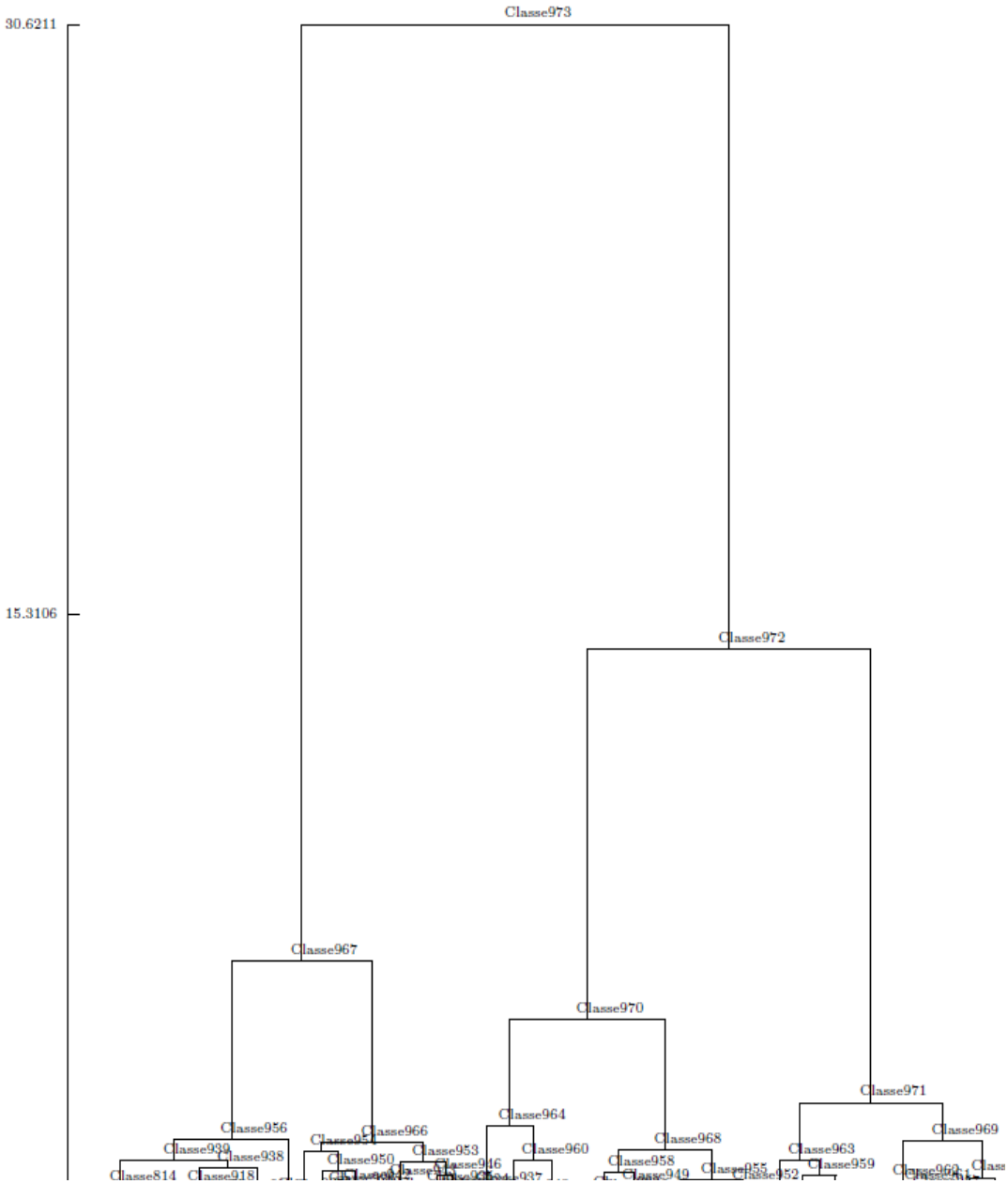


Fig. 10. Dendogram of the original set clustering.

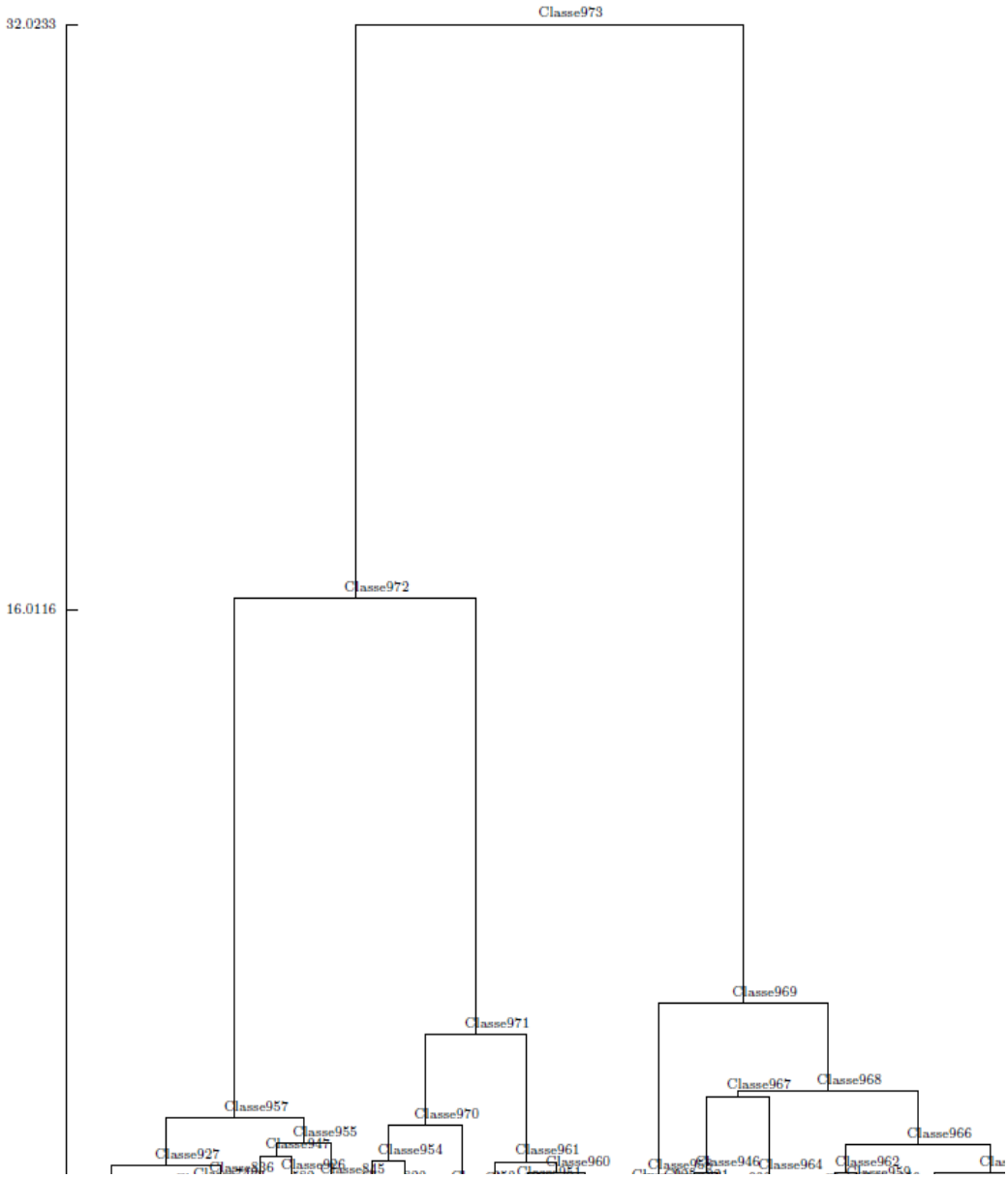


Fig. 11. Dendogram clustering of the semantic anonymized set.

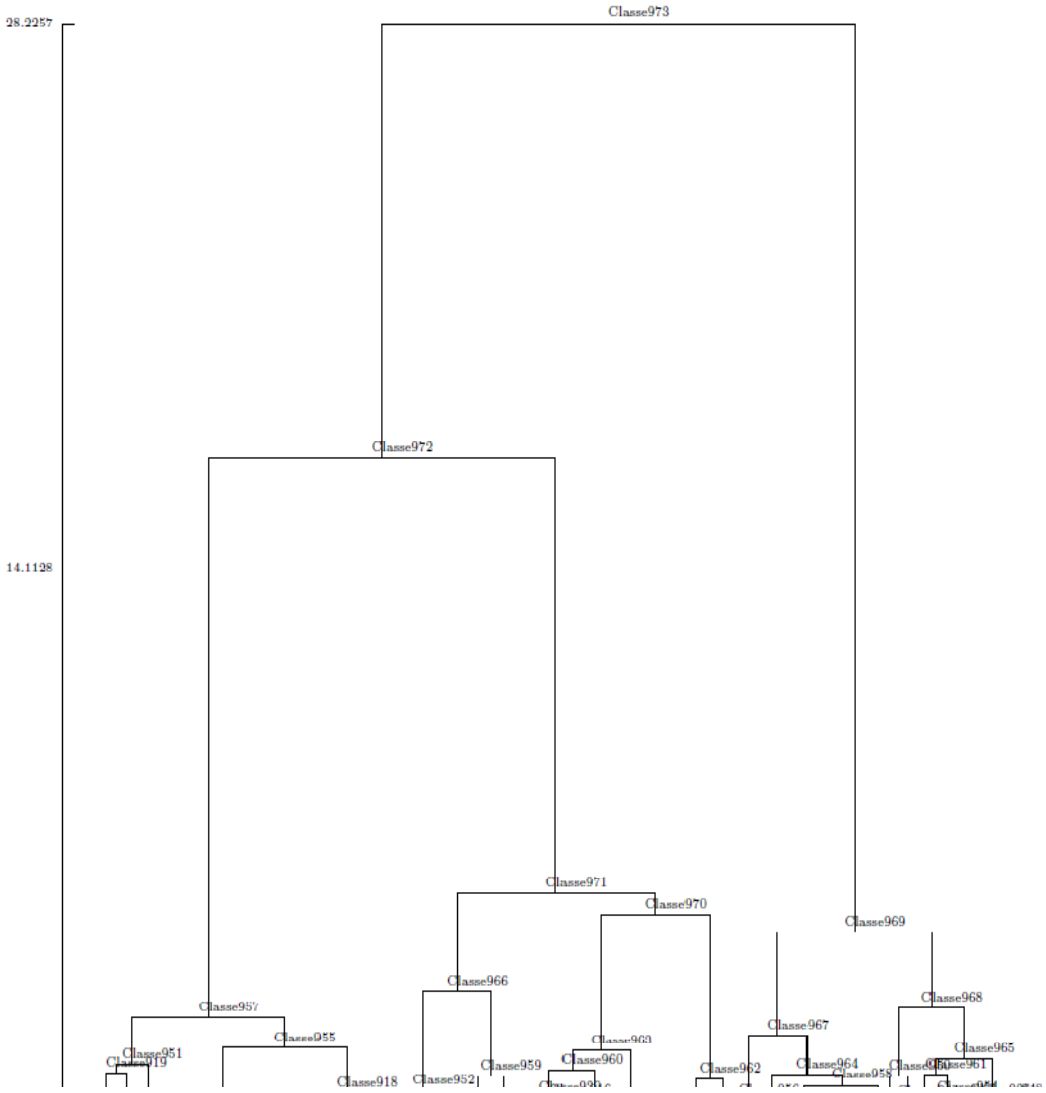


Fig. 12. Dendogram clustering of the distributional anonymized set.

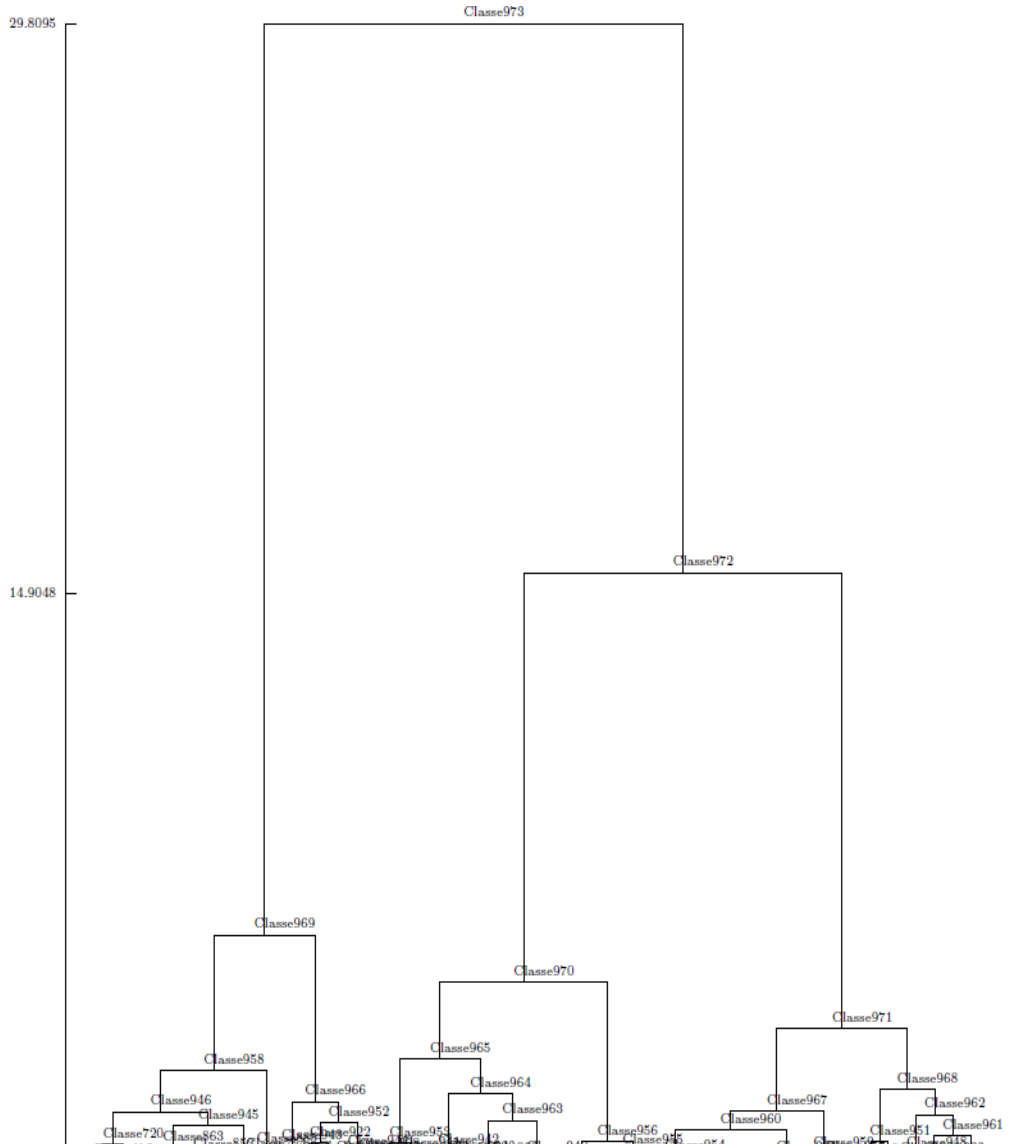


Fig. 13. Dendogram clustering of the VGH-based anonymized set

In order to make a more accurate analysis of the dendgormas at a lower level of the hierarchy, the clustering obtained in the three different configurations and the original dataset have been compared from a numerical point of view.

In order to make a more detailed analysis of the partitions that can be obtained in those trees we have quantified the differences between the clusters obtained from original data against both masking methods. Resulting clusters can be compared by means of the distance between partitions of the same set of objects as defined in [53]: considering two partitions of the same data set (in this case, the original an anonymized versions), being P_A a partition whose clusters are denoted

as A_i and P_B a partition whose clusters are denoted as B_j, the distance is defined as:

$$d_{part}\left(P_A, P_B = \frac{2 * I(P_A \cap P_B) - I(P_A) - I(P_B)}{I(P_A \cap P_B)}\right) \quad (23)$$

, where $I(P_A)$ is the average information of P_A which measures the randomness of the distribution of elements over the n classes of the partition (similarly for and $I(P_B)$), and $I(P_A \cap P_B)$ is the mutual average information of the intersection of two partitions. They are computed as

$$I(P_A) = -\sum_{i=1}^{n} P_i \log_2 P_i \quad (24)$$

$$I(P_B) = -\sum_{j=1}^{m} P_j \log_2 2P_j \quad (25)$$

$$I(P_A \cap P_B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log_2 P_{ij} \quad (26)$$

, where the probabilities of belonging to the clusters are $P_i=P(A_i)$, $P_j=P(B_j)$, and $P_{ij}=P(A_i \cap B_i)$. Distance values are normalized in the 0..1 interval, where 0 indicates identical clusters and 1 maximally different ones.

We have chosen even a finest cut level that permits to distinguish up to 14 different clusters (i.e. visitor's profiles) and keeping the optimization of the variance index. The distance between the partitions obtained from the original data and those obtained from the three masking approaches are summarized in Table 6.

Table 6. Distances between the different clustering results

Test	Distance
Original data vs. Anonymization based on WordNet	0.398
Original data vs. Anonymization based on VGH	0.515
Original data vs. Anonymization based on discernibility	0.560
Anonymization based on WordNet vs. Anonymization based on VGH	0.531
Anonymization based on WordNet vs. based on discernibility	0.589
Anonymization based on VGH vs. based on discernibility	0.623

From these results we can see how the ontology-based anonymization has given a dataset that retains better the semantics of the original data (i.e. less information loss) than the other approaches. Compared to the simpler VGH-based anonymization (0.398 vs. 0.515) we observe that even that Wordnet is a general-purpose ontology, it allows a better interpretation of input data. Due to the coarse granularity of VGHs, it is likely to suffer from high information loss. Moreover, they offer a rough and biased knowledge model compared to fine grained and widely accepted ontologies. So, VGHs, in addition to the cost of manually constructing them, offer a too simple structure which results in homogenous similarity values, making difficult a proper differentiation between terms.

Comparing the results of the ontology-based anonymization with distributional approaches, the difference is even bigger (0.398 vs. 0.56), showing that semantics play an important role in the preservation of data utility. In consequence, conclusions extracted from the analysis of ontology-based anonymized data would be more similar to those obtained from the original data when using the semantic approach presented in this paper.

It is also relevant to observe the big differences between clusters resulting from each anonymization schema, whose distance ranges from 0.531 to 0.623. This shows a high discrepancy in the way in which records are fused according to the different semantic backgrounds and quality metrics.

5.5 Record linkage

Data utility is an important dimension when aiming to anonymize data and minimize the information loss. However, from the privacy preserving point of view, disclosure risk should be also minimized. The latter may be measured as a function of the probability of re-identification of the masked dataset with respect to original data.

In order to evaluate the disclosure risk of both semantically and distributionally anonymized datasets, we have computed the level of record linkage (also named re-identification) [54] of the results. Record linkage (RL) is the task of finding matches in the original data from the anonymized results. The disclosure risk of a privacy preserving method can be measured as the difficulty of finding correct linkages between original and masked datasets. It is typically calculated as the percentage of correctly linked records [54]:

$$RL = \frac{\sum_{i=1}^{m} P_{rl}(r_i^A)}{m} \cdot 100 \quad (27)$$

, where the record linkage probability of an anonymized record $P_{rl}(r_i^A)$ is calculated as follow:

$$P_{rl}(r_i^A) = \begin{cases} 0 & if\ r_i \notin L \\ \frac{1}{|L|} & if\ r_i \in L \end{cases} \quad (28)$$

, where r_i is the original record, r_i^A is the anonymized record and L is the set of original records in D that match with $r_i^A (L \subseteq D)$. As we deal with textual features and value changes, record matching is performed by simple text matching of all individual attributes (in the same order). So, each r_i^A is compared to all records of the original dataset D by text matching, obtaining the set L of matching records. If r_i is in L, then, the probability of record linkage is computed as the probability of finding r_i in L (i.e. the number of records in L). On the contrary, if r_i is not in L, the record linkage probability is 0.

We have calculated the record linkage percentage for different levels of k-anonymity, comparing the original registers with respect to the semantic anonymization and afterwards with the distributional version of the method. The R_L probabilities are represented in Fig. 14.

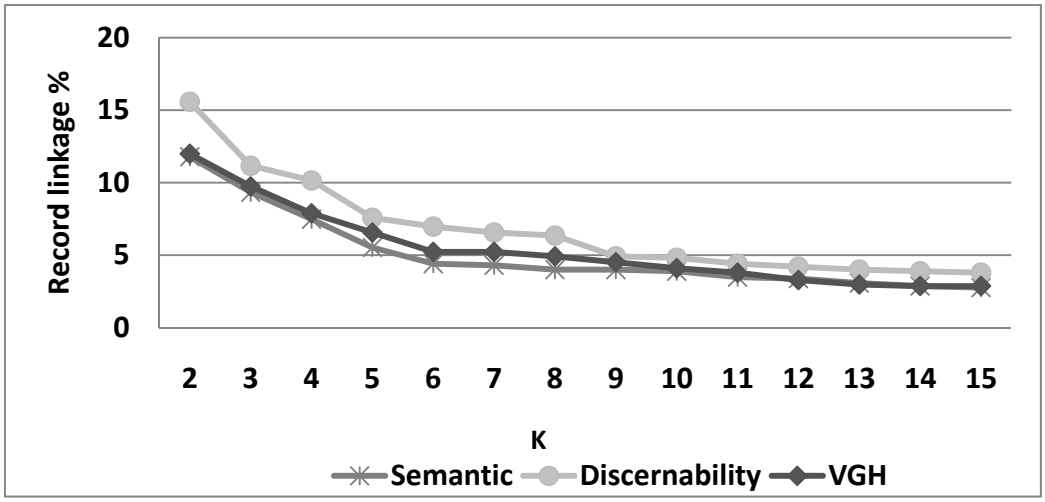


Fig. 14. Record Linkage percentage for semantic, VGH-based and dicernability-based anonymizations.

The three approaches follow a similar tendency, decreasing as k increases. It can also be seen that the degree of record linkage is quite stable for *k* values of 5 and above. The main difference is that our method gives lower probabilities of record re-identification than distributional and VGH-based approaches, especially for small values of *k*. This permits, in comparison to the distributional and VGH-based approaches, to decrease the *k*-anonymity degree (resulting in less information loss), while maintaining a comparable level of disclosure risk. If the level k is decreased, the utility of the data should increase, because the groups of indistinguishable records are smaller, keeping more variety in the dataset.

In conclusion, results show that an anonymization process focused on the preservation on data semantics does not contradicts the goal of a privacy preservation method which is to minimize the disclosure risk.

5.6 Execution time study

Finally, the last test is about the time needed for the execution of the masking process. This is an important component since usually the datafiles considered in National Statistical Offices are very large.

From a temporal perspective, executing our method over a 2.4 GHz Intel Core processor with 4 GB RAM, the run time of the anonymization process for the test dataset ranged from 1.2 to 1.6 seconds (according to the desired level of *k*-anonymity) as shown in Fig. 15. The pre-calculus of the semantic similarities between all value pairs of each attribute in the dataset lasted 6.33 minutes.

