Semantically-grounded construction of centroids for datasets with textual attributes

Sergio Martínez*, Aida Valls, David Sánchez

Department of Computer Science and Mathematics. Universitat Rovira i Virgili

Av. Països Catalans, 26. 43007. Tarragona, Catalonia (Spain)

{sergio.martinezl, aida.valls, david.sanchez}@urv.cat

Abstract

Centroids are key components in many data analysis algorithms such as clustering or microaggregation. They are understood as the central value that minimises the distance to all the objects in a dataset or cluster. Methods for centroid construction are mainly devoted to datasets with numerical and categorical attributes, focusing on the numerical and distributional properties of data. Textual attributes, on the contrary, consist on term lists referring to concepts with a specific semantic content (i.e., meaning), which cannot be evaluated by means of classical numerical operators. Hence, the centroid of a dataset with textual attributes should be the term that minimises the *semantic* distance against the members of the set. Semantically-grounded methods aiming to construct centroids for datasets with textual attributes are scarce and, as it will be discussed in this paper, they are hampered by their limited semantic analysis of data. In this paper, we propose a method that, exploiting the knowledge provided by background ontologies (like WordNet), is able to construct the centroid of multivariate datasets described by means of textual attributes. Special efforts have been put in the minimisation of the semantic distance between the centroid and the input data. As a result, our method is able to provide optimal centroids (i.e., those that minimise the distance to all the objects in the dataset) according to the exploited background ontology and a semantic similarity measure. Our proposal has been evaluated by means of a real dataset consisting on short textual answers provided by visitors of a natural park. Results show that our centroids retain the semantic content of the input data better than related works.

Keywords: Data analysis, centroid, clustering, semantic similarity, ontologies.

---

* Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain. Tel.: +34 977 256563; Fax: +34 977 559710;
E-mail: sergio.martinezl@urv.cat.

1. Introduction

Data analysis tools aim to discover new knowledge or extract useful conclusions from large databases (consisting on *r* records with *m* attributes). One of the most commonly applied technique is cluster analysis. It is devoted to build a set of groups of objects (clusters) according to the similarity between the attribute values describing the objects in the database. Different algorithms have been proposed in the past, such as hierarchical agglomerative methods, partitional methods or density-based methods, among others (Han, 2005).

Many clustering algorithms rely on a *central* value (called centroid or prototype) that best *represents* a dataset or cluster. In general, a centroid is defined as a typical example, basis, or standard for other elements of the same group (Han, 2005). In practice, a centroid of a dataset is a vector that encodes, for each attribute, the most representative value of the objects found in the dataset. Centroids are needed in methods such as the *Ward*'s unsupervised agglomerative clustering, in which the centroid is used to measure the inertia intra and inter clusters (Ward, 1963) (inertia is a function of the distance between the objects and the centroids). Similarly, the widely-used clustering algorithm *k-means* (based on iteratively partitioning a dataset), uses centroids to create new clusters, which are adjusted at each iteration step (Everitt et al., 2011). This centroid-based partitioning approach is especially interesting for data mining applications due to their efficiency in processing large data sets (Z. Huang, 1998).

In addition to clustering methods, centroids are also required in other tasks such as: classification of new objects into already defined clusters (T. Huang et al., 2010), detection of outliers (Shin et al., 2006), parameter selection (Maniruzzaman et al., 2002), text clustering (Zhang et al., 2010) or data classification and processing (Keikha et al., 2009; Wong & Hamouda, 2003; Yihui, 2012).

Data protection is another relevant field (closely related to database analysis) in which centroids are useful (Domingo-Ferrer, 2008). In order to preserve the privacy of individuals and institutions, databases must guarantee a certain level of privacy before being made accessible to third parties (Martinez et al., 2010; Martinez et al., 2011; Martínez et al., 2010). Privacy-preserving data mining algorithms use masking methods applying data transformations to hide the original data values (Domingo-Ferrer, 2008). In particular, microaggregation methods have proven to be very effective in keeping low the information loss inherent to data transformations while maintaining an acceptable disclosure risk (Domingo-Ferrer & Mateo-Sanz, 2002). In this approach, individual records are grouped into small clusters prior to publication. Then, a centroid for each cluster is calculated and published instead of the original values. Ideally, the group partition and the centroid should be constructed so that the information loss of the masked data with respect to the original dataset is minimised. This is commonly measured as the Sum of Squared Errors (SSE) of the objects with respect to the centroid, to maximise the contribution of objects that are further apart. Moreover, to prevent the disclosure of individual information, it is required that each cluster contains at least *k* records (fulfilling the *k*-anonymity property (Samarati & Sweeney, 1998)). Notice that this method is quite similar to clustering algorithms, with an additional requirement regarding the minimum size of the clusters. In fact, some variations of data mining clustering methods are commonly used, such as the *k-Ward* microaggregation (Domingo-Ferrer & Mateo-Sanz, 2002) or the *MDAV* method (which is similar to *k-means* clustering) (Domingo-Ferrer & Torra, 2005).

Formally, the centroid of a dataset or cluster is a representative value (or a tuple of values in multivariate datasets) that *minimises the distance* to all other elements in the set (Han, 2005). In accordance with this

definition, geometrically, the centroid of a group is an object located at the *centre* of the group. To compute the centroid, a simple approach consists on selecting one of the objects of the cluster, so that it minimises the distance with respect to all the members of the cluster. This kind of centroid is called Medoid (Park & Jun, 2009). More flexible approaches are based on constructing a new synthetic centroid using an averaging function, which is applied to all the attribute values of the objects in the dataset. In both cases, distance functions are needed.

The case of centroid construction for numerical databases has been widely studied, being the *mean* the most common averaging operator (e.g., Arithmetic, Geometric, Harmonic mean or other types described in (Torra & Narukawa, 2007)). As distance measure, the Euclidean is the most common one (Domingo-Ferrer & Mateo-Sanz, 2002; Domingo-Ferrer & Torra, 2005).

On the contrary to numbers, when dealing with non-numerical data such as categorical or textual attributes, modalities cannot be directly compared by means of arithmetic operators. As it will be discussed in section 2, in this case, the most frequent value (i.e., *mode*) is usually taken. (Ahmed et al., 2005; Bai et al., 2011; Cao et al., 2012; Chen & Wang, 2004; Zhang et al., 2010). When a total order is available for the set of values, the *median* can also be used (Torra, 2004). Traditional categorical attributes consist on a discrete enumeration of modalities (i.e., bounded vocabulary), so that the set of possible values is restricted and fixed a priori (e.g., country of birth). Nowadays, thanks to the success of the Information Society, unbounded textual attributes can be easily compiled (e.g., hobbies, preferred food, etc.). In this case, attribute values consist on a single word or a noun phrase (e.g., "*trekking*", "*photography of birds*" or "*classical dance*") configuring a potentially large set, because respondents' answers are not restricted. This kind of attributes are more challenging to manage and analyse because words or noun phrases refer to particular concepts (i.e., they have a semantic content). An adequate analysis of these values requires a semantic analysis to understand their meaning (Batet et al., 2011; Do & Kim, 2009). Hence, when constructing the centroid of a dataset with textual attributes, the similarities between their meanings (evaluated at a conceptual level), should be taken into consideration (e.g. for hobbies attribute, "*trekking"* value is more similar to "*jogging"* than to "*classical dance"*). Related works constructing centroids for textual attribute values typically omit (Domingo-Ferrer & Torra, 2005; Torra, 2004) or slightly consider (Abril et al., 2010; Guzman-Arenas et al., 2011) data semantics during their analysis, hampering the quality of the results (Batet et al., 2010; Martinez et al., 2011).

In this paper, we propose a new approach to construct centroids of datasets with textual attributes from a semantic point of view. Coherently to the centroid definition, our goal is to minimise the *semantic distance* between the centroid and all other values in a set (i.e., a cluster or the whole dataset). In fact, the preservation of data semantics is a crucial aspect in textual attributes because it is directly related to the utility of the centroid (Martinez et al., 2011), which directly influences the conclusions extracted from the data analysis. By evaluating the semantic distance, we are able to quantify and minimise the loss of semantic content derived from the substitution of a group of objects by its centroid.

To perform a semantic analysis, a predefined knowledge source is needed to map textual labels to concepts that permit to interpret their meaning (i.e., discovering semantic relationships and measuring their semantic similarity). In this work, we rely on widely used ontologies to support the data analysis and the centroid construction. Ontologies offer a formal, explicit and machine readable structuring of a set of concepts by means of a semantic network, where one or several hierarchies are defined, and semantic relations are explicitly

modelled as links between concepts (Cimiano, 2006). Nowadays, ontologies are widely used models for knowledge-based applications (Gaeta et al., 2009). Many ontologies have been created in the last years, from general purpose knowledge sources, such as WordNet (Fellbaum, 1998), to specific domain terminologies (e.g., medical sources like SNOMED-CT or MeSH (Nelson et al., 2001)).

On the contrary to most related works (discussed in section 2), which bound the centroid to values already contained in the input dataset (Domingo-Ferrer & Torra, 2005; Guzman-Arenas et al., 2011; Torra, 2004), we have put special efforts in maximising the flexibility of the centroid construction. In our approach, we exploit, as much as possible, the knowledge provided by the background ontology, resulting in finer grained and more accurate centroids.

In a nutshell, first, we present a method to semantically construct centroids for univariate data, which provides an optimal minimisation of the semantic distance between the input dataset and the centroid, according to the background data. Then, the method is generalised to multivariate data, ensuring the scalability when dealing with large and high dimensional multivariate datasets. Our method have been evaluated and compared against related works (being semantically-grounded or not) using real data containing textual attributes. The dataset has been evaluated as a whole and also in a clustered fashion. Results show that our method better retains the semantic of data, minimising the semantic distance between the obtained centroids and the input data.

