Under consideration for publication in Knowledge and Information Systems

Introducing semantic variables in mixed distance measures. Impact on hierarchical clustering

Karina Gibert[1], Aida Valls[2] and Montserrat Batet[2]

[1] Universitat Politècnica de Catalunya, BarcelonaTech

Barcelona, Spain; karina.gibert@upc.edu

[2] Dept. Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili

Tarragona, Spain

Abstract. Today it is well known that taking into account the semantic information available for categorical variables sensibly improves the meaningfulness of the final results of any analysis. The paper presents a generalization of mixed Gibert's metrics, which originally handled numerical and categorical variables, to include also semantic variables. Semantic variables are defined as categorical variables related with a reference ontology (ontologies are formal structures to model semantic relationships between the concepts of a certain domain).

The Superconcept-based distance is introduced to compare semantic variables taking into account the information provided by the reference ontology. A benchmark shows the good performance of Superconcept-based distance with respect to other proposals, taken from the literature, to compare semantic features. Mixed Gibert's metrics is generalized incorporating Superconcept-based distance.

Finally, two real applications based on touristic data shows the impact of the generalized Gibert's metrics in clustering procedures, and, in consequence, the impact of taking into account the reference ontology in clustering. The main conclusion is that the reference ontology, when available, can sensibly improve the meaningfulness of the final clusters.

Keywords: Clustering; Metrics; Numerical and Categorical variables; Semantic Data; Ontology

Received Jun 11, 2012
Revised Jan 17, 2013
Accepted May 09, 2013

1. Introduction

Describing the structure or obtaining knowledge of complex systems is known as a difficult task. Different research areas are dealing with this issue, such as Statistics, Artificial Intelligence, Information Systems and Data visualization. *Knowledge Discovery and Data Mining* (KDD) is a research area where all those fields interact in order to extract useful knowledge from data(**?**).

Clustering algorithms are well-known data mining methods and the more used for partitioning data into a certain number of homogeneous groups or clusters (**?**). In fact, we agree with the idea that a number of real applications in *KDD* either require a clustering process or can be reduced to it (**?**). Recent research has evidenced that about 60% of real data mining applications use clustering. Also, identification of classes is one of the basic methods used by human beings in structuring the world and that's why, many expert systems are indeed classifiers.

In the kernel of clustering algorithms, comparisons between objects are used to priorize the objects to be clustered first. Originally (**?**), clustering was conceived on numerical data, whose inherent geometric properties could be exploited to compare points, mainly by means of distances or related functions. This is the basis to guide the class construction.

However, in real applications, it is usual to get data matrices containing also some categorical variables as relevant as the numerical ones in the decision processes. As an example, the color of the water inside the bioreactor of a wastewater treatment plant, is as important as the concentration of nutrients, because both are quality indicators used by the head of the plant to make decisions about the proper treatment. The nutrients concentration values are obtained from biochemical analysis in the laboratory and when they go over some threshold, the biomass is incremented in the tank for a quicker degradation of organic matter. The color of the water is directly observed by the head of the plant, and when it is red, it indicates toxic alguee formation, which requires specific actions also. Such an scenario illustrates the importance of analyzing heterogeneous data matrices, where both numerical an non-numerical variables are included. In that case, particular attention is required to perform objects comparison.

Literature provides many references on the topic of analyzing what is named as messy data or heterogeneous data, depending on the contexts (**?**), (**?**), (**?**). Different strategies are proposed. Among them, the most popular ones are:

- to a-priori discretize all numerical variables and get an homogenous data matrix with only categorical variables. The main criticism to this practice is that discretization always implies a loose of information and can also involve a bias, both factors likely to impact in the robustness of final results.
- to define what is known as compatibility measure, which permit comparisons among heterogeneous objects. One of the main advantages of this approach is that interactions between numerical and non-numerical variables are taken into account.

Previous experiences (**?**), (**?**) show compatibility measures as the best approach for the particular context of clustering where global interaction among variables is relevant. Compatibility measures, as Gower coefficient (**?**), the generalization of Minkowski metrics by Ichino-Yaguchi (**?**) or the Gibert's mixed metrics (**?**), among others, often include different expressions according to the

type of every variable, and allow homogeneous treatment of different types of variables, by keeping them in their original form.

In most proposals, the compatibility measures contain a component for numerical variables and another for non-numerical, treating all non-numerical variables in the same way. Particularly most data analysis approaches treat all non-numerical (or categorical) variables under a syntactic point of view (based on equality/inequality of terms), or in the most sophisticated cases (**?**) introducing some weights related with the rarity of the terms (measurable through the χ^2-distance), which indicates que informational load of the term itself. Recently, in the field of KDD, some works address the topic of clustering only with categorical data and, even if they use more sophisticated criteria to compare objects, they still are based on syntactic fundamentals. The ROCK (**?**) algorithm groups first the more linked elements, a link being a common neighbor of both elements; neighborhood being determined upon a given threshold. CACTUS (**?**) provides a scalable algorithm which clusters the most interacting elements first, using complex co-occurrence related concepts to evaluate those interactions. Even in that case, the whole process is based in computing some frequencies and coincidences over the data matrix.

However, in many applications, part of the categorical variables describing an object can be semantically interpreted, as their values refer to concepts in a certain domain field. Think, for example of a variable reporting the *type of building*, where the degree of difference between *church* and *cathedral* should be smaller than the one between *church* and *office*, regardless how frequent are those terms in the data matrix or how often they co-occur in a document or in a data matrix. Furthermore, it may happen that additional semantic knowledge about these variables is available and formalized and might be interesting to take it into account to improve the comparison between objects and to improve cluster detection. Some works already point to that issue. The works (**?**) or (**?**) show the benefits of introducing ontologies to improve topic identification in text mining applications, particularly for document clustering. In (**?**) a discussion is provided about the added value of semantics with respect to the information provided by the algebraic structure involved in distances or similarities and their benefits on better finding underlying clusters in health-related applications.

Thus, integrating the semantic of the terms in a general compatibility measure to compare two heterogeneous objects is of high interest. It enables a better approach to the way of thinking of the experts and, in consequence, to get more coherence between the results and the prior knowledge of the phenomenon.

Finding a proper way to integrate the semantic of the terms in the context of clustering with heterogeneous objects is the main goal of this paper. This work introduces the concept of semantic variable to model those categorical variables for which additional semantic knowledge is available in a reference ontology. A generalization of the compatibility Gibert's mixed metrics(**?**) to also include semantic variables is provided. The generalization is based on the introduction of the *Superconcept-based distance*, which computes similarity over terms by taking into account the relationship of the compared terms in the reference ontology. The benefits of using generalized Gibert's mixed metrics for clustering heterogeneous data matrices with all numerical, categorical and semantic variables is addressed in some real applications.

The structure of the paper is the following: Section §2 provides related work, section §3 introduce the generalization of Giber't mixed metrics into a new measure to compare objects including semantic information, when available. Section

§4 studies different possibilities of introducing semantic information into the comparison between objects, focusing on similarity measures based on the taxonomical structure of a reference ontology. Section §5, presents the *Superconcept-based distance* as a new proposal for semantic variables, also based on their taxonomical structure. Section §6 discuss the convenience of metrical structure for clustering purposes and analyzes the metrical properties of the proposal. Section §7 is devoted to evaluate all the studied semantic measures against *Superconcept-based distance* using a standard benchmark based on human judgments. Section §8 evaluates the impact of using the extended Gibert's mixed metrics with the *Superconcept-based distance* in clustering with two case studies. Finally, the last section presents the conclusions.

2. Related work

As said before the main goal of this work is to improve the clustering results for heterogeneous matrices, by introducing semantic information for those categorical variables with available semantic models. This will be approached by extending the comparison measure between objects to include semantic variables.

In the literature, we can find different approaches to bring semantics to the object comparisons. The field of distance metrics learning (**?**), (**?**), (**?**) is mainly concerned with the idea of getting the final distances directly from experts and automatically learn the metrics that fits with those subjective similarities implicitly managed by experts. Basically, this approach can deal with any kind of variable together, since the expert implicitly takes into account both the type of variable and its meaning in his/her comparisons. Inter-experts coherence is important in this context for the robustness and consistency of the results. This approach is useful when experts are able to quantify similarities among objects, which is not always the case.

In the fuzzy approach (**?**), linguistic labels are introduced as a paradigm for including uncertainty into data representation itself, and the specific semantics of the terms is represented as membership functions.

Till now, none of the proposals directly applicable to cluster classical crisp heterogeneous data matrices is considering the semantics of the terms under comparison. This has been widely accepted by now, probably because of the lack of proper tools for representing and managing conceptual values together with their semantics. In our approach, the semantics, the expertise and the background domain knowledge is not expressed in a quantitative form and experts are not required to directly quantify similarities among terms, nor membership functions, nor a priori probability distributions (as it happens in Bayesian approaches).

Background domain knowledge is expressed by means of ontologies, a more expressive formalism that permits to express relationships among concepts under the logical paradigm, which sometimes is more natural to the expert. Ontologies have emerged in the last years as a fundamental tool for formalizing and representing domain knowledge. They offer a formal and explicit description of a shared conceptualization (**?**), providing a graphical model in which semantic interrelations are modeled as links between concepts. They are highly flexible and provide enough expressiveness for our purposes.

Recent research in the field of computational linguistics shows that using an

ontology as background knowledge can improve document clustering analysis (**?**) (**?**) (**?**) (**?**) (**?**). In this field the main objective is to interpret the text meaning by using the ontology and to establish an ordering among documents based on their similarities. Many references to the use of ontologies for improving document classification and clustering can be found. Most of them use Wordnet as reference ontology. A representative paper is (**?**), which uses textual data analysis techniques based on the corpus numerization (or the bag of words) and the principles of multivariate factorial analysis techniques. The ontology is used to enrich the corpus numerization as well as to compute prototypical documents for a class. The main difference with respect to our work is that we are working in a classical 2-dimensional data matrices, where objects are described by several variables. In our approach the ontology is used to refine the calculation of distances of some qualitative variables of the data matrix; those for which semantic information is available in a reference ontology: Then, the clustering can be applied to whatever kind of object (patients, tourists, villages, cars, etc) that can be described in a structured way by their values on a predefined set of variables numerical or qualitative variables.

In (**?**) the documents to be clustered are previously modeled by means of a graph structure which is build taken the ontology into account. Several measures are defined over those graph and they are used for the clustering process. In our approach the part of the objects described by means of semantic variables is also transformed to a more structured representation which takes into account the background ontology and the measures are also defined over the new representation. In our case, an extension of the adjacency matrix induced by the ontology is used, which records the existence of ascendent path between terms.

It is worth to note that, in most of these references, partitional clustering algorithms are used, mainly because they are low class of complexity algorithms, being most of them linear or almost linear. But in real data mining applications it is hard to establish in advance the number of clusters as required by these methods (**?**) (**?**). Hierarchical clustering methods avoid this kind of assumptions and permits to discover the number of real clusters from the self nature of the data set. Also, the family of density based methods, like DBSCAN (**?**) or OPTICS(**?**) could be suitable for this purpose, although in this work hierarchical clustering is used.

Although most references found in the literature regards to document classification and text mining, authors are convinced that the benefits of using ontologies to enrich data analysis are not restricted to computational linguistic problems.

In fact, ontologies have been recently used in data mining under several approaches, mainly in processes related to learning. Thus, ontologies have been introduced to improve classical machine learning classifiers (**?**) (**?**) (**?**) (**?**). Examples can be also find in decision trees learning (**?**), in neural networks (**?**), Bayesian networks (**?**) or even in case-based reasoning (**?**). In non-supervised methods, ontologies have been mainly used in the fields of Web mining (**?**) or gene analysis (**?**). Also they have been used to improve association rules mining, like in (**?**) (**?**), (**?**) (**?**) (**?**). In that case, the main issue is to reduce the number of association rules mined by different strategies, like using the ontology to prune the senseless associations (**?**), or getting more general associations that subsume big number of particular patterns (**?**), or to generate constraints to the construction of associations by reducing the searching space (**?**). Few references can be found introducing ontologies to guide the clustering process in data mining,

but there has been some contributions in this line in the fields of text mining (**?**), or in biomedicine dealing with genetic data (**?**) (**?**), (**?**). In (**?**) genes are clustered according to their expression patterns by using a semantic distance computed on the basis of the complete set of annotations of every gene obtained using the Gene Ontology; in the paper several measures are presented. Among them, Czearowski-Dice similarity (**?**) uses similar ideas to the ones presented on our work, but our proposal holds metrical properties, with some benefic implications on hierarchical clustering. In the field of clustering, the ontology is not used to reduce the number of patterns, like in association mining, but to improve the quality of the comparisons between the objects, in such a way that the final classes produce more compact classes from a conceptual point of view, by grouping more *conceptually-similar* objects.

All the related works are strictly based on the availability of semantic information about objects. However, when objects are described by means of a fixed set of interest variables, not always all these variables can be connected to some extra semantic knowledge base (i.e. an ontology). For this reason, in this work we have considered a more general approach in which:

a) a) data to be clustered is represented in a classical 2-dimensional data matrix, with a definite set of variables in columns
b) b) only part of the qualitative variables can be linked to available ontologies that permit the semantic interpretation of the variable values (i.e. its modalities are connected to concepts in the ontology)
c) c) classical numerical variables or measurements can also be part of the data matrix
d) d) some of the categorical variables do not have semantic extra information available and must be treated only from a syntactic point of view.

None of the referred works combines numerical, syntactic and semantic information in a single and integrated analysis in the context of clustering.

3. Mixed semantic metrics

In this paper, an approach to compare objects that are described with numerical, categorical and a set of semantically interpretable variables is presented. We are naming these variables as *semantic variables* as their values can be interpreted based on background knowledge and become *concepts* rather than simple *modalities*. Often, semantic variables can be extracted from a textual description of the object.

