TIN2009-11005
DAMASK

Data-Mining Algorithms with Semantic Knowledge

PROYECTO DE INVESTIGACIÓN
PROGRAMA NACIONAL DE INVESTIGACIÓN FUNDAMENTAL,
PLAN NACIONAL DE I+D+i 2008-2011
ÁREA TEMÁTICA DE GESTIÓN: Tecnologías informáticas

Deliverable D4

New techniques for semantic similarity measurement

Authored by
Aïda Valls, Universitat Rovira i Virgili

Work co-authors of the DAMASK project
Montserrat Batet, Universitat Rovira i Virgili
David Sánchez, Universitat Rovira I Virgili

Document information

project name:	DAMASK	
Project reference:	TIN2009-11005	
type of document:	Deliverable	
file name:	D4.pdf	
version:	2.0	
authored by:	A. Valls	01/08/2011
co-authored by	M. Batet, D. Sánchez	
released by:	A. Valls	01.08.2011
approved by:	Co-ordinator	Antonio Moreno

Document history

version	date	reason of modification
1.0	01.July.2011	Definition of a similarity measure for comparing conceptual and textual attributes.
2.0	01.August.2011	Addition of the semantic clustering section.

Table of Contents

1 Introduction

This document corresponds to Task 2 of the DAMASK project. Task T2 is focused on the goal O2 of the project: design of a clustering method based on ontologies. The inputs of this task are (1) a data matrix object × attribute (e.g. touristic destinations) and (2) a domain ontology. Based on those inputs, a method for automatically building clusters is needed. During the clustering, contextual knowledge provided by the domain ontology is used. Finally, an automatic interpretation process of the clusters is required, in order to obtain a semantic description of the clusters that can help the user in his/her decision making tasks.

This deliverable is the result of the subtask T2-3 developed from month 13 until month 21. The complete schedule of the tasks in the DAMASK project is given in Figure 1. Task 2-3 corresponds to the study of the adaptation of the traditional clustering algorithms to permit the use of semantic similarity measures based on ontologies and linguistic terms.

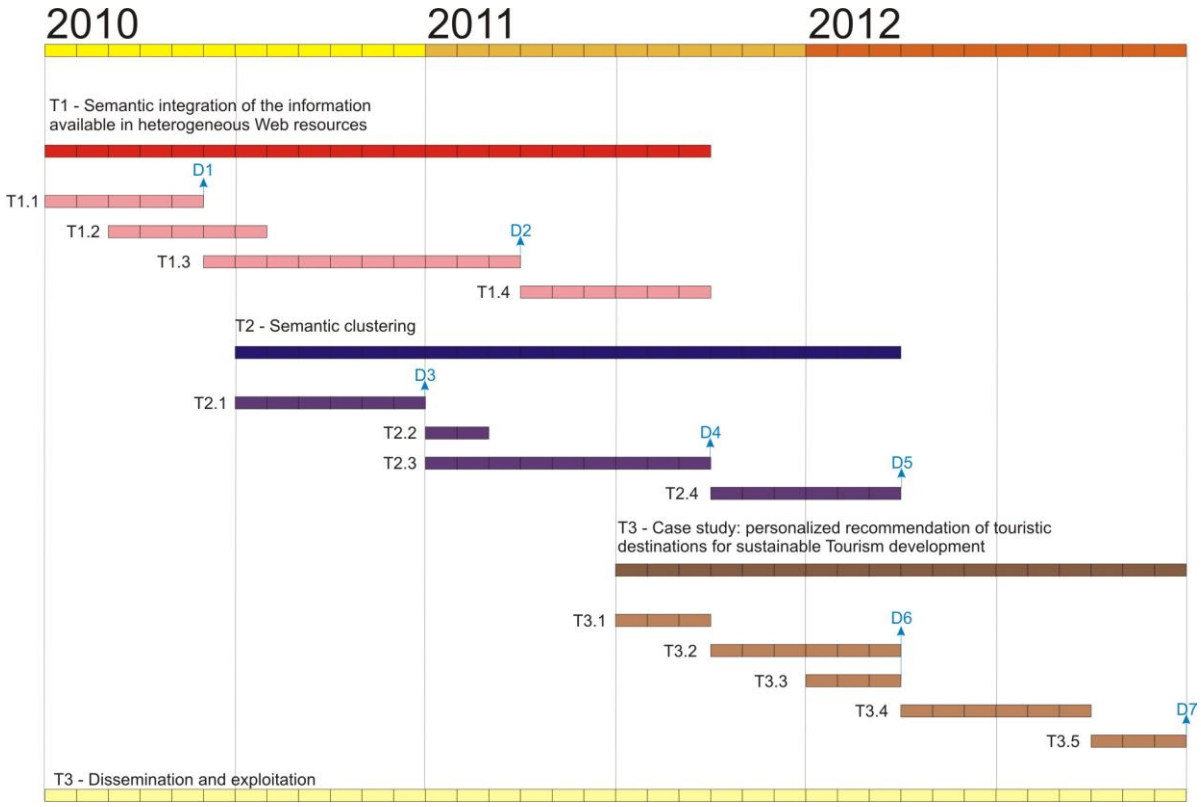

Figure 1: Tasks of DAMASK

An extensive analysis of the weak points of the existing semantic similarity measures was made in Task 2.1 and presented in deliverable D3. In this document we mainly concentrate on proposing some new semantic similarity measures for pairs of objects that solve the limitations of the previous approaches with respect to the improvement of the clustering of objects. A special section will explain how to include this semantic similarity measures into some clustering algorithms.

The work presented in this deliverable has been done mainly as part of the Ph.D. Thesis of Montserrat Batet having as advisors Dr. Aïda Valls (member of DAMASK project) and Dr. Karina Gibert (from the Universitat Politècnica de Catalunya in Barcelona, Spain). We kindly thank the collaboration of Karina Gibert with this project.

Several publications have been obtained as a result of the work done in this task. They can be found in the References section.

2 New semantic similarity measures based on ontologies

We have studied and contributed in the two main fields of semantic similarity computation, which were presented in Deliverable D3. These two fields are:

1. Ontology-based measures relying on:
    1.1. Edge-counting
    1.2. Features
    1.3. Information Content (corpora-dependent or intrinsic to an ontology)
2. Distributional measures based on:
    2.1. First-order co-occurrence
    2.2. Second-order co-occurrence (relying on corpora or on structured thesaurus glosses)

In this document we will present, first, an edge-counting method for semantic similarity computation is proposed. It is based on the exploitation of the taxonomic knowledge available in an ontology for the compared concepts. Its design aims to overcome some of the limitations and improve the accuracy of previous works based on edge-counting that only consider the minimum path between concepts.

Secondly, we will summarize our proposal for a new way to measure the information content (IC) by exploiting the Web information distribution instead of tagged. In this manner we aim to overcome the high data sparseness and the need of manual tagging that corpora-based IC computation models require. After that, in order to minimize the ambiguity of language and improve the accuracy of IC computation, we propose a method to contextualize the IC computation by exploiting the taxonomical knowledge available in an ontology.

Finally, with respect to distributional measures, in third place, we will present a contribution on collocation measures. Those measures are evaluated and compared in section 2.4 and their applicability for semantic clustering is analysed in section 3. A more comprehensive definition and evaluation of those proposals can be found in the papers indicated, as well as in the PhD thesis (Batet 2011).

2.1 SC: Superconcept-based distance

Path-based measures are computationally efficient since no pre-calculus is needed. However, due to their simplicity, they do not capture enough semantic evidence to provide assessments as reliable as other types of measures (as it will be shown in the evaluation section). Taking this into account, we have proposed a new similarity measure that can achieve a level of accuracy similar to corpus-based approaches but retaining the low computational complexity and lack of constraints of path-based measures (*i.e.*, no domain corpus is needed).

Analyzing the basic hypothesis of path-based methods, we can notice that these measures consider the minimum path length between a pair of concepts, which is the sum of taxonomical links between each of the concepts and their LCS. The path is composed, in addition to the LCS, of nodes corresponding to non-shared superconcepts (*i.e.*, subsumers of the evaluated terms), which are taken as an indication of distance. However, if one or both concepts inherit from several *is-a* hierarchies, all possible paths between the two concepts are calculated, but only the shortest one is kept. In consequence, the resulting path length does not completely measure the total amount of non-common superconcepts modelled in the ontology (*i.e.*, subsumers of a concept). Due to this reason, for complex and large taxonomies, covering thousands of interrelated concepts included in several overlapping hierarchies, and an extensive use of multiple inheritance (i.e. a concept is subsumed by several superconcepts), path-based measures waste a great amount of explicit knowledge.

The main idea of our proposal relies on taking into account all the available taxonomical evidence (*i.e.*, all the superconcepts) regarding the evaluated concepts (and not only the minimum path) could provide more accurate assessments.

Let us define the full concept hierarchy or taxonomy (H^C) of concepts (C) of an ontology as a transitive is-a relation $H^C \in C \times C$.

Let us define the set $\mathcal{A}(c_i)$ that contains the concept c_i and all the superconcepts (*i.e.*, ancestors) of c_i in a given taxonomy as:

$$\mathcal{A}(c_i)=\{c_j \in C \mid c_j \text{ is superconcept of } c_i\} \cup \{c_i\} \quad (1)$$

Let us represent the set of superconcepts $\mathcal{A}(c_i)$ by a binary vector $x_i = (x_{i1} \dots x_{in})$, being n the number of concepts of the ontology. Each element x_{ik} represents the existence of an is-a relation (considering its transitiveness) between c_i and c_k , $k = 1..n$, such as:

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

This vector provides a simple representation of a concept and its links in a given ontology and enables an easy analysis of the relation between a pair of concepts c_i and c_j , since it allows comparing all the shared and non-shared superconcepts of these concepts (not only the ones in the closest path).

Having the superconcepts represented in a vectorial form, one can define the distance between the concepts in terms of those vectors.

$$d(x_i,x_j) = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \quad (2)$$

Notice that this definition has a very clear interpretation in an algebraic way. 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 vice versa). Therefore, this expression is, in fact, equal to the number of non-shared superconcepts between c_i and c_j.

Based on this interpretation, the measure can be rewritten in terms of the set of superconcepts of $\mathcal{A}(c_i)$, providing a more compact expression, and more efficient to evaluate in the scope of ontologies with thousands of concepts.

$$d(c_i,c_j) = |\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)| \quad (3)$$

By considering concepts themselves in conjunction with the set of non-common superconcepts we are able to distinguish a pair of concepts that are siblings of an immediate superclass (*i.e.*, they are siblings and

share their complete sets of superconcepts) from identical concepts (for which the distance will be minimum).

