UNIVERSITAT ROVIRA I VIRGILI

ONTOLOGY BASED SEMANTIC CLUSTERING

Montserrat Batet Sanroma

ISBN: 9788469432327
Dipòsit Legal: T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Montserrat Batet Sanromà

ONTOLOGY BASED SEMANTIC CLUSTERING

PH.D. THESIS

Supervised by
Dr. Aida Valls and Dr. Karina Gibert

Department of
Computer Science and Mathematics

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2010

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI

Universitat Rovira i Virgili (URV)
Departament d'Enginyeria Informàtica i Matemàtiques
Escola Tècnica superior d'Enginyeria
Av. Països Catalans, 26 (Campus Sescelades)
43007 Tarragona
Telèfon: +34 977 559688
Fax: +34 977 559710
E-mail: aida.valls@urv.cat

FEM CONSTAR que aquest treball, titulat "Ontology based semantic clustering", que presenta na Montserrat Batet Sanromà per a l'obtenció del títol de Doctor, ha estat realitzat sota la direcció del a Dra. Aïda Valls al Departament d'Enginyeria Informàtica i Matemàtiques d'aquesta universitat i sota la direcció de la Dra. Karina Gibert Oliveras del Departament d'Estadística i Investigació Operativa de la Universitat Politècnica de Catalunya.

Tarragona, 13 de desembre de 2010

La directora de la tesi doctoral

Dra. Aïda Valls Mateu
Universitat Rovira i Virgili

La co-directora de la tesi doctoral

Dra. Karina Gibert Oliveras
Universitat Politècnica de Catalunya

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation (DAMASK project, *Data mining algorithms with semantic knowledge*, TIN2009-11005) and the Spanish Government (PlanE, Spanish Economy and Employment Stimulation Plan), the Universitat Rovira i Virgili (2009AIRE-04), and the European IST project K4Care: Knowledge Based Home Care eServices for an Ageing Europe (IST-2004-026968) project.

The author thanks the "Observatori de la Fundació d'Estudis Turístics Costa Daurada" and "Parc Nacional del Delta de l'Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)" for the data provided. The author acknowledges the collaboration of E. Fourier, D. Corcho, N. Maló and N. Corral in the data preparation. The author acknowledges to *KLASS guys* for providing the basic software, and in particular to Alejandro García-Rudolfh.

The author is also supported by a research grant provided by the Universitat Rovira i Virgili.

Finally, I would like to thank all the people who in one way or another have made possible this thesis: advisors, my colleagues, the lecturers of my department that have helped me, the staff of my department, my friends, and specially my family, who have supported me all this time.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Abstract

Clustering algorithms have focused on the management of numerical and categorical data. However, in the last years, textual information has grown in importance. Proper processing of this kind of information within data mining methods requires an interpretation of their meaning at a semantic level. In this work, the concept of semantic similarity is introduced to provide a formal framework for those data with additional semantic knowledge available. This is used as a basic piece to extend a clustering method that can interpret in an integrated manner, numerical, categorical and semantic data. The kernel of the research consisted in the definition of new semantic similarity measures that calculate the alikeness between words by exploiting available knowledge sources to better evaluate the similarity between semantic terms. Available knowledge is formalized by means of ontologies. Two new ways of compute semantic similarity are defined, based on 1) the exploitation of the taxonomical knowledge available on one or several ontologies and 2) the estimation of the information distribution of terms in the Web. An extension of a compatibility measure has been developed to introduce the proposals in the clustering method. The proposals have been extensively evaluated with benchmarked and real data with successful results. Experiments and applications show that a proper interpretation of textual data at a semantic level improves clustering results as well as the interpretability of the classifications.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Contents

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

List of Figures

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

List of Tables

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Chapter I

1 Introduction

Nowadays the extensive use of information and communication technologies provides access to a large amount of data (e.g. Wikipedia, questionnaires in an electronic way to a great number of people, etc). Usually, these resources provide qualitative data that may be semantically interpreted in order to extract useful knowledge. For example, dataset can contain attributes whose values have an associated semantics. For example, in a question about the "main hobby", the respondent can use short expressions like: photography of birds, jogging, dancing, trekking, horror films or play classical music.

The management of non-numerical data traditionally is a task typically associated to Artificial Intelligence methods. Data-mining techniques and in particular clustering algorithms were conceived for managing non-numerical data.

From the different methods included in the field of Data Mining, we have focused on knowledge discovery from data using clustering (Han and Kamber 2000). Clustering is a masterpiece in many data mining methodologies, because it builds a classification or partition into coherent clusters from unstructured data sets.

In fact, clustering has been used in many data mining problems, such as to build a structure of a complex data set, to reveal associations between objects or to make generalizations. Some exemplary problems illustrating the role of clustering in these tasks is given in (Mirkin 2005). Clustering methods have been practically applied in a wide variety of fields, ranging from engineering (*e.g.* pattern recognition, mechanical engineering, electrical engineering), computer sciences (*e.g.* web mining, spatial database analysis, image segmentation, privacy), life and medical sciences (*e.g.* genetics, biology, microbiology, palaeontology, psychiatry, pathology), to earth sciences (*e.g.* geography. geology, remote sensing), social sciences (*e.g.* sociology, psychology, archaeology, education), and economics (*e.g.* marketing, business)). Recently, new fields of application have increased the research on this topic, specially due to the developments in information retrieval and text mining, spatial database applications (Han, Kamber et al. 2001; Fan 2009; Monreale, Andrienko et al. 2010), Web applications (Cadez, Heckerman et al. 2003; Carpineto, Osinski et al. 2009; Kimura, Saito et al. 2010) and DNA analysis in computational biology (Tirozzi, Bianchi et al. 2007; Chen and Lonardi 2009; Romdhane, Shili et al. 2010), among others.

Classical clustering where, originally, developed in statistical context, and it was applied to numerical data. However, non-numerical (textual) data was also

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

introduced. This data, whose values are expressed with linguistic labels or terms, were exploited in form of categorical features. However, in this context, categorical data is treated at a syntactic level, and comparisons between their values have been limited to check the equality/inequality. In the most sophisticated cases the rarity of the terms (Benzecri 1973) or ordering relationships (if the variable is ordinal) are considered. So, no semantic interpretation of terms is done.

Fuzzy set theory (defined by Zadeh in 1968) introduced the concept of linguistic variables and provided mechanisms to work with them and represent their semantics by means of membership functions to fuzzy sets. From that point, several ways of semantic association of linguistic terms have been developed (Quevedo, Gibert et al. 1993; Herrera and Martínez 2000; Valls, Batet et al. 2009). At the end of the 90s it appeared the term "*computing with words*", which includes aspects related with those approximations (Zadeh and Kacprzyk 1999). Data mining methods have also incorporated that type of variables to their classical algorithms (some of them widely known like Fuzzy C-means (Bezdek 1981) or Fuzzy Expert Systems (Siler and Buckley 2005)). Even though some of those approximations allow dealing with linguistic information, and even providing a formal model to capture imprecision related with linguistic terms, they do not allow interpreting the sense of those terms.

With respect to the treatment of textual data, in the area of Computational Linguistics, it is worth mentioning the research related to the computation of semantic similarity measures in order to compare terms or concepts, using some knowledge base (e.g. a taxonomy\ontology, textual corpus, dictionaries, the Web, etc) (Turney 2001; Patwardhan and Pedersen 2006). These knowledge based measures allow a more accurate management of textual values at semantic level (e.g. for hobbies, *trekking* is more similar to *jogging* than to *dancing*). As a result, their application to clustering methods may improve the results when dealing with textual data.

As it has been stated, there is a lack of tools for treating qualitative information in the form on short texts from a semantic point of view. Such tools would be of great utility to open the possibility of using Data Mining methods that include this kind of values, which are becoming available thanks to the new technologies.

In fact, a new field of research called *Domain-Driven Data Mining* (DDDM) has recently appeared (Cao, Yu et al. 2010) that explores how to exploit semantic knowledge of a domain in order to improve the intelligent data analysis techniques. Then, the exploitation of the conceptual meaning of terms, when available, will certainly provide more accurate estimations of their degree of similarity (e.g. for hobbies, *trekking* is more similar to *jogging* than to *dancing*).

Summarizing, an adequate treatment of textual data improves the quality and the interpretability of the results of clustering. This establishes a new setting for data mining techniques that can be enhanced with the semantics tools developed in this work.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

1.1 Problem contextualization

Clustering is generally seen as the task of finding groups of similar individuals based on information found in data, which means that the data individuals in the same group are more similar to each other than to individuals in other groups. So, clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories) (Xu and Wunsch 2005).

The standard input of a clustering algorithm is a data matrix χ, where objects or individuals $I= \{1,..,n\}$ are described by the K features $X_1 \ldots X_K$.

Typically, objects are in rows, and columns contain the K features that describe them. Each object i can be represented as a vector x_i of K measurements (multidimensional vector), where each dimension refers to a feature (attribute, dimension, or variable) describing i.

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ik}, \ldots, x_{iK})$$

Thus, in the matrix (Figure 1) each cell x_{ik} (with $1 \leq i \leq n$ and $1 \leq k \leq K$) contains the value taken by object i for feature X_k.

$$\chi = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{iK} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{nK} \end{pmatrix}$$

Figure 1. Data matrix

A data matrix can contain different types of features. Classically, in a clustering context numerical features have been considered. From the 80s on, some works include also categorical features (Gibert and Cortés 1997).

- The values of *numerical* ones belong to $\mathbb{R}$ or $\mathbb{N}$ and have the classical arithmetic associated. For these features comparison operators can be directly taken from classical algebra.
- *Categorical* features take a set of symbolic values (named categories). Some subtypes of categorical variables can be distinguished: nominal, ordinal, or linguistic variables (composed by fuzzy sets), depending on the ordering relationships existing among their modalities or the imprecision modelling. Categorical features are traditionally treated at a syntactic level. Then, the comparison between two modalities is often based on their equality/inequality and on the rarity of those modalities. The addition of a total order between the

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

modalities permits to apply other ordinal operations in the comparisons (Walesiak 1999; Jajuga, Walesiak et al. 2003)..

Literature is abundant on proposals for calculating the distances or similarities for mixtures of numerical and categorical variables (Gibert, Nonell et al. 2005). However, the use of these features do not allows a semantic interpretation of data, in particular the sense of values of categorical data (i.e. terms) is not considered. So, the semantics of those objects whose representative data is extracted from resources which contain textual and qualitative data (e.g. the web and electronic questionnaires) cannot be properly exploited.

In the area of Computational Linguistics the assessment of the semantic similarity between terms or concepts is and active trend. Semantic similarity/relatedness quantifies how words extracted from documents ore textual descriptions are alike. Both, semantic similarity and relatedness, are based on the evaluation of the semantic evidence observed in a knowledge source (e.g. corpus, dictionaries, taxonomies/ ontologies (Studer, Benjamins et al. 1998a), the Web, etc.). Different approaches to compute semantic similarity between concepts can be disguised according to the techniques employed and the knowledge exploited to perform the assessment. First, there are unsupervised approaches in which semantics are inferred from the information distribution of terms in a given corpus (Landauer and Dumais 1997; Etzioni, Cafarella et al. 2005). Other trends exploit structured representations of knowledge as the base to compute similarities. There exist ontology-based measures which combine the knowledge provided by an ontology and the Information Content (IC) of the concepts that are being compared. IC measures the amount of information provided by a given term from its probability of appearance in a corpus (Resnik 1995). Without relying on a domain corpus, other approaches consider ontologies as a graph model in which semantic interrelations are modelled as links between concepts (Rada, Mili et al. 1989; Wu and Palmer 1994; Leacock and Chodorow 1998).

We can conclude that an adequate treatment of this type of data establishes a new setting for data mining techniques such as clustering and the development of tools to deal with the semantics of these terms in a clustering context is highly convenient to improve the quality and the interpretability of the results.

The design of a semantic clustering approach requires to considerer different points:

- Study how this kind of data will be introduced and exploited in the clustering process.
- Review semantic/relatedness measures in the literature in order to compare terms and define new ones.
- Although we are interested in the semantic interpretation of terms, an object *i* can be described by a combination of different types of features.
- A clustering algorithm must be generalized in order to take into account the semantic interpretation of data values, when available.

In order to introduce those terms whose sense must be semantically compared, a new type of feature X_k will be introduced, called *semantic feature*.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

In our approach different types of features will be considered: numerical, categorical and semantic features and a hierarchical clustering algorithm will be generalized to include semantic features. In the kernel of basic clustering algorithm numerical and qualitative features are considered and comparisons between objects represented by different types of features were done with a compatibility measure which permits a homogeneous treatment of all features, in spite of the heterogeneity of their scale of measurement. In our approach the same philosophy will be applied and each feature will be compared separately depending on its type and then integrated using a compatibility measure. As measures for numerical and categorical features are well defined, the compatibility measure will use some of the existent measures in the literature. For this reason we will focus on providing the way to perform comparisons between objects in the semantic case, in particular in the interpretation of semantic features in order to properly asses the similarity/distance between them, generalizing a hierarchical clustering algorithm(Gibert 1996) where numerical and categorical features are considered to include also semantic features .

1.2 Knowledge sources

As explained, semantic similarity measures can be classified according to the type of knowledge they exploit. Although, in this work many knowledge sources will be considered, we will focus our contributions in semantic clustering in the use of ontologies and the Web as a corpus.

1.2.1 Ontologies

In (Studer, Benjamins et al. 1998b) an ontology is defined as "a formal, explicit specification of a shared conceptualization". *Conceptualization* refers to an abstract model of some phenomenon in the world by having identified its relevant concepts. *Explicit* means that the type of concepts identified, and the constraints of their use, are explicitly defined. *Formal* refers to the fact that the ontology should be machine-readable. *Shared* reflects the notion that ontology captures consensual knowledge, that is, not a personal view of the target phenomenon of some particular individual, but one accepted by a group.

Neches et al. (Neches, Fikes et al. 1991) give a definition focused on the form of an ontology. An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary. In (Gruber 1993) an ontology is defined as explicit specifications of a conceptualization.

Ontologies are designed for being used in applications that need to process the content of information, as well as, to reason about it, instead of just presenting information to humans. They permit greater machine interpretability of content than

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

that supported by XML, RDF and RDF Schema (RDF-S), by providing additional vocabulary along with a formal semantics.

From a structural point of view (Stumme, Ehrig et al. 2003; Cimiano 2006), an ontology is composed by disjoint sets of *concepts, relations, attributes* and *data types. Concepts* are sets of real world entities with common features. *Relations* are binary associations between concepts. There exist inter-concept relations which are common to any domain and domain-dependant associations. *Attributes* represent quantitative and qualitative features of particular concepts, which take values in a given scale defined by the *data type*.

Concepts are classes organized in one or several *taxonomies*, linked by means of transitive *is-a* relationships (taxonomical relationships). Multiple inheritance (i.e. the fact that a concept may have several hierarchical ancestors or subsumers) is also supported.

Binary relations can be defined between concepts. In those cases, the concept in the origin of the relation represents the *domain* and those in the destination, the *range*. Those relationships may fulfil properties such as *symmetry* or *transitivity*.

Some standard languages have been designed to codify ontologies. They are usually declarative languages based on either first-order logic or on description logic. Some examples are KIF, RDF (Resource Description Framework), KL-ONE, DAML+OIL and OWL (Web Ontology Language) (Gómez-Pérez, Fernández-López et al. 2004). The most used are OWL (Bechhofer, van Harmelen et al. 2009) and RDF (Fensel 2000).

Ontologies can be classified in several forms. An interesting classification was proposed by Guarino (Guarino 1998), who classified types of ontologies according to their level of dependence on a particular task or point of view:

- *Top-level ontologies*: describe general concepts like space, time, event, which are independent of a particular problem or domain. Examples of top-level ontologies are: Sowa's (Sowa 1999) and Cyc's (Lenat and Guha 1990).
- *Domain-ontologies*: describe the vocabulary related to a generic domain by specializing the concepts introduced in the top-level ontology. There are a lot of examples of this type of ontologies in e-commerce UNSPSC, NAICS, I biomedicine SNOMED CT, MESH, etc.;
- *Task ontologies*: describe the vocabulary related to a generic task or activity by specializing the top-level ontologies.
- *Application ontologies*: they are the most specific ones. Concepts often correspond to roles played by domain entities. They have a limited reusability as they depend on the particular scope and requirements of a specific application. Those ontologies are typically developed ad hoc by the application designers (Batet, Isern et al. 2010; Valls, Gibert et al. 2010).

Domain ontologies, on one hand, are general enough to be required for achieving consensus between a wide community of users or domain experts and, on the other hand, they are concrete enough to present an enormous diversity with many different and dynamic domains of knowledge and millions of possible concepts to model. Being machine readable, they represent a very reliable and structured knowledge source.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Thanks to initiatives such as the Semantic Web, which brought the creation of thousands of domain ontologies (Ding, Finin et al. 2004), ontologies have been extensively exploited to compute semantic likeness.

However, an investigation of the structure of existing ontologies via the Swoogle ontology search engine (Ding, Finin et al. 2004) reveals that domain ontologies very occasionally model any semantic feature apart from *taxonomical relationships.*

1.2.2 The Web

In the last years, with the enormous growth of the Information Society, the Web has become a valuable source of information for almost every possible domain of knowledge. This has motivated many researches to start considering the Web as a valid repository for Information Retrieval and Knowledge Acquisition tasks.

So, the Web, thanks the huge amount of information available for every possible domain and its high redundancy, can be a valid knowledge source for similarity computation. In this sense, the amount and heterogeneity of information is so high that it can be assumed that the Web approximates the real distribution of information (Cilibrasi and Vitányi 2004), representing the hugest repository of information available (Brill 2003).

In many knowledge related tasks the use of statistical measures (e.g. co-occurrence measures) for inferring the degree of relationship between concepts is a very common technique when processing unstructured text (Lin 1998a). However, these techniques typically suffer from the *sparse data problem* (i.e. the fact that data available on words may not be indicative of their meaning). So, they perform poorly when the words are relatively rare (Sánchez, Batet et al. 2010b).

In that sense, the size and the redundancy of the Web has motivated some researches to consider it as a corpus from which extract evidences of word relationships. Some authors (Turney 2001; Brill 2003) have demonstrated the convenience of use a wide corpus as the Web to address the data sparse problem.

However, the analysis of such an enormous repository is, in most cases, impracticable. Here is where the use of web search engines (e.g. Google, Bing, Yahoo) can properly scale this high amount of information, obtaining good quality and relevant statistics. So, robust web-scale statistics about information distribution in the whole Web can be obtained in a scalable and efficient way from queries performed into a web search engine (Sánchez 2008) (Sánchez 2009).

These statistics about the presence of a certain query term in the Web can be computed efficiently from the estimated amount of returned results (see Figure 2).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

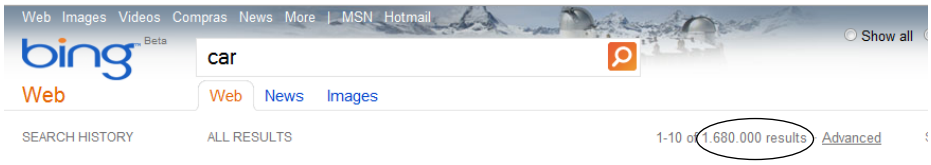


Figure 2. Number of results for a query term returned by Bing.

This is important to discover the relative frequencies of words. So, the probabilities of web search engine terms, conceived as the frequencies of page counts returned by the search engine divided by the number of indexed pages, approximate the actual relative frequencies of those search terms used in society (Cilibrasi and Vitányi 2004). This measure can give us an idea of the generality of a word. So, publicly available search engines can be a useful tool for extracting the required knowledge for computing similarity.

The use of web search engines for obtaining valuable statistics for information retrieval and knowledge acquisition has been applied previously by several authors (Turney 2001; Cilibrasi and Vitányi 2004; Etzioni, Cafarella et al. 2004) obtaining good quality results.

1.3 Semantic feature

We introduce here the concept of semantic feature as a key stone to exploit semantic knowledge in data mining in general and clustering in particular. The whole document builds contributions on the bases of this concept.

Definition 1. Given a feature X_k which takes values in a domain D $=\{c_1,c_2,..c_n\}$ and a set of knowledge sources ψ (which can be both ontologies and the Web), X_k is a *semantic feature* if:

$$\forall c_i \in D, \exists o \in \Psi \mid c_i \in o$$

Since the values of *semantic features* became concepts (i.e. they correspond to labels of concepts in the reference ontology) rather than simple modalities, this formal framework opens the door to perform comparisons between values from a semantic point of view and will be used as the basic piece for the contributions developed in this work.

1.4 Problem definition

So, this work will present a clustering approach in which the semantics of terms are considered when available, presenting a methodology to include semantic features in

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

clustering comparing them using a semantic similarity measures based on the exploitation of some knowledge source.

Then, we can define the problem as:

Given,

- A set of knowledge sources ψ (ontologies or the web).
- A set of n objects $I=\{1,..,n\}$
- A set χ of K features describing the objects of I, where $\chi=\zeta \cup Q \cup S$, where

$$\zeta = \{k : X_k \text{ is a numerical feature, } k=1:n_\zeta\}$$
$$Q = \{k : X_k \text{ is a categorical feature, } k=1: n_Q\}$$
$$S = \{k : X_k \text{ is a semantic feature, } k=1:n_S, X_k \text{ linked to } \psi\}$$

Such that $K= n_{\zeta+} n_{Q+} n_{S.}$

We, search a dendogram τ build over I by means of a hierarchical clustering. Such that:

- *The linkages between S and ψ are considered for building τ.*
- *Horizontal cuts of τ provide partitions of I semantically consistent with ψ.*

To solve that problem, the research has been developed addressing several goals described in section 1.5. In particular, this work will study the use of ontologies and the Web as corpus as knowledge sources.

1.5 Goals

The main goals of the present work are:

- Define forms to compare objects described by one or several semantic features. This is the most critical part of the research and requires:
  - To survey and compare different approaches to assess the similarity between a pair of terms or concepts. The main advantages and problems will be identified.
  - As said before, two knowledge sources will be considered. Thus we need:
    - To study the ontology structure in order to extract that knowledge that can provide a higher degree of similarity evidence between concepts. Considering the basic ontological concepts described in the previous section, we will focus on exploiting the taxonomical relationships between concepts.
    - To study the Web environment and the available web information retrieval tools in order to assess similarity using the Web as corpus.
- To make contributions of new ways of assessing the semantic similarity between terms by:

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

  - Proposing a new measure of semantic similarity based on the exploitation of the taxonomical knowledge of an ontology able to avoid some problems of existing measures and improve their performance.
  - Proposing a new approach of computing the Information Content (IC) of a term (related to the fact of considering the probability of appearance of terms in a corpus as evidence of their semantic content) from both the Web and ontologies, able to improve the performance of those IC-based similarity measures.
- To evaluate the accuracy of our proposals with respect to those similarity functions presented in the state of the art.
  - Results for several domains will be studied.
  - Computationally efficiency will be also evaluated.
  - Consider the problem of that similarity computation depends on the goodness of the background knowledge (e.g. on the completeness and homogeneity of ontologies).
- To address the case in which multiple ontologies are available. For that, we need:
  - To study different approaches to asses similarity using multiple ontologies.
  - To design a methodology to assess the similarity between terms or concepts using multiple ontologies able to obtain better results than those found in the literature.
  - To evaluate this methodology for integrating the knowledge of different ontologies in order to assess the similarity comparing it with approaches found in the literature.
- Identify which on the approaches presented for computing similarity is the best to be applied to the clustering.
- Generalize a previous compatibility measure designed for numerical and categorical features to include semantic features. This requires:
  - A study of how objects are compared when they are represented by different types of features.
  - Definition of a compatibility measure that can deal with numerical, categorical and those textual features for which semantic information is available.
  - Include our proposals of semantic features comparison in the compatibility measure.
- Modify a hierarchical clustering algorithm.
  - A review of different clustering methods will be done.
  - To accept data matrices with semantic features
  - To manage the associated knowledge source
  - To use the compatibility measure defined before.
- Developing an adapted methodology able to exploit the semantics of objects, named *Ontology based Semantic Clustering*.
- To apply the *Ontology based Semantic Clustering* to real data sets.
- To test the suitability and performance of the semantic clustering process, by evaluating the obtained results.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

1.6 Overview of this document

The rest of this document is organized as follows:

- Chapter 2: This chapter presents an extensive survey of most of the approaches dealing with similarity/relatedness estimation developed in recent years. After a deep analysis of their advantages and problems, two new measures have been presented: a new measure for computing semantic similarity exploiting the knowledge of an ontology and; a new approach for computing concept IC from the Web based on contextualizing the queries to a web search engine using an ontology. This chapter also introduces the evaluation and comparison of these new measures against related works considering a general domain an, the specific domain of biomedical terminology.
- Chapter 3: In this chapter related works focusing on ontology-based semantic similarity in a multi-ontology setting are introduced and analyzed. From this analysis different limitations of these works have detected. Moreover, this chapter describes in detail a new method to assess semantic similarity by exploiting multiple ontologies. Finally, a detailed comparison of our proposal against related works is done.
- Chapter 4: In this chapter a survey of clustering algorithms and a discussion of the different approaches is provided. Moreover, it is described the semantic clustering approach. Concretely, it is defined a compatibility measure used to compare objects whose features can be numerical, categorical and semantic, showing that the measures developed in the previous chapters are applied to the clustering by comparing the values of semantic features. Details of the clustering algorithm are also presented.
- Chapter 5: This chapter presents the evaluation of the semantic clustering approach. There are studied the consequences of including semantic features and ontologies to the clustering, and the dependence of the clustering results on the accuracy of the similarity. In addition, the presented clustering approach is evaluated with real data, analyzing the influence of using one or multiple ontologies when assessing similarity.
- Chapter 6: In this chapter, the contributions of this work have been applied to different fields. In this chapter some preliminary results of these applications are showed.
- Chapter 7: This chapter, contains a summary of the work, the main conclusions, and presents some lines of future research. The publications resulting from this work are listed.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Chapter II

2 Semantic similarity/relatedness

Most data mining process and clustering in particular requires the evaluation of data attributes in order to detect the degree of alikeness between records or individuals. On the contrary to numerical data, which can be directly and easily manipulated and compared by means of classical mathematical operators, the processing of qualitative data is a changeling task. Words are labels referring to concepts, which define their semantics. Semantic similarity is precisely the science that aims to estimate the alikeness between words or concepts by discovering, evaluating and exploiting their semantics. Due to semantics is an inherently human feature, methods to automatically calculate semantic similarity relies on evidences retrieved from one or several manually constructed knowledge sources. The goal is to mimic human judgments of similarity by exploiting implicit or explicit semantic evidences.