In this work background knowledge is represented in form of ontologies, as they are a powerful tool for semantic knowledge management, easy to process inside data analysis methods.

The proposal can be used in any application domain where the comparison between objects described with those types of variables is required, like, for example, analyzing the tourist interests of people, which involves numerical (age, yearly income), categorical (sex, original country) or semantic (hobbies, cultural preferences) relevant variables to describe the tourist itself.

Because of the variety of very complex domains in which it has been successfully applied — functional disability in elderly persons (**?**), dependency in schizophrenia (**?**), waste water treatment plants (**?**), neurorehabilitation in brain

damage (**?**) — and the good performance shown in cluster analysis (**?**) with respect to other proposals in the literature, the proposal presented in this work extends Gibert's mixed metrics (**?**) to incorporate semantic variables.

The standard input of a clustering algorithm is a data matrix with the values of K variables $X_1 \dots X_K$ observed over a set $\mathcal{I} = \{1, \dots n\}$ of individuals. Variables are in columns, while individuals in rows. Cells contain the value (x_{ik}), taken by individual $i \in \mathcal{I}$ for variable $X_k, (k = 1 : K)$. First the distance to be generalized is briefly introduced.

3.1. Extending Gibert's mixed metrics to semantic variables

In (**?**) Gibert introduces the mixed metrics as a weighting between the normalized Euclidean metrics for numerical variables and a rewriting of the χ^2 metrics for qualitative ones which do not require expansion to complete incidence tables. The Gibert's proposal is based on the idea that χ^2 metrics (**?**) upon qualitative variables is directly related with the quantity of information provided by the variable itself (**?**). In this work, Gibert's metrics is extended to a third type of variable, named *semantic variable*, by adding a third term to the original expression. The extended Gibert's mixed metrics is defined as:

$$d^2_{(\alpha,\beta,\gamma)}(i,i') = \alpha d^2_\zeta(i,i') + \beta d^2_\mathcal{Q}(i,i') + \gamma d^2_\mathcal{S}(i,i'), \;\; (\alpha,\beta,\gamma) \in [0,1]^3, \;\; \alpha+\beta+\gamma = 1 \tag{1}$$

, being $(\alpha, \beta, \gamma) \in [0,1]^3$, $\zeta = \{k : X_k \text{ numerical variable}, k = 1 : K\}$, $\mathcal{Q} = \{k : X_k \text{ categorical variable}, k = 1 : K\}$, $\mathcal{S} = \{k : X_k \text{ semantic variable}, k = 1 : K\}$, $d^2_\zeta(i,i')$ the normalized euclidean metrics for numerical variables, and $d^2_\mathcal{Q}(i,i')$ a rewriting of the χ^2 metrics to be computed directly on a symbolic representation of the categorical variable (see expression 3), $d^2_\mathcal{S}(i,i')$ measures the distance between i and i' only considering the semantic variables available in $\mathcal{S}$.

Thus

$$d^2_{(\alpha,\beta,\gamma)}(i,i') = \alpha \sum_{k\in\zeta} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \frac{\beta}{n^2_\mathcal{Q}} \sum_{k\in\mathcal{Q}} d^2_k(i,i') + \frac{\gamma}{n^2_\mathcal{S}} \sum_{\forall k\in\mathcal{S}} \delta^2_k(i,i') \tag{2}$$

where s_k^2 is the variance of numerical variable X_k; $n_\mathcal{Q} = card(\mathcal{Q})$ and $d^2_k(i,i')$ is the *contribution of categorical variable* X_k *to* $d^2_{(\alpha,\beta,\gamma)}(i,i')$ (see expression 3), $n_\mathcal{S} = card(\mathcal{S})$, $\delta^2_k(i,i')$ is the *contribution of semantic variable* X_k *to* $d^2_{(\alpha,\beta,\gamma)}(i,i')$.

In section 4.1 and 5, different possibilities for $\delta_k(i,i')$ will be provided, according to existent literature in computational semantics as well as our own contribution.

$$d^2_{\mathcal{Q}}(i,i')^2(i,i') = \begin{cases} 0, & \text{if } x_{ik} = x_{i'k} \\ \frac{1}{I_{k^i}} + \frac{1}{I_{k^{i'}}}, & \text{otherwise, for individuals} \\ \frac{(f_i^{k_s}-1)^2}{I^{k_s}} + \sum_{j\neq s}^{n_k} \frac{(f_i^{k_j})^2}{I^{k_j}}, & \text{if } x_{ik} = c_s^k, \text{and } i' \text{ is a class} \\ \sum_{j=1}^{n_k} \frac{(f_i^{k_j} - f_{i'}^{k_j})^2}{I^{k_j}}, & \text{in general case} \end{cases} \tag{3}$$

In expression (3), I^{k_j} represents the number of individuals of the sample that are in modality c_j^k; I_{k^i} is the number of individuals in the sample of the same modality as the element i for variable X_k; $f_i^{k_j} = \frac{I_i^{k_j}}{\sum_{j=1}^{n_k} I_i^{k_j}}$ represents the proportion of individuals from the i^{th} subclass satisfying $X_k = c_j^k$ and n_k is the number of modalities of variable X_k, which is qualitative. In (**?**) details on the mixed metrics are provided.

Under analog philosophy of Gibert's mixed metrics, and according to the principles of compatibility measures proposed by Anderberg, the contribution of a single variable to the final distance is different depending on its type and it can be computed per blocks, regarding the types of variable.

3.2. On the weighting indices α, β, γ

In fact, $d^2_{(\alpha,\beta,\gamma)}(i,i')$ is an infinite family of metrics where $(\alpha,\beta,\gamma) \in [0,1]^3$. In every particular application of this distance, concrete values of α,β,γ, must be chosen. With a similar argument as the one provided in (**?**), for hierarchical clustering purposes, and being α, β and γ the weight for numerical, categorical and semantic variables, it is enough to index the expression (1) with $(\alpha,\beta,\gamma) \in [0,1]^3$ and $\alpha+\beta+\gamma = 1$ since bounding the addition of weights defines an equivalence relationship over the set of achievable hierarchies and the same hierarchies available from $(\alpha,\beta,\gamma) \in \Re^3$ are found. This implies that at least two of the parameters must be determined and the third one is fixed consequently.

In (**?**) (**?**) some heuristic criteria are introduced to find acceptable values for the weighting constants α and β. Several real applications to complex domains (**?**) (**?**) (**?**) (**?**) showed a successful performance of the original proposal in front of other values, for the particular case of recognizing underlying classes on a given domain. Here, the criteria used in (**?**) for determining α,β are extended to semantic variables, including γ:

$$\alpha = \frac{a}{a+b+c} \quad \& \quad \beta = \frac{b}{a+b+c} \quad \& \quad \gamma = \frac{c}{a+b+c} \tag{4}$$

This guarantees that $(\alpha,\beta,\gamma) \in [0,1]^3$ and $\alpha+\beta+\gamma = 1$ as convenient for clustering purposes.

As the value of one of them approaches too much to 1, the distance will

behave giving maximum influence to the corresponding block of variables. For example, using $(\alpha, \beta, \gamma) = (0.9, 0.05, 0.05)$ means that the distances between objects will basically be determined by their similarities in numerical variables, and both qualitative and semantic variables will modify the final distance very few. Viceversa, using a value too close to zero implies to dismiss the information provided by the corresponding group of variables. Assuming that all the variables considered in the data matrix are equally relevant, it is reasonable to choose values for (α, β, γ) that balance the contribution of all types of variables in the final distance. According to this principle, the proposed values for *a, b, c* are the following:

$$a = \frac{n_\zeta}{d^2_{\zeta max^*}} \quad \& \quad b = \frac{n_\mathcal{Q}}{d^2_{\mathcal{Q} max^*}} \quad \& \quad c = \frac{n_\mathcal{S}}{d^2_{\mathcal{S} max^*}} \tag{5}$$

where $n_\zeta = card(\zeta)$, $n_\mathcal{Q} = card(\mathcal{Q})$ and $n_\mathcal{S} = card(\mathcal{S})$ and $d^2_{\zeta max^*}$ $d^2_{\mathcal{Q} max^*}$ and $d^2_{\mathcal{S} max^*}$ are the truncated maximums of the different sub-distances to provide robustness in front of multivariate outliers. In our proposal maximums are truncated to 95% but other possibilities could be considered as well.

This proposal has the following properties:

- The proposal gives to every sub-distance an importance proportional to the number of variables it represents. So,

  $$\alpha \propto n_\zeta \quad \& \quad \beta \propto n_\mathcal{Q} \quad \& \quad \gamma \propto n_\mathcal{S} \tag{6}$$

  This means that when the larger group of variables are qualitative, for example, the second term of the distance will have more contribution to the final distance, as there will be a bigger factor multiplying the single distance number, which in fact is representing a larger group of variables than the other terms.
- The proposal represents a balance among the different components of the final distance, since they are referred to a common interval. Dividing every term by the maximum value they can present, the three components will have equal influence on $d^2(i, i')$, since

  $$\alpha \propto \frac{1}{d^2_{\zeta max^*}} \quad \& \quad \beta \propto \frac{1}{d^2_{\mathcal{Q} max^*}} \quad \& \quad \gamma \propto \frac{1}{d^2_{\mathcal{S} max^*}} \tag{7}$$

- The proposal is robust to the presence of outliers, because it is considering truncated maximums. As outliers produce big distances with respect to the other objects they will not be taken as reference points for the quotient, and the other distances would not concentrate in a subinterval $[0,c_0]$, $c_0 \ll 1$, avoiding even numerical instability.
  Moreover, when outliers are not present, the truncated distances will be almost of the same range as $d^2_{\zeta max^*}$, $d^2_{\mathcal{Q} max^*}$, and $d^2_{\mathcal{S} max^*}$ respectively, and the real working interval $[0,c_0]$, $c_0 \approx 1$ will not imply a major change.

This is a proposal that use to provide clearly interpretable classes, but other possibilities can be considered as well. In (**?**) two other heuristic are proposed for α and β, one considering the inertia of the variables, the other considering the correlation between variables. In (**?**) the impact of the parameters for original Gibert's mixed metrics in final classes has been tested, including a comparison with the Ralambondrayni's proposals. On the one hand, it could be seen that small changes in the parameters do not provide big changes in the final classes,

since the different hierarchies are found when significant differences in the distances occur; and this requires big changes on the parameter values. On the other hand, the Gibert's proposal is the one that better recognized some data structures like filiform classes. Also, in (**?**) the use of this proposal in clustering has been compared with the use of other compatibility measures, like the Gower similarity coefficient (**?**) or the generalization of Minkowski metrics by Ichino-Yaguchi (**?**). It has been seen that the Gibert's mixed metrics using the present proposal for the parameters provided more balanced and robust classes, also easier to interpret. For this reason this proposal is the one used in this paper to compute the extended Gibert's metrics and the *Superconcept-based distance* introduced later on.

4. Comparisons using semantic variables

In the previous section, a mixed metrics that combines 3 distances has been proposed. In this section, the new component, the semantic distance, is addressed. As said before, the computation of the semantic similarity/distance between concepts has been a very active trend in computational linguistics. Proposals found in the literature can be divided according to the techniques employed and the knowledge exploited to perform the assessment:

- unsupervised approaches do not need external knowledge, because they compute the similarity from the information distribution of terms in a given corpus, using the degree of word co-occurrences (**?**); in particular, some approaches use the entire Web as a corpus (**?**).
- other approaches interpret terms by using structured representations of knowledge to compute the similarity, like pairwise constraints for object's relationships (**?**) or background classical Knowledge Bases (**?**) (**?**) (**?**) or ontologies (**?**) (**?**) (**?**).

It has been seen that the use of this additional background knowledge helps to improve the semantic coherence of the results of data mining (**?**) (**?**). Using ontologies instead of Knowledge Bases can be an advantage in some cases because recently, rich domain ontologies about very different fields are becoming available in Web repositories (i.e. Swoogle (**?**), OntoSelect (**?**)). Also, it may be easier to the expert to express his background knowledge in form of ontology than to explicitly formulate the relationships between variables by means of logic rules. Thus, this research is focused in those cases where an additional ontology is available, providing semantic information about some of the categorical variables of the data matrix.

Formally, an ontology O is composed by a set of concepts of the domain or classes C, which are *taxonomically* related by the transitive is-a relation $H^c \in C \times C$, called concept hierarchy or *taxonomy*. The ontology contains at least one taxonomical relationship and can also include other *non-taxonomic* relationships $R^* \in C \times C \times$ String. A subsumer or superconcept of a given concept c is another concept placed in a higher level of the hierarchy and connected with c by one or several is-a relationships.

4.1. Semantic similarity using ontologies

Taxonomical knowledge is the most common way of structuring domain knowledge and the minimum level of representation that can be expected from an ontology (**?**). A research of the structure of existing ontologies via the Swoogle ontology search engine (**?**) has shown that domain ontologies usually model only taxonomic relationships.

In the literature, a variety of methods can be found to exploit the subsumption hierarchy (or taxonomy) of concepts for terms comparisons. They can be divided in two main groups:

- methods exploiting the taxonomic structure of the ontology.
- methods that additionally introduce the *information content* (IC) by using the probability of appearance of the compared terms in a domain related corpus. They are based on the assumption that infrequent words are more informative than more frequent ones (**?**).