The rest of the paper is organised as follows. Section 2 reviews other methods for constructing centroids for non-numerical data, especially those exploiting some kind of background knowledge. Section 3 presents and formalises our centroid construction method. Section 4 tests our approach and discusses the results in comparison with related works. The final section contains the conclusions and several lines of future research.

2. Related work

When dealing with non-numerical data, we can distinguish two types of methods to construct centroids: those analysing data in a categorical fashion, which compare textual values according to their labels, and knowledge-based ones, which exploit a knowledge source to interpret data semantics.

Works considering textual data as categorical terms are based on Boolean (equality/inequality) operations applied to textual labels. In (Varde et al., 2006) it is proposed an approach called DesCond to extract a centroid for clusters of scientific input conditions. The centroid is selected from each cluster as a single object (in this case, this refers to all input conditions in a given experiment) such that it is the nearest neighbour to all other objects in the cluster. For this, the centroid is such value in the cluster that the sum of its distances to the rest of values of the cluster is the minimal. Because textual attributes are considered as categorical, the distance is defined as 0 if the attribute values are identical and 1 otherwise (Bai et al., 2011; Cao et al., 2012; Domingo-Ferrer & Torra, 2005; Torra, 2004). In (Domingo-Ferrer & Torra, 2005; Torra, 2004) authors propose a method for categorical microaggregation of confidential data (i.e., records with values linked to a particular individual) in order to ensure the privacy of the individuals before its publication. The microaggregated groups of records are substituted at the end of the algorithm by the centroid of the group. The centroid of textual attributes is selected as the value that most frequently occurs in the group (i.e., mode). In (Erola et al., 2010) authors also use a microaggregation-based masking method to protect query logs. To group and mask similar queries, it is

proposed a clustering algorithm based on finding similarities between queries by exploiting a taxonomy of topics. Then, for each cluster, a centroid consisting on a set of queries replaces all the queries in the cluster. The queries in the centroid are selected as those more frequently appearing in the cluster (i.e., mode). In (Greenacre & Hastie, 2010), authors use a similar strategy, classifying documents according to the most frequently appearing words. In (Zhang et al., 2010) authors propose a new method for document clustering. To make document clusters comprehensible, they assign the most frequent items in a cluster as the topic of the grouped documents. In (Bai et al., 2011), a new method is proposed to find the initial clusters centres for grouping algorithms dealing with categorical data. Authors select the most frequent attribute value (mode) as the cluster representative. In (Cao et al., 2012) it is proposed a dissimilarity measure for clustering categorical objects. Again, the mode is used as the criterion to select cluster representatives. In (T. Huang et al., 2010) authors proposed a supervised classification algorithm based on labelled training labels and local cluster centres. In order to avoid the interference of mislabelled data, authors select cluster centres so that they reflect the distribution of data (i.e. most frequent labels).

Due to the categorical treatment of textual data, the above-mentioned approaches can only evaluate data values as identical (i.e. maximum similarity) or not (i.e., minimum similarity). This naïve criterion can be improved by semantically analysing the data and, hence, being able to quantify more accurately the similarity between values. As stated in the introduction, this requires exploiting background knowledge to interpret the semantics.

In recent years, some authors started using knowledge sources to assist the construction of centroids. In (Abril et al., 2010) authors use the WordNet structured thesaurus (Pedersen et al., 2004) as ontology to assist the classification and masking of confidential textual documents. WordNet models and semantically interlinks more than 100,000 concepts referred by means of English textual labels. Authors exploit WordNet both to assist the classification process (in which relevant words are extracted from text and those are grouped according to the similarity of their meaning) and to select a centroid for each obtained cluster (which is used to mask confidential text). The Wu and Palmer's similarity measure (Wu & Palmer, 1994) is used to estimate the semantic alikeness between words by mapping them to WordNet concepts and computing the number of semantic links separating them. As a result, terms are clusterised according to their semantic similarity. The centroid of the resulting clusters, however, is selected as the Least Common Subsumer (LCS), which is the most concrete taxonomical ancestor found in WordNet for the terms found in the cluster. As a result, the centroid represents the semantic content that all the concepts referred in the cluster have in common. Even though term semantics are considered, the use of the LCS as centroid has some drawbacks. First, the presence of outliers (i.e., terms referring to concepts which are semantically far to the major part of the other elements in the cluster) will cause that the LCS becomes a very general concept (e.g., in the worst case, the root of the taxonomy). The substitution of cluster terms by such as general concept (e.g., entity, thing, abstraction, etc.) implies a high loss of semantic content. Moreover, the number of term repetitions is not considered during the centroid selection and hence, a scarce term will be considered as important as common ones, biasing the results. Those issues imply that the use of the LCS as the centroid does not minimise the semantic distance to all the elements in the cluster (incoherently to the centroid definition), resulting in a sub-optimal semantic loss.

In (Guzman-Arenas et al., 2011; Guzman-Arenas & Jimenez-Contreras, 2010) authors propose a generic way to calculate the centroid (called *consensus*) of a set of textual values. Their proposal exploits the knowledge

modelled in ad-hoc hierarchies that taxonomically link input values to measure their *confusion* (i.e., a measure of semantic alikeness). Confusion of two values *Conf (a,b)* refers to the confusion derived of using *a* instead of *b*. To measure this confusion, they count the number of descending links in the hierarchy from *a* to *b*, with the particularity that the confusion will be 0 if *a* is a descendent of *b* (i.e., the use of taxonomical subsumers is not penalised). The centroid or consensus of the set of textual values is selected as the term in the set that minimises the total confusion against the rest of elements of the set. This work incorporates better the notion of semantic centroid as a representative that minimises the semantic distance to all the other elements. However, as discussed above, the semantic distance derived from the substitution of a term by its subsumer (e.g., the root node, in the worst case) should also be quantified because it implies a noticeable loss of semantic content. Moreover, authors' approach is focused on very simple and overspecified taxonomies that must be constructed ad-hoc for each dataset because they only incorporate the values that appear in the input dataset. Hence, the quality of the results (i.e. the suitability of the selected centroid and the minimisation of the semantic distance) closely depends on the homogeneity, completeness and granularity of input values from the taxonomical point of view.

3. Ontology-based centroid construction

3.1. Preliminaries

In our approach, we follow the definition of centroid as the value or value tuple (in case of multivariate data) that minimises the distance against all the elements in a dataset/group. Formally, given a distance function *d* on the space of data, the centroid of a set of values $\{x_1,x_2,\ldots,x_n\}$ is defined as:

$$centroid(x_1,x_2,...,x_n) = arg\ min_c\left\{\sum_{i=1}^{n} d(c,x_i)\right\} \quad (1)$$

where *c* is a centroid candidate for the set of arguments.

As discussed in section 2, most related works dealing with textual attributes do not incorporate the notion of distance when constructing the centroid. Instead, some heuristics based on frequency analysis or on the notion of concept subsumption are used, assuming that those aid to minimise the distance. Considering the semantic nature of textual attribute values, we propose computing the distance relying on the estimation of their *semantic similarity*. Semantic similarity (Resnik, 1995) quantifies the taxonomical resemblance between a pair of terms by exploiting background knowledge sources from which semantic evidences are extracted. According to the type of source (e.g., ontology, corpora, tagged text, etc.) and the strategy followed to compute similarity, different families of measures can be identified, such as ontology-based (Leacock & Chodorow, 1998; Rada et al., 1989; Sánchez et al., 2012; Wu & Palmer, 1994), Information Content-based (Jiang & Conrath, 1997; Lin et al., 2010; Resnik, 1995; Sanchez & Batet, 2011; Sanchez et al., 2011) or distributional methods (Bollegala et al., 2007; Cilibrasi & Vitanyi, 2007; Sanchez et al., 2010; Turney, 2001). In this paper, we rely on similarity measures solely exploiting ontologies/taxonomies, which are characterised by their simplicity, scalability, lack of constraints and accuracy (Sánchez et al., 2012). In this context, the simplest way to estimate the semantic resemblance between two concepts (*a* and *b*) belonging to an ontology is to calculate the shortest *path length*

(i.e. the minimum number of links) connecting these elements via taxonomical relationships (Rada et al., 1989) (i.e., *is-a* relations).

$$dis_{path_length}(a,b) = \min \#\ of\ \ is-a\ links\ connecting\ a\ and\ b \tag{2}$$

This measure omits the fact that equally distant concept pairs located at different levels of the hierarchy should be evaluated in a different manner. It seems natural that the pair of concepts belonging to an upper level of the taxonomy should be considered less similar than those belonging to a lower level, because the former will have more general meanings than the later. Based on this premise, the similarity measure in (Wu & Palmer, 1994) also takes into account the depth of the concepts in the hierarchy. The negated form to obtain a dissimilarity is given in (3).

$$dis_{w\&p}(a,b) = 1 - \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \tag{3}$$

where N_1 and N_2 are the number of is-a links from *a* and *b*, 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.