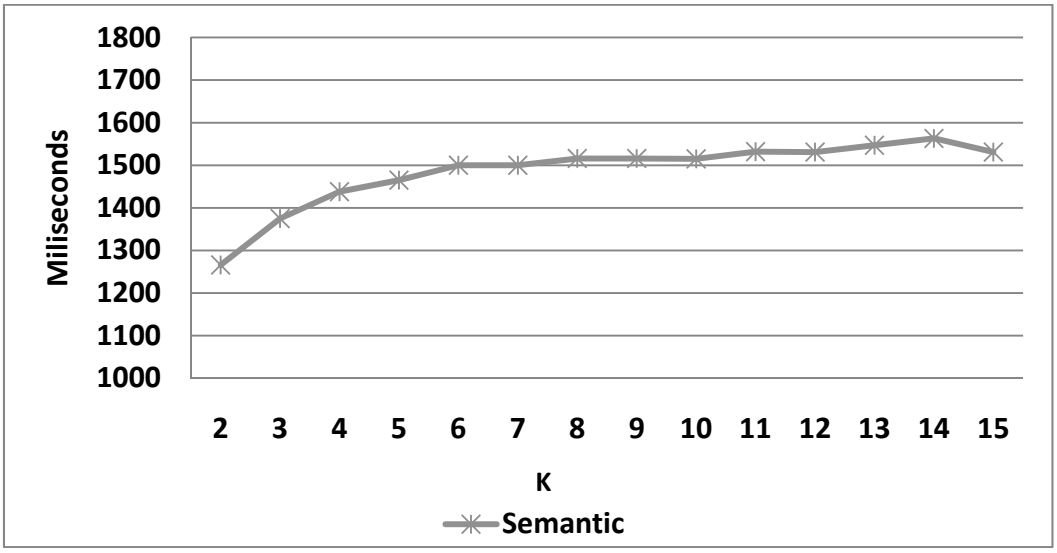


Fig. 15. Anonymization process runtime according to the level of k-anonymity

One can easily see how, as stated in section 4.4, similarity computation represents the most computationally expensive function, and how the minimization of the number of calculus results in a very noticeable optimization of runtime.

Run times are also much lower than those reported by related works that need several hours [10], [25] to perform the anonymization of the data, even for generalization schemas and very limited VGHs and bounded categorical data (3-4 levels of hierarchical depth and an average of a dozen of values [25]). On the contrary, we were able to mask much bigger and fine grained data in much less time while considering and big and wide ontologies like WordNet, with thousands of concepts and a maximum depth of 16 levels (as explained in section 3.1.1). This shows the scalability of our method for large and heterogeneous textual databases.

6 Conclusions

Anonymization of textual attributes deals with two, a priori, confronted aspects of information: on one hand, the minimization of the disclosure risk by fulfilling a desired level of k-anonymity and, on the other hand, the maximization of data utility in order to properly exploit them. Previous approaches neglected or very shallowly considered the semantic content of textual attributes. In this work we studied how Artificial Intelligence techniques can be used to improve those masking methods. As discussed in this paper, the meaning of data is an important dimension when aiming to make an analysis of the anonymized results to extract useful knowledge, as it is required in data mining, decision making or recommendation processes.

Most of previous generalization methods aggregate data by using ad-hoc hierarchical structures. Due to their limitations both from the semantic background and efficiency points of view, in this Master Thesis we have proposed an alternative way to aggregate the individually identifiable records into indistinguishable groups fulfilling k-anonymity by means of the substitution of semantically similar values.

This global masking method is based on the exploitation of wide and general ontologies in order to properly interpret the values from a conceptual point of view, rather than from a symbolic one. The algorithm uses several heuristics to guide the search on the set of possible value substitutions towards the preservation of the semantics of the dataset. This is a novel contribution of this work to the field of privacy preserving for databases.

Moreover, this non-exhaustive heuristic algorithm based on constrained value substitutions permitted to achieve a good scalability with regards to the size, heterogeneity and number of attributes of input data and with respect to the size, depth and branching factor of the ontology.

In addition to ensuring the applicability and scalability of the method when dealing with large and heterogeneous textual data, the use of ontologies avoids the need of constructing ad-hoc hierarchies according to data labels like VGH-based schemas. The construction of VGHs supposes a serious cost and limits the

applicability of the method. This drawback is solved with the use of available ontologies, such as WordNet, as it has been done in this work.

The method has been tested with real textual data obtained from visitors of a Catalan National Park. The results indicate that, in comparison with a classical approach based on the optimization of the distribution of the data, our approach better retains the quality and utility of data from a semantic point of view. This has been reflected with different evaluations, each of them, from a different point of view: comparing the quality of the anonymized datasets by the classic distributional model and our proposed semantic approach, evaluating the individual heuristic contributions and exploiting masked data with by means of a clustering process, for which we were able to obtain the most similar set of classes with respect to the original data. The goal of minimizing the disclosure risk has been studied with the evaluation the disclosure risk of the anonymized datasets, measuring the level of record linkages between the anonymized and original datasets. Finally, the computational viability of our proposal from a practical point of view has been evaluated studying also its execution time.

As a final conclusion, we can say that Artificial Intelligence techniques are useful for the adaptation of the existing methods for privacy preserving to deal with more complex data types, such as unbounded textual attributes.

7 Future work

In the future we would like to study the behavior of the method with respect to other ontologies, with different size and concreteness degrees (such as domain-specific ontologies, which could be exploited when input data refers to concrete domain terminology). We would also study the possibility of combining several ontologies as background knowledge in order to complement the knowledge modeled in each of them.

We also plan to study other anonymization methods for textual attributes from a semantic point of view. Micro-aggregation aimed to create fixed sized k-indistinguishable sets can be tackled from a semantic point of view by considering clustering techniques introduced in section 5.4. As a result of a semantically grounded semantic clustering, a sampling anonymization method can be developed by substituting a partition for a semantically similar representative record. Data swapping method can be also adapted in order to incorporate semantic knowledge in the anonymizing process by searching the most semantically similar record/value pairs to swap.

As the anonymization quality directly depends on the assessment of semantic similarity between words, we also plan to study others more complex semantic measures reported the literature. Additional semantic knowledge provided by WordNet can be also considered, exploiting other semantic relations (such as meronyms or coordinate terms) or glosses (i.e. related words extracted definitions and/or examples sentences). In this manner the space of value substitutions can be expanded.

We also believe that semantic similarity theory can be exploited in order to contribute the in area of evaluation of the disclosure control. Concretely, we plan to develop new methods of record linkage which are able to evaluate anonymized and original records from a semantic point of view by exploiting similar principles as those presented in this work.

8 References

1. Domingo-Ferrer, J. A survey of inference control methods for privacy-preserving data mining, in Privacy-Preserving Data Mining: Models and Algorithms, eds. C.C. Aggarwal and P.S. Yu, Advances in Database Systems, v.34, N.Y.: Springer Verlag, pp. 53--80 (2008)
2. Computational aspects of statistical confidentiality, 2009. European project IST-2000-25069 ESSproject, 8th FP, 2001-2009, http://neon.vb.cbs.nl/casc/..\casc\index.htm
3. Guarino, N. Formal Ontology in Information Systems. In Guarino N (ed) 1st Int. Conf. on Formal Ontology in Information Systems, pp. 3--15. IOS Press. Trento, Italy (1998)
4. Computational aspects of statistical confidentiality, 2009. European project IST-2000-25069 ESSproject, 8th FP, 2001-2009, http://neon.vb.cbs.nl/casc/..\casc\index.htm
5. HIPAA. Health insurance portability and accountability 2010. http://www.hhs.gov/ocr/hipaa/
6. European privacy regulations. http://ec.europa.eu/justice_home/fsj/privacy/index_en.htm
7. F. C. Statistical-Methodology. Report on statistical disclosure limitation methodology. Technical report, Statistical and Science Policy, Office of Information and Regulation Affairs, Office of Management and Budget (2005)
8. Agrawal, C., Yu, P.S (Eds) Privacy-preserving Data Mining: models and algorithms, Springer (2008).
9. Willenborg, L. and DeEaal T. Elements of Statistical Disclosure Control. Springer-Verlag, New York (2001)
10. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Wai-Chee Fu, A. Utility-based anonymization using local recoding, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, pp.785--790 (2006)
11. Sweeney, L. *k*-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) pp. 557--570 (2002)
12. Samarati, P., Sweeney, L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)
13. Domingo-Ferrer, J. Microaggregation for database and location privacy. In Next Generation Information Technologies and Systems, pp. 106--116 (2006)
14. T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. Manuscript (2005)
15. Gouweleeuw, J.M. Kooiman, P., Willenborg, L. C. R. J. and DeWolf, P. P. Post randomization for statistal disclousure control: Theory and implementation. Research paper no. 9731 (Voorburg: Statistics Netherlands) (1997)
16. A. C. Singh, F. Yu, and G. H. Dunteman. MASSC: A new data mask for limiting statistical information loss and disclosure. In H. Linden, J. Riecan, and L. Belsby, editors, Work Session on Statistical Data Confidentiality 2003, Monographs in Official Statistics, pages 373--394, Luxemburg, 2004. Eurostat.
17. Reiss, S. P. Practical data-swapping: the first steps. ACM Transactions on Database Systems, 9 pp. 20--37 (1984)

18. Domingo-Ferrer, J. and Torra, V. A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111--134, Amsterdam North-Holland. http://vneumann.etse.urv.es/publications/bcpi. (2001)
19. DeWaal, A. G. and Willenborg, L. C. R. J. Global recodings and local suppressions in microdata sets. In Proceedings of Statistics Canada Symposium'95, pp. 121--132, Ottawa. (1995)
20. Guo, L., Wu, X. Privacy preserving categorical data analysis with unknown distortion parameters, Transactions on Data Privacy, 2, (2009) 185-205.
21. T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. Manuscript (2005)
22. Bayardo, R. J., Agrawal, R. Data privacy through optimal *k*-anonymization. Proceedings of the 21st International Conference on Data Engineering (ICDE) pp. 217--228 (2005)
23. LeFevre, K., DeWitt, D. J. and Ramakrishan, R. Incognito: Efficient full-domain k-anomymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, IEEE (2006)
24. Iyengar, V. S. Transforming data to satisfy privacy constraints. Proceedings of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Minig (KDD), pp. 279--288 (2002)
25. Li, T., Li, N. Towards optimal *k*-anonymization. Data & Knowledge Engineering 65 pp. 22--39 (2008)
26. He, Y., Naughton, J.F. Anonymization of Set-Valued Data via Top-Down, Local Generalization. 35th Int. Conf. VLDB. Lyon, France (2009) Vol. 2. 934-945.
27. Terrovitis, M., Mamoulis, N. Kalnis, P. Privacy-preserving anonymization of set-valued data. In Proc. of VLDB (2008)
28. Fellbaum, C. WordNet: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press. (1998)
29. Eric G. Little, Galina L. Rogova. Designing ontologies for higher level fusion Information Fusion Volume 10, Issue 1, January (2009) 70-82
30. Mieczyslaw M. Kokar, Christopher J. Matheus, Kenneth Baclawski. Ontology-based situation awareness Information Fusion Volume 10, Issue 1, January (2009) 83-98
31. Cimiano, P. Ontology Learning and Population from Text. Algorithms, Evaluation and Applications. Springer-Verlag (2006)
32. L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi and J. Sachs, Swoogle: A Search and Metadata Engine for the Semantic Web, in: Proceedings of the thirteenth ACM international Conference on Information and Knowledge Management (CIKM04) (ACM Press, Washington, D.C., USA, 2004) 652-659.
33. Rada, R., Mili, H., Bichnell, E. Blettner, M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern pp. 17--30 (1989)
34. Leacock, C., Chodorow, M. Combining local context and WordNet similarity for word sense identification. In Fellbaum (ed.), WordNet: An electronic lexical database, pp. 265--283. MIT Press (1998)
35. Wu, Z., Palmer, M. Verb semantics and lexical selection. In Proceedings. 32nd annual Meeting of the Association for Computational Linguistics, pp. 133--138. New Mexico, USA (1994)

36. Pirró, Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content .LNCS 5332 (2009)
37. Bollegala, D., Matsuo, Y. and Ishizuka, M. WebSim: A Web-based Semantic Similarity Measure, in: The 21st Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2007) (Miyazaki, Japan, 2007) 757-766.
38. Jiang, J., Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In: proceedings of the International Conference on Research in Computational Linguistics (ROCLING X) Taiwan. Pp. 19--33 (1997)
39. Tversky, A. Features of similarity, Psychological Review, 84 (1977) 327-352
40. Rodriguez, M.A. and Egenhofer, M.J. Determining semantic similarity among entity classes from different ontologies, IEEE Transactions on Knowledge and Data Engineering, 15(2) (2003) 442–456
41. Petrakis, G., Varelas, G., Hliaoutakis, A. and Raftopoulou, R. X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies, Journal of Digital Information Management (JDIM), 4 (2006) 233-237
42. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Conference on Research in Computacional Linguistics (ROCLING X). pp. 448—453. Taiwan (1995)
43. Lin, D. An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning (ICML98). Madison. Wisconsin, USA pp. 296--304 (1998)
44. Patwardan, S., Pedersen, T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In Proceedings of EACL, Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together. Trento Italy pp. 1--8 (2006)
45. Pedersen, T., Patwardhan, S. and Michelizzi, J. WordNet. WordNet::Similarity – Measuring the Relatedness of Concepts. Hppt://search.cpan.org/dist/WordNet-Similarity. American Association for Artificial Intelligence (2004)
46. Rubenstein, H. and Goodenough, J. Contextual correlates of synonymy, Communications of the ACM, 8(10) (1965) 627-633
47. G. A. Miller and W. G. Charles, Contextual correlates of semantic similarity, Language and Cognitive Processes, 6(1) (1991) 1-28.
48. Domingo-Ferrer, J., Torra, V., Disclosure control methods and information loss for microdata, in: Confidentiality, disclosure and data access: Theory and practical applications for statistical agencies, Elsevier, p.91-110 (2001)
49. Porter. An algorithm for suffix stripping, Program, (1980) Vol. 14 no 3, 130-137.
50. Zengyou He, Xiaofei Xu, Shengchun Deng. k-ANMI: A mutual information based clustering algorithm for categorical data Information Fusion Volume 9, Issue 2, April (2008), Pages 223-233.
51. M. Batet, A. Valls, K. Gibert, Improving classical clustering with ontologies, in: Proceedings of the 4th World conference of the IASC, Japan, (2008) 137-146.
52. Ward, J.H. Hierarchical grouping to optimize an objective function, JASA, (1963) 58: 236-244.
53. López de Mántaras, R.: A distance-based Attribute Selection Measure for decision tree induction. Machine learning, 6 (1991) 81-92.
54. Torra, V., Domingo-Ferrer, J. Record Linkage methods for multidatabase data mining, in: Information Fusion in Data Mining. Springer (2003) 101-132.

Appendix A

Appendix A consists on a summary of four research papers that are the result of the work done in the Master Thesis. The summary includes: title, authors, abstract, conference or journal, dates and state of the papers. Three of them have been submitted and we are waiting for the answer, another one has been accepted in an international conference.

Title	Anonymizing Categorical Data with a Recoding Method based on Semantic Similarity
Authors	Sergio Martínez, Aida Valls, David Sánchez
Abstract	With the enormous growth of the Information Society and the necessity to enable access and exploitation of large amounts of data, the preservation of its confidentiality has become a crucial issue. Many methods have been developed to ensure the privacy of numerical data but very few of them deal with textual (categorical) information. In this paper a new method for protecting the individual's privacy for categorical attributes is proposed. It is a masking method based on the recoding of words that can be linked to less than *k* individuals. This assures the fulfillment of the *k*-anonymity property, in order to prevent the re-identification of individuals. On the contrary to related works, which lack a proper semantic interpretation of text, the recoding exploits an input ontology in order to estimate the semantic similarity between words and minimize the information loss.
Sent to	International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems - IPMU 2010 Dortmund - Germany
Type	Conference
Dates	Submission of papers: 27.01.2010 Notification of acceptance: 15.03.2010 Submission of final versions: 15.04.2010 Conference: 28.06. - 02.07.2010
State	Accepted

Title	Ontology-based anonymization of categorical values
Authors	Sergio Martínez, Aida Valls, David Sánchez
Abstract	The analysis of sensible data requires a proper anonymization of values in order to preserve the privacy of individuals. Information loss should be minimized during the masking process in order to enable a proper exploitation of data. Even though several masking methods have been designed for numerical data, very few of them deal with categorical (textual) information. In this case, the quality of the anonymized dataset is closely related to the preservation of semantics, a dimension which is commonly neglected of shallowly considered in related words. In this paper, a new masking method for unbounded categorical attributes is proposed. It relies on the knowledge modeled in ontologies in order to semantically interpret the input data and perform data transformations aiming to minimize the loss of semantic content. On the contrary to exhaustive methods based on simple hierarchical structures, our approach relies on a set of heuristics in order to guide and optimize the masking process, ensuring its scalability when dealing with big and heterogenous datasets and wide ontologies. The evaluation performed over real textual data suggests that our method is able to produce anonymized datasets which significantly preserve data semantics in comparison to apporaches based on data distribution metrics.
Sent to	The 7th International Conference on Modeling Decisions for Artificial Inteligence – MDAI 2010 Perpignan - France
Type	Conference
Dates	Submission of papers: 26.03.2010 Notification of acceptance: 10.06.2010 Submission of final versions: 25.06.2010 Conference: 27.10. - 29.10.2010
State	Under review

Title	Privacy protection of textual attributes through a semantic-based masking method
Authors	Sergio Martínez, David Sánchez, Aida Valls, Montserrat Batet
Abstract	Exploitation of microdata provided by statistical agencies can bring many benefits from the point of view of data mining. However, this data often refers to sensible information which can be directly or indirectly associated to individuals. A proper anonymization process is required to minimize the disclosure risk. Several masking methods have been developed for dealing with numerical data or bounded categorical values, but approaches tackling the anonymization of textual values are scarce and shallow. Due to the importance of textual data in Information Society, in this paper we present a new masking method aimed to anonymize unbounded textual values, based on the fusion of records with similar values to form groups of indistinguishable individuals. As the utility of textual information from the data exploitation point of view is closely related to the preservation of its meaning, our method relies on the structured knowledge representation given by ontologies. This domain knowledge is used to guide the masking process towards the merging that best preserves the semantics of the original data. Since textual data typically consist on large and heterogeneous value sets, our method focuses on providing a computationally efficient algorithm by relying on several heuristics instead of exhaustive searches. The method is evaluated with real data in a concrete data mining application consisting in solving a clustering problem. The method is also compared against more classical approaches, focused on the optimization of the value distribution of the dataset. Results show that a semantically-grounded anonymization preserves better the utility of data, both in theoretical and practical settings, offering a low the probability of record linkage. At the same time, it achieves a good scalability with regards to the size of input data.
Sent to	Information Fusion An International Journal on Multi-Sensor, Multi-Source Information Fusion Special issues on "Information fusion in the context of data privacy"
Type	Journal (Elsevier) – ISI JCR – Impact factor (2008): 2.057
Dates	Submission of papers: 30.04.2010
State	Under review

Title	The role of ontologies in the anonymization of textual variables
Authors	Sergio Martínez, David Sánchez, Aida Valls, Montserrat Batet
Abstract	The exploitation of sensible data associated to individuals requires a proper anonymization in order to preserve the privacy. Even though several masking methods have been designed for numerical data, very few of them deal with textual information. During the masking process, information loss should be minimized in order to enable a proper analysis of data with data mining methods. In the case of textual data, the quality of the anonymized dataset is closely related to the preservation of semantics, a dimension which has been only shallowly considered in some previous works, by using small and ad-hoc hierarchies of words. In this work we want to study the use of large and standard ontologies as the base to perform the anonymization of textual variables. We will evaluate the role of ontologies in preserving the utility of the anonymized information when a partition of the objects is done with unsupervised clustering methods. Results show that by exploiting detailed ontologies, one is able to improve the preservation of the data semantics in comparison to approaches based on ad-hoc structures and data distribution metrics.
Sent to	Tretzè Congrés International de l'Associació Catalana d'Intel·ligència Artificial – CCIA 2010 L'Espluga de Francolí - Tarragona
Type	Conference
Dates	Submission of papers: 24.05.2010 Notification of acceptance: 23.06.2010 Submission of final versions: 15.07.2010 Conference: 20.10. - 22.10.2010
State	Under review

Anonimizing Categorical Data with a Recoding Method based on Semantic Similarity

Sergio Martínez, Aida Valls, David Sánchez

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Avda. Països Catalans, 26, 43007 Tarragona, Spain
{sergio.martinezl, aida.valls, david.sanchez}@urv.cat

Abstract. With the enormous growth of the Information Society and the necessity to enable access and exploitation of large amounts of data, the preservation of its confidentiality has become a crucial issue. Many methods have been developed to ensure the privacy of numerical data but very few of them deal with textual (categorical) information. In this paper a new method for protecting the individual's privacy for categorical attributes is proposed. It is a masking method based on the recoding of words that can be linked to less than *k* individuals. This assures the fulfillment of the *k*-anonymity property, in order to prevent the re-identification of individuals. On the contrary to related works, which lack a proper semantic interpretation of text, the recoding exploits an input ontology in order to estimate the semantic similarity between words and minimize the information loss.

Keywords: Ontologies, Data analysis, Privacy-preserving data-mining, Anonymity, Semantic similarity.

1. Introduction

Any survey's respondent (i.e. a person, business or other organization) must be guaranteed that the individual information provided will be kept confidential. Statistical Disclosure Control discipline aims at protecting statistical data in a way that it can be released and exploited without publishing any private information that could be linked with or identify a concrete individual. In particular, in this paper we focus on the protection of microdata, which consists on values obtained from a set of respondents of a survey without applying any summarization technique (e.g. publishing tabular data or aggregated information from multiple queries) [1].

Since data collected from statistical agencies is mainly numerical, several different anonymization methods have been developed for masking numerical values in order to prevent the re-identification of individuals [1]. Textual data has been traditionally less exploited, due to the difficulties of handling non-numerical values with inherent semantics. In order to simplify its processing and anonymization, categorical values are commonly restricted to a predefined vocabulary (i.e. a bounded set of modalities). This is a serious drawback because the list of values is fixed in advance and, consequently, it tends to homogenise the sample. Moreover, the masking methods for categorical data do not usually consider the semantics of the terms (see section 2).

Very few approaches have considered semantics in some degree. However, they require the definition of ad-hoc structures and/or total orderings of data before anonymizing them. As a result, those approaches cannot process unbounded categorical data. This compromises their scalability and applicability. Approximate reasoning techniques may provide interesting insights that could be applied to improve those solutions [2]. As far as we know, the use of methods specially designed to deal with uncertainty has not been studied in this discipline until now.

In this work, we extend previous methods by dealing with unbounded categorical variables which can take values from a free list of linguistic terms (i.e. potentially the complete language vocabulary). That is, the user is allowed to write the answer to a specific question of the survey using any noun phrase. Some examples of this type of attributes can be "Main hobby" or "Most preferred type of food".

Unbounded categorical variables provide a new way of obtaining information from individuals, which has not been exploited due to the lack of proper anonymization tools. Allowing a free answer, we are able to obtain more precise knowledge of the individual characteristics, which may be interesting for the study that is being conducted. However, at the same time, the privacy of the individuals is more critical, as the disclosure risk increases due to the uniqueness of the answers.

