Contributions on semantic similarity and its applications to data privacy

Montserrat Batet and David Sánchez

Universitat Rovira i Virgili
Department of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans, 26
E-43007 Tarragona, Catalonia
{montserrat.batet, david.sanchez}@urv.cat

Abstract. Semantic similarity aims at quantifying the resemblance between the meaning of textual terms. Thus, it represents the corner stone of textual understanding. Given the increasing availability and importance of textual sources within the current context of Information Societies, a lot of attention has been put in recent years in the development of mechanisms to automatically measure semantic similarity and to apply them to tasks dealing with textual inputs (e.g. document classification, information retrieval, question answering, privacy-protection, etc.). This chapter offers describes and discusses recent findings and proposals published by the authors on semantic similarity. Moreover, it also details recent works applying semantic similarity to privacy protection of textual data.

Keywords: semantic similarity, privacy protection, ontologies, knowledge.

1 Introduction

The enormous development of the World Wide Web and the Information Societies has made available large amounts of electronic resources. Because many of these resources are of a textual nature, a great interest has been shown in recent years about the automated understanding of textual contents. The counter stone of textual understanding is the assessment of the *semantic similarity* between textual entities (e.g. words, phrases, sentences or documents).

Semantic similarity aims at assessing a score that quantifies the resemblance between the meanings of the compared entities, so that algorithms relying on such assessment (e.g. classification, clustering, etc.) can seamlessly manage textual information from a numerical perspective. Because data semantics is an inherently human feature, semantic similarity measures proposed in the literature exploit one or several human-tailored information or knowledge sources, which are exploited under different theoretical principles. According to such features and principles, the first part of this chapter offers a classification and a comparative discussion of recent findings and proposals published by the authors on semantic similarity.

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Due to its core importance and the need of dealing with textual inputs, semantic similarity has been applied in recent years in a variety of tasks, which include natural language processing, information management and retrieval, textual data analysis and classification or privacy-protection. Regarding the latter, in which privacy protection methods obfuscate sensitive information in order to guarantee the fundamental right to privacy of individuals while retaining a degree of data utility [1], data semantics are of utmost importance when dealing with textual inputs: they influence both the risk of disclosing confidential information due to semantic inferences and the utility of the protected output understood as the preservation of data semantics [2, 3]. Privacy protection mechanisms have been usually proposed for numerical inputs, thus focusing on the distributional and statistical features of data. However, neglecting data semantics hampered their applicability and accuracy with textual inputs. Fortunately, in recent years, there has been a growing interest in applying semantic technologies and, particularly, the notion of semantic similarity to offer a semantically-coherent and utility-preserving protection of textual data. The remainder of this chapter details some recent works focusing on such direction.

2 Semantic similarity

A plethora of semantic similarity approaches are currently available in the literature. This section offers a classification of the proposed approaches according to the theoretical principles on which they rely, highlighting their advantages and shortcomings under the dimensions of accuracy, applicability and dependency on external resources (which are summarised in Table 1). Recent findings reported by the authors on each of the approaches are described in more detail.

Table 1. Comparison between similarity calculus paradigms.

Measure type	Advantages	Drawbacks
Edge-counting	Simple	Low accuracy
Feature-based	Exploit all taxonomic ancestors to improve the accuracy	A detailed ontology is required
Extrinsic IC-based	Accurate	Suitable tagged corpora is needed
Intrinsic IC-based	Do not require corpora	A detailed ontology is required
1st order co-occurrence	No ontology is required	Compute relatedness rather than similarity
2nd order co-occurrence	Evaluate related terms that do not directly co-occur	Complexity

2.1 Ontology-based measures

Ontologies, which have been extensively exploited to compute semantic similarity, define the basic terminology and semantic relationships comprising the vocabulary of a topic area [4]. From a structural point of view, an ontology is composed by disjoint

sets of *concepts, relations, attributes* and *data types* [5, 6]. In an ontology, concepts are organised in one or several *taxonomies* and are linked by means of transitive *is-a* relationships (taxonomical relationships). Multiple inheritance (i.e. the fact that a concept may have several hierarchical ancestors or subsumers) are usually included.

Ontology-based measures estimate the similarity of two concepts according to the structured knowledge offered by ontologies. These measures can be classified into *Edge-counting measures*, *Feature-based measures* and *measures based on Information Content*.

Edge-counting measures. They evaluate the number of semantic links (typically is-a relationships) separating the two concepts in the ontology [7-10]. In general, edge-counting measures are able to provide reasonably accurate results when a detailed and taxonomically homogenous ontology is used [8]. They have a low computational cost (compared to approaches relying on textual corpora) and they are easily implementable and applicable.

However, they just evaluate the shortest taxonomical path between concept pairs as evidences of distance (i.e. the opposite to similarity). For example, in Fig. 1 the shortest path length between the concepts *Surfing* and *Sunbathing* is 2 (*Surfing – Beach Leisure – Sunbathing*). This is a drawback, because many ontologies (e.g. WordNet, SNOMED-CT or MeSH) incorporate multiple taxonomical inheritance that would result in several taxonomical paths connecting concept pairs, as shown in Fig. 1. These paths represent explicit knowledge that is omitted by edge-counting methods [11]. Because of their simplistic design, edge-counting measures usually provide a lower accuracy than other ontology-based methods [11].

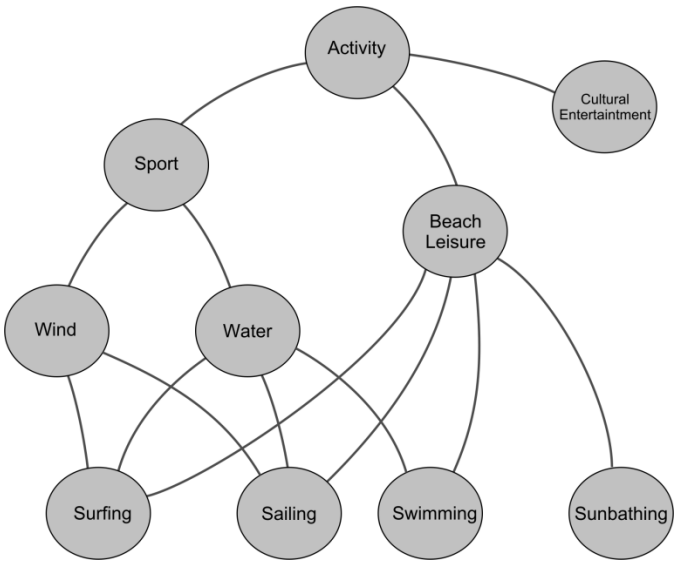


Fig. 1. Sample taxonomy about sports.