The measures from second group heavily depend on the availability of a reference corpus, as well as on the effective use of terms in real speeches to guarantee representativeness, which sometimes is difficult and works better in text mining applications and related fields. That is why the approach of first group, based on the exploitation of the ontology's geometrical model, is considered for Gibert's mixed metrics generalization. Some proposals from the literature are presented below. Most of them exploit the minimum path between a pair of concepts losing a lot of valuable information. In addition, they do not guarantee the metric properties, which are relevant for hierarchical clustering purposes. Under the aims of keeping metrical structure in the final proposal as well as taking into account as much taxonomical information available as possible, a new way to compute the distance between objects for semantic variables is proposed. The proposal is implemented in a system called *KLASS* (**?**) for application.

Path Length: In an is-a hierarchy, the simplest way to measure the distance between two concepts c_i and c_j is the shortest *path length* connecting these concepts (the minimum number of links) (**?**). This in fact fits on the classical definition of distance between nodes in a graph, from the mathematical point of view.

$$sim_{pL}(c_i, c_j) = minimum\ number\ of\ edges\ separating\ c_i\ and\ c_j \tag{8}$$

This measure, however, have sometimes difficult interpretation in the field of computational linguistics and several variations have been developed.

Wu and Palmer: (**?**). They propose a path length measure based on the depth of concepts in the hierarchy (9).

$$sim_{w\&p}(c_i, c_j) = \frac{2 * N_c}{N_i + N_j + 2 * N_c} \tag{9}$$

, where N_i and N_j are the number of is-a links from c_i and c_j to the LCS c respectively — being LCS the Least Common Subsumer or the first common ancestor in the ontology — and N_c is the number of is-a links from c to the root ρ of the ontology. Wu and Palmer similarity is analogous to the Dice coefficient considering the number of is-a links from c to ρ as what is common between c_i and c_j. It scores between 0 to 1(for maximally similar concepts).

Leacock and Chodorow: (**?**) proposed a measure that depends on the shortest path between two concepts (in fact, the number of nodes N_p from c_i to c_j) and the depth D of the taxonomy in which they occur (10).

$$sim_{l\&c}(c_i, c_j) = -\log N_p/2D \tag{10}$$

Concept Match: In previous measures, if the pair of concepts inherits from many is-a hierarchies, there exist many paths between a pair of concepts, but only the shortest one is considered. In that sense, another interpretation of these measures is possible, considering that the similarity is assessed from the minimum number of shared superclasses of the pair of concepts under comparison. Having this into account, Maedche and Zacharias (**?**) defined the Concept Match (CM) measure (11) based on the definition of the Upward Cotopy (UC) of a concept c_i, restricted to the set of upper concepts of c_i in a is-a hierarchy H^C and itself, denoted as $UC(c_i, H^C)$.

Concept Match considers a proportion between the number of common UC from the total of UC of both concepts.

$$sim_{CM}(c_i, c_j) = \frac{|UC(c_i, H^C) \cap UC(c_j, H^C)|}{|UC(c_i, H^C) \cup UC(c_j, H^C)|} \tag{11}$$

5. Superconcept-based Distance

Path length-based measures only consider the minimum path between a pair of concepts, omitting the rest of the taxonomical knowledge available in the ontology. For complex taxonomies with thousands of interrelated concepts by means of multiple hierarchies, this kind of measures waste a great amount of relevant information contained in the ontologies. Indeed, two concepts should be considered more similar as the number of hierarchies interrelating them increases. Thus, it seems reasonable that a measure taking into account the whole taxonomical hierarchy involving the evaluated concepts should provide more accurate similarity assessments. A proposal on this line is presented based on the well-known Euclidean distance.

Let us consider the set of superconcepts (ancestors in the taxonomy) of a concept c_i in a given Hierarchy H^C as all the concepts preceding c_i in some of the taxonomies of H^C, including c_i itself:

$$\mathcal{A}(c_i) = \{c_j \in C | c_j = c_i \vee c_j \text{ is ancestor or superconcept of } c_i \ \in H^C\}$$

From an algebraic point of view, $\mathcal{A}(c_i)$ can be represented by a binary vector $x_i = (x_{i1} \dots x_{in})$, being n the number of concepts of the ontology, and

$$x_{ik} = \begin{cases} 0, \text{if } c_k \notin \mathcal{A}(c_i) \\ \\ 1, \text{if } c_k \in \mathcal{A}(c_i) \end{cases}$$

Having a vectorial representation of the concepts, the distance between two concepts c_i, c_j can be defined as the Euclidean distance between the associated vectors x_i, x_j:

$$d_E(c_i, c_j) = d(x_i, x_j) = \sqrt{\sum_{i=k}^{n} (x_{ik} - x_{jk})^2}$$

In this case, this measure has a very clear interpretation. As the values in the vectors can only be 0 or 1, the difference $(x_{ik} - x_{jk})$ can only be equal to 1 if and only if c_k is a superconcept of c_i and it is not a superconcept of c_j (or viceversa). Therefore, $\sum_{k=1:n}(x_{ik} - x_{jk})^2$ is, in fact, equal to the number of non-shared superconcepts between c_i and c_j.

Based on this interpretation, the distance can be rewritten in terms of the set of superconcepts of c_i, $\mathcal{A}$, thus providing a more compact expression (12), which is more efficient for evaluation in the scope of the treated ontologies with thousands of concepts, and which do not require the explicit construction of the binary matrix associated to the ontology, too big and hard to manage in big ontologies.

$$d_E(c_i, c_j) = \sqrt{card\{\mathcal{A}(c_i) \cup \mathcal{A}(c_j)\} - card\{\mathcal{A}(c_i) \cap \mathcal{A}(c_j)\}} \tag{12}$$

Note that the distance d_E only considers the non-common information of two concepts but does not evaluate the amount of common information. So, it is not capable to distinguish between cases in which the number of common superconcepts is small (corresponding to general terms from upper levels of the ontology) from those cases in which the number of common superconcepts is high (corresponding to more specific terms at lower levels).

In order to take into account the number of common superconcepts, d_E is normalized by the total number of superconcepts of c_i and c_j. The sum of common and non-common superconcepts is $card\{\mathcal{A}(c_i) \cup \mathcal{A}(c_j)\}$. Consequently, the *Superconcept-based Distance* is defined as:

Definition: Superconcept-based Distance (SCD)

$$d_{\mathcal{SCD}}(c_i, c_j) = \sqrt{\frac{card\{\mathcal{A}(c_i) \cup \mathcal{A}(c_j)\} - card\{\mathcal{A}(c_i) \cap \mathcal{A}(c_j)\}}{card\{\mathcal{A}(c_i) \cup \mathcal{A}(c_j)\}}} \tag{13}$$

This definition introduce a desired penalization to those cases in which the number of shared superconcepts is too small. So, we are able to compare concepts on the basis of the ratio between the non-overlapping taxonomical knowledge versus the total number of ancestors.

In section 7, the results obtained with d_E distance and $d_{\mathcal{SCD}}$ are compared, showing that considering both the amount of common and non-common information between a pair of concepts givea a more accurate estimation of their semantic similarity. Our proposal is able to overpass the performance of other classical measures evaluated against a standard benchmark.

6. On metrical properties of measures for semantic variables

In section 4.1 different approaches to measure the semantic similarity have been presented. In order to introduce those measures in a global compatibility measure also considering numerical and categorical distances, first of all it is required to transform them from a similarity into a dissimilarity by means of $d = max_{sim} - sim$, where max_{sim} is the maximal value reached by sim, or $d = 1/sim$ when max_{sim} is not finite (**?**).

In general, all those measures will simply be dissimilarities, and not distances, so very often they will violate triangular inequality, due to the natural implicit uncertainty of linguistic data, which is related with not accurate expressions in natural language.

In the field of computational linguistics this is not a handicap in general (**?**), (**?**). So, metrical structure is not necessary *per se* for general purposes, and it can make sense to consider dissimilarity coefficients for either $d^2_\zeta(i,i')$, $d^2_\mathcal{Q}(i,i')$ and $d^2_\mathcal{S}(i,i')$ in a general context.

However, for the particular use of $d^2_{(\alpha,\beta,\gamma)}(i,i')$ in hierarchical clustering with Ward's method, it seems convenient to keep both metrical structure as well as quadratic form. Keeping the metrics structure guarantees the ultrametric properties of the resulting dendrogram. Keeping the distance as a combination of quadratic forms, the Huygens decomposition holds (**?**), what permits to use Ward's criterion, and it is directly related with interpretability of final results. If $d_{(\alpha,\beta,\gamma)}(i,i')$ is a metric, $d^2_{(\alpha,\beta,\gamma)}(i,i')$ is both a metric and quadratic form and Ward's method can be used with it.

As $d^2_\zeta(i,i')$, $d^2_\mathcal{Q}(i,i')$ and $d^2_\mathcal{S}(i,i')$ hold metrical properties over the spaces $\{X_k : k \in \zeta\}$, $\{X_k : k \in \mathcal{Q}\}$ and $\{X_k : k \in \mathcal{S}\}$ respectively, then $d^2_{(\alpha,\beta,\gamma)}(i,i')$ is a linear combination of metrical measures. In (**?**) it was already proved the metrical structure of the original Gibert's mixed metrics and $d_\zeta(i,i')$, $d_\mathcal{Q}(i,i')$ provided that $\alpha = 0 \Longrightarrow \zeta = \emptyset$ & $\beta = 0 \Longrightarrow \mathcal{Q} = \emptyset$. A natural extension of that condition to semantic variables guarantees the metric structure of the extended Gibert's mixed metrics, provided that $d^2_\mathcal{S}(i,i')$ is also a metrics:

$$\alpha = 0 \Longrightarrow \zeta = \emptyset \;\&\; \beta = 0 \Longrightarrow \mathcal{Q} = \emptyset \;\&\; \gamma = 0 \Longrightarrow \mathcal{S} = \emptyset \tag{14}$$

This is not a very restrictive constraint. When one of these weights is 0 the associated *subdistance* is not taken into account, and $d^2_{(\alpha,\beta,\gamma)}(i,i')$ then fails the identity property, unless the corresponding type of variables is also eliminated of the data matrix. The case (α,β,γ) = (0,0,0) is excluded because $d^2_{(0,0,0)}(i,i')$ is the constant function 0 which is nonsense in this context.

So, the metric structure of $d_\mathcal{S}$ guarantees metrical structure of $d^2_{(\alpha,\beta,\gamma)}(i,i')$ and good performance of clustering. Most of the proposals in the literature are only similarity coefficients. Path Length (**?**) also holds the triangular inequality and could be used to extend Gibert's mixed metrics for clustering.

Our proposal based on an Euclidean distance between binary vectors representing the hierarchical set of ancestors of the compared concepts (Eq. 12), keeps trivially all the properties of metrical structures, although it is expressed by means of an equivalent expression using classical set theory (**?**).

In this sense, we presume that the use of the SCD distance in a clustering context will perform better compared with other proposals because it considers more information of the ontology and it keeps metrical structure. This is analyzed in the next section.

7. On semantic comparisons' performance: Benchmark results

In this section, the behavior of the semantic measures presented above in front of a common data set is analyzed. There are some benchmarks available in the literature especially designed to evaluate the performance of semantic similarity coefficients, and used as standard in the field of computational linguistics. The most common way of evaluating similarity measures in semantic fields is by using a set of word pairs whose similarity has been assessed by a group of human

experts and computing their correlation with the results of the computerized measures.

Two cases are presented, one for general purpose terms using Wordnet as reference ontology, another a test regarding a specialized field, biomedicine, with a specialized ontology SNOMED-CT.

In both cases, the conclusion is that SCD measure is able to extract a robust semantic evidence from both general purpose ontologies like WordNet, or highly complex ontologies in specialized fields like biomedicine; the shared and non-shared superconcepts of the compared terms provide a more accurate estimation of the semantic distance than simple path lengths on the reference ontologies.

7.1. General purpose benchmark

Rubenstein and Goodenough (**?**) defined the first experiment in 1965 in which a group of 51 students, all native English speakers, assessed the similarity of 65 word pairs selected from ordinary English nouns, on a scale from 0 (semantically unrelated) to 4 (highly synonymous). Miller and Charles (**?**) re-created the experiment in 1991 by taking a subset of 30 noun pairs whose similarity was reassessed by 38 undergraduate students (Table 1). The correlation obtained with respect to Rubenstein and Goodenough's experiment was 0.97. Resnik (**?**) replicated again the same experiment in 1995, in this case, using 28 noun pairs and requesting two groups of human experts: 10 computer science graduate students and post-doc researchers. The correlation with respect to Miller and Charles results was 0.96. The average correlation over the graduate students and post-doc researchers was 0.884. This value is considered the upper bound to what one could expect from a machine computation on the same task (**?**). Thus, the performance of a semantic similarity measure is evaluated by means of the correlation with respect to the human judgments.

In this work, the ontology used is Wordnet (3.0), which contains words (nouns, verbs, adjectives and adverbs) of the English language. Dissimilarities were transformed into similarities by means of $sim = max_d - d$ (max_d being the maximal value of d), or $sim = 1/d$ when max_d is not finite (**?**).

Table 1 displays the similarity assigned to every pair of words by the measures presented in section 4.1. The first column shows the human ratings with a baseline correlation of 0,884, then Path Length, Wu and Palmer (WP), Leacock and Chodorow (LC), Concept Match (CM) similarities and d_E (Eq. 12) and the Superconcept-based Distance (Eq. 13), conveniently transformed to similarities for comparison with Resnik reference results.

It can be seen that Leacock and Chodorow and Wu and Palmer measures clearly outperform Path Length measure. The reason is that these measures explicitly take into account more information of the taxonomy than the depth of the ontology, considering the depth of the LCS (Leacock and Chodorow's proposal) and the relative depth between compared concepts and the LCS (Wu and Palmer). Wu and Palmer shows a correlation of 0.804 and Leacock and Chodorow of 0.829, very close to the upper bound (0.884) to what one could expect from a machine computation. Thus, they can be considered as an effective unsupervised way to assess concept's similarity.