In this paper, an anonymization technique for this kind of variables is proposed. The method is based on the replacement or recoding of the values that may lead to the individual re-identification. This method is applied locally to a single attribute. Attributes are usually classified as *identifiers* (that unambiguously identify the individual), *quasi-identifiers* (that may identify some of the respondents, especially if they are combined with the information provided by other attributes), *confidential* outcome attributes (that contain sensitive information) and *non-confidential* outcome attributes (the rest). The method proposed is suitable for quasi-identifier attributes.

In unbounded categorical variables, textual values refer to concepts that can be semantically interpreted with the help of additional knowledge. Thus, terms can be interpreted and compared from a semantic point of view, establishing different degrees of similarity between them according to their meaning (e.g. for hobbies, *treking* is more similar to *jogging* than to *dancing*). The estimation of semantic similarity between words is the basis of our recoding anonymization method, aiming to produce higher-quality datasets and to minimize information loss.

The computation of the semantic similarity between terms is an active trend in computational linguistics. That similarity must be calculated using some kind of domain knowledge. Taxonomies and, more generally ontologies [3], which provide a graph model where semantic relations are explicitly modelled as links between concepts, are typically exploited for that purpose (see section 3). In this paper we focus on similarity measures based on the exploitation of the taxonomic relations of ontologies.

The rest of the paper is organized as follows. Section 2 reviews methods for privacy protection of categorical data. Section 3 introduces some similarity measures based on the exploitation of ontologies. In section 4, the proposed anonymization method is detailed. Section 5 is devoted to evaluate our method by applying it to real data obtained from a survey at the National Park "*Delta del Ebre*" in Catalonia, Spain. The final section contains the conclusions and future work.

2. Related work

Categorical data is composed by a set of registers (i.e. records), each one corresponding to one individual, and a set of textual attributes, classified as indicated before (identifiers, quasi-identifiers, confidential and non-confidential). The anonymization or masking methods of categorical values are divided in two categories depending on their effect on the original data [4]:

- *Perturbative*: data is distorted before publication. They are mainly based on data swapping (exchanging the values of two different records) or the addition of some kind of noise, such as the replacement of values according to some probability distribution (PRAM) [5], [6] and [7].
- *Non-perturbative*: data values are not altered but generalized or eliminated [8], [4]. The goal is to reduce the detail given by the original data. This can be achieved with the local suppression of certain values or with the publication of a sample of the original data which preserves the anonymity. Recoding by generalization is also another approach, where several categories are combined to form a new and less specific value.

Anonymization methods must mask data in a way that disclosure risk is ensured at an enough level while minimising the loss of accuracy of the data, i.e. the information loss. A common way to achieve a certain level of privacy is to fulfil the *k-anonimity* property [9], so that each single value cannot be linked to less than *k* registers. On the other hand, low information loss guarantees that useful analysis can be done on the masked data.

With respect to recoding methods, some of them rely on hierarchies of terms covering the categorical values observed in the sample, in order to replace a value by another more general one.

Samariti and Sweeney [10] and Sweeney [9] employed a generalization scheme named Value Generalization Hierarchy (VGH). In a VGH, the leaf nodes of the hierarchy are the values of the sample and the parent nodes correspond to terms that generalize them. In this scheme, the generalization is performed at a fixed level of the hierarchy. The number of possible generalizations is the number of levels of the tree. Iyengar [11] presented a more flexible scheme which also uses a VGH, but a value can be generalized to different levels of the hierarchy; this scheme allows a much larger space of possible generalizations. Bayardo and Agrawal [12] proposed a scheme which does not require a VGH. In this scheme a total order is defined over all values of an attribute and partitions of these values are created to make generalizations. The problem is that defining a total order for categorical attributes is not straightforward.

T.Li and N. Li [13] propose three generalization schemes: Set Partitioning Scheme (SPS), in which generalizations do not require a predefined total order or a VGH; each partition of the attribute domain can be a generalization. Guided Set Partitioning Scheme (GSPS) uses a VGH to restrict the partitions that are generated. Finally, the Guided Oriented Partition Scheme (GOPS) includes also ordering restrictions among the values.

The main problem of the presented approaches is that either the hierarchies or the total orders are build ad-hoc for the corresponding data value set (i.e. categorical values directly correspond to leafs in the hierarchy), hampering the scalability of the

method when dealing with unbounded categorical values. Moreover, as hierarchies only include the categorical data values observed in the sample, the resulting structure is very simple and a lot of semantics needed to properly understand the word's meaning is missing. As a result, the processing of categorical data from a semantic point of view is very limited. This is especially critical in non-hierarchy-based methods, which do not rely on any kind of domain knowledge and, in consequence, due to their completely lack of word understanding, they have to deal with categorical data from the point of view of Boolean word matching.

3. Ontology-based semantic similarity

In general, the assessment of concept's similarity is based on the estimation of semantic evidence observed in a knowledge resource. So, background knowledge is needed in order to measure the degree of similarity between concepts.

In the literature, we can distinguish several different approaches to compute semantic similarity according to the techniques employed and the knowledge exploited to perform the assessment.

The most classical approaches exploit structured representations of knowledge as the base to compute similarities. Typically, subsumption hierarchies, which are a very common way to structure knowledge [3], have been used for that purpose. The evolution of those basic semantic models has given the origin to ontologies. Ontologies offer a formal, explicit specification of a shared conceptualization in a machine-readable language, using a common terminology and making explicit taxonomic and non-taxonomical relationships [14]. Nowadays, there exists massive and general purpose ontologies like WordNet [15], which offer a lexicon and semantic linkage between the major part of English terms (it contains more than 150,000 concepts organized into is-a hierarchies). In addition, with the development of the Semantic Web, many domain ontologies have been developed and are available through the Web [16].

From the similarity point of view, taxonomies and, more generally, ontologies, provide a graph model in which semantic interrelations are modeled as links between concepts. Many approaches have been developed to exploit this geometrical model, computing concept similarity as inter-link distance.

In an is-a hierarchy, the simplest way to estimate the distance between two concepts c_1 and c_2 is by calculating the shortest *Path Length* (i.e. the minimum number of links) connecting these concepts (1) [17].

$$dis_{pL}(c_1,c_2) = \min \#\ of\ is-a\ edges\ connecting\ c_1\ and\ c_2 \qquad (1)$$

Several variations of this measure have been developed such as the one presented by Wu and Palmer [18]. Considering that the similarity between a pair of concepts in an upper level of the taxonomy should be less than the similarity between a pair in a lower level, they propose a path-based measure that also takes into account the depth of the concepts in the hierarchy (2).

$$sim_{w\&p}(c_1, c_2) \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (2)$$

, where N_1 and N_2 are the number of is-a links from c_1 and c_2 respectively to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

Leacock and Chodorow [19] also proposed a measure that considers both the shortest path between two concepts (in fact, the number of nodes N_p from c_1 to c_2) and the depth D of the taxonomy in which they occur (3).

$$sim_{l\&c}(c_1, c_2) = -\log(N_p / 2D) \quad (3)$$

There exist other approaches which also exploit domain corpora to complement the knowledge available in the ontology and estimate concept's *Information Content* (IC) from term's appearance frequencies. Even though they are able to provide accurate results when enough data is available [20], their applicability is hampered by the availability of this data and their pre-processing. On the contrary, the presented measures based uniquely on the exploitation of the taxonomical structure are characterized by their simplicity, which result is a computationally efficient solution, and their lack of constraints as only an ontology is required, which ensures their applicability. The main problem is their dependency on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology [21]. In order to overcome this problem, classical approaches rely on WordNet's *is-a* taxonomy to estimate the similarity. Such a general and massive ontology, with a relatively homogeneous distribution of semantic links and good inter-domain coverage is the ideal environment to apply those measures [20].

4. Categorical data recoding based on semantic similarity

Considering the poor semantics incorporated by existing methods for privacy preserving of categorical values, we have designed a new local method for anonymization based on the semantic processing of, potentially unbounded, categorical values.

Aiming to fulfill the *k*-anonymity property but minimizing the information loss of textual data, it is proposed a recoding method based on the replacement of some values of one attribute by the most semantically similar ones. The basic idea is that, if a value does not fulfilling the *k*-anonymity, it will be replaced by the most semantically similar value on the same dataset. This decreases the number of different values. The process is repeated until the whole dataset fulfils the desired *k*-anonymity. The rationale for this replacement criterion is that if categorical values are interpreted at a conceptual level, the way to lead to the least information loss is to change those values in a way that the semantics of the record – at a conceptual level – is preserved. In order to ensure this, it is crucial to properly assess the semantic similarity/distance between categorical values. Path-length similarities introduced in the previous section have been chosen because they provide a good estimation of concept alikeness at a

very low computational cost [19], which is important when dealing with very large datasets, as it is the case of inference control in statistical databases [1].

As categorical data are, in fact, text labels it is also necessary to morphologically process them in order to detect different lexicalizations of the same concept (e.g. singular/plural forms). We apply a stemming algorithm to both text labels of categorical attributes and ontological labels in order to compare words from their morphological root.

The inputs of the algorithm are: a dataset consisting on a single attribute with categorical values (an unbounded list of textual noun phrases) and n registers (r), the desired level of k-anonymity and the reference ontology.

Algorithm Ontology - based recoding (dataset, k, ontology)
$r_i' := stem\ (r_i) \forall\ i\ in\ [1 \ldots n]$
while $(there\ are\ changes\ in\ the\ dataset)$ *do*
  for $(i\ in\ [1 \ldots n])$ *do*
    $m := count\ (r_j' = r_i') \forall\ j\ in\ [1 \ldots n]$
    if $(m < k)$ *then*
      $r'_{Max} := argMax\ (similarity\ (r_i', r_j', ontology)) \forall\ j\ in\ [1 \ldots n], r_i' \neq r_j'$
      $r_p' := r'_{Max}\ \forall\ p\ in\ [1 \ldots n], r_p' = r_i'$
    end if
  end for
end while

The recoding algorithm works as follows. First, all words of dataset are stemmed, so that, two words are considered equal if their morphological roots are identical. The process iterates for each register r_i of the dataset. First, it checks if the corresponding value fulfils the k-anonymity by counting its occurrences in the dataset. Those values which occur less than k times do not accomplish k-anonymity and should be replaced. As stated above, the ideal word to replace another one (from a semantic point-of-view) is the one that has the greatest similarity (i.e. the least distant meaning). Therefore, from the set of words that already fulfill the minimum k-anonymity, the most similar to the given one according to the employed similarity measure and the reference ontology is found and the original value is substituted. The process finishes when no more replacements are needed, meaning that the dataset fulfills the k-anonymity property.

It is important to note that, in our method, categorical values may be found at any taxonomical level of the input ontology. So, in comparison to hierarchical generalization methods introduced in section 2, in which labels are always leafs of the ad-hoc hierarchy and terms are always substituted by hierarchical subsumers, our method replaces terms for the nearest one in the ontology, regardless being a taxonomical sibling (i.e. the same taxonomical depth), a subsumer (i.e. a higher depth) or an specialization (i.e. lower depth), provided that those appear more frequently in the sample (i.e. they fulfill the k-anonymity).

5. Evaluation

In order to evaluate our method, we used a dataset consisting on textual answers retrieved from polls made by "*Observatori de la Fundació d'Estudis Turístics Costa Daurada*" at the Catalan National Park "*Delta del Ebre*". The dataset consists on a sample of the answers of the visitors to the question: *What has been the main reason to visit Delta del Ebre?*. As answers are open, the disclosure risk is high, due to the heterogeneity of the sample and the presence of uncommon answers, which are easily identifiable. The test collection has 975 individual registers and 221 different responses, 84 of them are unique (so they can be used to re-identify the individual), while the rest have different amount of repetitions (as shown in Table 1).

Table 1. Distribution of answers in the evaluation dataset (975 registers in total).

Number of repetitions	1	2	3	4	5	6	7	8	9	11	12	13	15	16	18	19	Total
Number of different responses	84	9	6	24	23	37	12	1	2	7	5	1	5	2	2	1	**221**
Total amount of responses	84	18	18	96	115	222	84	8	18	77	60	13	75	32	36	19	**975**

The three similarity measures introduced in section 3 have been implemented and WordNet 2.1 has been exploited as the input ontology. As introduced in section 2, WordNet has been chosen due to its general purpose scope (which formalizes in an unbiased manner concept's meaning) and its high coverage of semantic pointers. To extract the morphological root of words we used the Porter Stemming Algorithm [22].

Our method has been evaluated for the three different similarity measures presented in section 2, in comparison to a random substitution (i.e. each word is replaced by a random one which accomplishes the desired *k*-anonymity). Different levels of *k*-anonymity have been tested.

The quality of the anonymization method has been evaluated from two points of view. On one hand, we computed the information loss locally to the sample set. In order to evaluate this aspect we computed the Information Content (IC) of each individual of each categorical value after the anonymization process in relation to the IC of the original sample. IC of a categorical value has been computed as the inverse to its probability of occurrence in the sample (4). So, frequently appearing answers had less IC than rare (i.e. more easily identifiable) ones.

$$IC(c) = -\log p(c) \tag{4}$$

The average of the IC value for each answer is subtracted to the average IC of the original sample in order to obtain a quantitative value of information loss with regards to the distribution of the dataset. In order to minimize the variability of the random substitution, we averaged the results obtained for five repetitions of the same test. The results are presented in Figure 1.

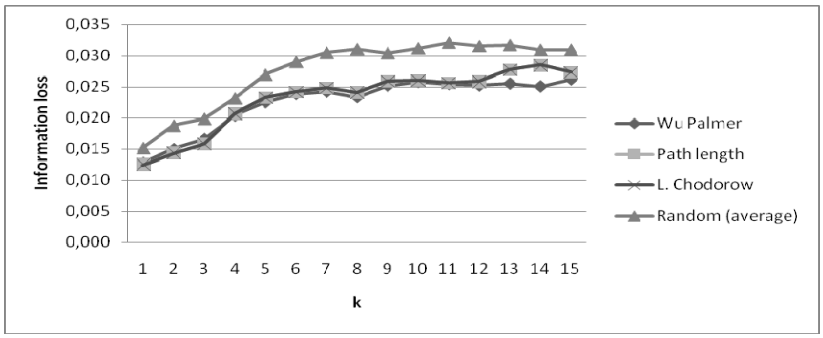


Fig 1. Information loss based on local IC computation.

To evaluate the quality of the masked dataset from a semantic point of view, we measured how different is the replaced value to the original one with respect to their meaning. This is an important aspect from the point of view of data exploitation as it represents a measure of up to which level the semantics of the original record are preserved. So, we computed the averaged semantic distance from the original dataset and the anonymized one using the Path Length similarity measure in WordNet. Results are presented in Figure 2.

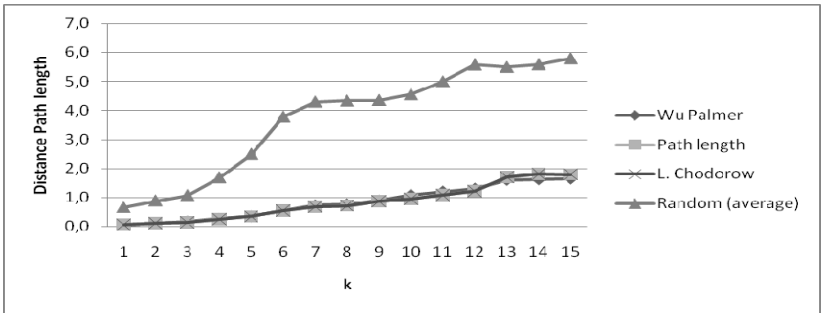


Fig. 2. Semantic distance of the anonymized dataset.

Analyzing the figures we can observe that our approach is able to improve the random substitution by a considerable margin. This is even more evident for a high *k*-anonymity level. Regarding the different semantic similarity measures, they provide very similar and highly correlated results. This is coherent, as all of them are based on the same ontological features (i.e. absolute path length and/or the taxonomical depth) and, even though similarity values are different, the relative ranking of words is very similar. In fact, Path length and Leacock and Chorodow measures gave identical results as the later is equivalent to the former but normalized to a constant factor (i.e. the absolute depth on the ontology). Evaluating the semantic distance in function of the level of *k*-anonymity one can observe a linear tendency with a very smooth growth. This is very convenient and shows that our approach performs well regardless the desired level of anonymization.

The local information loss based on the computation of the averaged IC with respect to the original dataset follows a similar tendency. In this case, however, the information loss tends to stabilize for k values above 9, showing that the best compromise between the maintenance of the sample heterogeneity and the semantic anonymization have been achieved with k=9. The random substitution performs a little worse, even though in this case the difference is much less noticeable (as it tends to substitute variables in a uniform manner and, in consequence, the original sample distribution tends to be maintained).

6. Conclusions

On the process of dataset anonymization it is necessary to achieve two main objectives: on one hand, to satisfy the desired k-anonymity to avoid the disclosure, preserving the confidentiality and, on the other hand, to minimize the information loss to maintain the quality of the dataset. This paper proposes a method of local recoding for categorical data, based on the estimation of semantic similarity between values. As the meaning of concepts is taken into account, the information loss can be minimized.

The method uses the explicit knowledge formalized in wide ontologies (like Wordnet) to calculate the semantic similarity of the concepts, in order to generate a masked dataset that preserves the meaning of the answers given by the respondents.

In comparison with the existing approaches for masking categorical data based on generalization of terms, our approach avoids the necessity of constructing ad-hoc hierarchies according to data labels. In addition, our method is able to deal with unbounded attributes, which can take values in a textual form.

The results presented show that with a level of anonymity up to 6, the semantics of the masked data is maintained 3 times more than with a naive approach. Classical information loss measure based on information content also shows an improvement of the ontology-based recoding method.

After this first study, we plan compare our method with the existing generalization masking methods mentioned in section 2, in order to compare the results of the different anonymization strategies. For this purpose, different information loss measures will be considered. Finally, we plan extend the method for global recoding, where different attributes are masked simultaneously.

Acknowledgements

Thanks are given to “Observatori de la Funcació d’Estudis Turístics Costa Daurada” and “Parc Nacional del Delta de l’Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)” for providing us the data collected from the visitors of the park. This work is supported the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02). Sergio Martínez Lluís is supported by the Universitat Rovira i Virgili predoctoral research grant.

References

1. Domingo-Ferrer, J. A survey of inference control methods for privacy-preserving data mining, in Privacy-Preserving Data Mining: Models and Algorithms, eds. C.C. Aggarwal and P.S. Yu, Advances in Database Systems, v.34, N.Y.: Springer Verlag, pp. 53--80 (2008)
2. Bouchon-Meunier, B., Marsala, C., Rifqi, M., Yager, R.R. Uncertainty and Intelligent Information Systems. World Scientific (2008)
3. Gómez-Pérez, A., Fernández-López, M., Corcho, O. Ontological Engineering, 2nd printing. Springer Verlag, pp. 79--84 (2004)
4. Willenborg, L. and DeEaal T. Elements of Statistical Disclosure Control. Springer-Verlag, New York (2001)
5. Guo, L., Wu, X. Privacy preserving categorical data analysis with unknown distortion parameters, Transactions on Data Privacy, 2, pp. 185--205 (2009)
6. Gouweleeuw, J.M. Kooiman, P., Willenborg, L. C. R. J. and DeWolf, P. P. Post randomization for statistal disclousure control: Theory and implementation. Research paper no. 9731 (Voorburg: Statistics Netherlands) (1997)
7. Reiss, S. P. Practical data-swapping: the first steps. ACM Transactions on Database Systems, 9 pp. 20--37 (1984)
8. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Wai-Chee Fu, A. Utility-based anonymization using local recoding, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, pp.785--790 (2006)
9. Sweeney, L. *k*-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) pp. 557--570 (2002)
10. Samarati, P., Sweeney, L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)
11. Iyengar, V. S. Transforming data to satisfy privacy constraints. Proceedings of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Minig (KDD), pp. 279--288 (2002)
12. Bayardo, R. J., Agrawal, R. Data privacy through optimal *k*-anonymization. Proceedings of the 21st International Conference on Data Engineering (ICDE) pp. 217--228 (2005)
13. Li, T., Li, N. Towards optimal *k*-anonymization. Data & Knowledge Engineering 65 pp. 22--39 (2008)
14. Guarino, N. Formal Ontology in Information Systems. In Guarino N (ed) 1st Int. Conf. on Formal Ontology in Information Systems, pp. 3--15. IOS Press. Trento, Italy (1998)
15. Fellbaum, C. WordNet: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press. (1998)
16. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, Swoogle, J.: A Search and Metadata Engine for the Semantic Web. In Proc. 13th ACM Conference on Information and Knowledge Management, pp. 652--659. ACM Press (2004)
17. Rada, R., Mili, H., Bichnell, E., Blettner, M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 9(1), 17--30 (1989)
18. Wu, Z., Palmer, M. Verb semantics and lexical selection. In Proc. 32nd annual Meeting of the Association for Computational Linguistics, pp. 133--138. New Mexico, USA (1994)
19. Leacock, C., Chodorow, M. Combining local context and WordNet similarity for word sense identification. In Fellbaum (ed.), WordNet: An electronic lexical database, pp. 265--283. MIT Press (1998)
20. Jiang, J., Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In Proc. Int. Conf. on Research in Computational Linguistics, pp. 19--33. Japan (1997)
21. Cimiano, P. Ontology Learning and Population from Text. Algorithms, Evaluation and Applications. Springer-Verlag (2006)
22. Porter. An algorithm for suffix stripping, Program, Vol. 14 no 3, pp. 130--137 (1980)

Ontology-based anonymization of categorical values

Sergio Martínez, David Sánchez, Aida Valls

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Avda. Països Catalans, 26, 43007 Tarragona, Spain
{sergio.martinezl, aida.valls, david.sanchez}@urv.cat

Abstract. The analysis of sensible data requires a proper anonymization of values in order to preserve the privacy of individuals. Information loss should be minimized during the masking process in order to enable a proper exploitation of data. Even though several masking methods have been designed for numerical data, very few of them deal with categorical (textual) information. In this case, the quality of the anonymized dataset is closely related to the preservation of semantics, a dimension which is commonly neglected of shallowly considered in related words. In this paper, a new masking method for unbounded categorical attributes is proposed. It relies on the knowledge modeled in ontologies in order to semantically interpret the input data and perform data transformations aiming to minimize the loss of semantic content. On the contrary to exhaustive methods based on simple hierarchical structures, our approach relies on a set of heuristics in order to guide and optimize the masking process, ensuring its scalability when dealing with big and heterogenous datasets and wide ontologies. The evaluation performed over real textual data suggests that our method is able to produce anonymized datasets which significantly preserve data semantics in comparison to apporaches based on data distribution metrics.

Keywords: Ontologies, Data analysis, Privacy-preserving data-mining, K-anonymity, Semantic similarity.

1. Introduction

Statistical agencies are an important source of information for intelligent data analysis and decision making. Those agencies collect responses of a set of individuals for which privacy must be guaranteed. So, before distributing the data, a masking method should be used in order to anonymize the data file and minimize the re-identification risk. The privacy level associated to masked data is typically related to the fulfilment of the *k-anonymity* property [16]. This property establishes that each anonymized record in a data set (i.e. a set of attribute values associated to an individual) has to be indistinguishable with at least *k*-1 other records within the same dataset, according to its individual attribute values.