Notice that as it is defined now the distance only considers the non-common knowledge of the two concepts. However, we are not able to distinguish concepts with very few or even no superconcepts in common from others with more shared knowledge. For example, as shown in Figure , the number of non-common superconcepts for the pair (c_1, c_2) and for the concepts (c_3, c_4) is equal, resulting in the same distance.

$$d(c_1, c_2) = |\mathcal{A}(c_1) \cup \mathcal{A}(c_2)| - |\mathcal{A}(c_1) \cap \mathcal{A}(c_2)| = 4 - 2 = 2$$

$$d(c_3, c_4) = |\mathcal{A}(c_3) \cup \mathcal{A}(c_4)| - |\mathcal{A}(c_3) \cap \mathcal{A}(c_4)| = 3 - 1 = 2$$

However, it makes sense that the distance between c_1 and c_2 is lower than the distance between c_3 and c_4 due to the higher amount of shared superconcepts of the pair (c_1,c_2). This is also related to the assumption formulated by some authors (Wu and Palmer 1994) who consider that pairs of concepts belonging to an upper level of the taxonomy (*i.e.*, they share few superconcepts) should be less similar than those in a lower level (i.e. they have more superconcepts in common). In order to take into account the amount of common information between a pair of concepts, we define our measure as the ratio between the amount of non-shared knowledge and the sum of shared and non-shared knowledge.

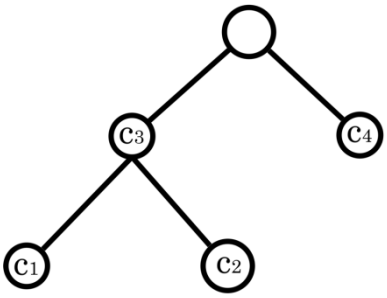

Figure 2. Taxonomy example

Definition 1: SuperConcept-based distance (SC)

$$SC(c_i, c_j) = \frac{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)|}{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)|} \qquad (4)$$

As a result, this definition introduces a desired penalization to those cases in which the number of shared superconcepts is small. Using the previous example, now the distance between concepts has changed to a better approximation of the real situation. The result is smaller as bigger is the common information, and vice versa.

$$SC(c_1, c_2) = \frac{4-2}{4} = \frac{2}{4} = 0.5$$

$$SC(c_3, c_4) = \frac{3-1}{3} = \frac{2}{3} = 0.66$$

On top of this definition of distance, we have developed two versions that introduce some additional properties.

- **Euclidean SC**

Considering that the vectorial representation of the concepts initially considered define an Euclidean space, it seems natural to define a measure of comparison as the Euclidean distance between their associated vectors x_i and x_j as:

$$d(c_i,c_j) = d(x_i,x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \qquad (5)$$

Rewriting the expression (35) in terms of sets and making the same normalization explained before, we have defined the Euclidean superconcept based distance as follows (Batet, Sánchez et al. 2009; Batet, Valls et al. 2010a).

Definition 2: Euclidian SuperConcept-based distance (SC_{Eu})

$$SC_{Eu}(c_i,c_j) = \sqrt{\frac{|{}^{\circ}\mathcal{A}(c_i) \cup {}^{\circ}\mathcal{A}(c_j)| - |{}^{\circ}\mathcal{A}(c_i) \cap {}^{\circ}\mathcal{A}(c_j)|}{|{}^{\circ}\mathcal{A}(c_i) \cup {}^{\circ}\mathcal{A}(c_j)|}} \qquad (6)$$

It is worth to note that it (6) fulfils the properties of a metric (*i.e.*, positivity, symmetry and triangular inequality) because it is no more than the well-known Euclidean distance. However, after the normalization process, these properties have to be analyzed, because we introduce a divisor that depends on the concepts compared. These properties have been studied considering that the concepts are obtained from an ontology, having some taxonomical relation. In fact, we have shown that there is a relation between the sets of shared and non-shared superconcepts that permits to prove the triangular inequality (Batet, Valls et al. 2010a).

- **Logarithmic SC**

Reinterpreting the distance proposed in (35) in terms of information theory, one can see the amount of shared and non-shared superconcepts as a measure of shared and non-shared information between concepts. In this context, a non-linear approach is considered the optimum for evaluating semantic features (Al-Mubaid and Nguyen 2006). So, we have introduced the inverted logarithm function, transforming the measures into a similarity (Batet, Sánchez et al. 2010; Batet, Sanchez et al. 2010b).

Definition 3: Logarithmic SuperConcept-based similarity (SC_{log})

$$SC_{log}(c_i,c_j) = -log_2 \frac{|{}^{\circ}\mathcal{A}(c_i) \cup {}^{\circ}\mathcal{A}(c_j)| - |{}^{\circ}\mathcal{A}(c_i) \cap {}^{\circ}\mathcal{A}(c_j)|}{|{}^{\circ}\mathcal{A}(c_i) \cup {}^{\circ}\mathcal{A}(c_j)|} \qquad (7)$$

Summarizing, our approach makes a more extensive use of the taxonomical information provided by the ontology because it takes into account all the available knowledge given by the different paths that connect the two concepts. We assume that, in this way, we will be able to improve the accuracy of the estimations by better capturing what is explicitly modelled in the ontology.

Computationally, our measure retains the simplicity of most path-based approaches, being much simpler than the calculus needed to estimate the information distribution in a corpus or to pre-process it.

2.1.1 Dealing with polysemic terms

The case of polysemic words has been also studied. Frequently, polysemic terms can be found in the databases that are going to be analysed with clustering techniques. This is due to the fact that the words available are not linked to concepts univocally. In fact, semantic similarity has been developed for any other kind of applications dealing with text, where the ambiguity of the words is also a matter of study (Budanitsky and Hirst 2006). For general ontologies such as WordNet, polysemic words correspond to several concepts (*i.e.*, one per word sense) which can be found by mapping words to concept synsets. A proper disambiguation of input terms may solve the ambiguity, assigning input words to unique ontological concepts. If the manual disambiguation is not possible, several pairs of concepts may be retrieved for a pair of polysemic words (in WordNet 2, polysemic nouns correspond to an average of 2.77 concepts[1]).

In previous works, polysemic words are tackled by retrieving all possible concepts corresponding to a term. Then, the similarity for each possible pair of concepts is computed and, as the final result, the maximum similarity value obtained is selected. The rationale for this criterion is that in order to evaluate the similarity between two non-disambiguated words (*i.e.*, no context is available), human subjects would pay more attention to their similarities (*i.e.*, most related senses) rather than their differences, as it has been demonstrated in psychological studies (Tversky 1977). Therefore, we have taken the same approach to solve this problem, taking the maximum similarity value obtained for all the possible combinations.

Definition 4: The generalized distance measure which is able to deal with polysemic terms is defined as:

$$SC_{generalized}(a,b) = \min_{\substack{\forall a' \in A \\ \forall b' \in B}} SC(a',b') \qquad (8)$$

where A is the set of concepts (*i.e.*, word senses) for the term a, and equally for term b. The same expression can be applied to the Euclidian and logarithmic versions of the measures (6) (7).

2.2 Contextualizing IC computation with ontologies

Measures based on the IC of concepts require tagged corpora in order to provide accurate assessments. The fact that available corpora typically consist of unstructured or slightly structured natural-language text implies that a certain degree of pre-processing is needed to extract implicit semantic evidence and to provide accurate results. In general, the more the pre-processing of the corpus is performed (in order to reduce noise or language ambiguity), the more accurate the results can potentially be. In fact, the size of corpora needed to provide good assessments is so big (millions of words) that their pre-processing introduces a serious computational burden. Therefore, even though a corpus-based approach may lead to accurate results, their dependency on data availability, suitability and pre-processing usually hampers their applicability.

To overcome these limitations, a completely unprocessed and massive corpus as the Web will be exploited to assess reliable estimations of concept appearance probabilities. In this manner we aim to minimize data sparseness observed for related works relying on high quality but reduced corpora. To solve the problems detected, we proposed a new way of computing IC from the Web, taking in a taxonomically coherent

[1] http://wordnet.princeton.edu/wordnet/man2.1/wnstats.7WN.html

manner (*i.e.*, monotonically increasing as concepts are specialized) and minimizing language ambiguity. This approach is based on the taxonomical knowledge given by ontologies.

The definition and evaluation is published in (Sánchez, Batet et al. 2009).

2.2.1 Web-based Information Content

We propose to estimate the IC of a concept from the Web with the ratio presented in Definition 5 (Sánchez, Batet et al. 2009; Sánchez, Batet et al. 2010a).

Definition 5: *Web-based Information Content* (*IC_IR*) of a concept '*a*' is defined as:

$$IC\_IR(a) = -\log_2 p_{web}(a) = -\log_2 \frac{H(a)}{M} \qquad (9)$$

where $p_{web}(a)$ is the probability of appearance of word '*a*' in a web resource. This probability is estimated from the Web hit counts returned by Web IR tool (denoted as H) when querying the term '*a*'. M is the total number of resources indexed by a Web search engine.

In this manner, the classical IC-based measures presented in deliverable D3 can be directly rewritten by incorporating the Web-based IC computation (IC_IR). Three examples are the following:

- Resnik (Resnik 1995) measure can be rewritten as follows:

$$sim_{res}\_IR(a,b) = IC\_IR\left(LCS(a,b)\right) = -\log \frac{H\left(LCS(a,b)\right)}{M} \qquad (10)$$

- Lin measure (Lin 1998) can be rewritten as follows:

$$sim_{lin}\_IR(a,b) = \frac{2 \times sim_{res}\_IR(a,b)}{\left(IC\_IR(a) + IC\_IR(b)\right)} = \frac{2 \times \left(-\log \frac{H\left(LCS(a,b)\right)}{M}\right)}{\left(-\log \frac{H(a)}{M} - \log \frac{H(b)}{M}\right)} \qquad (11)$$

- Jiang & Conrath distance measure (Jiang, Conrath 1997) can be rewritten as follows:

$$dis_{jcn}\_IR(a,b) = \left(IC\_IR(a) + IC\_IR(b)\right) - 2 \times sim_{res}\_IR(a,b) =$$
$$= \left(-\log \frac{H(a)}{M} - \log \frac{H(b)}{M}\right) - 2x\left(-\log \frac{H\left(LSC(a,b)\right)}{M}\right) \qquad (12)$$

2.2.2 Contextualized Information Content from the Web

After a large analysis of the results obtained with Web-based IC, we detected some weak points. Mainly, estimating *concept* probabilities from absolute *term* web hit counts without further manual processing can lead to very inaccurate results. Several issues related with language that affect to this estimation can be identified:

1) Absolute word usage in a corpus is a poor estimation of concept probability. This may lead to incoherent computation of the IC of the concept with respect to the underlying subsumption hierarchy. For example, the word *mammal*, as a subsumer of *dog* is much less frequent than the later in a general corpus like the Web. This may affect the monotony of the IC_IR associated to the taxonomy. This is usually solved in Resnik-based similarity measures by computing all individual occurrences of each concept and adding it to its subsumers. However, implementing this solution for the Web will lead to an enormous amount of web queries to recursively compute occurrences of all the concept's specializations, as well as, a heavy dependence on the corpus and ontology (re-computation will be needed to keep results up-to-date).
2) Language ambiguity may cause different problems: on one hand, different synonyms or lexicalizations of the same concept may result in different IC_IR values (e.g. *dog* is much more frequent than *canis*[2]), introducing bias. On the other hand, the same term may have different senses and, in consequence, correspondences to several concepts. In that last case, the computed term IC_IR will be the sum of IC_IR for all the associated concepts (e.g. IC of *dog* computed from a corpus includes appearances referring to a *mammal* and a *hot dog*, among other possible senses). In classical approaches (Resnik 1995; Hotho, Maedche et al. 2002) those problems were omitted by using a corpus tagged at a concept level based on WordNet *synsets*, rather than a word-level analysis. Therefore, they avoid potentially spurious results when only term (not concept) frequencies are used (Resnik 1995). In a more general approach, where the IC of a concept is computed from estimated term occurrences in the Web, ambiguity may cause inconsistencies if the context is not taken into consideration.