Feature-based measures. They rely on the amount of overlapping ontological features (e.g. taxonomic ancestors, concept descriptions, etc.) between the compared concepts [12-14]. Feature-based measures exploit more semantic evidences than edge-counting approaches, evaluating both the commonalities and the differences of concepts (e.g. common and different taxonomical ancestors). Since the additional knowledge helps to better differentiate concept pairs, they tend to be more accurate than edge-based measures [12].

However, since some feature-based approaches rely on semantic features other than taxonomical, like non-taxonomic relationships (e.g. meronymy) or concept descriptions (e.g. synsets, glosses, etc.), these measures can only be applied to a subset of the available ontologies, in which this kind of knowledge is available. In fact, domain ontologies often model semantic features rather than taxonomical relationships [15]. Another issue that hampers the applicability of feature-based measures as general purpose solutions is the fact that many of them [13, 14] incorporate ad-hoc weighting parameters that balance the contribution of each semantic feature.

To tackle these problems, in [11, 12], a coherent integration of taxonomic features is proposed, thus avoiding the need of weighting parameters while retaining a high accuracy. The approach in [12] assesses semantic similarity as a function of the amount of common and non-common taxonomic subsumers of concepts. Concretely, the similarity between two concepts c_1 and c_2 is measured according to the inverse non-linear ratio between their disjoint subsumers, as an indication of distance; this value is normalised by the total taxonomic subsumers of the concepts, because concept pairs that have many generalisations in common should be less distant than those sharing a smaller amount [12]. Formally, the semantic similarity measure is defined as (1), where $T(c_i)$ is the set of subsumers of the concept c_i, including itself.

$$sim\ (c_1, c_2) = -\log\left(1+\frac{|T(c_1)\cup T(c_2)|-|T(c_1)\cap T(c_2)|}{|T(c_1)\cup T(c_2)|}\right) \quad (1)$$

According to the sample taxonomy in Fig. 1, the number of disjoint subsumers between *Surfing* and *Sunbathing* counting themselves is 5 (*Surfing*, *Water*, *Wind*, *Sport*, *Sunbathing*), whereas the their total number of subsumers is 7 (*Surfing*, *Water*, *Wind*, *Sport*, *Sunbathing*, *Beach Leisure*, *Activity*). This results in a similarity value of $-log_2(1+(5/7))=-0.78$.

Measures based on Information Content. These measures rely on quantifying the amount of information (i.e. Information Content (IC)) that concepts have in common [16-18]. The IC of a concept states the amount of information provided by the concept when appearing in a context. On the one hand, the commonality between the concepts to compare is assessed from the taxonomic ancestors they have in common (which is referred as the *Least Common Subsumer*, LCS). For example, in Fig. 1, the LCS between *Swimming* and *Sunbathing* is *Beach Leisure*. On the other hand, the informativeness of concepts is computed either extrinsically from the concept occurrences in a corpus [16-18] or intrinsically, according to the number of taxonomical descendants and/or ancestors modelled in the ontology [19, 20].

In classical approaches [16-18] IC is computed extrinsically as the inverse of the appearance probability of a concept c in a corpus (2). Thus, infrequent terms are considered more informative than common or general ones.

$$IC(c) = -\log p(c) \tag{2}$$

However, textual corpora contain terms whereas ontologies model concepts, and, hence, a proper disambiguation and conceptual annotation of each word found in the corpus is required in order to accurately compute concept appearance probabilities [16]. Moreover, corpora should be large and heterogeneous enough to provide robust probabilities and avoid data sparseness (i.e. the fact that there are not enough data to extract valid conclusions about the distribution of terms). In addition, IC-based measures require that the probability of appearance of terms monotonically increases as one moves up in the taxonomy; thus, a concept is coherently considered more informative than any of its taxonomical ancestors and less informative than its descendants [16]. This requirement is fulfilled by computing the probability of appearance of a concept as the probability of the concepts and of any of its specialisations in the given corpus (3).

$$p(c) = \frac{\sum_{w \in W(c)} appearances(w)}{N} \tag{3}$$

where $W(c)$ is the set of words in the corpus whose senses are subsumed by c, and N is the total number of corpus words.

As a result, the background taxonomy must be as complete as possible (i.e. it should include most of the specialisations of a specific concept) to provide reliable results [21]. If either the ontology or the corpus changes, probability re-computations need to be recursively executed for the affected concepts. Scalability problems due to the need of manual annotation of corpora required to minimise language ambiguity hamper the applicability and accuracy of these measures [22].

To overcome these limitations, in recent years, some authors have proposed computing IC in an intrinsic manner by using only the knowledge structure modelled in an ontology [19, 23, 24]. These works assume that the taxonomic structure of ontologies is organised in a meaningful way, according to the principles of cognitive saliency [25]: concepts are specialised when they need to be differentiated from existing ones. In comparison to corpora-based IC computation models, intrinsic IC computation models consider that abstract ontological concepts with many hyponyms are more likely to appear in a corpus because they can be implicitly referred in text by means of all their specialisations. In consequence, concepts located at a higher level in the taxonomy with many hyponyms or leaves (i.e. specialisations) under their taxonomic branches would have less IC than highly specialised concepts (with many hypernyms or subsumers) located on the leaves of the hierarchy.

In a recent work [21], intrinsic IC calculus is improved by incorporating into the assessment additional semantic evidences extracted from the background ontology. The $p(c)$ is estimated as the ratio between the number of leaves in the taxonomical hierarchy under the concept c (as a measure of c's generality) and the number of taxonomical subsumers above c including itself (as a measure of c's concreteness) (4). It

is important to note that in case of multiple inheritance all the ancestors are considered. Formally:

$$IC(c) = -\log p(c) \cong -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max_leaves+1}\right) \quad (4)$$

The above ratio has been normalised by the least informative concept (i.e. the root of the taxonomy), for which the number of leaves is the total amount of leaves in the taxonomy (*max_leaves*) and the number of subsumers including itself is 1. To produce values in the range 0..1 (i.e. in the same range as the original probability) and avoid *log*(0) values, 1 is added to the numerator and denominator.