However, in order to preserve the utility of the values (i.e. to make the anonymized data as useful as possible from the analysis and data mining point of view), it is important that the anonymization method minimizes the information loss that is inherent to the masking process. This is measured by means of a quality metric. Up to

this moment, most of the attention has been paid to numerical data or bounded categorical attributes. The goal of the masking methods for numerical data was to maintain the statistical characteristics of the dataset [4]. For categorical attributes, which represent a discrete enumeration of modalities (i.e. bounded vocabulary), quality metrics are focused on maintaining the probability distribution of the values in the masked file. This has been criticized by several authors [19] as value distribution does not capture important dimensions of data utility. In fact, as categorical attributes typically represent concepts, their utility should be associated to the preservation of their inherent semantics. Omitting those semantics during the anonymization process can hamper the application of data analysis or decision making processes on those data, since the conclusions obtained can be significantly different from those obtained from the original data file.

In any case, with the success of the Information Society, textual data have grown both in size and importance. Those values can be obtained with traditional questionnaires where the user can answer with a short sentence or a noun phrase, such as "*Main hobby*" or "*Most preferred type of food*". This kind of attributes has a potentially unbounded set of values that represent a concept with a concrete semantic. Those attributes are more challenging than those corresponding to a limited set of modalities. In order to properly interpret and compare them, the similarities between their meaning, at a conceptual level, should be taken into consideration (e.g. for hobbies, *trekking* is more similar to *jogging* than to *dancing*).

Due to the ambiguity of human languages and the complexity and knowledge modelling, very few masking methods have considered the semantics of attribute values in some degree. In fact, many approaches [1, 15, 16] completely ignore this issue, dealing with textual data in a naïve way, proposing arbitrary suppressions or substitutions aimed to fulfil *k*-anonymy and preserve the distribution of the input data, but neglecting the importance of the meaning of the data. As it will be discussed in Section 2, even though there exist approaches exploiting knowledge structures during the anonymization, they consider semantics in a very shallow and ad-hoc manner and tackle the anonymization in an exhaustive manner, hampering their scalability and applicability as a general-purpose solution.

In order to overcome those limitations, in this paper we propose a new method of local anonymization for unbounded categorical attributes, which exploits ontologies [2] as knowledge background to support the anonymization process from a semantic point of view. Ontologies offer a formal, explicit and machine readable structuring of a set of concepts by means of a semantic network where multiple hierarchies are defined and semantic relations are explicitly modelled as links between concepts [6]. Thanks to initiatives such as the Semantic Web [3], many ontologies have been created in the last years, bringing the development of general purpose knowledge sources (such as WordNet [5] for English words), as well as specific domain terminologies (e.g. medical sources such as UMLS).

Due to the large size of general purpose ontologies (with respect to ad-hoc knowledge structured exploited in previous approaches [1, 7, 12, 15, 16]), our algorithm tackles the anonymization in an heuristic fashion, providing better scalability with respect to the size of the ontology and the input data than related works based on exhaustive search.

The rest of the paper is organized as follows. Section 2 reviews methods for privacy protection of categorical data that take into account some kind of semantic information. Section 3 introduces classical metrics aimed to measure data quality and present other ways of semantically measuring the information loss by exploiting ontologies. In section 4, the proposed anonymization method is detailed. Section 5 is devoted to evaluate our method by applying it to real data obtained from a survey at the National Park "*Delta del Ebre*" in Catalonia, Spain. The final section contains the conclusions and future work.

2. Related work

When categorical attributes are extended to the case of having a set of potentially unbounded textual modalities representing concepts, domain knowledge is required in order to properly evaluate them.

In the previous knowledge-based masking methods, the set of values of a categorical attribute are represented by means of Value Generalization Hierarchies (VGHs) [1, 7, 12, 15, 16]. In those cases, ad-hoc manually constructed tree-like structures are defined according to input data, where categorical labels represent leafs of the hierarchy and they are recursively subsumed by common generalizations. The masking process consists on substituting the original values by a more general one, obtained from the hierarchical structure. This generalization process decreases the number of distinct tuples and, in consequence, increases the level of *k*-anonymity. In general, for each value, different generalizations are possible according to the depth of the tree. Typically, the selection is made according to a quality metric that measures the information loss derived from the value substitution.

More in detail, in [11, 15, 16] authors propose a hierarchical scheme in which all values of an attribute are generalized to the same level of the VGH. The number of valid generalizations for an attribute is the height of the VGH for that attribute. The concrete generalization is selected by generating all the possible ones for each value and selecting the combination that provides the closest generalizations in all cases fulfilling the desired level of *k*-anonymity. In this case, the level of generalization is used as a measure of information loss.

Iyengar [8] presented a more flexible scheme which also uses a VGH, where each value of an attribute can be generalized to a different level of the hierarchy. This scheme allows a much larger space of possible generalizations. Again, for all values, all the possible generalizations fulfilling the *k*-anonymity are generated. Then, a genetic algorithm finds the optimization of a set of information loss metrics.

T. Li and N. Li [12] propose three generalization schemes. First, the Set Partitioning Scheme (SPS) represents an unsupervised approach in which each partition of the attribute domain represents a generalization. This supposes the most flexible generalization scheme but the size of the solution space grows enormously, meanwhile the benefits of a semantically coherent VGH are not exploited. The Guided Set Partitioning Scheme (GSPS) uses a VGH to restrict the partitions of the attribute domain and exploits the height of the lowest common ancestor of two values as a metric of semantic distance. Finally, the Guided Oriented Partition Scheme

(GOPS) adds ordering restrictions to the generalized groups of values to restrict even more the possible generalizations. In all three cases, all the possible generalizations allowed by the proposed scheme are constructed, selecting the one that minimizes the information loss (evaluated by means of the discernibility metric [1]).

He and Naughton [7] propose a partitioning algorithm in which generalizations are created in a Top-Down fashion and the best one, according to quality metric (Normalized Certainty Penalty [17]), is recursively refined. Xu et al. [19] proposes a Utility-based generalization algorithm. In this case, the method supports defining different "utility" functions for each attribute, according to the importance of each attribute.

All the approaches relying on a VGH present a series of drawbacks. On one hand, VGHs are manually constructed from the attribute value set of the input data. So, human intervention is needed in order to provide the adequate semantic background in which those algorithms rely. If input data values change, the VGH should be modified accordingly. Even though this fact may be assumable when dealing with reduced sets of categories (e.g. in [12] a dozen of different values per attribute are considered in average) this hampers the scalability and applicability of the approaches, especially when dealing with unbounded textual data (with hundreds or thousands of individual answers). On the other hand, the fact that VGHs are constructed from input data (which represents a limited sample of the underlying domain of knowledge), produces ad-hoc and small hierarchies with a much reduced taxonomical detail. It is common to observe VGHs with three or four levels of hierarchical depth whereas a detailed taxonomy (such as WordNet) models up to 16 levels [5]. From a semantic point of view, VGHs offer a rough and biased knowledge model compared to fine grained and widely accepted ontologies. As a result, the space for valid generalizations that a VGH offers would be much smaller than when exploiting an ontology. Due to the coarse granularity of VGHs, it is likely to suffer from high information loss due to generalizations. As stated above, some authors try to overcome this problem by making arbitrary generalizations, but this introduces a considerable computational burden and lacks of a proper semantic background. Moreover, the quality of the result would depend on the structure of the VGH that, due to its limited scope, offers a partial and biased view of the domain.

An alternative to the use of a VGH is proposed in Bayardo and Agrawal [1]. Their scheme is based on the definition of a total order over all the values of an attribute. According to this order, partitions are created in order to define different levels of generalizations. As a result, the solution space is exponentially large. The problem here is that the definition of a semantically coherent total order for categorical attributes is very difficult and nearly impossible for textual data. Moreover, the definition of a total order unnecessarily imposes constraints on the space of valid generalizations.

From the point of view of semantic understanding of the input data, in order to overcome the limitations of the presented methods, one may consider their application over a wide and detailed general ontology like WordNet.

WordNet [5] is a freely available lexical database that describes and organizes more than 100,000 general English concepts, which are semantically structured in an ontological fashion. WordNet contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (synsets), each expressing a distinct

concept (i.e. a word sense). Synsets are linked by means of conceptual-semantic and lexical relations such as synonymy, hypernymy (subclass-of), meronymy (part-of), etc. The result is a network of meaningfully related words, where the graph model can be exploited to interpret concept's semantics. Hypernymy is, by far the most common relation, representing more than an 80% of all the modeled semantic links. The maximum depth of the noun hierarchy is 16. Polysemous words present an average of 2.77 synsets (i.e. they belong to almost three different hierarchies) and up to 29 different senses (for the "line" word). Considering those dimensions, the size of the generalization space would be several orders of magnitude bigger than when using ad-hoc VGHs. However, as most of the presented approaches make generalizations in an exhaustive fashion, the generalization space is exponentially large according to the depth of the hierarchy, the branching factor and the values to evaluate. So, those approaches are computationally too expensive and hardly applicable in such a big ontology like WordNet.

In order to be able to exploit the semantic background provided by big ontologies like WordNet, we present a non-exhaustive heuristic value substitution which, bounding the search space according to the input data values and based on the theory of *semantic similarity* (see Section 3), is able to scale well in such a big ontology while minimizing the loss of semantics.

3. Quality metrics

As stated above, the goal of an anonymization method is finding a transformation of the original data, which satisfies *k-anonymity* while minimizing the information loss and, in consequence, maximizing the utility of the resulting data.

In the literature, various metrics have been proposed and exploited [1, 7, 8, 11, 12, 19] to measure the quality of anonymized data. Classical metrics, such as Dicernability Metric (DM) [1], evaluate the distribution of n records (corresponding to n individuals) into g groups of identical values, generated after the anonymization process. Concretely, (DM) assigns to each record a penalty based on the size of the group g_i to which it belongs after the generalization (1). A uniform distribution of values in equally sized groups (with respect to the original data) is the goal.

$$DM = \sum_{i=1}^{n} |g_i|^2 \tag{1}$$

However, metrics based on data distribution do not capture how *semantically similar* the anonymized set is with respect to the original data. As stated in the introduction, preservation of semantics when dealing with textual attributes is crucial in order to be able to interpret and exploit anonymized data. In fact, this aspect is, from the utility point of view, more important than the distribution of the anonymized dataset when aiming to describe or understand a record by means of its attributes.

In order to minimize the loss of semantics between original and anonymized datasets, we propose relying on the theory of *semantic similarity* [9]. Semantic similarity measures the taxonomical alikeness between words based on the semantic evidences extracted from one or several knowledge sources. As stated in section 2,

ontologies like WordNet offer wide and detailed views of knowledge domains and, in consequence, represent an ideal source from which computing semantic similarity [9]. As stated in the introduction, ontologies offer a graph model in which semantic interrelations are modeled as links between concepts. As a result, semantic similarity can be estimated as a function of the taxonomic inter-link distance.

In an is-a hierarchy, the simplest way to estimate the distance between two concepts c_1 and c_2 is by calculating the shortest *Path Length* (i.e. the minimum number of links) connecting these concepts (2) [14].

$$dis_{pL}(c_1, c_2) = \min \#\ of\ is-a\ edges\ connecting\ c_1\ and\ c_2 \quad (2)$$

However, this measure omits the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered as less similar that those belonging to a lower level, as they present different degrees of generality. Based on this premise Wu and Palmer's measure [18] also takes into account the depth of the concepts in the hierarchy (3).

$$sim_{w\&p}(c_1, c_2) \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (3)$$

, where N_1 and N_2 are the number of is-a links from c_1 and c_2 respectively to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

Based on the same principles Leacock and Chodorow [10] also proposed a measure that considers both the shortest path between two concepts (in fact, the number of nodes N_p from c_1 to c_2) and the depth D of the taxonomy in which they occur in a non-linear fashion (4).

$$sim_{l\&c}(c_1, c_2) = -\log(N_p / 2D) \quad (4)$$

Those measures will be exploited by our approach in order minimize the loss of semantics during the substitution of sensible values.

4. Ontology-based anonymization of categorical data

As discussed in section 2, exhaustive generalization methods are too expensive to be applicable over wide ontologies like WordNet. Moreover, the fact that values to anonymize correspond to leafs of the VGH implies that values are only substituted by more general ones (which unnecessarily imposes constraints on the space of valid generalizations).

Our approach, which aims to provide local anonymization of attribute values, tackles the problem in a different manner. Thanks to the wide coverage of WordNet, one would be able to map sensible values to ontological nodes which do not necessary represent leafs of a hierarchy. As a result, semantically related concepts can be retrieved going through the hierarchy/ies to which the value belongs. Moreover ontological hierarchies are designed in a much general and fine grained fashion than ad-hoc VGHs, according to the agreement of domain knowledge experts, not in

function on the input data. Those facts open the possibility of substituting sensible values by a much wider and knowledge-coherent set of semantically similar elements, including taxonomical subsumers (as done in generalization methods) but also with hierarchical siblings (with the same taxonomical depth) or specializations (located in a lower level). In fact, in many situations, an specialization may be more similar that a subsumer because, as stated in section 3, concepts belonging to lower levels of a hierarchy have less differentiated meanings due to their concreteness. As a result, the value change would result in less information loss and a higher preservation of data utility from a semantic point of view.

In order to ensure that value substitutions lead to the fulfillment of the desired degree of privacy, we should substitute each sensible value for another one that increases the level of *k*-anonymity. This implies that either value pairs are substituted for a new one which is "near" to both of them, or that one value is changed for another one already existing in the data set; in both cases, the goal is to make both values indistinguishable. It is important to note that, in all cases, the loss of semantic content would be equivalent: if all values of the dataset are semantically far, so are their related nodes, resulting in an inevitable high loss of semantics either by changing them for the nearest node to both of them or by substituting one for the other. As the first option would lead to an enormous set of possible substitutions according to all the semantically related concepts available in the ontology for each sensible value, we opted for the second strategy. As a result, the space of valid substitutions is bounded to the number of *different* values available in the dataset.

The most appropriate value to which a non anonymous one should be substituted is the one that minimizes the semantic distance with respect to the original. So, semantic similarity metrics introduced in Section 3 (which explore and quantify the distance of ontological nodes in the semantic network) can be used to select the substitution and minimize the loss of semantic content. As a result of a value replacement, the number of different values is decreased and the *k*-anonimity is increased. The process is repeated until the whole dataset fulfills the desired *k*-anonymity level.

As we are dealing with values represented by text labels, it is also necessary to morphologically process them in order to detect different lexicalizations of the same concept (e.g. singular/plural forms). We apply a stemming algorithm to both text labels of categorical attributes and ontological labels in order to be able to map values to ontological concepts and to detect conceptually equivalent values in the dataset.

Notice that the order in which the values to be replaced are selected may produce affect the anonymization. The generation of the optimum result implies generating all possible substitution iterations for all sensible values and picking the order that maximizes the quality of the result set. As unbounded textual attributes may usually correspond to a high number of different answers, many of them being unique, the amount of values not fulfilling the *k*-anonymity would be high. Consequently, as the cost of generating all the possible combinations is O(n!), it is computationally too expensive. In order to ensure the scalability of our approach, we implemented several heuristics that aim to select, at each step, the substitution that would likely maximize the quality of the result.

The first heuristic consists on selecting the value with the lowest number of repetitions in the original set (i.e. the more identifiable). The motivation is that those values would require a higher number of substitutions in order to fulfill the desired *k*-

anonymity level. In case of a tie (e.g. several unique values, which would be very common with free text attributes), the algorithm selects the value for which its best substitution (according to the quality metric) leads to the minimum semantic information loss (according to the same quality metric), aiming to maximize the quality of the result dataset. Finally, if several replacements imply the same information loss (which would be quite rare), the algorithm selects the value for which the *k*-anonymity level resulting from that change is lower. Again, values which are more difficult to anonymize are prioritized, as they require more substitutions.

Formally, the algorithm has the following inputs: D, a set of n categorical values for a single attribute (i.e. an unbounded list of textual noun phrases, each one referring to an ontological concept), the desired level of *k*-anonymity and the ontology, o.

```
1   Ontology-based local anonymization (D, k, o)
2      D' := stem (D)
3      D' := rank by number of repetitions (D')
4      v := first value (D')
5      while (number of repetitions (v, D') < k) do
6         V := values with the same number of repetitions (v, D')
7         V_max := set of values with the maximum similarity (D', V)
8         v' := value with minimum resulting k-anonymity (D', V_max)
9         D' := replace all occurrences of the value in the set (D', v')
10        D' := rank by number of repetitions (D')
11        v := first register (D')
12     end while
13  end
```

The algorithm works as follows. First, all words of the attribute dataset are stemmed, so that, two words are considered equal if their morphological roots are identical (line #2). The set is ascending ranked according to the number of value repetitions; then, the first value (v) is the register with the lowest *k*-anonymity (line #4). It checks if the corresponding value fulfils the *k*-anonymity according to the number of repetitions (line #5). If *k*-anonymity is fulfilled, the entire set will be anonymized. Otherwise, the value should be replaced. The algorithm selects all the values the same minimum number of repetitions (line #6) and finds another value in the dataset with results in the maximum semantic similarity according to the quality metric (introduced in section 3) (line #7). If several substitutions are equally optimum, it is selected the value whose replacement results in the lowest *k*-anonymity level (i.e. repetitions) (line #8). Finally, all the occurrences in the dataset for that value are substituted (line #9) and the dataset is reordered. The process finishes when no more replacements are needed, because the dataset is *k*-anonymous.

The most computationally expensive function corresponds to the calculation of the semantic similarity between value pairs, executed p^2 times in the line #7, being p the number of different labels in the attribute ($p \leq n$, being n the total number of attribute values). In the worst case, when the main loop (line #5) ends, this calculation is executed $p^2 \cdot p \cdot k = p^3 \cdot k$ times. However, as the total set of different values are known a priori and do not change during the masking process (unlike generalization methods), it is possible to pre-calculate and store the similarities between all of them. This avoids repeating similarity measuring calculus for already evaluated value pairs. In this manner, the calculation of the similarity measure is executed a priori only p^2

times and, as it will illustrated in the evaluation section, the execution of the algorithm stays in the range of milliseconds for hundred-sized datasets.

It is important to note that the computational cost of our algorithm uniquely depends on the number of different labels, unlike the related works that depend on the total size of the dataset and on the depth and branching factor of the hierarchy (which represent an exponentially large generalization space).

5. Evaluation

We have evaluated the proposed method by applying it to a dataset consisting on textual answers to the question "*What has been the main reason to visit Delta del Ebre?"* retrieved from polls made by "*Observatori de la Fundació d'Estudis Turístics Costa Daurada*" at the Catalan National Park "*Delta del Ebre*".

The dataset consists on a set of textual and unbounded answers regarding user preferences expressed by means of a noun phrase (with one or several words). As answers are open, the disclosure risk is high and, therefore, individuals are easily identifiable. The dataset is composed by 975 individual registers, with 221 different responses, being 84 of them unique. Note that this sample represents a much wider and heterogeneous test bed than those reported in related works [12], which are focused on bounded categorical values.

As those answers correspond to general and widely used concepts (i.e. sports, beach, nature, etc.) all of them have been found in WordNet 2.1, corresponding to one or several synsets. We used the Porter Stemming Algorithm [13] to extract the morphological root of words and to detect semantically equivalent answers.

We evaluated our approach from two points of view. First, we measured the contribution of the designed heuristics in guiding the substitution process towards minimizing the information loss form a semantic point of view (as detailed in section 4). We used Wu and Palmer, Leacock and Chodorow and Path Length measures (see section 3) as quality metrics.

As baseline, we implemented a naïve substitution method that consists on replacing each sensible value by a random one from the same dataset. Following the same basic algorithm presented in section 4, each random change would increase the level of k-anonymity; the process ends when all values are anonymized. Values are ordered alphabetically, in order to avoid depending on the initial order of data. The results obtained for the random substitution are the average of 5 executions.

We compared our heuristic approach against the random substitution for different levels of k. To evaluate the quality of the masked dataset from a semantic point of view, we measured how semantically similar the replaced values are, in average, with respect to the original ones. We computed the averaged difference of semantics between original and anonymized sets using the Wu and Palmer's (Fig. 1) and Path Length (Fig. 2) measures.

Analyzing the figures, we can observe that our approach is able to improve the random substitution by a considerable margin. This indicates the usefulness and necessity of a heuristic substitution aimed to minimize the semantic content loss of the original dataset. This is even more noticeable for a high k level. Evaluating the

semantic distance in function of the desired level of *k*-anonymity, one can observe a linear tendency with a very smooth growth. This is very convenient and shows that our approach performs well regardless the desired level of anonymization. Regarding the different semantic similarity measures, they provide very similar and highly correlated results. This is coherent, as all of them are based on the same ontological features (i.e. absolute path length and/or the taxonomical depth) and, even though similarity values are different, the relative ranking of words is very similar. In fact, Path length and Leacock and Chorodow measures gave identical results as the later is equivalent to the former but normalized to a constant factor (i.e. the ontology depth).

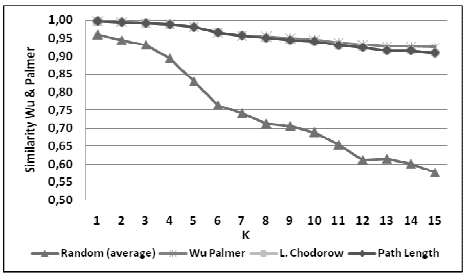


Fig. 1. Semantic similarity of the anonymized dataset

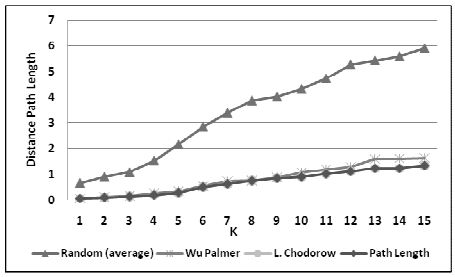


Fig. 2. Distance Path Length of the anonymized dataset

On the other hand, in order to show the importance of a semantically focused anonymization, we simulated the effect that a more traditional making schema, aimed to preserve the distribution of the masked dataset (as stated at the beginning of section 3), will represent on the resulting dataset. This has been done by using the Discernability metric (eq. 1) in our algorithm instead a semantic similarity measure as a quality metric, in order to guide the substitution process. Both approaches (semantic, based on Wu and Palmer's measure, and distributional, based on Discernability metric) have been compared by evaluating the semantic loss of the anonymized dataset (for different levels of *k*). Again, this loss is computed as the semantic similarity with respect to the original data by means of the Wu and Palmer's measure (see Fig. 3).