In (Sánchez, Batet et al. 2010b) several detailed examples of these situations were reported.

In order to avoid these incorrect results, we have redefined the way in which concept probabilities for IC computation are estimated from the Web. in this section we present a new way to coherently compute concept's IC from word's Web hit counts for similarity assessment using a reduced number of queries (Sánchez, Batet et al. 2010b).

We propose to compute the probability of appearance of a concept from the Web hit counts in a scalable manner, by contextualizing the concept appearances in the scope of its subsumer. The hypothesis is that the hit count of an explicit query of the co-occurrence of the concept and its subsumer provides a better estimation of the probability of appearance in the Web than the query of the concept alone. This relies on the fact that the explicit appearance of a concept's subsumer in the same context as the concept is considered as an explicit evidence of the correct word sense, aiding to minimize language ambiguity.

From the technical point of view, Web search engines natively support word co-occurrences from especially formulated queries (using logic operators such as AND or +). Using this feature, we force the co-occurrence between the subsumer (e.g. *mammal*) and each of the subsumed terms (e.g. *dog*) in the web query

[2] Occurrence of the word *dog* is 201 millions, while *canis* is 2 millions, computed from Bing (Nov. 9th, 2008)

ensuring that the IC_IR of the subsumed term (computed as *H(dog AND mammal)*) is higher than its subsumer (computed as *H(mammal)*). It is important to note that, in the case in which a concept is represented by several words (e.g. *persian cat*), double quotes should be used to maintain the context.

In addition, contextualizing the search aids to minimize ambiguity of absolute word appearance counts. For example, computing the occurrence of the term *dog* (referred as an *mammal*) in a corpus may give an idea of the word's appearance probability considering all its possible senses (i.e. associated concepts like *animal*, but also *fast food*); however, forcing the occurrence of *dog* and *mammal* (being *mammal* the LCS of *dog* and another concept such as *cat*) will introduce additional contextual information about the preferred word sense. Obviously, this implies a reduction of the corpus evaluated for the statistical assessment (i.e. only explicit co-occurrences are considered) and a subestimation of the real concept probability. Certainly, there will be many documents referring to the concept of *dog as a mammal* which will not explicitly include the word *mammal* in the text. However we hypothesize that, on one hand, considering the enormous size of the Web, data sparseness problems are minimized (Brill 2003). On the other hand, from the similarity computation point of view, the comparison of subestimated probabilities of the concepts will lead to more accurate assessments than probabilities based on absolute word occurrences.

Notice that using this approach we implicitly consider that each document of a corpus is typically using each word (which represents a web *hit* in a search engine) unambiguously. Disambiguation of term appearances at a document level is based on the observation that words tend to exhibit only one sense in a given discourse or document (context). This fact was tested by (Yarowsky 1995) on a large corpus (37.232 examples), obtaining a very high precision (around 99%).

From the similarity computation point of view, we propose that the subsumer used to contextualize web queries is the LCS of the pair of evaluated concepts in a given taxonomy. In this manner, we define the *Web-based Contextualized Information Content (CIC_IR)* for a pair of concepts as follows (Sánchez, Batet et al. 2010b):

Definition 6: For any pair of concepts *a* and *b* contained in a taxonomy *T*, the *Web-based Contextualized Information Content (CIC_T_IR)* of *a* with respect to *b* is:

$$CIC_T\_IR(a_b) = -\log_2 p_{web}(a_b) = -\log_2 \frac{H(a\ AND\ LCS_T(a,b))}{M} \qquad (13)$$

where $p_{web}(a_b)$ is the subestimated probability of concept *a* in the Web when computing its similarity against *b*. The least common subsumer, $LCS_T(a,b)$, is obtained from the taxonomy *T* which contains *a* and *b*. Then, the probability is computed from the Web hit counts returned by a search engine when querying the terms *a* and $LCS_T(a,b)$ at the same time (using *AND* or '+' logic operators). *M* is the total number of resources indexed by the web search engine.

Equally, for *b* with respect to *a:*

$$CIC_T\_IR(b_a) = -\log_2 p_{web}(b_a) = -\log_2 \frac{H(b\ AND\ LCS_T(a,b))}{M} \qquad (14)$$

As stated above, this is a subestimation of concept's probability. Note that the presented formula is different to the conditioned probability of the term with respect to the *LCS* (i.e. *p(a|LCS(a,b)) = hits(a AND LCS(a,b))/hits(LCS(a,b))*). The conditioned probability calculation, due to the denominator, will introduce the recursive problem of LCS concept probability estimation from absolute word hit counts, which we try to avoid.

With the proposed approach there is a relation between the original IC_IR expression and the contextualized version defined above.

Proposition 1. The IC_IR of the subsumer is always inferior to the CIC_T_IR of its subsumed terms.

$$IC\_IR(LCS(a,b)) \leq \min(CIC_T\_IR(a_b), CIC_T\_IR(b_a)) \quad (15)$$

This guarantees that the subsumer will be more general -less informative- than its specializations, because the IC of the specializations are computed in the context of the documents covering the subsumer. In consequence, from the similarity computation point of view, IC values will be *taxonomically coherent*.

It is important to note that, with this method, only one web query is needed to estimate the IC of each evaluated concept. So, the cost for a given pair of concepts with one LCS in common is constant. In addition, modifications in the taxonomy, which may affect Resnik-like IC computation (like adding a new sibling to the taxonomic specialization of a given subsumer), does not influence the calculation of CIC_T_IR. In consequence, our approach is more scalable and more independent to changes in the knowledge base.

Two examples of contextualized semantic similarity measures are given below.

- The Web-based contextualized version of the Lin similarity measure ($sim_{lin}\_CIC_T\_IR$) between concepts *a* and *b* contained in the taxonomy *T* is defined as follows:

$$sim_{lin\_CIC_T\_IR}(a,b) = \frac{2 \times IC\_IR(LCS_T(a,b))}{(CIC_T\_IR(a_b) + CIC_T\_IR(b_a))} = \frac{2 \times \left(-\log_2 \frac{H(LCS_T(a,b))}{M} \right)}{\left(-\log_2 \frac{Hs(a\ AND\ LCS_T(a,b))}{M} - \log_2 \frac{H(b\ AND\ LCS_T(a,b))}{Ms} \right)} \quad (16)$$

- The Web-based contextualized version of the Jiang & Conrath measure ($dis_{lin}\_CIC_T\_IR$) for concepts *a* and *b* contained in the taxonomy *T* is defined as follows:

$$dis_{jcn\_CIC_T\_IR}(a,b) = (CIC_T\_IR(a_b) + CIC_T\_IR(b_a)) - 2 \times IC\_IR(LCS_T((a,b))) = \left(-\log_2 \frac{H(a\ AND\ LCS_T((a,b)))}{M} - \log_2 \frac{H(b\ AND\ LCS_T((a,b)))}{M} \right) - 2 \times \left(-\log_2 \frac{H(LCS_T((a,b)))}{M} \right) \quad (17)$$

The evaluation section shows that the performance of both measures is greatly improved by the inclusion of CIC_T_IR because, even though concept probabilities have been subestimated, they are based in less ambiguous Web occurrences (Batet 2011).

2.2.3 Dealing with polysemy and synonymy

IC-based measures tackle polysemy by in the same manner as edge-counting approaches. For polysemic cases the strategy will be the same as presented before: all the LCSs available through the several taxonomic paths are retrieved, the similarity measure is computed for each of them and highest value (or lowest for dissimilarity) is taken.

In the case of synonyms (i.e. different textual forms are available for the same concept) one may consider to add the hit counts for the queries constructed with the available LCS synonyms. For example, being *dog* and *canis* synonyms of the subsumer of *terrier*, we can compute *H(terrier AND dog NOT canis) + H(terrier AND canis NOT dog) + H(terrier AND canis AND dog)*. However, in cases with a large set of synonyms (which is common in WordNet), a large amount of queries are needed, because they must include all the possible synonym combinations, as well as, a considerable number of keywords (resulting in a query which length may be not supported by typical web search engines). In addition, the final value will accumulate a considerable error derived from the individual errors inherent to the *estimated* hit counts provided by the search engine. Finally, this will make the similarity results dependant on the synonym coverage of each concept. Instead, we opted to consider each LCS synset synonym individually, computing the similarity value for each one and taking as a result the highest one (the lowest for dissimilarity measures). In this way, the LCS would correspond to the word that best contextualizes the queries (i.e. the less ambiguous textual form). During the research, we observed that this strategy leads to more accurate results than considering the sum of synonyms hit counts.

2.3 Analyzing collocation measures

As stated in the introduction, there exist other measures which seek for the co-occurrence between terms in order to estimate their correlation. In this case, they are completely unsupervised, as no background knowledge (a part from an unprocessed corpus) is employed. Those measures have been applied in similarity estimation based on the relation that exists between term co-occurrence in a corpus and their similarity (Spence and Owens 1990).

In order to statistically assess the degree of correlation and, as stated above, the similarity between words, standard collocation functions have been proposed. Formally, they are defined in the following way:

$$c_k(a,b) = \frac{p(ab)^k}{p(a)p(b)} \qquad (18)$$

, being *p(a)* the probability that the word *a* occurring within the text and *p(ab)* the probability of co-occurrence of words *a* and *b*. Here, the collocation of *a* and *b* is defined as the comparison between the probability of observing *a* and *b* together with respect to observing them independently. If *a* and *b* are statistically independent, the probability that they co-occur is given by the product *p(a)p(b)*. If they are not independent, and they have a tendency to co-occur in a corpus, *p(ab)* will be greater than *p(a)p(b)*. Therefore the ratio between *p(ab)* and *p(a)p(b)* is a measure of the degree of statistical dependence between *a* and *b* (Turney 2001).

