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)

Abstract

Using microdata provided by statistical agencies has many benefits from the data mining point of view. However, such data often involve sensitive information that can be directly or indirectly related to individuals. An appropriate anonymisation process is needed to minimise the risk of disclosure. Several masking methods have been developed to deal with continuous-scale numerical data or bounded textual values but approaches to tackling the anonymisation of textual values are scarce and shallow. Because of the importance of textual data in the Information Society, in this paper we present a new masking method for anonymising unbounded textual values based on the fusion of records with similar values to form groups of indistinguishable individuals. Since, from the data exploitation point of view, the utility of textual information 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. Because textual data typically consist of large and heterogeneous value sets, our method provides a computationally efficient algorithm by relying on several heuristics rather than exhaustive searches. The method is evaluated with real data in a concrete data mining application that involves solving a clustering problem. We also compare the method with more classical approaches that focus on optimising the value distribution of the dataset. Results show that a semantically grounded anonymisation best preserves the utility of data in both the theoretical and the practical

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

setting, and reduces the probability of record linkage. At the same time, it achieves good scalability with regard to the size of input data.

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

1 Introduction

Statistical agencies generally provide summarised data generated from a collection of responses given by a set of individuals. Therefore, because responses are not directly published, an individual's privacy can be easily guaranteed. Privacy preserving techniques must ensure that an intruder cannot infer any individual's information from these summarised data [1] but this information may be not useful enough if a detailed analysis of the responses is needed. Many intelligent data mining techniques reveal interesting knowledge from sample data such as user profiles, tendencies and user behaviours. Such techniques require microdata, i.e. detailed individual information corresponding to an individual subject's response values. In this case, the data to be protected consists of a set of *m* records (corresponding to *m* individuals), each represented by a type with the values of *n* attributes (or variables).

Because of the potential benefits of exploiting microdata, new masking techniques are being developed to minimise the risk of re-identification when this information is made available [2]. From the point of view of privacy protection, data attributes are classified into four types: identifiers (which unambiguously identify the individual); quasi-identifiers (which may identify some of the respondents, especially if they are combined with the information provided by other attributes); confidential outcome attributes (which contain sensitive information); and non-confidential outcome attributes (the rest). The goal of statistical disclosure control is to prevent the link of confidential information to unique individuals. Identifiers (such as ID card numbers) are directly removed from the dataset. Quasi-identifiers do not link to specific respondents if they are considered separately but the problem arises if they are considered in groups. This is more problematic as the dataset includes a larger number of variables, thus resulting in unique value combinations and increasing the risk of re-identification. One way to achieve a certain level of anonymity and lower the risk of re-identification on a set is to satisfy the *k*-anonymity property [3]. This establishes that each record in a dataset must be indistinguishable with at least *k-1* other records within the same dataset, according to its individual attribute values.

To satisfy the *k*-anonymity property, micro-aggregation masking methods have been designed to build groups of *k* indistinguishable registers by substituting the original values with a prototype. This data transformation results in a loss of information, which is a function of the differences between the original and masked datasets. These differences may compromise the utility of the anonymised data

from the data mining point of view. Ideally, the masking method should minimise the information loss and maximise data utility. We can distinguish between *global* anonymisation methods, in which all identifier or quasi identifier attributes are considered and anonymised at the same time (i.e. the records will satisfy *k*-anonymity) and *local* methods, in which each attribute is anonymised independently (i.e. each attribute will satisfy *k*-anonymity individually). In this latter case, the information loss of the whole dataset cannot be minimised because local transformations optimise individual attributes but not record's information loss.

In the past, many micro-aggregation methods were designed to build groups of continuous-scale 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 optimised by retaining a set of statistical characteristics [2]. However, extending these methods to categorical attributes (particularly textual categories or even open textual answers) is not straightforward because of the limitations on defining appropriate aggregation operators for textual values, which have a restricted set of possible operations. Moreover, textual attributes may have a large and rich set of modalities if the individuals are allowed to give responses in textual form. Because of the characteristics of this kind of values and the ambiguity of human language, defining suitable aggregation operators is even more difficult. Semantics play a crucial role in properly interpreting these data but this dimension is often ignored in the literature. In fact, some authors [3], [4], [5] deal with these data as a bounded set of textual categories for which suppressions or substitutions are executed in order to satisfy *k*-anonymity without taking into account the semantics of the values. The quality of masked data is typically considered by preserving the distribution of input data. Although data distribution is a dimension of data utility, we agree with other authors [6] that retaining the semantics of the dataset is more important if the aim is to draw substantive conclusions from data analyses.