Recently, in (Sánchez et al., 2012) it is proposed a new measure that, relying on similar principles as ontology-based approaches, aims to improve edge-counting measures by evaluating additional taxonomic knowledge modelled in ontologies. Instead of basing the assessment only on the length of the minimum path, authors evaluate, in a non-linear way, the number of non-common subsumers between the compared concepts as an indication of distance. This value is normalised by the complete set of subsumers of both concepts (3).

$$dis_{\log sc}(a,b) = \log_2\left(1 + \frac{\left|T(a) \cup T(b)\right| - \left|T(a) \cap T(b)\right|}{\left|T(a) \cup T(b)\right|}\right) \tag{4}$$

where *T(a)* is the set of taxonomic subsumers of the concept *a*, including itself. The advantage of this measure is that it implicitly considers all taxonomic paths between concept pairs (which appear due to multiple taxonomic inheritance), while retaining the efficiency and scalability of path-based measures.

In the proposed centroid construction method, we denote $dis_sem_O(a,b)$ to any semantic distance for the pair of concepts *(a,b)* measured using the ontology *O*.

A second relevant difference of our approach compared to related works concerns the search space explored to select the centroid. In general, knowledge-based approaches have a potential advantage with respect to methods dealing with data in a categorical fashion. When selecting the centroid according to the frequency of repetitions (i.e., mode) the search space of possible centroids is limited to the set of different values found in the cluster. As a result the centroid value is discretised according to the input values. In ontology/taxonomy-based approaches, on the contrary, the search space can be potentially extended to all the concepts modelled in the ontology (e.g., taxonomical subsumers of terms contained in the cluster) and, hence, the centroid can be synthetically constructed from a finer grained set of candidates. This advantage, however, is either slightly considered in related works (Abril et al., 2010), restricting the centroid to the LCS, or even not exploited at all (Guzman-Arenas et al., 2011; Guzman-Arenas & Jimenez-Contreras, 2010), in cases in which the background taxonomy only incorporates terms found in the input dataset. To expand the search space and obtain fine-

grained centroids, we propose relying on detailed ontologies like WordNet (Fellbaum, 1998). We map input labels to concepts in WordNet so that the hierarchical trees where each one is located can be explored to retrieve possible centroid candidates (e.g. complete sets of taxonomical ancestors). This strategy, combined with the estimation of the semantic similarity between terms, will help to minimise even more the semantic distance and propose more accurate centroids.

3.2. Constructing the centroid for univariate data

In this section, the case of univariate datasets is studied first, to facilitate the definition of the concepts and notation involved in our proposal. After that, in section 3.3, the method will be generalised to multivariate data.

Let us take an input dataset with a single textual attribute that can be represented as $V=\{<v_1,\omega_1>,\ldots,<v_{n,},\omega_n>\}$, where v_i is each distinct textual label and ω_i is its number of occurrences in the dataset. Let us take an ontology O containing and semantically modelling all v_i in V. The first step of our method consists on mapping the values in V to concepts in O, so that semantically related concepts can be extracted from O following the semantic relationships. In this work, we focus the analysis on taxonomical relationships because those are the most structure-building ones and, hence, the most usually available in ontologies. In fact, an investigation of the structure of existing ontologies via the Swoogle ontology search engine (Ding et al., 2004) has shown that available ontologies very occasionally model non-taxonomic knowledge. By analysing taxonomic relationships we retrieve new concepts (i.e. taxonomical ancestors) that become centroid candidates for the values in V. Following a similar premise as in (Abril et al., 2010; Guzman-Arenas et al., 2011), we assume that taxonomical subsumers of a term (including itself) are valid representatives of the term. The set of candidates is given in the *minimum subsumer hierarchy* ($H_O(V)$) that goes from the concepts corresponding to the values in V to the Least Common Subsummer (LCS) of all the values. We omitted taxonomical ancestors of the LCS because those will always be more general than the LCS (i.e., more semantically distant) and, hence, worse centroid candidates. Similarly, taxonomical specialisations below the bound defined by the minimum hierarchy are also omitted because they will be also more distant to the other values. Formally:

Definition 1. The *Least Common Subsumer* (LCS) of a set of textual values V in an ontology O ($LCS_O(V)$) is the most deep (i.e., specific) taxonomical ancestor that all the terms in V have in common for the ontology O.

Definition 2. The *set of taxonomical subsumers* ($S_{LCSO}(v_i)$) between a certain v_i in V and $LCS_O(V)$ is defined as the set of concepts found in the ontology O that connect via taxonomic relationships v_i and $LCS_O(V)$, including themselves. Note that in ontologies with multiple taxonomical inheritance, several paths can be found between v_i and $LCS_O(V)$; all of them are included in $S_{LCSO}(v_i)$.

Definition 3. The *minimum subsumer hierarchy* ($H_O(V)$) extracted from the ontology O corresponding to all the values in V is defined as the union of all the concepts in $S_{LCSO}(v_i)$ for all v_i.

$$H_O(V) = \bigcup_{i=1}^{n} \{ S_{LCS_O}(v_i) \} \tag{5}$$

where n is the cardinality of V.

Example 1. As illustrative example, consider a univariate dataset of patients, where each record corresponds to an individual and the unique attribute describes his/her condition, so that the following set of textual values can be extracted: V_1 = {<*colic, 1*>, <*lumbago, 3*>, <*migraine, 2*>, <*pain, 1*>, <*appendicitis, 1*>, <*gastritis ,1*>}. By mapping these values to concepts found in the background ontology O (WordNet –WN-, for this example) and applying Definition 3, we are able to extract the minim hierarchy H_{WN}, shown in the Fig. 1.

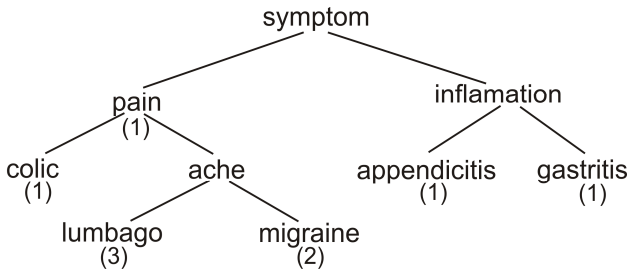


Fig. 1. Minimum subsumer hierarchy H_{WN} for the set V_1, extracted from WordNet. Numbers in parenthesis represent the number of repetitions of each value.

Applying Definition 1, the LCS between a pair of values such as *lumbago* and *migraine* is LCS_{WN}({*lumbago, migraine*})=*ache*; whereas the LCS of the entire set V_1 is $LCS_{WN}(V_1)$=*symptom*. Applying Definition 2, the ancestors of a value such as *migraine* are S_{LCSWN}(*migraine*)={*migraine, ache, pain, symptom*}.

Definition 4. All the concepts c in H_O are the *centroid candidates* for V.

Following Example 1, the centroid candidates of V_1 are those in H_{WN}: {*colic, lumbago, migraine, appendicitis, gastritis, pain, ache, inflammation, symptom*}.

Note that, according to the values in V and the structure of the background ontology, centroid candidates may be taxonomical ancestors of v_i, but also specialisations or siblings are found (e.g., in Example 1, the candidate ache is a specialisation of pain and a sibling of colic). This configures a larger search space of centroid candidates and a more flexible approach to construct the centroid than related works that limit to the values found in V or to their LCS (Abril et al., 2010; Guzman-Arenas et al., 2011).

From the set of centroid candidates, and following the centroid definition (eq. 1), the term c in H_O that minimises the semantic distance to all the v_i in V, considering also their number of repetitions in the dataset (ω_i), will be selected as the final centroid. It is important to consider the number of repetitions of individual values so that the centroid is adapted to the data distribution (i.e., values with a high number of repetitions should weight

more during the centroid selection than scarcer ones). Relying on semantic similarity measures introduced in 3.1 over the same background ontology, we evaluate each centroid candidate according to its aggregated distance to all input values.

Definition 5. The semantic distance (sd_O) computed from the ontology O between a centroid candidate c_j and all the elements in V is defined as the sum of the semantic distances between c_j and each $v_i \in V$, weighted by the number of repetitions of v_i in the dataset (ω_i):

$$sd_O(c_j, V) = \sum_{i=1}^{n} (\omega_i \cdot dis_sem_O(c_j, v_i)), \forall v_i \ in \ V \quad (6)$$

where the function dis_sem_O can be any semantic measure (like those presented in section 3.1), and n is the cardinality of V.

Finally, we can define the semantic centroid of a set of textual attributes as follows.

Definition 6. The *centroid* of a set of textual values V in the ontology O is defined as the concept c_j belonging to $H_O(V)$ that minimises the –weighted– semantic distance sd_O with respect to all the values of in V:

$$centroid_O(V) = \{\arg\min_{\forall c_j \in H_O(V)} (sd_O(c_j, V))\} \quad (7)$$

When more than one candidate minimises the distance, all of them would be equally representative, and any of them can be selected as the final centroid.

To illustrate the procedure for obtaining the centroid, let us take Example 1 and select the path length distance measure (eq. 2) for its simplicity. Taking the values in V_1, Table 1 shows the weighted semantic distance (sd_O, see eq. 6) for each centroid candidate (first column) in H_{WN}. In this case, the candidate "*ache*" is the concept that minimises the weighted semantic distance with respect to all values in V_1. Hence, applying Definition 6, $centroid_{WN}(V_1)=\{ache\}$.

Table 1.

Weighted semantic distance (sd_o, last column) obtained for each centroid candidate (first column) in Example 1. Inner columns show the path length distance between each centroid candidate and the values belonging to V_1 (numbers in parenthesis are the amount of repetitions, ω).