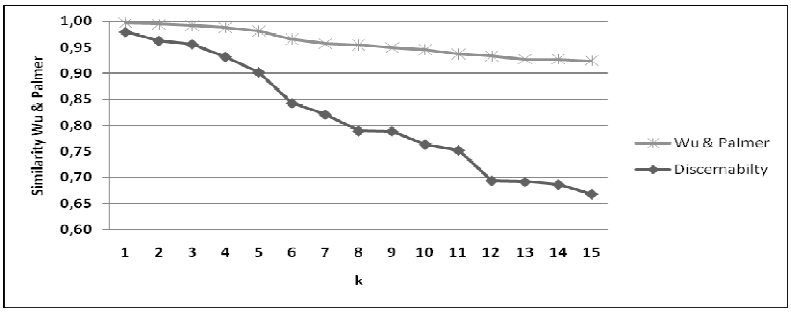


Fig. 3. Semantic similarity for our method with respect to a distributional metric

The figure shows that the optimization of dataset distribution and the preservation of information semantics are not correlated. In fact, there exists a very noticeable semantic loss in the resulting dataset for *k* values above 5. As stated in the

introduction, the utility of textual information from the data analysis point of view is highly dependent on its semantics. One can see that classical approaches focused on providing uniform groups of masked values may significantly modify dataset's meaning, hampering their exploitation.

From a temporal perspective, executing our method over a 2.4 GHz Intel Core processor with 4 GB RAM, the runtime of the anonymization process ranged from 0.7 to 1.3 seconds (according to the desired level of *k*-anonymity) as shown in Fig. 4. The pre-calculus of the semantic similarities between all value pairs of the dataset lasted 2.24 minutes. One can easily see how, as stated in section 4, similarity computation represents the most computationally expensive function, and how the minimization of the number of calculus results in a very noticeable optimization of runtime. Runtimes are also much lower than those reported by related works (several hours [12, 19]) based on generalization schemas and very limited VGHs and bounded categorical data (3-4 levels of depth and an average of a dozen of values [12]). This shows the scalability of our method when applied with large and heterogeneous textual data and big and wide ontologies like WordNet.

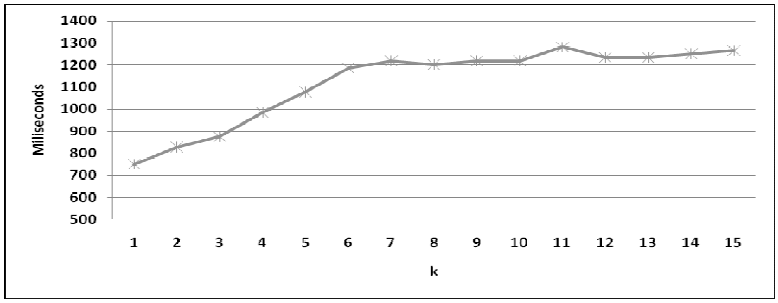


Fig. 4. Anonymization process runtime according to the level of k-anonymity

6. Conclusions

Categorical anonymization deals with two, a priori, confronted aspects of information: on the one hand, avoiding disclosure by fulfilling a desired level of *k*-anonymity and, on the other hand, maximization of data utility in order to properly exploit them. Previous approaches neglected or very shallowly considered the semantic content of textual data. As discussed in this paper, the *meaning* of information is an important dimension when aiming to apply analyses and mining processes over anonymized data.

This paper proposes a local masking method for unbounded categorical data based on the exploitation of wide and general ontologies aimed to preserve the semantics of the dataset. Special care has been put in ensuring the scalability of the method when dealing with large and heterogeneous datasets (which are very common when involving text attributes) and big ontologies like WordNet. By enabling the exploitation of those already available ontologies we avoid the necessity of constructing ad-hoc hierarchies according to data labels like VGH-based schemas, which supposes a serious cost and limits the method's applicability.

As future lines of research, we plan to extend our method to global anonymization of complete registers, where different attributes should be masked simultaneously.

References

1. Bayardo, R. J., Agrawal, R. Data privacy through optimal *k*-anonymization. Proceedings of the 21st International Conference on Data Engineering (ICDE) pp. 217--228 (2005)
2. Cimiano, P. Ontology Learning and Population from Text. Algorithms, Evaluation and Applications. Springer-Verlag (2006)
3. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, Swoogle, J.: A Search and Metadata Engine for the Semantic Web. In Proc. 13th ACM Conference on Information and Knowledge Management, pp. 652--659. ACM Press (2004)
4. Domingo-Ferrer, J. A survey of inference control methods for privacy-preserving data mining, in Privacy-Preserving Data Mining: Models and Algorithms, eds. C.C. Aggarwal and P.S. Yu, Advances in Database Systems, v.34, N.Y.: Springer Verlag, pp. 53--80 (2008)
5. Fellbaum, C. WordNet: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press. (1998)
6. Guarino, N. Formal Ontology in Information Systems. In Guarino N (ed) 1st Int. Conf. on Formal Ontology in Information Systems, pp. 3--15. IOS Press. Trento, Italy (1998)
7. He, Y., Naughton, J.F. Anonymization of Set-Valued Data via Top-Down, Local Generalization. 35th Int. Conf. VLDB. V.2 pp. 934--945 Lyon, France (2009)
8. Iyengar, V. S. Transforming data to satisfy privacy constraints. Proceedings of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Minig (KDD), pp. 279--288 (2002)
9. Jiang, J., Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In Proc. Int. Conf. on Research in Computational Linguistics, pp. 19--33. Japan (1997)
10. Leacock, C., Chodorow, M. Combining local context and WordNet similarity for word sense identification. In Fellbaum (ed.), WordNet: An electronic lexical database, pp. 265--283. MIT Press (1998)
11. Lefevre, K., DeWitt, D.J., Ramakrishnan, R. Mondrian Multidimensional K-Anonymity. Proceedings on the 22nd Int. Conf. on Data Engineering ICDE pp. 25 (2006)
12. Li, T., Li, N. Towards optimal *k*-anonymization. Data & Knowledge Engineering 65 pp. 22--39 (2008)
13. Porter. An algorithm for suffix stripping, Program, Vol. 14 no 3, pp. 130--137 (1980)
14. Rada, R., Mili, H., Bichnell, E., Blettner, M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 9(1), 17--30 (1989)
15. Samarati, P., Sweeney, L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)
16. Sweeney, L. *k*-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) pp. 557--570 (2002)
17. Terrovitis, M., Mamoulis, N. Kalnis, P. Privacy-preserving anonymization of set-valued data. In Proc. of VLDB (2008)
18. Wu, Z., Palmer, M. Verb semantics and lexical selection. In Proc. 32nd annual Meeting of the Association for Computational Linguistics, pp. 133--138. New Mexico, USA (1994)
19. Xu, J., Wang, W., Pei, J., Wang X., Shi, B., Wai-Chee Fu, A. Utility-Based Anonymization for Privacy Preservation with Less Information Loss. ACM SIGKDD Explorations Newsletter V.8 I.2 pp. 21--30 (2006)

Privacy protection of textual attributes through a semantic-based masking method

Sergio Martínez, David Sánchez, Aida Valls*, Montserrat Batet

Department of Computer Science and Mathematics. Universitat Rovira i Virgili
Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) research group
Av. Països Catalans, 26. 43007. Tarragona, Catalonia (Spain)

* Corresponding author. E-mail: aida.valls@urv.cat. Phone: +34 977559688, Fax: +34 977559710.

Summary

Exploitation of microdata provided by statistical agencies can bring many benefits from the point of view of data mining. However, this data often refers to sensible information which can be directly or indirectly associated to individuals. A proper anonymization process is required to minimize the disclosure risk. Several masking methods have been developed for dealing with numerical data or bounded categorical values, but approaches tackling the anonymization of textual values are scarce and shallow. Due to the importance of textual data in Information Society, in this paper we present a new masking method aimed to anonymize unbounded textual values, based on the fusion of records with similar values to form groups of indistinguishable individuals. As the utility of textual information from the data exploitation point of view is closely related to the preservation of its meaning, our method relies on the structured knowledge representation given by ontologies. This domain knowledge is used to guide the masking process towards the merging that best preserves the semantics of the original data. Since textual data typically consist on large and heterogeneous value sets, our method focuses on providing a computationally efficient algorithm by relying on several heuristics instead of exhaustive searches. The method is evaluated with real data in a concrete data mining application consisting in solving a clustering problem. The method is also compared against more classical approaches, focused on the optimization of the value distribution of the dataset. Results show that a semantically-grounded anonymization preserves better the utility of data, both in theoretical and practical settings, offering a low the probability of record linkage. At the same time, it achieves a good scalability with regards to the size of input data.

Keywords: Privacy protection, anonymity, ontologies, semantic similarity, fusion of textual data.

1 Introduction

Statistical agencies generally provide summarized data generated from a collection of responses given by a set of individuals. In this way, subject's privacy can be easily guaranteed because responses are not directly published, so privacy preserving techniques must only take into account that an intruder cannot infer individual's information from these summarized data [1]. However, this kind of information may be not useful enough if a detailed analysis of the responses is desired. In fact, many intelligent data mining techniques can help to discover interesting knowledge from sample data such as user profiles, tendencies and user behaviours. This kind of data analysis requires the detailed individual information, which corresponds to subject's response values (known as *microdata*). In this latter case, the protection the data to be protected consists on a set of *m* records (corresponding to *m* individuals), each one represented by a type with the values of *n* attributes (or variables).

Due to the potential benefits of exploiting microdata, new masking techniques are being developed to minimize the risk or re-identification when this information is made available [2]. From the privacy preserving point of view, data attributes are classified into 4 types: identifiers (that unambiguously identify the individual), quasi-identifiers (that may identify some of the respondents, especially if they are combined with the information provided by other attributes), confidential outcome attributes (that contain sensitive information) and non-confidential outcome attributes (the rest). The first two types are the most critical ones, especially when they are associated to confidential data, as they lead to the re-identification of individuals. Even though identifiers (such as id-card numbers) can be directly removed from the dataset, quasi-identifiers are more problematic as they increase as the dataset includes a larger number of variables, resulting in unique value combinations and, in consequence, easing the re-identification. In order to avoid the presence of registers with unique value combinations, the *k-anonymity* property should be fulfilled [3]. It establishes that each record in a dataset has to be indistinguishable with at least *k-1* other records within the same dataset, according to its individual attribute values. So, the *k* value establishes the degree of desired anonymity.

In order to fulfil the *k*-anonymity property, micro-aggregation making methods have been designed aiming to build groups of *k* indistinguishable registers by substituting the original values with a prototype. Obviously, this process results in a loss of information which may compromise the utility of the anonymized data from the data mining point of view. Ideally, the masking method should minimize this loss and maximize data utility according to a certain metric. We can distinguish between *global* anonymization methods in which all identifier or quasi identifier attributes are considered and anonymized at the same time (i.e. records will fulfil *k*-anonymity) and *local* ones in which each attribute is anonymized independently (i.e. each attribute will fulfil *k*-anonymity

individually). In the latter case, the information loss of the whole dataset is not optimized because the transformations only have a local view of the problem.

Statistical agencies provide numerical and non-numerical data. In the past, many micro-aggregation methods have been designed for building groups of numerical data [2]. Numbers are easy to manage and compare; so, the quality of the resulting dataset from the utility point of view can be optimized by retaining a set of statistical characteristics [2]. However, the extension of these methods to categorical attributes is not straightforward, because of the limitations on defining appropriate aggregation operators for categorical values, which have a restricted set of possible operations. Moreover, categorical attributes may have a potentially large and rich set of modalities if the individuals are allowed to give responses in textual form. Due to the nature of this kind of values and the ambiguity of human languages, the definition of appropriate aggregation operators is even more difficult. Word semantics play a crucial role in the proper interpretation of this data, a dimension which is commonly ignored in the literature. In fact, some authors [3], [4], [5] deal with this data as a bounded set of categories for which suppressions or substitutions are executed in order to fulfil *k*-anonymity without having into account the semantics of the values. The quality of masked non-numerical data is typically considered by preserving the distribution of input data. Even though data distribution is a dimension of data utility, we argue, as it has been stated by other authors [6] that retaining the semantics of the dataset play a more important role when one aims to extract conclusions by means of data analysis.

Semantic interpretation of textual attribute values for masking purposes requires the exploitation of some sort of structured knowledge sources which allow a mapping between words and semantically interrelated concepts. As it will be discussed in Section 2, some approaches have incorporated some sort of background knowledge during the masking process. However, due to the lightweight and ad-hoc nature of that knowledge and the shallow semantic processing of data, they hamper their applicability as a general-purpose solution. On the contrary, we argue that the use of well-defined general purpose semantic structures, as ontologies, will allow a better interpretation of data [7], [8]. Ontologies are formal and machine readable structures of shared conceptualisations of knowledge domains, expressed by means of semantic relationships [9]. Thanks to initiatives such as the Semantic Web [10], many ontologies have been created in the last years, from general purpose ones, such as WordNet [11] (for English words), to specific domain terminologies (e.g. medical sources such as SNOMED-CT [12] or MeSH [13]).

Moreover, as it will be also stated in Section 2, related works typically tackle the anonymization in an exhaustive manner, defining an exponentially large search space of value substitutions. As a result, the scalability of the method is compromised specially when dealing with unbounded textual

attributes. In fact, those attributes are more challenging than a small and pre-defined set of modalities, which are typically considered in the literature [5], [10], [14], [15]. However, by incorporating free textual answers in traditional questionnaires, we are able to obtain more precise knowledge of the individual characteristics, which may be interesting for the posterior study of the dataset. At the same time, the privacy of the individuals is more critical, as the disclosure risk increases due to the uniqueness of the answers. This has been argued in some previous works [16] in which we proposed a simple algorithm to mask textual attributes individually.

In order to overcome the limitations identified in related works, in this paper we propose a global masking method for unbounded textual values. It is based on the merging of quasi-identifier values of the input records, which permits to build groups of indistinguishable registers with multiple textual attributes in a way in which *k*-anonymity is fulfilled. The method relies on the well-defined semantics provided by big and widely used ontologies like WordNet. This permits to properly interpret words' meaning and maximize the quality of the anonymized data from the semantic point of view. The aim is that the conclusions that may be inferred from the masked dataset by means of data analysis methods would be the most similar to those obtained from the original data. Due to potentially large size of ontologies (with respect to ad-hoc knowledge structured exploited in previous approaches [3], [4], [5], [15], [17]) and the fact of dealing with potentially unbounded textual attributes, we propose a non-exhaustive heuristic approach which provides better scalability with respect to the size of the ontology and the input data than related works. Our proposal will be evaluated both from theoretical and practical sides by applying our method to real data and comparing the results of out method with another masking approach based on the optimization of data distribution.

The rest of the paper is organized as follows. Section 2 reviews the methods for privacy protection of categorical data, focusing on those that take into account some kind of semantic knowledge. Section 3 discusses the exploitation of ontologies for data anonymization purposes and details the proposed method, describing the semantic foundations in which it relies, the designed heuristics and the expected computational cost. Section 4 is devoted to test our method by applying it to real data obtained from a survey at the National Park "Delta del Ebre" in Catalonia (Spain). It evaluates the method under the dimensions of data utility preservation and minimization of disclosure risk. The final section contains the conclusions and future work.

2 Related works

As stated above, masking of categorical data is not straightforward due to the textual nature of attribute values. Some basic works consider categorical data as enumerated terms for which only

boolean word matching operations can be performed. On the one hand, we can find methods based on data swapping (which exchange values of two different records) and methods that add of some kind of noise (such as the replacement of values according to some probability distribution done in PRAM [18], [19]). On the other hand, other authors [3], [5] perform local suppressions of certain values or select a sample of the original data aimed to fulfil *k*-anonymity while maintaining the information distribution of input data.

Even though those methods are effective in achieving a certain degree of privacy in an easy and efficient manner, they fail to preserve the meaning of the original dataset, due to their complete lack of semantic analysis. Due to this reason, in recent years, some authors have incorporated some kind of knowledge background to the masking process.

In previous knowledge-based masking methods, the set of values of each categorical attribute of the input records in the dataset are represented by means of *Value Generalization Hierarchies* (VGHs) [3], [4], [5], [14], [15], [17], [20]. Those are ad-hoc and manually constructed tree-like structures defined according to a given input dataset, where categorical labels of an attribute represent leafs of the hierarchy and they are recursively subsumed by common generalizations. The masking process consists on, for each attribute, substituting several original values by a more general one, obtained from the hierarchical structure associated to that attribute. This generalization process decreases the number of distinct tuples in the dataset and, in consequence, increases the level of *k*-anonymity. In general, for each value, different generalizations are possible according to the depth of the tree. The concrete substitution is selected according to a metric that measures the information loss of each substitution with regards to the original data.

More in detail, in [3], [5], [20] authors propose a global hierarchical scheme in which all values of each attribute are generalized to the same level of the VGH. The number of valid generalizations for each attribute is the height of the VGH for that attribute. For each attribute, the method picks the minimal generalization which is common to all the record values for that attribute. In this case, the level of generalization is used as a measure of information loss.

Iyengar [14] presented a more flexible scheme that also uses a VGH, where a value of each attribute can be generalized to a different level of the hierarchy in different steps. This scheme allows a much larger space of possible generalizations. Again, for all values and attributes, all the possible generalizations fulfilling the *k*-anonymity are generated. Then, a genetic algorithm finds the optimum one according to a set of information loss metrics measuring the distributional differences with regards to the original dataset.

T. Li and N. Li [15] propose three global generalization schemes. First, the *Set Partitioning Scheme* (SPS) represents an unsupervised approach in which each possible partition of the attribute values represents a generalization. This supposes the most flexible generalization scheme but the size of the solution space grows enormously, meanwhile the benefits of a semantically coherent VGH are not exploited. The *Guided Set Partitioning Scheme* (GSPS) uses a VGH per attribute to restrict the partitions of the corresponding attribute and uses the height of the lowest common ancestor of two values as a metric of semantic distance. Finally, the *Guided Oriented Partition Scheme* (GOPS) adds ordering restrictions to the generalized groups of values to restrict even more the set of possible generalizations. Notice that in the three cases, all the possible generalizations allowed by the proposed scheme for all attributes are constructed, selecting the one that minimizes the information loss (evaluated by means of the discernibility metric [4]).

On the contrary to global methods introduced above, He and Naughton [17] propose a local partitioning algorithm in which generalizations are created for an attribute individually in a Top-Down fashion. The best combination, according to quality metric (Normalized Certainty Penalty [21]), is recursively refined. Xu et al. [6] also proposes a local generalization algorithm based on individual attribute utilities. In this case, the method defines different "utility" functions for each attribute, according to their importance. Being local methods, each attribute is anonymized independently, resulting in a more constrained space of generalizations (i.e. it is not necessary to evaluate generalization combinations of all attributes at the same time). However, the optimization of information loss for each attribute independently does not imply that the result obtained is optimum when the whole record is considered. As stated in the introduction, non necessary generalizations would be typically done in a local method as each attribute should fulfil *k*-anonymity independently.

All the approaches relying on VGHs present some drawbacks. On one hand, VGHs are manually constructed from each attribute value set of the input data. So, human intervention is needed in order to provide the adequate semantic background in which those algorithms rely. If input data values change, VGHs should be modified accordingly. Even though this fact may be assumable when dealing with reduced sets of categories (e.g. in [15] a dozen of different values per attribute are considered in average), this hampers the scalability and applicability of the approaches, especially when dealing with unbounded textual data (with potentially hundreds or thousands of individual answers). On the other hand, the fact that VGHs are constructed from input data (which represents a limited sample of the underlying domain of knowledge), produces ad-hoc and small hierarchies with a much reduced taxonomical detail. It is common to observe VGHs with three or four levels of hierarchical depth whereas a detailed taxonomy (such as WordNet) models up to 16 levels [11]. From a semantic point of view, VGHs offer a rough and biased knowledge model compared to fine grained and widely accepted ontologies. As a result, the space for valid generalizations that a VGH offers

would be much smaller than when exploiting an ontology. Due to the coarse granularity of VGHs, it is likely to suffer from high information loss due to generalizations. As stated above, some authors try to overcome this problem by trying all the possible generalizations exhaustively, but this introduces a considerable computational burden and lacks of a proper semantic background. Finally, the quality of the results heavily depends on the structure of VGHs that, due to their limited scope, offer a partial and biased view of each attribute domain.

An alternative to the use of VGHs is proposed in Bayardo and Agrawal [4]. Their scheme is based on the definition of a total order over all the values of each attribute. According to this order, partitions are created to define different levels of generalization. As a result, the solution space is exponentially large. The problem here is that the definition of a semantically coherent total order for categorical attributes is very difficult and nearly impossible for unbounded textual data. Moreover, the definition of a total order unnecessarily imposes constraints on the space of valid generalizations.

3 Exploiting ontologies for anonymizing textual attributes

As stated in the introduction, in order to overcome the limitations presented by related works caused by their dependency on ad-hoc knowledge structures and their shallow semantic analysis, one may consider the exploitation of a wide and detailed general ontology like WordNet. In this case, attribute values (i.e. words) can be mapped to ontological nodes (i.e. concepts) via simple word-concept label matching so that the hierarchical tree to which each textual value belongs can be explored to retrieve possible generalizations.

WordNet [11] is a freely available lexical database that describes and organizes more than 100,000 general English concepts, which are semantically structured in an ontological fashion. It contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (*synsets*), each expressing a distinct concept (i.e. a word sense). Synsets are linked by means of conceptual-semantic and lexical relations such as synonymy, hypernymy (subclass-of), meronymy (part-of), etc. The result is a network of meaningfully related words, where the graph model can be exploited to interpret concept's semantics. Hypernymy is, by far, the most common relation, representing more than an 80% of all the modelled semantic links. The maximum depth of the noun hierarchy is 16. Polysemous words present an average of 2.77 synsets (i.e. they belong to almost three different hierarchies) and up to 29 different senses (for the "line" word).

Considering those dimensions, the use of WordNet instead of VGHs as semantic background for data anonymization would result in a generalization space which size would be several orders of magnitude bigger. In fact, as most of the related works make generalizations in an exhaustive fashion,

the generalization space is exponentially large according to the depth of the hierarchy, the branching factor, the values and the number of attributes to consider. So, those approaches are computationally too expensive and hardly applicable in such a big ontology like WordNet.

In order to be able to exploit the advantages that big ontologies like WordNet provide with respect to semantics, we present a heuristic global masking method based on the fusion of values of semantically similar records. In our method, each non-anonymous record in the input dataset will be iteratively substituted by another one according to a semantically-grounded metric (see section 3.1) until, by repetition, the desired degree of *k*-anonymity is fulfilled. As we bound the search space for possible substitutions to the number of different records in the input data, our method is able to scale well in such a big ontology regardless the total number of attributes, while minimizing the loss of semantics thanks to the semantically-driven substitution process. Moreover, on the contrary to related works based only on the substitution of the sensible values for more general ones, in our method, other semantically similar concepts (such as hierarchical siblings or specializations) would be also considered.

3.1 Guiding the masking of data

As stated above, the goal of an anonymization method is finding a transformation of the original data, which satisfies *k*-anonymity while minimizing the information loss and, in consequence, maximizing the utility of the resulting data. In order to guide the masking process towards the transformation that would result in the minimum information loss, a metric that evaluates, according to a certain dimension, the difference between the original data and the data resulting from each transformation is needed.