The semantic interpretation of textual attribute values for masking purposes requires the exploitation of some sort of structured knowledge sources that allow mapping between words and semantically interrelated concepts. As we will describe in Section 2, some approaches have incorporated rudimentary background knowledge during the masking process. However, due to the lightweight and overspecified nature of that knowledge and the shallow semantic processing of data, such approaches do not provide a general solution. We argue that using well-defined general purpose semantic structures such as ontologies will help to better interpret textual data [7], [8], thus enabling a more accurate anonymisation of textual values and minimising information loss from a semantic point of

view. Ontologies are formal and machine-readable structures of shared conceptualisations of knowledge domains expressed by means of semantic relationships [9]. They have been successfully applied in many areas that deal with textual resources[10] and knowledge management [11]. Thanks to initiatives such as the Semantic Web [12], many ontologies have been created in the last few years from general purpose ones, such as WordNet [13] (for English words), to specific domain terminologies (e.g. medical sources such as SNOMED-CT [14] or MeSH [15]).

As we also show in section 2, related works usually tackle anonymisation in an exhaustive manner, defining an exponentially large search space of possible value substitutions. As a result, the scalability of the method is compromised, especially when we are 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], [12], [16], [17]. 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 later studying the dataset. At the same time, the privacy of the individuals is more critical because the disclosure risk increases due to the uniqueness of the answers. This has been argued in previous works [18] in which we proposed a simple algorithm to mask textual attributes individually.

To overcome these limitations of related works, in this paper we propose a global masking method for unbounded textual values. This method is based on the merging of quasi-identifier values of the input records, which permits groups of indistinguishable registers to be built with multiple textual attributes so that *k*-anonymity is satisfied. The method relies on the well-defined semantics provided by large and widely used ontologies such as WordNet. This ensures the proper interpretation of the meanings of words and maximises the quality of the anonymised data from the semantic point of view. The aim is to make the conclusions that can be inferred from the masked dataset using data analysis methods be the most similar to those obtained from the original data. Because of the potentially large size of ontologies (compared to tailor-made knowledge structures used in previous approaches [3], [4], [5], [17], [19]) and the fact that we are dealing with potentially unbounded textual attributes, we propose a non-exhaustive heuristic approach that provides better scalability than related works with regard to the size of the ontology and the input data. We will evaluate our proposal from both the theoretical and the practical points of view by applying it to real data and comparing the results of our method with another masking approach that is based on the optimisation of data distribution.

The rest of the paper is organised as follows. In section 2 we review the methods for the privacy protection of textual categorical data by focusing on those that take into account some kind of semantic knowledge. In section 3 we discuss the exploitation of ontologies for data anonymisation purposes and describe the proposed method, including the semantic foundations on which it relies, the

designed heuristics and the expected computational cost. In section 4 we test our method by applying it to real data obtained from a survey at the Delta de l'Ebre National Park in Catalonia (Spain). We evaluate the method on the basis of data utility preservation and the minimisation of disclosure risk. In the final section we present our conclusions and future work.

2 Related works

As we stated above, the masking of textual categorical data is not straightforward. Some basic works consider textual 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 some kind of noise (such as replacing values according to some probability distribution used by PRAM [20], [21]). Others [3], [5] perform local suppressions of certain values or select a sample of the original data aimed at satisfying *k*-anonymity while maintaining the information distribution of input data.

Though these methods achieve a certain degree of privacy in an easy and efficient manner, due to their complete lack of semantic analysis they fail to preserve the meaning of the original dataset. In recent years, therefore, some authors have incorporated a kind of knowledge background to the masking process.

In previous knowledge-based masking methods, the set of values of each attribute of the input records in the dataset are represented by *Value Generalisation Hierarchies* (VGHs) [3], [4], [5], [16], [17], [19], [22]. VGHs are manually constructed tree-like structures defined according to a given input dataset, where labels of an attribute represent leaves of the hierarchy and are recursively subsumed by common generalisations. The masking process involves, for each attribute, substituting several original values by a more general one obtained from the hierarchical structure associated with that attribute. This generalisation process decreases the number of distinct tuples in the dataset and therefore increases the level of *k*-anonymity. In general, for each value, different generalisations are possible according to the depth of the tree. In practice, the substitution is selected according to a metric that measures the information loss of each substitution compared to the original data.

In [3], [5], [22], the authors propose a global hierarchical scheme in which all the values of each attribute are generalised to the same level of the VGH. The number of valid generalisations for each attribute is the height of the VGH for that attribute. For each attribute, the method picks the minimal

generalisation that is common to all the record values for that attribute. In this case, the level of generalisation is used as a measure of information loss.

Iyengar [16] presented a more flexible scheme that also uses a VGH, in which a value of each attribute can be generalised to a different level of the hierarchy iteratively. This scheme allows a much larger space of possible generalisations. Again, for all values and attributes, all the possible generalisations satisfying *k*-anonymity are generated. A genetic algorithm then finds the optimum one according to a set of information loss metrics measuring the distributional differences with respect to the original dataset.

T. Li and N. Li [17] propose three global generalisation schemes. First, the *Set Partitioning Scheme* (SPS) represents an unsupervised approach in which each possible partition of the attribute values represents a generalisation. This provides the most flexible generalisation scheme but the size of the solution space grows enormously while 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 generalised groups of values to restrict the set of possible generalisations even more. Notice that in all three cases, all the possible generalisations allowed by the proposed scheme for all attributes are constructed and the one that minimises the information loss [4] is selected.