As discussed in [20, 21] this approach represents an improvement to previous ones [19, 24] in that it can differentiate concepts with the same number of hyponyms/leaves but different degrees of concreteness (expressed by the number of subsumers that normalises the numerator). Moreover, it can also consider the additional knowledge modelled by means of multiple inheritance relationships. Finally, it prevents the dependence on the granularity and detail of the inner taxonomical structure by relying on taxonomic *leaves* rather than complete sets of hyponyms.

Intrinsic IC-based approaches overcome most of the problems observed for corpus-based IC approaches (specifically, the need of corpus processing and data-sparseness). Moreover, they achieve a similar, or even better accuracy than corpus-based IC calculation when applied over detailed and fine grained ontologies [21]. However, for small or very specialised ontologies with a limited taxonomical depth and low branching factor, the resulting IC values could be too homogenous to enable a proper differentiation of concepts' meanings [21].

2.2 Semantic similarity from multiple ontologies

The main drawback of all the above ontology-based similarity measures is their complete dependency on the degree of coverage and detail of the input ontology [5]. Hence, ontology-based measures require a unique, rich and consistent knowledge source with a relatively homogeneous distribution of semantic links and good inter-domain coverage to work properly [23]. However, this is sometimes hard to achieve since, in many domains, knowledge is spread through different ontologies.

To tackle this problem, some authors [13, 26] focused on exploiting multiple ontologies for similarity assessment. The use of multiple ontologies provides additional knowledge that helps to improve the similarity estimation and to solve cases in which terms are not contained in a unique ontology [26].

Semantic similarity assessment from multiple ontologies is challenging because different ontologies may present significant differences in their levels of detail, granularity and semantic structures, making the comparison and integration of similarities computed from such different sources difficult [26]. In some approaches [13, 14], the two ontologies are simply connected by creating a new node which is a direct subsumer of their roots. This avoids the problem of knowledge integration but poorly captures possible commonalities between ontologies. Other authors [26] base their

proposal in the differentiation between *primary* and *secondary* ontologies, so that secondary ontologies are connected to concepts with the same textual label in the primary ontology. Since such work relies on absolute path lengths to compute similarity (which depend on the size of each ontology), the authors scale path lengths taking as reference the size of the primary ontology.

In any case, the core task in multi-ontology similarity assessment is the discovery of equivalent concepts between the different ontologies, which can be used as bridges between the ontologies and thus, enabling a standard similarity calculus from the integrated structure [13, 26-28]. In the following we detail recent works proposed by the authors on multi-ontology similarity assessment that propose solutions for different similarity calculus paradigms.

A multi-ontology semantic similarity method based on ontological features. The method presented in [28] considers all the possible situations according to which ontology the compared concepts belong, and computes similarities according to the feature-based approach formalised in eq. (1). Three cases are distinguished:

- *Case 1*: If the pair of concepts occurs in a unique ontology, the similarity is computed like in a mono-ontology setting (using e.g. eq. (1)).
- *Case 2*: If the two concepts appear at the same time in several ontologies, each one modelling knowledge in a different but overlapping way, the similarity calculus will depend on the different levels of detail or knowledge representation accuracy of each ontology [13]. Given the nature of the ontology engineering process, and the psychological implications of a human assessment of the similarity, two premises can be derived. First, ontological knowledge modelling is the result of a manual and explicit engineering process performed by domain experts. However, because of the bottleneck that characterises manual knowledge modelling, ontological knowledge is usually partial and incomplete [29]. As a result, if two concepts appear to be semantically distant, one cannot ensure if this is an implicit indication of semantic disjunction between the compared concepts or the result of partial or incomplete knowledge modelling. Second, as demonstrated in psychological studies [30], humans pay more attention to common than to differential features of the compared entities. As a conclusion of the two previous premises, given a pair of concepts appearing in different ontologies, the method in [28] considers the highest similarity score as the most reliable estimation and, thus, computes the similarity as follows:

$$sim(c_1, c_2) = \max_{\forall O_i \in O | c_1, c_2 \in O_i} sim_{O_i}(c_1, c_2) \quad (5)$$

- *Case 3*: If each of the two concepts belongs to a different ontology, each one modelling the knowledge from a different point of view, the set of shared and non-shared subsumers from the ontologies are assessed as follows: the set of shared subsumers for c_1 belonging to the ontology o_1 and c_2 belonging to the ontology o_2, $(T_{O1}(c_1) \cap T_{O2}(c_2))$ is composed by those subsumers of c_1, and c_2 with the same label, and also the set of ancestors of these terminological equivalent subsumers.

$$T_{O_1}(c_i) \cap T_{O_2}(c_i) = \bigcup_{\forall c_i \in ES} (T_{O_1}(c_i) \cup T_{O_2}(c_i)) \quad (6)$$

where $T_{O1}(c_1)$ and $T_{O2}(c_2)$ are defined as the set of subsumers of the concept c_1 belonging to the ontology o_1, including the concept c_1, and the set of subsumers of the concept c_2 belonging to the ontology o_2, including the concept c_2. The set of terminologically equivalent superconcepts (*ES*) is defined as:

$$ES = \{ c_i \in T_{O_1}(c_1) \mid \exists c_j \in T_{O_2}(c_2) \wedge c_i \equiv c_j \} \quad (7)$$

Notice that "≡" means a terminological match (i.e. identical concept labels). The remaining elements in $T_{O1}(c_1) \cup T_{O2}(c_2)$ are considered as non-common subsumers. Once the set of common and non-common subsumers has been defined, the similarity measure defined in eq. (1) can be directly applied.