Centroid candidates	colic (1)	lumbago (3)	migraine (2)	appendicitis (1)	gastritis (1)	pain (1)	sd_O
colic	0	3	3	4	4	1	24
lumbago	3	0	2	5	5	2	19
migraine	3	2	0	5	5	2	21
appendicitis	4	5	5	0	2	3	34
gastritis	4	5	5	2	0	3	34
pain	1	2	2	3	3	0	17
ache	**2**	**1**	**1**	**4**	**4**	**1**	**16**
inflammation	3	4	4	1	1	2	27
symptom	2	3	3	2	2	1	22

The fact that all the centroid candidates are evaluated in order to minimise the distance to all values in *V*, produces optimal results with respect to the background ontology. It is important to note that, as shown in Example 1, neither the LCS of *V* (*symptom* in Example 1) (Abril et al., 2010) nor the most frequently appearing value in *V* (*lumbago* with 3 appearances in Example 1) (Domingo-Ferrer & Torra, 2005; Torra, 2004) necessarily minimise that distance. In fact, the use of the LCS as centroid for non-uniformly distributed data values (both with respect to their frequency of appearances, but also to their distribution through the hierarchy H_{WN}), typically results in a high semantic distance ($sd_{WN}(symptom, V_1) = 22$). For Example 1, the optimal centroid (*ache*) is the result of commonly appearing term (*lumbago*, with 3 appearances) and an unbalanced distribution of terms within the hierarchy (i.e., the *pain* branch includes more concepts than the *inflammation* one).

To show the influence of data value repetitions during the centroid construction, let us consider the following example:

Example 2. Given the set $V_2 = \{<colic,1>, <lumbago,1>, <migraine,1>, <appendicitis,1>, <gastritis,1>\}$ with uniformly distributed values and being $H_{WN}(V_2) = \{colic, lumbago, migraine, appendicitis, gastritis, pain, ache, inflammation, symptom\}$, the weighted semantic distance for each candidate is shown in Table 2.

Table 2.

Weighted semantic distance (sd_o, last column) obtained for each centroid candidate (first column) in Example 2. Inner columns show the path length distance between each centroid candidate and the values belonging to V_2 (numbers in parenthesis are the amount of repetitions, ω).

Centroid candidates	Colic (1)	Lumbago (1)	Migraine (1)	Appendicitis (1)	Gastritis (1)	sd_O
colic	0	3	3	4	4	14
lumbago	3	0	2	5	5	15
migraine	3	2	0	5	5	15
appendicitis	4	5	5	0	2	16
gastritis	4	5	5	2	0	16
pain	**1**	**2**	**2**	**3**	**3**	**11**
ache	2	1	1	4	4	12
inflammation	3	4	4	1	1	13
symptom	2	3	3	2	2	12

In this last case, the centroid is only influenced by the number of different terms found at each taxonomical branch. Because the hierarchy is unbalanced, the LCS does not minimise the semantic distance. In fact, the candidate "*pain*" is the concept that minimises the weighted semantic distance with respect to all values in V_2 being, hence, the centroid of the dataset.

3.3. Constructing the centroid for multivariate data

Real-world datasets are usually multivariate, with several attributes describing each object. In this section, we generalise the previous method to deal with multivariate textual datasets.

We consider object attributes as independent variables because they potentially belong to different domains of knowledge (i.e., they could define a different set of modalities or even they could be mapped to different ontologies). Consequently, for each attribute, an individual centroid should be constructed so that is minimises the distance to all the values in that attribute. The final centroid for a multivariate dataset is a *tuple* with the same cardinality as the number of attributes in the dataset. Note that because multivariate data is considered as a set of univariate attributes the search space of final centroid has a polynomial size ($|H_1|+\ldots+|H_m|$).

Formally, let us take a multivariate dataset with n objects and m textual attributes, the set MV of value tuples can be extracted: $MV=\{<\{v_{11},\ldots,v_{1m}\},\omega_1>,\ldots,<\{v_{n1},\ldots,v_{nm}\},\omega_n>\}$, where $v_i=\{v_{i1},\ldots,v_{im}\}$ is an object with m textual attributes and ω_i is its number of occurrences in the dataset.

Considering them as independent attributes, let A_j be the set of distinct values in the jth attribute in MV so that $A_j=\{<a_1,\psi_1>,\ldots,<a_q,\psi_q>\}$, where q is the cardinality of A_j, a_i is each distinct textual value for the jth attribute and ψ_i is its number of occurrences. Then, the centroid construction method presented in section 3.2 is applied to each univariate dataset A_j obtaining ($centroid_O(A_j)$. Concretely, applying Definition 3 over each A_j, a *minimum subsumer hierarchy* for each A_j will be extracted (H^j_O, covering all values in the jth attribute) configuring the set of centroid candidates for each attribute. The final centroid is the tuple of all attribute centroids, as follows.

Definition 7. The *centroid* of a set of multivariate data MV in the ontology O is defined as the tuple in which the jth value corresponds to the $centroid_O(A_j)$, being A_j the set of distinct values for the jth attribute in MV:

$$centroid_O(MV) = \{centroid_O(A_1), centroid_O(A_2), ..., centroid_O(A_m)\} \quad (8)$$

where, m is the number of attributes in MV.

Note that this multivariate centroid corresponds to the optimal solution with respect to the minimisation of the global semantic distance to all objects in the dataset. By global distance we refer to the arithmetic mean of the partial distances for each attribute centroid, as follows:

$$sdm_O(centroid_O(MV), MV) = \sum_{i=1}^{n}(\omega_i \sum_{j=1}^{m} \frac{dis_sem_O(centroid_O(A_j), v_{ij})}{m}) \quad \forall v_i \text{ in } MV \quad (9)$$

Hence, owing that each component of the centroid tuple ($centroid_O(A_j)$) minimises the distance of an attribute, the tuple also minimises the aggregated distance to all attributes.

Example 3. As illustrative example consider a dataset with two attributes describing the conditions and treatments of a set of patients, so that the following set of textual values can be extracted: MV_1 = *{<{colic, antibiotic}, 1>, <{lumbago, rehabilitation}, 3>, <{migraine, aspirin}, 2>, <{appendicitis, hospitalisation}, 1>, <{gastritis, codeine}, 1>, <{lumbago, codeine}, 2>, <{colic, hospitalisation}, 1>}*. Then, two sets of distinct textual values with their local number of occurrences are extracted: A_1 = *{<colic,2>, <lumbago,5>,*

<migraine,2>, <appendictits,1>, <gastritis,1>}, for the first attribute and A_2 = *{<antibiotic,1>, <rehabilitation,3>, <aspirin,2>, <hospitalisation,2>, <codeine,3>}* for the second one.

The *minimum subsumer hierarchies* H^1_{WN} and H^2_{WN} extracted from WordNet (and labelled with ψ_i) are shown in Fig. 2 and 3 respectively. The centroid candidates for the first attribute are those in H^1_{WN} = *{colic, lumbago, migraine, appendicitis, gastritis, pain, ache, inflammation, symptom}*, whereas for the second one are those in H^2_{WN} *={rehabilitation, antibiotic, aspirin, codeine, hospitalisation, therapy, medication, analgesic, medical_care}*. Note that in this case, 9+9=18 centroid candidates will be evaluated.

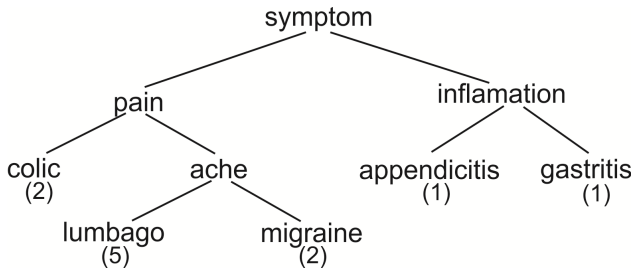


Fig. 2. Minimum hierarchy H^1_{WN} for A_1, extracted from WordNet, with the number of occurrences of each value.

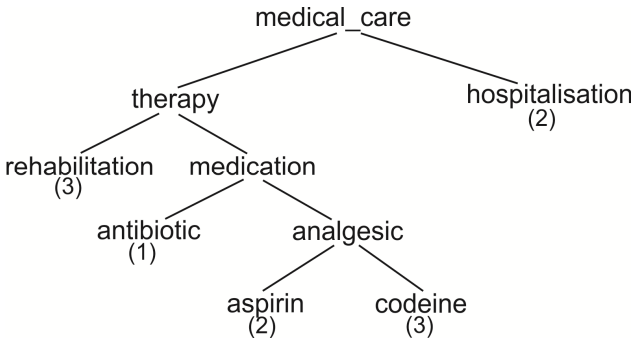


Fig. 3. Minimum hierarchy H^2_{WN} for A_2, extracted from WordNet, with the number of occurrences of each value.

Applying Definition 6, for each set H^j_O of centroid candidates, the centroid of the *j*th attribute ($centroid_O(A_j)$) is constructed. Using the path length-based measure (eq. 2) as semantic similarity, the results for A_1 and A_2 are shown in Tables 3 and 4, respectively.

Table 3.

Weighted semantic distance (sd_o, last column) obtained for each centroid candidate (first column) in A_1 of Example 3. Inner columns show the path length distance between each centroid candidate and the values belonging to A_1 (numbers in parenthesis are the amount of repetitions, ψ).

Centroid candidates	colic (2)	lumbago (5)	migraine (2)	appendicitis (1)	gastritis (1)	sd_O
colic	0	3	3	4	4	29
lumbago	3	0	2	5	5	20
migraine	3	2	0	5	5	26
appendicitis	4	5	5	0	2	45
gastritis	4	5	5	2	0	45

pain	1	2	2	3	3	22
ache	**2**	**1**	**1**	**4**	**4**	**19**
inflammation	3	4	4	1	1	36
symptom	2	3	3	2	2	29

Table 4.