A second group of measures consider more complete information from the ontology: the whole set of ancestors of the compared concepts. This is the case of Concept Match, SCD and its predecessor d_E. It can be seen that taking into

account the whole subsumer's hierarchy, also outperform the Path Length measure. The Concept Match measure give a correlation of 0.802, very close to the Wu and Palmer results. Notice that d_E achieves a correlation of 0.814. Moreover, the SCD (a normalized d_E) provides some improvement to the similarity assessment (from 0.814 to 0.839). Although the difference is not very large, the *Superconcept-based distance*, proposed in this paper, provides the highest correlation regarding expert's judgments (0.839). The results show that SCD performs slightly better also in those data sets. To confirm these results, a second analysis in a specialized domain has been performed.

7.2. Biomedical benchmark

For the biomedical domain, Pedersen et al. (**?**), in collaboration with Mayo Clinic experts, created a set of 30 word pairs referring to medical disorders. Their similarity was assessed in a scale from 1 to 4 by a set of 9 medical coders who knew the notion of semantic similarity and a group of 3 physicians who were experts in the area of rheumatology. For each pair of terms, the averaged scores for each group of experts is presented in Table 2. The correlation between physician judgements was 0.68 and between the medical coders was 0.78.

We used these data to evaluate the semantic measures presented in this paper, using SNOMED-CT as the domain ontology. The term pair *"chronic obstructive pulmonary disease" - "lung infiltrates"* was excluded from the test, as the later term was not found in the SNOMED-CT terminology.

As some of the measures involved in the test compute similarity (Wu and Palmer, Leacock and Chodorow and Concept-match) and others evaluate dissimilarity (Path Length and Superconcept-based distance), for a consistent comparison, all the results have been converted into similarity values. So, $sim(c_i) = max_d - d(c_i)$, where max_d is the maximal value that can be obtained by the measure d (**?**). In this case, max_d corresponds to 2*maximum depth of any taxonomical branch in SNOMED-CT. Note that this conversion does not affect the result of the evaluation, since a linear transformation of the values will not change the magnitude of the resulting correlation coefficient.

The correlations between the results of the different compared measures with respect to the human expert scores (including physicians, coders and the averaged scores of both) are presented in Table 3.

The correlation between human experts (0.68 for physicians and 0.78 for coders) represent an upper bound for a computerized approach. Taking this into account, it can be seen that Path Length-based measures offer a limited performance with correlations smaller than 0.45 and 0.59 respectively. Poor results are obtained when estimating semantic similarity from the minimum inter-concept path in complex domain ontologies, such as SNOMED-CT, where multiple paths between concepts from several overlapping taxonomies are available.

On the other hand, similarities computed with measures using much more ontological knowledge (the whole subsumer's hierarchy) correlate much better than Path Length-based ones. The improvement is of almost a 20% (0.33 of Path Length vs 0.56 of Concept Match) with respect to the upper bound. Furthermore, the SCD measure has the best performance compared against the others and it is quite close to the correlation between human manual evaluation: 0.589 vs 0.68 in the case of physicians and 0.744 vs 0.78 with respect to medical coders.

Even using a wide ontology like SNOMED-CT, classical approaches based on

Table 1. Comparative using Resnik experiment.

		Comparative					
Word pair	**Human ratings**	**Path Length**	**WP**	**LC**	**CM**	d_E	**SCD**
automobile car	3.9	32	1.0	5.0	1.0	5.657	1.0
jewel gem	3.5	32	1.0	5.0	1.0	5.657	1.0
voyage journey	3.5	31	0.947	4.0	0.910	4.657	0.698
lad boy	3.5	31	0.910	4.0	0.910	4.657	0.698
shore coast	3.5	31	0.889	4.0	0.833	4.657	0.592
madhouse asylum	3.6	31	0.947	4.0	0.910	4.657	0.698
wizard magician	3.5	32	1.0	5.0	1.0	5.657	1.0
noon midday	3.6	32	1.0	5.0	1.0	5.657	1.0
stove furnace	2.6	23	0.471	1.678	0.357	2.657	0.198
fruit food	2.1	23	0.308	1.678	0.25	2.657	0.134
cock bird	2.2	31	0.947	4.0	0.910	4.657	0.698
crane bird	2.1	29	0.857	3.0	0.769	3.924	0.520
implement tool	3.4	31	0.714	4.0	0.875	3.657	0.646
monk brother	2.4	31	0.6	4.0	0.917	4.657	0.711
implement crane	0.3	28	0	2.678	0.6	3.657	0.367
brother lad	1.2	28	0.461	2.678	0.667	3.657	0.423
car journey	0.7	15	0.118	0.83	0.053	1.414	0.030
oracle monk	0.8	25	0.461	2.0	0.533	3.011	0.317
rooster food	1.1	17	0.118	1.1	0.118	1.784	0.061
hill coast	0.7	28	0.6	2.678	0.5	3.657	0.293
graveyard forest	0.6	24	0.444	1.83	0.2	2.193	0.105
slave monk	0.7	28	0.6	2.678	0.667	3.657	0.423
forest coast	0.6	27	0.444	2.415	0.231	2.657	0.134
wizard lad	0.7	28	0.6	2.678	0.667	3.657	0.423
smile chord	0.1	22	0.286	1.54	0.231	2.340	0.123
magician glass	0.1	25	0.4	2.0	0.286	2.494	0.156
string noon	0.0	21	0.154	1.415	0.154	2.340	0.080
voyage rooster	0.0	9	0.0	0.415	0.042	0.861	0.021
Correlation	0.884	0.670	0.804	0.829	0.802	0.814	0.839

Table 2. Set of 30 medical term pairs with associated averaged expert's similarity scores (extracted from Pedersen et al.)

Term 1	Term 2	Phys.	Coder
Renal failure	Kidney failure	4.0	4.0
Heart	Myocardium	3.3	3.0
Stroke	Infarct	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Delusion	Schizophrenia	3.0	2.2
Congestive heart failure	Pulmonary edema	3.0	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2.0
Diarrhea	Stomach cramps	2.3	1.3
Mitral stenosis	Atrial fibrillation	2.3	1.3
Chronic obstructive pulmonary disease	Lung infiltrates	2.3	1.9
Rheumatoid arthritis	Lupus	2.0	1.1
Brain tumor	Intracranial hemorrhage	2.0	1.3
Carpal tunnel syndrome	Osteoarthritis	2.0	1.1
Diabetes mellitus	Hypertension	2.0	1.0
Acne	Syringe	2.0	1.0
Antibiotic	Allergy	1.7	1.2
Cortisone	Total knee replacement	1.7	1.0
Pulmonary embolus	Myocardial infarction	1.7	1.2
Pulmonary fibrosis	Lung cancer	1.7	1.4
Cholangiocarcinoma	Colonoscopy	1.3	1.0
Lymphoid hyperplasia	Laryngeal cancer	1.3	1.0
Multiple sclerosis	Psychosis	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0
Rectal polyp	Aorta	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0
Depression	Cellulitis	1.0	1.0
Varicose vein	Entire knee meniscus	1.0	1.0
Hyperlipidemia	Metastasis	1.0	1.0

Table 3. Correlations obtained for each measure against Physicians, Coders and both

Measure	Physician	Coder	Both
Path Length	0.33	0.395	0.386
Wu and Palmer	0.293	0.364	0.353
Leacock and Chodorow	0.453	0.585	0.548
CM	0.56	0.685	0.656
SCD	0.589	0.744	0.7

Path Length have shown a poor performance. Due to the inherent complexity of taxonomical links modeled in that ontology, with relationships of multiple inheritance between concepts, the computation of the minimum path between a pair of concepts only represents a partial view of the modeled knowledge. Taking into account the ration between shared and non shares superconcepts as well as the multiple inheritance, as SCD does, helps to better evaluate similarities between semantic terms, and the results provided by SCD are those closer to human judgements. Provided that SCD ($d_{\mathcal{S}}$) seems to adequately evaluate word's similarity, we propose to use it in clustering processes, for its previously mentioned metrical properties.

8. On the impact of including semantic variables in clustering

In this section, our proposal is used for clustering in two real applications in order to show the significant improvements achieved when semantic variables and reference ontologies are included in the data analysis.

The two case studies refer to the Tourist field. In fact, the recreational and tourist activities have a growing importance in relation to economic development (**?**). For that reason getting any kind of knowledge about the characteristics of the visitors of different tourist destinations is of great importance for planning, improving facilities and increasing the economic potential of an area.

In the particular field of Tourism, which will be studied in this section, the World Tourism Organization (UNWTO) has developed the *Thesaurus on Tourism and Leisure Activities*; however that is not available in the ontological languages, so it is not machine readable. Up to now, the ontology-based systems developed in this field rely on specific-purpose ontologies, designed and built ad-hoc for each particular system. For instance in (**?**) an ontology is defined for covering concepts about what activities one can do, when and where are they developed. In (**?**) the mobility of tourists in a recreational area is studied with the support of a Tourist Mobility Behavior ontology. Other ontologies include a larger taxonomy of types of activities (**?**) (**?**) (**?**), including information about opening times or admission fees such as in (**?**). Modular ontologies facilitate the integration of the portions of different ontologies that are relevant for some specific application (**?**) (**?**).

In other cases, the ontology is tailored to the characteristics of a particular territory, such as the Jeju travel ontology focused on the Jeju volcanic island in Korea (**?**), or the ontology developed for the typical activities in the Catalan Mediterranean area (**?**).

However, none of those ontologies include concepts related to the motivations of the visitors, which is the content of the second case study. Only the e-Tourism ontology defined in (**?**) includes a small set of terms related to this issue. Unfortunately, they do not cover the large diversity of terms we have in the variables of our target dataset. That's the reason why we decided to work with a general-purpose ontology: WordNet. The fact of using an standard ontology guarantees that personal biasses are not propagated to the results, allowing a more neutral analysis of the improvement of introducing the semantic component.

The methods proposed in the previous sections have been implemented and integrated in the software KLASS (**?**). The hierarchical clustering is performed using the SCD distance with the values for α, β and γ proposed in section 3.2.

8.1. Tourist destinations case

The first case study refers to the identification of typical touristy city destinations. A data matrix with 23 cities from all over the world was considered. Each city is represented by a vector of 9 variables extracted from Wikipedia: *i)* population (numerical); *ii)* land area (numerical); *iii)* continent (categorical); *iv)* city ranking, categorical (country capital, state capital, city or village); *v)* country (France, Italy, Usa, Canada, Venezuela, Cuba, Spain, France, Andorra, Switzerland, Portugal, Australia); *vi)* language (French, Italian, English, Spanish, Catalan, German, Portuguese); *vii)* geographical situation (valley, plain, is-

land, coast, island, mountain range, mountain, lake, archipelago); *viii)* major city interest (cathedral, basilica, business, shopping center, government structure, office building, basilica, monument, historical site, church, mosque, recreational structure, ski resort, tourism, viewpoint, theater); *ix)* and major geographical interest (river, coast, bay, lake, mountain, beach, volcano, cliff, crater, ocean).

A hierarchical clustering based on the Ward criterion and the generalized Gibert's mixed metrics using SCD has been used in this study. Hierarchical clustering is appropriated, in this case, because the number of classes can be decided a posteriori. The cities have been clustered under two different approaches:

1. ignoring the semantic contribution of semantic variables and treating them as simple categorical variables, and
2. using a reference ontology for better treatment of semantic variables.

8.1.1. Clustering without semantic information

In this case, semantic variables are treated as categorical, that is, their semantics is not considered and the original Gibert's mixed metrics (which uses Chi-squared distance for categorical variables) is applied. Fig. 1 (left) shows the dendrogramm resulting from clustering (a binary hierarchical tree showing the sequence of aggregations performed by the algorithm; objects are placed in the bottom of the tree, classes are the internal nodes of the tree and increasing height of classes is related with the decreasing internal homogeneity). Apart from a trivial cut in two classes, which is not informative enough, the dendrogram seems to recommend a cut in 8 classes, which results in tree singletons (Interlaken, Montreal and Sydney), 3 classes of two cities C10={Havana, Caracas}, C14={PontaDelgada, Funchal}, C7={LosAngeles, NewYork} and the rest of cities divided in two bigger groups of 7 cities, one of them (C13) containing all the Spanish cities considered in the study. Class Panel graph was used for interpretation (**?**).

Fig. 2 (up) shows the Class Panel Graph for this partition (conditional distributions of variables versus the classes are placed side-by-side in a panel providing global perspective of class specificities). With this information, the following descriptions of the clusters are inferred:

- Interlaken is the only city near a lake with a sky resort and German-speaking.
- Montreal is the state capital of Quebec in Canada (North America), it is placed in an island and is interesting for its relative proximity to big lakes. The speaking language is French. In addition, it concentrates much office buildings, according to be the second largest city in Canada.
- Sidney is the largest city in Australia with more than 4 millions population. It is the state capital of New South Wales. It is situated near the coast and it is English-speaking. It has 5 theaters and the Sydney's Opera House.
- Class14 is composed by state (autonomous region) capitals from Portugal, they are located in islands or archipielagos. The spoken language is Portuguese. Their main interests are the historical site and craters in Ponta Delgada, and the viewpoints and cliffs in Funchal.
- Class10 is composed by country capitals of South America Spanish speaking.
- Class7 is composed by state capitals in USA. They are located either in islands or near the coast. However, one of their interest are their bays. New York City is the leading center of banking, finance and communication in USA, and Los Angeles, in addition, have some well-known shopping areas.
- Class13 is composed by 7 Spanish cities of different sizes. The spoken language is Catalan or Spanish. They have a wide diversity of interests.
- Class12 is the most heterogeneous one. It contains 7 either country capitals or villages from different countries and continents, with a wide diversity of cultural or geographical interests.