A multi-ontology semantic similarity method based on IC. On the contrary to feature-based similarities, as stated above, IC-based measures rely on the ability to discover an appropriate *Least Common Subsumer* (LCS) that subsumes the meaning of the concepts belonging to different ontologies, and the ability to coherently integrate IC values computed from different ontologies. In [31] a multi-ontology semantic similarity method based on IC is presented, which also considers the three cases detailed above:

- *Case 1*: Both concepts appear in a unique ontology, so that the LCS can be retrieved unequivocally from it and the similarity can be computed in the same manner as in a mono-ontology scenario.
- *Case 2*: If both concepts appear in several ontologies at the same time, several LCSs can be retrieved. In this case, it is necessary to decide which LCS is the most suitable to compute inter-concept similarity. Following the same premises derived for the previous method, the *most specific* LCS from those retrieved from the overlapping set of ontologies is considered. In terms of IC, the *most specific* LCS corresponds to the LCS candidate with the maximum IC value:

$$LCS(c_1, c_2) = \underset{\forall o \in O | c_1, c_2 \in o}{\arg\max} (IC(LCS_o(c_1, c_2))) \quad (8)$$

  where $LCS_o(c_1,c_2)$ is the *LCS* between c_1 and c_2 for the ontology $o \in O$.
- *Case 3*: If each concept belongs to a different ontology, the set of subsumers of each concept is retrieved. Then, both sets are compared to find *equivalent* subsumers (i.e. those with the same textual label). As a result, the two ontologies can be connected by a set of equivalent concepts and the LCS for the concept pair can be retrieved as the *least common equivalent subsumer*:

$$LCS(c_1, c_2) = Least_Common_Equivalent_Subsumer(c_1, c_2) \quad (9)$$

  where the *Least_Common_Equivalent_Subsumer* is a function that terminologically compares all the subsumers of c_1 in o_1 and c_2 in o_2, and selects the most specific common one. In any case, the IC value of the retrieved LCS will be different when

computing it from an ontology than from the other. In this case the maximum IC value is selected.

$$IC(LCS(c_1, c_2)) = \max_{o \in \{o_1, o_2\}} IC_o(LCS(c_1, c_2)) \quad (10)$$

A method to discover semantically equivalent LCS between different ontologies. Finally, in [27] a method is proposed that complements the strict terminological matching of subsumers, in which the previous approaches rely, with a structural similarity function that aims at discovering semantically similar but not necessarily terminologically identical subsumers. This process is illustrated in Fig. 2.

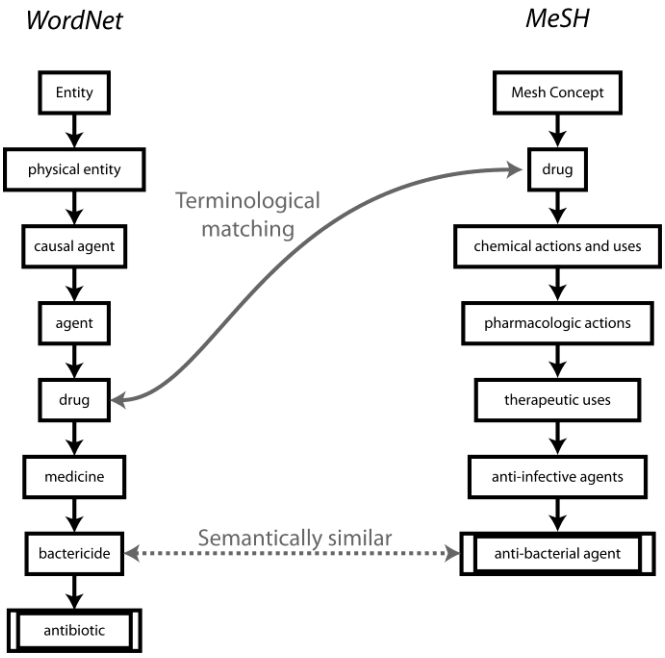


Fig. 2. Taxonomies for *antibiotic* (in the WordNet ontology) and *anti-bacterial agent* (in the MeSH ontology).

Essentially, the method assesses the similarity of subsumer pairs according to their *semantic overlap* and their *structural similarity*. The former is computed according to the number of hyponyms that subsumer pairs have in common.

$$sem_overlap(s_1, s_2) = \frac{|total_hypo_{O_1}(s_1) \cap total_hypo_{O_2}(s_2)|}{\sqrt{|total_hypo_{O_1}(s_1)| \times |total_hypo_{O_2}(s_2)|}} \quad (11)$$

where $total_hypo_O(s_i)$ is the complete set of hyponyms of the subsumer s_i in the ontology O, and the intersection ($\cap$) between both sets of hyponyms is defined as the set of concepts that are terminologically equivalent. The motivation under this score is the fact that hyponyms of a concept summarise and bind its meaning and enable to differentiate it from other ones. Interpreting this principle in an inverse manner, the fact that two subsumers (each one modelled in a different ontology) share a certain amount of hyponyms, gives us an evidence of similarity.

Finally, *structural similarity* relies on the average *semantic overlap* of the immediate neighbourhood of the compared subsumers. The idea is to assess the similarity between two subsumers according to the semantic overlap between their pairwise set of immediate ancestors and specialisations. Once all of the possible pairs of subsumers are evaluated, the pair with the highest structural similarity is selected as the LCS. In this manner, concepts with non-strictly identical labels but similar meanings could be matched, thus enabling a more accurate assessment of semantic similarity in a multi-ontology setting.

2.3 Distributional similarities

Distributional approaches for similarity calculus only use textual corpora to infer concept semantics. They are based on the assumption that words with similar distributions have similar meanings [32]. Thus, they assess term similarities according to their co-occurrence in corpora. Because words may co-occur due to different kinds of relationships (i.e. taxonomic and non-taxonomic), distributional measures capture the more general notion of semantic *relatedness* in contrast to *similarity*, which is understood strictly as taxonomic resemblance. Distributional approaches can be classified into first order and second order co-occurrence measures.

First order co-occurrence measures. They assume that related terms have a tendency to co-occur, and measure their relatedness directly from their explicit co-occurrence [33-35].

These measures usually rely on large raw corpora like the Web, from which term co-occurrence and, thus, similarity are estimated from the page-count provided by a Web Search Engine when querying both terms. Compared to IC-based measures relying on tagged corpora, distributional measures require neither any knowledge source nor manual annotation to support the assessment. Thanks to the coverage offered by a corpus as large and heterogeneous as the Web, distributional measures can be applied to terms that are not typically considered in ontologies such as named entities or recently minted terms.