The most typical forms of collocation functions are the *Symmetric Conditional Probability* (SCP), defined as c_2 (Ferreira da Silva and Lopes 1999) and the *Pointwise Mutual Information* (PMI), defined as $log_2 c_1$ (Church et al. 1991). In the latter case, the measure can be expressed in terms of the IC for *a* when we observe *b* and IC for *b* when we observe *a* (17).

$$PMI(a,b) = \log_2 \frac{p(ab)}{p(a)p(b)} = \big(IC(a) + IC(b)\big) - IC(ab) \qquad (19)$$

Considering the Web as a valuable corpus from which compute reliable statistics about information distribution, PMI was adapted by (Turney 2001) to approximate concept probabilities using the hit counts of a web search engine. The equation is specified as follows:

$$PMI\_IR(a,b) = \log_2 \frac{\frac{hits(a\ \ AND\ \ b)}{total\_webs}}{\frac{hits(a)}{total\_webs} \times \frac{hits(b)}{total\_webs}} \qquad (20)$$

However, from previous investigations (Downey et al. 2007), SCP have outperformed PMI by a large margin in the task of assessing similarity values for pairs of words, as it is less dependent on the order of magnitude of occurrence values. In the same manner as Turney (Turney, 2001), SCP can be adapted to compute concept probabilities from web hit counts (Downey et al. 2007).

$$SCP\_IR(a,b) = \frac{\left(\frac{hits(a\ \ AND\ \ b)}{total\_webs}\right)^2}{\frac{hits(a)}{total\_webs} \times \frac{hits(b)}{total\_webs}} = \frac{\left(hits(a\ \ AND\ \ b)\right)^2}{hits(a) \times hits(b)} \qquad (21)$$

Even though both measures have been applied to the task of evaluating concept relatedness (Downey et al. 2007; Etzioni et al. 2005), due to their lack of semantics, they offer a limited performance (Lemaire and Denhère 2006). This is caused by the inaccurate concept probability estimation from absolute word hit counts. In addition, decontextualized term co-occurrences in a document may be indicative of *relatedness* (Patwardhan and Pedersen 2006) but not necessarily of *semantic similarity* (Lemaire and Denhière 2006).

2.3.1 Contextualizing collocation measures

In order to overcome the presented problems of Web-based collocation measures due to their lack of semantics, we have proposed a solution consisting on estimating concept probabilities by means of the taxonomy information, using the CIC_T_IR measure. The idea is again to contextualize the queries using the LCS of the terms evaluated. In order to properly assess the concepts co-occurrence from the Web in a contextualized manner, we extend the CIC_T_IR definition (Definition 6) in the following way:

Definition 7. For any pair of concepts *a* and *b* contained in a taxonomy *T*, the *Web-based Contextualized Information Content (CIC_T_IR)* of the co-occurrence of *a and b* is:

$$CIC_T\_IR(ab) = -\log_2 p_{web}(ab) = -\log_2 \frac{hits(a\ AND\ b\ AND\ LCS_T(a,b))}{total\_webs} \qquad (22)$$

, where $p_{web}(ab)$ is the probability of co-occurrence of concepts *a*, *b*, estimated from the co-occurrence of words *a*, *b* and $LCS_T(a,b)$ in the Web, which is computed from the web hit counts of a web search engine.

This function permits to minimize term co-occurrence ambiguity, using the information of the taxonomic structure. In fact, co-occurrence will be biased by the ontological knowledge to the taxonomical side due to the additional semantics provided by the inclusion of the subsumer.

In order to introduce CIC_T_IR to collocation measures, we have rewritten the classical collocation definition in terms of IC by including the log_2 function. More details are given in (Sánchez, Batet et al. 2010b).

Definition 8. Given the concepts *a* and *b*, the collocation measure expressed in terms of concept's IC is defined as follows:

$$c_k\_IC(a,b) = \log_2 \frac{p(ab)^k}{p(a)p(b)} = \left(IC(a) + IC(b)\right) - k \times IC(ab) \qquad (23)$$

This new version can be directly contextualized by means of the CIC_T_IR calculation, which takes into account the information provided by the taxonomy *T* that includes *a* and *b*.

Definition 9. Given the concepts *a* and *b*, their *Web-based Contextualized Collocation measure* is defined as:

$$c_k\_CIC_T\_IR(a,b) = \left(CIC_T\_IR(a_b) + CIC_T\_IR(b_a)\right) - k \times CIC_T\_IR(ab) =$$
$$= \log_2 \frac{\left(\frac{hits(a \ \ AND \ \ b \ \ AND \ \ LCS_T(a,b))}{total\_webs}\right)^k}{\frac{hits(a \ \ AND \ \ LCS_T(a,b))}{total\_webs} \times \frac{hits(b \ \ AND \ \ LCS_T(a,b))}{total\_webs}} \qquad (24)$$

As explained, PMI and SCP are two common forms of c_k that are used for concept similarity estimation. Using the presented notation, c_1_ CIC_T_*IR* will correspond to *PMI*_CIC_T_*IR* and c_2_CIC_T_*IR* will correspond to *SCP*_CIC_T_*IR*. It is expected that the SCP will have better results also in this version as it has offered the best performance in its original form.

- PMI computation when introducing CIC_T_IR for concepts *a* and *b* in the taxonomy *T* is defined as follows:

$$PMI\_CIC_T\_IR(a,b) = \log_2 \frac{\frac{hits(a \ \ AND \ \ b \ \ AND \ \ LCS_T(a,b))}{total\_webs}}{\frac{hits(a \ \ AND \ \ LCS_T(a,b))}{total\_webs} \times \frac{hits(b \ \ AND \ \ LCS_T(a,b))}{total\_webs}} \qquad (25)$$

- IC-based SCP computation when introducing CIC_T_IR for concepts *a* and *b* in the taxonomy *T* is defined as follows:

$$SCP\_CIC_T\_IR(a,b) = \log_2 \frac{\left(hits(a \ \ AND \ \ b \ \ AND \ \ LCS_T(a,b))\right)^2}{hits(a \ \ AND \ \ LCS_T(a,b)) \times hits(b \ \ AND \ \ LCS_T(a,b))} \qquad (26)$$

Note that the *total_webs* constant is simplified as it is common to the numerator and denominator.

The mathematical operations of this similarity measure were studied in (Sánchez, Batet et al. 2010b), proving that it fulfills symmetry, maximality and positiveness.
Moreover, in the same paper, the case of polysemic and synonym terms was considered. We proposed a generalized version of the collocation-based functions for the case of multiple subsumers and synonyms available in the taxonomy *T*.

$$c_k\_CIC_T\_IR(a,b) = \max_{L \in S(a,b)} \left(\log_2 \frac{\left(\frac{hits(a \ \ AND \ \ b \ \ AND \ \ L)}{total\_webs}\right)^k}{\frac{hits(a \ \ AND \ \ L)}{total\_webs} \times \frac{hits(b \ \ AND \ \ L)}{total\_webs}} \right) \qquad (27)$$

,where *S(a,b)* is the set of textual form (synonyms) of all the LCS that subsume *a* and *b* in the given taxonomy *T*.

2.4 Tests

The evaluation of the semantic similarity measures explained in this deliverable was done following the same procedure explained in Deliverable D3.

It consists on using data benchmarks consisting of word pairs whose similarity was assessed by a set of humans. The goal is to obtain a high correlation between the similarity values given by the humans and the estimated similarity calculated with some of the measures proposed.

If the two rating sets are exactly the same, the correlation coefficient is 1, whereas 0 means that there is no relation. Spearman's and Pearson's correlations coefficients have been commonly used in the literature; both are equivalent if the ratings sets are ordered (which is the case). They are also invariant to linear transformations which may be performed over the results such as a change between distance and similarity by changing the sign of the value or normalizing values in a range.

We have performed test with general purpose benchmarks such as the one defined by Rubenstein and Goodenough (Rubenstein and Goodenough, 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 (Miller and Charles, 1991) re-created the experiment in 1991 by taking a subset of 30 noun pairs whose similarity was reassessed by 38 undergraduate students. Resnik (Resnik, 1995) replicated again the same experiment in 1995, in this case, requesting 10 computer science graduate students and post-doc researchers to assess similarity. Finally, Pirro (Pirró, 2009) replicated and compared the three above experiments in 2009, involving 101 human subjects, both English and non-English native speakers.

We have taken the correlation values reported by related works for Rubenstein and Goodenough's and Miller and Charles' benchmarks in order to compare our results. The support ontology is WordNet.

WordNet (Fellbaum 1998) is a domain-independent and general purpose ontology/thesaurus that describes and organizes more than 100,000 general English concepts, which are semantically structured in an ontological fashion. It contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (*synsets*), each expressing a distinct concept (*i.e.*, a word sense). Synsets are linked by means of conceptual-semantic and lexical relations such as synonymy, hypernymy (is-a), six types of meronymy (part-of), antonymy, complementary, etc. The backbone of the network of words is the subsumption hierarchy which accounts more than an 80% of all the modelled semantic links, with a maximum depth of 16 nodes. The result is a network of meaningfully related words, where the graph model can be exploited to interpret the meaning of the concept.

WordNet 2 is the most common version used in related works. In cases in which the original authors used an older version (WordNet 2 was released in July 2003), we took a more recent replication of the measure evaluation performed by another author in order to enable a fair comparison. As a result, we picked up results reported by authors in papers published from 2004 to 2009.

We have also performed some tests on biomedical benchmarks. In the last few years, the amount of clinical data that is electronically available has increased rapidly. Digitized patient health records and the vast amount of medical and scientific documents in digital libraries have become valuable resources for research. However, most of these information sources are presented in unprocessed and heterogeneous textual formats. Semantic technologies play an important role in this context enabling a proper interpretation of this information.

Pedersen *et al.* (Pedersen, Pakhomov et al. 2007) stated the necessity of having objectively scored datasets that could be used as a direct means of evaluation in the biomedical domain. Thus, they created, in collaboration with Mayo Clinic experts, a benchmark referring to medical disorders. The similarity between

term pairs was assessed by a set of 9 medical coders who were aware of the notion of semantic similarity and a group of 3 physicians who were experts in the area of rheumatology. After a normalization process, a final set of 30 word pairs were rated with the average of the similarity values provided by the experts in a scale between 1 and 4. The correlation between the physicians was 0.68, whereas the correlation between medical coders achieved a value of 0.78.

Pedersen *et al.* (Pedersen, Pakhomov et al. 2007) used that benchmark to evaluate most of the measures based on path length and information content, and their own context vector measure, by exploiting SNOMED CT as the domain ontology and the Mayo Clinical Corpus and Thesaurus as corpora. Al-Mubaid and Nguyen (Al-Mubaid and Nguyen 2006) also used that benchmark and SNOMED CT to evaluate path-based measures considered in this document.