Weighted semantic distance (sd_o, last column) obtained for each centroid candidate (first column) in A_2 of Example 3. Inner columns show the path length distance between each centroid candidate and the values belonging to A_2 (numbers in parenthesis are the amount of repetitions, ψ).

Centroid candidates	Rehabilitation (3)	Antibiotic (1)	Aspirin (2)	Codeine (3)	Hospitalisation (2)	sd_O
rehabilitation	0	3	4	4	3	29
antibiotic	3	0	3	3	4	32
aspirin	4	3	0	2	5	31
codeine	4	3	2	0	5	29
hospitalisation	3	4	5	5	0	38
therapy	1	2	3	3	2	24
medication	**2**	**1**	**2**	**2**	**3**	**23**
analgesic	3	2	1	1	4	24
medical_care	2	3	4	4	1	31

As a result, the centroid of A_1 is $centroid_{WN}(A_1)=\{pain\}$ and the centroid of A_2 is $centroid_{WN}(A_2)=\{medication\}$. Hence, applying Definition 7, the *centroid* of MV_1 is $centroid_{WN}(MV_1) = \{centroid_{WN}(A_1), centroid_{WN}(A_2)\} = \{pain, medication\}$.

Note that the global distance (eq. 9) for our centroid is $sdm_O(\{ache, medication\}, MV_1) = 21$, whereas other criteria, such as the LCS (for which the centroid is $sdm_O(\{symptom,medical_care\}, MV_1) = 30$) and the mode (for which the centroid is $sdm_O(\{lumbago,rehabilitation\}, MV_1) = 24.5$) would give worse results.

4. Evaluation

In this section, we evaluate the centroid construction strategy proposed in section 3. Our method is compared against those proposed by related works (being or not semantically-grounded). In order to numerically quantify the quality of the centroid proposed by each method, we compute the aggregated semantic distance between the centroid and each element in the input dataset as in eq. 9. The dis_sem_o function has been assessed with the measure presented in eq. 4, because it better quantifies the semantic differences between terms (Sánchez et al., 2012). In this manner, we are able to objectively quantify the loss of semantic content resulting from the substitution of a dataset or cluster by its centroid and, hence, the quality of the centroid. This is an important aspect from the point of view of data exploitation as it represents a measure of up to which level the semantics of the original values are preserved in the centroid.

4.1. Evaluation data

As evaluation dataset we used a set of answers provided by visitors of the *Delta de l'Ebre* Catalan National Park. Each object in the dataset corresponds to an individual and consists on a tuple (i.e. record) with two textual answers expressed as a noun phrase, corresponding to the two main reasons to visit the park. The textual values for these two attributes are widely used concepts mainly related to the visitor's hobbies, interests and activities (i.e. sports, beach, fishing, ecology, etc.) which can be directly found in WordNet.

The dataset comprises 975 records corresponding to the individual answers. Analysing the dataset distribution, we observe a total of 211 different response combinations, being 118 of them unique. The tuple with the highest amount of repetitions appears 115 times. Fig. 4 details the distribution of data according to the number of repetitions (i.e., frequency of appearance).

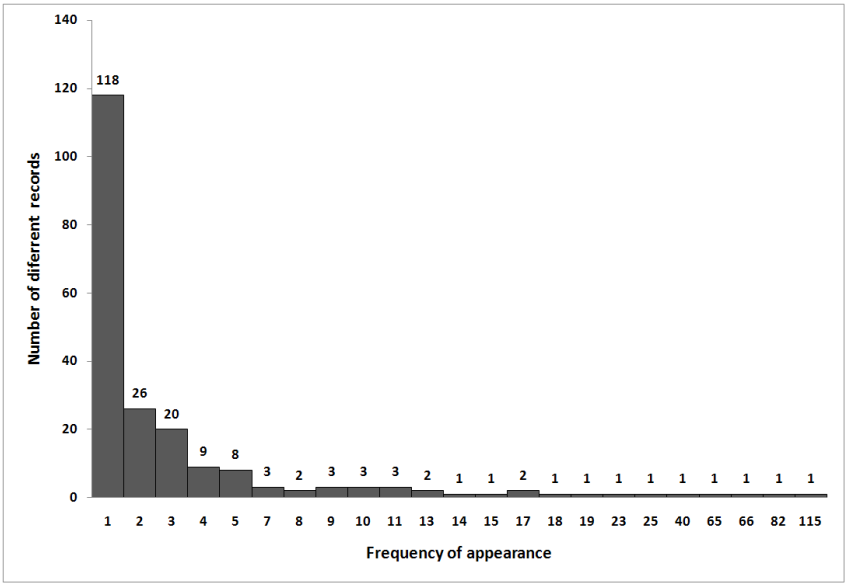


Fig. 4. Dataset value tuple distribution.

As stated in the introduction, the centroid is typically needed either when selecting a "central" value for all the input data from which data classification algorithms (e.g., K-means, MDAV, etc.) can be executed, or when extracting a "representative" for each of the clusters resulting from the classification process. Even though the centroid can be obtained in both cases in the same manner, the characteristics of the data are different in each case. Due to the goal of a classification algorithm (which is to minimise the distance between the elements in each cluster), one can expect that when considering the whole input set, data distribution would be more spread, with less homogenous and more distant values than when evaluating clusters individually. As a result, in the former case, the centroid calculation will be more affected by the presence of outliers and heterogeneously distributed values, a circumstance that will negatively influence some methods more than others.

In order to consider these two well-differentiated cases during the evaluation and to analyse the behaviour of each method in each circumstance, we configured two different scenarios. In the first one (*Scenario1*), the whole dataset described above will be analysed, constructing the centroid for all the 975 records. In the second one (*Scenario2*), a ontology-based hierarchical clustering algorithm (Batet et al., 2008), aimed to build a partition of the individuals according to their preferences, has been applied to the dataset. In this clustering method,

ontologies are exploited to guide the comparison of textual features, assessing their resemblance using semantic similarity measures. According to these similarities, an iterative aggregation of objects is performed based on Ward's clustering method (Ward, 1963). As a result, a hierarchical classification of non-overlapping sets of objects is obtained (i.e. a dendogram). The height of the internal nodes in the resulting dendogram is related to the variance of the clusters. Then, a cutting level in the dendogram is selected so that the cohesion of the resulting groups and the distinguishability among clusters is maximised. In this case, as a result of the clustering process, seven clusters of records have been obtained. Their size and data distribution are shown in Fig. 5.

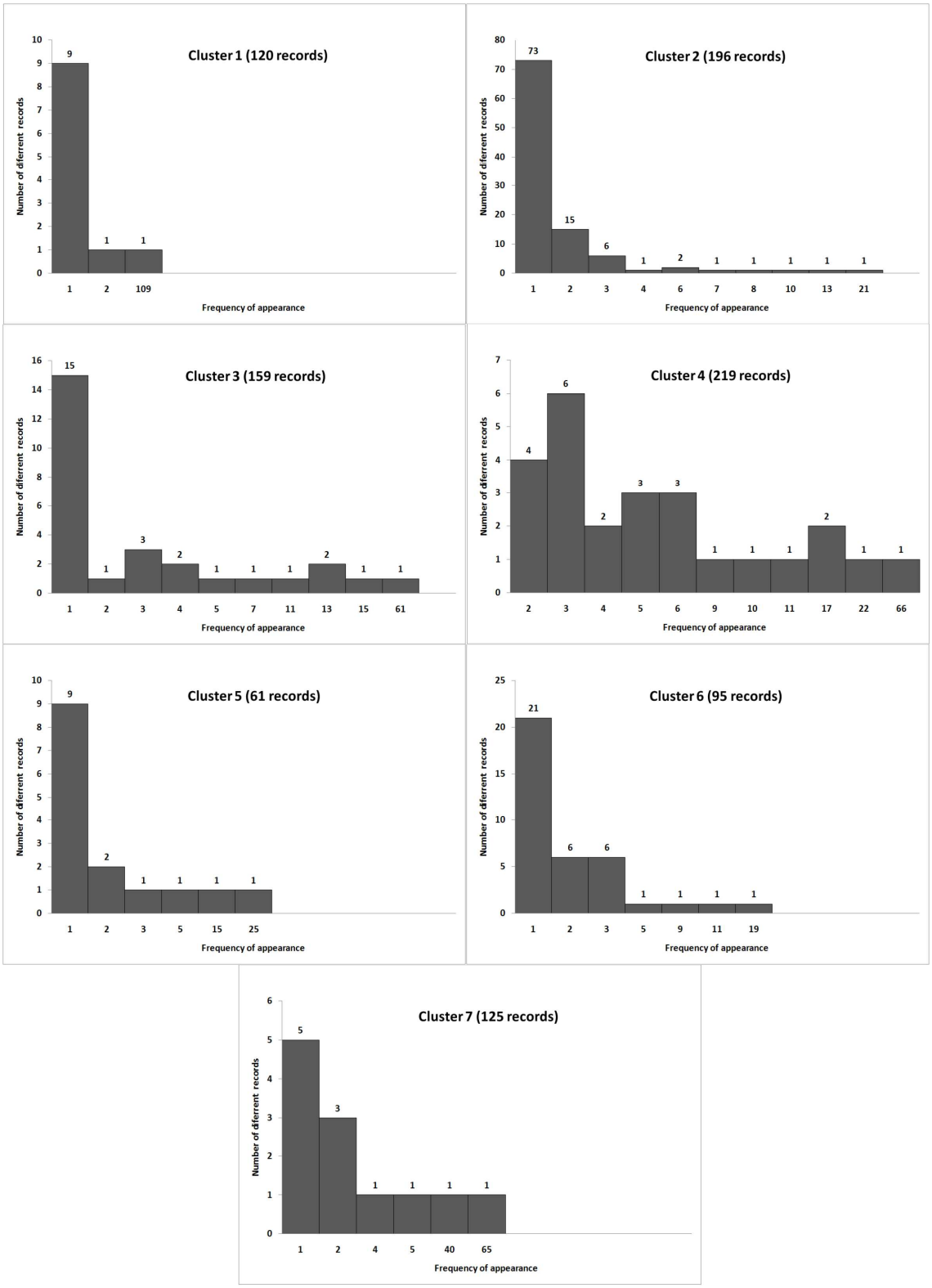


Fig. 5. Value tuple distribution for each cluster.