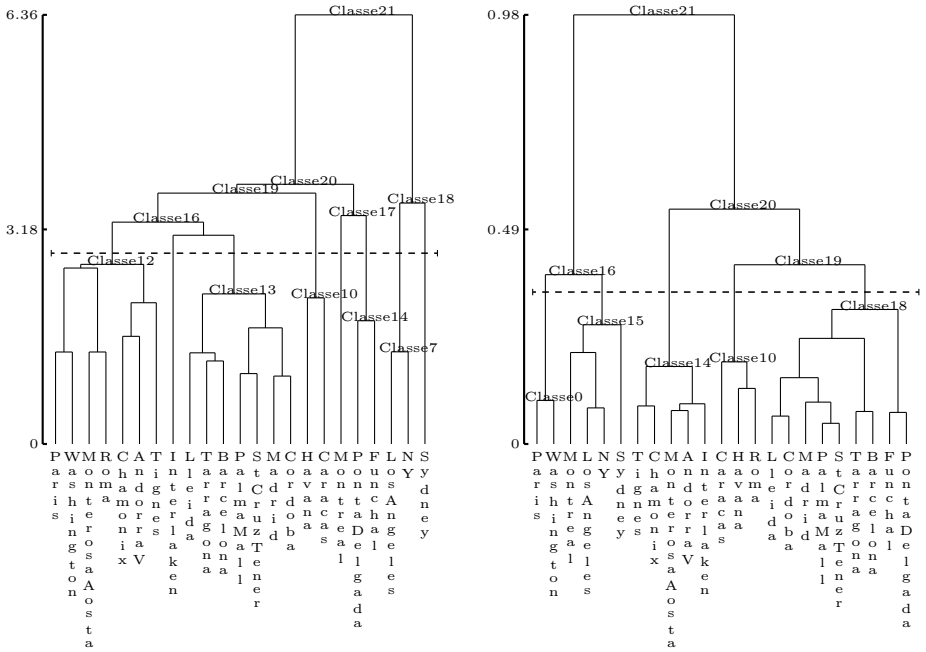

Fig. 1. *left)*: Dendrogram without ontologies; *right)*: Dendrogram using ontologies.

It seems that country and language directed the grouping and monuments, geography or situation have not influenced very much the partition. Consequently, the final grouping is not taking into account that cities in the coast might have more in common that those for skying, for example.

For better comparison with the results obtained when considering the ontological information, a cut in 4 classes has been also analyzed (see Fig. 1). In this case, classes contain cities very heterogeneous among them. As usual in real complex domains, there is a very big class of 15 cities quite heterogeneous which seems to share all type of cities. After that, classes of two or three cities appear and it is difficult to understand the underlying criteria for such a division (for example, Montreal is added to the class of Ponta Delgada and Funchal, which seems to make no sense at all).

8.1.2. Clustering with semantic information

In this case, 4 variables were treated as semantic using the WordNet ontology in the similarity assessment: country, language, geographical situation and major interest. Continent and city ranking are treated as categorical. Hierarchical clustering with Ward's criterion and the generalized Gibert's mixed metrics using SCD was used, since it is the one that showed better performance in the experiments presented in previous section. Fig. 1 (right) shows the resulting dendrogram, quite different from Fig. 1 (left) and producing groups more balanced in size.

The structure of the tree was studied and the Calinski-Harabatz index optimized to find the most suitable number of classes. A 5-classes cut is selected. In this case, the interpretation of clusters, made from the class panel graph (see figure 2), looks more coherent:

- Class10 has country capitals from Latin cultures (Cuba, Venezuela, Italy) speaking Romance languages with religious architecture as main interest.
- Class0 contains country capitals from Atlantic cultures (France and USA) located in valleys near a river.
- Class15 corresponds to big cities. All of them are state capitals of North America or Australia, located in islands or near the coast. The main interests are business or shopping (Theatre for Sydney), and the spoken language is English (French in Montreal) such as New York or Los Angeles.
- Class14 contains European small cities, all of them located near big mountains. The main interests are ski and recreational infrastructures.
- Class18 contains Iberian cities (Spain and Portugal). Most of them small cities in the coast or islands, which can have volcanoes or craters (Funchal and Ponta Delgada), except Madrid and Cordoba, in plain, and Lleida in valley. Their main interests are religious monuments or other historical sites. All cities speaks romance language and many are placed near the sea.

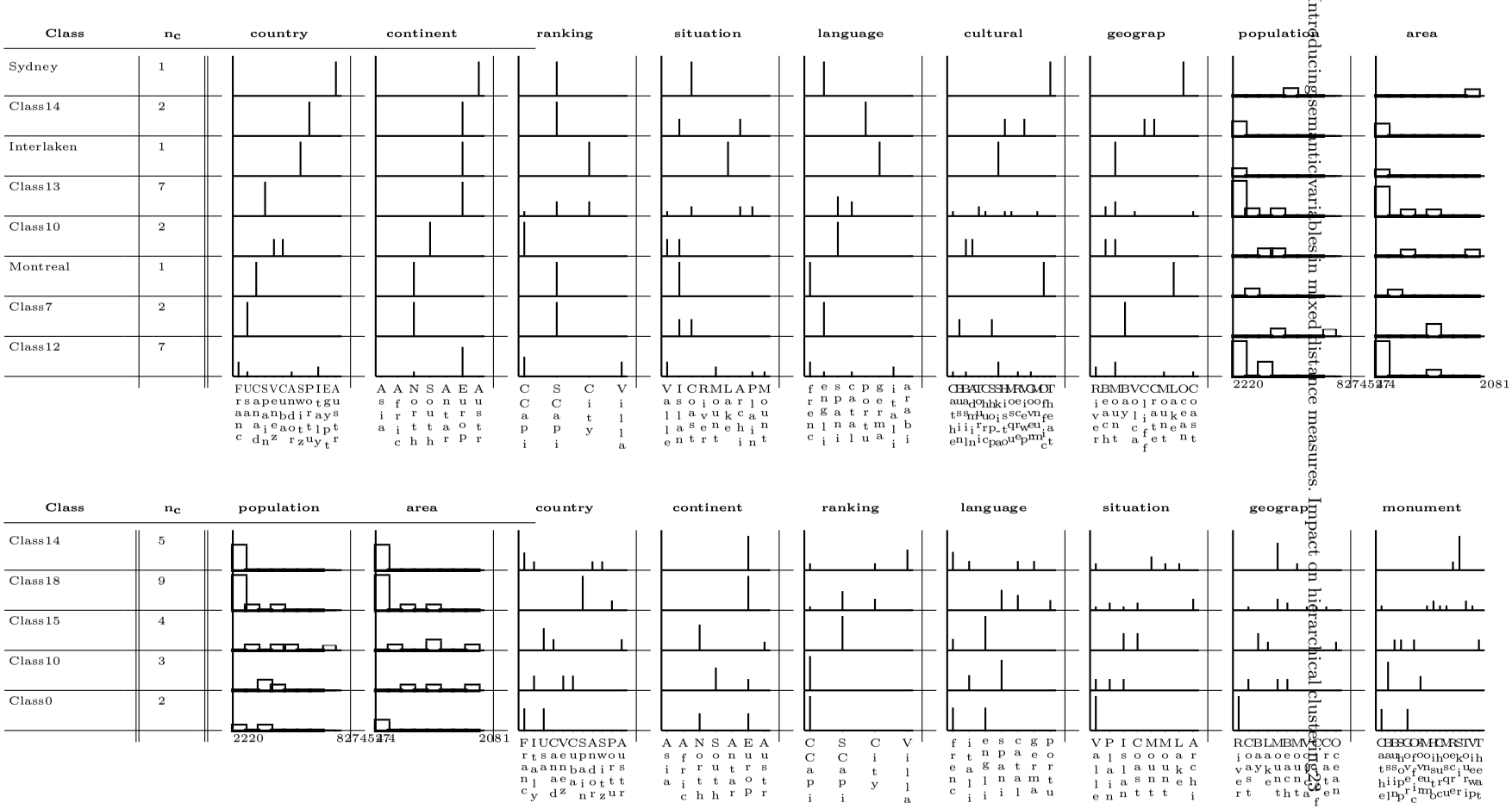

Fig. 2. up: Class panel graph of 8-class partition without considering ontologies.; down Class panel graph of 5-class partition considering ontologies.

Table 4. Cross experiments.

Categ \ Onto	C0	C10	C15	C18	C14	Total
C12	2	1	0	0	4	7
C7	0	0	2	0	0	2
Montreal	0	0	1	0	0	1
C10	0	2	0	0	0	2
C13	0	0	0	7	0	7
Interlaken	0	0	0	0	1	1
C14	0	0	0	2	0	2
Sydney	0	0	1	0	0	1
useful	2	3	4	9	5	23

Fig. 3. Geographical distribution of clusters using ontologies. Class0 is represented by a square, Class10 is represented by a cross, Class18 by a circle, Class15 by a star, and Class14 by a triangle.

Here, the meaning of the classes is clearer and more compact, and the underlying clustering criteria is a combination of several factors, as location, geography and main interests, which responds better to a multivariate treatment of the cities. Table 4 crosses the results of both partitions considering or not semantic information. This table shows that some countries moved to a more appropriate cluster when considering semantics, like Washington, which moved from Class16 (European cities) to a cluster with country and state capitals most of them in North America. Figure 3 shows the geographical distribution of clusters.

8.2. The Delta Ebre Natural Park case

In this section a second real application with a bigger data set is presented.

Natural parks have increased their importance as a tourism destination in the recent decades. In 2004, the Observatori de la Fundació d'Estudis Turístics Costa Daurada conducted a study of the visitors of the Ebre Delta Natural Park (Spain), with the funding of the Spanish Research Agency. The Ebre Delta is one of the largest wetlands areas in the Western Mediterranean; it is considered a Bird Special Protection Area and receives many tourists each year (about 300.000).

8.2.1. The Dataset of Park Visitors

The data was obtained with a questionnaire made to 975 visitors to Ebre Delta Natural Park between July and September 2004. A questionnaire was designed in order to determine the main characteristics of the tourism demand and the recreational uses of this natural area. It consisted of 17 closed-ended nominal questions, 5 numerical questions and 2 questions that evaluate the satisfaction of the visitor with a fixed numerical preference scale (Likert-type). The questions

Table 5. Frequencies of the reported values of features first and second reason.

	1rst rsn		2nd rsn			1rst rsn		2nd rsn	
Linguistic Value	Freq	%	Freq	%	Linguistic Value	Freq	%	Freq	%
Nature	339	17,4	211	11,4	Loyalty	8	0,4	12	0,6
Relaxation	146	7,5	222	11,4	Business	6	0,3	1	0,1
Beach	125	6,4	45	2,3	Education	5	0,3	3	0,2
Wildlife	61	3,1	88	4,5	Familiar tourism	5	0,3	5	0,3
Landscape	49	2,5	31	1,6	Walking	5	0,3	3	0,2
Culture	46	2,4	39	2,0	By chance	4	0,2	1	0,1
Second residence	45	2,3	9	0,5	Fishing	3	0,2	2	0,1
Visit	40	2,1	20	1,0	Photography	2	0,1	1	0,1
Sightseeing	20	1,0	6	0,3	Recommendation	2	0,1	3	0,2
Holidays	19	1,0	6	0,3	Before disappearance	1	0,1	2	0,1
Sports	13	0,7	19	1,0	Bicycling	1	0,1	2	0,1
Tranquillity	10	0,5	13	0,7	Clime	1	0,1	2	0,1
Others	10	0,5	6	0,3	Ecotourism - Birds	0	0,0	2	0,1
Gastronomy	9	0,5	3	0,2	Missing value	0	0,0	218	11,2

are about demographic and socio-economical aspects of the visitor (f.i. origin, age, sex or level of studies), aspects of the trip organization (f.i. previous information, material), characteristics of the visit (f.i. means of transport or activities done in the park) and, finally, the interests and satisfaction degrees on different features of the park. From this set of variables, two groups of interest have been defined (**?**): 4 variables that define the tourist profile (origin, age group, accompanying persons and social class) and 6 that model the trip profile (previous planning, reasons for trip, accommodation, length of stay and loyalty). We performed a proper descriptive analysis for data cleaning. Table 5 shows the values frequencies of features reporting the first and second reasons to come to Ebre Delta.

In (**?**), techniques of dimensionality reduction were used to find visitor's profiles, in order to improve the management of the area according to a better knowledge of the kind of people that visits the park and their main interests. In particular, a multivariate homogeneity analysis was carried out. Two dimensions were selected for the analysis, keeping a 30% and 26% of variance respectively. In the interpretation phase, it was seen that Dimension 1 can discriminate among the variables relating to type of accommodation, length of stay and reason for the trip. It shows the degree of involvement of the tourist with the nature. The second dimension is determined by the type of group and by age and shows the degree of involvement with the services, such as accommodation. It is important to note that the reasons for visiting the park play a role in both dimensions, being, at the end, the major factor used to distinguish the two main big groups of tourists. From that, five clusters (Table 6) of visitors were identified, from which the two first groups include a total of 83.9 % of the individuals. In (**?**)

Table 6. Typology of visitors to Ebre Delta Natural Park presented in (Clave et al 2007)

Class	%	Description
Ecotourism	44,6	Main interests: nature, observation of wildlife, culture and sports. Stay mainly in rural establishments and campgrounds. They are youths (25-24) coming from Catalonia and the Basque Country. First time.
Beach Tourism	39,3	Main interests: beach, relaxation, walk, family tourism. Family tourism, staying in rental apartments or second home. They come from Spain and overseas. Middle-class people with ages between 35-64. More loyalty (long and frequent visits).
Residents	11,0	Visitors from Aragon and Tarragona. Some of them have a second home, or friends and family living there. Nature is just an added value.
Youths	3,6	Mainly from Valencia, with ages between 15 and 24. They come with friends and quite frequently.
Educational Professional	1,5	Professional and educational interests. Mainly school groups.

it was concluded that the rest of groups were really small and targeted to a very reduced group of visitors. For this reason, only the two main groups, corresponding to EcoTourism and BeachTourism, were characterized and discussed and Chi-square independence test was performed to show the significant difference between those two profiles regarding different variables.