SNOMED CT (*Systematized Nomenclature of Medicine, Clinical Terms*) is an ontological/terminological resource distributed as part of the UMLS (*Unified Medical Language System*) of the US National Library of Medicine. It is used for indexing electronic medical records, ICU monitoring, clinical decision support, medical research studies, clinical trials, computerized physician order entry, disease surveillance, images indexation and consumer health information services. It contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into 18 overlapping hierarchies: clinical findings, procedures, observable entities, body structures, organisms, substances, pharmaceutical products, specimens, physical forces, physical objects, events, geographical environments, social contexts, linkage concepts, qualifier values, special concepts, record artifacts, and staging and scales. Each concept may belong to one or more of these hierarchies by multiple inheritance (e.g. euthanasia is an event and a procedure). Concepts are linked with approximately 1.36 million relationships. In such a complete domain ontology, *is-a* relationships have been exploited to estimate term similarity, even though much of the taxonomical knowledge explicitly modelled is still unexploited. The SNOMED CT-ontology, has proven to have very good concept coverage of biomedical terms (Lieberman, Ricciardi et al. 2003; Penz, Brown et al. 2004; Spackman 2004) and it has been adopted as reference terminology by some countries (e.g. UK, USA, Spain), and some organizations (e.g. UK's National Health Services, ASTM International's Committee E31 on Healthcare Informatics, Federal Drug Administration) (Cornet and Keizer 2008).

2.5 Discussion

All the semantic similarity measures proposed have been compared to related works as explained in the previous section. Here we report briefly the results. More details can be found in the indicated publications.

2.5.1 Superconcept-based distance

Compared to other ontology-based measures, it is interesting to note that our approach's accuracy surpasses the basic edge-counting approaches. In general, in complex and detailed ontologies like WordNet, where multiple taxonomical paths can be found connecting concept pairs (overlapping hierarchies), path-based measures waste explicitly available taxonomical knowledge as only the minimum path is considered as an indication of distance. Only the Li *et al.*,'s measure is able to achieve a very similar accuracy when the appropriate scaling parameters are empirically chosen. Feature-based approaches' correlations are also surpassed, even though they are based on other non-taxonomical features and weighting parameters. This shows that taxonomical knowledge plays a more relevant role in stating term similarity than other more scarce features which are typically poorly considered in available ontologies.

The same situation is repeated for corpus-based IC measures, showing that the exploitation of high quality taxonomical knowledge available in ontologies provides even more reliable semantic evidences than un-

structured textual resources. This is coherent to what it is observed for approaches computing IC in an intrinsic manner, which, conceptually, follow a very similar principle as our approach. In their case, similarity is computed as a function on the number of hyponyms whereas in our case it is estimated as a function of overlapping and non-overlapping hypernyms. The only case in which they surpass our measure's correlation is when IC is computed as Zhou *et al.*'s (Zhou, Wang et al. 2008), in which a weighting parameter is introduced to optimize the assessment.

It is worth to note that in the test with biomedical benchmarks, context vector measures significantly surpass the SC method. High correlations are obtained when a huge amount of data (1 million clinical notes) is used to create the vectors. One can see how the accuracy of the measure decreases when a narrower corpus is used. This dependency on the corpus size implies that the amount of processing needed to create the vectors from such an amount of data is not negligible. Moreover, the highest correlation is only obtained when using a particular subset of data, which corresponds to the descriptions of diagnostics and treatments.

From the runtime point of view, in our approach only the set of subsumers (with dozens instead of thousands of elements) must be compiled. As a result, our approach is more efficient, general and easily applicable for different domains and ontologies, and it does not need training.

Summarizing, our approach is able to provide a high accuracy without any dependency on data availability, data pre-processing or tuning parameters for a concrete scenario. As it only relies on the most commonly available ontological feature (is-a), our measure ensures its generality as a domain-independent proposal. At the same time, it retains the low computation complexity and lack of constraints of edge-counting measures as it only requires retrieving, comparing and counting ontological subsumers. This ensures its scalability when it must be used in data mining applications, which may require dealing with large sets of terms (Batet, Valls et al. 2008; Armengol 2009).

Compared to other approaches based on taxonomical knowledge, the exploitation of the whole amount of unique and shared subsumers seems to give solid semantic evidence of semantic resemblance. First, the distinctive features implicitly include information about the different paths connecting the pair of terms. In the same manner, the depth of the Least Common Subsumers of those concepts is implicitly included in the set of shared subsumers (*i.e.*, the deeper the LCS, the higher the amount of common features). Other features that have been identified in the literature, such as relative taxonomical densities and branching factors, are also implicitly considered, being all of them useful dimensions to assess semantic similarity.

As any other ontology-based measure, the final accuracy will depend on the detail, completeness and coherency of taxonomical knowledge. Moreover, most of the improvements achieved by our approach are derived from the fact that similarity is estimated from the total set of subsumers considering the different taxonomical hierarchies. Due to the definition of the measure as a ratio, we can use this measure also in small domain-specific or even application ontologies built for a very specific problem.

2.5.2 Contextualized information content

In our experiments, the results of the proposed modifications to IC-based similarity measures ($sim_{lin}\_CIC_T\_IR$, $dist_{jcn}\_CIC_T\_IR$,) have been compared against their original forms computed also from the Web (Sánchez, Batet et al. 2010a; Sánchez, Batet et al. 2010b). In all cases, we have used the Web hit counts to estimate probabilities and compute concept's IC. This compares the contextualized and non-contextualized web-based concept probability assessment. The performance of each measure is also evaluated by computing the correlation of the values obtained for each word pair against the human ratings. All the measures have been tested using the Web as corpus. We have also ensured the same conditions, executing the tests at the same moment (to minimize variance due to web-IR estimation changes).

Classical IC-based measures perform poorly when only absolute word occurrences are used to assess concept probabilities (i.e. no tagged corpus is available). The inclusion of the contextualized version of IC

computation in Lin and Jiang & Conrath improves the results, due to the additional context. As a result, they clearly outperform the basic versions, almost doubling the correlation value.

Although the contextualized approach obtains subestimations of the real observations, it has provided good results. This shows, on the one hand, that even reducing the size of the corpus, the Web provides enough resources to extract reliable conclusions. On the other hand, the calculated probabilities, even subestimated, lead to better similarity assessments due to the minimized ambiguity. It is important to note that this approach is able to provide taxonomically coherent IC estimations with a constant -low- number of web queries for non-polysemic ontologies. Resnik-like approaches would require and exponential amount of calculus according to concept's branching factor of specializations, hampering the scalability of the approach. For polysemic cases, the number of queries is linear to the number of LCS available for the pair of evaluated concepts.

2.5.3 Contextualized collocation measures

In our experiments, the results of the proposed modifications to Resnik-based and collocation-based similarity measures (sim_{lin}_CIC_T_IR, $dist_{jcn}$_CIC_T_IR, PMI_CIC_T_IR, $Dis_{SCP\_\ CICT\_IR}$) have been compared against their original forms. In all cases, we have used the Web hit counts to estimate probabilities and compute concept's IC.

Analyzing the values, in general, we can say that they tend to outperform IC-based measures when using the Web as a corpus. Considering that they do not require any background taxonomy, they are an effective unsupervised way to assess concept's relatedness. Under the same conditions, SCP outperforms PMI by a considerable margin. Introducing the contextualized taxonomy-based IC computation to collocation measures, we observe clear improvements. As stated before, the added knowledge biases the corpus statistical analysis towards the correct word sense and guides the occurrence analysis to the taxonomic side. As expected, SCP-based function outperforms again its PMI counterpart. This is an expected improvement obtained at the cost of requiring a background taxonomy. In a more fair comparison (as both measures exploit a taxonomy). However, SCP does not include the ambiguous estimation of LCS's IC, which results in a more accurate assessment. At the end, SCP_CIC_T_IR have been able to obtain a correlation which is only a 7% worse than the original Resnik-based measures. This shows the reliability of Web-based statistics when language ambiguity is tackled.

3 Semantic similarity measures into clustering algorithms

In Deliverable D3 a description and classification of clustering methods was presented. Two main families of methods are distinguished: hierarchical and partitional.

When the purpose of the clustering is mainly knowledge discovery (i.e. find new structures with a data mining process), hierarchical algorithms offer better features. On the one hand, hierarchical algorithms are more versatile than partitional algorithms. The hierarchical representation provides very informative descriptions and visualization for the potential data clustering structures. Moreover, the dendrogram obtained as a result of the clustering provides the inner structure of the data set. They are also non-order-dependent, which guarantees stability of results by permutations of the rows of data matrix. On the other hand, the hierarchical process requires a major time complexity, quadratic in most cases, what makes some of these algorithms prohibitive for massive data sets analysis. Another disadvantage is that they suffer from their inability to perform adjustments once the merging decision is made (i.e. once an object is assigned to a cluster, it will not be considered again), which means that hierarchical algorithms are not capable of correcting a possible previous misclassification.

When the purpose of the clustering is to find some partition of the objects regarding their similarities, partitional algorithms are used. The time and space complexities of the partitional algorithms are typically lower than those of the hierarchical algorithms (Day 1992). In particular, partitional methods have advantages in applications involving large data sets for which the construction of a dendrogram is computationally prohibitive (Jain, Murty et al. 1999) (Everitt, Landau et al. 2001). The main drawback of partitional algorithms is how to make the choice of the number of desired output clusters. In the case of hierarchical methods, they do not require the number of clusters to be known in advance as the final clustering results are obtained by cutting the dendrogram at different levels, and several criteria provide the best level where to cut. Partitional algorithms suffer from the problem of getting trapped in a local optimum and therefore being dependent on the initialization. Some approaches to find a global optimum introduce additional parameters that can be adjusted to the characteristics of the domain.

Both approaches (partitional and hierarchical) share some components that are the clue for merging or splitting the objects. The main one is the measure of the distance or dissimilarity between a pair of individuals.

In this project, a data matrix with different types of values will be considered (see Deliverable XX). We will consider numerical, nominal (i.e. non-ordered categorical values) and semantic features. Semantic features are an extension of categorical features, which have a non fixed and large set of possible values, without any order or scale of measurement defined between terms.

In Deliverable D3 the dissimilarity and distance functions for numerical and categorical data were presented. In this document, in section XX, new semantic similarity functions have been proposed.

To combine different types of attributes into a single measure we will use the approach of Compatibility Measures (see the introduction in Deliverable D3). It permits the combination of the contribution of numerical, nominal and semantic features into a global function. After the definition of this compatibility operation, any of the semantic similarity functions could be used to deal with the comparison of semantic values (i.e. terms corresponding to concepts). Some preliminary work in such a compatibility measure has already been proposed by the research team of this project in (Batet et al., 2008).

We propose the following expression for the compatibility measure:

$$d^2{}_{(\alpha,\beta,\gamma)}(i,i') = \alpha \sum_{k\in\varsigma} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \frac{\beta}{n_Q^2} \sum_{k\in Q} d_k^2(i,i') + \frac{\gamma}{n_S^2} \sum_{k\in S} ds_k^2(i,i') \qquad (28)$$

The first component corresponds to the Euclidean distance used for numerical data, the second component is the contribution of categorical feature according to the Chi-squared metrics. In particular, we have used a decomposition of the χ^2 metrics calculation prosed in (Gibert, Nonell et al. 2005). Finally, the third component $ds_k^2(i,i')$ corresponds to the contribution of semantic feature, which can be computed using any of the exiting measures presented before.