It is important to note that two different concepts, which are often confounded, can be found in the literature. On one hand, *semantic similarity* states how taxonomically near two terms are, because they share some aspects of their meaning (*e.g.*, *dogs* and *cats* are similar to the extend they are mammals; and *bronchitis* and *flu* are similar because both are disorders of the respiratory system). On the other hand, the more general concept of *semantic relatedness* does not necessary relies on a taxonomic relation (*e.g.*, *car* and *wheel* or *pencil* and *paper*); other non taxonomic relationships (*e.g.*, meronymy, antonymy, functionality, cause-effect, etc.) are also considered (e.g., *diuretics* help in the treatment of *hypertension).*Both are based on the evaluation of the semantic evidence observed in a knowledge source (such as textual corpus, dictionaries, taxonomies/ontologies, etc.)).

Semantic similarity/relatedness computation has many direct and relevant applications. In particular, it has been mainly studied in computational linguistics because the assessment of semantic likeness between words helps to interpret textual data. As it has been demonstrated in psychological experiments (Goldstone 1994), it acts as a fundamental organizing principle by which humans organize and classify objects. Some basic natural language processing tasks such as word sense disambiguation (Resnik 1999; Patwardhan, Banerjee et al. 2003), synonym detection (Lin 1998b), document categorization or clustering (Cilibrasi and Vitányi 2006; Aseervatham and Bennani 2009), automatic language translation (Cilibrasi and Vitányi 2006) or automatic spelling error detection and correction (Budanitsky and Hirst 2001) rely on the assessment of words' semantic resemblance. Direct applications can be found in the knowledge management field, such as thesauri generation (Curran 2002), information extraction or ontology learning (Sánchez and

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Moreno 2008a; Sánchez and Moreno 2008b), in which new terms related to already existing concepts, should be acquired from textual resources. The Semantic Web is an especially relevant application area, when dealing with automatic annotation of Web pages (Cimiano, Handschuh et al. 2004), community mining (Mika 2007), and keyword extraction for inter-entity relation representation (Mori, Ishizuka et al. 2007).

Similarity estimation between texts has also an important role in the classification and structuring of textual resources such as digital libraries (Sánchez and Moreno 2007), in which resources should be classified according to the similarity of their main topics (expressed as textual signatures), and in information retrieval (IR) (Lee, Kim et al. 1993; Hliaoutakis, Varelas et al. 2006; Ratprasartporn, Po et al. 2009), in which similar or related words can be used to expand user queries and improve recall (Sahami and Heilman 2006). It is also exploited in the elaboration of methods for integrating the knowledge of different data bases into unique queries, where equivalent concepts must be identified (Schallehn, Sattler et al. 2004).

Applied domains such as biomedicine, chemistry or engineering have been especially considered by the research community (Hliaoutakis 2005; Al-Mubaid and Nguyen 2006; Morbach, Yang et al. 2007; Pedersen, Pakhomov et al. 2007; Armengol 2009; Pirró 2009) due to the proliferation and importance of terminology. In this case, similarity assessment can aid to discover semantically equivalent terms corresponding to different lexicalizations, synonyms, abbreviations or acronyms of the same concept. This is of great interest in healthcare in order to be able to retrieve the desired information from a literature data base, especially tasks such as patient cohort identification (Bichindaritz and Akkineni 2006; Pedersen, Pakhomov et al. 2007). Authors have also applied them to discover similar protein sequences (Lord, Stevens et al. 2003) or to the automatics indexing and retrieval of biomedical documents (e.g. in PubMed digital Library)(Wilbu and Yang 1996).

Other fields of applications are appearing. Privacy preserving techniques can also take advantage of semantic knowledge to mask data in order that it cannot be linked to personal and confidential values. Anonymization based on semantic knowledge is proposed in (Martínez, Sánchez et al. 2010a). Some video and image understanding techniques are also based on the semantic interpretation of the textual features referred to the images for indexing or searching purposes (Allampalli-Nagaraj and Bichindaritz 2009). Semantic filtering of multimedia content needs to discover the relationships that exist between semantic concepts. In (Naphade and Huang 2001), some relevant concepts may not be directly observed in terms of media features, but are inferred based on their semantic likeness with those that are already detected.

In the context of this work, the proliferation of textual data referring to user descriptions (*e.g.*, polls or questionnaires), enables word similarity measurement contribute to develop specific data mining algorithms that take into account the semantics of the values. In particular, taking into account the semantics of the terms for detecting profiles or preferences must constitute a great improvement in the conceptual coherence of these profiles and preferences. This is one of the main reasons for introducing semantic similarity/relatedness into clustering or classification techniques (Batet, Valls et al. 2008; Chen, Garcia et al. 2009) aiding the development of decision support systems.

Despite its usefulness, robust measurement of semantic similarity/relatedness between textual terms remains a challenging task (Bollegala, Matsuo et al. 2007).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Many works have been developed in the last years, especially with the increasing interest on the Semantic Web. Proposed methods aim to automatically assess a numerical score between a pair of terms according to the semantic evidence observed in one or several knowledge sources, which are used as semantic background. According to the concrete knowledge sources exploited for the semantic assessment (*e.g.*, ontologies, thesaurus, domain corpora, etc.) and the way to use them, different families of methods can be identified.

The objective of this chapter is, on the one hand, to compare the different semantic similarity/relatedness paradigms and discuss their performance and, on the other hand, to present our contributions in this area and show their validity.

In section 2.1, we survey, classify and compare most of the approaches dealing with similarity/relatedness estimation developed in recent years. We have collected much more measures than related works (Seco, Veale et al. 2004; Patwardhan and Pedersen 2006; Petrakis, Varelas et al. 2006; Bollegala, Matsuo et al. 2007; Wan and Angryk 2007; Zhou, Wang et al. 2008; Bollegala, Matsuo et al. 2009), offering an updated and more detailed review and comparison of the expected performance of these measures both from theoretical and practical points of view. Concretely, for each family of functions, we identify their main advantages and limitations under the dimensions of expected accuracy, computational complexity, dependency on knowledge sources (type, size, structure-dependency and pre-processing) and parameter tuning.

After the analysis of the existing measures, in section 2.2 we make two proposals of new methods for computing term similarity by means of the exploitation of the taxonomical knowledge available in an ontology. In section 2.2.1 it is presented a new pure ontology-based semantic similarity that considers semantic evidence typically omitted by other approaches. This measure aims to rival more complex approaches in terms of accuracy, but having a low computational cost. Section 2.2.2 presents a new way of computing concept IC from the Web by means of contextualizing its assessment using an ontology. Its application in Resnik-based measures is studied.

Section 2.3, evaluates and compares the results obtained by our measures against those reported by related works. The analysis is done in two settings: first, taking a general domain and, second, considering the specific domain of biomedical terminology. The results show that our proposals provide a higher accuracy without having some of the limitations identified on some of the existing ones. In both cases, standard benchmarks have been used, in order to enable an objective comparison that sustains the conclusions obtained.

The last section summarizes the chapter and presents some conclusions.

2.1 Semantic similarity/relatedness: state of the art

According to the type of knowledge exploited to extract semantic evidences and the principles in which similarity estimation relies, measures can be grouped in several families of functions. In this section, we survey, review and compare them according to the following classification:

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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 section we try to maintain the notation of the original papers, if it is necessary we will adapt it to our context in the following chapters.

2.1.1 Ontology-based measures

As stated in the introduction, ontologies provide a formal specification of a shared conceptualization (Guarino 1998). Being machine readable and constructed from the consensus of a community of users or domain experts, they represent a very reliable and structured knowledge source. Due to this reason, and thanks to initiatives such as the Semantic Web, which brought the creation of thousands of domain ontologies (Ding, Finin et al. 2004), ontologies have been extensively exploited to compute semantic likeness. WordNet (Fellbaum 1998) (more details in section 2.3.2.1) is a paradigmatic example typically exploited in the literature, as it consists on a domain-independent and general purpose thesaurus that describes and organises in an ontological fashion more than 100,000 English terms.

In this section, we cover approaches completely or partially relying on ontologies to compute semantic similarity/relatedness.

2.1.1.1 Edge counting-based measures

The first family of measures exploits the geometrical model of semantic pointers provided by ontologies. In fact, ontologies can be seen as a directed graph in which concepts are interrelated mainly by means of taxonomic (is-a) and, in some cases, non-taxonomic links (Sánchez 2010). Input terms are mapped to ontological concepts by means of their textual labels. A straightforward method to calculate the similarity between terms is to evaluate the minimum *Path Length* connecting their corresponding ontological nodes via is-a links (Rada, Mili et al. 1989). The longer the path, the semantically farther the terms are.

Let us define $path(a,b)=\{l_1,....l_k\}$ as a set of links connecting the terms a and b in a taxonomy. Let $|path(a,b)|=k$ be the length of this path. Then, considering all the possible paths from a to b (where i is the number of paths), their semantic distance as defined by (Rada, Mili et al. 1989) is (1). This fact fits on the classical definition of distance between nodes in a graph, from the mathematical point of view.

$$dis_{rad}(a,b) = \min_{\forall i} \left| path_i(a,b) \right| \quad (1)$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

This measure, however, have sometimes a difficult interpretation in the field of computational linguistics. So, several variations and improvements of this edge-counting approach have been proposed. On the one hand, in addition to this absolute distance between terms, Wu and Palmer (Wu and Palmer 1994) considered that the relative depth in the taxonomy of the concepts corresponding to the evaluated terms is an important dimension, because concept specializations become less distinct as long as they are recursively refined. So, equally distant pairs of concepts belonging to an upper level of the taxonomy should be considered less similar than those belonging to a lower level. Wu and Palmer's measure counts the number of is-a links (N_1 and N_2) from each term to their Least Common Subsumer (LCS) (*i.e.*, the most concrete taxonomical ancestor that subsumes both terms) and also the number of is-a links from the LCS to the root (N_3) of the ontology (2).

$$sim_{w\&p}(a,b)\frac{2\times N_3}{N_1+N_2+2\times N_3} \tag{2}$$

Based on the same principle, Leadcock and Chodorow (Leacock and Chodorow 1998) also proposed a non-linear measure that considers both the number of nodes N_p separating the ontological nodes corresponding to terms *a* and *b*, included themselves, and the depth *D* of the taxonomy in which they occur (3).

$$sim_{l\&c}(a,b)=-\log(N_p/2D) \tag{3}$$

Li *et al*., (Li, Bandar et al. 2003) also proposed a similarity measure that combines the shortest path length and the depth of the ontology in a non-linear function (4).

$$sim_{li}(a,b) \;= e^{-\alpha \min_{\forall i}|path_i(a,b)|}\cdot\frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}} \tag{4}$$

where *h* is the minimum depth of the LCS in the hierarchy and $\alpha \geq 0$ and $\beta > 0$ are parameters scaling the contribution of the shortest path length and depth, respectively. Based on the data of a benchmark, authors stated that the optimal parameters for the measure with respect to a concrete set of human judgements were: $\alpha = 0.2$; $\beta = 0.6$. However, this is just an empirical finding for a specific setting. It lacks a theoretical basis and should not be generalized.

Al-Mubaid and Nguyen (Al-Mubaid and Nguyen 2006) proposed a measure that combines the minimum *path length* and the *taxonomical depth*. First, they define clusters for each of the branches in the hierarchy with respect to the root node. They measure the common specificity of two terms by substracting the depth of their LCS from the depth D_c of the cluster (5).

$$CSpec(a,b)=D_c-depth(LCS(a,b)) \tag{5}$$

Then, the common specificity is used to consider that the pairs of concepts found in lower levels should be considered more similar than the ones found in higher levels, following Wu and Palmer's approach. So, the proposed distance measure (*sem*) is defined as follows (6):

$$dis_{sem}(a,b)=\log((\min_{\forall i}\left|path_i(a,b)\right|-1)^{\alpha}\times(CSpec)^{\beta}+k) \tag{6}$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

where $\alpha>0$ and $\beta>0$ are the contribution factors of the path length and the common specify features and k is a constant. Authors use k=1 because with $k\geq1$ they proved that the distance is positive. Moreover, in their experiments, they give an equal weight to the contribution of the two components (path length and common specify) by using $\alpha=\beta=1$.

Both Li *et al*., and Al-Mubaid and Nguyen approaches are often considered in the literature (Petrakis, Varelas et al. 2006; Pirró 2009) as *hybrid* approaches, as they combine several structural characteristics (such as path length, depth and local density) and assign weights to balance the contribution of each component to the final similarity value. Even though their accuracy for a concrete scenario (see evaluation section) is higher than more basic edge-counting measures, they depend on the empirical tuning of weights according to the ontology and input data.

Hirst and St-Onge (Hirst and St-Onge 1998) extended the notion of taxonomical edge-counting by considering also non-taxonomic semantic links in the path (*full_path*). All types of relations found in WordNet together with rules that restrict possible semantic chains are considered, along with the intuition that the longer the path and the more changes in relation's direction, the lower the likeness. The following path directions are considered: upward (such as *hypernymy* and *meronymy*), downward (such as *hyponymy* and *holonymy*) and horizontal (such as *antonymy*). The resulting formula is (7)

$$sim_{h\&s}(a,b) = C - full\_path(a,b) - k \times turns(a,b) \qquad (7)$$

where C and k are constants (C = 8 and k = 1 are used by the authors), and *turns(a, b)* is the number of times the path's direction changes.

Due to the non-taxonomic nature of some of the relations considered during the assessment, Hirst and St-Onge's measure captures a more general sense of *relatedness* than of taxonomical *similarity*, assessed by the approaches detailed above.

The main advantage of the presented measures is their simplicity. They only rely on the geometrical model of an input ontology whose evaluation requires a low computational cost (in comparison to approaches dealing with text corpora, see section 2.1.2). However, several limitations hamper their performance.

In general, any ontology-based measure depends on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology (Cimiano 2006). So, they require rich and consistent ontologies like WordNet to work properly (Pirró 2009). So, general massive ontologies such as WordNet, with a relatively homogeneous distribution of semantic links and good inter-domain coverage, are the ideal environment to apply those measures (Jiang and Conrath 1997).

For the concrete case of taxonomic path-based measures, they only consider the shortest path between concept pairs. However, wide and detailed ontologies such as WordNet incorporate multiple taxonomical inheritance, resulting in several taxonomical paths which are not taken into account. Other features also influencing the concept semantics, such as the number and distribution of common and non-common taxonomical ancestors, are not considered either. As a result, by taking only the minimum path between concepts, many of the taxonomical knowledge explicitly modelled in the ontology is omitted.

Another problem of path-based measures typically admitted (Wan and Angryk 2007; Bollegala, Matsuo et al. 2009) is that they rely on the notion that all links in the

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

taxonomy represent a uniform distance. In practice, the semantic distance among concept specializations/generalizations in an ontology depends on the degree of granularity and taxonomic detail implemented by the knowledge engineer.

2.1.1.2 Feature-based measures

Feature-based methods try to overcome the limitations of path-based measures regarding the fact that taxonomical links in an ontology do not necessary represent uniform distances. This fact is addressed by considering the degree of likeness between sets of features that are built to describe the two terms compared. As a result, they are more general than the previous ones.

So, on the contrary to edge-counting measures which, as stated above, are based on the notion of minimum path distance, feature-based approaches assess similarity between concepts as a function of their properties. This is based on Tversky's model of similarity, which, derived from set theory, takes into account common and non-common features of compared terms. Common features tend to increase similarity and non-common ones tend to diminish it (Tversky 1977). Formally, let be $\Psi(a)$ and $\Psi(b)$ the sets of features of terms a and b respectively, let $\Psi(a) \cap \Psi(b)$ be the intersection between those two sets of features, and $\Psi(a) \setminus \Psi(b)$ is the set of all elements which are members of $\Psi(a)$ but not members of $\Psi(b)$. Then, the similarity between a and b can be computed as a function of $\Psi(a) \cap \Psi(b)$, $\Psi(a) \setminus \Psi(b)$ and $\Psi(b) \setminus \Psi(a)$ as (8).

$$sim(a,b) = \alpha \cdot F(\Psi(a) \cap \Psi(b)) - \beta \cdot F(\Psi(a) \setminus \Psi(b)) - \gamma \cdot F(\Psi(b) \setminus \Psi(a)) \quad (8)$$

where F is a function that reflects the importance of a set of features, and α, β and γ are parameters that weight the contribution of each component.

The information provided by the input ontology is exploited as features. For WordNet, concept synonyms (*i.e.*, *synsets*, which are sets of linguistically equivalent words), definitions (*i.e.*, *glosses*, containing textual descriptions of word senses) and different kinds of semantic relationships can be considered.

In Rodriguez and Egenhofer (Rodríguez and Egenhofer 2003), the similarity is computed as the weighted sum of similarities between synsets, meronyms and neighbour concepts (those linked via semantic pointers) of evaluated terms (9).

$$sim_{r\&e}(a,b) = w \cdot S_{synsets}(a,b) + u \cdot S_{meronyms}(a,b) + v \cdot S_{neighborhoods}(a,b) \quad (9)$$

where w, u and v weight the contribution of each component, which depends on the characteristics of the ontology. Meronyms refer to matching of concepts via part-of relationships.

In Tversky (Tversky 1977) concepts and their neighbours (according to semantic pointers) are represented by synsets. The similarity (10) is computed as:

$$sim_{tve}(a,b) = \frac{|A \cap B|}{|A \cap B| + \gamma(a,b)|A \setminus B| + (1-\gamma(a,b))|B \setminus A|} \quad (10)$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

where *A*, *B* are the synsets for concepts corresponding to *a* and *b*, *A\B* is the set of terms in *A* but not in *B* and *B\A* the set of terms in *B* but not in *A*. The term |A| is the cardinality of the set A. Finally, *γ(a, b)* is computed as a function of the depth of *a* and *b* in the taxonomy as follows (11):

$$\gamma(a,b) = \begin{cases} \dfrac{depth(a)}{depth(a)+depth(b)}, & depth(a) \le depth(b) \\ 1 - \dfrac{depth(a)}{depth(a)+depth(b)}, & depth(a) > depth(b) \end{cases} \quad (11)$$

In Petrakis *et al.*, (Petrakis, Varelas et al. 2006) a feature-based function called *X-similarity* relies on the matching between synsets and concept glosses extracted from WordNet (*i.e.*, words extracted by parsing term definitions). They consider that two terms are similar if the synsets and glosses of their concepts and those of the concepts in their neighbourhood (following semantic links) are lexically similar. The similarity function is expressed as follows:

$$sim_{X-Similarity}(a,b) = \begin{cases} 1, if\ S_{synsets}(a,b) > 0 \\ \max\{S_{neighborhoods}(a,b), S_{glosses}(a,b)\}, if\ S_{synsets}(a,b) = 0 \end{cases}$$

(12)

where $S_{neighborhoods}$ is calculated as follows:

$$S_{neighborhoods}(a,b) = \max \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (13)$$

where each different semantic relation type (*i.e.*, *is-a* and *part-of* in WordNet) is computed separately (*i* denotes the relation type) and the maximum (joining all the synsets of all concepts up to the root of each hierarchy) is taken. $S_{glosses}$ and $S_{synsets}$ are both computed as:

$$S(a,b) = \max \frac{|A \cap B|}{|A \cup B|} \quad (14)$$

where *A* and *B* denote synsets or glosses sets for terms *a* and *b*.

Feature-based measures exploit more semantic knowledge than edge-counting approaches, evaluating both commonalties and differences of compared concepts. However, by relying on features like glosses or synsets (in addition to taxonomic and non-taxonomic relationships), those measures limit their applicability to ontologies in which this information is available. Only big ontologies/thesaurus like WordNet include this kind of information. In fact, a research of the structure of existing ontologies via the Swoogle ontology search engine (Ding, Finin et al. 2004) reveals that domain ontologies very occasionally model other semantic features apart from taxonomical relationships.

Another problem is their dependency on the weighting parameters that balance the contribution of each feature because a lack of clear criteria to assign values (like the

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

hybrid approaches introduced in section 2.1.1.2). In all cases, those parameters should be tuned according to the nature of the ontology and even to the evaluated terms. This hampers their applicability as a general purpose solution. Only the proposal by Petrakis (Petrakis, Varelas et al. 2006) does not depend on weighting parameters, as the maximum similarity provided by each single feature is taken. Even though this adapts the behaviour of the measure to the characteristics of the ontology and to the knowledge modelling, the contribution of other features is omitted if only the maximum value is taken at each time.

2.1.1.3 Information Content-based measures

Also acknowledging some of the limitations of edge-counting approaches, Resnik (Resnik 1995) proposed to complement the taxonomical knowledge provided by an ontology with a measure of the information distribution of concepts computed from corpora. He exploited the notion of Information Content (IC) associating to each concept of the taxonomy the corresponding probability of appearance, which is computed from their occurrences in a given corpus. Concretely, the IC of a term *a* is computed as the inverse of its probability of occurrence, $p(a)$ (15). In this manner, infrequent words are considered more informative than common ones.

$$IC(a) = -\log P(a) \tag{15}$$

According to Resnik, semantic similarity depends on the amount of shared information between two terms, a dimension which is represented by *their LCS* in an ontology. The more specific the subsumer is (higher IC), the more similar the terms are, as they share *more information*. Two terms are maximally dissimilar if a LCS does not exist (*i.e.*, in terms of edge-counting, it would not be possible to find a path connecting them). Otherwise, their similarity is computed as the IC of the LCS (16).

$$sim_{res}(a,b) = IC(LCS(a,b)) \tag{16}$$

One of the problems of Resnik's proposal is that any pair of terms having the same LCS results in exactly the same semantic similarity. Both Lin (Lin 1998b) and Jiang and Conrath (Jiang and Conrath 1997) extended Resnik's work by also considering the IC of each of the evaluated terms.

Lin enounced that the similarity between two terms should be measured as the ratio between the amount of information needed to state their commonality and the information needed to fully describe them. As a corollary of this theorem, his measure considers, on the one hand, commonality in the same manner as Resnik's approach and, on the other hand, the IC of each concept alone (17).

$$sim_{lin}(a,b) = \frac{2 \times sim_{res}(a,b)}{(IC(a) + IC(b))} \tag{17}$$

The measure proposed by Jiang and Conrath is based on quantifying, in some way, the length of the taxonomical links as the difference between the IC of a concept and its subsumer. When comparing term pairs, they compute their distance by substracting the sum of the IC of each term alone from the IC of their LCS (18).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$dis_{j\&c}(a,b) = (IC(a) + IC(b)) - 2 \times sim_{res}(a,b) \quad (18)$$

The coherence of the IC computation with respect to the taxonomical structure is an aspect that should be ensured in order to maintain the consistency of the similarity computation.

It is important to note that IC-based measures need, in order to behave properly, that the probability of appearance *p* monotonically increases as one moves up in the taxonomy (*i.e.*, $\forall\, c_i \mid c_j$ *is hypernym of* c_i => $p(c_i) \leq p(c_j)$). This will ensure that the subsumer's IC must be lower than its specializations. This is achieved by computing *p(a)* as the probability of encountering *a* and any *specializations* of *a* in the given corpus. In practice, each individual occurrence of any word in the corpus is counted as an occurrence of each taxonomic class containing it (19) (Resnik 1995). So, this approach forces the recursive computation of all the appearances of the subsumed terms to obtain the IC of the subsumer.

$$p(a) = \frac{\sum_{w \in W(a)} count(w)}{N} \quad (19)$$

where $W(a)$ is the set of words in the corpus whose senses are subsumed by *a*, and *N* is the total number of corpus words that are present in the taxonomy.