In contrast to the global methods introduced above, He and Naughton [19] propose a local partitioning algorithm in which generalisations are created for an attribute individually in a Top-Down fashion. The best combination, according to a quality metric (Normalised Certainty Penalty [23]), is recursively refined. Xu et al. [6] also propose a local generalisation 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 anonymised independently, which results in a more constrained space of generalisations (i.e. it is not necessary to evaluate generalisation combinations of all attributes at the same time). However, the optimisation 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, local methods typically lead to unnecessary generalisations as each attribute has to satisfy *k*-anonymity independently.

All the approaches that rely on VGHs have several drawbacks. Firstly, VGHs are manually constructed from each attribute value set of the input data. Human intervention is therefore needed in order to provide a suitable semantic background on which those algorithms rely. If input data values change, VGHs must be modified accordingly. Although this fact may be assumable when dealing

with reduced sets of values (e.g. in [17] on average a dozen different values per attribute are considered), this hampers the scalability and applicability of the approaches, especially when dealing with unbounded textual data (with potentially hundreds or thousands of individual answers). Secondly, the fact that VGHs are constructed from input data, which represent a coarse picture of the underlying domain of knowledge, produces overspecified and small hierarchies with a much reduced taxonomical detail. It is common to find VGHs with three or four levels of hierarchical depth, whereas a detailed taxonomy (such as WordNet) models up to 16 levels [13]. From a semantic point of view, VGHs offer a rough and overspecified knowledge model compared to fine-grained and widely accepted ontologies. As a result, the space for valid generalisations offered by a VGH would be much smaller than when exploiting an ontology. The coarse granularity of VGHs makes them likely to suffer from high information loss due to generalisations. As stated above, some authors try to overcome this problem by trying all the possible generalisations exhaustively, but this introduces a considerable computational burden and still lacks a proper semantic background. Finally, the quality of the results heavily depends on the structure of VGHs that, due to their limited scope and overspecified nature, offer a partial view of each attribute domain.

An alternative to using 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 generalisation. As a result, the solution space is exponentially large. The problem with this approach is that defining a semantically coherent total order for non-numerical attributes is very difficult and almost impossible for unbounded textual data. Moreover, the definition of a total order, compared with a multi-level hierarchy, limits the space of valid generalisations.

3 Exploiting ontologies for anonymising textual attributes

As we stated in the introduction, to overcome the limitations of the above VGH-based approaches, we can consider using a broad and detailed general ontology such as WordNet. With such ontologies, 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 generalisations.

WordNet [13] is a freely available lexical database that describes and organises 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 a concept's semantics. Hypernymy is by far the most common relation, representing over 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 word "line").

Considering those dimensions, using WordNet instead of VGHs as the semantic background for data anonymisation would result in a generalisation space several orders of magnitude larger. In fact, as most of the related works make generalisations in an exhaustive fashion, the generalisation space is exponentially large according to the depth of the hierarchy, the branching factor, the values and the number of attributes to consider. These approaches are therefore computationally too expensive and difficult to apply in such a big ontology like WordNet.

To be able to exploit the advantages that large ontologies like WordNet have over semantics, we present a heuristic global masking method that is based on the fusion of values of semantically similar records. In our method, each non-k-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 satisfied. As we bind the search space for possible substitutions to the number of different records in the input data, our method scales well in such a large ontology regardless of the total number of attributes and minimises the loss of semantics thanks to the semantically driven substitution process. Moreover, unlike the VGH-based approaches based on substituting sensitive values for more general ones, in our method, other semantically similar concepts (such as hierarchical siblings or specialisations) would also be considered.

3.1 Guiding the masking of data

As stated above, the goal of an anonymisation method is to find a transformation of the original data that satisfies *k*-anonymity while minimising the information loss and therefore maximising the utility of the resulting data. To guide the masking process towards the transformation that would result in the minimum information loss, a metric that evaluates the difference between the original data and the data resulting from each transformation is needed.

In the literature, various metrics have been exploited [3], [6], [16], [17], [19], [22]. Classical metrics, such as the Discernibility Metric (DM) [4], are used to evaluate the distribution of *m* records (corresponding to *m* individuals) into *c* groups of identical values, generated after the anonymisation process. Specifically, DM assigns to each record a penalty based on the size of the group g_i to which it belongs after the generalisation (1). A uniform distribution of values in equally sized groups would optimise this metric.

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

However, metrics based on data distribution do not capture how semantically similar the anonymised set is with respect to the original data. As we stated in the introduction, when dealing with textual attributes preserving semantic information is crucial to interpreting and exploiting anonymised data. In fact, from the utility point of view this aspect is more important than the distribution of the anonymised dataset when we wish to describe or understand a record by means of its attributes (this will be tested in the evaluation section).