According to the principles of compatibility measures proposed by Anderberg (Anderberg 1973), the contribution of a single feature to the final distance is different depending on its type and it can be computed per blocks, regarding the types of the considered variables. So, this expression permits to associate a weight to each component, giving different importance to numerical, categorical and semantic attributes.

With respect to the semantic similarity measure, any of the existing functions can be incorporated into this compatibility measure. The advantages and drawback of each approach were discussed in Deliverable D3, and they have also been reviewed in section 2.

Path-based approaches are quite adequate because they only relies on taxonomic ontological knowledge, lacking of corpora-dependency or parameter-tuning, being also computationally efficient. In particular, the results reported by the Superconcept-based measure (*SC*) suggest a promising accuracy, improving the correlations reported by most of other ontology-based and corpora-based approaches, while minimizing the constraints that may hamper their applicability both from the computational efficiency and resource-dependency points of view.

After this study, the SC measure has been selected to make a first analysis of the behaviour of semantic similarities in clustering.

Comparing the two versions of *SC*, named SC_{log} and SC_{Eu}, we have observed that the one framed in the information theory, which uses the inverted logarithm to smooth the evaluation of common an non-common subsumers provides the best accuracy. However this functions is not a distance (i.e. it does not fulfil the properties of a metric: positivity, symmetry and triangular inequality. On the other hand, SC_{Eu} is a metric as proved in (Batet, Valls et al. 2010a). These properties have been studied considering that the concepts are obtained from an ontology, having some taxonomical relation. In fact, we have shown that there is a relation between the sets of shared and non-shared superconcepts that permits to prove the triangular inequality.

When this comparison between terms is done in the context of decision making, it is worth to know the metrical properties of the measure, because it may have implications on the results that will be obtained. For example, for the particular case of hierarchical clustering, it is interesting to maintain the ultrametric properties of the dendograms and the Huygens theorem of decomposition of inertia, which are directly related with the interpretability of the final results. If the comparison measure is a distance, these properties are guaranteed. However, in the field of computational linguistics, the comparison measures proposed usually do not fulfill the triangle inequality property, being only similarities but not distances. Thus, this fact should be taken into account in each application.

3.1 The clustering method

In the DAMASK project we will study the two main approaches to clustering, selecting one of their most representative algorithms.

3.1.1 K-means

The most important algorithm for partitional clustering is called *k-means*. It is present in most statistical software packages (see Deliverable D3). The *k-means* is the most well-known centroid algorithm (Forgy 1965; MacQueen 1967). The method attempts to find a number *k* of clusters fixed a priori, which are represented by its centroid. Method *k*-means uses the squared error criterion (MacQueen 1967) as criterion function. The steps of this clustering algorithm are the following:

```
A priori:  Determine the number K of partitions
Step 1) Select k seeds, one per class: randomly
        or based on some prior knowledge, and consider them as cluster cen-
        troids.
Step 2) Assign each object to the nearest cluster (i.e. the cluster proto-
        type or centroid) on the basis of the similarity function.
Step 3) Recalculate the centroid of each cluster based on the current par-
        tition.
Step 4) Repeat steps 2)-3) until there is no change for each cluster or
        when a number of beforehand defined iterations is done.
```

Notice that in Step 2, the compatibility function proposed before can be used to find the nearest cluster to any of the prototypes, which are represented as the objects (with numerical, categorical and semantic features).

In this algorithm a second component must be studied: the prototype construction. Prototypes or centroids are fictitious objects that represent the members of a cluster. In general, a centroid is defined as a typical example, basis, or standard for other elements of the same group. In practice, a centroid of a dataset is a vector that encodes, for each attribute, the most representative value of the objects found in the dataset. Methods for centroid construction are mainly devoted to numerical and categorical datasets, focusing on the numerical and distributional properties of data. Semantic attributes, on the contrary, consist on labels referring to concepts with a specific semantic content (i.e., meaning), which cannot be evaluated by means of classical numerical operators. Hence, the centroid of a cluster with semantic attributes should be the concept that best represents/preserves the semantics of the original terms. Semantically-grounded methods aiming to construct centroids for textual datasets are scarce, they are mainly limited to the use of the LCS (Least Common Subsummer) of the terms that is found in an ontology (Abril, et al., 2010, Erola, el al., 2010).

3.1.2 Ward's method

We have focused on agglomerative methods that have been more used and are present in most of the software packages for clustering (see Deliverable D3).

Agglomerative clustering starts with *n* clusters each of them including exactly one object and then consecutive merge operations are followed out to end-up with a single cluster including all individuals. This follows a bottom-up approach. The general agglomerative clustering can be summarized by the following procedure:

```
Step 1) Start with n singleton clusters and calculate the distances matrix between
        clusters.
Step 2) Search the pair of clusters that optimizes the aggregation criterion, which
        is a function of the distances matrix, and combine in a new cluster.
Step 3) Update the distances matrix by computing the distances between the new clus-
        ter and the remaining clusters to reflect this merge operation.
Step 4) Repeat steps 2)-3) until all objects are in the same cluster.
```

The aggregation criterion is a function of the distances matrix and the different aggregation criteria provide the different hierarchical clustering methods. The *Minimum-variance loss* or *Ward's method* (Ward 1963) has some mathematical properties that are interesting. This criterion consists of taking the clusters that minimize the loss in the inertia between classes and aggregate them. Since the inertia between classes is related with the information contents of the data set in the sense of Shanon theory (demonstrated by Benzecri (Benzecri 1973) this method obtains a more optimum partition of the objects. Moreover, when the comparison between objects is performed by means of a distance, the Huygens theorem of decomposition of inertia holds and a recursive expression can be used to calculate the loss in the between-classes inertia due to a certain merge between two clusters, on the basis of the inertia within those two clusters.

3.2 Experimentation

In order to study the improvement of semantic information in knowledge extraction, we have selected a hierarchical clustering algorithm. As explained, this approach exhibits some good properties in order to discover the hierarchical relations among objects. In order to facilitate the testing and analysis of the results, we have taken a software system named KLASS.

The semantic similarity measure SC and the compatibility measure proposed have been integrated into the KLASS software system (Gibert and Cortés 1998; Gibert, Nonell et al. 2005) with the collaboration of Dr. Karina Gibert, responsible of the KLASS software. In fact, the experimentation that will be explained in this section has been part of the Ph.D Thesis of Montserrat Batet, directed by Dr. Aïda Valls and Dr. Karina Gibert.

KLASS was specially designed for integral knowledge discovery from databases (KDD) by combining statistical and Artificial Intelligence tools. *KLASS* provides, among others, tools for descriptive data analysis, sequential data analysis, clustering, classes interpretation (in this thesis the most used one will be Class Panel Graph) and *reporting*, offering a friendly graphical interface. A Class Panel Graph is compact graphical displaying of the conditional distributions of the variables against the classes which evidences the particularities of classes and contributes effectively to quick understanding of the meaning of them (Gibert, Garcia-Rudolph et al. 2008). *KLASS* includes different clustering algorithms, as the reciprocal neighbours hierarchical algorithm with the Ward's criterion, which we have used for experimentation. It may graphically represent the resulting dendrogram and it can recommend the final number of classes using a heuristic based on Calinski-Harabaz criterion (Calisnski-Harabaz 1974). At the moment, this tool has been modified to include semantic features and obtain some first results to study the effect of this component in the quality of the clusters obtained.

As a first study, we have analysed the influence of including semantic information in clustering and the influence of using different semantic similarity measures in the results of clustering. The case of touristic city destinations has been considered, including 2 numerical, 2 categorical and 5 semantic features. We specially designed a data set with a balance between non-semantic and semantic information. Next section gives more details about this experiment.

This dataset of touristic destinations has been also used to perform an analysis of the influence of changing the semantic similarity measure. From the wide range of available semantic similarity approaches, we have centred our study on those measures that are based on the taxonomical exploitation of the ontology, as argued before.

The results have shown that those similarities that correlate better with human ratings in a standard benchmark, also provide more accurate and refined clusters (Batet, Valls et al. 2010b).This is an interesting result because it indicates that simple tests based on correlations between pairs of words can be performed to evaluate the similarity measures before incorporating them into a more complex and time-consuming clustering algorithm. For this reason, we have performed the following tests using the Superconcept-based distance, which obtained the highest correlations at an acceptable runtime.

In the next experiment, we tested our approach with real data using a dataset obtained from a survey done to the visitors of a Natural Protected Park. This data was given by our partners in the DAMASK project: the Parc Científic Tecnològic de Turisme i Oci (PCTTO).

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 that receives many tourists each year (about 300.000). It is considered a Bird Special Protection Area. The data was obtained with a questionnaire made to 975 visitors to Ebre Delta Natural Park between July and September 2004. The 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 analysis and results are available in some publications (Batet, Valls et al. 2010c; Batet, Gibert et al. 2011), as well as in the Ph.D. Thesis documentation (Batet, 2011).

The results show that a semantic clustering approach is able to provide a partition of objects that considers the meaning of the textual values and, thus, the result is more interpretable and permits to discover semantic relations between the objects. The method has produced a more equilibrated grouping that non-semantic approaches and provides useful knowledge about the characteristics of the visitors. In short, 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 neighbourhood that visits the Ebre Delta Natural Park for its natural interest. Moreover, the clusters are able to also identify 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 behaviour with respect to the Park.

3.3 Case study: tourist destinations

For this study we considered a reduced dataset of cities which are tourist destinations. WordNet has been used for the semantic similarity calculation. The data matrix contains 23 cities from all over the world. Each city is represented with a vector of 9 features extracted from Wikipedia: two numerical features (population and land area), two categorical features (continent and city ranking) and five qualitative features (country, language, geographical situation, major city interest and geographical interest).

The cities have been clustered under two different approaches:

- treating all the variables as simple categorical variables, as available before this research, and
- taking advantage of the additional semantic information about the features, so managing as semantic variables, using WordNet ontology for better treatment of semantic variables.

Differences in the results will be directly assignable to the benefit of using the generalized compatibility measure in the clustering process, i.e. to the fact that the semantics provided by a background ontology is taken into the account in the process.

Table 1. Feature values for the cities.

Feature	Values
City ranking	country capital, state capital, city or village
Geographical situation	valley, plain, island, coast, island, mountain range, mountain, lake, archipelago
Major city interest	cathedral, basilica, business, shopping center, government structure, office building, basilica, monument, historical site, church, mosque, recreational structure, ski resort, tourism, viewpoint, theatre
Geographical interest	river, coast, bay, lake, mountain, beach, volcano, cliff, crater, ocean
Continent	Asia, Africa, North America, South America Antarctica Europe, Australia
Country	France, Usa, Canada, Spain, Venezuela, Cuba, Andorra, Switzerland, Portugal, Italy Egypt, Australia
Language	French, English, Spanish, Catalan, Portuguese, Germany, Italian, Arabic
Population	1..10000000
Land area	0..5000

3.3.1 Clustering without considering semantic information

In this case, semantic features are treated as ordinary categorical features, which uses Chi-squared distance for categorical variables. So, the different features are considered as: numerical (population, land area), categorical (continent, city ranking, country, language, geographical situation, major city interest, major geographical interest).