As a result, an accurate computation of concept probabilities requires a proper disambiguation and annotation of each noun found in the corpus. If either the taxonomy or the corpus changes, re-computations are needed to be recursively executed for the affected concepts. So, it is necessary to perform a manual and time-consuming analysis of corpora and resulting probabilities would depend on the size and nature of input corpora.

Only words in small to moderate size corpora can be manually tagged with the sense for each word in the corpus. This arises an important drawback: the sparseness of the data that will support the similarity assessment, which may not be enough to extract valid conclusions about the information distribution.

Moreover, the background taxonomy must be as complete as possible (*i.e.*, it should include most of the specializations of each concept) in order to provide reliable results. Partial taxonomies with a limited scope may not be suitable for this purpose. All those aspects limit the scalability and applicability of those approaches.

2.1.1.3.1 Intrinsic computation of IC

Considering the limitations of IC-based approaches due to their dependency on corpora, some authors tried to intrinsically derive IC values from an ontology. Those works rely on the assumption that the taxonomic structure of ontologies like WordNet is organized in a meaningful way, according to the principle of *cognitive saliency* (Blank 2003). This states that humans specialise concepts when they need to differentiate them from already existing ones. So, concepts with many hyponyms (*i.e.*, specializations) provide less information than the concepts at the leaves of the hierarchy. From the Information theory point of view, they consider that abstract

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

ontological concepts appear more probably in a corpus as they subsume many other ones. In this manner, they estimate the probability of appearance of a concept and in consequence, the amount of information that a concept provides, as a function of the number of hyponyms and/or their relative depth in the taxonomy.

Seco *et al*., (Seco, Veale et al. 2004) and Pirró and Seco (Pirró and Seco 2008) base IC calculations on the number of hyponyms. Being *hypo(a)* the number of hyponyms (i.e. descendents) of the concept *a*, *p*(*a*) the probability of occurrence of *a* and *max_nodes* the maximum number of concepts in the taxonomy, they compute the IC of a concept in the following way (20):

$$IC_{seco}(a) = -\log(p(a)) = 1 - \frac{\log(hypo(a)+1)}{\log(max\_nodes)} \quad (20)$$

The denominator (corresponding to the root on the taxonomy, that is the most informative concept) ensures that IC values are normalized in the range 0..1.

This approach only considers hyponyms of a given concept in the taxonomy; so, concepts with the same number of hyponyms but different degrees of generality appear to be equally similar. In order to tackle the problem, and in the same manner as for edge-counting measures, Zhou *et al*., (Zhou, Wang et al. 2008) proposed to complement the hyponym-based IC computation with the relative depth of each concept in the taxonomy. Then, the IC of a concept is computed as (21):

$$IC_{zhou}(a) = k\left(1 - \frac{\log(hypo(a)+1)}{\log(max\_nodes)}\right) + (1-k)\left(\frac{\log(deep(a))}{\log(max\_depth)}\right) \quad (21)$$

Here, *hypo* and *max_nodes*, which have the same meaning as eq. 20. *deep(a)* corresponds to the depth of the concept *a* in the taxonomy and *max_depth* is the maximum depth of the taxonomy. Factor *k* adjusts the weight of the two features involved in the IC assessment. Authors proposal is *k*=0.5.

Both ways of computing intrinsically the IC have been applied directly on the similarity functions proposed by Resnik, Lin and Jiang and Conrath. Those approaches overcome most of the problems observed for corpus-based IC approaches (specifically, the need of corpus processing and their high data-sparseness). As it will be shown in the evaluation section, the results indicate that they have a similar, or even better, accuracy than corpus-based IC calculation when applied over WordNet. However, they require big, detailed and fine grained taxonomies/ontologies in order to enable an accurate estimation of a concept's IC. For small or very specialized ontologies with a limited taxonomical depth and low branching factor, the resulting IC values between concepts would be too homogenous to enable a proper differentiation.

2.1.2 Distributional approaches

On the contrary to ontology-based measures, distributional approaches only use text corpora as source to infer the semantics of the terms. They are based on the

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

assumption that words with similar distributional properties have similar meanings (Waltinger, Cramer et al. 2009); so, they infer semantic likeness from word co-occurrences in text corpora. As words may co-occur due to many different reasons, distributional measures capture the more general sense of *relatedness* in contrast to taxonomically-based *similarity* measures.

According to the way in which distributional resemblance is determined, one may distinguish two different approaches. On one hand, some authors measure similarity from direct word co-occurrence in text (first order co-occurrence). On the other hand, other authors estimate relatedness as a function of the similarity of the contexts in which words occur (second order co-occurrence). In this section, we survey the main proposals of each kind.

2.1.2.1 First order co-occurrence

First order co-occurrence approaches rely on the principle that the more frequently two words appear together, the higher their relatedness. This follows the simple cognitive principle that people would judge two words as similar because they are exposed to them simultaneously (Lemaire and Denhière 2006). This is a controversial principle which has been questioned by some authors, but there are many works based on it. We will briefly present some of them and we will discuss advantages and drawbacks at the end of the section

Being completely corpora-dependant, the choice of input data is crucial for these methods. In order to extract reliable conclusions, the corpus should be as representative as possible with regards to the real social-scale information distribution. For practical reasons, the analysis is restricted to textual sources, mainly due to the fact that people usually learn words from texts (Landauer and Dumais 1997). As a result, the text corpora size and heterogeneity are important dimensions for being able to capture global-scale knowledge.

The Web, being the biggest electronic repository currently available, created from the interaction of a big community of users, represents one of the best options to apply those measures(Sánchez and Moreno 2008a). In fact, unsupervised models perform better when statistics are obtained from the Web rather than from other large corpora (Keller and Lapata 2003).

First order approaches estimate relatedness as a function of the probability of co-occurrence of two terms in relation to individual probabilities. As computing absolute term appearances in the Web is very time consuming, authors associate probabilities to page counts provided by Web search engines. It is important to note that those engines estimate the number of appearances of a given query in individual documents but not the total amount of appearances (*e.g.*, in case of several appearances per document).

Pointwise Mutual Information (PMI) was one of the first functions to be adapted to the Web to compute term appearance probabilities from the Web page count (Turney 2001). It is defined as the comparison between the probability of observing *a* and *b* together (estimated from the *page count* of the query '*a* AND *b*') and observing them independently (estimated from the *page count* when querying *a* and *b* alone). Turney work is based on the idea that when *a* and *b* are not statistically independent, they will

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

have a tendency to co-occur (which is the case of words in a corpus) and the numerator will be greater than the denominator. Therefore, the resulting ratio (22) is considered as a valid a measure of the degree of statistical dependency between *a* and *b* (Turney 2001).

Let *H*(*a*) and *H*(*b*) denote the page count (*i.e.*, hits) provided by a search engine when querying '*a*' and '*b*' , respectively. Let *H*(*a*,*b*) be the page count when the query is '*a* AND *b*'. Let *M* be the total number of pages indexed by the search engine.

$$PMI(a,b) = -\log\left(\frac{\frac{H(a,b)}{M}}{\frac{H(a)}{M}\frac{H(b)}{M}}\right) \tag{22}$$

Cilibrasi and Vitanyi (Cilibrasi and Vitányi 2006) proposed a distance between words based on Information Theory, using also the page counts of Web search engines. It is defined as the normalized information distance between two words. The function, named *Normalised Google Distance* (NGD) is defined as follows (23):

$$NGD(a,b) = \frac{\max(\log H(a), \log H(b)) - \log H(a,b)}{\log M - \min(\log H(a), \log H(b))} \tag{23}$$

Bollegala (Bollegala, Matsuo et al. 2007) adapted several classical co-occurrence measures: Jaccard (24), Overlap (Simpson) (25), Dice (26) and the mentioned PMI in a similar way as Turney did (27).

$$WebJaccard(a,b) = \begin{cases} 0 & if\ H(a,b) \le \lambda \\ \dfrac{H(a,b)}{H(a)+H(b)-H(a,b)} & otherwise \end{cases} \tag{24}$$

$$WebOverlap(a,b) = \begin{cases} 0 & if\ H(a,b) \le \lambda \\ \dfrac{H(a,b)}{\min(H(a),H(b))} & otherwise \end{cases} \tag{25}$$

$$WebDice(a,b) = \begin{cases} 0 & if\ H(a,b) \le \lambda \\ \dfrac{2H(a,b)}{H(a)+H(b)} & otherwise \end{cases} \tag{26}$$

$$WebPMI(a,b) = \begin{cases} 0 & if\ H(a,b) \le \lambda \\ \log_2\left(\dfrac{\frac{H(a,b)}{M}}{\frac{H(a)}{M}\frac{H(b)}{M}}\right) & otherwise \end{cases} \tag{27}$$

In order to minimize the influence of noise existing in Web data, they set each coefficient to zero if the page count for the query *a* AND *b* is less than a threshold (λ

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

= 5 was used in (Bollegala, Matsuo et al. 2007)). This omits some cases of random co-occurrences and misspelled terms. *M* is estimated as 10^{10} according to the number of indexed pages reported by Google in 2007.

Instead of using the absolute value of page counts for a given query, Chen *et al.*, (Chen, Lin et al. 2006) rely on the amount of co-occurrences observed in some, apparently, more reliable resources. They propose to use the short texts presented by the search engine in the first positions of the results list, called snippets. Snippets are brief windows of text extracted by a search engine around the query term in a document and provide, in a direct manner, a local context for the queried term. Snippet processing is very efficient when compared to the cost of accessing and downloading individual web documents. For two terms *a* and *b*, they collect a fixed number of snippets provided by the search engine when querying each term. Then, they count the number of occurrences of *a* in the snippets of *b f(a@b)* and vice-versa *f(b@a)*. The two values are combined in a non-linear fashion to compute their relatedness (with a function called CODC).

$$CODC(a,b) = \begin{cases} 0 & if \;\; f(a@b) = 0 \quad or \quad f(b@a) = 0 \\ e^{\log\left(\frac{f(b@a)}{f(a)} \times \frac{f(a@b)}{f(b)}\right)^{\alpha}} & otherwise \end{cases} \tag{28}$$

where *f* represents the number of occurrences of the corresponding term in the top *N* snippets returned by the search engine when querying the term. The constants α=0.15 and *N*=600 were used in their experiments (Chen, Lin et al. 2006).

This approach heavily depends on the Web search engine ranking algorithm and the fact that only a subset of snippets can be processed (*i.e.*, most search engines only provide access to the first 1000 web resources for a given query). Therefore, there is no guarantee that the evidence needed to support the semantic assessment for a pair of terms is contained in the top-ranked snippets. As a result, even though this method is able to provide relatively reliable results for common and related terms, it suffers from high data sparseness due to the locality of the analysis.

In a more elaborated approach, Bollegala *et al.*, (Bollegala, Matsuo et al. 2007) also relied on snippets obtained when querying both terms, *a* and *b*, at the same time. Snippets are used as co-occurrence context, and lexical patterns (n-grams in a window from 2 to 5 words), evidencing the co-occurrence of *a* and *b*, are extracted. The most reliable patterns according to a predefined list are selected, and the number of their appearances is normalized. They create a feature vector using 200 patterns and the web scores for *a* and *b* computed from the functions: Web-Dice (26), Web-overlap (25), Web-Jaccard (24) and Web-PMI (27) stated above. The vector is created for a pre-tagged set of synonym and non-synonym word pairs and a Support Vector Machine (SVM) is trained accordingly. The trained SVM is then used to classify new word pairs using the same vector-based procedure. Semantic relatedness (referred with the name *SemSim*) is computed as the posterior probability *Prob(F|synonymous)* that the obtained feature vector *F* belongs to the synonymous-word class (29).

$$SemSim(a,b) = Prob(F \mid synonymous) \tag{29}$$

In Bollegala *et al.*, (Bollegala, Matsuo et al. 2009) the same authors modified their measure by: *1)* introducing an algorithm to select the most reliable lexical patterns

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

according to a set of semantically related words which are used as training data, and *2)* clustering semantically related patterns into groups in order to overcome data sparseness of a fine-grained pattern list and reduce the number of training parameters. As a result, two words are represented by a feature vector defined over the clusters of patterns. Semantic relatedness is computed as the Mahalanobis (Everitt, Landau et al. 2001) distance between the points of the feature vectors.

It is important to note that Bollegala *et al.*,'s supervised measures rely on pre-tagged data and trained classifiers. This introduces many limitations such as the fact that manually tagged training data should be available and that this data should be general and big enough to avoid the risk that the classifier could be over-fitted by them. As a result, the same problems noted for IC corpus-based measures can be noted in this case.

In general, the main advantage of co-occurrence-based approaches is that, relying uniquely on the Web, they do not need any knowledge source to support the assessment. Thanks to the Web coverage of almost any written word, they can be applied to terms that are not typically considered in ontologies such as named entities. However, their unsupervised nature and their reliance on search engine page counts introduce several drawbacks. Firstly, word co-occurrence estimated by page-counts omits the semantic dimension of the co-occurrence. Words may co-occur because they are taxonomically related, but also because they are antonyms or by pure chance. So, page counts (without considering the relative positions of words in the document) give a rough estimation of statistical dependency. Secondly, page counts deal with words rather than concepts (on the contrary to ontological features). Due to the ambiguity of language and the omission of word context analysis during the relatedness assessment, polysemy and synonymy may negatively affect the estimation of concept probability by means of word appearance frequency. Polysemic words associated to a concept causes that their page counts contain a combination of all their senses. Moreover, the presence of synonyms for a given concept causes that word page counts underestimate the real concept probability. Finally, as stated above, page counts may not be necessarily equal to word frequency because the queried word might appear several times on a Web resource. Due to these reasons, some authors have questioned the usefulness of page counts alone as a measure of relatedness (Bollegala, Matsuo et al. 2007).

Some authors (Lemaire and Denhière 2006; Bollegala, Matsuo et al. 2007) also questioned the effectiveness of relying on first order co-occurrences as a measure of relatedness. Studies on large corpora gave examples of strongly associated words that never co-occur (Lund and Burgess 1996). This situation is caused, in many cases, by the fact that both words tend to co-occur with a third one. Psycholinguistics researchers have shown that, in those cases, the association between two words is done by means of a third word (Livesay and Burgess 1998). This is called a *second-order co-occurrence* (Lemaire and Denhière 2006), which is precisely the principle of the approaches reviewed in the following section.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2.1.2.2 Second order co-occurrence

Second order co-occurrence measures are based on the principle that two words are similar to the extent that their contexts are similar (Harris 1985). The definition of context may vary from one measure to another and might be considered a small or large window around a word occurrence or an entire document.

A classical approach based on this principle is Latent Semantic Analysis (LSA) (Deerwester, Dumais et al. 2000). It consists on compiling a term context matrix containing the occurrences of each word in each context. A Singular Value Decompositions (SVD) process is performed to enhance the differences between reliable and unreliable extractions. Considering word context as vectors, the final distance between words is computed as the cosine of the angle between them.

The most common approach in this type of measures is to construct a co-occurrence vectors that represents the contextual profile of a concept. These vectors are built by extracting contextual words (within a fixed window of context) from a corpus of textual documents covering the evaluated concepts. Again, these vectors capture a more general sense of concept likeness, not necessarily reduced to taxonomical similarity but also to inter-concept relatedness.

Using the Web as corpus, Sahami and Heilman (Sahami and Heilman 2006) computed the likeness between two terms by means of snippets returned when querying those terms in a search engine. Authors process each snippet and represent it as a TF-IDF (inverted term frequency in a document) weighted word vector. The centroid of the set of vectors obtained by querying each term is defined, and the relatedness between two terms is computed as the inner product between the corresponding centroids.

Even though using the Web as a corpus and search engines as middlewares has several advantages derived from the Web's size and heterogeneity, some authors have criticized their usefulness as a support for relatedness computation. In fact, while semantic relatedness is inherently a relation on concepts, Web-based approaches measure a relation on words (Budanitsky and Hirst 2006) . Big-enough sense-tagged corpora is needed to obtain reliable concept distributions from word senses, much like corpus-based IC measures needed in the past. However, due to the nature of the Web, it is not feasible to have such tagged corpora, at least until the Semantic Web (Berners-Lee, Hendler et al. 2001) becomes a reality. Moreover, ontology-based measures rely on pre-defined knowledge sources manually created by human experts, which one may consider to be true and unbiased. As stated above, commercial bias, spam, noise and data sparseness are problems that may affect distributional measures when using the Web as corpus.

In order to overcome those problems, some authors preferred to apply distributional hypotheses over more reliable corpora. Concept glosses from wide thesaurus like WordNet were exploited. Words appearing in a gloss are likely to be more relevant for the concept's meaning than text drawn from a generic corpus and, in consequence, may represent a more reliable context. Based on the distributional hypothesis, the author defend the idea that if two terms have similar glosses (*i.e.*, their textual descriptions overlap), they are likely to have similar meanings (Banerjee and Pedersen 2003).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Banerjee and Pedersen (Banerjee and Pedersen 2003) presented the Extended Gloss Overlap (EGO) measure (30), which determines the relatedness of terms as a function of the overlap of their WordNet glosses. As synset glosses in WordNet tend to be rather short, they extended the gloss by including example sentences (also provided by WordNet) and glosses of related concepts directly linked by means of a semantic relation.

$$EGO(a,b) = score(gloss(a), gloss(b)) + score(hyper(a), hyper(b)) + \\ + score(hypo(a), hypo(b)) + score(hyper(a), gloss(b)) + score(gloss(a), hyper(b)) \quad (30)$$

where *score()* is the function that find the phrases that overlap between two glosses and returns an score as defined in (Banerjee and Pedersen 2003); *hypo(a)* and *hyper(a)* represent respectively hyponyms and hypernyms of *a* in the given ontology.

Patwardhan and Pedersen (Patwardhan and Pedersen 2006) also used extended WordNet glosses as corpora to retrieve co-occurrence information for term contexts, creating gloss vectors (GV). Gloss vectors are constructed considering gloss words that are not a stop word and whose occurrence is above a minimum frequency. Due to the size of WordNet and the extension of glosses (which consist on approximately 1.4 million words once low frequency and stop words are removed), vectors are defined in a space of 20,000 dimensions. The relatedness between two words is defined as the cosine of the angle between gloss vectors (31).

$$GV(a,b) = \frac{\vec{v}_a \cdot \vec{v}_b}{\left|\vec{v}_a\right| \cdot \left|\vec{v}_b\right|} \quad (31)$$

where $\vec{v}_a$ and $\vec{v}_b$ are the context vectors corresponding to *a* and *b* respectively.

The Gloss Vector measure presents some advantages over the Extended Gloss Overlap, as the later looks for exact string overlaps as a measure of relatedness. Gloss Vector does not rely on exact matches by using vectors that capture the contextual representation of concepts.

Wan and Angryk (Wan and Angryk 2007) identified some weaknesses of Patwardhan and Pedersen's measure and proposed and new Context Vector measure based on a similar principle. They used related synsets instead of glosses to augment the context vector of a term. They join different elements: (1) the term synset, (2) synsets having direct semantic relations to the concerned term synset and (3) all direct and inherited hypernyms. In order to limit the vector space, they remove senses with a frequency of appearance lower than a threshold. Finally, the cosine of the angle between vectors is used as relatedness measure as in the previous formula (30).

As it will be discussed in the evaluation section, the use of reliable glosses instead of the Web as corpora, results in a significant improvement of accuracy. However, the computational complexity is a factor that hampers those measures since the creation of context vectors in such a big dimensional space is difficult. Moreover, the quality of the words used as the dimensions of these vectors greatly influences the accuracy of the results. Big differences were observed by the authors (Patwardhan and Pedersen 2006) when changing the frequency cut-off for scarce senses. Finally, by relying on WordNet glosses, those measures are hardly applicable to other ontologies

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

in which glosses or textual descriptions are typically omitted (Ding, Finin et al. 2004). In fact, Pedersen *et al.*, (Pedersen, Pakhomov et al. 2007) applied the Gloss Vector measure to the biomedical domain by exploiting the SNOMED-CT repository as ontology. Due to the lack of concept glosses, they required a time-consuming process of manual compilation and processing of a large set of medical diagnoses from which to extract term descriptions. In that case, the algorithm parameters, such as the choice and size of corpora, had a very notorious influence in the results.

2.2 Contributions on semantic similarity

In this section, we present two different approaches to the computation of semantic similarity that overcome some of the problems identified in the study presented above.

First, a new pure ontology-based 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.

Second, a new way to measure the information content (IC) by exploiting the Web information distribution instead of tagged corpora is proposed. 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. Using this approach, the modified versions of some IC-based similarity measures are presented.

Those measures are evaluated and compared in section 2.3 and their applicability for semantic clustering is analysed in chapter 4.

2.2.1 A new measure to compute the semantic similarity

From the study of similarity measures described in previous sections, the following conclusions can be extracted. From the applicability point of view, those measures that only exploit the geometrical model of the ontology (*e.g.*, path based) are the most adequate ones as no pre-calculus or pre-processing is needed, which makes them more computationally efficient. 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 propose a new similarity measure that can achieve a level of accuracy similar to corpus-based approaches but retaining the low

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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. So, it seems reasonable that a measure that takes 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:

$$\mathcal{A}(c_i)=\{c_j \in C \mid c_j \text{ is superconcept of } c_i\} \cup \{c_i\} \quad (32)$$

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.

$$d(x_i, x_j) = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \quad (33)$$

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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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.

$$d(c_i, c_j) = \left|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)\right| - \left|\mathcal{A}(c_i) \cap \mathcal{A}(c_j)\right| \quad (34)$$

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. 3, 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).

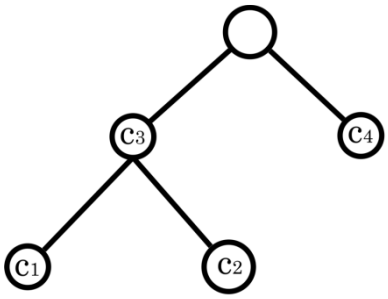


Figure. 3. Taxonomy example

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.

Definition 2: SuperConcept-based distance (SC)

$$SC(c_i, c_j) = \frac{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)|}{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)|} \quad (35)$$

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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.

Euclidian 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:

$$d(c_i, c_j) = d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (36)$$

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

$$SC_{Eu}(c_i, c_j) = \sqrt{\frac{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)|}{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)|}} \quad (37)$$

It is worth to note that it (36) 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). Details are given in Annex A.

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).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Definition 4: Logarithmic SuperConcept-based similarity (SC_{log})

$$SC_{log}(c_i, c_j) = -log_2 \frac{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)|}{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)|} \quad (38)$$

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.

2.2.1.1 Dealing with polysemic terms

In this section we study the problem of polysemic words that can be found in the databases that are going to be analysed with clustering techniques (as it is the goal of this thesis). 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 5: The generalized distance measure which is able to deal with polysemic terms is defined as:

$$SC_{generalize\ d}(a, b) = \min_{\substack{\forall a' \in A \\ \forall b' \in B}} SC(a', b') \quad (39)$$

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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 (37) (38).

2.2.2 A new approach to compute IC from the Web

As it has been explained in section 2.1.1.3, 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.

In this section we propose a method to overcome these limitations. In order to minimize the corpus dependency and, in consequence, the coverage limitations of IC-based measures, we will not rely on pre-processed data. In fact, 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.

In order to achieve reliable similarity estimations from the Web without manual pre-tagging or explicit disambiguation, our approach proposes a new way of computing concept IC from word appearances in text in a taxonomically coherent manner (*i.e.*, monotonically increasing as concepts are specialized) and minimizing language ambiguity.

2.2.2.1 Computing IC from a general corpus: the Web

The corpus-dependency of IC-based measures (when IC is computed from a corpus and not intrinsically) introduces limitations about the applicability of the measures as general purpose similarity assessor (a corpus with a wide coverage with respect to the domain of the evaluated terms must be available). Data sparseness (i.e. the fact that not enough tagged data is available for certain concepts to reflect an appropriate semantic evidence) is the main problem (Brill 2003). Moreover, the preparation of the input data for each domain in a format which can be exploited (i.e. data pre-processing or filtering) is typically required.

Ideally, the robustness of the semantic evidence may be increased by using a bigger and more general corpus like the Web. The Web can be considered as a social-scale general purpose corpus for its size and heterogeneity. Its main advantages are:

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

(1) its free and direct access and (2) its size. The Web offers more than 1 trillion of accessible resources which are directly indexed by web search engines[2] (compared to the 100 passages of SemCor (Miller, Leacock et al. 1993)).

It has been demonstrated (Brill 2003) the convenience of using such a wide corpus to improve the sample quality for statistical analysis. Concretely, the amount and heterogeneity of information in the Web are so high that it can statistically approximate the real distribution of information (Cilibrasi and Vitányi 2006).