In the literature, various metrics have been proposed and exploited [3], [6], [14], [15], [17], [20]. Classical metrics, such as Discernibility Metric (DM) [3], are used to evaluate the distribution of *m* records (corresponding to *m* individuals) into *g* groups of identical values, generated after the anonymization process. Concretely, DM assigns to each record a penalty based on the size of the group g_i to which it belongs after the generalization (1). A uniform distribution of values in equally sized groups would optimize this metric.

$$DM = \sum_{i=1}^{n} |g_i|^2 \tag{1}$$

However, metrics based on data distribution do not capture how semantically similar the anonymized set is with respect to the original data. As stated in the introduction, preservation of semantics when dealing with textual attributes is crucial in order to be able to interpret and exploit anonymized data. In fact, this aspect is, from the utility point of view, more important than the distribution of the

anonymized dataset when aiming to describe or understand a record by means of its attributes (this will be tested in the evaluation section).

In order to minimize the loss of semantics between original and anonymized datasets, we rely on the theory of *semantic similarity* [22]. Semantic similarity measures the taxonomical alikeness between words based on the semantic evidences extracted from one or several knowledge sources. In the literature, we can distinguish different approaches to compute semantic similarity according to the techniques employed and the knowledge exploited to perform the assessment. The most classical approaches exploit the graph model of structured representations of knowledge as the base to compute similarities. Typically, subsumption hierarchies and, more generally, ontologies, have been used for that purpose, as they provide a directed graph in which semantic interrelations are modelled as links between concepts. Many edge-counting approaches have been developed to exploit this geometrical model, computing word similarity as a function of concept inter-link distance [23], [24], [25]. There exist other approaches which also exploit domain corpora to complement the knowledge available in the ontology and estimate concept's Information Content (IC) from term's appearance frequencies. Even though the latter are able to provide accurate estimations when enough data is available [22], their applicability is hampered by the availability of this data and their pre-processing. On the contrary, edge-counting measures introduced above are characterized by their simplicity (which result is a computationally efficient solution), and their lack of constraints (as only an ontology is required) which ensures their applicability. Due to these reasons, we will rely on edge-counting metrics to guide the masking process in order to maximize the semantic similarity between the original data and those resulting from the masking of record tuples.

In order to provide accurate results, edge-counting measures rely on WordNet's is-a taxonomy to estimate the similarity. Such a general and massive ontology, with a relatively homogeneous distribution of semantic links and good inter-domain coverage is the ideal environment to apply those measures [22].

The simplest way to estimate the semantic distance (i.e. the inverse to similarity) between two ontological nodes (*c1* and *c2*) is by calculating the shortest Path Length (i.e. the minimum number of links) connecting these elements (2) [23].

$$distance_{Path\_length}(c_1, c_2) = \min \#\ of\ is-a\ edges\ connecting\ c_1\ and\ c_2 \quad (2)$$

In order to normalize this distance, Leacock and Chodorow [24] divided the path length between two concepts (N_p) by the maximum depth of the taxonomy (D) in a non-linear fashion (3). The function is inverted to measure similarity.

$$similarity_{l\&c}(c_1, c_2) = -\log(N_p / 2D) \quad (3)$$

However, those measures omit the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered as less similar than those belonging to a lower level, as they present different degrees of generality. Based on this premise Wu and Palmer's measure [25] also takes into account the depth of the concepts in the hierarchy (4).

$$similarity_{w\&p}(c_1, c_2) \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (4)$$

, where N_1 and N_2 are the number of is-a links from c_1 and c_2 respectively to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

As Wu and Palmer's measure incorporates more semantic features that the other measures (i.e. absolute path length normalized by relative depth in the taxonomy) we have taken it as the metric to measure semantic similarity during the anonymization process.

3.2 An ontology-based method to mask textual attributes

Our method addresses the problem of masking a subset of the textual attributes of the input record set in a global manner. As it has been said in the introduction, four different types of attributes are distinguished: identifiers, quasi-identifiers confidential and non-confidential. Only the first two may lead to the re-identification of individuals. Identifiers are directly removed from the dataset because they refer to values that are unique for each individual (e.g. personal identification number or social security number). As a consequence, the masking process would be applied over tuples of textual quasi-identifier attributes.

As explained above, exhaustive generalization methods are computationally too expensive to be applicable with unbounded textual attributes and large ontologies like WordNet. Moreover, the fact that values to anonymize correspond to leafs of the VGH implies that values are only substituted by more general ones (which unnecessarily imposes constraints on the space of valid transformations).

Our approach deals with the global masking process in a different manner. Thanks to the wide coverage of WordNet, one would be able to map textual attribute values into ontological nodes which do not necessary represent leafs of a hierarchy. As a result, semantically related concepts can be retrieved going through the ontological hierarchy/ies to which the value belongs. Those ontological hierarchies are designed in a much general and fine grained fashion than ad-hoc VGHs and, according to the agreement of domain knowledge experts, not in function on the input data. Those facts open the possibility of substituting values by a much wider and knowledge-coherent set of semantically similar elements. In order to ensure the scalability with regards on the ontology size and the input data, we

bound the space of valid value changes to the set of value combinations that are present in the input dataset. When changing a value of a record for another, one may represent a taxonomical subsumer to the other (which is the only case covered by generalization method) but also a hierarchical siblings (with the same taxonomical depth) or a specialization (located in a lower level). In fact, in many situations, a specialization may be more similar than a subsumer because, as stated in section 3.1, because concepts belonging to lower levels of a hierarchy have less differentiated meanings due to their higher concreteness. As a result, the value change would result in less information loss and a higher preservation of data utility from a semantic point of view. This is an interesting characteristic and an improvement over the more restricted data transformations supported by VGH-based generalization methods.

In a nutshell, the method proposed is based on the fusion of quasi-identifier values of each record with the values of another record. In order to select the value that minimizes the information loss resulting from the data substitution, a semantic metric (section 3.1) is used to select the most similar one. As a result of the fusion, quasi-identifier values for both records (the one to anonymize and the most semantically similar one) will take the same values and will become indistinguishable; so, the *k*-anonymity level for both records will increase. By repeating the process iteratively for each non anonymous record according to a certain value of *k*-anonymity, the input dataset will be anonymized.

In order to formally present the method, we introduce some definitions.

Let us take an $m \times n$ data matrix, D, where each of the m rows corresponds to the record of a different respondent and each of the n columns is a textual quasi-identifier attribute. Let us name D^A the anonymized version of D. And let us define the records belonging to the original data matrix as $r_i = \{r_{i1},...,r_{in}\}$ and the records of the anonymized version as $r_i^A = \{r_{i1}^A,...,r_{in}^A\}$, where r_{ij} and r^A_{ij} are attribute values for each record.

Definition 1. A set of *indistinguishable records* with respect to a given record r_i is defined as $I(r_i) = \{r_k \,|\, r_{kj} = r_{ij} \; \forall j = 1..n\}$. That means that two records are indistinguishable if they have exactly the same value for all of their quasi identifier attributes. Let us call $\Psi = \{I_1,...,I_p\}$, the set formed by sets of indistinguishable records.

Definition 2. A set indistinguishable records I_l is considered *anonymous* (*A*) iff $|I_l| \geq k$ (i.e, it contains at least *k* elements, where *k* is the level of anonymity). Then, $\Lambda = \{A_1,..., A_q\}$ is the group of anonymous sets of records built from the dataset *D*.

Definition 3. The *similarity between two records* r_i and $r_k \in D$ is defined as the mean of the semantic similarity of each of their attribute values as follows:

$$record\_similarity(r_i, r_k) = \frac{\sum_{j=1}^{n} sim_{sem}(r_{ij}, r_{kj})}{n} \quad (5)$$

, where for each attribute value pair, the function sim_{sem} can be any of the semantic similarity measures presented in section 3.1. As stated before, in this paper, we choose Wu & Palmer similarity (eq. 4) for testing purposes.

Definition 4. Let us consider a record r_i such that $\forall A_l \in \Lambda,\ r_i \notin A_l$ (i.e. it is not anonymous). Then, we maximum similarity with regards to any other record available in D will represent the *quality of the best data transformation* for that record.

$$best\_quality(r_i) = \max(record\_similarity(r_i, r_k)) \qquad \forall r_k \in D \quad (6)$$

Definition 5. The *minimum degree of anonymity achievable* with the fusion of the values of a record r_i with respect to any other record r_k available in D is given by:

$$min\_achievable\_anonymity(r_i) = \min(|I(r_i) \cup I(r_k)|) \qquad \forall r_k \in D \quad (7)$$

Definition 6. The *quality* of D^A with regard to D from a semantic point of view is defined as the inverse of the information loss derived from the transformation of D in its anonymized version D^A. Information loss is usually given by the absolute difference [26], so the quality is measured in terms of semantic similarity (sim_{sem}).

$$semantic\_quality(D^A) = \sum_{i=1}^{m}\sum_{j=1}^{n} sim_{sem}(r_{ij}, r^A{}_{ij}) \quad (8)$$

This value can be normalized in the range of the sim_{sem} values by dividing it by the total number of records (m) and the total number of attributes (n)

$$norm\_semantic\_quality(D^A) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} sim_{sem}(r_{ij}, r^A{}_{ij})}{m * n} \quad (9)$$

Based on a semantic similarity measure, which evaluates the quality of the best data transformation, our method aims to find the best value fusion between records that leads to a partition formed by anonymized record sets (i.e. $\forall r_i \in D\ \exists A_l \in \Lambda, r_i \in A_l$). The optimum anonymous partition is the one that maximizes the utility of the data, by preserving the meaning of the values. In our case, this is a

partition that minimizes the information loss from a semantic point of view, which is calculated with eq. 9.

As noted in section 2, finding the optimum anonymous partition requires the generation of all the possible value fusions for all the non-anonymous records, which has an exponential cost. In order to ensure the scalability of our approach, we opted for a greedy algorithm which selects, at each iteration, a set of indistinguishable records (I_l) and finds a feasible value fusion. However, with an uninformed approach, the quality of the result would depend on the selection of the records at each step. To solve this, an exhaustive method that tests all the combinations can be used, with a factorial cost with respect to the number of non-anonymous records. This approach is again computationally too expensive because, as records are defined by unbounded textual attributes, they usually correspond to a high number of combinations, many of them being unique, leading to a high amount of records not fulfilling *k*-anonymity. In order to ensure the scalability of the method and guide the anonymization towards a minimization of information loss, we have designed several heuristics (H) that permit the select, at each iteration, the best set of indistinguishable records (I_l) to transform:

H_1) From D, select the group of sets of indistinguishable records $S_1 \subseteq \Psi$ whose record value tuples have the lowest number of repetitions in the original set. That is the ones with minimum $|I_i|$, which correspond to the least anonymous ones.

H_2) From S_1, select a subset $S_2 \subseteq S_1$ that contains sets of indistinguishable records for whom the best merging of values leads to the minimum semantic information loss. The aim is to maximize the quality of the anonymized dataset of the result at each iteration. That is the $I(r_i)$ with maximum $best\_quality(r_i)$.

H_3) From S_2, select the subset $S_3 \subseteq S_2$ for which the minimum achievable degree of anonymity of their records (after the transformation) is lower. That is the $I(r_i)$ that minimize $min\_achievable\_anonymity(r_i)$. In this way, the records that are more difficult to anonymize are prioritized, as they will require more value fusions.

Those criteria are applied in the order indicated above. In this way, if the set S_1 obtained with H_1 contains more than one element, we apply H_2 to S_1. In the same way, if the resulting set S_2 obtained with H_2 has not a unique element then H_3 is applied. Through tests performed over real data, those three criteria are enough to obtain a unique $I(r_i)$ whose values are merged with the ones of the $I(r_k)$ that allows the maximization of $best\_quality(r_i)$, increasing the *k*-anonymity level of both $I(r_i)$ and $I(r_k)$. However, if using those three criteria it was not possible to find a unique I, a random one in S_3 would be selected.

Algorithmically, the method works as follows:

Algorithm

```
Inputs: D (dataset), k (level of anonymity)
Output: D^A (a transformation of D that fulfils the k-anonymity level).
1    D^A := D
2    min_repetitions := min |I(r_i)| for all r_i ∈ D^A
3    while (min_repetitions < k) do
4       S_1 := set of I(r_i), r_i ∈ D^A with |I(r_i)|=min_repetitions
5       S_2 := set of I(r_i) ∈ S_1 with maximum best_quality(r_i)
6       S_3 := set of I(r_i) ∈ S_2 with minimum min_achievable_anonymity(r_i)
7       Take an I(r_i) randomly from S_3
8       Find a I(r_k), r_k ∈ D^A so that r_k = argmax(record_similarity(r_i, r_k))
9       for all (r_i ∈ I(r_i)) do
10         r_ij := r_kj    ∀j=1..n
11      min_repetitions := min |I(r_i)| for all r_i ∈ D^A
12   end while
13   output D^A
```

As a result of the iterative process, a dataset in which all records are at least *k*-anonymous is obtained (i.e. $\forall r_i \in D \; \exists A_l \in \Lambda, r_i \in A_l$).

With this method, the cost of the anonymization is O(p^3), being p the number of different records in the dataset ($p \leq m$). In fact, the computationally most expensive step is the calculation of the semantic similarity between all the pairs of different records that is required in step #5 in order to find the subset with maximum *best_quality*(r_i). Since each record has n values, this operation requires to execute $n \cdot p^2$ times the semantic similarity between a pair of single values. In the worst case, we require p iterations to build the valid partition (loop in line #3), so the final cost of the algorithm is $n \cdot p^2 \cdot p = n \cdot p^3$ times, with n being a relative small number when compared with p, because the set of quasi-identifier attributes is usually small.

For big datasets, where p can be large due to the unbound nature of values, the scalability is more critical. For this reason we have optimized the implementation. Notice that the semantic similarity between records is measured in line #5 to calculate *best_quality*(R) and again in line #8 to find the most similar record, and repeated at each iteration. As the set of different attribute values and distinct

record tuples is known a priori and does not change during the masking process (unlike for generalization methods), it is possible to pre-calculate and store the similarities between all of them. This avoids repeating the calculus of the similarity for already evaluated value pairs and, more generally, register pairs. In this manner, the calculation of the similarity measure is executed a priori only $n \cdot p^2$ times, leading to an efficiency for the most expensive function of $O(p^2)$. As it will illustrated in the evaluation section, with this modification the execution of the algorithm stays in the range of milliseconds for hundred-sized datasets.

It is important to note that the computational cost of our algorithm uniquely depends on the number of different tuples (p), unlike the related works that depend on the total size of the dataset (m), and on the depth and branching factor of the hierarchy (which represent an exponentially large generalization space of substitutions to evaluate).

4 Evaluation

We have evaluated the proposed method by applying it to a dataset consisting on answers to polls made by the "Observatori de la Fundació d'Estudis Turístics Costa Daurada" at the Catalan National Park "Delta de l'Ebre". Visitants were requested to respond several questions regarding the main reasons and preferences when visiting the park. Each record, which corresponds to an individual, includes a set of textual answers expressed by means of a noun phrase (with one or several words). Due to the variety of answers, the disclosure risk is high and, therefore, individuals are easily identifiable. So, we consider textual answers as quasi identifiers which should be anonymized.

The dataset is composed by 975 individual records, for which we considered two textual attributes per record as quasi-identifiers. From the combination of those attributes, a total of 211 different responses were identified, being 118 of them unique (i.e. identifiers). Fig. 1 shows the distribution of values for the pair of attributes according to their degree of repetition. Note that this sample represents a much wider and heterogeneous test bed than those reported in related works [5], [15], which are focused on bounded categorical values.

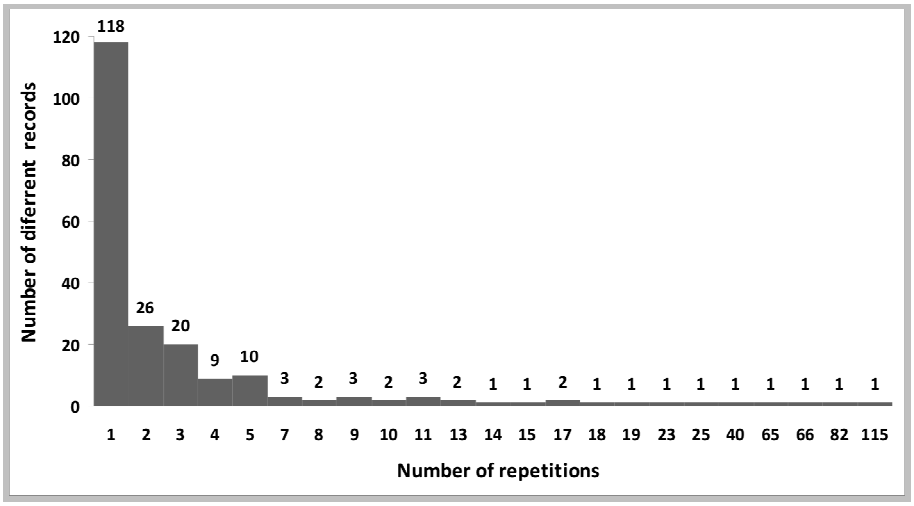


Fig. 1. Attribute distribution according to answer repetitions

The answers values for those two attributes are general and widely used concepts (i.e. sports, beach, nature, etc.), all of them have been found in WordNet 2.1, which permits to use this ontology for performing the semantic similarity measurement. However, as we are dealing with values represented by text labels, it was necessary to morphologically process them in order to detect different lexicalizations of the same concept (e.g. singular/plural forms). We apply the Porter Stemming Algorithm [27] to both text labels of attributes and ontological labels in order to extract the morphological root of words and to be able to map values to ontological concepts and to detect conceptually equivalent values in the dataset.

4.1 Evaluation of the heuristics

In this section, we evaluate the contribution of each of the designed heuristics in guiding the substitution process towards minimizing the information loss form a semantic point of view (as detailed in section 3). The quality of the masked dataset has been evaluated by measuring the information loss according to how semantically similar the masked values are, in average, with respect to the original ones. Information loss has been computed and normalized as defined in eq. 9. The same evaluation was repeated for different levels of *k*-anonymity.

In order to show the contribution of each heuristic in minimizing the information loss of the results, we replaced the heuristic substitution by a naïve replacement that changes each sensible record by a random one from the same dataset. Following the same basic algorithm presented in section 3, each random change would increase the level of *k*-anonymity until all records are anonymized. For the random substitution, records are ordered alphabetically, in order to avoid depending on the initial order of data. The results obtained for the random substitution are the average of 5 executions. The three heuristics proposed in section 3.2, were gradually introduced instead of the random substitution,

in a way that permits to quantify the contribution of each one in the results' quality. The results of this test are shown in Fig. 2: considering no heuristic at all, only the first one, only the first and the second one, and all three together.

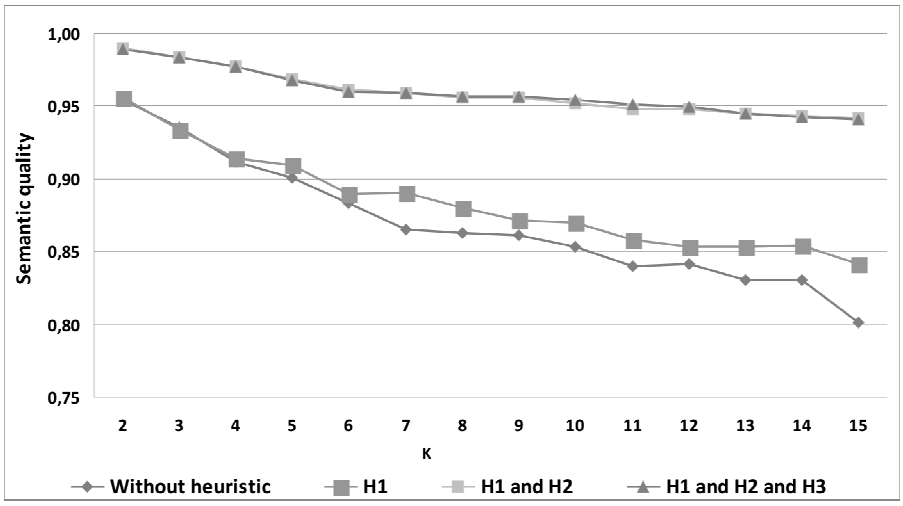


Fig. 2. Contribution of each heuristic to the anonymized dataset quality

Results reflected in Fig. 2 are coherent to what it is expected from the design of each heuristic. The first one, which only re-orders input data according to the degree of record repetition in order to prioritize the less anonymous records, produces a slight improvement over the complete random substitution. The second one, which incorporates the semantic similarity function as a metric to guide the value fusion process towards the minimization of the semantic loss produces the most significant improvement. The incorporation of the third heuristic produces a very slight improvement in some situations, as it is only executed in case of tie (i.e. when there exists several replacements with an equal value of maximum similarity, which is a quite scarce situation).

As a result of the heuristic fusion process, our approach is able to improve the naïve replacement by a considerable margin. This is even more noticeable for a high *k*-anonymity level (above 5), when using the three heuristics we clearly outperform the semantic loss of the random version. This is very convenient and shows that our approach performs well regardless the desired level of privacy protection.

4.2 Comparing semantic and distributional approaches

In order to show the importance of a semantically focused anonymization, we compared it with a more traditional schema, focused on the distributional characteristics of the masked dataset (as stated at the beginning of section 3.1). This has been done by using the Discernibility metric (eq. 1) in our algorithm instead of the Wu and Palmer's measure as metric, in order to guide the masking process.

Both semantic and distributional approaches have been compared by evaluating the semantic difference between the original and masked dataset as stated in eq. 9 - see Fig. 3- and also by computing the Discernibility penalty of the results with respect to the original data (as stated in eq. 1, section 3.1) - see Fig. 4-

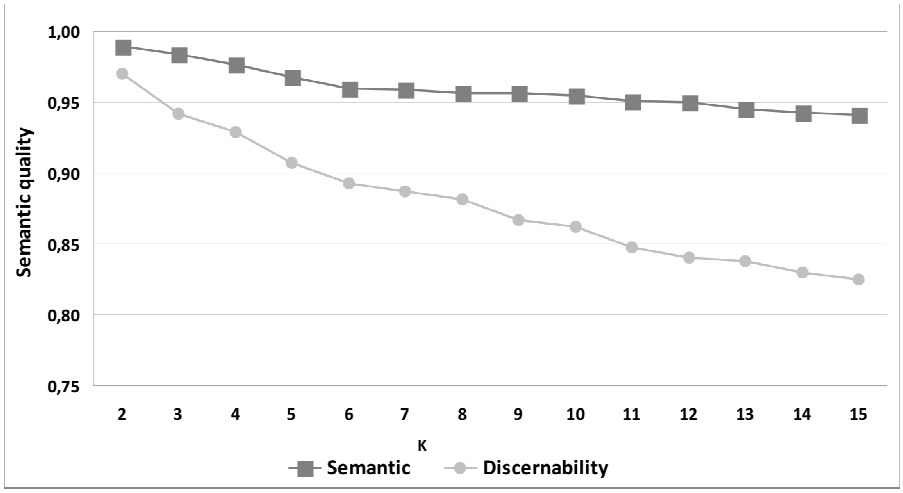


Fig. 3. Similarity against original data for semantic and distributional anonymizations.