To minimise the loss of semantic information between original and anonymised datasets, we rely on the theory of *semantic similarity* [24]. Semantic similarity measures the taxonomical likeness of words based on the semantic evidence extracted from one or several knowledge sources. In the literature, several approaches to computing semantic similarity can be identified according to the techniques employed and the knowledge used to perform the assessment. Classical approaches exploit the graphical nature of structured representations of knowledge as the basis for computing similarities. Typically, subsumption hierarchies and, more generally, ontologies have been used for this 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 [25], [26], [27]. Other approaches also exploit domain corpora to complement the knowledge available in the ontology and estimate a concept's Information Content (IC) from a term's frequency of appearance [28]. Though the latter approaches provide accurate estimations when enough data is available [24], their applicability is hampered by the availability and pre-processing of these data. In contrast, the edge-counting measures introduced above are characterised by their simplicity (which results in a computationally efficient solution) and their lack of constraints (as only an ontology is required), which ensures their applicability. For these reasons, we will rely on edge-counting metrics to guide

the masking process in order to maximise the semantic similarity between the original data and those resulting from the masking of record tuples.

To provide accurate results, edge-counting measures use WordNet's 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 in which to apply those measures [24].

The simplest way to estimate the semantic distance (i.e. the inverse to similarity) between two ontological nodes (*c1* and *c2*) is to calculate the shortest Path Length (i.e. the minimum number of links) connecting these elements (2) [25].

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

To normalise this distance, Leacock and Chodorow [26] 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, these 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 because they present different degrees of generality. Based on this premise, Wu and Palmer's measure [27] 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. This ranges from 1 (for identical concepts) to 0.

As Wu and Palmer's measure incorporates more semantic features than the other measures (i.e. absolute path length normalised by relative depth in the taxonomy), we have taken it as the metric to measure semantic similarity during the anonymisation 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 we stated in the introduction, four 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.

Unlike the exhaustive generalisation methods based on the VGHs analysed above, our approach deals differently with the global masking process. Thanks to the wide coverage of WordNet, we would be able to map textual attribute values into ontological nodes that do not necessarily represent leaves of a hierarchy. As a result, semantically related concepts can be retrieved by going through the ontological hierarchy/ies to which the value belongs. These ontological hierarchies are designed in a much more general and fine-grained fashion than VGHs and, according to the agreement of domain knowledge experts, not as a function of the input data. This opens the possibility of substituting values by a much wider and knowledge-coherent set of semantically similar elements. To ensure scalability with regard to ontology size and input data, we bind the space of valid value changes to the set of value combinations that are present in the input dataset. When changing one value of a record for another, we can substitute a taxonomical subsumer with another one (this is the only case covered by the generalisation method) but also with a hierarchical sibling (with the same taxonomical depth) or a specialisation (located at a lower level). In fact, in many situations a specialisation can be more similar than a subsumer because, as stated in section 3.1, due to their higher specificity concepts belonging to lower levels of a hierarchy have less differentiated meanings. As a result, the value change would result in a higher preservation of the semantic of data. This is an interesting characteristic and an improvement on the more restricted data transformations supported by VGH-based generalisation methods.

Briefly, the proposed method is based on the fusion of quasi-identifier values of each record with the values of another record. To select the value that minimises 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 anonymise and the most semantically similar one) will take the same values and become indistinguishable; therefore, 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 anonymised.

To formally present the method, we introduce several 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 anonymised 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 anonymised 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: $I(r_i) = \{r_k \,|\, r_{kj} = r_{ij}\ \forall j = 1..n\}$. This 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 of 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:

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

where, for each attribute value pair, the function sim_{sem} can be any of the semantic similarity measures presented in section 3.1. As we stated earlier, 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, the maximum similarity with regard to any other record available in D will represent the *quality of the best data transformation* for that record.

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

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:

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

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 anonymised version D^A. Information loss is usually given by the absolute difference [29], so quality is measured in terms of semantic similarity (sim_{sem}).

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

This value can be normalised 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)

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

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 anonymised 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 maximises the utility of the data by preserving the meaning of the values. In our case, this is a partition that minimises 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. 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 which are unique, thus leading to a large number of records that do not satisfy k-anonymity. To ensure the scalability of the method and guide the anonymisation

process towards the minimisation of information loss, we have designed several heuristics (*H*) that ensure the selection, at each iteration, of the best set of indistinguishable records (I_l) to transform:

H1) 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. These are those with minimum $|I_i|$, which correspond to the least anonymous ones.

H2) 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 maximise the quality of the anonymised dataset of the result at each iteration. This is the $I(r_i)$ with maximum $best\_quality(r_i)$.

H3) 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 the lowest. This is the $I(r_i)$ that minimises $min\_achievable\_anonymity(r_i)$. In this way, the records that are more difficult to anonymise are prioritised, since they will require more value fusions.

These 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 does not have a unique element, then H_3 is applied. Through tests performed on real data, these three criteria are enough to obtain a unique $I(r_i)$ whose values are merged with those of the $I(r_k)$ that allows the maximisation of $best\_quality(r_i)$, thus increasing the *k*-anonymity level of both $I(r_i)$ and $I(r_k)$. However, if when using these 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 satisfies 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 anonymisation is O(p^3), where p is 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, which is required in step #5 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, where n is a relatively small number compared to p because the set of quasi-identifier attributes is usually small.