The problem is that the analysis of such an enormous repository for computing concept appearances is impracticable. However, the availability of massive Web Information Retrieval tools can help in this purpose. The frequency of page counts returned by the search engine divided by the number of indexed pages can be used to estimate the probability of appearance of a term. In fact, (Cilibrasi and Vitányi 2006) claim that the probabilities of Web search engine terms approximate the relative frequencies of those searched terms as actually used in society. So, exploiting Web Information Retrieval (IR) tools and concept's usage at a social scale as an indication of its generality, one can estimate the concept probabilities from web hit counts (Turney 2001). As stated in section 2.1.2, some authors (Turney 2001; Cilibrasi and Vitányi 2006), have exploited web information distribution in order to evaluate word relatedness in an unsupervised fashion (i.e. no domain knowledge is employed).

We propose to estimate the IC of a concept from the Web with the ratio presented in Definition 6 (Sánchez, Batet et al. 2009; Sánchez, Batet et al. 2010a; Sánchez, Batet et al. 2010b).

Definition 6: *Web-based Information Content* (*IC_IR*) of a concept '*a*' is defined as:

$$IC\_IR(a) = -\log_2 p_{web}(a) = -\log_2 \frac{H(a)}{M} \quad (40)$$

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, IC-based measures presented in section 2.1.1.3 can be directly rewritten by incorporating the Web-based IC computation (IC_IR).

Based on that, Resnik measure can be rewritten as follows.

$$sim_{res}\_IR(a,b) = IC\_IR(LCS(a,b)) = -\log\frac{H(LCS(a,b))}{M} \quad (41)$$

Lin measure can be rewritten as follows.

[2] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$sim_{lin}\_IR(a,b) = \frac{2 \times sim_{res}\_IR(a,b)}{\left(IC\_IR(a) + IC\_IR(b)\right)} = \frac{2 \times \left(-\log \frac{H\left(LCS(a,b)\right)}{M}\right)}{\left(-\log \frac{H(a)}{M} - \log \frac{H(b)}{M}\right)} \quad (42)$$

Jiang & Conrath distance measure can be rewritten as follows.

$$dis_{jcn}\_IR(a,b) = \left(IC\_IR(a) + IC\_IR(b)\right) - 2 \times sim_{res}\_IR(a,b) =$$
$$= \left(-\log \frac{H(a)}{M} - \log \frac{H(b)}{M}\right) - 2x\left(-\log \frac{H\left(LSC(a,b)\right)}{M}\right) \quad (43)$$

However, 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, as shown in Figure 4, 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. As mentioned in section 2.1.1.3, 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*[3]), 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). As mentioned in section 2.1.1.3, 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

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

An example

In order to illustrate the poor estimation of the IC_IR, and its effects in the similarity assessment, let us consider the taxonomy presented in Figure 4 and the estimated term Web appearances obtained from queries performed over the Bing[4] web search engine (more details on the convenience of using this search engine will be provided in the evaluation section).

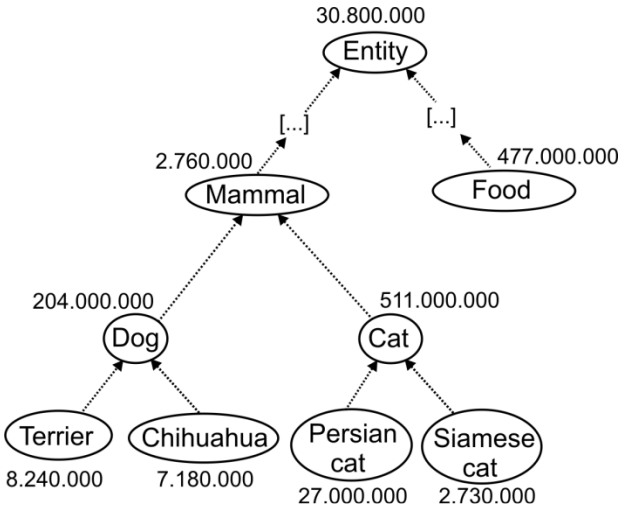


Figure 4. Portion of an example taxonomy (with *Entity* as the root concept) and occurrence values of concept's terms in the Web computed from Bing hit count [Accessed: Nov. 9th, 2008].

We have applied the Resnik similarity measure to this taxonomy introducing *IC_IR* (eq. 40), using as concept appearances the Web term hit count presented in Figure 4. The final concept probabilities are obtained by dividing term appearances by the total amount of indexed resources in the corpus, which in the case of the Web we have considered to be 10^{12}. The similarity between some pairs of concepts has been estimated. First two dog breeds are compared: *Terrier* and *Chihuahua*. Then, the similarity between *Chihuahua* and *Persian Cat* is calculated. In the latter case, the most specific concept that generalizes them is *Mammal*, according to the ontology given in Figure 4. The following results have been obtained:

$$sim_{res}\_IR(terrier, chihuahua) = IC\_IR(LSC(terrier, chihuahua)) = IC\_IR(dog) =$$

$$= -\log_2 \frac{hits(dog)}{total\_webs} = -\log_2 \frac{204\times10^6}{10^{12}} = 12.25$$

$$sim_{res}\_IR(chihuahua, persian\_cat) = IC\_IR(LSC(chihuahua, persian\_cat)) =$$

$$= IC\_IR(mammal) = -\log_2 \frac{hits(mammal)}{total\_webs} = -\log_2 \frac{2{,}76\times10^6}{10^{12}} = 18.46$$

[4] Bing search engine (http://www.bing.com).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

in consequence, we will erroneously conclude that

$$sim_{res}\_IR(chihuahua, terrier) < sim_{res}\_IR(chihuahua, persian\_cat)$$

, because, contrarily to what it is expected in a subsumption hierarchy, *IC_IR(mammal) > IC_IR(dog).*

The non-monotonic IC_IR values affect even more to measures in which the IC of the LCS is compared against the IC of the evaluated concepts, producing incorrect results (out of range values). Concretely, evaluating the similarity using Lin measure and IC_IR, we obtain the following results:

$$sim_{lin}\_IR(terrier, chihuahua) = \frac{2\times(IC\_IR(dog))}{(IC\_IR(terrier)+IC\_IR(chihuahua))} =$$

$$= \frac{2\times\left(-\log_2\left(\frac{204\times10^6}{10^{12}}\right)\right)}{-\log_2\frac{8.24\times10^6}{10^{12}}-\log_2\frac{7.18\times10^6}{10^{12}}} = 0.72$$

$$sim_{lin}\_IR(chihuahua, persian\_cat) = \frac{2\times(IC\_IR(mammal))}{(IC\_IR(chihuahua)+IC\_IR(persian\_cat))} =$$

$$= \frac{2\times\left(-\log_2\left(\frac{2.76\times10^6}{10^{12}}\right)\right)}{-\log_2\frac{7.18\times10^6}{10^{12}}-\log_2\frac{27\times10^6}{10^{12}}} = 1.14$$

As the similarity between *chihuahua* and *persian cat* has been incorrectly assessed (with a value above 1), we will erroneously conclude that

$$sim_{lin}\_IR(terrier, chihuahua) < sim_{lin}\_IR(chihuhaua, persian\_cat)$$

Applying IC_IR to Jiang & Conrath dissimilarity measure, the same problem appears:

$$dis_{jcn}\_IR(terrier, chihuahua) =$$

$$= (IC\_IR(terrier)+IC\_IR(chihuahua))-2\times IC\_IR(dog) = 9.45$$

$$dis_{jcn}\_IR(chihuahua, persian\_cat) =$$

$$= (IC\_IR(chih)+IC\_IR(pers))-2\times IC\_IR(mammal) = -4.66$$

In this case, the incorrectly assessed dissimilarity between *Chihuahua* and *Persian cat* results in a negative value, erroneously concluding that

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$dis_{jcn\_IR}(terrier, chihuahua) > dis_{jcn\_IR}(chihuahua, persian\_cat)$$

To solve these problems, in the next section, we propose a new way of estimating the information content of concepts from Web hit counts.

2.2.2.2 Contextualized Information Content from the Web

In order to avoid the incorrect results shown in the previous section we have redefined the way in which concept probabilities for IC computation are estimated from the Web. Applying Resnik's approach to the Web (i.e. recursively adding specialized concept appearances to their subsumers) as shown in section 2.2.2.1, introduces problems about scalability (because a large number of web queries is required), as well as, a heavy dependence to both the ontology and corpus modifications. In order to tackle those issues, 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 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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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 7: 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:

$$CIC_T\_IR(a_b) = -\log_2 p_{web}(a_b) = -\log_2 \frac{H(a\ AND\ LCS_T(a,b))}{M} \quad (44)$$

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:*

$$CIC_T\_IR(b_a) = -\log_2 p_{web}(b_a) = -\log_2 \frac{H(b\ AND\ LCS_T(a,b))}{M} \quad (45)$$

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.

$$IC\_IR(LCS(a,b)) \leq \min(CIC_T\_IR(a_b), CIC_T\_IR(b_a)) \quad (46)$$

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*.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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.

2.2.2.2.1 Applying CIC_T_IR to IC-based measures

As Resnik similarity measure only considers the occurrence of the LCS in a corpus and not the IC of the evaluated concepts, CIC_T_IR cannot be directly applied. For measures like Lin or Jiang & Conrath, which evaluate the difference between the IC of subsumed terms against their LCS, the introduction of CIC_T_IR can aid to obtain a taxonomically coherent, less ambiguous an more accurate similarity assessment from the Web. More details will be given in the evaluation section. The proposed contextualized versions of Lin and Jiang & Conrath functions are defined below (Sánchez, Batet et al. 2010b).

Definition 8: 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:

$$sim_{lin\_CIC_T\_IR}(a,b) = \frac{2 \times IC\_IR(LCS_T(a,b))}{(CIC_T\_IR(a_b) + CIC_T\_IR(b_a))} = \\ = \frac{2 \times \left(-\log_2 \frac{H(LCS_T(a,b))}{M} \right)}{\left(-\log_2 \frac{Hs(a\ AND\ LCS_T(a,b))}{M} - \log_2 \frac{H(b\ AND\ LCS_T(a,b))}{Ms} \right)} \quad (47)$$

Definition 9: 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:

$$dis_{jcn\_CIC_T\_IR}(a,b) = (CIC_T\_IR(a_b) + CIC_T\_IR(b_a)) - 2 \times IC\_IR(LCS_T((a,b)) = \\ = \left(-\log_2 \frac{H(a\ AND\ LCS_T((a,b))}{M} - \log_2 \frac{H(b\ AND\ LCS_T((a,b))}{M} \right) - 2 \times \left(-\log_2 \frac{H(LCS_T((a,b))}{M} \right) \quad (48)$$

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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2.2.2.3 Dealing with polysemy and synonymy

Typical domain ontologies are unambiguous (Dujmovic and Bai 2006) (i.e. a unique LCS represented by one textual form is available for any pair of concepts). However, general purpose ontologies, such as WordNet, typically implement *polysemy* by representing several *is-a* relationships for the same concept and *synonymy* by associating a list of semantically equivalent terms to each sense (*synsets*). In the former case, several LCS may exist for different taxonomical classifications of a given pair of terms; in the latter case, several textual forms for each LCS may be available.

IC-based measures tackle polysemy by in the same manner as edge-counting approaches, using the *Most Specific Common Subsumer* (MSCS) which corresponds to the LCS with the highest IC value (Resnik 1995) (*i.e.* for a pair of terms, they consider the pair of most similar senses represented by the MSCS). They take all the possible subsumers, compute the similarity for each of them and take the maximum value as final result (the minimum for dissimilarity measures). Synonyms associated to LCSs are not a problem in their approach because the background corpus used by those measures incorporates frequencies of concepts (WordNet *synsets*) rather than words.

In the framework proposed in this work, in which a general ontology and corpus can be also used, these two issues must be considered. For polysemic cases the strategy will be the same as presented in section 2.2.2.3: 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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2.3 Evaluation

This section presents an evaluation of the similarity measures reviewed in this work as well as our proposed measures. The objective of this section is twofold: (1) to objectively compare the different semantic similarity/relatedness paradigms and discuss their performance and (2) show the validity of our proposals when compared with the state of the art.

In this section three experiments are presented: (1) the performance of *SC* in a general domain is analyzed (Section 2.3.2); the performance of *SC* in the biomedical domain is analyzed (section 2.3.3); (3) the performance of the IC_T_IR in a general domain is presented (section 2.3.4).

In all cases, first we present the resources used in the evaluation: the used standard benchmarks and ontologies. Then, we provide the results of the tested measures and a discussion about them.

2.3.1 Evaluation criteria

As stated in (Bollegala, Matsuo et al. 2009), an objective evaluation of the accuracy of a semantic similarity function is difficult because the only criteria that can be used to evaluate the correctness of similarity is the human evaluation.

In order to enable fair comparisons, several authors have created evaluation benchmarks consisting of word pairs whose similarity were assessed by a set of humans (Rubenstein and Goodenough 1965; Miller and Charles 1991; Resnik 1995; Pedersen, Pakhomov et al. 2007).

As a result, correlation values obtained against those benchmarks (*i.e.*, the human judgments and the results of the computerized assessment) can be used to numerically quantify the closeness of two ratings sets. When, the human judgments and the computerized results are equal the correlation coefficient is 1. Spearman's and Pearson's correlations coefficients have been commonly used in the literature; both are equivalent if the rating 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. This enables a fair and objective comparison against different approaches.

We have adopted this evaluation procedure, using the Pearson's correlation (49) coefficient to compare the different similarity measures.

$$Pearson\_correlation(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (49)$$

where *X, Y* are variables, *cov(X,Y)* is the covariance and σ_X σ_Y are their respective standard deviations.

The correlation values obtained against different benchmarks created for the study of these measures will be compared.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2.3.2 Evaluation of SC in a general domain

When the terms to be compared do not belong to any specific domain (e.g. medicine, computers, marketing), we refer to them as *general domain words* or domain-independent analysis.

In such a framework, general-purpose resources are used. The most well-known ontology is WordNet, which will be explained in this section. Corpus-based similarity measurements are applied using SemCor, a hand tagged corpus consisting of 100 passages of the Brown Corpus. The Brown Corpus was compiled in the 1960s by Henry Kucera and W. Nelson Francis at Brown University, Providence, Rhode Island as a general corpus (text collection) in the field of corpus linguistics. It is a carefully compiled selection of current American English, totalling about a million words drawn from a wide variety of sources. The Corpus consists of 500 samples, distributed across 15 genres in rough proportion to the amount published in 1961 in each of those genres.

For second-coocurrence measures based on glosses, Patawardhan and Pedersen (Patwardhan and Pedersen 2006) created vectors from term glosses extracted from WordNet. Considering WordNet glosses as a corpus of contexts one obtains about 1.4 million words, which should be processed in order to create the context vectors.

2.3.2.1 Benchmarks and ontologies

Several authors created evaluation benchmarks consisting of word pairs whose similarity were assessed by a set of humans. Rubenstein and Goodenough (Rubenstein and Goodenough 1965) defined the first experiment in 1965 in which a group of 51 students, all native English speakers, assessed the similarity of 65 word pairs selected from ordinary English nouns, on a scale from 0 (semantically unrelated) to 4 (highly synonymous) (see Table 1). 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 (Table 2). The correlation obtained with respect to Rubenstein and Goodenough's experiment was 0.97. 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 (see Table 3). The correlation with respect to Miller and Charles results was 0.96.

It is interesting to see the high correlation obtained between the experiments even though being performed in a period of more than 30 years and with heterogeneous sets of people. This means that similarity between the selected words is stable over the years, making them a reliable source for comparing measures.

In fact, the benchmarks of Rubenstein and Goodenough and Miller and Charles have become *de facto* standard tests to evaluate and compare the accuracy of similarity measures.

Moreover, in order to test SC_{Eu} and SC_{log} measures WordNet ontology is used as background ontology and the presented benchmarks (in particular Miller & Charles (Miller and Charles 1991) and Rubenstein & Goodenough (Rubenstein and

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Goodenough 1965) ones). 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.

Table 1. Rubenstein and Goodenough's benchmark

Pairs of words	**Average Human Ratings**	**Pairs of words**	**Average Human Ratings**
cord smile	0.02	car journey	1.55
rooster voyage	0.04	cemetery mound	1.69
noon string	0.04	glass jewel	1.78
fruit furnace	0.05	magician oracle	1.82
autograph shore	0.06	crane implement	2.37
automobile wizard	0.11	brother lad	2.41
mound stove	0.14	sage wizard	2.46
grin implement	0.18	oracle sage	2.61
asylum fruit	0.19	bird crane	2.63
asylum monk	0.39	bird cock	2.63
graveyard madhouse	0.42	food fruit	2.69
glass magician	0.44	brother monk	2.74
boy rooster	0.44	asylum madhouse	3.04
cushion jewel	0.45	furnace stove	3.11
monk slave	0.57	magician wizard	3.21
asylum cemetery	0.79	hill mound	3.29
coast forest	0.85	cord string	3.41
grin lad	0.88	glass tumbler	3.45
shore woodland	0.90	grin smile	3.46
monk oracle	0.91	serf slave	3.46
boy sage	0.96	journey voyage	3.58
automobile cushion	0.97	autograph signature	3.59
mound shore	0.97	coast shore	3.60
lad wizard	0.99	forest woodland	3.65
forest graveyard	1.0	implement tool	3.66
food rooster	1.09	cock rooster	3.68
cemetery woodland	1.18	boy lad	3.82
shore voyage	1.22	cushion pillow	3.84
bird woodland	1.24	cemetery graveyard	3.88
coast hill	1.26	automobile car	3.92
furnace implement	1.37	midday noon	3.94
crane rooster	1.41	gem jewel	3.94
hill woodland	1.48		

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 2. Miller and Charles' benchmark

Pairs of words	Average Human Ratings	Pairs of words	Average Human Ratings
car automobile	3.92	lad brother	1.66
gem jewel	3.84	journey car	1.16
journey voyage	3.84	monk oracle	1.1
boy lad	3.76	cemetery woodland	0.95
coast shore	3.7	food rooster	0.89
asylum madhouse	3.61	coast hill	0.87
magician wizard	3.5	forest graveyard	0.84
midday noon	3.42	shore woodland	0.63
furnace stove	3.11	monk slave	0.55
food fruit	3.08	coast forest	0.42
bird cock	3.05	lad wizard	0.42
bird crane	2.97	chord smile	0.13
tool implement	2.95	glass magician	0.11
brother monk	2.82	noon string	0.08
crane implement	1.68	rooster voyage	0.08

Table 3. Resnik's benchmark

Pairs of words	Average Human Ratings	Pairs of words	Average Human
automobile car	3.9	implement crane	0.3
jewel gem	3.5	brother lad	1.2
voyage journey	3.5	car journey	0.7
lad boy	3.5	oracle monk	0.8
shore coast	3.5	rooster food	1.1
madhouse asylum	3.6	hill coast	0.7
wizard magician	3.5	graveyard forest	0.6
noon midday	3.6	slave monk	0.7
stove furnace	2.6	forest coast	0.6
fruit food	2.1	wizard lad	0.7
cock bird	2.2	smile chord	0.1
crane bird	2.1	magician glass	0.1
implement tool	3.4	string noon	0.0
monk brother	2.4	voyage rooster	0.0

Polysemous words present an average of 2.77 synsets (i.e. they belong to almost three different hierarchies) and up to 29 different senses (for the word "line").

The latest version of WordNet is 3. However, as most of the reviewed related works used version 2 (as that was the latest available at that moment), we have used this version in our tests.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2.3.2.2 Results and discussion

The values obtained with SC_{Eu} and SC_{log} have been correlated to human judgements and reposted to Table 4.

So, we have taken the correlation values originally reported by related works for Rubenstein and Goodenough's and Miller and Charles's benchmarks (when available) and reported in Table 4. In the case in which a concrete measure depends on certain parameters (such as weights or corpora selection/processing) the best correlation value reported by the authors' experiments according to optimum parameter tuning was compiled. It is important to note that, even though some of them rely on different knowledge sources (such as tagged corpora or the Web), all ontology-based ones use WordNet. WordNet 2 is the most common version used in related works. In cases in which 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. In order to evaluate our approaches under the same conditions WordNet 2 is also used.

From the results reported in Table 4, in the following, we have made a comparative analysis of the different similarity measures explained in this section. Together with their accuracy, their main advantages, drawbacks and applicability conditions will be discussed. Those are key aspects when considering the practical use of the measures, as stated in the introduction.

For ontology-based measures, the basic path length measure (Rada, Mili et al. 1989) presents the lowest accuracy (0.59) due to the fact that the absolute lengths of the paths between two concepts may not accurately represent their specificity. This is the case of WordNet, since concepts higher in the hierarchy are more general than those lower in the hierarchy (Pirró 2009). As a result, other edge-counting approaches also exploiting the relative depth of the taxonomy (Wu and Palmer (Wu and Palmer 1994), Leadcock and Chodorow (Leacock and Chodorow 1998)) offer a higher accuracy (0.74). It is remarkable the correlation values obtained by Li (Li, Bandar et al. 2003) and Al-Mubaid and Nguyen (Al-Mubaid and Nguyen 2006), which combine the length of the path with the depth of the concepts in a weighted and non-linear manner. However, they rely on empirical parameters whose values have been experimentally determined to optimize the accuracy for the evaluated benchmark, hampering their generality. Hirst and St-Onge (Hirst and St-Onge 1998) present a similar behaviour, also relying on tuning parameters but, in this case, using non-taxonomic relationships that consider a more general concept of relatedness.

Feature-based methods try to overcome the limitations of path-based measures by considering different kinds of ontological features. The problem, which has been also noted for some edge-counting measures, is their dependence on the parameters introduced to weight the contribution of each feature (for the approaches of Rodriguez and Egenhofer (Rodríguez and Egenhofer 2003) and Tversky (Tversky 1977) approaches). Correlation values are, however, very similar to those offered by edge-counting measures (0.71-0.74) in these benchmarks (except for (Rada, Mili et al. 1989)) . This can be motivated by the fact that they rely on concept features, such as synsets, glosses or non-taxonomic relationships which have secondary importance in ontologies like WordNet in comparison with taxonomical features. In fact, those kinds

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

of features are scarce in ontologies (Ding, Finin et al. 2004), which causes those approaches to be based on partially modelled knowledge. As a result, those measures, even being more complex, are not able to significantly outperform the state of the art of edge-counting measures.

Table 4. Correlation values for each measure. From left to right: authors, measure type, correlation with Miller and Charles's benchmark, correlation with Rubenstein and Goodenough's benchmark and reference in which those correlations where reported.

Measure	**Type**	**M&C**	**R&G**	**Evaluated in**
Rada *et al*., (*path length*)	Edge-counting	0.59	N/A	(Petrakis, Varelas et al. 2006)
Wu and Palmer	Edge-counting	0.74	N/A	(Petrakis, Varelas et al. 2006)
Leacock and Chodorow	Edge-counting	0.74	0.77	(Patwardhan and Pedersen 2006)
Li *et al*.,	Edge-counting	0.82	N/A	(Petrakis, Varelas et al. 2006)
Al-Mubaid and Nguyen (*sem*)	Edge-counting	N/A	0.815	(Al-Mubaid and Nguyen 2009)
Hirst and St-Onge	Edge-counting	0.78	0.81	(Wan and Angryk 2007)
Rodriguez and Egenhofer	Feature	0.71	N/A	(Petrakis, Varelas et al. 2006)
Tversky	Feature	0.73	N/A	(Petrakis, Varelas et al. 2006)
Petrakis *et al*., (*X-similarity*)	Feature	0.74	N/A	(Petrakis, Varelas et al. 2006)
Resnik	IC (corpus)	0.72	0.72	(Patwardhan and Pedersen 2006)
Lin	IC (corpus)	0.7	0.72	(Patwardhan and Pedersen 2006)
Jiang and Conrath	IC (corpus)	0.73	0.75	(Patwardhan and Pedersen 2006)
Resnik (IC computed as Seco *et al*.,)	IC (intrinsic)	N/A	0.829	(Zhou, Wang et al. 2008)
Lin (IC computed as Seco *et al*.,)	IC (intrinsic)	N/A	0.845	(Zhou, Wang et al. 2008)
Jiang and Conrath (IC computed as Seco *et al*.,)	IC (intrinsic)	N/A	0.823	(Zhou, Wang et al. 2008)
Resnik (IC computed as Zhou *et al*.,)	IC (intrinsic)	N/A	0.842	(Zhou, Wang et al. 2008)
Lin (IC computed as Zhou *et al*.,)	IC (intrinsic)	N/A	0.866	(Zhou, Wang et al. 2008)
Jiang and Conrath (IC computed as Zhou *et al*.,)	IC (intrinsic)	N/A	0.858	(Zhou, Wang et al. 2008)
Normalized Google Distance	1st ord. co-occ.	0.205	N/A	(Bollegala, Matsuo et al. 2009)
Web-Jaccard	1st ord. co-occ.	0.259	N/A	(Bollegala, Matsuo et al. 2007)
Web-Overlap	1st ord. co-occ.	0.382	N/A	(Bollegala, Matsuo et al. 2007)
Web-Dice	1st ord. co-occ.	0.267	N/A	(Bollegala, Matsuo et al. 2007)
Web-PMI	1st ord. co-occ.	0.548	N/A	(Bollegala, Matsuo et al. 2007)
Chen *et al*., (CODC)	1st ord. co-occ.	0.693	N/A	(Bollegala, Matsuo et al. 2007)
Bollegala *et al*., 2007 (*SemSim*)	1st ord. co-occ.	0.834	N/A	(Bollegala, Matsuo et al. 2007)
Bollegala *et al*., 2009	1st ord. co-occ.	0.867	N/A	(Bollegala, Matsuo et al. 2009)
Latent Semantic Analysis	2n ord. (Web)	0.72	N/A	(Seco, Veale et al. 2004)
Sahami and Heilman	2n ord. (Web)	0.579	N/A	(Bollegala, Matsuo et al. 2007)
Banerjee and Pedersen (*Extended Gloss Overlap*)	2n ord. (WordNet)	0.81	0.83	(Patwardhan and Pedersen 2006)
Patwardhan and Pedersen (*Gloss Vector*)	2n ord. (WordNet)	0.91	0.9	(Patwardhan and Pedersen 2006)
Wan and Angryk (*Context Vector*)	2n ord. (WordNet)	0.80	0.83	(Wan and Angryk 2007)
SC_{Eu} (eq. 37)	**Ontology-based**	**0.81**	**0.84**	(Batet, Valls et al. 2010a)
SC_{log} (eq. 38)	**Ontology-based**	**0.85**	**0.86**	(Batet, Sanchez et al. 2010a)

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

For IC-based measures, we observe that the approaches relying on an intrinsic computation of IC (based on the number of concept hyponyms) clearly outperform the approaches relying on corpora (0.72 vs. 0.84, in average). This is very convenient as corpora dependency seriously hampers the applicability of classical IC measures. The difference between both ways of computing IC is caused by two factors. Firstly, the data sparseness problem that appears when relying on tagged corpora (which would be necessary small due to manual tagging) to obtain accurate concept appearance frequencies. Secondly, the fact that WordNet's taxonomy is detailed and fine-grained, which enables an accurate estimation of a term's generality as a function of its number of hyponyms. With regard to the performance of each measure, Lin (Lin 1998b) tends to improve Resnik (Resnik 1995) when IC is computed intrinsically, as the former is able to differentiate terms with identical LCS but different taxonomical depths. With regard to the way in which the intrinsic IC is computed, more complex approaches also exploiting relative depth and relying on weighting parameters (Zhou *et al*., (Zhou, Wang et al. 2008)) offer the highest accuracy (0.86).