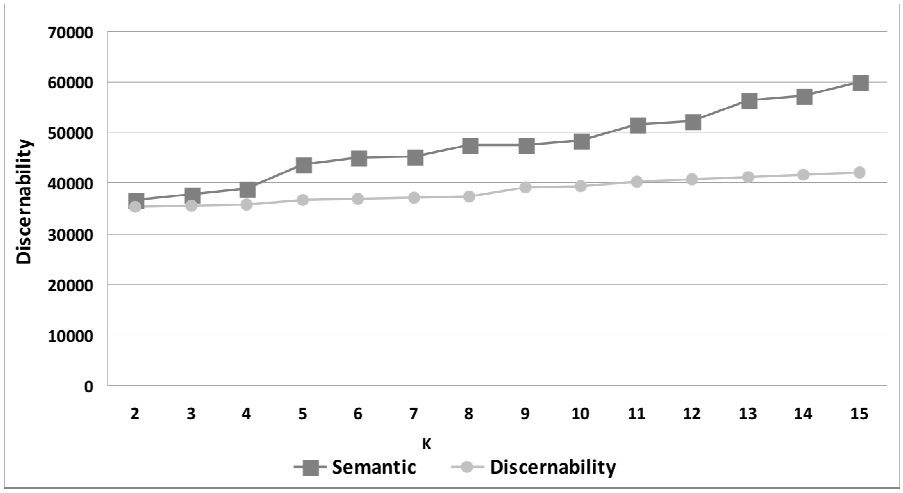


Fig. 4. Discernibilty penalty against original data for semantic and distributional anonymizations.

The figures show that the optimization of the dataset distribution and the preservation of records' semantics are not correlated. In fact, there exists a very noticeable semantic loss in the resulting dataset for *k*-anonymity values above 5 for the distributional approach. As stated in the introduction, the utility of textual information from the data analysis point of view is highly dependent on its semantics. One can see that classical approaches focused on providing uniform groups of masked values may significantly modify dataset's meaning, hampering their exploitation.

4.3 Evaluation of data utility for semantic clustering

In order to evaluate the hypothesis that a semantic-driven anonymization retains better the utility of the original data than distributional approaches from the data exploitation point of view, we next compared the utility of the dataset resulting from both approaches in a concrete data mining application.

As stated in the introduction, data acquired by statistical agencies are of great interest for data analysis in order to, for example, extract user profiles, detect preferences or perform recommendations [2]. Data mining and, more concretely, clustering algorithms are widely used for organizing and classifying data into a number of homogenous groups. Even though clustering algorithms have been traditionally focused on numerical data or bounded categorical data, the increase in volume and importance of textual data have motivated authors in developing semantically grounded clustering algorithms [28].

In [29] it is presented a hierarchical clustering algorithm which is able to interpret and compare both numerical and textual features of objects. In a similar manner as in the present work, ontologies are exploited as the base to map textual features to semantically comparable concepts. Then, concepts' alikeness is assessed by means of semantic similarity measures. According to those similarities, an iterative aggregation process of objects is performed based on the Ward's method [30]. As a result, a hierarchical classification of non-overlapping sets of objects is constructed from the evaluation of their individual features. The height of the internal nodes in the resulting dendogram reflects the distance between each pair of aggregated elements.

By means of this algorithm, and using WordNet as the background ontology, we evaluated the utility of data from the semantic clustering point of view. We compare the clusters obtained from the original dataset against those resulting from the execution of the clustering process, both for distributional (i.e. discernibility-based) and semantic (i.e. Wu and Palmer's similarity-based) anonymization procedures. A *k*-anonymity level of 5 has been chosen for this comparison, as it is a moderate privacy level that would be able to still retain data utility.

By quantifying the differences between the clusters obtained from original data against both masking methods, we are able to conclude which one retains better the semantics and, in consequence, the utility of data. Resulting clusters can be compared by means of the distance between partitions of the same set of objects as defined in [31]: considering two partitions of the same data set (in this case, the original an anonymized versions), being P_A a partition whose clusters are denoted as A_i and P_B a partition whose clusters are denoted as B_j, the distance is defined as:

$$d_{Part}(P_A, P_B) = \frac{2 * I(P_A \cap P_B) - I(P_A) - I(P_B)}{I(P_A \cap P_B)} \quad (10)$$

, where $I(P_A)$ is the average information of P_A which measures the randomness of the distribution of elements over the n classes of the partition (similarly for and $I(P_B)$), and $I(P_A \cap P_B)$ is the mutual average information of the intersection of two partitions. They are computed as

$$I(P_A) = -\sum_{i=1}^{n} P_i \log_2 P_i \quad (11)$$

$$I(P_B) = -\sum_{i=1}^{m} P_j \log_2 2P_j \quad (12)$$

$$I(P_A \cap P_B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log_2 P_{ij} \quad (13)$$

, where the probabilities of belonging to the clusters are $P_i=P(A_i)$, $P_j=P(B_j)$, and $P_{ij}=P(A_i \cap B_i)$. Distance values are normalized in the 0..1 interval, where 0 indicates identical clusters and 1 maximally different ones.

The distance between the original clusters and those obtained from both masking approaches are the following.

Table 1. Distances between the different clustering results

	Distance
Original data vs. Semantic anonymization	0.26
Original data vs. Distributional anonymization	0.57
Semantic vs. Discernibility anonymizations	0.56

It is easy to see how a semantically-driven anonymization results in a dataset that better retains the semantics of the original data (i.e. less information loss) than distributional approaches, with a distance in the resulting classification with respect to the original data of 0.26 and 0.57, respectively. In consequence, conclusions extracted from the analysis of semantically anonymized data would be more similar to those obtained from the original data when using the approach presented in this paper. It is also relevant to observe the big differences between clusters resulting from each anonymization schema, whose distance is a significant 0.56. This shows a high discrepancy in the way in which records are fused according to the different quality metrics. This result is coherent to what was observed in section 4.2, in which semantic and distributional anonymizations were significantly uncorrelated.

4.4 Record linkage

Data utility is an important dimension when aiming to anonymize data and minimize the information loss. However, from the privacy preserving point of view, disclosure risk should be also minimized. The latter may be measured as a function of the probability of re-identification of the masked dataset with respect to original data.

In order to evaluate the disclosure risk of both semantically and distributionally anonymized datasets, we computed the level of *record linkage* (also named re-identification) [32] of the results. Record linkage (RL) is the task of finding matches in the original data from the anonymized results. The disclosure risk of a privacy preserving method can be measured as the difficulty of finding correct linkages between original and masked datasets. It is typically calculated as the percentage of correctly linked records [32]:

$$RL = \frac{\sum_{i=1}^{m} P_{rl}(r_i^{\,A})}{m} \cdot 100 \tag{14}$$

, where the record linkage probability of an anonymized record $P_{rl}(r_i^{A})$ is calculated as follow:

$$P_{rl}(r_i^{\,A}) = \begin{cases} 0 & if\ r_i \notin L \\ \dfrac{1}{|L|} & if\ r_i \in L \end{cases} \tag{15}$$

, where r_i is the original record, $r_i^{\,A}$ is the anonymized record and L is the set of original records in D that match with $r_i^{\,A}$ ($L \subseteq D$). As we deal with textual features and value changes, record matching is performed by simple text matching of all individual attributes (in the same order). So, each $r_i^{\,A}$ is compared to all records of the original dataset D by text matching, obtaining the L set of matching records. If r_i is in L, then, the probability of record linkage is computed as the probability of finding r_i in L (i.e. the number of records in L). On the contrary, if the r_i is not in L, the record linkage probability is 0.

We have calculated the record linkage percentage for different levels of *k*-anonymity, comparing the original registers with respect to the semantic anonymization and afterwards with the distributional version of the method. The RL probabilities are represented in Fig. 5.

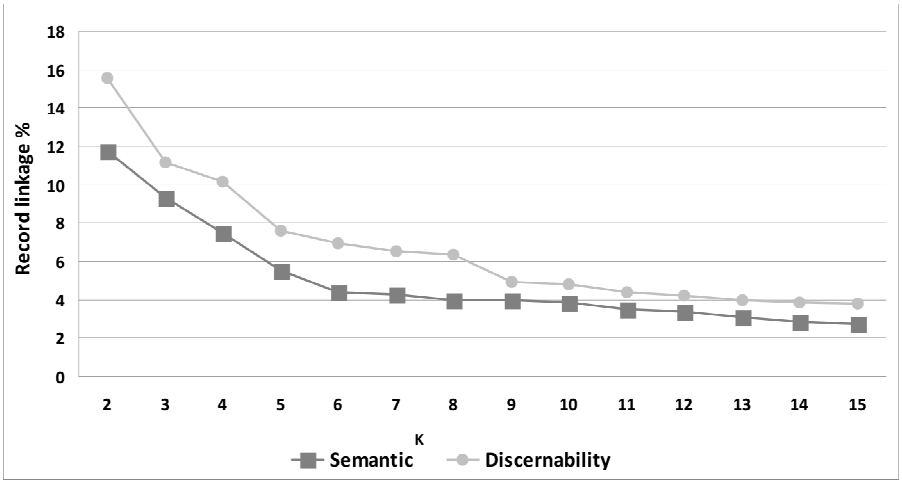


Fig. 5. Record Linkage percentage for semantic and dicernability-based anonymizations.

Both approaches follow a similar tendency, decreasing as *k* increases. It can also be seen that the degree of record linkage is quite stable for *k* values of 5 and above. The main difference is that our method gives lower probabilities of record re-identification than a distributional approach, especially for small values of *k*. This permits, in comparison to the distributional approach, to lower the *k*-anonymity degree (resulting in less information loss), while maintaining a comparable level of disclosure risk.

In conclusion, results show that an anonymization process focused on the preservation on data semantics does not contradicts the goal of a privacy preservation method which is to minimize the disclosure risk.

4.5 Execution time study

From a temporal perspective, executing our method over a 2.4 GHz Intel Core processor with 4 GB RAM, the run time of the anonymization process for the test dataset ranged from 1.2 to 1.6 seconds (according to the desired level of *k*-anonymity) as shown in Fig. 6. The pre-calculus of the semantic similarities between all value pairs of each attribute in the dataset lasted 6.33 minutes.

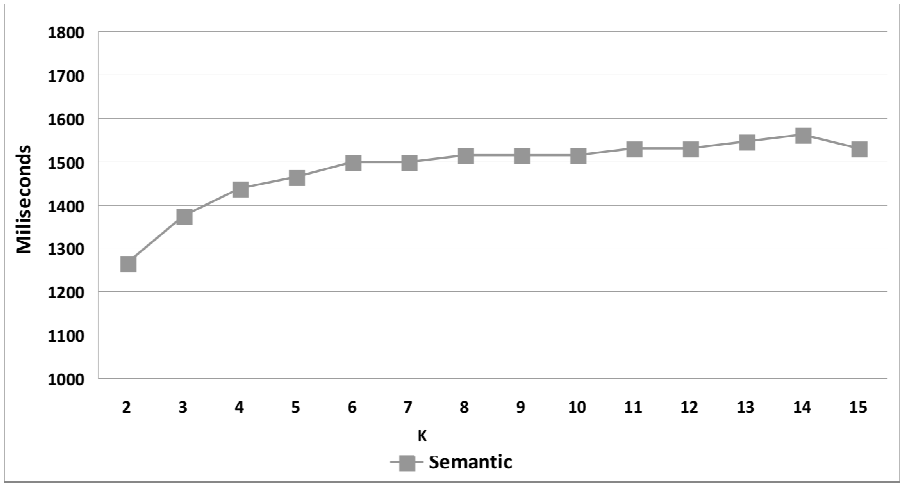


Fig. 6. Anonymization process runtime according to the level of k-anonymity

One can easily see how, as stated in section 3.2, similarity computation represents the most computationally expensive function, and how the minimization of the number of calculus results in a very noticeable optimization of runtime.

Run times are also much lower than those reported by related works that need several hours [6], [15] to perform the anonymization of the data, even for generalization schemas and very limited VGHs and bounded categorical data (3-4 levels of hierarchical depth and an average of a dozen of values [15]). On the contrary, we were able to mask much bigger and fine grained data in much less time while considering and big and wide ontologies like WordNet, with thousands of concepts and a maximum depth of 16 levels (as explained in section 3). This shows the scalability of our method for large and heterogeneous textual databases.

5 Conclusions

Anonymization of textual attributes deals with two, a priori, confronted aspects of information: on one hand, the minimization of the disclosure risk by fulfilling a desired level of *k*-anonymity and, on the other hand, the maximization of data utility in order to properly exploit them. Previous approaches neglected or very shallowly considered the semantic content of textual attributes. As discussed in this paper, the meaning of data is an important dimension when aiming to make an analysis of the anonymized results to extract useful knowledge, as it is required in data mining, decision making or recommendation processes.

Micro-aggregation is the most common masking method applied to non-numerical data [Domingo-Ferrer]. It builds groups of *k* similar registers and substitutes them by their prototype to assure the *k*-

anonymity property. However, the application of this method to textual attributes is not straightforward because of the limitations on defining appropriate averaging operators for this kind of unbounded values. Most of related works aggregate data by using a generalization approach relying on ad-hoc hierarchical structures. Due to their limitations both from the semantic background and efficiency points of view, in this paper, we have proposed an alternative way to aggregate the individually identifiable records into indistinguishable groups fulfilling *k*-anonymity by means of the fusion of semantically similar values.

This global masking method is based on the exploitation of wide and general ontologies in order to properly interpret the values from a conceptual point of view, rather than from a symbolic one. The algorithm uses several heuristics to guide the search on the set of possible value fusions towards the preservation of the semantics of the dataset. This has been demonstrated with different tests, performed with real textual data obtained from visitors of a Catalan National Park. The results indicate that, in comparison with a classical approach based on the optimization of the distribution of the data, our approach better retains the quality and utility of data from a semantic point of view. This has been reflected when exploiting masked data with by means of a clustering process, for which we were able to obtain the most similar set of classes with respect to the original data.

Finally, special care has been put in ensuring the applicability and scalability of the method when dealing with large and heterogeneous textual data. By enabling the exploitation of already available ontologies we avoid the necessity of constructing ad-hoc hierarchies according to data labels like VGH-based schemas, which supposes a serious cost and limits the method's applicability. In addition, the non-exhaustive heuristic algorithm based on constrained value substitutions permitted to achieve a good scalability with regards to the size, heterogeneity and number of attributes of input data and with respect to the size, depth and branching factor of the ontology.

In the future we would like to study the behaviour of the method with respect to other ontologies, with different size and concreteness degrees (such as domain-specific ontologies, which could be exploited when input data refers to concrete domain terminology). We would also study the possibility of combining several ontologies as background knowledge in order to complement knowledge modelled for each of them.

Acknowledgements

Thanks are given to "Observatori de la Funcació d'Estudis Turístics Costa Daurada" and "Parc Nacional del Delta de l'Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)" for providing us the data collected from the visitors of the park. This work is supported the Spanish

MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02). Sergio Martínez Lluís is supported by the Universitat Rovira i Virgili predoctoral research grant.

References

[1] Giessing, S., Survey on methods or tabular data protection in Argus, in: Privacy in Statistical Databases. Lecture Notes in Computer Science, 3050 (2004) 1-13.
[2] Domingo-Ferrer, J. A survey of inference control methods for privacy-preserving data mining, in Privacy-Preserving Data Mining: Models and Algorithms, eds. C.C. Aggarwal and P.S. Yu, Advances in Database Systems, v.34, N.Y.: Springer Verlag, (2008) 53-80.
[3] Sweeney, L. *k*-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5) (2002) 557-570.
[4] Bayardo, R. J., Agrawal, R. Data privacy through optimal *k*-anonymization. Proceedings of the 21st International Conference on Data Engineering (ICDE) (2005) 217-228.
[5] Samarati, P., Sweeney, L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)
[6] Xu, J., Wang, W., Pei, J., Wang X., Shi, B., Wai-Chee Fu, A. Utility-Based Anonymization for Privacy Preservation with Less Information Loss. ACM SIGKDD Explorations Newsletter V.8 I.2 (2006) 21-30.
[7] Eric G. Little, Galina L. Rogova. Designing ontologies for higher level fusion Information Fusion Volume 10, Issue 1, January (2009) 70-82
[8] Mieczyslaw M. Kokar, Christopher J. Matheus, Kenneth Baclawski. Ontology-based situation awareness Information Fusion Volume 10, Issue 1, January (2009) 83-98
[9] Cimiano, P. Ontology Learning and Population from Text. Algorithms, Evaluation and Applications. Springer-Verlag (2006)
[10] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, Swoogle, J.: A Search and Metadata Engine for the Semantic Web. In Proc. 13th ACM Conference on Information and Knowledge Management, ACM Press (2004) 652-659.
[11] Fellbaum, C. WordNet: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press. (1998)
[12] Spackman KA, Campbell KE and Cote RA. Snomed-RT: A reference terminology for healthcare. Journal of the American Medical Informatics Association (Special issue) (1997) 640-644.
[13] Nelson SJ, Johnston D and Humphreys BL. Relationships in Medical Subject Headings. Relationships in the Organization of Knowledge, Kluwer Academic Publishers, New York, (2001) 171-184
[14] Iyengar, V. S. Transforming data to satisfy privacy constraints. Proceedings of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Minig (KDD), (2002) 279-288.
[15] Li, T., Li, N. Towards optimal *k*-anonymization. Data & Knowledge Engineering 65 (2008) 22-39.
[16] Martinez S. Sanchez D. Valls A. Anonymizing Categorical Data with a Recoding Method based on Semantic Similarity. In: Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based System IPMU. Dortmund, Germany (2010)
[17] He, Y., Naughton, J.F. Anonymization of Set-Valued Data via Top-Down, Local Generalization. 35th Int. Conf. VLDB. Lyon, France (2009) Vol. 2. 934-945.
[18] Guo, L., Wu, X. Privacy preserving categorical data analysis with unknown distortion parameters, Transactions on Data Privacy, 2, (2009) 185-205.
[19] Gouweleeuw, J.M. Kooiman, P., Willenborg, L. C. R. J. and DeWolf, P. P. Post randomization for statistal disclousure control: Theory and implementation. Research paper no. 9731 (Voorburg: Statistics Netherlands) (1997)
[20] Lefevre, K., DeWitt, D.J., Ramakrishnan, R. Mondrian Multidimensional K-Anonymity. Proceedings on the 22nd Int. Conf. on Data Engineering ICDE 25 (2006)
[21] Terrovitis, M., Mamoulis, N. Kalnis, P. Privacy-preserving anonymization of set-valued data. In Proc. of VLDB (2008)

[22] Jiang, J., Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In Proc. Int. Conf. on Research in Computational Linguistics, Japan (1997) 19-33.
[23] Rada, R., Mili, H., Bichnell, E., Blettner, M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 9(1), (1989) 17-30.
[24] Leacock, C., Chodorow, M. Combining local context and WordNet similarity for word sense identification. In Fellbaum (ed.), WordNet: An electronic lexical database, MIT Press (1998) 265-283.
[25] Wu, Z., Palmer, M. Verb semantics and lexical selection. In Proc. 32nd annual Meeting of the Association for Computational Linguistics, New Mexico, USA (1994) 133-138.
[26] Domingo-Ferrer, J., Torra, V., Disclosure control methods and information loss for microdata, in: Confidentiality, disclosure and data access,
[27] Porter. An algorithm for suffix stripping, Program, (1980) Vol. 14 no 3, 130-137.
[28] Zengyou He, Xiaofei Xu, Shengchun Deng. k-ANMI: A mutual information based clustering algorithm for categorical data Information Fusion Volume 9, Issue 2, April (2008), Pages 223-233.
[29] M. Batet, A. Valls, K. Gibert, Improving classical clustering with ontologies, in: Proceedings of the 4th World conference of the IASC, Japan, (2008) 137-146.
[30] Ward, J.H. Hierarchical grouping to optimize an objective function, JASA, (1963) 58: 236-244.
[31] López de Mántaras, R.: A distance-based Attribute Selection Measure for decision tree induction. Machine learning, 6 (1991) 81-92.
[32] Torra, V., Domingo-Ferrer, J. Record Linkage methods for multidatabase data mining, in: Information Fusion in Data Mining. Springer (2003) 101-132.

The role of ontologies in the anonymization of textual variables

Sergio MARTÍNEZ[1], David SÁNCHEZ, Aida VALLS and Montserrat BATET
Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) research group
Departament d'Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili
Av. Països Catalans, 26. 43007. Tarragona (Spain)

Abstract. The exploitation of sensible data associated to individuals requires a proper anonymization in order to preserve the privacy. Even though several masking methods have been designed for numerical data, very few of them deal with textual information. During the masking process, information loss should be minimized in order to enable a proper analysis of data with data mining methods. In the case of textual data, the quality of the anonymized dataset is closely related to the preservation of semantics, a dimension which has been only shallowly considered in some previous works, by using small and ad-hoc hierarchies of words. In this work we want to study the use of large and standard ontologies as the base to perform the anonymization of textual variables. We will evaluate the role of ontologies in preserving the utility of the anonymized information when a partition of the objects is done with unsupervised clustering methods. Results show that by exploiting detailed ontologies, one is able to improve the preservation of the data semantics in comparison to approaches based on ad-hoc structures and data distribution metrics.

Keywords. Privacy preserving data mining, statistical disclosure control, ontologies, semantic similarity.

Introduction

Privacy Preserving Data Mining is a new field that has appeared with the increase of the ability to store data from individuals [1]. Third parties can be interested in performing statistical or data mining analysis on this data to obtain new knowledge about the users. Before distributing the data to third parties, a masking method should be used in order to anonymize the data file and minimize the disclosure of private and sensible information (i.e. re-identification risk). The privacy level is typically related to the fulfilment of the *k-anonymity* property [19]. This property establishes that each anonymized record in a data set (i.e. a set of attribute values associated to an individual) has to be indistinguishable with at least *k*-1 other records within the same dataset, according to its individual attribute values.

In order to make the anonymized data as useful as possible from the data mining point of view, it is needed to preserve the utility of the values, which is achieved if the information loss inherent to the masking process is minimized. This is measured by

[1] Corresponding Author.

means of some quality index. Up to this moment, most of the attention has been paid to numerical data. The goal of the masking methods for numerical data is to maintain the statistical characteristics of the dataset [6].

Textual data is typically considered in a simplistic manner as categorical values, representing them as a discrete enumeration of modalities (i.e. bounded vocabulary). Due to the lack of any semantic background, quality metrics in this case are focused on maintaining the probability distribution of the values in the masked file. This has been criticized by several authors [22] as value distribution does not capture important dimensions of data utility. In fact, as textual attributes represent concepts, their utility should be associated to the preservation of their inherent semantics. Moreover, textual data retrieved, for example, from free answers of individuals, represent an unbounded set of values that correspond to concepts with a concrete semantics. In order to properly interpret and compare them, the similarities between their semantics should be taken into consideration.

Due to the ambiguity of human languages and the complexity and knowledge modelling, very few masking methods have considered the semantics of attribute values in some degree. In fact, many approaches [4][18][19] completely ignore this issue, dealing with textual data in a naïve way, proposing arbitrary suppressions or substitutions aimed to fulfil *k*-anonymy and preserve the distribution of the input data. As it will be discussed in section 1, other approaches are based on specific knowledge structures built ad-hoc for the anonymization process [4][10][14][18][19]. Their main drawback is that the masking method makes an exhaustive search on these structures, which hampers their scalability and applicability with big and heterogeneous datasets. To overcome this limitation, in [16] we proposed a non-exhaustive masking method, based on a heuristic search on the knowledge model and a substitution of textual values.