When comparing the data distribution of each individual cluster against the whole dataset (Fig. 4 vs. Fig. 5), it is easy to realise about the higher homogeneity of the data in the clusters. It is also interesting to observe the differences between clusters, which range from highly homogenous ones (e.g. *Cluster1*, with a dominant value

tuple with 109 repetitions) to less compact ones (e.g. *Cluster6* or *Cluster2*, with more spread value tuples). These last ones tend to incorporate a higher number of outliers that can hardly fit in other clusters due to their higher cohesion.

4.2. Implemented methods

Having a multivariate dataset for testing, our multivariate semantic centroid (SC) proposal tested. Each of the three semantic measures introduced in section 3.1 (path (Rada et al., 1989), W&P (Wu & Palmer, 1994) and LogSC (Sánchez et al., 2012)) have been applied to eq. 9, that is, during the centroid construction as goal function.

In addition, for evaluation purposes, the following strategies applied by related works have been also implemented:

- *Random centroid* (RC): a random record of the input dataset is selected as the centroid. This naïve strategy is used as baseline to compare against other methods.
- *Most frequently appearing centroid* (MFC): the record in the dataset with the highest amount of repetitions (e.g., 115 when considering the whole dataset) is considered as the centroid (i.e., mode). This is a common strategy employed by non-semantic approaches like (Bai et al., 2011; Cao et al., 2012; Domingo-Ferrer & Torra, 2005; Erola et al., 2010; Torra, 2004; Zhang et al., 2010) in which textual data are evaluated in a categorical fashion.
- *Least Common Subsumer centroid* (LCSC): this method relies on a background taxonomy/ontology and selects, as centroid, the concept that subsumes all the different values found in the dataset. When dealing with multivariate data, the subsumer is a tuple so that each value individually subsumes all the values of an attribute. According to the input data and the background ontology, the resulting centroid may or may not belong to input data. This is also a common strategy used by semantically-grounded methods like (Abril et al., 2010).
- *Consensus-based centroid* (CC): reproducing the method proposed in (Guzman-Arenas et al., 2011; Guzman-Arenas & Jimenez-Contreras, 2010), the centroid is selected as the record in the dataset that minimises the average *confusion* to all other records. A background taxonomy/ontology is also used to map the objects' values to concepts so that the confusion can be computed as the number of descending links in the hierarchy from *a* to *b* when substituting the former with the latter. Note that, in the original method, the background ontology is constructed according to the individual values in the input dataset. This hampers the method when input values cannot be taxonomical organised among them (e.g., when most values correspond to leaves in the taxonomy). To enable a more favourable situation against our methods, we used the same background ontology to compute the confusion between values.

In all cases in which an ontology is required as background knowledge, we used WordNet to retrieve candidate centroids and/or compute semantic similarity measures.

4.3. Evaluation with *scenario 1*

The first scenario considers the whole dataset (i.e., 975 records). We quantify the centroid's accuracy in preserving data semantics by computing the aggregated semantic distance (using eq. 9 and the logSC measure - eq. 4 (Sánchez et al., 2012)) between the centroid obtained with each of the methods listed above and all the objects in the dataset. Results are presented in Fig. 6.

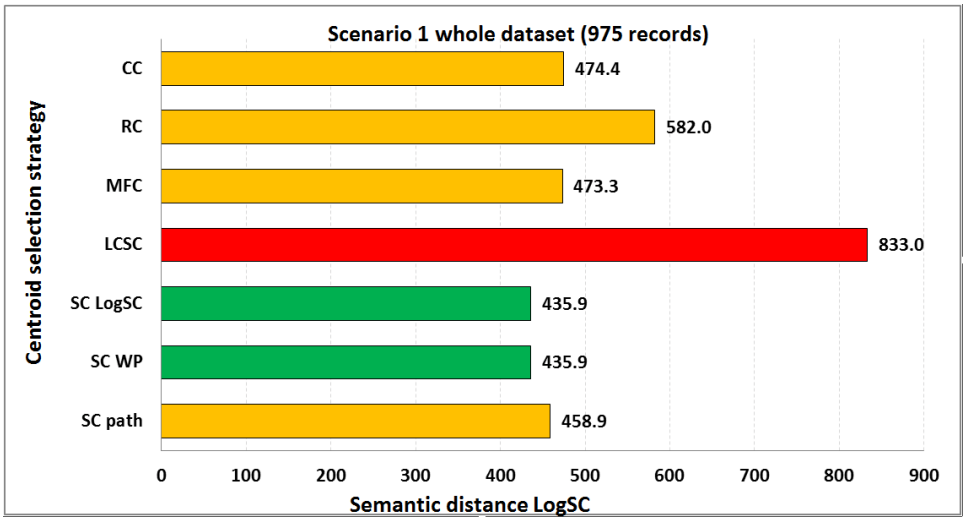


Fig. 6. Comparison of centroid construction strategies.

Analysing the results, we observe that our method (SC) is able to improve related works in all cases. Regarding the different semantic measures used to measure the distance used during the construction of the centroid (eq. 6), LogSC and W&P measures give identical results, whereas the path-based measure provided slightly worse results. As stated in section 3.1, measures relying only on the taxonomical path length provide less accurate results than other strategies because the path omits semantic evidences explicitly available in the ontology that are relevant in similarity estimation (such as the relative depth of concepts in the taxonomy or the presence of multiple inheritance relations).

Compared to related works, we observe that the worst method was the one using the LCS of the whole data as the centroid (LCSC). This method does not quantify the semantic distance between values and does not consider their distribution (i.e., number repetitions). It only assumes that the most specific common ancestor to all the values will serve as a good representative for all the objects. As discussed in section 2, however, this method is hampered by the presence of value outliers that are semantically far to most other values in the dataset, forcing the generalisation of the LCS to very abstract (and, hence, semantically distant) concepts. Even though values in the evaluated dataset are relatively homogenous (see Fig. 6), the variability of responses (with 118 answers unique answers) forced the generalisation to the root concept of the ontology (i.e., entity) as the LCS. Hence, the centroid is semantically distant to most values in the dataset, resulting in a high loss of semantics.

The other semantically grounded related work, based on measuring the *confusion* between input values (CC), significantly improved the above method by numerically quantifying the cost of substituting a value with a specialisation. However, the fact that this quantification is based only on the path (i.e., number of descending

levels in the taxonomy), that only input values are considered as centroid candidates, and that the cost of replacing a value with a subsumer is not considered, provided worse results than our methods. Note that in this case, the method is favoured by the fact that a large ontology as WordNet is used to compute confusion values, rather than a constrained taxonomy constructed ad-hoc from input values.

Finally, methods based on data distribution performed reasonably well thanks to the characteristics of data. On the one hand, the MFC method, based on selecting the most frequently appearing value tuple, is favoured by the fact that a tuple with a significant 115 repetitions is found. Hence, by selecting this tuple as centroid it is achieved a zero semantic distance to 115 records of the dataset. Moreover, this tuple corresponds to nature-related reasons that are similar to other frequently appearing tuples, resulting in a low semantic distance. As a result, the aggregated semantic distance is implicitly minimised. A similar situation is observed for the naïve strategy (RC) that, thanks to the presence of numerous sets of frequently appearing tuples, it increases the chance to randomly select a tuple that results in a low semantic distance. These behaviours, however, depend very closely on the characteristics of the input data, in particular, on the fact that predominant values exist. When dealing with more scattered and less cohesive datasets, results are likely to worsen.

Regarding the computational cost of the different centroid calculation strategies, in an univariate setting, we observe that our method (SC) scales as $O(|H|\cdot n)$, where $|H|$ is the number of concepts in the minimum hierarchy modelling values in the input dataset (i.e., the number of centroid candidates), and n is the number of *distinct* values in the dataset (to which, for each centroid candidate, the weighted sum of distances should be computed). Note that, when dealing with categorical data, n is usually much lower than the size of the dataset. This is because, as shown in the data distribution analysis, categorical data usually present a limited set of modalities that tend to repeat. Since multivariate datasets (with m attributes) are considered as sets of univariate attributes, the cost is maintained at a polynomial size $O(|H_1|\cdot n_1+\ldots+|H_m|\cdot n_m)$. In comparison, the *consensus-based centroid* (CC) selects the record in the dataset that minimises the *confusion* against all other records. Confusion corresponds to the number of descending links between records in the background ontology, which in the worst case would be the maximum taxonomical depth of H. For univariate data, this scales $O(h\cdot r)$, where r is the number of records in the dataset and h is the maximum taxonomical depth of H. For multivariable datasets, this results in $O(h_1\cdot r+\ldots+h_m\cdot r)$, where h_i is the maximum depth of the H_i for the i^{th} attribute. Even though $h_i<|H_i|$, as stated above, typically $r>n_i$. As a result, the final cost is similar to that of our method. The *Least Common Subsumer centroid* (LCSC) simply picks the LCS generalising all the values of each attribute, resulting in a cost of $O(|H|)$ for univariate data and $O(|H_1|+\ldots+|H_m|)$ for multivariate datasets. Hence, the expected runtime is lower than that of our method but at the cost of sensibly worse results. Finally, the non-semantic approach (*Most frequently appearing centroid* (MFC)) offers the lowest cost ($O(n)$, being n the number of distinct values/tuples in univariate/multivariate datasets), at the cost of lacking both distance and semantic optimisations.