With regard to distributional approaches, unsupervised approaches relying on direct term co-occurrences computed from Web page counts (Web-Jaccard, Web-Overlap, Web-PMI, Web-Dice and NGD) offer a limited performance (between 0.2 and 0.54). Uncontextual Web page-counts are not accurate enough to estimate reliable term resemblance due to ambiguity and noise of word Web occurrences. On the contrary, Chen *et al*., (Chen, Lin et al. 2006), and Bollegala *et al*. works (Bollegala, Matsuo et al. 2007; Bollegala, Matsuo et al. 2009) exploit snippets as contexts in which terms co-occur. In these experiments, we can see that this approach produces a less ambiguous estimation of term co-occurrence (due to their likeness) and better accuracy (0.69 for Chen *et al*.,'s approach). Bollegala *et al*.,'s works offer a noticeably high accuracy (0.83-0.86) as they rely on a supervised classifier (trained SVM) and lexical patterns to distinguish from highly similar co-occurrent words (such as synonyms) from less related ones. Even though those methods can be applied to terms that are not typically covered by ontologies (such as named entities), their dependency on manually tagged data and trained classifiers compromise their applicability.

Distributional approaches based on second order co-occurrences computed from the Web (such as LSA) improve the results of unsupervised first order approaches (0.72 vs. 0.54). Second order co-occurrences are able to capture non-directly co-occurrent words (such as synonyms) that, even though being highly related, typically co-occur by means of a third word. When a highly reliable and structured corpus such as WordNet glosses is used instead of the more general and noisy Web, the accuracy is significantly improved. In this manner, gloss vector and gloss overlap-based approaches (Banerjee and Pedersen 2003; Patwardhan and Pedersen 2006; Wan and Angryk 2007) are able to obtain correlation values among 0.8 and 0.91 in these tests. In fact, the Gloss Vector approach reported the highest correlation values ever achieved for the evaluated benchmarks (0.91 and 0.9). It is worth noting that the Context Vector measure (Wan and Angryk 2007), even aiming to overcome some of the theoretical limitations observed by the authors for the Gloss Vector measure, obtained a lower correlation (0.91 vs. 0.8). However, the Gloss Vector accuracy heavily depends on the way in which contexts are built. Authors (Patwardhan and Pedersen 2006) reported a high variability on the results according to the filtering

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

policy (*i.e.*, stop words removal and TF-IDF-based cut-offs) applied to words extracted from concept glosses. As a result, the maximum correlation value is obtained under a carefully tuned setting. The accuracy lowered down to 0.7 when TF-IDF cut-offs were modified in the authors' experiments. Another limitation is caused by their reliance on concept glosses. When this information is not directly available (which is the usual case in ontologies), word vectors are more difficult to build, requiring the compilation and processing of reliable corpora. The same authors (Pedersen, Pakhomov et al. 2007) discussed the difficulties and dependency on corpora and parameter tuning of their measure when applied to the domain of Biomedicine. These dependencies limit the applicability of those measures in concrete domains.

2.3.2.2.1 SC accuracy in a general domain

The similarity measures presented in this thesis (eq. 37 and 38) have also been compared with related works. We can see that their correlation results improve the accuracy of most related works. It is worth to note that the measures' correlation is only significantly surpassed by Gloss Vector (Patwardhan and Pedersen 2006) and the supervised SVM-based measures of Bollegala *et al*., (Bollegala, Matsuo et al. 2007; Bollegala, Matsuo et al. 2009). However, the correlation achieved by the Gloss Vector measure is obtained after it is tuned for optimal performance. Moreover, as stated above, Bollegala *et al*.,'s approaches, being distributional, exploit a very different principle to compute similarity, as corpora are used instead of ontological knowledge. They also present some limitations as either tagged examples and very reliable and carefully processed glosses are needed. As a result, their generality is compromised by either requiring training data or parameter tuning. On the contrary, our approach is much less constrained, as it does not depend on any tuning parameter and it only relies on taxonomical knowledge which is the most commonly available in ontologies.

From the runtime point of view, the computational complexity of Gloss Vector or Bollegala *et al*.,'s measures is much higher than our approach, as either on-line web searches and trained SVM-based classifiers are needed, or large amounts of data (*i.e*., WordNet glosses) are considered and processed (resulting in a very large dimensional space). 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.

Compared to other ontology-based measures, it is interesting to note that our approach's accuracy surpasses the basic edge-counting approaches (0.85 vs. 0.74). 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 (0.85 vs. 0.74), even though they are based on other non-taxonomical features and

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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 (0.85 vs. 0.73) showing that the exploitation of high quality taxonomical knowledge available in ontologies provides even more reliable semantic evidences than unstructured 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. Moreover, Lin as well as Jiang and Conrath proposals compute IC intrinsically following the same principle as feature-based measures: similarity is proportional to feature overlapping (in their case, represented by the IC of the LCS) and inversely proportional to the differences (in their case, the IC of each individual term). So, if the IC is computed from taxonomical evidences (*i.e.*, number of hyponyms) it makes sense that their correlation values are similar as those of our approach. 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.

Comparing the two versions of our proposal, we observe 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 (0.86 vs 0.84 and 0.85 vs. 0.81). This is coherent to what was discussed above, as an interpretation form the information theory perspective is consistent with the premises that intrinsic IC computation models exploit (which precisely obtained some of the highest accuracies among related works).

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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.

2.3.3 Evaluation of SC in the biomedical domain

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.

In the biomedical field, similarity computation can improve the performance of information retrieval from biomedical sources (Hliaoutakis, Varelas et al. 2006; Pedersen, Pakhomov et al. 2007) and may ease the integration of heterogeneous clinical data (Sugumaran and Storey 2002).

Pedersen *et al.* (Pedersen, Pakhomov et al. 2007) and Al-Mubaid and Nguyen (Al-Mubaid and Nguyen 2006; Al-Mubaid and Nguyen 2009) adapted some of the existing similarity measures to the biomedical domain by exploiting the biomedical ontology SNOMED CT (*Systematized Nomenclature of Medicine, Clinical Terms*).

In order to use IC measures based on corpora, general purpose corpus like SemCor cannot be used to estimate the information distribution of terms in the biomedical domain due to its reduced coverage of medical terms (Melton, Parsons et al. 2006). In this sense, Pedersen *et al.* (Pedersen, Pakhomov et al. 2007) proposed the use of the Mayo Clinic Corpus of Clinical Notes as a domain corpus together with the SNOMED CT taxonomy. The Mayo Clinic Corpus consists of 1,000,000 clinical notes collected over the year 2003 which cover a variety of major medical specialties at the Mayo Clinic. Clinical notes have a number of specific characteristics that are not found in other types of discourse, such as news articles or scientific medical articles found in MEDLINE. They are generated in the process of treating a patient and contain the record of the patient-physician encounter. Notes were transcribed by trained personnel and structured according to the reasons, history, diagnostic, medications and other administrative information. Patient's history, reason for visit and diagnostic related notes were used as the domain-specific data corpus from which IC-based measures can be computed.

With respect to the calculation of the semantic similarity for Gloss Vector measures, in the biomedical field, Pedersen *et al* adapted the *EGO* measure using the mentioned Mayo Clinic Corpus of Clinical Notes and the Mayo Clinic Thesaurus. They extracted context words and term descriptions from these resources, within a context window of one line of text. The Mayo Clinic thesaurus is a source of clinical problem descriptions that have been collected in the Mayo Clinic (i.e. the equivalent to WordNet glosses). It contains 16 million diagnostic phrases expressed in natural language classified in over 21,000 categories. Authors took these phrases to generate quasi-definitions (term descriptions) for terms found in SNOMED CT, after a pre-

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

processing stage aimed to reduce noise and redundancy of natural language text. The context words of the terms found in the descriptions extracted from the Clinical Notes repository were aggregated to get the context vector of a concept.

So, we will also evaluate our approach in this specific domain.

2.3.3.1 Benchmarks and ontologies

Several authors (Hliaoutakis, Varelas et al. 2006; Lee, Shah et al. 2008; Matar, Egyed-Zsigmond et al. 2008) have created *ad hoc* datasets to evaluate their approaches, framed on concrete research projects or oriented to particular ontologies. 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 (see Table 5). 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[5] 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, including their own proposal (*sem*); in this case, results were only compared against coders' ratings because they considered them to be more reliable than physicians' judgments.

Fortunately, there are a number of relevant biomedical ontologies, knowledge repositories and structured vocabularies that model and organize concepts in a comprehensive manner. Well-known examples are MeSH[6] (Medical Subject Headings) for indexing literature, the ICD taxonomy (International Classification of Diseases) for recording causes of death and diseases, and SNOMED CT[7].

Several authors (Al-Mubaid and Nguyen 2006; Pedersen, Pakhomov et al. 2007; Al-Mubaid and Nguyen 2009) have applied some of the classical similarity computation paradigms to medical data by exploiting SNOMED CT and/or clinical data. Some authors compared different similarity measures using SNOMED CT on particular datasets (Caviedes and Cimino 2004; Lee, Shah et al. 2008; Matar, Egyed-Zsigmond et al. 2008). In the context of a concrete application, semantic similarity has been used for document clustering (Melton, Parsons et al. 2006; Aseervatham and Bennani 2009). Although SNOMED CT is usually the ontology used, other authors

[5] Note that the pair *"chronic obstructive pulmonary disease" - "lung infiltrates"* was excluded from the test as the latter term is not found in the SNOMED CT terminology.

[6] http://www.ncbi.nlm.nih.gov/mesh

[7] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

exploited the MeSH ontology to compute the similarity assessment between words (Rada, Mili et al. 1989; Caviedes and Cimino 2004; Hliaoutakis, Varelas et al. 2006; Petrakis, Varelas et al. 2006; Pirró 2009).

Table 5. Set of 30 medical term pairs with averaged experts' similarity scores (extracted from (Pedersen, Pakhomov et al. 2007)).

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

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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).

2.3.3.2 Results and discussion

This section evaluates and compares the results obtained by our approach against those reported by other similarity functions when applied to the biomedical domain (Batet, Sánchez et al. 2009; Batet, Sánchez et al. 2010).

In order to enable an objective comparison between our proposal and other measures in the biomedical domain, we have also used the benchmark of Pedersen *et al.* and the SNOMED CT ontology to evaluate the accuracy of our measure. Correlation values obtained for our measure and correlations reported by related works (for those cases in which classical measures were evaluated in the biomedical context) with respect to both sets of human experts are presented in Table 6. Note that, for the context vector measure, four different tests are reported, changing two of its most influential parameters: corpus size (1 million or 100,000 clinical notes) and corpus selection (considering only the diagnostic section of clinical notes or all the sections of the document).

The first observation indicates that path length-based measures offered a limited performance, with correlations smaller than 0.36 and 0.66 respectively. This shows that limited accuracy is obtained when estimating semantic similarity only from the minimum inter-link path. In complex domain ontologies, such as SNOMED CT, where multiple paths between concepts constructed from several overlapping taxonomies are available, this approach wastes a lot of explicitly available knowledge. In fact, the measure with the best accuracy (0.66 for coders) is the one proposed by Al Mubaid which base the assessment both in the path length and in the relative depth of the concepts. This captures more knowledge than measures based only on the absolute path and the global depth of the ontology.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 6. Correlation values obtained for each measure against ratings of physicians, coders and both.

Measure	Type	Physicians	Coders	Both	Evaluated in
Rada	Edge-counting	0.36	0.51	0.48	(Pedersen, Pakhomov et al. 2007)
Wu and Palmer	Edge-counting	N/A	0.29	N/A	(Al-Mubaid and Nguyen 2006)
Leacock and Chodorow	Edge-counting	0.35	0.50	0.47	(Pedersen, Pakhomov et al. 2007)
Li et al.	Edge-counting	N/A	0.37	N/A	(Al-Mubaid and Nguyen 2006)
Al Mubaid	Edge-counting	N/A	0.66	N/A	(Al-Mubaid and Nguyen 2006)
Resnik	IC (corpus)	0.45	0.62	0.55	(Pedersen, Pakhomov et al. 2007)
Lin	IC (corpus)	0.60	0.75	0.69	(Pedersen, Pakhomov et al. 2007)
Jiang and Conrath	IC (corpus)	0.45	0.62	0.55	(Pedersen, Pakhomov et al. 2007)
Context vector (1 million notes, diagnostic section)	2n order co.	0.84	0.75	0.76	(Pedersen, Pakhomov et al. 2007)
Context vector (1 million notes, all sections)	2n order co.	0.62	0.68	0.69	(Pedersen, Pakhomov et al. 2007)
Context vector (100,000 notes, diagnostic section)	2n order co.	0.56	0.59	0.60	(Pedersen, Pakhomov et al. 2007)
Context vector (100,000 notes, all sections)	2n order co.	0.41	0.53	0.51	(Pedersen, Pakhomov et al. 2007)
SC_{Eu} (eq. 37)	**Ontology-based**	**0.59**	**0.74**	**0.70**	**This work**
SC_{log} (eq.38)	**Ontology-based**	**0.60**	**0.79**	**0.73**	**This work**

With regards to IC-based measures, in general, they are able to improve the results of path-length approaches. The maximum correlations are 0.6 for the physicians and 0.75 for the coders. Moreover, the minimum correlation for coders is 0.62, which outperforms path length results (with the exception of the one proposed by Al Mubaid). The fact of relying on high-quality domain corpora (i.e. clinical notes) allows complementing the taxonomical knowledge extracted from the ontology with additional semantic evidence, given by the distribution of the information of the concept in domain corpora. However, the applicability of these measures is hampered by the dependency on the availability and adequacy of domain data with respect to the evaluated concepts.

For the context vector measure, four cases were studied by the authors (Pedersen, Pakhomov et al. 2007), changing the corpus size and corpus selection. The correlation strongly depends on the amount and quality of the background corpus (with values between 0.51 and 0.76 considering the average of both sets of experts). The best accuracy (correlations of 0.84 for the physicians and 0.75 for the coders) is achieved under particular circumstances: 1 million notes involving only the diagnostic section. In this case, due to the fact that term definitions are extracted from high-quality corpora and due to the enormous size of the information sources, the obtained context vectors can adequately define the evaluated terms, enabling accurate estimates. However, for other corpus configurations, the accuracy of the measure decreases noticeably, at levels even below path-based approaches like the one proposed by Al Mubaid. In fact, it drops to correlations of 0.41 for physicians and 0.53 for coders when 100,000 notes involving all sections are used.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

In general, all the measures, except the context-vector-based ones, correlate better with coders than with physicians. On one hand, this is motivated by the amount of discrepancies observed in the physicians' ratings, which correlate lower than those of coders' (0.68 vs 0.78). On the other hand, the way in which human experts interpret concept likeness also influences the results. During the construction of the original data set, medical coders were requested to reproduce the classical Rubenstein & Goodenough (Rubenstein and Goodenough 1965) and Miller & Charles (Miller and Charles 1991) benchmarks in order to ensure that coders understood the instructions and the notion of similarity. However, physicians rated the pairs of concepts without pre-training or external influences. As a consequence, medical coders' ratings, which are trained in the use of hierarchical classifications, seem to reproduce better the concept of (taxonomic) *similarity* whereas physicians' ratings seem to represent a more general concept of (taxonomic and non-taxonomic) *relatedness*. These intuitions are coherent with the fact that the context vector measure estimates *relatedness*, whereas the other ontology-based measures estimate *similarity*.

In addition, for the tests with context vector measure, the data corpus used to create vectors was constructed by physicians of the same clinic; so, it is biased towards the way in which physicians interpret and formalize knowledge. As stated by Pedersen *et al.* (Pedersen, Pakhomov et al. 2007), these clinical notes may reflect implicit relations between concepts which were taken into consideration during the ratings and which are not explicitly indicated in a more general ontology such as SNOMED CT. Again, it makes sense that all the similarity measures correlate better with the less biased coders' ratings. In contrast, the unique relatedness measure considered in this review (context vector), which exploits the data composed by the same type of professionals which rated the benchmark, behaves in the inverse manner.

2.3.3.2.1 SC accuracy in the biomedical domain

Compared to other approaches the correlation values obtained by our measures for the evaluated benchmark are, in all cases, higher than those reported for path-based measures. It is particularly interesting to see how, being a pure ontology-based approach, our proposal reports higher correlations than some IC-based measures (concretely the ones defined by Resnik and Jiang & Conrath); only Lin's measure reports correlation values similar to ours. This shows that the exploitation of all the taxonomical knowledge available in the ontology provides comparable or even more semantic evidence than other approaches exploiting additional data sources. As it has been already said, the set of common and non-common superconcepts considered by our proposal incorporates, in an indirect manner, evidence of all the possible taxonomical paths between concepts, relative depths of branches and the relative densities of the involved taxonomical branches. As stated during the review of the related work, in other ontology-based approaches these semantic features are only partially considered, obtaining less accurate assessments. The context vector measure, however, offered a comparable or even better correlation with regards to our approach, when the complete amount of data and/or the diagnostic notes were used. In fact, it reported the highest correlation value (0.84) for physicians when using all the diagnostic notes.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Our measure obtains correlations which are comparable to the correlation between human experts: 0.60 vs 0.68 in the case of physicians, and 0.79 vs 0.78 with respect to medical coders. Analyzing in detail the different correlation values obtained with respect to the physicians' and the coders' ratings, one can notice important differences between the similarity measures. Again, the information theoretic version of our approach provides the highest correlations (0.60, 0.79 and 0.73) demonstrating again the convenience of adopting the non-linear logarithm based function in comparison to the geometrically founded Euclidian one.

Finally, we have analyzed the situation in which the context vector measures significantly surpass the correlation obtained with our new method. Correlation values higher than 0.6 (with respect to physicians) 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. As stated by Pedersen *et al.* (Pedersen, Pakhomov et al. 2007), this section contains more closely related terms than others which involve more noisy data. In consequence, as stated above, the choice, size and processing of the corpora used with the context vector measure is critical to achieve a good accuracy. This requires making a number of informed choices a priori in order make the measure behave as best as possible for a concrete situation and domain.

On the contrary, our measure, which is based only on an ontology, is able to provide a comparatively high accuracy without any dependency on data availability and pre-processing (which would hamper its applicability) and, at the same time, retains the low computational complexity and lack of constraints of path-based measures.

2.3.4 Evaluation of the contextualized Information Content approach

In this section, the presented approaches to compute the IC of a term, both in a contextualized and in a no contextualized way, are tested and compared. We are particularly interested in evaluating the increase in accuracy when introducing the contextualized queries in comparison with the uncontextual versions. The comparison against experiments based on tagged corpora are less relevant, as we are using different knowledge sources.

In the analysis we have used the following resources: WordNet as background ontology from which extract LCS of the pairs of words, the Resnik's benchmark, (being the more modern and the one proposed to evaluated IC-based measures, see section 2.3.2.1), and finally, in order to obtain term appearances from the Web, we use the search engine Bing[8] (the decision about the search engine is explained in the next section).

[8] http://www.bing.com/

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2.3.4.1 On the Web search engine

Even though other search engines can be used (Google, for example, offers a higher IR recall (Dujmovic and Bai 2006)), we found inconsistencies in the co-occurrence estimation in some of them which may produce unexpected results and compromise the similarity properties. Some examples of problematic cases for Google are provided in Table 7. In that case, quite different hit counts are obtained for equivalent web queries. Moreover, we observed a high variability in the hit counts for tests performed within a short period of time (days). Contrarily, Bing has provided consistent results during the different tests. Minimal variations in hit counts (also shown in Table 7) have been observed, mainly motivated by the use of different cached data from one query to another or changes in the IR database.

In any case, other Web search engines may be also suitable if they provide coherent results. As discussed in (Sánchez 2008), although the absolute occurrence values for a specific query may be quite different from one search engine to another, the final similarity values tend to be very similar as they are based in relative functions.

Table 7. Hit count returned by Google and Bing for equivalent queries [accessed: May 26th, 2009].

Terms to evaluate	Equivalent web queries	Google Web hit count	Bing Web hit count
dog cat	dog cat	173.000.000	72.600.000
	cat dog	28.800.000	70.300.000
	“dog” “cat”	30.900.000	72.600.000
	dog AND cat	26.100.000	72.600.000
dog dog	dog dog	449.000.000	208.000.000
	dog AND dog	374.000.000	208.000.000
	dog	396.000.000	209.000.000
	“dog”	332.000.000	208.000.000
cat dog mammal	cat dog mammal	277.000	519.000
	dog cat mammal	403.000	519.000
	cat mammal dog	1.740.000	519.000
	dog mammal cat	1.740.000	519.000
	mammal dog cat	404.000	516.000
	mammal cat dog	277.000	519.000

2.3.4.2 Results and discussion

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

word pair against the human ratings (see section 2.3.2.1) employed as baseline (Resnik 1995). 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) and, for the case of polysemic WordNet concepts, using the generalized versions presented in section 2.2.2.3.

Table 8 shows the list of correlation values of each similarity measure for each pair of words. The values of the first three rows correspond to non-contextualized measures presented in 2.2.2.1. The values in **bold** correspond to the contextualized measures proposed in this thesis (section **¡Error! No se encuentra el origen de la referencia.**). For each row, the correlation of the similarity values against the human ratings is provided as an indication of the result's quality.

Table 8. Correlation factors obtained for the evaluated measures.

Measure	Correlation
Resnik_IR	0.403
Lin_IR	0.357
Jcn_IR	0.364
Lin_ CIC_T _IR	**0.665**
JCN_ CIC_T_IR	**0.678**

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 inaccurate estimation derived from the language ambiguity and the lack of taxonomic coherence in the IC computation hamper the final results (correlation values range from 0.35 to 0.4). Lin and Jiang & Conrath are the most handicapped by the latter issue due to their explicit comparison between concept's IC and their subsumer (correlations are below 0.4).

Comparatively, remembering the results obtained using the SemCor (Miller, Leacock et al. 1993) as corpus (see Table 4 of section 2.3.2.2), Resnik measure obtained a correlation of 0.72. Jiang & Conrath measure obtained a correlation among 0.73 and 0.75. Both are quite near to the human upper bound of 0.884 computed by Resnik replication (Resnik 1995). However, this quality is heavily associated to the accurate frequencies computed from the limited manually disambiguated corpus, tagged according to WordNet synsets. As shown in our tests, the lack of this tagged data gives much lower accuracy.

The inclusion of the contextualized version of IC computation in Lin and Jiang & Conrath, due to the additional context. As a result, they clearly outperform the basic versions, almost doubling the correlation value (0.67 vs. 0.36). In this case, even though concept probabilities have been sub-estimated, the monotonic coherence of IC computation with respect to the taxonomic structure and the minimized ambiguity of word occurrences certainly improve the results.