However, these domain-specific structures are manually built by experts with the single purpose of anonymization. In this paper, we want to study the possibility of using more general and knowledge structures as ontologies, which have been built from the consensus of multiple experts. Ontologies offer a formal, explicit and machine readable structuring of a set of concepts by means of a semantic network where multiple hierarchies are defined and semantic relations are explicitly modelled as links between concepts [8]. Thanks to initiatives such as the Semantic Web [5], many ontologies have been created in the last years, bringing the development of general purpose knowledge sources (such as WordNet [7] for English words), as well as domain terminologies (e.g. medical sources such as UMLS).

To study the behaviour of a general ontology with respect to ad-hoc knowledge structures, we will use real textual data obtained from a survey at the National Park "*Delta del Ebre*" in Catalonia, Spain. The quality of the anonymized dataset will be measured according to its utility for data mining. In particular, unsupervised clustering will be performed on the original and the anonymized versions of the data to evaluate the preservation of the semantics. From the results obtained, an analysis of the advantages and drawbacks of each approach will be done, studying specially the role of ontologies for data privacy preserving in data mining.

The rest of the paper is organized as follows. Section 1 reviews methods for privacy protection of categorical data. Section 2 discusses the exploitation of ontologies to aid the masking of textual variables. In section 3, the ontology-based anonymization method is detailed. Section 4 is devoted to evaluate our method and compare it with related works from the data exploitation point of view. The final section contains the conclusions and future work.

1. Related work

As stated above, masking of categorical data is not straightforward due to the textual nature of attribute values. Some basic works consider categorical data as enumerated terms for which only Boolean word matching operations can be performed. We can find methods based on data swapping (which exchange values of two different records) and methods that add of some kind of noise (such as the replacement of values according to some probability distribution [9]). Other authors [18][19] perform local suppressions of certain values or select a sample of the original data aimed to fulfill *k*-anonymity while maintaining the information distribution of input data. In the literature, various metrics have been proposed to measure this data distribution, such as the Dicernability Metric (DM) [4], which evaluates the distribution of n records (corresponding to n individuals) into g groups of identical values, generated after the anonymization process. Concretely, DM assigns to each record a penalty based on the size of the group g_i to which it belongs after the generalization (1). A uniform distribution of values in equally sized groups (with respect to the original data) is the goal.

$$DM = \sum_{i=1}^{n} |g_i|^2 \tag{1}$$

Even though those methods are effective in achieving a certain degree of privacy in an easy and efficient manner, they fail to preserve the meaning of the original dataset, due to their lack of semantic analysis. Due to this reason, in recent years, some authors have incorporated some kind of knowledge background to the masking process.

In previous knowledge-based masking methods, the set of values of each categorical attribute of the input records in the dataset are represented by means of *Value Generalization Hierarchies* (VGHs) [4][6][10][11][13][14][19]. Those are ad-hoc and manually constructed tree-like structures defined according to a given input dataset, where categorical labels of an attribute represent leafs of the hierarchy and they are recursively subsumed by common generalizations.

The masking process consists on, for each attribute, substituting several original values by a more general one, obtained from the hierarchical structure associated to that attribute. This generalization process decreases the number of distinct records in the dataset and, in consequence, increases the level of *k*-anonymity. In general, for each value, different generalizations are possible according to the depth of the tree. The concrete substitution is selected according to a metric that measures the information loss of each substitution with regards to the original data. This metric can be either distributional (as stated above) or based on the minimization of the generalization level of the value substitution with regards to the VGH [13][18][19].

Notice that in this approach VGHs are manually constructed from the set of possible values of each attribute of the input data file. So, human intervention is needed in order to provide the adequate semantic background in which those algorithms rely. If the values in the input data file change, VGHs should be modified accordingly. Even though this fact may be assumable when dealing with reduced sets of categories (e.g. in [14] a dozen of different values per attribute are considered in average), this hampers the scalability and applicability of the approaches, especially when dealing with unbounded textual data (with potentially hundreds or thousands of individual answers).

Moreover, the fact that VGHs are constructed from a given input data produces ad-hoc and small hierarchies. It is common to observe VGHs with three or four levels of hierarchical depth whereas a detailed taxonomy (such as WordNet) models up to 16 levels [7]. So, the space for valid generalizations in a VGH is small and permits to design an anonymization process that considers all the possible generalizations exhaustively. Finally, the quality of the results heavily depends on the structure of VGHs that, due to their limited scope, offer a partial and biased view of each attribute domain. For this reason we want to study the possibility of using ontologies instead of VGHs.

2. Exploiting ontologies to preserve the semantics of masked data

As it has been said, the presented approaches make generalizations in an exhaustive fashion. If the knowledge structures (i.e. VGHs) are large and detailed, the computational cost increases exponentially according to the depth of the hierarchy, the branching factor and the number of values to evaluate. In order to overcome this limitation, we designed a non-exhaustive heuristic method (detailed in the next section) which, bounding the search space according to the input values, is able to scale well in big knowledge structures.

We rely on the theory of *semantic similarity* [12] to compare textual values from a semantic point of view using those taxonomical representations of the concepts. Semantic similarity measures the taxonomical alikeness between words based on the semantic evidences extracted from one or several knowledge sources. In an *is-a* hierarchy, the simplest way to estimate the distance between two concepts c_1 and c_2 is by calculating the length of the shortest *path* (i.e. the minimum number of links) connecting these concepts [17].

However, the minimum path length measure omits the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered as less similar that those belonging to a lower level, as they present different degrees of generality. Based on this premise Wu and Palmer's measure [21] also takes into account the depth of the concepts in the hierarchy (2).

$$sim_{W\&P}(c_1, c_2) \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \tag{2}$$

, where N_1 and N_2 are the number of is-a links from c_1 and c_2 respectively to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

As stated above, ontologies offer a graph model in which semantic interrelations are modeled as links between concepts, in consequence, represent an ideal source from which computing this kind of semantic similarity [12], because the taxonomical relations are better modeled than in VGHs. In this case, attribute values (i.e. words) can be mapped to ontological nodes (i.e. concepts) via simple word-concept label matching so that the hierarchical tree to which each textual value belongs can be explored to retrieve possible generalizations.

In particular, WordNet [7] is a freely available lexical database that describes and organizes more than 100,000 general English concepts, which are semantically

structured in an ontological fashion. WordNet contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (synsets), each expressing a distinct concept (i.e. a word sense). Synsets are linked by means of conceptual-semantic and lexical relations such as synonymy, hypernymy (subclass-of), meronymy (part-of), etc. The result is a network of meaningfully related words, where the graph model can be exploited to interpret concept's semantics. Hypernymy is, by far the most common relation, representing more than an 80% of all the modeled semantic links. The maximum depth of the noun hierarchy is 16. Polysemous words present an average of 2.77 synsets (i.e. they belong to almost three different hierarchies).

Due to the high level of detail and generality of WordNet, in this paper, we will exploit it as the background ontology to assist the anonymization process. It is important to note that, considering WordNet's dimensions, the size of the generalization space would be several orders of magnitude bigger than when using ad-hoc VGHs. So, special care should be put in the anonymization method in order to ensure its scalability.

3. A method for ontology-based anonymization of textual variables

In this section we briefly present a non-exhaustive heuristic method (based on [16]) aimed to offer good scalability in large taxonomies and relying on the concept of semantic similarity to guide the masking process.

Our approach makes a global anonymization of the values of the records of a set of individuals. In order to ensure that value substitutions lead to the fulfillment of the desired degree of privacy, we should substitute sensible values for other ones that increase the level of k-anonymity. This implies that the values of a record that could be de-identified are substituted for (i) new ones, which are semantically "near" to them, or (ii) for another one already existing in the data set. In both cases, the goal is to make both records indistinguishable. As the first option would lead to an enormous set of possible substitutions according to all the semantically related concepts available in an ontology, we opted for the second strategy. As a result, the space of valid substitutions is bounded to the number of different records in the dataset.

The most appropriate record to substitute a non anonymous one is the record that minimizes the semantic distance with respect to the original. So, the semantic similarity measure introduced in section 2 is used to select the best candidate and make the substitution that minimizes the loss of semantic content. We measure the similarity between two records (*r, r'*) as the mean of the individual similarities of each attribute value (4).

$$sim(r, r') = \frac{\sum_{i}^{\#atributes} sim(v_i, v'_i)}{\#\,atributes} \tag{3}$$

where v_i and v'_i are the values of the attribute *i* of the records *r* and *r'* respectively.

As a result the substitution of the values of a sensible record by others, the number of different records is decreased and the *k*-anonymity is increased. The process is repeated until the whole dataset fulfills the desired *k*-anonymity level.

Notice that the order in which the records are selected may produce different anonymization results. The generation of the optimum result implies creating all possible substitution iterations for all sensible records and picking the order that maximizes the quality of the result set. As unbounded textual attributes may usually correspond to a high number of different answers, many of them being unique, the amount of records not fulfilling the *k*-anonymity would be high. Consequently, the cost of generating all the possible combinations is computationally too expensive.

In order to ensure the scalability of our approach, we implemented several heuristics that aim to select, at each step, the substitution that would likely maximize the quality of the result.

The algorithm has the following steps:

1. First, the set of records is ascending ranked according to the number of record repetitions.
2. When the first record (*r*) is the register with the lowest *k*-anonymity, we check if the corresponding record fulfils the *k*-anonymity according to the number of repetitions.
3. If *k*-anonymity is fulfilled, the entire set is anonymized at the desired level and the algorithm stops (step 8). Otherwise, we proceed with step 4.
4. We find the records with the minimum number of repetitions because they would require a higher number of substitutions in order to fulfill the desired *k*-anonymity level. So, the algorithm selects all the records (*R*) with the same minimum number of repetitions.
5. We evaluate the semantic similarity (according to the metric introduced in section 2) between all the records in *R* with respect the rest of records available in the dataset. As a result, the set of records (*Rmax*) with the same maximum similarity against any other record is selected. In this manner it aims to select those records whose value substitution would lead to the minimum loss of semantics.
6. If several substitutions are equally optimum, it is selected the record (*r* in *Rmax*) whose replacement results in the lowest *k*-anonymity level (i.e. repetitions). Again, records which are more difficult to anonymize are prioritized, as they require more substitutions.
7. Finally, all the occurrences in the dataset for that record value tuple (*r*) are substituted by the selected one, and the process is repeated (going to step 2).
8. The process finishes when no more replacements are needed, because the dataset is *k*-anonymous.

4. Evaluation

In this section, we evaluate the role of ontologies in aiding the anonymization process in comparison to more simple approaches, based on ad-hoc VGHs, and approaches without any kind of semantic background, based on optimizing data distribution. The quality of the anonymization will be evaluated according to the utility of the masked data in data mining, in particular in a semantic clustering algorithm.

The masked dataset consisting on answers to polls made by the "*Observatori de la Fundació d'Estudis Turístics Costa Daurada*" at the Catalan National Park "Delta de l'Ebre". Visitants were requested to respond several questions regarding the main reasons and preferences when visiting the park. Each record, which corresponds to an individual, includes a set of textual answers. Due to the variety of answers, the disclosure risk is high and, therefore, individuals are easily identifiable. So, we consider textual answers as attributes that may lead to the de-identification of individual, leading to the disclosure of sensible and confidential information.

The dataset is composed by 975 individual records, for which we considered two sensible textual attributes per record. From the combination of those attributes, a total of 211 different responses were identified, being 118 of them unique (i.e. identifiers). Note that this sample represents a much wider and heterogeneous test bed than those reported in related works [14][18], which are focused on bounded categorical values. The answer values for those two attributes are general and widely used concepts (i.e. sports, beach, nature, etc.), all of them have been found in WordNet 2.1, which permits to use this ontology for performing the semantic similarity measurement.

The dataset has been masked with the method detailed in section 3 in three different configurations:

- Using WordNet 2.1 as ontology and the Wu & Palmer similarity (3) to guide the anonymization process. This will show the performance of a semantically grounded anonymization process in the preservation of data semantics.
- Using an ad-hoc VGH (see Figure 1), constructed according to the labels in which textual attributes are expressed in the dataset instead of WordNet. The same similarity metric as above is maintained. This will potentially show the limitations introduced by the use of simple and ad-hoc VGHs (as discussed in section 1) with regards to the semantic interpretation of data.
- No semantics are employed. The anonymization process is guided by a metric aimed to optimize the data distribution of the masked data. The discernibility measure (1) introduced in section 1 is used.

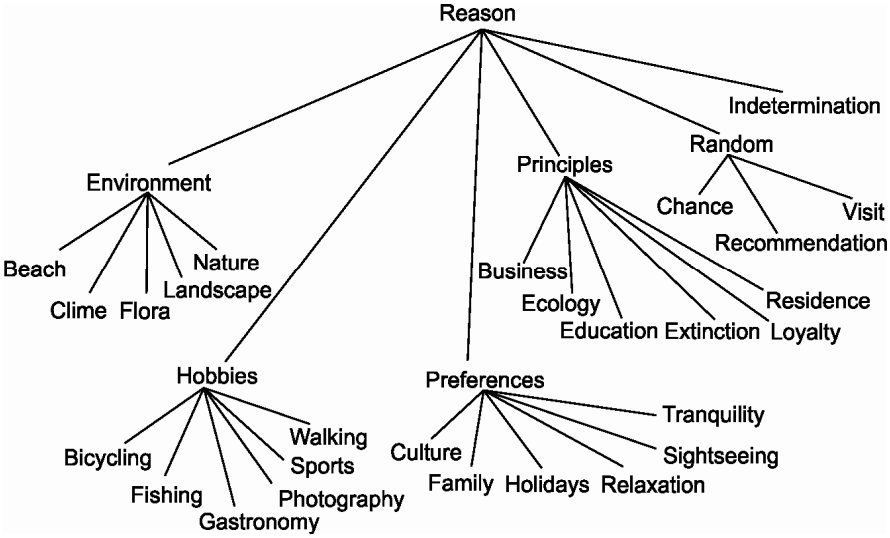


Figure 1. VGH constructed according to textual labels of sensible attributes.

In order to measure the quality of the masked dataset resulting from the application of the three different approaches, we employed an unsupervised clustering algorithm. The generation of clusters from data set is a well-known data mining task that permits to extract useful knowledge about a domain. Clustering is part of more complex task such as inference of rules or definition of user profiles. In [1][3] it was presented a hierarchical agglomerative clustering algorithm that is able to interpret and compare both numerical and textual features of objects. In this method ontologies are used to semantically compare the textual features of the objects that are being analyzed. According to those similarities, an iterative aggregation process of objects is performed based on the Ward's method [20]. As a result, a hierarchical classification of non-overlapping sets of objects is constructed, obtaining a dendrogram. This tree can be cut at a certain level to obtain a partition of the objects in clusters.

In this study, the clustering has been performed using both WordNet and the manually constructed VGH as background ontologies. It is important to note that the algorithm is able to evaluate, in an integrated fashion, several input ontologies. The method uses the more appropriate ontology at each moment, according to the ontology that better assess the alikeness between a given pair of terms. In this way, the partition into clusters obtained retains better the semantics of the textual attributes than other categorical clustering approaches [3].

By quantifying the differences between the clusters obtained from original data against the three masking approaches, we are able to conclude which one retains better the semantics and, in consequence, the utility of the data from a data mining point of view. Resulting clusters can be compared by means of the distance between partitions of the same set of objects as defined in [15]: considering two partitions of the same data set (in this case, the original an anonymized versions), being P_A a partition whose clusters are denoted as A_i and P_B a partition whose clusters are denoted as B_j, the distance is defined as:

$$distance(P_A, P_B) = \frac{2 * I(P_A \cap P_B) - I(P_A) - I(P_B)}{I(P_A \cap P_B)} \quad (4)$$

, where $I(P_A)$ is the average information of P_A which measures the randomness of the distribution of elements over the n classes of the partition (similarly for and $I(P_B)$), and $I(P_A \cap P_B)$ is the mutual average information of the intersection of two partitions.

The distance between the clusters obtained from the original data and those obtained from the three masking approaches are summarized in Table 1.

Table 1. Distances between the different clustering results

Test	Distance
Original data vs. Anonymization based on WordNet	0.398
Original data vs. Anonymization based on VGH	0.515
Original data vs. Anonymization based on discernibility	0.560
Anonymization based on WordNet vs. Anonymization based on VGH	0.531
Anonymization based on WordNet vs. Anonymization based on discernibility	0.589
Anonymization based on VGH vs. Anonymization based on discernibility	0.623

From these results we can see how the ontology-based anonymization has given a dataset that retains better the semantics of the original data (i.e. less information loss) than the other approaches. Compared to the simpler VGH-based anonymization (0.398

vs. 0.515) we observe that even that Wordnet is a general-purpose ontology, it allows a better interpretation of input data. Due to the coarse granularity of VGHs, it is likely to suffer from high information loss. Moreover, they offer a rough and biased knowledge model compared to fine grained and widely accepted ontologies. So, VGHs, in addition to the cost of manually constructing them, offer a too simple structure which results in homogenous similarity values, making difficult a proper differentiation between terms.

Comparing the results of the ontology-based anonymization with distributional approaches, the difference is even bigger (0.398 vs. 0.56), showing that semantics play an important role in the preservation of data utility. In consequence, conclusions extracted from the analysis of ontology-based anonymized data would be more similar to those obtained from the original data when using the semantic approach presented in this paper.

It is also relevant to observe the big differences between clusters resulting from each anonymization schema, whose distance ranges from 0.531 to 0.623. This shows a high discrepancy in the way in which records are fused according to the different semantic backgrounds and quality metrics.

5. Conclusions

In this paper we have studied masking methods to preserve the privacy of the individuals in textual attributes. Previous approaches neglected or very shallowly considered the semantics when masking textual data. This compromised the utility of information for data mining tasks, as the *meaning* of information is an important dimension when analysis must be made on anonymized data sets.

This paper proposes the exploitation of big and detailed ontologies when masking unbounded categorical data. Special care has been put in ensuring the scalability of the method when dealing with large and heterogeneous datasets (which are very common when involving text attributes) and big ontologies like WordNet. By enabling the exploitation of those already available ontologies we avoid the necessity of constructing ad-hoc hierarchies according to data labels like VGH-based schemas, which introduced several limitations both related to the cost of constructing them, and regarding to the shallow semantic background they represent.

The results show that the partition produced with the data anonymized with WordNet is more similar to the one obtained with the original data file. In fact, thanks to the wide coverage of WordNet, we can map sensible values to ontological nodes which do not necessary represent leafs of a hierarchy. As a result, semantically related concepts can be retrieved going through the hierarchy/ies to which the value belongs. Moreover ontological hierarchies are designed in a much general and fine grained fashion than ad-hoc VGHs, according to the agreement of domain knowledge experts, not in function on the input data. Those facts open the possibility of substituting sensible values by a much wider and knowledge-coherent set of semantically similar elements, including taxonomical subsumers (as done in generalization methods) but also with hierarchical siblings (with the same taxonomical depth) or specializations (located in a lower level).

In conclusion, the use of general ontologies for masking textual data has been proven to be a feasible and adequate solution that overcomes some of the drawbacks of previous approaches.

Acknowledgements

Thanks are given to "Observatori de la Fundació d'Estudis Turístics Costa Daurada" and "Parc Nacional del Delta de l'Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)" for the data used in the test. This work is supported the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02). Sergio Martínez Lluís and Montserrat Batet are supported by predoctoral research grant from the Universitat Rovira i Virgili.

References

[1] Agrawal, C., Yu, P.S (Eds) Privacy-preserving Data Mining: models and algorithms, Springer (2008).
[2] Batet, M., Valls, A., Gibert, K. Improving classical clustering with ontologies, in: Proceedings of the 4th World conference of the IASC, Japan, (2008), 137-146.
[3] Batet, M., Sánchez, D. Valls, A., Gibert, K.: Exploiting Taxocomical Knowledge to Compute Semantic Similarity: An Evaluation in the Biomedical Domain. In: Trends in Applied Intelligent Systems. 23rd Int. Conf. on Industrial Engineering and other Applications. IEA/AIE Córdoba, Spain, Springer (2010).
[4] Bayardo, R. J., Agrawal, R. Data privacy through optimal k-anonymization. Proceedings of the 21st International Conference on Data Engineering (ICDE), (2005), 217-228.
[5] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, Swoogle, J.: A Search and Metadata Engine for the Semantic Web. In Proc. 13th ACM Conference on Information and Knowledge Management, ACM Press, (2004), 652-659.
[6] Domingo-Ferrer, J. A survey of inference control methods for privacy-preserving data mining, in Privacy-Preserving Data Mining: Models and Algorithms, eds. C.C. Aggarwal and P.S. Yu, Advances in Database Systems, v.34, N.Y.: Springer Verlag, (2008), 53-80.
[7] Fellbaum, C. WordNet: An Electronic Lexical Database. Cambridge, Massachusetts: MIT Press. 1998.
[8] Guarino, N. Formal Ontology in Information Systems. In Guarino N (ed) 1st Int. Conf. on Formal Ontology in Information Systems. IOS Press. Trento, Italy, (1998), 3-15.
[9] Guo, L., Wu, X. Privacy preserving categorical data analysis with unknown distortion parameters, Transactions on Data Privacy, 2, (2009) 185-205.
[10] He, Y., Naughton, J.F. Anonymization of Set-Valued Data via Top-Down, Local Generalization. 35th Int. Conf. VLDB. V.2. Lyon, France, (2009), 934-945.
[11] Iyengar, V. S. Transforming data to satisfy privacy constraints. Proceedings of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Minig (KDD), (2002), 279-288.
[12] Jiang, J., Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In Proc. Int. Conf. on Research in Computational Linguistics, Japan, (1997), 19-33.
[13] Lefevre, K., DeWitt, D.J., Ramakrishnan, R. Mondrian Multidimensional K-Anonymity. Proceedings on the 22nd Int. Conf. on Data Engineering ICDE, (2006), 25.
[14] Li, T., Li, N. Towards optimal k-anonymization. Data & Knowledge Engineering 65. Pp. 22-39 (2008).
[15] López de Mántaras, R.: A distance-based Attribute Selection Measure for decision tree induction. Machine learning, 6, (1991), 81-92.
[16] Martinez S., Valls A., Sanchez D. Anonymizing Categorical Data with a Recoding Method based on Semantic Similarity. In: Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based System IPMU. Dortmund, Germany, (2010).
[17] Rada, R., Mili, H., Bichnell, E., Blettner, M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 9(1), (1989), 17-30.
[18] Samarati, P., Sweeney, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, (1998).
[19] Sweeney, L. k-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5), (2002), 557-570.
[20] Ward, J.H. Hierarchical grouping to optimize an objective function, JASA, (1963), 236-244.
[21] Wu, Z., Palmer, M. Verb semantics and lexical selection. In Proc. 32nd annual Meeting of the Association for Computational Linguistics, New Mexico, USA, (1994), 133-138.
[22] Xu, J., Wang, W., Pei, J., Wang X., Shi, B., Wai-Chee Fu, A. Utility-Based Anonymization for Privacy Preservation with Less Information Loss. ACM SIGKDD Explorations Newsletter V.8 I.2, (2006), 21-30.