However, their reliance on raw textual corpora is also a drawback. First, term co-occurrences estimated by page-counts omit the type of semantic relationship that motivated the co-occurrence. Many words co-occur because they are taxonomically related, but also because they are antonyms or by pure chance. Thus, page counts give a rough estimation of statistical dependency. Second, page counts deal with words rather than concepts. Due to the ambiguity of language and the omission of word context analysis, polysemy and synonymy may negatively affect the estimation of similarity/relatedness. Finally, page counts may not necessarily be equal to word frequen-

cies because queried words might appear several times in a Web resource. Due to these reasons, some authors have questioned the usefulness of page counts alone as a measure of relatedness [35]. Other authors [36] also questioned the effectiveness of relying just on first order co-occurrences because studies on large corpora gave examples of strongly associated words that never co-occur [35].

Second order co-occurrence measures. They estimate relatedness as a function of the co-occurrence of words appearing in the context in which the terms to evaluate occur [37-39].

These measures were developed to tackle the situation in which closely related words do not necessarily co-occur [36]. In such approaches, two words are considered similar or related to the extent that their contexts are similar [40]. Some approaches build contexts of words from a corpus of textual documents or the Web [41]. However, some authors have criticised their usefulness to compute relatedness because, while semantic relatedness is inherently a relation on concepts, Web-based approaches measure a relation on words [42]. Indeed, big-enough sense-tagged corpora are needed to obtain reliable concept distributions from word senses. Moreover, commercial bias, spam and noise are problems that may affect distributional measures when using the Web as corpus.

To minimise these problems, some authors exploited concept descriptions (*glosses*) offered by structured thesaurus like WordNet [39]. They argue that words appearing in a *gloss* are likely to be more relevant for the concept's meaning than text drawn from a generic corpus and, in consequence, glosses may represent a more reliable context. The use of glosses instead of the Web as corpora, results in a significant improvement of accuracy [39]. However, those measures are hampered by their computational complexity, since the creation of context vectors in such a big dimensional space is difficult. Moreover, because they rely on WordNet glosses, those measures are hardly applicable to other ontologies in which glosses or textual descriptions are typically omitted [15].

3 Applications to data privacy

Within the current context of Information Societies, governments, public administration and organisations usually require the interchange or release of potentially sensitive information. Because of the confidential nature of such information, appropriate data protection measures should be taken by the responsible organisations in order to guarantee the fundamental right to privacy of individuals. To guarantee such privacy, data protection/sanitisation methods obfuscate or remove sensitive information. This process necessarily incurs some loss of information, which may hamper data utility. Because the protected data should still be useful for third parties, data protection should balance the trade-off between privacy and data utility [1].

Traditionally, data protection has been performed manually, in a process by which a set of human experts rely on their knowledge to detect and minimise the risks of disclosure [43]. Semantics are in fact crucial in data privacy because they define the way by which humans (sanitisers, data analysts and also potential attackers) understand and manage information. Semantics thus directly influence the practical disclo-

sure and also the analytical utility of the protected data, because they provide the *meaning* of data [2, 44]. However, manual protection efforts can hardly cope with the current needs of privacy-preserving information disclosing [45].

In recent years, a plethora of automated protection mechanisms have been proposed [46]. However, most of them manage data from a pure statistical/distributional perspective and neglect the importance of semantics [2, 3]. Moreover, because of their numerically-oriented design, current solutions mainly focus on structured databases and/or their implementations mostly assume univalued and numerical attributes [1], which can be easily evaluated, compared and transformed by using standard arithmetical operators. However, a large amount of the sensitive data that is involved nowadays in data releases (i.e. Open Data [47]) is of unstructured and textual nature. For such data, standard protection methods are either non-applicable, or neglect data semantics. The limitations of current works with regard to data semantics hamper both the practical privacy and the analytical utility of the protected outputs.

In this section, we discuss some recent works on data privacy that, by relying on the theory of semantic similarity and exploiting structured knowledge bases, aim at overcoming the limitations of pure statistical and distributional methods when dealing with textual data.

3.1 Design of semantic operators

Privacy preserving methods usually transform input data to make it less detailed or more homogenous, so that the chance of disclosing sensitive information is minimised. Typical mechanisms employed in the protection of structured databases (i.e. microdata) are:

- Data microaggregation, which consists on making groups of similar individuals (represented by records in a data base) and replacing them by a prototypical value, thus making them indistinguishable from the other members of the group.
- Noise addition, which adds a degree of uncertainty to the input data proportional to the range of such data, in order to lower the probability of an unequivocal disclose of sensitive information.
- Data recoding, which replace individual values by, usually, generalised versions in order to make them indistinguishable.
- Sampling, which selects a subset of individuals as representatives for the whole dataset.
- Data swapping, which reciprocally replaces attribute values of similar individuals (records) in order to avoid unequivocal disclosure.

Most of the above mechanisms require a set of basic operators in order to compare and transform input data. Specifically:

- A distance measure is needed to detect which individuals are most similar in order to group/microaggregate/swap their values and minimise the loss of information resulting from the subsequent data transformation.
- An averaging operator is usually needed to select a prototypical value as the central point of a sample or to sort such sample according to their distance to the most central value.

- The variance of a sample of values is relevant when measuring, for example, the magnitude of the noise to be added to distort input values.

When dealing with numerical data, the standard arithmetical operators used to measure distances, compute arithmetic averages, standard deviations or variances can be straightforwardly applied. However, when managing textual data, such operators do not make sense. To tackle this problem, in [44, 48, 49], the authors propose several operators analogous to arithmetical ones that can manage textual data and that are both semantically and mathematically coherent. Essentially, they rely on the notion of semantic similarity to compute the resemblance between textual term pairs. By means of these operators one is able:

- To select the average value of a sample or the centroid as the value that minimises the semantic distance to all other values of the sample.
- To sort data according to their distance to the least central value (i.e. the most marginal one [49]).
- To measure the variance of a sample as the standard variance between their aggregated semantic distances.

Empirical experiments conducted in [44, 48, 49] show how the proposed operators are able to better capture the meaning of input textual data than pure distributional methods and, thus, to better preserve the utility of the output when applied to privacy-protection methods. Moreover, in [49], the proposed operators are proven to be coherent from a mathematical perspective, which is relevant when those are applied to existing algorithms designed for numerical data (e.g. hierarchical clustering [50]).