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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.

2.4 Summary

In this chapter, an up-to-date survey and review of the semantic similarity measures that can be used to estimate the resemblance between terms has been provided. A total of 31 individual measures have been reviewed and grouped in four different families.

The main contributions presented in this chapter are:

(1) A similarity measure that considered all the superconcepts of terms in a given ontology (Batet, Sánchez et al. 2009; Batet, Sánchez et al. 2010).

(2) A contextualized version for the computation of the IC from the Web and its application to some IC-based similarity measures (Sánchez, Batet et al. 2010b).

With respect to the first contribution, this measure only relies on taxonomic ontological knowledge, lacking of corpora-dependency or parameter-tuning, being also computationally efficient. The chapter has analysed the advantages and problems of both related works and our proposal, with the aim of giving some insights on their accuracy, applicability, dependencies and limitations. In addition, a complete comparison of all these measures has been presented, considering two different scenarios: a non-specific domain and the biomedical domain.

In particular, the results reported by the new measure 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.

About the second contribution, it has been studied the behaviour of IC-based approaches when IC is computed form the Web. We have shown that applying the classic approach over a massive unprocessed corpus like the Web resulted in very inaccurate similarity assessments, as absolute word occurrences provided a poor estimation of concept probabilities. In order to minimize this problem and the problem about data-sparseness of corpora and to maintain the scalability of the approach, a contextualized version of IC computation which seeks for explicit word co-occurrences between evaluated concepts and their LCS has been proposed. Then, this approach has been applied to IC-based measures.

The modified versions of similarity measures (contextualizing the IC computation) are able to provide results for virtually any possible concept contained

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

in WordNet (as far as they are indexed by web search engines) or any ontology (i.e. an ontology containing domain-specific classes or even instances not considered in WordNet). Moreover, this is done in a Web-scalable manner without any kind of manual intervention or pre-processing. In consequence data sparseness problems which may appear with rare concepts are greatly minimized and the generality of the measures is improved.

Some advantages of the approach of computing IC from the Web is that any possible concepts can be evaluated if the term is contained in WordNet. However, queries to search engines have a high temporally cost. In addition, we have found inconsistencies in the co-occurrence estimation in some search engines which may produce unexpected results. A high variability in the hit counts for tests performed within a short period of time is also observed.

The SC_{log} measure provides a clearly improvement of the results when compared against standard benchmarks. This measure has a low computational cost because only an ontology is exploited. It does not depend on tuning parameters. Different to the IC approach from the Web this measure is not affected by the ambiguity problem. In addition, as a great amount of the Web resources are in English, the statistics obtained from the Web for a term will depend on the language used to query it. On the contrary, the knowledge represented in an ontology is language-independent; so, ontology-based measures are not affected by the language labels used to refer to concepts.

For all these reasons SC_{log} will be used in order to compare terms in our semantic clustering approach (chapter 4).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Chapter III

3 Semantic similarity using multiple ontologies

As explained in chapter 2, similarity estimation is based on the extraction of semantic evidences from one or several knowledge resources. In this chapter we will focus on the measures that use several ontologies as background knowledge.

Pure ontology-based measures work in an automatic and very efficient manner, and provide quite accurate results when a well detailed and taxonomically homogenous ontology is available (Leacock and Chodorow 1998). Their main drawback is the complete dependency on the degree of coverage and detail of the input ontology for the evaluated concepts. In the literature, most ontology-based similarity measures only use a unique ontology as background knowledge. However, in many domains, there are different ontologies which provide partial and/or overlapping views of the same knowledge.

Multiple ontologies provide additional knowledge (Al-Mubaid and Nguyen 2009) that may help to improve the similarity estimation and to solve cases in which terms are not represented in a certain ontology (i.e. missings). This is especially important in domains such as biomedicine in which several big and detailed ontologies are available (*i.e.*, MeSH and SNOMED) offering overlapping and complementary knowledge for the same domain. Even in this case, general purpose ontologies such as WordNet offer a relatively good coverage of some specific domains and may aid to improve similarity assessments in a multi-ontology setting. In consequence, the exploitation of multiple input sources may lead to a better coverage and more robust similarity estimations.

The main issue regarding multiple ontologies exploitation is the variation among the level of detail and granularity of different ontological representations, which makes difficult the comparison and integration of similarities computed from different ontologies (Al-Mubaid and Nguyen 2009). As it will be detailed in the section 3.1, very little work has been done regarding the development of similarity methods that use multiple ontologies in an integrated way.

In this chapter we present a new similarity computation method exploiting several ontologies. The method relies on the ontology-based similarity measure presented in section 2.2.1, extending its definition to a multi-ontology setting.

In a multi-ontology setting, our methodology would be able to improve even more the similarity estimation presented in the previous chapter. On the one hand, it permits to estimate the similarity when some term is missing in a particular ontology but can

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

be found in another one. On the other hand, in case of overlapping knowledge (*i.e.*, ontologies covering the same pair of terms), it will be also able to increase the accuracy by selecting the most reliable similarity estimation from those computed from the different ontologies. A heuristic has been designed to tackle this task by measuring the amount of semantic evidences observed for each ontology. It is worth to note that, on the contrary to previous works, our approach does not make any assumption regarding the suitability and coverage of the input ontologies for the evaluated terms, and it operates in an unsupervised way during the semantic integration. Finally, our approach has been designed to avoid the problems of the correct integration of the numerical scale of the similarities computed from sources with different granularity, an issue which represented a serious problem in previous attempts based on absolute ontological paths (Al-Mubaid and Nguyen 2009).

The objectives of this chapter are the following:

1. To review related works focusing on ontology-based semantic similarity in a multi-ontology setting and discuss their performance.
2. To propose a method to exploit multiple ontologies as input.
3. To evaluate our approach with several ontologies and compare its accuracy against a mono-ontology setting and with regards related works.

In section 3.1 an introduction of related works focusing on ontology-based semantic similarity in a multi-ontology setting is presented. In section 3.2, it is described the proposed method to exploit several ontologies to assess semantic similarity. Section 3.3 presents the results obtained for several standard benchmarks and provides a detailed comparison against related works. The final section summarizes the chapter and presents the conclusions of this work.

3.1 Semantic similarity using multiple ontologies: state of the art

In the past, the general approach to data integration has been to map the local terms of distinct ontologies into a existent single one (Tversky 1977; Gangemi, Pisanelli et al. 1998; Guarino 1998; Weinstein and Birmingham 1999) or to create a new shared ontology by integrating existing ones (Mena, Kashyap et al. 1996; Bergamaschi, Castano et al. 1998; Gangemi, Pisanelli et al. 1998). However, manual or semi-automatic ontology integration represents a really changeling problem, both from the cost and scalability point of view (requiring the supervision of an expert) and the difficulty to treat with overlapping concepts and inconsistencies across ontologies (Rodríguez and Egenhofer 2003).

Tackling the problem from a different point of view, Rodríguez and Egenhofer (Rodríguez and Egenhofer 2003) compute the similarity between classes as a function of some ontological features and the degree of generalization between classes (the path distance between the pair of classes and their least common subsumer) into the same or different ontologies. When the term pair belongs to different ontologies they approximate the degree of generalization between classes by considering that these ontologies are connected by a new imaginary root node that subsumes the root classes of these two ontologies. A problem of their approach is that it relies in many ontological features (such as attributes, synonyms, meronyms and other kind or non-

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

taxonomic relationships) which are rarely found in ontologies. In fact, an investigation of the structure of existing ontologies via the Swoogle ontology search engine (Ding, Finin et al. 2004) has shown that domain ontologies very occasionally model non-taxonomical knowledge. Moreover, the method (Rodríguez and Egenhofer 2003) does not consider the case in which the term pair is found in several ontologies (a very common situation as it will be shown during the evaluation). In consequence, it omits the problem of selecting the most appropriate assessment and/or to integrate overlapping sources of information. In addition, the integration of different ontologies is very simple and does not consider the case in which ontologies share subsumers that could be used as bridging classes. Finally, the evaluation of the method does not use a standard benchmark, making difficult a direct comparison against other approaches.

A very recent and more general proposal covering all possible situations which may appear in a multi-ontology setting is presented by Al-Mubaid and Nguyen (Al-Mubaid and Nguyen 2009). They apply it to the UMLS framework, in which concepts are dispersed across several overlapping ontologies and terminologies (such as MeSH or SNOMED). Authors propose a methodology to exploit those several knowledge sources using the path-based *SemDist* distance introduced section 2.1.1.1. They rely on a pre-defined "primary" ontology (the rest are considered as "secondary") that acts as the master in cases in which concepts belong to several ontologies and that is used to normalize similarity values. As a result, they propose different strategies according to the situation in which the compared concept pair appear. Concretely, if both concepts appear in the primary ontology, the similarity is computed exclusively from that source (even in the case that they also appear in a secondary ontology). If both concepts only appear in a unique secondary ontology, obviously, the similarity is computed from that source. A more interesting case arises when concepts appear in several secondary ontologies. Authors propose a heuristic to choose from which of those ontologies the similarity should be computed, based on the degree of overlapping with respect to the primary ontology and the degree of detail of the taxonomy (granularity). Finally, if a concept is uniquely found in an ontology (e.g. the primary) and the other in another one (e.g. a secondary one), they temporally "connect" both ontologies by finding "common nodes" (i.e. a subsumer representing the same concept in any of the ontologies) and considering the result as a unique ontology.

A problem that authors face is the fact that different ontologies may have different granularity degrees (i.e. depth and branching factor for a certain taxonomical tree). So, being their measure based on absolute path distances between concepts, the similarity computed for each term pair from different ontologies will lead to a different similarity scale which cannot be directly compared. So, they propose a method to scale similarity values (both in the case in which the concept pair belongs to a unique secondary ontology or when it belongs to different ontologies - both secondary, or one primary and the other secondary - which are "connected") taking as reference the predefined primary ontology. They scale both the *Path* and *CSpec* (see section 2.1.1.1) features to the primary ontology according to difference in the depth with respect to the primary ontology. For example, in the simpler case in which both concepts belong to a unique secondary ontology, *Path* and *CSpec* are computed as stated in eq. 50 and 51 respectively, and they compute the similarity using eq. 5.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$Path(a,b) = Path(a,b)_{\sec ondary\_onto} x \frac{2D_1 - 1}{2D_2 - 1} \quad (50)$$

$$CSpec(a,b) = CSpec(a,b)_{\sec ondary\_onto} x \frac{D_1 - 1}{D_2 - 1} \quad (51)$$

where D_1 and D_2 are the depths of the primary and secondary ontologies, respectively. Hence, (D_1 -1) and (D_2 -1) are the maximum common specificity values of the primary and secondary ontologies respectively, and ($2D_1$ -1) and ($2D_2$ -1) are the maximum path values of two concept nodes in the primary and secondary ontologies, respectively.

This approach has two drawbacks. On one hand, as the proposed method bases the similarity assessment on the *minimum path* connecting concept pairs, it omits much taxonomical knowledge already available in the ontology, as it has been argued in chapter 2. On the other hand, the cross-ontology method is hampered by the fact that a primary ontology should be defined a priori by the user in order to normalize similarity values. First, this scaling process, motivated by the fact of basing the similarity on absolute values of semantic features (i.e. path and depth) of a concrete ontology, results in a complex casuistic to be considered during the similarity assessment. Second, it assumes that, in all situations, the primary ontology will lead to better similarity estimation than secondary ones, which is a hard assumption that could not be true, particularly when no clear criteria are provided to select the primary ontology. In fact, there may be situations in which, for a pair of concepts appearing both in the primary ontology and, in addition, in one or several secondary ontologies, similarity estimation from a secondary one may lead to better results because its knowledge representation is more accurate or more detailed for a particular subset of terms. Even though, they evaluate the approach using standard benchmarks and widely available general and biomedical ontologies. Experiments regarding the influence in the results of selecting one ontology or another as the primary are missing.

3.2 A new method to compute similarity from multiple ontologies

In order to overcome the limitations of existing methods, our approach will tackle the problem from another point of view. Taking the similarity measure introduced in section 2.2.1 (SC_{log}) (which provided the highest accuracy in the different tests) we propose a more general multi-ontology similarity computation method covering all the possible situations tackled by (Al-Mubaid and Nguyen 2009) but, differently from (Rodríguez and Egenhofer 2003), we only rely on taxonomical knowledge which is usually available in domain ontologies. Moreover, instead of relying on one primary ontology as (Al-Mubaid and Nguyen 2009) do, we will exploit all available ontologies individually and then, we will assess which of them provides the best estimation, without requiring to scale numerical values because they are ratios between shared and non shared superconcepts. This will potentially lead to better

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

results because the punctual benefits in the estimation obtained from each ontology are exploited.

One of the premises of our method (Batet, Valls et al. 2010c) is to consider that all input ontologies are equally important. This avoids the limitations introduced by the dependency of a pre-selected primary ontology. As a result, our multi-ontology similarity computation methodology is simplified to three cases (instead of five proposed in (Al-Mubaid and Nguyen 2009)), according to whether or not each or both concepts (c_1 and c_2) belong to any of the considered ontologies.

Case 1: Concepts a and b appear together in a single ontology.

If the pair of concepts occurs in a unique ontology, the similarity is computed like in a mono-ontology setting, using *SC* measure proposed in section 2.2.1.

Case 2: Concepts a and b both appear in more than one ontology.

In this case, both concepts appear in several ontologies, each one modeling domain knowledge in a different but overlapping way. As a result, the computation of similarity is influenced by the different levels of detail or knowledge representation accuracy of each ontology (Rodríguez and Egenhofer 2003). So, it is necessary to decide, according to a heuristic, which ontology provides a better estimation of the inter-concept similarity. Considering the nature of the ontology engineering process, and the psychological implications of a human assessment of the similarity, two premises can be enounced.

First, ontological knowledge (*i.e.* a concept or relation) needs to be explicitly modelled by knowledge experts. In consequence, the fact that a pair of concepts obtains a high similarity score is the result of considering common knowledge modelled through an explicit ontology engineering process. However due to the knowledge modeling bottleneck, which typically affects manual approaches, ontological knowledge is usually partial and incomplete (Gómez-Pérez, Fernández-López et al. 2004). As a result, if two concepts appear to be semantically far (*i.e.* they do not share knowledge), one cannot ensure if this is an implicit indication of semantic disjunction or the result of partial or incomplete knowledge. So, we can conclude that explicitly modelled knowledge gives more semantic evidence than the lack of it.

Second, psychological studies have demonstrated that human subjects pay more attention to similar than to different features during the similarity assessment (Tversky 1977; Krumhansl 1978). Thus, we assume that non-common characteristics between entities are less important than common ones.

The method proposed in this case is the following:

Given a pair of concepts appearing in different ontologies,

1. we calculate the semantic similarity between them (computed using eq. 38) in all the ontologies and
2. then we keep the highest similarity score obtained.

$$sim(a,b) = \max_{i \in O} SC_{\log_i}(a,b) \tag{52}$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

, being *O* the set of ontologies that contain both *a* and *b*.

Notice that with SC_{log}, a high score is obtained if the number of common ancestors is large. Following the previous premises, the highest score is obtained from the ontology that incorporates the largest amount of explicit evidences of relationship between terms (explicitly modelled by the knowledge engineer).

With this method, for each pair of concepts belonging to several ontologies, the final similarity values may be taken from a different ontology. This will correspond to cases in which a particular ontology provides a more accurate modeling of the two concepts, regardless its global degree of granularity or detail in comparison to the other ontologies. Our heuristic will exploit the punctual benefits offered by each ontology for the given pair of concepts in cases in which overlapping knowledge is available on the contrary to approaches which always rely on a primary ontology.

Moreover, as the SC_{log} measure provides *relative* similarity values, normalized to the granularity degree of each ontology, the comparison between the results obtained from different ontologies for the same pair of concepts does not require additional scaling.

Case 3: None of the ontologies contains the concepts a and b simultaneously

In this case, each of the two concepts belongs to a different ontology, each one modeling the knowledge from a different point of view. This case faces the problem of exploiting, in some way, different ontologies in order to measure the similarity across them. In that sense, the computation of similarity across ontologies can be only achieved if they share some components (Rodríguez and Egenhofer 2003).

As stated in section 3.1, some approaches tackled this problem by merging different ontologies in a unique one, introducing a high computational and human cost and dealing with the difficulties inherent to the treatment of ambiguous overlapping concepts and to the avoidance of inconsistencies.

From a different point of view, as introduced in section 3.1, Al-Mubaid and Nguyen (Al-Mubaid and Nguyen 2009) base their proposal in the differentiation between a primary and secondary ontologies, connecting the secondary one to the primary by joining all the equivalent nodes (i.e. those with concepts represented by the same *label*). These nodes are called *bridges*. In their proposal the LCS of the pair of concepts in two ontologies is redefined as the LCS of the concept belonging to the primary ontology and one of the bridge nodes.

$$LCS_n(a,b) = LCS(a, bridge_n) \tag{53}$$

Then, the path is computed through the two ontologies via the LCS and the bridge node, and the similarity is assessed. Again, as the ontologies have different granularity degrees, it is needed to normalize the calculation. Then, it is measured the path between the concept of the primary ontology and the LCS, and the path between the concept of the secondary ontology and the LCS *scaled* with respect to the primary ontology dimension.

In Rodríguez and Egenhofer (Rodríguez and Egenhofer 2003) the two ontologies are simply connected by creating a new node (called *anything*) which is a direct ancestor of their roots.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Considering the nature of our measure, in which the set of common and non-common superconcepts is exploited, the approach proposed in (Rodríguez and Egenhofer 2003) implies the loss of all the potentially common superconcepts. Moreover, differently from (Al-Mubaid and Nguyen 2009), where only the minimum path length to the LCS is computed, we need to have a complete view of the taxonomical structure above the evaluated concepts (including all taxonomical relationships in cases of multiple inheritance) .

Due to these reasons, we propose a new method to assess the similarity across different ontologies. It is based on calculating the union of the set of superconcepts of *a* and *b* in each ontology and on finding equal concepts between them. This allows knowing the amount of common and non-common knowledge in a cross-ontology setting. It is important to note that, on the contrary to (Al-Mubaid and Nguyen 2009), where all the bridge nodes of the ontology are considered (introducing a high computational burden in big ontologies), we only evaluate the superconcepts of each concept.

The method which detects equivalences between superconcept sets is based on the theory of ontology *alignment*. This consists of finding relationships between entities belonging to different ontologies, but preserving the original ontologies (Noy and Musen 1999). In our approach, a terminological method is used, in which concept *labels* are compared to find equivalent nodes. Equivalent superconcepts are those with the same textual label considering, if available, their synonyms.

In addition to terminologically equivalent superconcepts, it is also logic to consider that all their subsumers are also common, regardless having or not an identical label. In fact, each of the evaluated concepts inheriting from terminologically equivalent superconcepts recursively inherits from all the superconcepts' subsumers. Summarizing, the set of shared superconcepts for *a* belonging to the ontology O_1 and *b* belonging to the ontology O_2, is composed by those superconcepts of *a* and *b* with the same label, and also all the subsumers of these equivalent superconcepts.

Formally, $A_{O1}(a)$ is the set of superconcepts of concept *a* (including *a*) in the is-a hierarchy H^C_{O1} of concepts (C_{O1}) in ontology O_1, and $A_{O2}(b)$ is the set of superconcepts of concept *b* (including *b*) in the hierarchy H^C_{O2} of concepts (C_{O2}) in ontology O_2, defined as:

$$A_{O1}(a) = \{c_i \in C_{O1} \mid c_i \text{ is superconcept of } a\} \cup \{a\} \quad (54)$$

$$A_{O2}(b) = \{c_j \in C_{O2} \mid c_j \text{ is superconcept of } b\} \cup \{b\} \quad (55)$$

Then, the set *terminologically equivalent superconcepts* (*ES*) in $A_{O1}(a) \cup A_{O2}(b)$ is defined as:

$$ES = \{c_i \in A_{O_1}(a) \text{ and } c_j \in A_{O_2}(b) \mid c_i = c_j\} \quad (56)$$

Finally, the set of common superconcepts ($A_{O1}(a) \cap A_{O2}(b)$) is composed by the elements in *ES* and all the ancestors of the elements in *ES*.

$$\bigcup_{\forall c_i \in ES} (A_{O_1}(c_i) \cup A_{O_2}(c_i)) \quad (57)$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

The remaining elements in $A_{O1}(a) \cup A_{O2}(b)$ are considered as non-common superconcepts.

Having the set of common and non-common superconcepts, we are able to apply the basic similarity measure.

If *a* and/or *b* belong to several ontologies (e.g. *a* belongs to O_1 and O_3; and *b* belongs to O_2 and O_4), the described alignment process and the similarity computation are executed for each combination of ontology pairs (e.g. O_1 - O_2, O_1 - O_4, O_3 - O_2 and O_3 - O_4). Following the heuristic introduced for *Case 2*, the highest similarity value obtained will be taken as the final result.

Note, that when if the pair of compared concepts appear both in a single ontology, then is not considered the case in which concept *a* is in an ontology an concept *b* in contained in another ontology.

So, summarizing this method (which we call MSC_{log}) and denoting as Θ the set of ontologies and $SC_{log\_o}(a,b)$ the computation of SC_{log} over an ontology $o \epsilon \Theta$:

$$MSC_{\log}(a,b) = \begin{cases} SC_{\log\_o}(a,b) & \text{if } a \text{ and } b \text{ appear toghether in a single ontology} \\ \max\limits_{\{o \in \Theta | a,b \in o\}} SC_{\log\_o}(a,b) & \text{if } a \text{ and } b \text{ appear in several ontologies} \\ \max\limits_{\substack{\{o_1 \in \Theta | a,b \in o_1 \\ o_2 \in \Theta | a,b \in o_2\}}} -\log 2 \dfrac{|A_{O_1}(a) \cup A_{O_2}(b)| - |\bigcup\limits_{\forall c_i \in ES} (A_{O_1}(c_i) \cup A_{O_2}(c_i))|}{|A_{O_1}(a) \cup A_{O_2}(b)|} & \text{otherwise} \end{cases}$$

(58)

Using this approach, we are able to maintain the properties of the underlying measure (i.e. maximization of the taxonomical knowledge exploited in the assessment) but integrating the knowledge from different ontologies in a seamless way. The numerical scale of the similarity values is also maintained regardless of the input ontology, because the results are implicitly normalized by the size of the corresponding superconcept sets.

3.3 Evaluation

As stated in section 2.3.1, the most common way of evaluating similarity measures consists of using standard benchmarks of word pairs whose similarity has been assessed by a group of human experts. The correlation of the similarity values obtained by the computerized measures against human similarity ratings is calculated. If the correlation approaches to 1, this will indicate that the measure properly approximates the judgments of human subjects.

The accuracy of the SC_{log} measure introduced in section 2.2.1, was shown in section 2.3.2 for a general domain and in section 2.3.3 for the biomedical domain. As stated above, this measure is the base of our multi-ontology similarity computation

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

method. Through the different tests it has been shown that it was able to provide some of the highest correlated values, thanks to the exploitation of additional taxonomical knowledge.

In this section, the accuracy of the similarity measure method in a multi-ontology setting (MSC_{log}) is evaluated and compared against related works using different benchmarks and ontologies, which are described below.

The case of biomedicine will be the focus of our evaluation. As stated above, the biomedical domain is especially representative, as it is very prone to the development of big and detailed ontologies and knowledge structures. The UMLS repository (Unified Medical Language System) is a paradigmatic example of global initiatives bringing structured knowledge representations and detailed ontological structures. It includes many biomedical ontologies and terminologies (MeSH, SNOMED, ICD, etc.). Those ontologies are also characterized by their high level of detail, classifying concepts in several overlapping hierarchies, which make explicit a great amount of taxonomical knowledge) offering overlapping and complementary knowledge for the same domain. Therefore, biomedicine is a particularly interesting domain of application for the multi-ontology method that we propose.

The section is organized as follows. First, the resources used in order to perform the evaluation will be described. In sections 3.3.1, the standard benchmarks of word pairs and the ontologies used are described. Second, we will show and discuss the results obtained in the evaluation performed (section 3.3.2). In particular, in section 3.3.2.3 our approach is compared against related works.

3.3.1 Benchmarks and ontologies

Focusing the benchmarks available in the biomedical domain, we have analyzed two different biomedical datasets: the ones created by Pedersen and Hliaoutakis (Hliaoutakis 2005; Pedersen, Pakhomov et al. 2007).

The first one has already been used in the evaluation of our similarity measures. This dataset (see section 2.3.3.1) was created by Pedersen *et al.* in collaboration with Mayo Clinic experts, and is composed by a set of word pairs referring to general medical disorders (Pedersen, Pakhomov et al. 2007).

The second biomedical benchmark was proposed by Hliaoutakis (Hliaoutakis 2005). This dataset is composed by a set of 36 word pairs extracted from the MeSH repository (see Table 9). The similarity between word pairs was also assessed by 8 medical experts from 0 (non similar) to 1 (perfectly similarity).