8.3. Clustering Park visitors without extra ontological knowledge

In this section, tourist profiles of visitors to the Ebre Delta is found by means of clustering based on the well-known Ward's criterion. In order to be able to compare our study with the previous one, we have taken into consideration the same subset of attributes, formed by 4 variables that define the tourist profile (origin, age group, accompanying persons and social class) and 6 that model the trip profile (previous planning, first reason to come, second reason to come, accommodation, length of stay and loyalty). Since the previous study did not make use of intelligent data analysis, we first have performed the clustering using traditional treatment of categorical features. The classic mixed Gibert's metric (**?**) has been used as compatibility measure for the clustering. In this experiment, age group, length of stage and loyalty are taken as numerical features, while origin, accompanying persons, social class, previous planning, accommodation, and reason 1 and 2 for the trip are taken as categorical. The dendrogram for this experiment with 8 classes is shown in Fig. 4. In this case, as usual in real complex domains, there is a very big class quite heterogeneous which seems to share all type of visitors (Table 7). Neither the second reason nor the social class discriminates at all among the classes. The main problem here is that class 964 concentrates 817 visitors, which represents the 83.8% of the total sample size and is quite heterogeneous. Thus, although some differences can be seen in other profiles, these are just referring marginal groups of the population and the information provided by this clustering is not very useful. This is a common effect

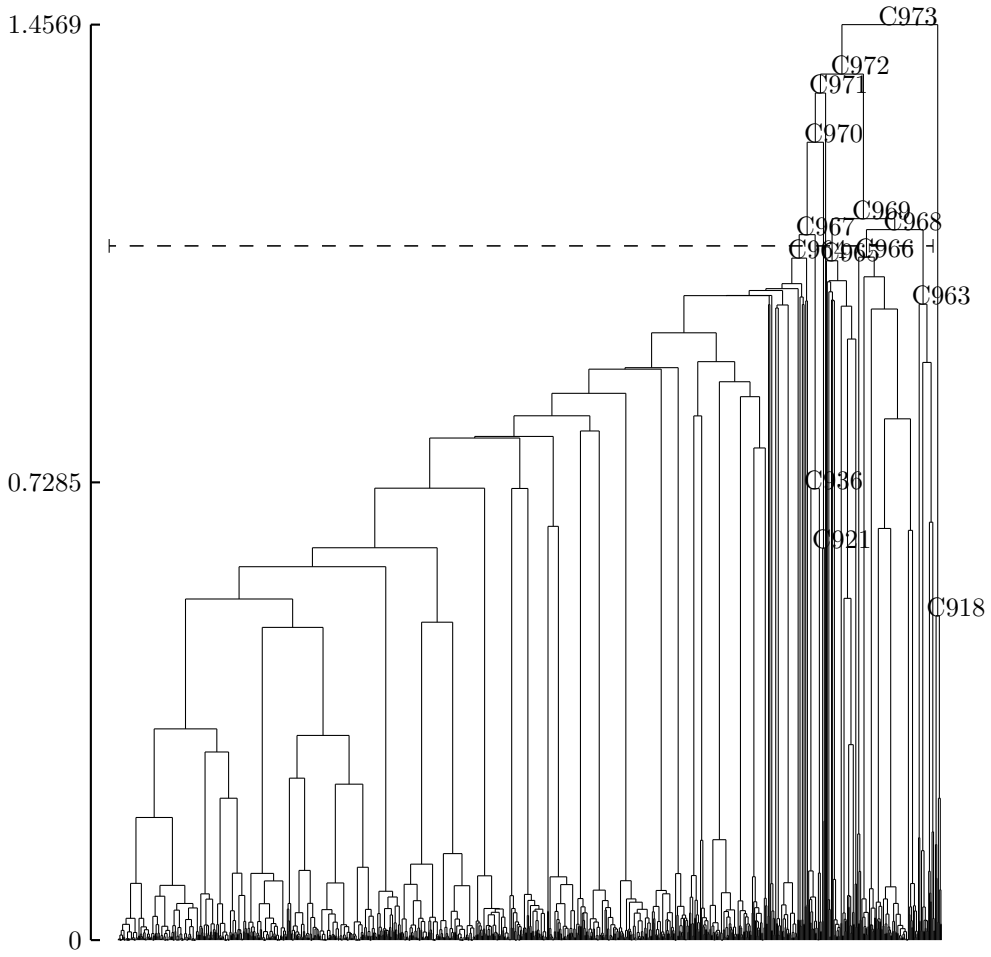

Fig. 4. Dendogram with categorical features (8 classes)

of clustering when many variables are used. Our hypothesis is that this effect could be minimized by introducing the semantics of the terms in the clustering process.

8.4. Introducing semantic information into the clustering

The semantic clustering has been applied by considering all textual variables (origin, accompanying persons, social class, previous planning, accommodation, first reason to come, and second reason) as semantic variables and using the metrics proposed before. The well known WordNet (**?**) ontology is exploited in order to estimate the semantic similarity.

From the results of this experiment, a cut in 8 classes is recommended for

Table 7. Typology of visitors to the Ebre Delta Natural Park with categorical features.

Class	n_c	Description
C864	1	Single outlier visitor
C963	20	Long stage, between 35-early 40s years, 52% stays at second home, 75% are Catalan people, it is not clear the first reason to come (some of them come for walking).
C966	72	Long stage, between 35-early 40s, 68% is at home, half Spanish, half Catalan, have a second residence near the park
C965	37	Long stage, higher fidelity, around 46 years, 65% home, 78% Catalan, their main interest is gastronomy
C921	4	Long stage, more fidelity, between 35-early 40s, 50% goes to the hotel, an important part makes reservation, part of the foreigners concentrated in this group 25% of the class are foreigners, they come for recommendation of other people, 50% Catalan, main interests: relaxation or landscape
C936	16	Shorter stage, between 35-early 40s, almost 60 % home, 80% Catalan, main interests: nature or business
C918	8	Youngs, under 30s, 50% stay in camping, 75% makes reservation, 50% Spanish, education tends to be first reason
C964	817	Shorter stage, occasional visit, between 35-early 40ss, mainly hotel, 63% Catalan, main reasons: nature, landscape and sightseeing

its interpretability. The dendrogram for this experiment is shown in Figure 5. This time, classes are more equilibrated that in the previous experiment (table 7, see table 8). With this semantic clustering we obtain the richest typology of visitors. Here, different targets of visitors are clearly identified, from the group of older people that comes only for the beach and makes long stays, to the group of young people from the neighborhood that visits the Ebre Delta Natural Park for its natural interest. Moreover, the clusters provide differences in these groups, mainly based on their origin, differentiating between foreigners, national and regional visitors, and also based on the preparation of the trip (visitors who have a reservation from visitors that have not). Finally, we discover a group that uses camping as staying form, which determines that this kind of visitors has a specific behavior with respect to the Park.

8.5. Comparison of the three data analysis

The data analysis made on the visitors to the Ebre Delta National Park in (**?**) was a pure statistical multivariate approach, consisting in projecting the data in a new artificial space of factorial components which preserves as much information of the complete data set as possible. The success in the results when using these techniques depends, in general, on the experience of the data analyst, who must be able to find proper interpretation of the selected dimensions. In this case, an interpretation for the first two dimensions was found and used to define the profiles. On the other hand, for this particular application, the data analysis technique used provided a rather unbalanced partition with two very big groups and some very small ones. Experts considered these two big groups, disregarding some interesting information contained in the other ones. However, the total variance represented by the two first dimensions considered is 56%,

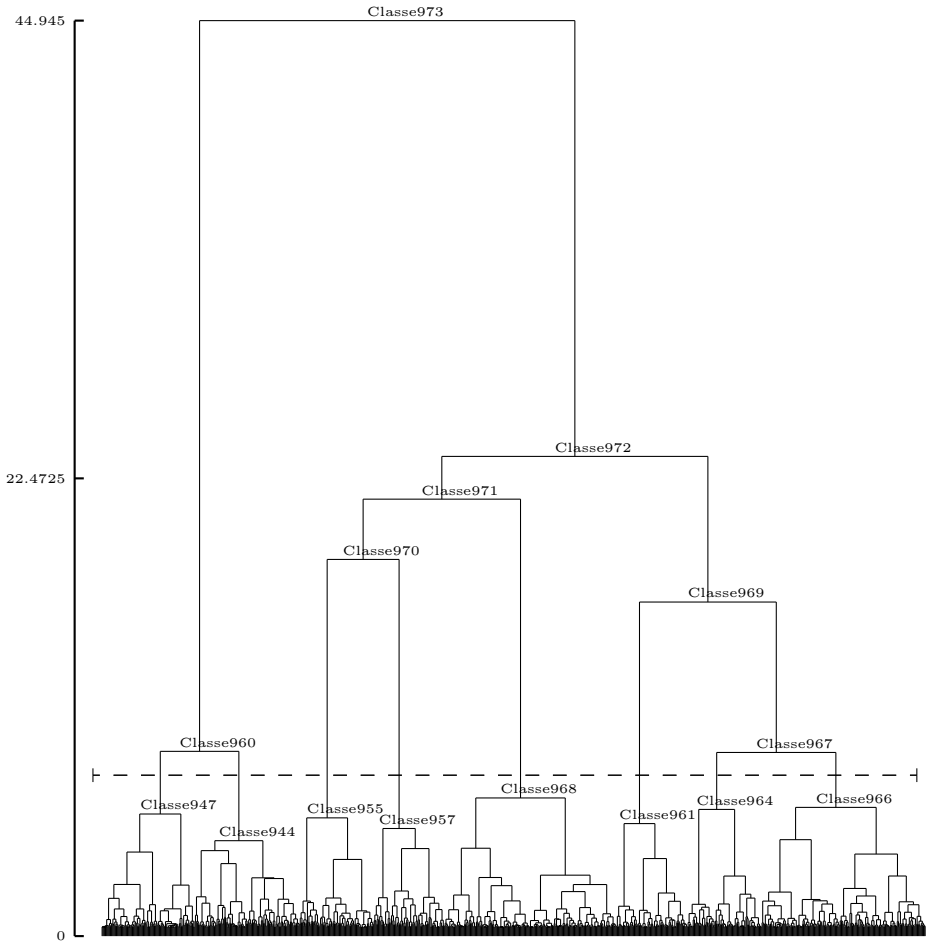

Fig. 5. Dendrogram with semantic features (8 classes)

which means that 44% of the information contained in the data set is missed. This is a rather good result when qualitative variables are used with this technique in real applications. However, the disregarded information is so important that can seriously affect the correspondence with reality. Since most of the variables were categorical, the standard techniques used, worked as usual under a pure syntactic approach, where simple binary comparisons between modalities were performed, only distinguishing equal or different responses to each question of the survey, so leading to a very poor estimation of the real similarity between different responses.

Our proposal is using clustering techniques to find the tourist profiles. In particular, unsupervised clustering algorithms are adequate to study the relationships among objects and to define a grouping with high intra-cluster homogeneity and high inter-cluster distinguishability, allowing a qualitative analysis

Table 8. Typology of visitors to the Ebre Delta Natural Park with semantic features.

Class	n_c	Description
C947	110	The 81% comes for nature, but also for relax (35%), they use mainly hotels and rural establishments (79%), they have a reservation (95%)
C966	194	They come for relax (36%), visit the family (14.4%), but the second reason is mainly nature (35%), they have no hotel, they stay at home or at a family house (68,5%), and they have no reservation (99%), this is a group of young people leaving in the area, which repeat the visits more than others.
C968	203	Short stage, around 2 days, they clearly come for nature reasons (91.6%) and second for relax and wildlife (43.6%), they are in hotels or apartments (44.6 %) although they have not reservation, mainly Catalan and Spanish
C955	88	The first reason for coming is heterogeneous (nature, relaxation, beach, landscapes), the second is nature, they stay in a camping (90%), the half have a reservation, mainly Catalan and Spanish but also concentrates a big proportion of foreigners
C944	124	Relax and beach (46%) are the first reasons for coming and second is nature (40%), they stay at hotels, cottages(72%), and have reservation (88%). This is a group of slightly older people programming the stay in hotel or apartment, looking for relax or beach
C964	88	Wildlife and the landscape are the first reasons for coming (67%), but also for culture (19.5%) and the second reason is nature, they are mainly in hotel (54%). They are mainly Catalan or Spanish.
C957	84	Stay longer, slightly older that the rest, nature (38%) and beach (16%) are the main interest and second main interest is wildlife, most of them are foreigners with a second home, or that stay in an apartment.
C961	84	They all come for beach, their secondary interests are equally relaxation and nature, they live near the park and their visit is improvised, the stage is longer.

of the structural relationships between the concepts expressed by data. The analysis is performed directly on the original variables space, guaranteeing the direct interpretability of the results. However, when they are applied in categorical variables, the same restrictions mentioned before appear, making difficult the establishment of differences between the objects. This can be seen in the results obtained with the partition of the data using a classical approach with only numerical and categorical features. The clustering generates a big class (with 83.8% of the tourists) and other 7 small classes. These results do not determine a typology of tourists with class-specificities allowing the park managers to improve recreational uses planning. Therefore, although the interpretation of the small classes is possible (see Table 7), from the manager point of view, this partition is useless because the majority of visitors belongs to the same profile.