Figure 3 shows the dendrogram resulting from clustering. 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 three 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. The following descriptions of the clusters can be inferred:

- Interlaken is the only city near a lake with a ski 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 million population. It is the state capital of New South Wales. It is situated near the coast and it is English-speaking. It has 5 theatres and the Sydney's Opera House.
- Class14 is composed by state (autonomous region) capitals from Portugal, they are located in islands or archipelagos. 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 elements either country capitals or villages from different countries and continents, with a wide diversity of cultural or geographical interests.

Although the results are quite coherent, 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 skiing, for example.

For better comparison with the results obtained when considering the ontological information, a cut in 5 classes has been also analyzed. In this case, classes contain cities very heterogeneous among them. As usual in real complex domains (Gibert, Garcia-Rudolph et al. 2008) there is a very big class of 15 cities quite heterogeneous which seems to share all type of cities (Batet, Valls et al. 2008). 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).

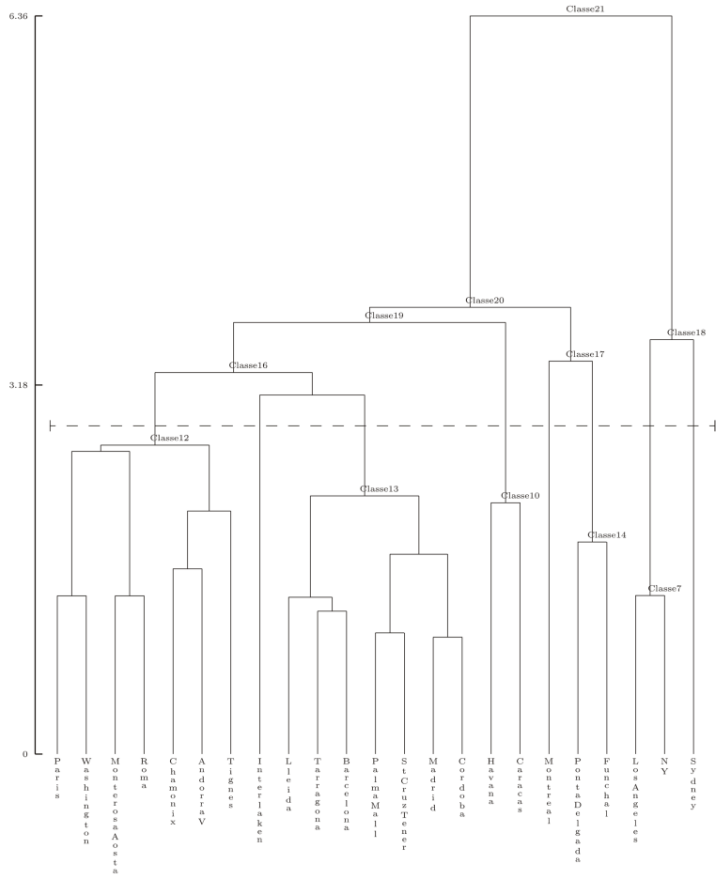

Figure 3. Dendogram without considering ontologies.

3.3.2 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. The proposal presented in section 2.1 has been used to compare de values of semantic features. Figure 4 shows the resulting dendrogram, quite different from Figure 3 and producing groups more equilibrated in size.

After studying the structure of the tree, a 5-classes cut is selected. In this case, the interpretation of clusters 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 not located in mountains, 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.

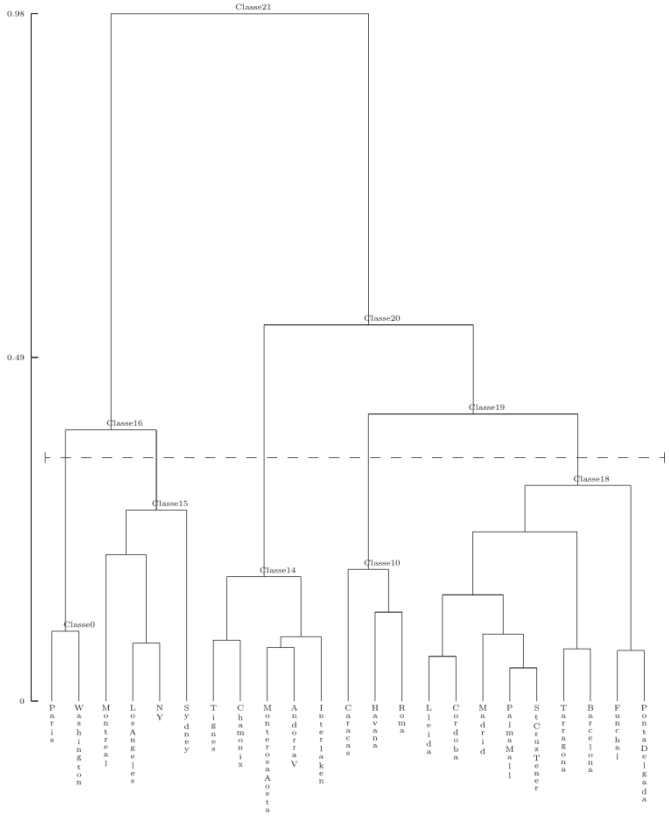

Figure 4. Dendrogram using ontologies.

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 reminds more to a multivariate treatment of the cities.

3.4 Discussion

In this section we have seen how semantic similarity measures can be integrated into the clustering algorithms by means of a compatibility measure that is capable of combining the contribution of a set of numerical, categorical and semantic features.

The case of the *Ward's clustering* method has been studied because it assures that the clusters obtained exhibit some interesting properties, which permits to obtain an optimum partition of the objects under some conditions. Moreover, the fact of building a hierarchical structure (i.e. dendogram) permits to have a visual representation of the relations between the objects at different levels, which is much more informative than having a single partition, like in the *k-means* algorithm.

The case of touristic city destinations has been considered, including 2 numerical, 2 categorical and 5 semantic features. This data set had a balance between non-semantic and semantic information. A first study let us know that those similarities that correlate better with human ratings in a standard benchmark, also provide more accurate and refined clusters. This is an interesting result because it indicates that simple tests based on correlations between pairs of words can be performed to evaluate the similarity measures before incorporating them into a more complex and time-consuming clustering algorithm.

A second study has been done comparing the results obtained when textual data is treated as categorical or as semantic data. Since the dataset is small, we can easily visualize the dendograms using KLASS software (as well as extract some statistics and other representation forms of the clusters). These results have been reported in this deliverable, illustrating that the clusters obtained when the data is interpreted with ontologies are much more clear and coherent than when only using the traditional categorical methods.

4 Final conclusions and remarks

From the research conducted in task 2-3, which includes the developed similarity measures and their application into a clustering method, we can extract the following conclusions:

- The knowledge representation provided by ontologies is a powerful tool to assess the semantic similarity between terms or concepts. Although there are different approaches, the results obtained with the ontology-based similarity measure presented in section 2.2 (called Superconcept-based distance, *SC*) show that it is able to improve most of related works when evaluated using standard benchmarks.
- The Web can be considered a valid corpus from which extract statistics of the real distribution of terms in the society. In particular, available information retrieval tools (Web search engines) can be exploited in order to compute the information content of words. The presented approach to compute the IC from the Web in a contextualized manner (section 2.3) permits to redefine classical measures based on the IC of terms in a way that they can be applied to the Web instead of domain corpora. This overcomes data sparseness problems caused by their reliance on –reduced- tagged corpora.
- The use of the Web to contextualize collocation measures (section 2.4) also leads to better results than the related works.
- From the different proposals presented, we have selected the *SC* similarity measure to be used in the clustering process. The reasons are: its high accuracy, stability, its low computational cost, the fact that it does not depend on tuning parameters and the availability of large and detailed general-purpose and domain-specific ontologies.
- Semantic clustering can be achieved using a compabilitity measure. The use of a compatibility measure in order to compare objects described by different features allows the analysis of the different values, maintaining their original nature, without making any transformation. So, there is no a priori loss of information produced by previous transformations and avoids taking previous arbitrary decisions that could bias results.
- Semantic clustering is able to provide a partition of objects that considers the meaning of textual terms. The results obtained in the tests show that a semantic clustering approach is able to provide a partition of objects that considers the meaning of the textual values and, thus, the result is more interpretable and permits to discover semantic relations between the objects. The method enriches the results and provides useful knowledge about the characteristics of the objects.

Since this task has been done before starting the implementation of the Personalized Recommender System (task 3), we have used an statistical software to test the algorithms developed in this task. In particular the KLASS software was selected because it has been specially designed to deal with knowledge discovery with advanced AI techniques. KLASS considers only hierarchical clustering methods, so it has not been possible to test the behaviour of partitional algorithms like k-means. From the tests done, we have seen that the time consumption of the Ward's method can be a problem for the problem addressed in this project: the dynamic recommendation of touristic destinations. For this reason, we have started to consider the use of k-means.

The advantages of the k-means algorithm are its simplicity and its time complexity (can be used to cluster large data sets). The stopping criterion usually needs a small number of iterations making this algorithm very efficient. In fact, it has been used in very large data sets. Although there are also some disadvantages, they can be partially solved with some variants that have been already proposed (see deliverable D3). With respect to the parameters needed a priori: the number of clusters and the initial prototypes, we think that they

are not critical in the case of building groups of similar touristic destinations, because we can build some stereotypes from statistical data. We have already done some work on this direction together with our partners of the PCTTO, see (Borràs et al. 2011).

Finally, it is worth to note that both the developed semantic similarity and the semantic clustering tools are general enough to be integrated into other data analysis techniques. The unique requirement is to have, at least, one ontology associated to the semantic features. In this regards, we have seen in this work that ontologies are becoming available in many different domains. Moreover, as the comparisons between the values of semantic features are done by means of an ontology-based similarity measure, without relying on the availability of a domain corpus and any kind of parameter tuning, their applicability in different tasks and scenarios is guaranteed.

5 References

Abril, D., Navarro-Arribas, G. and Torra, V (2010). Towards semantic microaggregation of categorical data for confidential documents, in: Proceedings of the 7th international conference on Modeling decisions for artificial intelligence, Springer-Verlag, Perpignan, France, 266-276.

Al-Mubaid, H. and H. A. Nguyen (2006). A cluster-based approach for semantic similarity in the biomedical domain. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006 New York, USA, IEEE Computer Society, 2713–2717.

Anderberg, M. R. (1973). Cluster analysis for applications. New York, Academic Press Inc.

Armengol, E. (2009). "Using explanations for determining carcinogenecity in chemical compounds." Engineering Applications of Artificial Intelligence **22**(1): 10-17.