Those domain dependent benchmarks will be also complemented in some tests with the general purpose benchmarks presented in section 2.3.2.1.

Moreover, in order to evaluate our proposal both in a mono and multi ontology setting, we have used WordNet (Fellbaum 1998) as domain independent ontology, and SNOMED CT and MeSH as domain-specific biomedical ontologies.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 9. Set of 36 medical term pairs with averaged experts' similarity scores (extracted from (Hliaoutakis 2005)).

Term 1	Term 2	Human ratings (averaged)
Anemia	Appendicitis	0.031
Otitis Media	Infantile Colic	0.156
Dementia	Atopic Dermatitis	0.060
Bacterial Pneumonia	Malaria	0.156
Osteoporosis	Patent Ductus Arteriosus	0.156
Amino Acid Sequence	Antibacterial Agents	0.155
Acq. Immunno. Syndrome	Congenital Heart Defects	0.060
Meningitis	Tricuspid Atresia	0.031
Sinusitis	Mental Retardation	0.031
Hypertension	Kidney Failure	0.500
Hyperlipidemia	Hyperkalemia	0.156
Hypothyroidism	Hyperthyroidism	0.406
Sarcoidosis	Tuberculosis	0.406
Vaccines	Immunity	0.593
Asthma	Pneumonia	0.375
Diabetic Nephropathy	Diabetes Mellitus	0.500
Lactose Intolerance	Irritable Bowel Syndrome	0.468
Urinary Tract Infection	Pyelonephritis	0.656
Neonatal Jaundice	Sepsis	0.187
Anemia	Deficiency Anemia	0.437
Psychology	Cognitive Science	0.593
Adenovirus	Rotavirus	0.437
Migraine	Headache	0.718
Myocardial Ischemia	Myocardial Infarction	0.750
Hepatitis B	Hepatitis C	0.562
Carcinoma	Neoplasm	0.750
Pulmonary Stenosis	Aortic Stenosis	0.531
Failure to Thrive	Malnutrition	0.625
Breast Feeding	Lactation	0.843
Antibiotics	Antibacterial Agents	0.937
Seizures	Convulsions	0.843
Pain	Ache	0.875
Malnutrition	Nutritional Deficiency	0.875
Measles	Rubeola	0.906
Chicken Pox	Varicella	0.968
Down Syndrome	Trisomy 21	0.875

As explained in section 2.3.2.1, WordNet is a freely available lexical database that describes and structures and amount of general English concepts. In order to properly compare the results, we use WordNet version 2 in our tests as it is the same version used in the related works. SNOMED CT (see section 2.3.3.1) is one of the largest sources included in the Unified Medical Language System (UMLS) that includes most of the medical concepts, including them in one or several conceptual hierarchies.

The Medical Subject Headings (MeSH) ontology is mainly a hierarchy of medical and biological terms defined by the U.S National Library of Medicine to catalogue books and other library materials, and to index articles for inclusion in health related

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

databases including MEDLINE. It consists of a controlled vocabulary and a hierarchical tree. The controlled vocabulary contains several different types of terms such as Descriptors, Qualifiers, Publication Types, Geographic and Entry terms. MeSH descriptors are organized in a tree which defines the MeSH Concept Hierarchy. In the MeSH tree there are 16 categories, with more than 22,000 terms appearing on one or more of those categories.

3.3.2 Similarity evaluation in a multi-ontology setting

In order to evaluate the proposed methodology, we have performed several tests combining the ontologies introduced in the previous section.

As it has been said, the tests have been done using the biomedical benchmarks proposed by Pedersen *et al.* (Pedersen, Pakhomov et al. 2007) (using physicians', coders' and both ratings) and Hliaoutakis (Hliaoutakis 2005). With these datasets and the ontologies MeSH, SNOMED CT and WordNet, we can find the different cross-ontology cases detailed in section 3.2. It is important to note that WordNet, even being a general-purpose ontology covers most of the terms considered in those benchmarks, enabling an interesting comparison of the benefits that one can achieve when incorporating different kind of ontologies to the similarity assessment.

3.3.2.1 Evaluation with missing terms

In the first experiment we have taken all word pairs of each of both benchmarks and we have evaluated them with several mono and multi-ontology configurations:

- SNOMED CT, MeSH and WordNet in an independent way (mono-ontology setting).
- All the combinations of pairs of those ontologies.
- All the three ontologies at the same time.

Note that the term pair *"chronic obstructive pulmonary disease"* - *"lung infiltrates"* of Pedersen *et al*. benchmark was excluded from the test bed as the latter term was not found in any of the three ontologies. From the remaining 29 pairs, all of them are contained in SNOMED CT, 25 of them are found in MeSH and 28 in WordNet. For the Hliaoutakis' benchmark, all the 36 word pairs are found in MeSH and WordNet, but only 35 are contained in SNOMED CT. Those values indicate that there will be some cases in which one of the words is missing in some of the ontologies but found in another. In those cases, a multi-ontology setting will potentially lead to a better accuracy in the results, because it is able to calculate the similarity of those missing terms from the combination of multiple ontologies.

In order to enable a fair comparison with regards to the multi-ontology setting, we substitute the missing evaluations by the average similarity for the benchmark's word pairs found for the given ontology. This introduces a proper penalization in the correlation when missing word pairs appear in a mono-ontology setting. The

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

correlation values obtained for all benchmarks and for all ontology combinations are shown in Table 10.

Table 10. Correlation values obtained by the proposed method for Pedersen *et al*.'s benchmark (Pedersen, Pakhomov et al. 2007) (with 29 word pairs) for the rations of physicians, coders and both and for Hliaoutakis' benchmark (Hliaoutakis 2005) (with 36 pairs).

Ontologies	Pedersen Physicians	Pedersen Coders	Pedersen Both	Hliaoutakis
SNOMED CT	0.601	0.788	0.727	0.557
MeSH	0.562	0.769	0.694	0.749
WordNet	0.535	0.747	0.669	0.611
SNOMED CT + WordNet	0.624	0.799	0.744	0.727
MeSH + WordNet	0.596	0.790	0.724	0.770
SNOMED CT + MeSH	0.642	0.817	0.762	0.740
SNOMED CT + MeSH + WordNet	0.656	0.825	0.773	0.787

Analysing the results, we can extract several conclusions. First, we can observe a surprisingly good accuracy when using WordNet as ontology (especially for the Pedersen's benchmark), which correlation values are only marginally worse than those obtained from the medical ontologies. In fact, for the Hliaoutakis' benchmark (which was designed from MeSH terms), WordNet is able to improve the correlation obtained with SNOMED CT alone. This shows that WordNet, even being a general purpose ontology, offers a good coverage of relatively common biomedical terms, possibly because parts of the WordNet taxonomy have been taken from UMLS.

Secondly, we can observe that, in most situations, the use of several ontologies to compute similarity leads to an improvement in the accuracy in contrast to the use of each ontology individually. It is particularly interesting to see how the addition of WordNet to each medical ontology qualitatively improves the results. For the Pedersen's benchmark: from 0.69 to 0.72 for MeSH and from 0.72 to 0.74 for SNOMED CT. For the Hliaoutakis' benchmark: from 0.75 to 0.77 for MeSH and from 0.56 to 0.73 for SNOMED CT. This means that, at least, parts of the WordNet taxonomy represent better the semantics inherent to the evaluated terms or even covers more terms (resulting in a higher similarity as assumed by the method described in section 3.2). In any case, the relative improvement obtained from the combination of the two medical ontologies (SNOMED CT and MeSH) leads to a higher accuracy in most situations. This is reasonable due to the biomedical nature of evaluated word pairs. Finally, the combination of all ontologies provides the highest correlation in all cases (e.g. the correlation obtained against the Pedersen's medical coders is 0.825 when using all the ontologies vs. 0.788 when using only SNOMED CT, 0.769 when using MESH and 0.747 when using WordNet). This result shows that the more knowledge available, the better the estimations will be. This is motivated both for the solving of missing values and thanks to the selection of the most accurate assessment from those provided by the different overlapping ontologies.

From a general point of view, we can also observe that our method (and in consequence the underlying similarity measure) correlates better with coders than with physicians for the Pedersen's benchmark. On one hand, this is motivated by the higher amount of discrepancies observed in physician ratings, which correlate lower than coders (Pedersen *et al*. reported a correlation between human subjects of 0.68 for

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

physicians and 0.78 for coders). On the other hand, coders, due to their training and skills, were more familiar than physicians to the concept of hierarchical classifications and semantic *similarity* (on the contrary to *relatedness*) which lead to a better correlation with the design principles of our similarity approach.

3.3.2.2 Evaluation without missing terms

Considering that, in the experiments reported above, some of the tests presented missing terms whose similarity estimation hampered the final correlation values in a mono-ontology setting, we ran an additional battery of tests considering only word pairs appearing in all the ontologies. In this manner, our method will always face the situation described in case 2 of section 3.2, in which it should select the best assessment from those provided by several overlapping ontologies. So, in addition to the benefits obtained by solving missing cases evaluated above, in the current tests, we will evaluate the absolute performance of the proposed heuristic, which takes the assessment with the highest similarity as the final result (as described in section 3.2).

In this case, only 24 of the 29 word pairs of Pedersen's benchmark and 35 of 36 word pairs of Hliaoutakis' benchmark have been found in all the ontologies. In order to quantify the differences between each individual ontology, we first computed the correlation between the similarity values obtained for each one with respect to the others. For the Hliaoutakis' benchmark the correlation between the similarity computed for each word pair in SNOMED CT with respect to the same word pairs when evaluated in MeSH was 0.636, between WordNet and SNOMED CT was 0.505 and between WordNet and MeSH was 0.630. The relatively low correlation values show a discrepancy on the way in which knowledge is represented in each ontology for this benchmark (especially for WordNet with respect to SNOMED CT) and, in consequence, a higher variance on the similarity values obtained for each ontology with respect to the same pair of words. For the Pedersen's benchmark the correlation between ontologies were much higher and constant: 0.914 for SNOMED with respect to MeSH, 0.914 for WordNet with respect to SNOMED CT and 0.904 for WordNet with respect to MeSH. In this case, the three ontologies model Perdersen's words in a very similar manner and the differences between ratings (and the potential improvement after the heuristic is applied in a multi-ontology setting) would be less noticeable than for the Hliaoutakis' benchmark.

The re-evaluation of the same scenarios introduced in the previous section and the comparison of the similarity values with respect to human ratings results in the correlation values shown in Table 11.

Analyzing the results new conclusions arise. On one hand, those ontologies which, for the previous tests, presented a high amount of missing concepts, now offer a higher correlation as they are not hampered by missing term pairs (e.g. 0.782 vs. 0.769 for MeSH with Pedersen's coders). In other cases in which all the word pairs were available, the correlation is lower, showing that some accurately assessed word pairs were removed (e.g. 0.777 vs. 0.788 for SNOMED CT with Pedersen's coders).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 11. Correlation values obtained by the proposed method for Pedersen *et al.*'s benchmark (Pedersen, Pakhomov et al. 2007) (with 24 word pairs) for the ratings of physicians, coders and both and for Hliaoutakis' benchmark (Hliaoutakis 2005) (with 35 pairs).

Ontologies	Pedersen Physicians	Pedersen Coders	Pedersen Both	Hliaoutakis
SNOMED CT	0.588	0.777	0.717	0.558
MeSH	0.588	0.782	0.716	0.7496
WordNet	0.541	0.745	0.6745	0.610
SNOMED CT + WordNet	0.610	0.787	0.734	0.727
MeSH + WordNet	0.580	0.775	0.708	0.772
SNOMED CT + MeSH	0.615	0.798	0.744	0.740
SNOMED CT + MeSH + WordNet	0.638	0.812	0.760	0.786

We can see again that the combination of several ontologies leads to better results in most cases. Even though, the increase in the correlation is lower than for the first battery of tests because there are no missing concepts to solve (e.g. 0.787 vs. 0.799, 0.775 vs. 0.79 and 0.798 vs. 0.817 for Pedersen's coders for SNOMED CT + WordNet, MeSH + WordNet and SNOMED CT + MeSH respectively). Only the combination of MeSH and WordNet provided a slightly lower correlation than when using MeSH alone (for the Perdersen's benchmark), even though it significantly improves WordNet's correlation alone.

In the same manner as in the previous tests, the combination of all the three ontologies leads to the best results, showing the benefits of integrating the assessments from different ontologies even when missing word pairs are not considered. Moreover, the correlation improvement is also coherent with the differences observed between each ontology. For example, the assessments based on SNOMED CT for the Hliaoutakis' benchmark improved from 0.558 to 0.727 when WordNet is also used; as stated above, the correlation between those ontologies was 0.505 (*i.e.* they model Hliaoutakis' words in a noticeable different manner). For the case of the similarity based on SNOMED CT for the Perdersen's coders, the results slightly improved from 0.777 to 0.787 when WordNet is included; as stated above, both ontologies presented a high correlation of 0.914 indicating that Perdersen's words are modelled in a very similar way and, in consequence, the potential benefits of combining them would be less noticeable.

All those results show that the proposed heuristic which selects the most appropriate assessment for overlapping terms (*i.e.* the one with the highest similarity) behaves as hypothesized in section 3.2.

3.3.2.3 Comparison against related work

Finally, in order to directly compare our method in a multi-ontology setting with the one proposed by (Al-Mubaid and Nguyen 2009) (which represents the most recent and complete related work), we reproduced their biggest test. In that case, the Rubenstein and Goodenough benchmark was joined to Pedersen's and Hliaoutakis' biomedical benchmarks individually and to both of them at the same time. Note that, with regards to the human ratings, only those provided by the medical coders for the

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Pedersen's benchmark were used. The reason argued by Al-Mubaid and Nguyen was that medical coders' judgments were more reliable than physicians' ones because more human subjects were involved (9 coders vs. 3 physician) and because the correlation between coders were higher than between physicians (0.78 vs. 0.68).

In (Al-Mubaid and Nguyen 2009), the set of word pairs resulting from joining the two and three benchmarks were evaluated against the combination of MeSH and WordNet in first place, and against SNOMED CT and WordNet in second place. WordNet was selected as the primary ontology in all their tests. Obviously, being the Rubenstein and Goodenough word pairs general terms, they can only be found in WordNet, whereas the rest can be found in both WordNet and the medical ontologies in an overlapping way. It is important to note that the human rating scores of the benchmarks of Pedersen *et al*., Hliaoutakis and Rubenstain and Goodenough have to be converted to a common scale in order to properly compute the final correlation value.

In a first experiment, we used WordNet and MeSH ontologies. As stated above, 25 out of 30 pairs of Pedersen's benchmark and all the 36 pairs of Hliaoutakis's benchmark were found in MeSH. Following the experiment performed in (Al-Mubaid and Nguyen 2009), missing word terms were removed. Their correlation values in comparison with those obtained in our test are shown in Table 12.

Table 12. Correlation values obtained when joining Rubenstain and Goodenough (R&G) benchmark (65 words) with Pedersen *et al.*'s benchmark (with 24 pairs) (only coders' ratings are considered), and Hliaoutakis' benchmark (with 36 pairs) using MeSH and WordNet.

Method	Ontologies	*R&G + Ped. (Coders)*	*R&G + Hliaoutakis*	*R&G + Ped. (Coders)+ Hliaoutakis*
Al-Mubaid & Nguyen (Al-Mubaid and Nguyen 2009)	MESH + WordNet	0.808	0.804	0.814
Our Method	MESH + WordNet	**0.848**	**0.825**	**0.830**

In the second experiment, WordNet and SNOMED CT ontologies were used. Again, 29 out of 30 pairs of Pedersen *et al*.'s benchmark and 35 out of 36 pairs in Hliaoutakis' benchmark were found in SNOMED CT. Missing word pairs were removed. The results are shown in Table 13.

Table 13. Correlation values obtained when joining Rubenstain and Goodenough (R&G) benchmark (65 words) with Pedersen *et al.*'s benchmark (with 29 pairs) (only coders' ratings are considered) and Hliaoutakis' benchmark (with 35 pairs) using SNOMED CT and WordNet.

Method	Ontologies	*R&G + Ped.(Coders)*	*R&G + Hliaoutakis*	*R&G +Ped. (Coders) + Hliaoutakis*
Al-Mubaid & Nguyen (Al-Mubaid and Nguyen 2009)	SNOMED CT + WordNet	0.778	0.700	0.757
Our Method	SNOMED CT + WordNet	**0.850**	**0.811**	**0.816**

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Analyzing both tables, in all cases, our method is able to improve the results reported in (Al-Mubaid and Nguyen 2009). This is motivated both from the highest accuracy of the underlying similarity measure (in comparison with path-based ones) and because of the method to select the most appropriate assessment for overlapping word pairs. In this last case, the fact of relying on a primary ontology implies that Al-Mubaid and Nguyen's method, in many situations, omits potentially more accurate assessments which could be obtained from ontologies considered as secondary. On the contrary, our approach evaluates each word pair and ontology individually and homogenously (which avoids the necessity of pre-selecting a primary one). This exploits the punctual benefits that each one may provide with regards to knowledge modeling. In summary these results support the hypothesis introduced in section 3.2 about the benefits that our approach is able to provide not only in cases in which the pair of concepts belongs to a unique ontology, but also with multiple and overlapping ontologies.

3.4 Summary

In this chapter, we have studied different approaches to compute the semantic similarity from different ontologies. Several limitations were identified, such as the fact of relying on a predefined primary ontology (omitting the benefits that secondary ontologies may provide), the necessity to scale the results in order to compare them and the complexity of the casuistic during the integration of partial results.

As a result of this analysis, we proposed a new method (Batet, Valls et al. 2010c) for computing the semantic similarity from multiple ontologies, which avoids most of the limitations and drawbacks introduced above. All the situations in which concepts may appear in several ontologies are considered and summarized into three cases. A solution for integrating partial or overlapping knowledge in a way in which the semantic assessment could be improved is also proposed. Several advantages were achieved, such as the transparent management of input ontologies, the direct comparison of partial results and the exploitation of the punctual benefits offered by individual ontologies, regarding to the amount of explicitly modelled semantic evidences. In this manner, the final similarity value is the result of an assessment of the best similarity estimation obtained from all the input ontologies.

Part of those benefits and the improvements in the accuracy observed in the evaluation are a consequence of the design of the measure employed to assess the similarity between a pair of concepts, which is the new ontology-based similarity proposed in section 2.2.1.

As shown in the evaluation, our proposal clearly outperforms the previous works. General purpose ontologies (WordNet) and overlapping biomedical ones (SNOMED CT and MeSH) and several standard benchmarks have been used in order to enable an objective comparison with the results reported in related works. On one hand, the underlying measure provided the highest correlation with respect to human evaluations, when compared to other ontology-based measures. On the other hand, our method of computing similarity from multiple ontologies was able to increase even more the accuracy by solving missing cases and by heuristically selecting the

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

best individual assessment. As a result, in most cases, the accuracy is increased with regards to mono-ontology assessments. It was also interesting to observe that even a general ontology such as WordNet was able to improve the results obtained by individual medical ontologies. In fact, the highest correlation corresponded to the case in which all the ontologies were exploited.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Chapter IV

4 Clustering

Clustering is of great importance in many real applications in KDD (Knowledge Discovery and Data Mining) because they can either require a clustering process or can be reduced to it (Nakhaeizadeh 1996).

In this chapter we will focus our work in enlarging the scope of clustering methods in such a way that they can profit of the benefits of having background additional knowledge about the domain. In particular, we will extend a hierarchical clustering algorithm to the possibility of including also *semantic features* together with numerical and categorical variables.

As stated in the introduction, the extensive use of information and communication technologies provides access to a large amount of data that may be semantically interpreted. For example, questionnaires can contain questions whose response has an associated semantics, like questions about the "*Main hobby*". Even though they can be treated as simple strings, and for extension be managed as categorical values, an adequate treatment of this type of data should improve the quality and the interpretability of the clustering results.

Different to categorical values, now values represent concepts rather than simple modalities. Literature is abundant on proposals for calculating the distance for numerical and categorical variables (*e.g.* (Gower 1971; Ichino and Yaguchi 1994; Gibert, Nonell et al. 2005)). In chapter 2 and 3 it was shown how the use of one or several ontologies would certainly improve the comparison between semantic terms with regards to what can be done with classical categorical features. It has been seen how, taking into account the conceptual meaning of the qualitative terms, when available, will certainly provide more accurate estimations of their degree of similarity. As a consequence, applying the results obtained in the first part of this research to the clustering processes should also have benefits on having a better identification of the clusters than non-semantic clustering.

This chapter focuses on the introduction of the concept of semantic features into the clustering methods. The chapter presents a clustering method that can deal with heterogeneous data matrices which include simultaneously numerical, categorical and *semantic features.* In the kernel of the clustering algorithm, comparisons between objects are done by means of a compatibility measure which takes into account the different types of feature and permits to make a homogeneous treatment of all features, in spite of the heterogeneity of their scale of measurement.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

The chapter is organized as follows:
Section 4.1 provide a survey of clustering algorithms and a discussion of the different approaches, highlighting the advantages and drawbacks.

In section 4.2 our contributions in semantic clustering are presented. Section 4.2.1 presents the compatibility measure used to compare objects which features can be numerical, categorical and semantic, and section 4.2.2 presents details of the clustering algorithm.

The final section presents the conclusions.

4.1 Survey of clustering algorithms

The goal of this section is to provide a review of clustering techniques. Clustering methods are unsupervised techniques that aim to discover the underlying structure of a data set. This approach must be distinguished from *Classification* or *supervised methods*, which learn how to assign instances to predefined classes or categories. In the latter model, the classifier is trained using data from the different classes. So, a (training or learning) set of labelled objects is used to build a classifier for the categorization of future observations. A third typology is denoted as *semi-supervised clustering*. These algorithms try to improve the results of the unsupervised methods adding some extra knowledge of the experts. Those methods explore different approaches to guide the clustering process, like the introduction of different types of constraints (Basu, Davidson et al. 2008; Huang, Cheng et al. 2008) (e.g. cluster size balancing, pairwise constraints for object's relationships) or the use of domain-dependent rules (Gibert and Cortés 1998; Valls, Batet et al. 2009; Gibert, Rodríguez-Silva et al. 2010). It has been seen that the use of this additional background knowledge helps to improve the coherence of the obtained results.

In this work we are interested in clustering methods, because we want to address problems where there is no information about the structure of the data, so the goal is to discover a possible classification of objects. So, the rest of this document is focused on this type of methods.

There are many families of clustering algorithms. According to the properties of the generated clusters, the most well-known and extensively used families can be divided into hierarchical clustering and partitional clustering. In section 4.1.3 other families are presented.

- **Partitional clustering** are methods that provide a division of the set of data objects into a partition (non-overlapping subsets such that each data object is exactly in one subset). Most of the times, it attempts to find a c-partition of I where c is a pre-specified number indicating the amount of desired clusters ($c \leq n$).
- **Hierarchical clustering** methods attempt to construct a tree-like nested structure over I. These methods create a hierarchical decomposition of the given data set, producing a binary tree known as a *dendogram* (see section 4.1.2).

In the rest of the section these two approaches are reviewed.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

4.1.1 Partitional clustering

Partitional clustering assigns a set of *n* objects into *c* clusters with no hierarchical structure where each group must contain at least one object and each object must belong to one group. It is important to note that in this clustering approach, most proposals require the number of clusters to be found as an input. Clusters are usually build done on the basis of some specific criterion, so one of the important factors in partitional clustering is the *criterion* function (Hansen and Jaumard 1997).