4.4. Evaluation with s*cenario 2*

The second scenario evaluates, individually, each of the seven clusters in which the dataset has been classified, as described in section 4.1. The semantic distance of the obtained centroids for each cluster using the

different strategies to build centroids is shown in Fig. 7. In addition, we have calculated the accumulated sum of distances for all the clusters and each strategy.

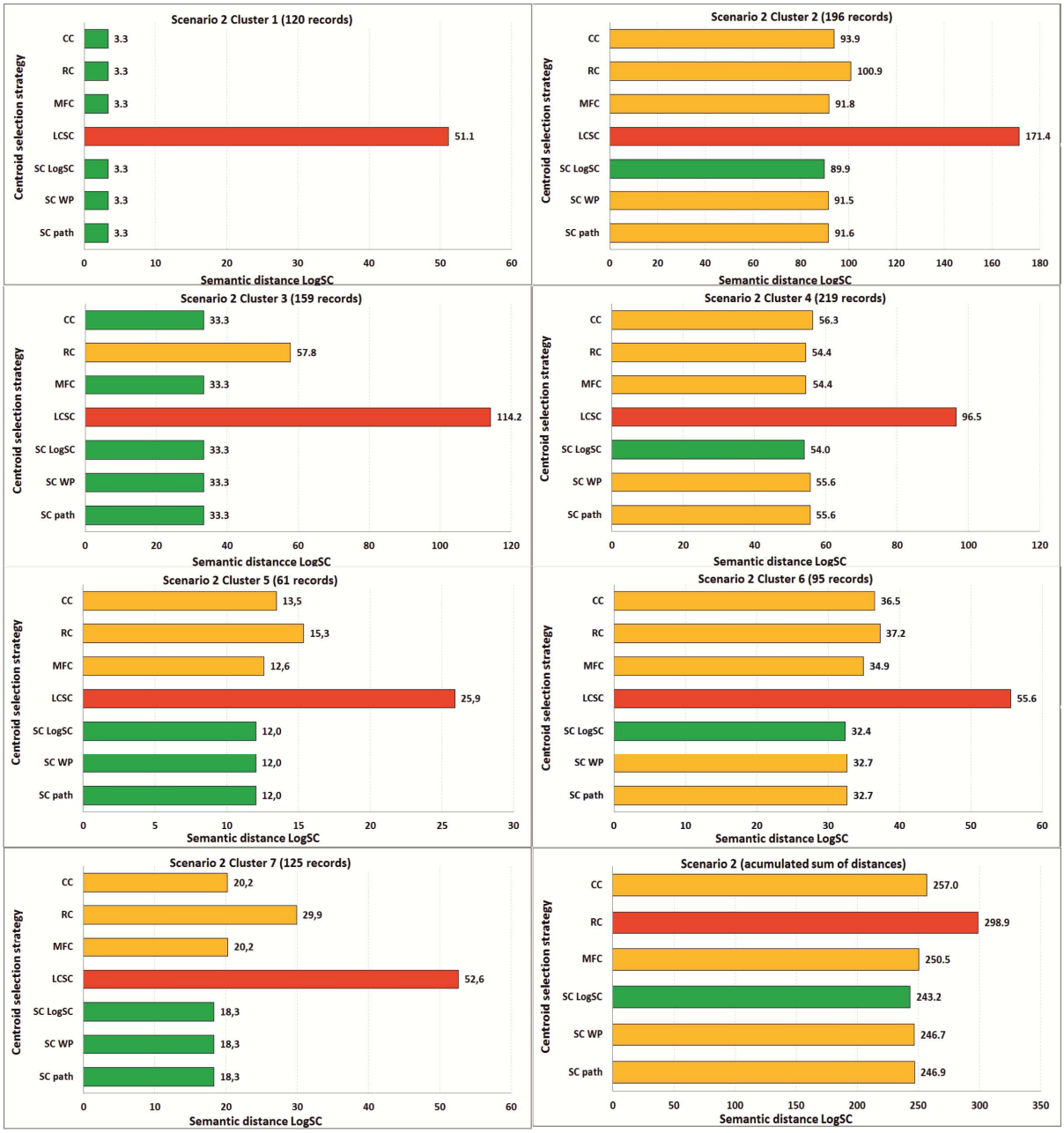


Fig. 7. Semantic distances between each cluster and its centroid.

The consideration of clustered data allows analysing the differences observed between the different centroid construction strategies according to the characteristics of each cluster. On the one hand, when dealing with highly homogenous clusters (like *Cluster1*), all approaches, except the LCSC, tend to provide the same – optimum- results. In fact, analysing the frequency distribution of data in Cluster1 (Fig. 7), we observe a clear predominant value (with 109 repetitions in a cluster with 120 elements). As a result, this value is selected as the centroid of the cluster by all strategies except the LCSC, which selects a more abstract taxonomical ancestor that also subsumes scarce values. This is the most favourable scenario for methods based on data distribution

because the most frequent tuple is the optimum representative. *Cluster3* also presents a similar behaviour, even though the RC approach has not selected the most frequent tuple and, hence, the distance is higher. In this case, the chance to randomly select the optimal centroid is lower than with *Cluster1* due to the lower predominance of the most frequent tuple (the probability of selecting randomly the mode is p=0.98 for the *cluster1* and p=0.38 for the *cluster3*). On the other hand, when dealing with less uniform clusters with no predominant values (like *Cluster5*, *Cluster6* or *Cluster7*), only our proposal is able to obtain optimal centroids. In this case, the selection of a taxonomical subsumer results in a lower semantic distance than the most frequent value tuple. This shows that distributional methods are not ideal when dealing with less cohesive datasets, because centroids are selected only from those values in the input dataset (a characteristic also shared with CC). Hence, their accuracy heavily depends on the distribution of the objects. Only our methods and the LCSC expand the search space to other concepts extracted from the background ontology. However, as stated above, the LCSC strategy is hampered by the presence of outliers and the lack of frequency analysis, providing the worst results from the bunch.

In addition, we have calculated the accumulated sum of distances for all 7 clusters (with their corresponding 7 centroids) to have a measure of the overall loss of semantics when the objects are substituted by their respective centroid, according to the clustering assignment. We observe that our proposal provides the best results.

5. Conclusions and future work

Centroids are usually needed in data analysis algorithms, both for discovering the central value of a dataset from which the data aggregation can be performed and also for constructing representative values of resulting clusters. When dealing with numerical data (i.e., continuous values) it is straightforward to compute the optimum value that minimises the distance to all the objects in the dataset by means of averaging operators. Textual attributes, on the contrary, represent discrete elements that should be interpreted according to their semantic content. The preservation of data semantics is of outmost importance when performing data transformation over textual data (Martinez et al., 2011; Sánchez et al., 2012; Sanchez et al., 2010) in order to ensure their utility, which is precisely the aim of data analysis methods. Hence, the accuracy of these methods closely depends on the goodness of the constructed centroids.

This paper proposes a new approach to construct centroids of datasets with textual attributes with special emphasis on the preservation of data semantics. It has also been generalised to the case of multivariate data while ensuring the scalability of the centroid construction with large multivariate datasets. On one hand, the use of large ontologies like WordNet as background knowledge allowed us to retrieve and evaluate a larger set of centroid candidates in comparison with previous works (Bai et al., 2011; Domingo-Ferrer & Torra, 2005; Guzman-Arenas et al., 2011; T. Huang et al., 2010; Torra, 2004; Zhang et al., 2010), which restricted centroid candidates to values found in the input dataset. As a result our method is able to pick more accurate centroids and, hence to better retain the semantics of data. The only drawback was the slightly higher computational cost, derived from the fact that more candidates need to be evaluated to optimise the results. On the other hand, semantic similarity measures are used in conjunction with data distribution to select the candidate that best minimises the semantic distance to all the values in the dataset. This ensures an optimal preservation of data

semantics according to the background knowledge and the similarity measure. In comparison, other semantically-grounded methods (Abril et al., 2010; Guzman-Arenas et al., 2011) fail at minimising the distances between records and the centroid due to the suboptimal (Guzman-Arenas et al., 2011) or even null (Abril et al., 2010) semantic assessment. The evaluation also showed that the, apparently coherent, strategy of using the LCS as the centroid (Abril et al., 2010) provided the worst results due to the need of generalising outliers. This shows the convenience of taking into account data distribution in conjunction with their semantics as our method does. In fact, methods based on using the most frequent tuple as the centroid (Bai et al., 2011; Cao et al., 2012; Domingo-Ferrer & Torra, 2005; Erola et al., 2010; Torra, 2004; Zhang et al., 2010) performed reasonably well due to they tend to implicitly consider the distributional distances (but not the semantic distances).

As future work, we plan to extend the evaluation to other datasets with more heterogeneous attributes. Moreover, we would like to study the case of dealing with attributes belonging to different domains of knowledge, using different domain ontologies for each one (e.g. large medical ontologies like SNOMED-CT). As shown in section 3.3, the minimal subsumer hierarchies are independent for each attribute and, hence, each one can be constructed from different ontologies. We also plan to apply our method to concrete data analysis algorithms dealing with textual attributes such as semantic clustering (Batet et al., 2008) and masking methods focused on the anonymisation of textual microdata (Martinez et al., 2010; Martinez et al., 2011). By analysing the results obtained by those algorithms we would be able to quantify the influence of the centroid construction in data utility.