For large datasets, where p can be large because of the unbound nature of values, the scalability is more critical. For this reason we have optimised 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 is 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 generalisation methods), the similarities between all of them can be pre-calculated and stored. This avoids recalculating the similarity for already evaluated value pairs and, more generally, register pairs. In this way, the similarity measure is calculated *a priori* only $n \cdot p^2$ times, improving the efficiency with respect to the most expensive function of O(p^2). As we illustrate in the evaluation section, with this modification the execution of the algorithm stays in the range of milliseconds for large datasets.

Note that the computational cost of our algorithm uniquely depends on the number of different tuples (p), unlike related works, which depend on the total size of the dataset (m) and on the depth and

branching factor of the hierarchy (which represent an exponentially large generalisation space of substitutions to evaluate).

4 Evaluation

We evaluated our method by applying it to a dataset consisting of answers to polls made by the *Observatori de la Fundació d'Estudis Turístics Costa Daurada* at the Delta de l'Ebre Catalan National Park. Visitors were asked to respond to several questions on their main reasons and preferences for visiting the park (see an extract in Table 1). Each record, which corresponds to an individual, includes a set of textual answers expressed as a noun phrase (with one or several words). Because of the variety of answers, the disclosure risk is high and individuals are easily identifiable. We therefore consider textual answers as quasi identifiers that should be anonymised.

Table 1. Extract of sample microdata used for evaluation. The last two rows are textual attributes masked with our approach.

Age	**Gender**	**Duration (in days) of the visit to the park**	**Number of companion**	**Origin**	**Reason for visiting the park**	**Main activities during the visit to the park**
23	M	1	2	Spain	nature	fishing
26	M	3	1	Spain	landscape	sports
45	F	3	2	Belgium	sports	bicycling
56	M	1	0	France	nature	culture
54	F	2	0	Spain	nature	fishing
26	F	5	3	France	fishing	nature
45	F	1	1	Spain	relaxation	culture
30	M	2	0	Holland	holidays	visit
37	F	2	3	Spain	second residence	beach

The dataset comprises 975 individual records and 26 variables. Two of the variables are unbounded textual attributes (the last two columns of Table 1). Considering these two attributes to be quasi-identifiers, we find a total of 211 different responses, 118 of which were unique (i.e. identifying a single person). Notice that if the person is identified, some confidential data may be released, such as the age or the number of accompanying persons (see Table 1). Fig. 1 shows the equivalence class structure defined by the values of the pair of attributes considered in this study. Note that this sample represents a much wider and more heterogeneous test bed than those reported in related works [5], [17], which focused on bounded textual values.

Fig. 1. Attribute distribution according to answer repetitions

The answer values for these two attributes are general and widely used concepts (i.e. sports, beach, nature, etc.). All of them are found in WordNet 2.1, which allows this ontology to be used to perform the semantic similarity measurement. However, as we are dealing with values represented by text labels we had to morphologically process them in order to detect different lexicalisations of the same concept (e.g. singular/plural forms). We applied the Porter Stemming Algorithm [30] to both text labels of attributes (e.g. *sports*) and ontological labels (e.g. *sport*) in order to extract the morphological root of words (e.g. *sport*) and be able to map values to ontological concepts and detect conceptually equivalent values in the dataset (e.g. *relaxation=relax* as the morphological root of both words is *relax*).

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 minimising the information loss from a semantic point of view (as described 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, on average, compared to the original ones. Information loss has been computed and normalised as defined in eq. 9. The same evaluation was repeated for different levels of *k*-anonymity.

To show the contribution of each heuristic in minimising the information loss of the results, we replaced the heuristic substitution by a naïve replacement that changes each sensitive record by a random one from the same dataset. Using the same basic algorithm presented in section 3, each random change will increase the level of *k*-anonymity until all records are anonymised. 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 five executions. The three heuristics proposed in section 3.2 were gradually introduced instead of the random substitution in a way that enables the contribution by each heuristic to the resultant quality to be quantified. The results of this test are shown in Fig. 2, where: no heuristic at all is considered; only the first one is considered; only the first one and the second one are considered; all three are considered together.

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

The results illustrated in Fig. 2 are consistent with 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 prioritise the less anonymous records, leads to a slight improvement on the complete random substitution. The second one, which incorporates the semantic similarity function as a metric to guide the value fusion process towards the minimisation of the semantic loss, leads to the most significant improvement. Incorporating the third heuristic leads to a very slight improvement in some situations as it is only executed in case of a tie (i.e. when there are several replacements with an equal value of maximum similarity, which is quite unusual).

As a result of the heuristic fusion process, our approach considerably improves the naïve replacement. This is even more noticeable for a high *k*-anonymity level (above 5); when the three heuristics were used, we clearly outperformed the semantic loss of the random version. This is highly convenient and shows that our approach performs well regardless of the desired level of privacy protection.