3.2 Adaptation of privacy protection mechanisms to textual attributes

Once a set of semantically and mathematically coherent operators is available, they can be used to adapt existing protection mechanisms so that the semantics of textual data are also considered.

Within the context of structured databases (e.g. census databases), in [3, 49, 51], the authors rely on the basic semantic similarity calculus and on the ability to compute the centroid of a sample of textual values in order to apply a well-known microaggregation algorithm (MDAV: Maximum Distance to the Average Vector [52]) to textual data in a semantically-coherent way. Specifically, in [49], it is shown how such a process is trivial if the appropriate operators are available and in [3] it is shown how the inherent characteristics of textual data (which are discrete and usually take values from a finite set of categories) can be considered in the algorithm design to better retain the utility of the protected output.

In [53], it is shown how a resampling algorithm meant for numerical data can also deal with textual values by relying on semantically coherent sorting and averaging operators. Even though data resampling usually incurs in a higher information loss than microaggregation it is also significantly more efficient, as it is empirically shown in [44], which may be preferable when protecting large data sets.

Finally, in [54] a new recoding method based on iteratively replacing semantically similar values is proposed. Several heuristics designed to preserve data semantics as much as possible and make the process more efficient are also defined.

In [44], an empirical study of the three above-described mechanisms is performed, comparing them accordingly to the dimensions of utility preservation, disclosure risk and computational efficiency, respectively.

All the above mechanisms focus on making input data more homogenous and indistinguishable in order to fulfil the *k*-anonymity privacy guarantee [55] over structured databases, that is, the fact that each individual (record) contained in a data set (database) is indistinguishable from, at least, k-1 other individuals; thus, the practical probability of re-identification of an individual is reduced to 1/k.

There exist other privacy models which offer different and more robust privacy guarantees. The most well-known is the ε-differential privacy model [56], which requires the protected data to be insensitive to modifications of *one* input record with a probability depending on ε. In this manner, an attacker would not be able to unequivocally disclose the confidential information of a specific individual with a probability depending on ε. To achieve this guarantee, practical enforcements focusing on numerical data add noise to the attribute values in a magnitude so that the protected output becomes insensitive (according to the ε parameter) to a modification of an input record.

In [57, 58], the authors propose mechanisms to achieve the ε-differential privacy guarantee for textual attributes in structured databases. The proposed method relies on a modified version of the MDAV algorithm that is applied to microaggregate the input values to reduce the sensitivity of data. Then, instead of adding numerical noise (which does not make sense for textual values), an exponential mechanism is used to replace the microaggregated values by the prototypes of each group (i.e. centroids) in a probabilistic way. The probability calculus picks the most probable centroid according to its degree of semantic centrality towards the other elements of the group, which is computed as detailed above [49]. In this manner, the degree of uncertainty and, thus, of information loss (from a semantic perspective) of differentially-private outputs is significantly reduced in comparison with the standard mechanism based on Laplacian noise [56].

3.3 Protection of semi-structured data

The above mechanisms and privacy models focus on relational data bases, in which individuals are represented by means of records with a finite set of univalued attributes. In such cases, it is quite straightforward to compare record pairs, since they have the same set of attributes and cardinality.

There exist, however, data sets containing transactional data in which individuals' information is represented by lists of items with variable and usually high cardinality (e.g. lists of diagnoses, query logs of users of a web search engine, etc.). Moreover, such data sets usually contain textual values. The protection of such data sets has been usually performed by generalising values to a common abstraction [59]. This process, however, severely hampers data utility due to the loss of granularity of the generalisation process, especially when generalising outlying values.

To tackle this issue, in [51, 60] the authors adapt the MDAV microaggregation method, which was originally designed for univalued numerical records, to achieve *k*-anonymity in transactional data sets with textual values. To do so, the authors first propose different mechanisms to aggregate the semantic similarities of sets of values

with different cardinalities, so that the MDAV algorithm can be applied like in the univalued scenario. After that, an especially designed aggregation methodology is proposed so that the prototypes of each microaggregated group capture both the semantics and the distributional features of the set of transaction lists that they are aggregating. The evaluation performed over a real set of user query logs shows that the use of semantic similarity measures results in a better preservation of data utility than purely distributional approaches.

3.4 Sanitisation of unstructured text

In the area of document redacting and sanitisation, input data consists on unstructured textual documents (e.g. medical visit outcomes, e-mails) whose contents must be protected according to the kind of information that should not be revealed (e.g. sensitive diseases, religious orientations, addresses, etc.). The challenge here is that no a priori sets of quasi-identifiers can be defined because, potentially, any combination of words of any cardinality may disclosure sensitive information. Thus, two tasks are usually performed: i) detection of terms that, individually (e.g. proper nouns) or, due to their co-occurrence (e.g. treatments and symptoms of sensitive diseases), may disclose sensitive data, and ii) protection of such terms, either by simple removal (redaction) or generalisation (i.e. sanitisation). The latter is more desirable because it better preserves the semantics of the original document and, thus, its analytical utility.

Within this area, in [61] an automatic method is proposed to detect sensitive terms individually, by using their degree of informativeness to measure the amount of sensitive semantics that they disclose. Then, these sensitive terms are replaced by generalisations obtained from a knowledge base. In [62] this method is extended in order to detect sets of terms that, because of their co-occurrence, may disclosure more sensitive information via semantic inference. In this latter work, first-order distributional similarity measures computed from the Web (as corpora) are used to quantify the degree of disclosure that combinations of terms enable with regard to a sensitive one (e.g. combinations of treatments and symptoms with respect to a sensitive disease). This is done as a function of their semantic relatedness that, as stated in the previous chapter, is captured by the distributional measures.

In [63, 64] the authors also focus on the anonymisation of textual documents. They rely on document vector spaces, which are normally used in information retrieval systems, to characterise a document as a vector of terms with frequency-based weights. Then, such vectors are microaggregated under the *k*-anonymity model using their cosine-distance as similarity measure.

3.5 Evaluating the semantic data utility

In general, the data utility of anonymised data is retained if the same conclusions can be extracted from the analysis of the original and the protected data sets. When dealing with textual data, such utility should be understood from a semantic perspective [2, 3].