Results are much better in the semantic-based approach proposed in this paper, where the clustering method is able to compare the values of the features in terms of their semantics, relating them to concepts in a given ontology. As it has been described in section 8.4, the partition obtained with this semantic-based clustering generates 8 clusters of more homogeneous dimension. This is an important fact, since now we can identify typologies of visitors that represent a significant proportion of the total number of visitors. From the dendrogram

in Figure 5, it can also be seen that we have obtained clusters with high cohesion, which means that the distances between the members of the cluster are quite small in comparison with their distances with objects outside the cluster. Moreover, if the level of partition is increased, then the cohesion of the clusters decreases quickly, which also indicates that the clusters are well defined.

This clustering is coherent with the grouping made by (**?**) using multivariate analysis, because the variables about the reasons for visiting the park have a great influence in the formation of the groups. Interests on nature, beach and relax are present in different classes. However, thanks to the semantic interpretation of the concrete textual values provided by the respondents, we have been able to identify that visitors interested in nature are similar to those interested in wildlife. The system has been also able to identify the similarity between hotels and cottages and between second homes and familiar houses. This proofs that the estimation of the relative similarities among objects in terms of the meaning of the terms improves the final grouping. In this way, the two types of visitors identified in statistical analysis as Ecotourism and Beach Tourism (mainly guided by the variable First Reason to visit the park) have now been refined as follows:

- Ecotourism: visitors that stay in hotels and apartments for relax (C938), visitors with familiars or a second residence (C966), Catalan and Spanish visitors interested in wildlife (C965) and tourists interested in culture (C956).
- Beach tourism: older people staying in hotels or apartments looking for relax and people that live near the park and go to the beach quite frequently.
  Notice that this is a more rich classification that establishes clear profiles of visitors with different needs. This is according with the hypothesis of the experts that suggested that the park attracts highly different types of visitors. After this study, the manager may study different actions according to the different types of demand.

9. Discussion and Conclusions

The exploitation of data from a semantic point of view establishes a new setting for data mining methods, particularly in contexts where heterogeneous variables appear. This paper reports the possibility of improving the comparison between pairs of objects by using reference ontologies, when available. This permits to take into account the semantics associated to categorical values. For those cases in which the ontology is not available, or not reliable, the original Gibert's mixed metrics can be used, distinguishing only between qualitative and quantitative variables. Thus, our proposal is currently oriented to those particular domains in which background well established ontologies are available. Specific domain ontologies developed by international committees, currently accepted as standards in some fields, can be a good knowledge source. This is the case of SNOMED-CT for biomedicine, the YAGO ontology that covers entities, persons and organizations, FOAF (Friend of a Friend) ontology describing relations between people, among others, as mentioned in section (§8). However in the field of Tourism (considered in this paper), there is not an ontology yet sufficiently large to cover the domain nor sufficiently consensued to be accepted as an stardard (§8). In this context, the introduction of general purpose ontologies, like Wordnet, overcomes the blind syntactic approach of original Gibert's mixed metrics.

Obviously, the improvement on the results is directly related with the quality

of the reference ontology itself, in the same way as the quality of the analysis results is directly related with the quality of data (this makes the preprocessing step crucial for the whole knowledge discovery process). The selection of the right ontology is well known as a complex issue, as usual in all knowledge-based systems, and a complete new experimental design should be designed to study the impact of the ontology in an exhaustive case study, with different datasets and different structural problems. However, for the particular methodology presented in the paper, related to the use of an ontology to improve a clustering process, and considering some previous works and the experiences presented in the paper, it seems that clustering only numerical variables provides poorer results than clustering heterogeneous data matrices (**?**); this, in turn, seems a poorer solution than using general purpose ontologies for some qualitative variables, as shown in the results presented for Ebre Delta Natural Park 8; finally, it seems that using domain-specific domain ontologies may provide even richer results, provided, of course, that the quality of the ontology has been properly tested a priori. For this particular context we would recommend to include a general purpose ontology when available, unless a sound and well accepted specific ontology is available. In this case, both can be considered, just to avoid lack of terms in the specific ontology (**?**). More work is in progress to verify if this is a general property of the method or it also holds in some specific kind of problems.

Although great part of the paper has been focused on semantic variables, which allow the semantic interpretation by means of ontologies, it is necessary to remind that our approach is able to deal also with classical numerical variables or measurements and categorical variables. None of the referred works combines numerical, syntactic and semantic information in a single and integrated analysis in the context of clustering. In this regards, the general framework of compatibility measures (**?**), considering numerical and categorical variables together has been extended to also consider what has been named *semantic variable*. This permits to introduce in the analysis the different kind of variables in its original form, simplifying the data preparation step and avoiding arbitrary decisions on transforming data that can produce non-desired biases, or mask some relationships among variables.

The *Superconcept-based distance* (SCD) has been introduced as a new proposal for computing the similarity/distance between semantic variables. It exploits the geometrical structure of the reference ontology. The *Superconcept-based distance* is an Euclidean distance computed on a binary representation of the taxonomic ontology structure, normalized to take into account the relative importance of the non common information versus the total information of the compared concepts. The explicit construction of the underlying binary matrix mentioned above is n^2, (n being the number of terms in the ontology). An equivalent rewriting of the original expression has been found as a function of common superconcepts of the compared terms, which are directly provided by the ontology itself. This avoids the explicit construction of the binary matrix, reducing the complexity to $2n$ in the worst case.

Several benchmarks support that SCD performs better than other proposals from the literature. Two experiments have been included in this paper, one regarding a general purpose ontology (Wordnet) and another with specialized ontology in the biomedical field (SNOMED-CT). In both cases, measures based only on the minimum path length between concepts provide poor results, whereas improvements are found as more taxonomic information is considered. That is why SCD is the one better correlating with human judgements. This correlation

with human judgements are used as performance indicators, under the assumption that the distance provided by the experts is coherent with the structure represented in the domain ontologies.

SCD has also the advantage that do not relay in a domain corpus to compute semantic evidence, like measures based on co-occurrences. This is specially interesting when no corpus exists or data is unavailable for privacy reasons, frequent situation in biomedicine. Also, SCD do not require available experts providing subjective quantitative similarities as for distance metric learning approach. Performance of SCD only depends on the quality of the ontology itself. As discussed above, some bias might be introduced if the ontology is not complete (**?**), but it is clear that experts feel more confortable providing relationships between concepts in form of ontologies rather than quantifying similaritites. In the worst case, the paper shows that general purpose ontologies like Wordnet can be always used with good results.

The paper proposes to extend the Gibert's mixed metrics (**?**) by introducing a new term for semantic variables, based on the *Superconcept-based distance*. This proposal fits with the idea of defining compatibility measures to analyze heterogeneous matrices, already established by Anderberg in the 70s. In our approach, convex combinations of distances are used as compatibility measures, with an implicit assumption that all the variables have a similar importance for the analysis. For those applications where it is relevant to assign different weights to the variables, a further extension with weighted subdistances should be considered. As usual in convex combinations, the extended Gibert's metrics represents, in fact an infinite family of distances, indexed by three parameters (α, β, γ). For hierarchical clustering purposes, only elements bounded by $\alpha + \beta + \gamma = 1$ are considered,since this guarantees to find all possible hierarchies. An specific proposal to choose the parameter values is presented and justified in section §3.2. The SCD by itself, as well as the extended Gibert's metrics, are suitable to be used in any distance based method, from clustering, to case-based reasoning or multivariate analysis, provided that a reference ontology is available to involve the semantics of the terms into the analysis.

In this work we focused on hierarchical clustering, which is highly related with our current research and applications. Improvements of taking into account the semantics of the terms in hierarchical clustering processes have been tested with two different data sets. The first case study contained a reduced set of tourist city destinations. The second case study presents an application to a real survey done to about 1000 visitors of a Natural Protected Park in Catalonia. In both case studies clustering considering the semantic variables as ordinary categorical variables, with original Gibert's mixed metrics and dismissing the available reference ontology is compared with a second clustering taking into account WordNet ontology for semantic variables and using SCD. In both cases Ward criterion is used since it tends to provide more interpretable results than other hierarchical clustering criteria. Better results are obtained using semantic variables. Being the role of the ontology the only factor changing between the two approaches compared, the improvements observed in second results can be directly assigned to the introduction of the ontology in the process.

Original Gibert's metrics use χ^2-distance for categorical variables, which only considers equality *vs* inequality of terms scaled by their rarity. Extending the Gibert's metrics with a semantic term, makes for example mountain and valley more similar than mountain and beach, this transporting semantics to the final clusters, which consider the meaning of the words. Thus, the result becomes

more interpretable. Clusters are also improved from a structural point of view, as more balanced clusters, in both size or semantic compactness, are obtained. In the partitions obtained without considering ontologies, clusters look more heterogeneous in terms of the distinguishability of classes and understanding of class particularities.

These improvements are the key to move the results of a clustering process to a real decision-making process in the target domain. Unfortunately, too often, very well structured clusters (from a technical point of view) are never used to support decisions in the target domain, because end-users cannot understand well their meaning. Improving clustering algorithms to guarantee that classes are built regarding *also* to the semantics in the domain field contributes to bridge this gap. Another important issue in this line is to define a complete automatic process to produce the final interpretation of classes, trying to produce directly understandable descriptions of classes for the end-user, thus completing the whole knowledge discovery process. We are currently working in some proposals related to this field. In (**?**) concepts are associated to classes containing their relevant issues. In (**?**) automatic construction of traffic light panels is proposed as a visual symbolic abstraction of the class particularities to assist the comprehension process of classes. The automatic interpretation of classes is still an open problem and requires further efforts to establish the best methodology.

Certainly we tested the effect of considering reference ontologies by means of a semantic distance in a very particular hierarchical clustering method and we cannot ensure this will be extendable to other type of algorithms. However, we guess this improvement might be observed in other distance based methods, since the effect of the ontology is only modifying the distances among objects and is not related with the particular hosting method in which it is introduced. Further research is required to verify that using semantic similarities to take into account reference ontologies in distance based metrics improves the results for the meaningfulness point of view.

Acknowledgements. This work is partially supported by the Spanish Ministry of Science and Innovation (DAMASK, TIN2009-11005) in the Spanish Government PlanE (Spanish Economy and Employment Stimulation Plan). Montserrat Batet has been supported by a research grant provided by the Universitat Rovira i Virgili. The testing part has been possible thanks to the data provided by "Observatori de la Fundació d'Estudis Turístics Costa Daurada" and "Parc Nacional del Delta de l'Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)". Thanks to S. Clavé for his close collaboration. The authors also acknowledge the collaboration of E. Fourier, D. Corcho, N. Malé and N. Corral in the data preparation.

References

[1] M. R. Anderberg. *Cluster Analysis for Aplications*. Monographs and Textbooks on Probability and Mathematical Statistics. Acad. Press, Inc., NY, 1973.

[2] Mihael Ankerst and Markus M. Breunig abd Hans-Peter Kriegel abd Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM SIGMOD international conference on Management of data*, pages 49–60, 1999.

[3] R. Annichiarico, K. Gibert, and *e. a.* Qualitative profiles of disability. *JRRD*, 41(6A):835–845, 2004.

[4] S. Anton-Clavé, M. G. Nel-lo, and A. Orellana. Coastal tourism in natural parks. an analysis of demand profiles and recreational uses in coastal protected natural areas. *Revista Turismo y Desenvolvimiento*, 7-8:9–81, 2007.

[5] Cláudia Antunes. Onto4ar: a framework for mining association rules. In *Proceedings ECML/PKDD07*, 2007.
[6] Elena Baralis, Luca Cagliero, Tania Cerquitelli, Paolo Garza, and Marco Marchetti. Casmine: providing personalized services in context-aware applications by means of generalized rules. *Knowledge and Information Systems*, 28:283–310, 2011.
[7] M. Batet, A. Valls, and K. Gibert. A distance function to assess the similarity of words using ontologies. In *Proc. XV ESTYLF'10*, pages 561–566, 2010.
[8] Montserrat Batet, David Sánchez, Aida Valls, and Karina Gibert. Semantic similarity estimation from multiple ontologies. *Applied Intelligence*, 38:29–44, 2013.
[9] J.P. Benzécri. *Pratique de l'analyse des données. Analyse des Correspondances, Expose Elementaire*, volume 1. Paris: Dunod., 1980.
[10] A. Bernstein, F. Provost, and S. Hill. Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):503 – 518, april 2005.
[11] Emmanuel Blanchard, Mounira Harzallah, and Pascale Kuntz. A generic framework for comparing semantic similarities on a subsumption hierarchy. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikolaos M. Avouris, editors, *Procedeings of 18th European Conference on Artificial Intelligence (ECAI)*, volume 178, pages 20–24, Patras, Greece, 2008. IOS Press.
[12] C. Breen, L. Khan, and A. Ponnusamy. Image classification using neural networks and ontologies. In *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pages 98 – 102, sept. 2002.
[13] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5:6–6, 2004.
[14] Paul Buitelaar and et al. Ontoselect: A dynamic ontology library with support for ontology selection. In *Proc. of the Int'l Semantic Web Conf.*, 2004.
[15] Jorge Cardoso. Developing an owl ontology for e-tourism. In Jorge Cardoso and AmitP. Sheth, editors, *Semantic Web Services, Processes and Applications*, volume 3 of *Semantic Web and Beyond*, pages 247–282. Springer US, 2006.
[16] Luigi Ceccaroni, Ulises Cortés, and Miquel Sánchez-Marré¨. Ontowedss: augmenting environmental decision-support systems with ontologies. *Environmental Modelling & Software*, 19(9):785 – 797, 2004. ¡ce:title¿Environmental Sciences and Artificial Intelligence¡/ce:title¿.
[17] H. Cespivova, J. Rauch, Svatek V., Kejkula M., and Tomeckova M. Roles of medical ontology in association mining crisp-dm cycle. In *ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO 2004)*, 2004.
[18] C. Chemudugunta, P. Smyth, and M. Steyvers. Combining concept hierarchies and statistical topic models. In *Proc. CIKM)*, pages 1469–1470, 2008.
[19] Chang Choi, Miyoung Cho, Junho Choi, Myunggwon Hwang, Jongan Park, and Pankoo Kim. Travel ontology for intelligent recommendation system. In *Modelling Simulation, 2009. AMS '09. Third Asia International Conference on*, pages 637 –642, may 2009.
[20] P. Cimiano. *Ontology Learning and Population from Text. Algorithms, Evaluation and Appli-cations*. Springer-Verlag, 2006.
[21] William R. Dillon and Matthew Goldstein. *Multivariate Analysis: Methods and Applications*. Wiley, 1984.
[22] Li Ding and et al. Swoogle: A search and metadata engine for the semantic web. In *Proc. XIIIth ACM int'l CIKM04*, pages 652–659, NY, 2004. ACM Press.
[23] Rezende Domingues. Using ontologies to facilitate the analysis of association rules. In *ECML/PKDD07 Workshop on Knowledge Discovery and Ontologies*, 2005.
[24] D. Downey and et al. Locating complex named entities in web text. In *Proc. 20th IJCAI*, pages 2733–2739, 2007.
[25] Faezeh Ensan and Weichang Du. A knowledge encapsulation approach to ontology modularization. *Knowledge and Information Systems*, 26:249–283, 2011.
[26] M. Epler Wood. Ecotourism: Principles, practices and policies for sustainability. *United Nations Publications*, 2002.
[27] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In AAAI Press, editor, *KDD'96*, pages 226–231, 1996.
[28] Jianping Fan, Yuli Gao, and Hangzai Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *Image Processing, IEEE Transactions on*, 17(3):407 –426, march 2008.