Batet, M. (2011) Ontology-based semantic clustering, Ph.D. Thesis, Universitat Rovira i Virgili, Tarragona, Spain.

Batet, M., K. Gibert and A. Valls (2008). The Data Abstraction Layer as knowledge provider for a medical multi-agent system. Knowledge Management for Health Care Procedures, From Knowledge to Global Care, AIME 2007 Workshop K4CARE 2007, Amsterdam, The Netherlands, July 7, 2007, Revised Selected Papers. D. Riaño, Springer-Verlag. **LNAI 4924:** 86-100.

Batet, M., K. Gibert and A. Valls (2011). Semantic Clustering Based On Ontologies: An Application to the Study of Visitors in a Natural Reserve (in press). 3th International Conference on Agents and Artificial Intelligence, Rome, Italy.

Batet, M., D. Isern, L. Marin, S. Martínez, A. Moreno, D. Sánchez, A. Valls and K. Gibert (2010). "Knowledge-driven delivery of home care services." Journal of Intelligent Information Systems (in press).

Batet, M., S. Martínez, A. Valls and K. Gibert (2009). Customization of an agent-based medical system. Artificial Intelligence Research and Development, Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence, CCIA 2009, October 21-23, 2009, Vilar Rural de Cardona (El Bages), Cardona, Spain. S. Sandri, M. Sánchez-Marré and U. Cortés, IOS Press. **202:** 242-251.

Batet, M., D. Sánchez and A. Valls (2010). "An ontology-based measure to compute semantic similarity in biomedicine (in press)." Journal of Biomedical Informatics.

Batet, M., D. Sanchez, A. Valls and K. Gibert (2010a). Exploiting Taxonomical Knowledge to Compute Semantic Similarity: An Evaluation in the Biomedical Domain. Trends in Applied Intelligent Systems. 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, LNAI 6096, Springer**:** 274-283.

Batet, M., D. Sanchez, A. Valls and K. Gibert (2010b). Exploiting Taxonomical Knowledge to Compute Semantic Similarity: An Evaluation in the Biomedical Domain. Trends in Applied Intelligent Systems. 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, Cordoba, Spain, June 1-4, 2010, Proceedings, Part I. N. García-Pedrajas, F. Herrera, C. Fyfe, J. M. Benítez and M. Ali, Springer. **LNAI 6096:** 274-283.

Batet, M., D. Sánchez, A. Valls and K. Gibert (2009). Ontology-based semantic similarity in the biomedical domain. Workshop Intelligent Data Analysis in Biomedicine and Pharmacology, IDAMAP 2009 at 12th Conference on Artificial Intelligence in Medicine, AIME 2009 Verona, Italy, 41-46.

Batet, M., A. Valls and K. Gibert (2008). Improving classical clustering with ontologies. 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, IASC 2008, Yokohama, Japan, International Association for Statistical Computing, 137-146.

Batet, M., A. Valls and K. Gibert (2010a). A distance function to assess the similarity of words using ontologies. Proceeding of the XV congreso español sobre tecnologías y lógica fuzzy, Huelva, 561-566.

Batet, M., A. Valls and K. Gibert (2010b). Performance of Ontology-Based Semantic Similarities in Clustering. Artificial Intelligence and Soft Computing, 10th International Conference, ICAISC 2010, Zakopane, Poland, June 13-17, 2010, Part I. L. Rutkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh and J. M. Zurada. Zakopane. Poland, Springer-Verlag. **LNAI 6113:** 281–288.

Batet, M., A. Valls, K. Gibert and D. Sánchez (2010c). Semantic clustering using multiple ontologies. Artificial intelligence research and development. Proceedings of the 13th International Conference on the Catalan Association for Artificial Intelligence. R. Alquézar,

Benzecri, J. P. (1973). L'analyse des donnees, Paris: Dunod.

A. Moreno and J. Aguilar. Amsterdam, IOS Press**:** 207-216.

Borràs, J., De la Flor, J., Pérez, Y, Moreno, A., Valls, A., Isern, D., Orellana, A., Russo, A., Anton-Clavé, S., (2011) SigTur/E-Destination: a system for the management of complex tourist regions, Proc ENTER conference, Innsbruck (Austria), 39-50.

Brill, E. (2003). Processing Natural Language without Natural Language Processing. 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, CICLing 2003, Mexico City, Mexico, Springer Berlin / Heidelberg, 360-369.

Budanitsky, A. and G. Hirst (2006). "Evaluating wordnet-based measures of semantic distance." Computational Linguistics **32**(1): 13-47.

Calinski, R. B. and J. Harabasz (1974). "A dendrite method for cluster analysis." Comm. in Statistics **3**: 1–27.

Cornet, R. and N. F. Keizer (2008). "Forty years of SNOMED: a literature review." BMC Medical Informatics and Decision Making **8(Suppl 1):S2**.

Day, W. H. E. (1992). Complexity theory: An introduction for practitioners of classification. Clustering and Classification. P. Arabie and L. Hubert. River Edge, NJ, World Scientific Publishing Co., Inc.

Erola, A., Castella-Roca, J., Navarro-Arribas, G. and Torra, V. (2010). Semantic microaggregation for the anonymization of query logs, in: Proceedings of the 2010 international conference on Privacy in statistical databases, Springer-Verlag, Corfu, Greece 127-137.

Etzioni, O., M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld and A. Yates (2005). "Unsupervised named-entity extraction form the Web: An experimental study." Artificial Intelligence **165**: 91-134.

Everitt, B. S., S. Landau and M. Leese (2001). Cluster Analysis. London, Arnold.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, Massachusetts, MIT Press. More information: http://wordnet.princeton.edu.

Forgy, E. (1965). "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications." Biometrics **21**: 768–780.

Gibert, K. and U. Cortés (1998). "Clustering based on rules and Knowledge Discovery in ill-structured domains." Computación y Sistemas, Revista Iberoamericana de Computación **1**(4): 213-227.

Gibert, K., A. García-Rudolph and G. Rodríguez-Silva (2008). "The Role of KDD Support-Interpretation Tools in the Conceptualization of Medical Profiles: An Application to Neurorehabilitation." Acta Informatica Medica **16**(4): 178-182.

Gibert, K., R. Nonell, J. M. Velarde and M. M. Colillas (2005). "Knowledge Discovery with clustering: impact of metrics and reporting phase by using KLASS." Neural Network World **15**(4): 319-326.

Hotho, A., A. Maedche and S. Staab (2002). "Ontology-based Text Document Clustering."

Jain, A. K., M. N. Murty and P. J. Flynn (1999). "Data clustering: a review." ACM Computing Surveys **31**(3): 264-323.

Jiang, J. J. and D. W. Conrath (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. International Conference on Research in Computational Linguistics, ROCLING X, Taipei, Taiwan, 19-33.

Lemaire, B. and G. Denhière (2006). "Effects of High-Order Co-occurrences on Word Semantic Similarities." Current Psychology Letters - Behaviour, Brain and Cognition **18**(1): 1.

Lieberman, M., T. Ricciardi, F. Masarie and K. Spackman (2003). The use of SNOMED CT simplifies querying of a clinical data warehouse. AMIA Annual Symposium Proceedings, 910.

Lin, D. (1998b). An Information-Theoretic Definition of Similarity. 15th International Conference on Machine Learning (ICML98), Madison, Wisconsin, USA, Morgan Kaufmann, 296-304.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium on Mathematical Statistics and Probability, 281–297.

Matar, Y., E. Egyed-Zsigmond and S. Lajmi (2008). KWSim: Concepts Similarity Measure. 5th French Information Retrieval Conference en Recherche d'Infomations et Applications, CORIA 2008 475-482, Université de Renne, 475-482.

Miller, G. A. and W. G. Charles (1991). "Contextual correlates of semantic similarity." Language and Cognitive Processes **6**(1): 1-28.

Murty, M. N. and G. Krishna (1980). "A computationally efficient technique for data clustering." Pattern Recogn. **12**: 153–158.

Patwardhan, S. and T. Pedersen (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. EACL 2006 (European Association for Computational Linguistics) Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy, 1-8.

Pedersen, T., S. Pakhomov, S. Patwardhan and C. Chute (2007). "Measures of semantic similarity and relatedness in the biomedical domain." Journal of Biomedical Informatics **40**(3): 288-299.

Petrakis, E. G. M., G. Varelas, A. Hliaoutakis and P. Raftopoulou (2006). "X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies." Journal of Digital Information Management (JDIM) **4**: 233-237.

Pirró, G. (2009). "A semantic similarity metric combining features and intrinsic information content." Data & Knowledge Engineering **68**(11): 1289-1308

Resnik, P. (1995). Using Information Content to Evalutate Semantic Similarity in a Taxonomy. 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Montreal, Quebec, Canada, Morgan Kaufmann Publishers Inc. , 448-453.

Rubenstein, H. and J. Goodenough (1965). "Contextual correlates of synonymy." Communications of the ACM **8**(10): 627-633.

Sánchez, D., M. Batet and A. Valls (2009). Computing knowledge-based semantic similarity from the Web: an application to the biomedical domain. Knowledge Science, Engineering and Management. Third International Conference, KSEM 2009, Vienna, Austria, November 25-27, 2009. Proceedings. D. Karagiannis and Z. Jin, Springer Berlin / Heidelberg. **LNAI 5914:** 17-28.

Sánchez, D., M. Batet and A. Valls (2010a). "Web-Based Semantic Similarity: An Evaluation in the Biomedical Domain." International Journal of Software and Informatics **4**(1): 39-52.

Sánchez, D., M. Batet, A. Valls and K. Gibert (2010b). "Ontology-driven web-based semantic similarity." Journal of Intelligent Information Systems **35**(3): 383-413.

Spackman, K. (2004). "SNOMED CT milestones: endorsements are added to already-impressive standards credentials." Healthcare Informatics **21**(9): 54-56.

Tirozzi, B., D. Bianchi and E. Ferraro, Eds. (2007). Introduction to computational neurobiology and clustering. Series on Advances in Mathematics for Applied Sciences. Singapore, World Scientific Publishing.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. 12th European Conference on Machine Learning, ECML 2001, Freiburg, Germany, Springer-Verlag, 491-502.

Tversky, A. (1977). "Features of Similarity." Psycological Review **84**(4): 327-352.

Ward, J. H. (1963). "Hierarchical Grouping to Optimize an Objective Function." Journal of the American Statistical Association **58**: 236-244.

Wu, Z. and M. Palmer (1994). Verb semantics and lexical selection. 32nd annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, Association for Computational Linguistics, 133 -138.

Yarowsky, D. (1995). Unsupervised Word-Sense Disambiguation Rivalling Supervised Methods. 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, 189-196.

Zhou, Z., Y. Wang and J. Gu (2008). A New Model of Information Content for Semantic Similarity in WordNet. Second International Conference on Future Generation Communication and Networking Symposia, FGCNS 2008, Sanya, Hainan Island, China, IEEE Computer Society, 85-89.