4.2 Comparing semantic and distributional approaches

To show the importance of a semantically focused anonymisation, we compared it with a more traditional schema that focused on the distributional characteristics of the masked dataset (as stated at the beginning of section 3.1). This was done by using the Discernibility Metric (eq. 1) in our algorithm instead of Wu and Palmer's measure to guide the masking process. Both semantic and distributional approaches were 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 anonymisations.

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

The figures show that the optimisation of the dataset distribution does not imply better preservation of the records' semantics. In fact, there is a noticeable semantic loss in the resulting dataset for *k*-anonymity values above 5 for the distributional approach. As we stated in the introduction, the utility of textual information from the data analysis point of view strongly depends on its semantics. We can see that classical approaches that focus on providing uniform groups of masked values may significantly modify a dataset's meaning, thus hampering its exploitation.

4.3 Evaluation of data utility for semantic clustering

In order to evaluate the hypothesis that, from the data exploitation point of view, a semantic-driven anonymisation retains the utility of the original data better than distributional approaches, we then 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 interesting for data analysis in order, for example, to extract user profiles, detect preferences or perform recommendations [2]. Data mining and, more specifically, clustering algorithms are widely used for organising and classifying data into homogenous groups. Although clustering algorithms have traditionally focused on continuous-scale numerical or bounded categorical data, the increase in volume and the importance of unbounded textual data have motivated authors to develop semantically grounded clustering algorithms [31].

In [32] a hierarchical clustering algorithm is presented that can interpret and compare both numerical and textual features of objects. In a similar approach to that used in the present study, ontologies are exploited as the base to map textual features to semantically comparable concepts. The likenesses of the concepts are then assessed using semantic similarity measures. According to these similarities, an iterative aggregation process of objects is performed based on Ward's method [33]. A hierarchical classification of non-overlapping sets of objects is therefore 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.

With this algorithm, and using WordNet as the background ontology, we evaluated the utility of data from the semantic clustering point of view. We compared the clusters obtained from the original dataset with 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) anonymisation procedures. A k-anonymity level of 5 was chosen for this comparison because it is a moderate privacy level that allows the retention of data utility.

By quantifying the differences between the cluster set obtained from original data versus those obtained for both masking methods, we determined which one best retains the semantics and, therefore, the utility of data. The resulting cluster sets can be compared using the distance between partitions of the same set of objects as defined in [34]: considering two partitions (i.e. cluster sets) of

the same data set (in this case, the original and anonymised versions), where P_A is a partition whose clusters are denoted as A_i, and P_B is 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 that measures the randomness of the distribution of elements over the n classes of the partition (similarly for $I(P_B)$), and $I(P_A \cap P_B)$ is the mutual average information of the intersection of two partitions. These are computed as

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

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

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

where the probabilities of belonging to the clusters are $P_i=P(A_i)$, $P_j=P(B_j)$, and $P_{ij}=P(A_i \cap B_i)$.
This distance evaluates whether the objects have been distributed in the same clusters when two different partitions (original and anonymised) are compared. Distance values are normalised in the [0..1] interval, where 0 indicates that both partitions have identical clusters and 1 indicates that the partitions are maximally different.

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

Table 2. Distances between the clustering results

	Distance
Original data vs. Semantic anonymisation	0.26
Original data vs. Distributional anonymisation	0.57
Semantic vs. Discernibility anonymisations	0.56

Table 2 shows how a semantically driven anonymisation produces a dataset that retains the semantics of the original data better than distributional approaches (the distances in the resulting classification with respect to the original data are 0.26 and 0.57, respectively). Conclusions drawn from analysis of semantically anonymised data would therefore be more similar to those from the original data when the approach presented in this paper is used. As we stated in the introduction, this shows that semantics play an important role in the preservation of data utility. Note also the large differences between clusters resulting from each anonymisation schema, whose distance is a significant 0.56. This

shows a high discrepancy in the way records are fused according to the different quality metrics. This result is consistent with that observed in section 4.2, where semantic and distributional anonymisations provided significantly different results.

4.4 Record linkage

Data utility is an important dimension when aiming to anonymise data and minimise information loss. From the point of view of privacy protection, however, disclosure risk should be also minimised. The latter can be measured as a function of the probability of reidentifying the masked dataset with respect to original data.

To evaluate the disclosure risk of both semantically and distributionally anonymised datasets, we computed the level of *record linkage* (also named re-identification) [35] of the results. Record linkage (RL) is the task of finding matches in the original data from the anonymised results. The disclosure risk of a privacy-preserving method can be measured as the difficulty in finding correct linkages between original and masked datasets. This is typically calculated as the percentage of correctly linked records [35]:

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

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

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

where r_i is the original record, $r_i^{\,A}$ is the anonymised record, and L is the set of original records in D that match with $r_i^{\,A}$ ($L \subseteq D$). As we are dealing with textual features and value changes, record matching is performed by simple text matching all individual attributes (in the same order). Therefore, each $r_i^{\,A}$ is compared to all records of the original dataset D by text matching, thus 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). 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 and compared the original registers with respect to the semantic anonymisation and then with the distributional version of the method. The RL probabilities are illustrated in Fig. 5.