In [51, 54], the authors propose a semantically-grounded method to evaluate the *information loss* (and thus, the degree of utility preservation) of privacy-protected outputs. They rely on semantic clustering algorithms [65] able manage textual data

and build clusters according to the semantic similarity between textual entities. To measure the information loss resulting from the data transformation performed during the anonymisation process, the authors quantify the differences between the cluster set obtained from original data against that obtained from the masked version. Because such cluster sets are a function of data semantics, their resemblance is a function of the preservation of data semantics during the protection process and, thus, of data utility (i.e. similar cluster sets indicate that similar analytical conclusions can be extracted from both the original and masked datasets).

In [63] the semantic data utility is evaluated from an information retrieval perspective, by performing specific "utility queries" and computing standard metrics of precision and recall.

3.6 Measuring the semantic disclosure risk

The practical privacy achieved by a protection method is measured as the risk of re-identification of the original records. This is usually quantified as the percentage of records in the original data set that can be correctly matched from the protected output, that is, the percentage of Record Linkages (RL).

When dealing with numerical data, RL is usually measured according to the number of correct matches between the records whose values are the least distant. In order to apply the same process when protecting textual data sets, the notion of semantic similarity can be used instead of the usual arithmetical distance. In [66], the authors propose a semantically-grounded RL method that quantifies the similarity between the semantics of record values, and perform an analysis according to the kind of features of the knowledge structure exploited to guide that assessment. In this work, the use of semantic similarities to guide the linkage process produced a higher number of correct linkages than the standard non-semantic approach (which is just based on comparing value labels), thus providing a more realistic evaluation of disclosure risk.

In [44, 51], the semantic RL method was also used to evaluate the practical privacy of the semantically-grounded protection mechanisms discussed above.

In [63] risk evaluation was considered from the information retrieval perspective by formulating queries containing risky terms.

4 Conclusions

In this chapter we have classified and discussed recent works proposed by the authors in the area of semantic similarity. Available solutions offer a wide spectrum of tools and methods to cover the different needs (e.g. accuracy, efficiency, etc.) and availability of external resources (e.g. ontologies, tagged or raw corpora) of specific application scenarios.

A relevant application scenario of semantic similarity is the protection of textual data. Even though privacy protection algorithms have been traditionally designed for numerical data, in recent years, there is a growing interest in applying them to textual data. Semantic similarity has a crucial role in such scenarios, because it is the key to enable the semantically coherent comparisons and transformations of data that privacy-protection algorithms require. This chapter summarised recent approaches in this

direction, which either adapt well-known protection algorithms to structured textual data sets (i.e. tabular data) and/or add support for less structure textual data, such as transactional data sets or unstructured textual documents. Data semantics has been also considered during the evaluation of the protected outputs. Empirical experiments reported in those works have shown that by carefully considering semantics during the data protection process, the analytical utility of the protected outputs can be better preserved in comparison with purely statistical or distributional approaches.

Acknowledgements

Authors are solely responsible for the views expressed in this chapter, which do not necessarily reflect the position of UNESCO nor commit that organisation. This work was partly supported by the European Commission under FP7 project Inter-Trust, by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, ICWT TIN2012-32757, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, CO-PRIVACY TIN2011-27076-C03-01 and BallotNext IPT-2012-0603-430000) and by the Government of Catalonia (under grant 2009 SGR 1135).

References

1. Domingo-Ferrer, J.: A Survey of Inference Control Methods for Privacy-Preserving Data Mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining, pp. 53-80. Springer (2008)
2. Torra, V.: Towards knowledge intensive data privacy. Proceedings of the 5th international Workshop on data privacy management, pp. 1-7. Springer-Verlag (2011)
3. Martínez, S., Sánchez, D., Valls, A.: Semantic adaptive microaggregation of categorical microdata. Computers & Security 31, 653-672 (2012)
4. Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.R.: Enabling Technology for Knowledge Sharing. AI Magazine 12, 36-56 (1991)
5. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag (2006)
6. Stumme, G., Ehrig, M., Handschuh, S., Hotho, S., Madche, A., Motik, B., Oberle, D., Schmitz, C., Staab, S., Stojanovic, L., Stojanovic, N., Studer, R., Sure, Y., Volz, R., Zacharia, V.: The karlsruhe view on ontologies. Technical report, University of Karlsruhe, Institute AIFB, Germany (2003)
7. 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, 17-30 (1989)
8. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd annual Meeting of the Association for Computational Linguistics, pp. 133 -138. Association for Computational Linguistics, (1994)
9. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, pp. 265-283. MIT Press (1998)
10. Li, Y., Bandar, Z., McLean, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Transactions on Knowledge and Data Engineering 15, 871-882 (2003)
11. Batet, M., Sánchez, D., Valls, A.: An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 44, 118-125 (2011)

12. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications 39, 7718-7728 (2012)
13. Rodríguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering 15, 442–456 (2003)
14. Petrakis, E.G.M., Varelas, G., Hliaoutakis, A., Raftopoulou, P.: X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. Journal of Digital Information Management 4, 233-237 (2006)
15. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: A Search and Metadata Engine for the Semantic Web. In: thirteenth ACM international conference on Information and knowledge management, CIKM 2004, pp. 652-659. ACM Press, (2004)
16. Resnik, P.: Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In: 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, pp. 448-453. Morgan Kaufmann Publishers Inc., (1995)
17. Jiang, J.J., Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: International Conference on Research in Computational Linguistics, ROCLING X, pp. 19-33. (1997)
18. Lin, D.: An Information-Theoretic Definition of Similarity. In: Fifteenth International Conference on Machine Learning, ICML 1998, pp. 296-304. Morgan Kaufmann, (1998)
19. Seco, N., Veale, T., Hayes, J.: An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In: 16th European Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, pp. 1089-1090. IOS Press, (2004)
20. Sánchez, D., Batet, M.: A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge. International Journal on Semantic Web and Information Systems 8, 34-50 (2012)
21. Sánchez, D., Batet, M., Isern, D.: Ontology-based Information Content computation. Knowledge-based Systems 24, 297-303 (2011)
22. Sánchez, D., Batet, M., Valls, A., Gibert, K.: Ontology-driven web-based semantic similarity. Journal of Intelligent Information Systems 35, 383-413 (2009)
23. Pirró, G.: A semantic similarity metric combining features and intrinsic information content. Data & Knowledge Engineering 68, 1289-1308 (2009)
24. Zhou, Z., Wang, Y., Gu, J.: A New Model of Information Content for Semantic Similarity in WordNet. In: Second International Conference on Future Generation Communication and Networking Symposia, FGCNS 2008, pp. 85-89. IEEE Computer Society, (2008)
25. Blank, A.: Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology. In: Eckardt, R., von Heusinger, K., Schwarze, C. (eds.) Words and Concepts in Time: towards Diachronic Cognitive Onomasiology, pp. 37-66. Mouton de Gruyter, Berlin, Germany (2003)
26. Al-Mubaid, H., Nguyen, H.A.: Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 39, 389-398 (2009)
27. Sánchez, D., Solé-Ribalta, A., Batet, M., Serratosa, F.: Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. Journal of Biomedical Informatics 45, 141-155 (2012)
28. Batet, M., Sánchez, D., Valls, A., Gibert, K.: Semantic similarity estimation from multiple ontologies. Applied Intelligence 38, 29-44 (2013)
29. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering. Springer-Verlag (2004)
30. Tversky, A.: Features of Similarity. Psycological Review 84, 327-352 (1977)