Acknowledgements

We would like to thank the Observatori de la Fundació d'Estudis Turístics Costa Daurada and the Delta de l'Ebre National Park for providing the data. This work was partly funded by the Spanish Government through the projects CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES" and eAEGIS TSI2007-65406-C03-02, and by the Government of Catalonia under grant 2009 SGR 1135 and 2009 SGR-01523. Sergio Martínez Lluís is supported by a research grant of the Ministerio de Educación y Ciencia (Spain).

References

Abril, D., Navarro-Arribas, G., & Torra, V., 2010. Towards semantic microaggregation of categorical data for confidential documents, Proceedings of the 7th international conference on Modeling decisions for artificial intelligence. Publishing, Perpignan, France, pp. 266-276.

Ahmed, A., Bora, B., & Bhattacharyya, D.K., 2005. A Frequency Vector Based Hierarchical Clustering Algorithm for Mixed Type of Data, ADCOM. Publishing.

Bai, L., Liang, J., & Dang, C. 2011. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. *Knowl.-Based Syst.*, *24*, 785-795.

Batet, M., Sanchez, D., & Valls, A. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *J Biomed Inform*, *44*, 118-125.

Batet, M., Valls, A., & Gibert, K., 2008. Improving classical clustering with ontologies, Proceedings of the 4th World conference of the IASC. Publishing, Japan, pp. 137-146.

Batet, M., Valls, A., & Gibert, K., 2010. Performance of ontology-based semantic similarities in clustering, Proceedings of the 10th international conference on Artificial intelligence and soft computing: Part I. Publishing, Zakopane, Poland, pp. 281-288.

Bollegala, D., Matsuo, Y., & Ishizuka, M., 2007. Measuring semantic similarity between words using web search engines, Proceedings of the 16th international conference on World Wide Web. Publishing, Banff, Alberta, Canada, pp. 757-766.

Cao, F., Liang, J., Li, D., Bai, L., & Dang, C. 2012. A dissimilarity measure for the k-Modes clustering algorithm. *Knowl.-Based Syst., 26*, 120-127.

Cilibrasi, R.L., & Vitanyi, P.M.B. 2007. The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng., 19*, 370-383.

Cimiano, P. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc.

Chen, P., & Wang, Y. 2004. Efficient clustering algorithm for categorical and mixed typed attributes. *Computer Eng. Appl., 1*, 190-191.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., & Sachs, J., 2004. Swoogle: a search and metadata engine for the semantic web, Proceedings of the thirteenth ACM international conference on Information and knowledge management. Publishing, Washington, D.C., USA, pp. 652-659.

Do, H.-J., & Kim, J. 2009. Clustering Categorical Data Based on Combinations of Attribute Values. *Int. J. Innovative Comp Inf Control, 5*, 4393-4405.

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

Domingo-Ferrer, J., & Mateo-Sanz, J.M. 2002. Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Trans. on Knowl. and Data Eng., 14*, 189-201.

Domingo-Ferrer, J., & Torra, V. 2005. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. *Data Min. Knowl. Discov., 11*, 195-212.

Erola, A., Castella-Roca, J., Navarro-Arribas, G., & Torra, V., 2010. Semantic microaggregation for the anonymization of query logs, Proceedings of the 2010 international conference on Privacy in statistical databases. Publishing, Corfu, Greece, pp. 127-137.

Everitt, B.S., Landau, S., Leese, M., & Stahl, D. 2011. *Cluster Analysis, 5th Edition*.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Gaeta, M., Orciuoli, F., & Ritrovato, P. 2009. Advanced ontology management system for personalised e-Learning. *Knowl.-Based Syst., 22*, 292-301.

Greenacre, M., & Hastie, T. 2010. Dynamic visualization of statistical learning in the context of high-dimensional textual data. *Web Semantics: Science, Services and Agents on the World Wide Web, 8*, 163-168.

Guzman-Arenas, A., Cuevas, A.-D., & Jimenez, A. 2011. The centroid or consensus of a set of objects with qualitative attributes. *Expert Syst. Appl., 38*, 4908-4919.

Guzman-Arenas, A., & Jimenez-Contreras, A. 2010. Obtaining the consensus and inconsistency among a set of assertions on a qualitative attribute. *Expert Syst. Appl., 37*, 158-164.

Han, J. 2005. *Data Mining: Concepts and Techniques* 2nd. ed. Morgan Kaufmann, cop., San Francisco (Calif.).

Huang, T., Yu, Y., Guo, G., & Li, K. 2010. A classification algorithm based on local cluster centers with a few labeled training examples. *Knowl.-Based Syst., 23*, 563-571.

Huang, Z. 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov., 2*, 283-304.

Jiang, J.J., & Conrath, D.W., 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, International Conference Research on Computational Linguistics (ROCLING X). Publishing, p. 9008.

Keikha, M., Khonsari, A., & Oroumchian, F. 2009. Rich document representation and classification: An analysis. *Knowl.-Based Syst., 22*, 67-71.

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

Lin, J.-L., Wen, T.-H., Hsieh, J.-C., & Chang, P.-C. 2010. Density-based microaggregation for statistical disclosure control. *Expert Syst. Appl., 37*, 3256-3263.

Maniruzzaman, M., Chaves, J., McGee, C., Ma, S., & Jr, R.S., 2002. CHTE quench probe system: a new quenchant characterization system, 5th International Conference on Frontiers of Design and Manufacturing (ICFDM). Publishing, p. 619.

Martinez, S., Sanchez, D., & Valls, A., 2010. Ontology-Based Anonymization of Categorical Values, in: V. Torra, Y. Narukawa & M. Daumas (Eds.), Modeling Decisions for Artificial Intelligence. Publishing, pp. 243-254.

Martinez, S., Sanchez, D., Valls, A., & Batet, M. 2011. Privacy protection of textual attributes through a semantic-based masking method. *Inf. Fusion, DOI: 10.1016/j.inffus.2011.03.004*.

Martínez, S., Sánchez, D., Valls, A., & Batet, M., 2010. The Role of Ontologies in the Anonymization of Textual Variables, Proceeding of the 2010 conference on Artificial Intelligence Research and Development: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence. Publishing, pp. 153-162.

Nelson, S.J., Johnston, D., & Humphreys, B.L., 2001. Relationships in Medical Subject Headings, in: K.A. Publishers (Ed.), Relationships in the Organization of Knowledge. Publishing, New York, pp. 171-184.

Park, H.-S., & Jun, C.-H. 2009. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl., 36*, 3336-3341.

Pedersen, T., Patwardhan, S., & Michelizzi, J., 2004. WordNet::Similarity: measuring the relatedness of concepts, Demonstration Papers at HLT-NAACL 2004. Publishing, Boston, Massachusetts, pp. 38-41.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. 1989. Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern., 19*, 17-30.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy, Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1. Publishing, Montreal, Quebec, Canada, pp. 448-453.

Samarati, P., & Sweeney, L. 1998. Protecting Privacy when Disclosing Inf.: k-Anonymity and Its Enforcement through Generalization and Suppression. *Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory*.

Sanchez, D., & Batet, M. 2011. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *J Biomed Inform, 44*, 749-759.

Sanchez, D., Batet, M., & Isern, D. 2011. Ontology-based information content computation. *Knowl.-Based Syst., 24*, 297-303.

Sánchez, D., Batet, M., Isern, D., & Valls, A. 2012. Ontology-based semantic similarity: A new feature-based approach. *Expert Syst. Appl., 39*, 7718-7728.

Sanchez, D., Batet, M., Valls, A., & Gibert, K. 2010. Ontology-driven web-based semantic similarity. *J. Intell. Inf. Syst., 35*, 383-413.

Shin, K., Abraham, A., & Han, S.-Y., 2006. Enhanced Centroid-Based Classification Technique by Filtering Outliers, Proceedings of TSD. Publishing, pp. 159-163.

Torra, V., 2004. Microaggregation for Categorical Variables: A Median Based Approach, in: J. Domingo-Ferrer & V. Torra (Eds.), Privacy in Statistical Databases. Publishing, pp. 518-518.

Torra, V., & Narukawa, Y. 2007. *Modeling Decisions*.

Turney, P.D., 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, Proceedings of the 12th European Conference on Machine Learning. Publishing, pp. 491-502.

Varde, A.S., Rundensteiner, E.A., Ruiz, C., Brown, D.C., Maniruzzaman, M., & Sisson, R.D., 2006. Designing semantics-preserving cluster representatives for scientific input conditions, Proceedings of the 15th ACM international conference on Inf. & knowledge management. Publishing, Arlington, Virginia, USA, pp. 708-717.

Ward, J.H. 1963. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc., 58*, 236-244.

Wong, S.V., & Hamouda, A.M.S. 2003. The development of an online knowledge-based expert system for machinability data selection. *Knowl.-Based Syst., 16*, 215-229.

Wu, Z., & Palmer, M., 1994. Verbs semantics and lexical selection, the 32nd annual meeting on Association for Computational Linguistics. Publishing, Las Cruces, New Mexico, pp. 133-138.

Yihui, L. 2012. Dimensionality reduction and main component extraction of mass spectrometry cancer data. *Knowl.-Based Syst., 26*, 207-215.

Zhang, W., Yoshida, T., Tang, X., & Wang, Q. 2010. Text clustering using frequent itemsets. *Knowl.-Based Syst., 23*, 379-388