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

Both approaches follow a similar trend, i.e. RL probability decreases as *k* increases. We can also see that the degree of record linkage is quite stable for *k* values of 5 and over. The main difference is that our method gives lower probabilities of record re-identification than a distributional approach, especially for small values of *k*. Compared to the distributional approach, this allows the degree of *k*-anonymity to be lowered (resulting in less information loss) while a comparable level of disclosure risk is maintained.

In conclusion, these results show that an anonymisation process that is focused on the preservation of data semantics does not contradict the goal of a privacy preservation method, i.e. to minimise the risk of disclosure.

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 anonymisation process for the test dataset ranged from 1.2 to 1.6 seconds (according to the desired level of *k*-anonymity) (see Fig. 6). The pre-calculation of the semantic similarities between all value pairs of each attribute in the dataset took 6.33 minutes.

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

We can clearly see how, as stated in section 3.2, similarity computation is the most computationally expensive function and how minimising the number of calculations noticeably optimises runtime. Run times are also much lower than those reported by related works that required several hours [6], [17] to perform the anonymisation of the data even for generalisation schemas, very limited VGHs and bounded categorical data (3-4 levels of hierarchical depth and an average of a dozen values [17]). In contrast, we were able to mask much bigger and fine grained data in much less time while considering large and wide ontologies such as 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

Anonymisation of textual attributes deals with two *a priori* conflicting aspects of information: on the one hand, the minimisation of the disclosure risk by fulfilling a desired level of *k*-anonymity and, on the other hand, the maximisation of data utility in order to properly exploit the data. Previous approaches neglected or only shallowly considered the semantic content of textual attributes. As we have discussed in this paper, the meaning of data is an important dimension when analysing the anonymised results to extract useful knowledge since it is required in data mining, decision making and recommendation processes.

Micro-aggregation is the most common masking method applied to categorical data [29]. It builds groups of *k* similar registers and substitutes them by their prototype to assure *k*-anonymity. However, applying this method to textual attributes is not straightforward because of the limitations on defining appropriate averaging operators for this kind of unbounded values. Most related works aggregate data using a generalisation approach that relies on tailor-made hierarchical structures. Because of 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 that satisfy *k*-anonymity through 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 several tests conducted with real textual data from visitors to a Catalan National Park. Our results indicate that, compared with a classical approach based on optimisation of the distribution of the data, ours retains the quality and utility of data better from a semantic point of view. This was illustrated when we exploited masked data using a clustering process. The partitions generated from the original dataset and the anonymised data are more similar with our semantic method than with classical approaches.

Finally, we have taken special care to ensure 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 need to construct tailor-made hierarchies according to data labels such as VGH-based schemas, which suppose a high cost and limit the method's applicability. Moreover, the

non-exhaustive heuristic algorithm based on constrained value substitutions achieved a good scalability with regard to the size, heterogeneity and number of attributes of input data and to the size, depth and branching factor of the ontology.

In future work we will study how the method behaves with other ontologies with different sizes and granularities (such as domain-specific ontologies, which may be exploited when input data refer to concrete domain terminology). We will also study the possibility of combining several ontologies as background knowledge in order to complement knowledge modelled for each of them.

Acknowledgements

We would like to thank the Observatori de la Fundació d'Estudis Turístics Costa Daurada and the Delta de l'Ebre National Park (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya; Department of the Environment and Housing of the Autonomous Government of Catalonia) for providing the data collected from the visitors to the Park. This work is supported by the Spanish Ministry of Education and Science (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02). Sergio Martínez Lluís is supported by a predoctoral research grant of the Universitat Rovira i Virgili.

References

[1] S. Giessing, Survey on Methods for Tabular Data Protection in ARGUS, in: J. Domingo-Ferrer, V. Torra (Eds.) Privacy in Statistical Databases, Springer Berlin / Heidelberg, 2004, pp. 519-519.

[2] J. Domingo-Ferrer, A Survey of Inference Control Methods for Privacy-Preserving Data Mining, in: C.C. Aggarwal, P.S. Yu (Eds.) Privacy-Preserving Data Mining, Springer US, 2008, pp. 53-80.

[3] L. Sweeney, k-anonymity: a model for protecting privacy, Int. J. Uncertain Fuzziness Knowl-Based Syst., 10 (2002) 557-570.

[4] R.J. Bayardo, R. Agrawal, Data Privacy through Optimal k-Anonymization, in: Proceedings of the 21st International Conference on Data Engineering, IEEE Computer Society, 2005, pp. 217-228.

[5] P. Samarati, L. Sweeney, 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] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A.W.-C. Fu, Utility-based anonymization for privacy preservation with less information loss, SIGKDD Explor. Newsl., 8 (2006) 21-30.