31. Sánchez, D., Batet, M.: A Semantic Similarity Method Based on Information Content Exploiting Multiple Ontologies. Expert Systems with Applications 40, 1393–1399 (2013)
32. Waltinger, U., Cramer, I., TonioWandmacher: From Social Networks To Distributional Properties: A Comparative Study On Computing Semantic Relatedness. In: Thirty-First Annual meeting of the Cognitive Science Society, CogSci 2009, pp. 3016-3021. Cognitive Science Society, (2009)
33. Turney, P.D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: 12th European Conference on Machine Learning, ECML 2001, pp. 491-502. Springer-Verlag, (2001)
34. Cilibrasi, R.L., Vitányi, P.M.B.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19, 370-383 (2006)
35. Bollegala, D., Matsuo, Y., Ishizuka, M.: A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, pp. 803–812. ACL and AFNLP, (2009)
36. Lemaire, B., Denhière, G.: Effects of High-Order Co-occurrences on Word Semantic Similarities. Current Psychology Letters - Behaviour, Brain and Cognition 18, 1 (2006)
37. Banerjee, S., Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: 18th International Joint Conference on Artificial Intelligence, IJCAI 2003, pp. 805-810. Morgan Kaufmann, (2003)
38. Wan, S., Angryk, R.A.: Measuring Semantic Similarity Using WordNet-based Context Vectors. In: IEEE International Conference on Systems, Man and Cybernetics, SMC 2007, pp. 908 - 913. IEEE Computer Society, (2007)
39. Patwardhan, S., Pedersen, T.: Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In: EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together, pp. 1-8. (2006)
40. Harris, Z.: Distributional structure. In: Katz, J.J. (ed.) The Philosophy of Linguistics, pp. 26–47. Oxford University Press, New York (1985)
41. Sahami, M., Heilman, T.D.: A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In: 15th International World Wide Web Conference, WWW 2006, pp. 377 - 386. ACM Press, (2006)
42. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of semantic distance. Computational Linguistics 32, 13-47 (2006)
43. MRA Health Information Services, http://mrahis.com/blog/mra-thought-of-the-day-medical-record-redacting-a-burdensome-and-problematic-method-for-protecting-patient-privacy/
44. Martínez, S., Sánchez, D., Valls, A.: A semantic framework to protect the privacy of electronic health records with non-numerical attributes. Journal of Biomedical Informatics 46, 294-303 (2013)
45. http://www.osti.gov/opennet
46. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., Wolf, P.P.d.: Statistical Disclosure Control. Wiley (2013)
47. Auer, S.R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web, pp. 722. (2007)
48. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. Knowledge-Based Systems 35, 160-172 (2012)
49. Domingo-Ferrer, J., Sánchez, D., Rufian-Torrel, G.: Anonymization of nominal data based on semantic marginality. Information Sciences 242, 35-48 (2013)
50. Batet, M.: Ontology based semantic clustering. AI Communications 24, 291-292 (2011)
51. Batet, M., Erola, A., Sánchez, D., Castellà-Roca, J.: Utility preserving query log anonymization via semantic microaggregation. Information Sciences 242, 49-63 (2013)

52. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11, 195-212 (2005)
53. Martínez, S., Sánchez, D., Valls, A.: Towards k-Anonymous Non-numerical Data via Semantic Resampling. In: Information Processing and Management of Uncertainty (IPMU), pp. 519-528. (2012)
54. Martínez, S., Sánchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. Information Fusion 13, 304-314 (2012)
55. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI International Report (1998)
56. Dwork, C.: Differential Privacy. In: 33rd International Colloquium ICALP, pp. 1-12. Springer, (2006)
57. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: Enhancing Data Utility in Differential Privacy via Microaggregation-based k-Anonymity. VLDB Journal (in press) (2014)
58. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: Improving the utility of differentially private data releases via k-anonymity. In: 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. (2013)
59. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. In: VLDB Endowment, pp. 115-125. (2008)
60. Batet, M., Erola, A., Sánchez, D., Castellà-Roca, J.: Semantic Anonymisation of Set-valued Data. In: 6th International Conference on Agents and Artificial Intelligence, pp. 102-112. (2014)
61. Sánchez, D., Batet, M., Viejo, A.: Automatic general-purpose sanitization of textual documents. IEEE Transactions on Information Forensics and Security 8, 853-862 (2013)
62. Sánchez, D., Batet, M., Viejo, A.: Minimizing the disclosure risk of semantic correlations in document sanitization. Information Sciences 249, 110-123 (2013)
63. Nettleton, D.G., Abril, D.: Document sanitization: Measuring search engine information loss and risk of disclosure for the wikileaks cables. In: International Conference on Privacy in Statistical Databases, pp. 308-321 (2012)
64. Abril, D., Navarro-Arribas, G., Torra, V.: Towards a private vector space model for confidential documents. In: 28th Annual ACM Symposium on Applied Computing, pp. 944-945 (2013)
65. Batet, M.: Ontology-based Semantic Clustering. AI Communications 24, 291-292 (2011)
66. Martínez, S., Sánchez, D., Valls, A.: Evaluation of the Disclosure Risk of Masking methods Dealing with Textual Attributes. International Journal of Innovative Computing, Information and Control 8, 4869-4882 (2012)