[29]U. Fayyad and *et alt. Advances in KDD and Data Mining*, chapter From Data Mining to Knowledge Discovery: An overview. AAAI/MIT Press., 1996.
[30]C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press. More information: http://www.cogsci.princeton.edu/ wn/, Cambridge, Massachusetts, 1998.
[31]V. Ganti, J. Gehrke, and R. Ramakrishnan. Cactus: Clustering categorical data using summaries. In *Proc. 5th ACM SIGKDD Int. Conf on Knowledge Discovery in Data Mining*, pages 73–83, 1999.
[32]AnaCristinaB. Garcia, Cristiana Bentes, RafaelHeitorC. Melo, Bianca Zadrozny, and ThadeuJ.P. Penna. Sensor data analysis for equipment monitoring. *Knowledge and Information Systems*, 28:333–364, 2011.
[33]Angel Garcia-Crespo, Jose Luis Lopez-Cuadrado, Ricardo Colomo-Palacios, Israel Gonzalez-Carrasco, and Belen Ruiz-Mezcua. Sem-fit: A semantic based expert system to provide recommendations in the tourism domain. *Expert Systems with Applications*, 38(10):13310 – 13319, 2011.
[34]K. Gibert and U. Cortés. Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing*, 4(3):251–266, 1997.
[35]K. Gibert and U. Cortés. Clustering based on rules and knowledge discovery in ill-structured domains. *Computación y Sistemas.*, 1(4):213–227, abril 1998.
[36]K. Gibert, Garcï¿½a-Rudolph, and et al. Response to tbi-neurorehabilitation through an ai& stats hybrid kdd methodology. *Medical Archives*, 62(3), 2008.
[37]K. Gibert, Nonell, and et al. Kdd with clustering: impact of metrics and reporting phase by using klass. *Neural Net.World*, 15(4):319–326, 2005.
[38]K. Gibert and R. Nonell. Impact of mixed metrics on clustering. *Lecture Notes on Computer Science*, 2905:464–471, Nov 2003.
[39]K. Gibert and R. Nonell. Pre and post-processing in klass. In *iEMSs 2008 Procs.*, pages 1965–1966. iEMSs, 2008.
[40]K. Gibert, G. Rodríguez-Silva, and I. Rodríguez-Roda. Knowledge discovery with clustering based on rules by states: A water treatment application. *Environmental Modelling and Software*, 25:712–723, 2010.
[41]K Gibert, L Salvador-Carulla, and C Garcï¿½a Alonso. Integrating clinicians, knowledge and data: Expert-based cooperative analysis in medical decision support. *Health Research Policy and Systems*, in press, 2010.
[42]K. Gibert, Z. Sonicki, and J. C. Martín. Impact of data encoding and thyroids dysfunctions. *Studies in Health Tech. and Informatics*, 90:494–498, 2002.
[43]Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering*. 2nd printing. Springer-Verlag. ISBN: 1-85233-551-3, 2004.
[44]J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–874, 1971.
[45]J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–874, 1971.
[46]S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. In *Proc. 15th Int. Conf. on Data Engineering*, pages 512–521, 1999.
[47]EM. Helsper and LC. van der Gaag. Building bayesian networks through ontologies. In *Proceedings of ECAI2002*, pages 680–684, 2002.
[48]A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 541 – 544, nov. 2003.
[49]Y. Huang and L. Bian. A bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the internet. *Expert Systems with Applications*, 36(1):933 – 943, 2009.
[50]Yvo I. Biomedical ontologies and text mining for biomedicine and healthcare: a survey. *Journal of computing science and engineering*, 2(2):109–136, 2008.
[51]M. Ichino and H. Yaguchi. Generalized Minkowski Metrics for Mixed feature-type data analysis. *IEEE Tr. on SMC*, 22(2):146–153, 1994. April.
[52]Meng X. J., Chen Q. C., and Wang X. L. A tolerance rough set based semantic clustering method for web search results. *Information Technology Journal*, 8(4):453–464, 2009.
[53]L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang. Ontology-based distance measure for text clustering. In *SIAM SDM workshop on text mining*, Bethesda, Maryland, USA, 2006.
[54]D. Vrecko K. Gibert, D. Conti. Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. *Environmental Engineering and Management Journal*, 11(5):931–944, 2012.

[55]Rasmus Knappe. *Measures of Semantic Similarity and Relatedness for Use in Ontology-based Information Retrieval*. PhD thesis, Roskilde University, DN, 2005.
[56]C. Lamsfus, C. Grun, A. Alzua-Sorzabal, and H. Werthner. Context-based matchmaking to enhance tourists' experience. *Journal for the Informatics Professional*, 203:17 – 23, 2010.
[57]Claudia Leacock and Martin Chodorow. *Combining local context and WordNet similarity for word sense identification*, chapter WordNet: An electronic lexical database, pages 265–283. MIT Press, 1998.
[58]B Lemaire and G Denhière. Effects of high-order co-occurrences on word semantic similarities. *Current Psychology Letters - BBC*, 18(1), 2006.
[59]Alexander Maedche and Valentin Zacharias. Clustering ontology-based metadata in the semantic web. In ., volume 2431 of *LNCS*, pages 348–360, London, UK, 2002. Springer-Verlag.
[60]G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
[61]I. Minguez, D. Berrueta, and L. Polo. Cruzar: An application of semantic matchmaking to e-tourism. In *Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications. IGI Global*, pages 255–271, 2010.
[62]Antonio Moreno, Aida Valls, David Isern, Lucas Marin, and Joan Borras. Sigtur/e-destination: Ontology-based personalized recommendation of tourism and leisure activities. *Engineering Applications of Artificial Intelligence*, 26(1):633 – 651, 2013.
[63]G. Nakhaeizadeh. Classification as a subtask of of Data Mining experiences form some industrial projects. In *Proc. IFCS, vI*, pages 17–20, march 1996.
[64]K. Ovaska, M. Aakso, and S. Hautaniemi. Fast gene ontology based clustering for microarray experiments. *BioData Mining*, 1(11), 2008. open access doi:10.1186/1756-0381-1-11.
[65]G. Pandey, C. L. Myers, and V. Kumar. Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics*, I:0:42, 2009.
[66]Paul Pavlidis, Jie Qin, Victoria Arango, JohnJ. Mann, and Etienne Sibille. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research*, 29:1213–1222, 2004.
[67]T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40:288–299, 2007.
[68]A. Pérez-Bonilla and K. Gibert. Automatic generation of conceptual interpretation of clustering. In *Progress in Pattern Recognition, Image analysis and Applications. LNCS*, volume 4756, pages 653–663. Springer, 2007.
[69]R. Rada, H. Mili, E. Bichnell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans. SMC*, 9(1):17–30, 1989.
[70]Dnyanesh Rajpathak, Rahul Chougule, and Pulak Bandyopadhyay. A domain-specific decision support system for knowledge discovery using association and text mining. *Knowledge and Information Systems*, 31:405–432, 2012.
[71]Henri Ralambondrainy. *A clustering method for nominal data and mixture of numerical and nominal data. Clasification and Related Methods of Data Analysis*. H.H.Bock, Elsevier Science Publishers, B.V. (North-Holland), 1988.
[72]Kotagiri Ramamohanarao, P. Radha Krishna, and et al. *Advances in Databases: Concepts, Systems and Applications DASFAA*, volume 4443. ., 2007.
[73]Chiara Renso, Miriam Baglioni, Jose António Macedo, Roberto Trasarti, and Monica Wachowicz. How you move reveals who you are: understanding human behavior by analyzing trajectory data. *Knowledge and Information Systems*, pages 1–32, 2012.
[74]P. Resnik. Using information content to evalutate semantic similarity in a taxonomy. In *Proc. IJCAI 95*, pages 448–453, Montreal, Canada, 1995.
[75]H. Rubenstein and J. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
[76]M. Ruiz-Montiel and J.F. Aldana. Semantically enhanced recommender systems. In *Proceedings of the OTM 2009 conference, workshop on The Move To Meaningful Internet Systems*, pages 604 – 609, 2009.
[77]David Sánchez, Montserrat Batet, Aida Valls, and Karina Gibert. Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems*, 35:383–413, 2010.
[78]Pinar Senkul and Suleyman Salin. Improving pattern quality in web usage mining by using semantic information. *Knowledge and Information Systems*, 30:527–541, 2012.
[79]K. Shin and A. Abraham. *IDEAL 2006, LNCS*, chapter Two Phase Semi-supervised clustering using background knowledge, pages 707–712. Springer-Verlag, 2006.

[80]R.R. Sokal and P.H.A. Sneath. *Principles of numerical taxonomy*. Freeman, San Francisco, 1963.
[81]S. Song, Z. Guo, and P. Chen. Fuzzy document clustering using weighted conteptual model. *Information technology journal*, 10(6):1178–1185, 2011.
[82]M. Steyvers, P. Smyth, and C. Chemuduganta. Combining background knwoledge and learned topics. *Topics in Congnitive Science*, 3:18–47, 2011.
[83]M. Thangamani and P. Thangaraj. Integrated clustering and feature selection shceme for text documents. *Journal of Computer Science*, 6(5):536–541, 2010.
[84]Nicki Tiffin, Janet F. Kelso, Alan R. Powell, Hong Pan, Vladimir B. Bajic, and Winston A. Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, 33(5):1544–1552, 2005.
[85]A. Tversky. Features of similarity. *Phychological Review*, 84:327–352, 1977.
[86]A. Valls, M. Batet, and E. Lopez. Using experts rules as background knowledge in the clusdm methodology. *EJOR*, 193(3):864–875, 2009.
[87]F. Wang, J. Sun, and S. Ebadollahi. Integrating distance metrics learned from multiple experts and its application in inter-patient similarity assessment. *Proc. 11th SIAM International Conference on Data Mining (SDM)*, pages 59–70, 2011.
[88]Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proc. 32nd ACL*, pages 133 –138, New Mexico, USA, 1994.
[89]E.P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. *NIPS*, pages 505–512, 2002.
[90]SY. Yang, PC. Liao, and CS. Ho. An ontology-supported case-based reasoning technique for faq proxy service. In *Proceedings of the 17th Int. Conf. on Software Engineering and Knowledge Engineering*, pages 639–644, 2005.
[91]L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, 8:199–249, 1975.
[92]O. Zamir and O. Etzioni. Grouper: a dynamic clustering itnerface to web search results. *Computers and Networks*, 31:1361–1374, 1999.
[93]Jun Zhang, Adrian Silvescu, and Vasant Honavar. Ontology-driven induction of decision trees at multiple levels of abstraction. In Sven Koenig and RobertC. Holte, editors, *Abstraction, Reformulation, and Approximation*, volume 2371 of *Lecture Notes in Computer Science*, pages 316–323. Springer Berlin Heidelberg, 2002.

Author Biographies

Karina Gibert bla bla bla.

insert photo

Aida Valls is a lecturer at the Department of Computer Science and Mathematics in Universitat Rovira i Virgili (URV). She received the PhD in computer science from the Technical University of Catalonia in 2002. Her research interests include multiple criteria decision making, recommender systems, data mining and privacy preserving. Her work is mainly focused on the treatment of linguistic and semantic information. She has participated in several Spanish and EU research projects, with applications in Tourism, Environment Risk Management and Health Care. She is the author of more than 80 papers in international journals and conferences. She is currently the Head of the PhD Program in Computer Science at URV.

insert photo

Montserrat Batet is a Researcher at the University Rovira i Virgili's Computer Science and Mathematics Department. He received a Ph.D. on Artificial Intelligence from URV in 2011. Her research interests are semantic similarity, semantic clustering and privacy protection of textual data. She has been involved in several research projects (National and European), and published more than 30 papers including 14 in high quality international journals.

Correspondence and offprint requests to: Karina Gibert, Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, C/ Jordi Girona, Barcelona, Spain. Email: karina.gibert@upc.edu