[7] E.G. Little, G.L. Rogova, Designing ontologies for higher level fusion, Inf. Fusion, 10 (2009) 70-82.

[8] M.M. Kokar, C.J. Matheus, K. Baclawski, Ontology-based situation awareness, Inf. Fusion, 10 (2009) 83-98.

[9] P. Cimiano, Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer-Verlag New York, Inc., 2006.
[10] D. Sánchez, D. Isern, M. Millan, Content annotation for the semantic web: an automatic web-based approach, Knowl. Inf. Syst., (2010) 1-26.
[11] A. Valls, K. Gibert, D. Sánchez, M. Batet, Using ontologies for structuring organizational knowledge in Home Care assistance, Int. J. Med. Inform., 79 (2010) 370-387.
[12] L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, 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, ACM, Washington, D.C., USA, 2004, pp. 652-659.
[13] C. Fellbaum, WordNet: An Electronic Lexical Database (Language, Speech, and Communication), The MIT Press, 1998.
[14] K.A. Spackman, K.E. Campbell, R.A. Cote, SNOMED RT: a reference terminology for health care, Proc AMIA Annu Fall Symp, (1997) 640-644.
[15] S.J. Nelson, D. Johnston, B.L. Humphreys, Relationships in Medical Subject Headings, in: K.A. Publishers (Ed.) Relationships in the Organization of Knowledge, New York, 2001, pp. 171-184.
[16] V.S. Iyengar, Transforming data to satisfy privacy constraints, in: KDD, ACM, 2002, pp. 279-288.
[17] T. Li, N. Li, Towards optimal k-anonymization, Knowl. Data Eng., 65 (2008) 22-39.
[18] S. Martínez, A. Valls, D. Sánchez, Anonymizing Categorical Data with a Recoding Method Based on Semantic Similarity, in: E. Hüllermeier, R. Kruse, F. Hoffmann (Eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications, Springer Berlin Heidelberg, 2010, pp. 602-611.
[19] Y. He, J. Naughton, Anonymization of Set-Valued Data via Top-Down, Local Generalization, in: VLDB '09: Proceedings of the Thirtieth international conference on Very large data bases, VLDB Endowment, Lyon, France, 2009.
[20] L. Guo, X. Wu, Privacy Preserving Categorical Data Analysis with Unknown Distortion Parameters, Trans. Data Privacy, 2 (2009) 185-205.
[21] J.M. Gouweleeuw, P. Kooiman, L.C.R.J. Willenborg, P.P. DeWolf, Post randomization for statistal disclousure control: Theory and implementation, in, Voorburg: Statistics Netherlands, 1997.
[22] K. LeFevre, D.J. DeWitt, R. Ramakrishnan, Mondrian Multidimensional K-Anonymity, in: Proceedings of the 22nd International Conference on Data Engineering, IEEE Computer Society, 2006, pp. 25.
[23] M. Terrovitis, N. Mamoulis, P. Kalnis, Privacy-preserving anonymization of set-valued data, Proc. VLDB Endow., 1 (2008) 115-125.
[24] J.J. Jiang, D.W. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, in: International Conference Research on Computational Linguistics (ROCLING X), 1997, pp. 9008.
[25] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Trans. Syst. Man Cybern., 19 (1989) 17-30.

[26] C. Leacock, M. Chodorow, Combining local context with WordNet similarity for word sense identification, in: WordNet: A Lexical Reference System and its Application, 1998.

[27] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, Las Cruces, New Mexico, 1994, pp. 133-138.

[28] D. Sánchez, M. Batet, A. Valls, K. Gibert, Ontology-driven web-based semantic similarity, J. Intell. Inf. Syst., (2009) 1-31.

[29] V. Torra, J. Domingo-Ferrer, Disclosure control methods and information loss for microdata, in: P. Doyle, J.I. Lane, J.J.M. Theeuwes, L.V. Zayatz (Eds.) Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies, Elsevier, 2001, pp. 91-110.

[30] M.F. Porter, An algorithm for suffix stripping, in: Readings in information retrieval, Morgan Kaufmann Publishers Inc., 1997, pp. 313-316.

[31] Z. He, X. Xu, S. Deng, k-ANMI: A mutual information based clustering algorithm for categorical data, Inf. Fusion, 9 (2008) 223-233.

[32] M. Batet, A. Valls, K. Gibert, Improving classical clustering with ontologies, in: Proceedings of the 4th World conference of the IASC, Japan, 2008, pp. 137-146.

[33] J.H. Ward, Hierarchical Grouping to Optimize an Objective Function, J. Am. Stat. Assoc., 58 (1963) 236-244.

[34] R.L. De Mántaras, A Distance-Based Attribute Selection Measure for Decision Tree Induction, Mach. Learn., 6 (1991) 81-92.

[35] V. Torra, J. Domingo-Ferrer, Record Linkage methods for multidatabase data mining, in: V. Torra (Ed.) Information Fusion in Data Mining, Springer, 2003.