Partitioning methods are divided into two major subcategories:

- The *centroid-based* algorithms represent each cluster by using the centre of gravity of the objects, with an artificially created prototype. This approach has the problem of defining a method for generating this prototype, which is usually based on calculating some sort of average of the values of the objects. The definition of an averaging function hampers the application to non-numerical variables. Different methods for aggregating values and build the prototype have been defined for categorical variables, such as Huang (Huang 1998) and Gupta *et al.* (Gupata, Rao et al. 1999). If an ordinal relation can be defined on the categorical values, then specific averaging operators are defined (Godo and Torra 2000). Other solution consists in using median operators to build the prototype. (Domingo-Ferrer and Torra 2003; Beliakov, Bustince et al. 2010). Diday (Diday 2000) provides a method to calculate the centroid when both qualitative and qualitative variables are used and proposes a compatibility measure to work with in partitional methods. This problem is not specific for partitional methods and can also be found in hierarchical methods. In Gibert (Gibert and Cortés 1998) an alternative proposal to calculate the centroid when both qualitative and quantitative variables coexists is presented, which could also be useful for partitional methods, provided that the proper mixed metrics is used.
- The *medoid-based* algorithms represent each cluster by means of the object of the cluster whose average dissimilarity to all the objects in the cluster is minimal, *i.e.* it is a most centrally located point in the cluster. This approach avoids the problem of calculating an artificial prototype. It only requires the definition of a distance between objects and any compatibility measure could be used in this case

The most important algorithm for partitional clustering is called k-means. It is present in most statistical software packages. Several variations of this algorithm can be found. They are reviewed in the following.

k-means is the most well-known centroid algorithm (Forgy 1965; MacQueen 1967). K-means 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 centroids.
```

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011


```
Step 2) Assign each object to the nearest cluster
        (i.e. the cluster prototype or centroid)
        on the basis of the criterion function.
Step 3) Recalculate the centroid of each cluster
        based on the current partition.
Step 4) Repeat steps 2)–3) until there is no
        change for each cluster or when a number
        of beforehand defined iterations is done.
```

The advantages of this algorithm are its simplicity and its time complexity-class (can be used to cluster large data sets). The stopping criterion usually needs a small number of iterations making this algorithm very efficient. The disadvantages of this method are: (1) it is sensitive to the selection of the initial seeds of classes and there is no efficient method for identifying neither the initial seeds nor the number of clusters. Usually, the strategy followed is to run the algorithm iteratively using different random initializations. However, some authors studied the initialization of the method (Kaufman and Rousseeuw 1990; Mirkin 2005). (2) The iterative procedure of k-means cannot guarantee convergence to a global optimum (minimum global variance, although it can guarantee the minimum variance inside of a cluster or local optimum). (3) Due to its initial randomness, obtaining the same results for all the executions cannot be guaranteed. (4) K-means is sensitive to outliers and noise because even if an object is quite far away from the cluster centroid, it is still forced to be in the cluster, which distorts the cluster shapes. (5) It is order-dependent, as most partitional methods, which means that composition of final classes hardly depends on the order in which the elements in the target data set are processed, which means that results are not robust to permutations on the rows of the data matrix.

However, there are variants of the k-means which solve some of these limitations. In the following we briefly mention them:

- *PAM* (Kaufman and Rousseeuw 1990) (partitioning around medoids) is an early k-medoid algorithm that uses the data points (medoids) as the cluster prototypes avoiding the effect of outliers. PAM has a drawback that it works inefficiently for a large data set due to its time complexity (Han, Kamber et al. 2001).
- *CLARA* (Kaufman and Rousseeuw 1990) was developed to solve the problem of a large data set. CLARA applies the PAM to sampled objects instead of all objects.
- *ISODATA* algorithm (iterative self-organizing data analysis technique) (Ball and Hall 1965): it is a variation of the k-means that employs a technique of merging and splitting clusters. A cluster is split when its variance is above a pre-specified threshold, and two clusters are merged when the distance between their centroids is below another pre-specified threshold. So, ISODATA can estimate the number of clusters with these merging and splitting procedures. ISODATA considers the effect of outliers in clustering procedures, but results hardly depends on the variance threshold used.
- *GKA* (genetic -means algorithm)(Krishna and Murty 1999): it is designed in order to avoid getting stuck in a local optimum, it can find a global optimum. The method hybridizes the well-known genetic algorithm based on selection, mutation and crossover with classical gradient descend algorithms used in clustering. In fact, k-means is used to as a crossover operator.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- The *k-modes algorithm* (Huang 1998): it uses the simple matching coefficient measure to deal with categorical attributes.
- The *k-prototypes algorithm* (Huang 1998): this algorithm uses the definition of a combined dissimilarity measure and integrates the k-means and k-modes algorithms to allow clustering of instances described by mixed attributes.
- The *X-means algorithm* (Pelleg and Moore 2000): this method automatically finds the number of clusters by using a binary k-means, combined with internal validity indices. At each step a k-means with K = 2 is executed to find a division in two clusters. If the split increases the overall value given by the internal validity indices, the cluster is split and the binary k-means continues execution, recursively. If it is no possible to divide any cluster obtaining an improved validity index, the algorithm stops and takes the current partition as result.

4.1.2 Hierarchical clustering

Hierarchical clustering (HC) algorithms organize data into a hierarchical structure according to a proximity matrix. A distance matrix is a n x n symmetric matrix defined from a data set with n input objects whose (i,j)th element represents the distance between the ith and jth objects.

- The result of the clustering is a hierarchical classification of the objects following a taxonomy of is-a relations (i.e. class-subclass), known as *dendogram*. The *root* node of the dendrogram represents the whole data set χ and each *leaf* node is a single object i; the rest of intermediate nodes correspond to nested clusters that group more similar objects as near the bottom of the tree they appear. The tree is a taxonomy with is-a relations. Overlapping between clusters is not admitted. The internal nodes of the dendogram have an associated numerical value, named *level index*, indicating the degree of homogeneity of the objects included in the cluster represented by that node, which is related with the intra-cluster cohesion. The set of classes is found by means of a horizontal cut of the dendrogram, which determines a set of non-connected subtrees of the dendrogram. The leafs of each subtree determine the elements of every class and they constitute a partition of the data set.

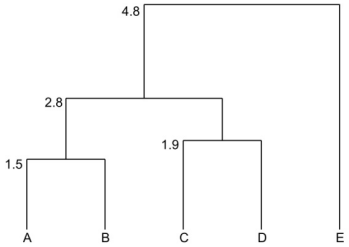


Figure 5. Dedogram.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

So objects belong to a set of nested clusters, which decrease homogeneity as the cluster is placed higher on the tree. The final partition is obtained by means of an horizontal cut of the dendrogram There is a lot of work to define criteria for determining the optimum level of the cut (*e.g.* (Dunn 1974; Rousseeuw 1987)). One of the most frequently used is the Calinski-Harabasz index that maximizes the ratio between the inertia between- and within- clusters (Calinski and Harabasz 1974).

It is worth to note that, if the function used to measure the distances between objects is a metric (i.e. it fulfils the triangle inequality), the dendrogram keeps the ultrametrics structure and Huygens theorem of decomposition of inertia holds (Dillon and Goldstein 1984), which is the basis for the Calinkski-Harabaz criteria to decide the final number of classes.

HC algorithms are mainly classified according on the sense of construction of the dendogram as *agglomerative clustering* and *divisive clustering*.

4.1.2.1 Agglomerative clustering

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.

This type of clustering is the most used and fulfils the properties of sequentiality and exclusivity, also known as SAHN (Sneath and Sokal 1973) (*Sequential, Agglomerative, Hierarchic and Nonoverlapping*).

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 cluster 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 simplest and most popular methods are:

- *Single linkage* (Sneath 1957): the distance between two clusters is determined as the minimum of the distances between all pairs of objects in both clusters. This method produces a reduction of the objects' space since it is taking the minimum distance at each step. One interesting consequence is that small changes between a pair of objects do not significantly modify the dendogram, meaning that the process is non-sensitive to small variations. Single-linkage is sensitive to noise and outliers. Its tendency is to produce straggly or elongated clusters (chaining effect).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- *Complete linkage* technique (Sorensen 1948): the distance between two clusters is determined as the maximum of all pairwise distances objects in the two clusters. This approach produces an expansion of the object's space. Complete linkage is less susceptible to outliers and noise. An interesting property is that it can break large clusters and produces compact clusters (Baeza-Yates 1992). On the contrary to Single Linkage, this method is conservative because all pairs of objects must be related before the objects can form a cluster. In general this algorithm produces more useful hierarchies in many applications than Single linkage (Jain and Dubes 1988).
- *Average linkage:* the distance between two clusters is computed as the average of the distance among all the objects of the two clusters. The average can be calculated in different ways, but the most common is to use the arithmetic mean. Another version weights each object according to the number of elements of the cluster to which it belongs. In this case it is called "*group average*"(Sokal and Michener 1958). There are other ways to assign weights to objects, such as depending on how the objects have been successively incorporated to the cluster.
- *Centroid linkage:* this approach considers also an artificial object that is built as the prototype of a cluster. The distance between two clusters is defined as the distance between their centroids. The centroid is calculated using some aggregation operator, usually the arithmetic mean, on each of the attributes that describe the objects.
- *Median linkage*: the distance between two clusters is based on an artificial point that is taken as the median of the objects in the cluster (Gower 1967). This solves a drawback of the centroid approach, because if two clusters with very different size are fused, the centroids will have different degrees of representativeness with respect to their clusters. Considering that the centroid of the new cluster will lie along the median of the triangle defined by the clusters that are forming a new group and an external one, the median is proposed for the similarity computation.
- *Minimum-variance loss* or *Ward's method*: (Ward 1963):the clusters that minimize the loss in the inertia between classes are aggregated. 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 provides obtains a more optimum partition of the objects. 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.
- *Reciprocal-neighbours linkage:* A pair of reciprocal neighbour's clusters is merged in a new cluster at every iteration (Rham 1980). This is particularly interesting because it is cheaper than quadratic.

Those methods presented until now are the most well-known hierarchical clustering techniques. In spite of their differences, they share a common way of construction of the dendogram. Lance and Williams (Lance and Williams 1967) proposed a parameterized updating formula to calculate distances between a new cluster and existing points, based on the distances prior to forming the new cluster, provided that the classical Euclidean metrics was used in a context where strictly

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

numerical variables are used. This recursive approach avoids the recalculation of the distances with respect to all the objects that belong to the compared clusters . This formula has three parameters, and each of the clustering methods can be characterized by its own set of Lance-Williams parameters (see Table 14).

- Using the notation of Lance-Williams, let d_{ij} be the distance between points *i* and *j* and let $d_{k(ij)}$ be the updated distance of point *k* to the newly formed cluster (*ij*). Thus, d_{ij} is a within cluster distance and $d_{k(ij)}$ becomes a distance between clusters. The recursive formula is defined as:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha j d_{kj} + \beta d_{ij} + \lambda | d_{ki} - d_{kj} | \quad (59)$$

- The α, β and λ variables are the parameters that define the linkage process. The following table shows the values of these parameters for the methods presented before. The n_i values refer to the number of elements in cluster *i*. One feature of the recurrence formula is that any hierarchical clustering scheme which satisfies the relation will also possess a unique set of parameter values.

Table 14. Hierarchical Algorithms

Method	α_i	α_j	β	λ	Monotonic/ Ultrametric
Single Linkage	1/2	1/2	0	-1/2	Yes
Complete Linkage	1/2	1/2	0	1/2	Yes
Group Average	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0	Yes
Arithmetic Average	1/2	1/2	0	0	Yes
Centroid	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	$-n_in_j/(n_i+n_j)^2$	0	No
Median	1/2	1/2	-1/4	0	No
Minimum Variance (Ward)	$(n_i+n_k)/(n_i+n_j+n_k)$	$(n_j+n_k)/(n_i+n_j+n_k)$	$-n_k/(n_i+n_j+n_k)$	0	Yes

The n_i values refer to the number of elements in cluster i.

However, previous works shown (Gibert, Nonell et al. 2005) that extending the Lance-Williams recursive updating of distances to the general case of heterogeneous data matrices and non-metrical approaches is extremely complex and recursive implementations hardly limit the flexibility to incorporate new comparisons between objects and new feature types into the clustering algorithms, and it is preferable keep the original distances matrix for internal calculations in spite of some extra computing time (not too critical)

Agglomerative clustering methods can also be divided according to the way of representing the clusters:

- *graph methods*: consider all points of a pair of clusters when calculating their inter-cluster distance.
- *geometric methods*: use geometric centers of the clusters in order to determine the distance between them.

Table 15. Classification of the hierarchical methods presented in this section.

Graph methods	Single linkage, complete linkage and average linkage.
Geometric methods	Centroid linkage, median linkage and Ward's method and reciprocal neighbours

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

As said before, most of these algorithms are quadratic; some are cheaper, but always greater than linear, like partitional methods. In recent years, with the requirement for handling large-scale data, new hierarchical techniques have appeared with the aim to minimize the computational cost of the classical algorithms and achieve scalability. Some examples include:

- *BIRCH* (Zhang, Ramakrishnan et al. 1996) (Balanced Iterative Reducing and Clustering using Hierarchies) is an incremental and hierarchical clustering algorithm for very large databases. The two main building components in the Birch algorithm are a hierarchical clustering component, and a main memory structure component. Birch uses a *main memory* (of limited size) data structure called *CF tree*. The tree is organized in such a way that (i) leafs contain current clusters, and (ii) the size of any cluster in a leaf is not larger than *R*. Initially, the data points are in a single cluster. As the data arrives, a check is made whether the size of the cluster does not exceed *R*. If the cluster size grows too big, the cluster is split into two clusters, and the points are redistributed. The tree structure also depends on the branching parameter *T*, which determines the maximum number of children each node can have.
- *CURE* (Clustering Using REpresentatives) (Guha, Rastogi et al. 2001): it represents a cluster by a fixed number *h* of points scattered around it. The distance between two clusters used in the agglomerative process is equal to the minimum of distances between two scattered representatives. This approach is an adaptation of the Single Linkage method but working with representatives. Therefore, CURE takes a middle-ground approach between the graph (all-points) methods and the geometric (one centroid) methods. CURE is capable of finding clusters of different shapes and sizes, and it is insensitive to outliers. CURE was designed to work with numerical values.
- *ROCK (Guha, Rastogi et al. 2000)* (The RObust Clustering using linKs) clustering algorithm is based on links between data points, instead of distances when it merges clusters. These links represent the relation between a pair of objects and their common neighbours. The notion of links between data helps to overcome the problems with distance based coefficients. For this reason, this method is extended to non-metric similarity measures that are relevant in situations where a domain expert/similarity table is the only source of knowledge. ROCK works with categorical features.
- *RCH* (Relative hierarchical clustering) considers both the internal distance (distance between a pair of clusters which may be merged to yield a new cluster) and the external distance (distance from the two clusters to the rest), and uses their ratio to decide the proximities (Mollineda and Vidal 2000).
- *SBAC* (similarity-based agglomerative clustering) which was developed by Li and Biswas (Li and Biswas 1999) extends agglomerative clustering techniques to deal with both numeric and nominal data. It employs a mixed data measure scheme that pays extra attention to less common matches of feature values (Li and Biswas 2002).
- *CHAMELEON* (Karypis, Han et al. 1999). It uses dynamic modelling in cluster aggregation. It uses a connectivity graph corresponding to the K-nearest neighbour model of sparsification of the proximity matrix, so that the edges of the k most similar points to any given point are preserved, and the rest are pruned.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

CHAMELEON has two stages. In the first stage small tight clusters are built to ignite the second stage. In the second stage an agglomerative process is performed.

- Hierarchical clustering of subpopulations, which assumes a known distribution for the variables for the subpopulations of the dataset. The dissimilarity between any two populations is defined as the likelihood ratio statistic which compares the hypothesis that the two subpopulations differ in the parameter of their distributions to the hypothesis that they do not. A general algorithm for the construction of a hierarchical classification is presented in (Ciampi, Lechevallier et al. 2008).

4.1.2.2 Divisive clustering

The divisive approach proceeds in a top-down manner. Initially, the entire data set belongs to a unique cluster and a procedure successively divides it until all clusters are singleton clusters.

The crucial points of this type of clustering are (1) the definition of a coherence function in order to select the next cluster to split, and (2) the definition of the splitting function. The former can be resolved by calculating the variance in the cluster and selecting the cluster with the highest variance for splitting, or simply detecting the largest cluster for splitting. About the latter, the splitting task usually consists on putting the data points into two different clusters. This type of methods has been less exploited because the agglomerative approach is more efficient, and they use to collapse as soon as a new cluster contains small data. However, they may become more popular in the near future where massive data sets are more and more available.

Some divisive clustering algorithms are:

- *Bi-Section-Kmeans*: this clustering algorithm is an extension of the basic k-means, that divides one cluster in two (for k=2) at each step. Different criteria can be used to establish the division. The process ends when the desired number of clusters has been generated.
- *DIANA(DIvisive ANAlysis)*(Kaufman and Rousseeuw 1990): is a heuristic method that consists on considering only a part of all the possible divisions at each step. Consists of a series of iterative steps to move the objects to the closest splinter. The splinter is initialized with the object that is farthest from the others.
- *MONA* (monothetic analysis) (Kaufman and Rousseeuw 1990). When all the features are used together the algorithm is called polythetic. Otherwise, it is called monothetic, because only one feature is considered at each step. In (Kaufman and Rousseeuw 1990) this approach is used with binary features, where the similarity is computed through association measures.
- *DIVCLUS-T* (Chavent, Lechevallier et al. 2007) is a divisive hierarchical clustering algorithm based on a monothetic bipartitional approach allowing the dendrogram of the hierarchy to be read as a decision tree. It is designed for either numerical or categorical data. Like the Ward algorithm, it is based on the minimization of the inertia criterion. However, it provides a simple and natural interpretation of the clusters.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

4.1.3 Clustering techniques in AI

The Statistics field has been the pioneer in developing most of the techniques explained in the previous section. With the appearance of the Artificial Intelligence discipline in 1950s, new clustering methods based on symbolic treatment of data were studied. Those methods rely on logic models instead of algebraic criteria, as it has been done before. The main representative is conceptual clustering, developed in the 1980 in conceptual clustering. Cluster formation is not only driven by the inherent structure of the data that drives cluster formation, but also the description language based on logics, usually limited to feature conjunction. This language is used to build a model that determines which data belong to each cluster. This means that each cluster is considered as a model that can be described intensionally, rather than as a collection of points assigned to it. A popular method for categorical data is COBWEB (Fisher 1987), which uses simple distribution functions to build the model of the clusters. Conceptual clustering is closely related to data clustering. In fact, COBWEB is a hierarchical algorithm that uses incremental learning instead of following divisive or agglomerative approaches.

Several other clustering approaches can be found in the literature. A good survey is done in the book of Xu et al. (Xu and Wunsch 2005). Here we present a summarized list of those techniques:

- *Densities-Based Clustering*: Here a cluster is understood as a dense region of objects that is surrounded by a region of low density. A known algorithm of this type is DBSCAN (Ester, Kriegel et al. 1996), which assigns the points that are close enough in the same cluster. Likewise, any border point that is close enough to a core point is put to the same cluster as the core point. However, noisy points are discarded producing a non-complete clustering.
- *Neural Networks-Based Clustering*: Here objects are represented as neurons, these neurons increases the neighbourhood in some regions creating clusters and decrease it with other neurons. Some examples of this kind of algorithms are LVQ (Learning vector quantization), SOFMs (Self-Organized Feature Maps) and ART (Adaptative Resonance Theory (Kohonen 1997)).
- *Graph Theory-Based Clustering*: Here the data are represented as a graph where the nodes are objects and the links represent connections between objects. Then a cluster is defined as a group of objects that are connected between them but that have not connections with objects outside the group. A well-known graph-theoretic divisive clustering algorithm is based on the construction of the *minimal spanning tree* (MST) of the data (Zahn 1971), and then deleting the MST edges with the largest lengths to generate clusters.
- *Kernel-based clustering*: The basis of this approach is that with a nonlinear transformation of a set of objects into a higher-dimensional space, one can easily find a linear separation of these objects into clusters. So, the goal is to change the space of representation of the objects. However, building a nonlinear mapping in the transformed space is usually a time-consuming task. This process can be avoided by calculating an appropriate inner-product kernel. The most common kernel functions include polynomial kernels and Gaussian radial basis functions (RBFs) and sigmoid kernels (Corchado and Fyfe 2000).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- *Fuzzy clustering*: while traditional clustering approaches generate partitions where each data object belongs to one and only one cluster, *fuzzy* clustering extends this notion to associate each data object with every cluster using a membership function (Zadeh 1965). Larger membership values indicate higher confidence in the assignment of the object to the cluster. The most widely used algorithm is the Fuzzy C-Means (FCM) algorithm (Sato, Sato et al. 1997), which is based on k-means. FCM attempts to find the most characteristic point in each cluster, which can be considered as the "center" of the cluster and, then, the grade of membership of each instance to the clusters. Many extensions of FCM are still being developed, such as (Hamasuna, Endo et al. 2010).

4.1.4 General comments of the different approaches

In this section we make an analysis of the different clustering methods introduced in this document, focusing mainly in hierarchical and partitional techniques. Several observations are commonly done about these techniques, which are really relevant for the user in order to select the appropriate methodology for a particular problem.

The first concern is about the fact that the selection of the clustering algorithm determines some characteristics of the clusters that are obtained. For example, centre-based algorithms as the k-means will produce compact and spherical groups, and do not perform well if the real classes from the target domain are not spherical. Hierarchical methods organize groups on a multi-level groups and subgroups structure, providing alternatives of few general classes or more specific classes, which can be interesting in some particular applications. If a partition is generated from the dendrogram, clusters with different characteristics are obtained (see details in section 4.1.2). Other characteristics are obtained in density classification methods, which form groups according to the objects density; therefore, they do not limit the size of the group and very heterogeneous forms of groups can be modelled. If the user has some knowledge about the form of the clusters, then the selection of the clustering algorithm must be done according to this knowledge. For example, the single-link clustering algorithm works well on data sets containing non-isotropic clusters including well-separated, chain-like, and concentric clusters, whereas a typical partitional algorithm such as *k*-means works well only on data sets having isotropic clusters (Nagy 1968).

If there is no information about the form of the clusters, which is the most frequent situation in practical applications, one may consider the advantages and disadvantages of each of the different techniques. 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 provides the inner structure of the data set, so the number of classes can be determined as a result of the process, rather than blindly imposed a priori. They are also non-order-dependent, which guarantees stability of results by permutations of the rows of data matrix. But this also implies a major class-complexity, quadratic in most cases, what makes some of these algorithms prohibitive for massive data sets analysis. On the other hand, the time and space complexities of

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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. However, their main disadvantage is that they suffer from their inability to perform adjustments once the splitting or 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. This rigidity is useful because it leads to a smaller computational cost, since it does not have to worry for a combinatorial number of possible options, but it makes this kind of algorithms not useful for incremental data sets, since it requires the whole data set at the beginning. Moreover, most of the hierarchical methods have the tendency to form spherical shapes and the reversal phenomenon, in which the normal hierarchical structure is distorted.

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, for which there are no theoretical guidelines to select the most effective value.

It can be observed that both approaches have advantages and disadvantages in different aspects. In that sense, it is possible to develop hybrid algorithms that exploit the good features of both categories (Murty and Krishna 1980). Finally, in crisp clustering methods, clusters are not always well-separated. Fuzzy clustering overcomes the limitations of hard classification. However, a problem with fuzzy clustering is that it is difficult to define the membership values of the objects.

As a final and general remark, one can observe that there is not a best algorithm for all the cases. Depending on the purpose of the clustering, the most suitable approach must be selected.

4.2 Semantic clustering approach

In this work we will focus on a hierarchical clustering approach, chosen after analysing the pros and cons explained in the previous section (Batet, Valls et al. 2008; Batet, Valls et al. 2010b; Batet, Valls et al. 2010c; Batet, Gibert et al. 2011). Hierarchical clustering generates a tree-like classification of objects (a dendrogram), which can be analysed at different levels of generality to obtain different partitions. This feature is important in applications where the definition of the number of clusters in advance is not possible.

Hierarchical methods have a critical component: the way of measuring the distance or dissimilarity between a pair of objects or individuals. In fact, the distance between individuals is a clue to decide which individuals will form a new cluster and to determine the structure of the dendrogram.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

In this section, a methodology able to exploit the semantic content of terms when used in unsupervised clustering methods is presented. This semantic interpretation will be introduced to the clustering by considering *semantic features* in addition to numerical and categorical features.

The proposal is based on the classical general clustering algorithm presented before:

```
Given a data set I and
Given a proximity function between a pair of objects d(i,j)
1.  Create a cluster for each object i in I
2.  Compute the proximity matrix between all the clusters
    D(nxn)=(d_ij= d(i,j))
3.  Repeat
4.        Merge the closest two clusters
5.    Update the proximity matrix to reflect the proximity
      between the new     cluster and the original clusters
6.  Until: only one clusters remains
```

As said before, the main differences in hierarchical algorithms are the merging criterion (Single Linkage, Centroid, Ward, etc.) and the way in which the proximity (or distance) between two clusters is computed. With respect to the former, the Ward criterion has been chosen, as it will be explained in section 4.2.2. With respect to the latter, the type of criteria usually determines the way of calculating the proximity/distance.

When heterogeneous types of values must be taken into consideration in a joint way, three main approaches can be used: (Anderberg 1973)

1. The transformation of the values into a common domain (e.g. discretization of numerical variables, or mapping the data into a new space using projection algorithms (Anderberg 1973; Jain and Dubes 1988).
2. Separate the variables upon their type and then reduce the analysis to the dominant type, which is usually determined by the group with the largest number of variables, or depending on some background knowledge on the structure of the domain.
3. The use of compatibility measures that combine different expressions according to the type of each of the variables (Gower 1971; Anderberg 1973; Gowda and Diday 1991; Ichino and Yaguchi 1994; Ralambondrainy 1995; Gibert and Cortés 1997).

The third approach allows the analysis of the different values maintaining the original scales, having several advantages in comparison with the other two approaches: (1) data is analyzed in its original nature, (2) there is no a priori loss of information produced by previous transformations (i.e. discretization of numerical features), (3) all the variables are taken into account in the clustering process, (4) it avoids taking previous arbitrary decisions that could bias results and (5) interactions between different type of variables are exploited in the analysis (Gibert, Nonell et al. 2005).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

In order to take advantage of the potential of semantic similarity measures, a compatibility measure is needed to combine the contribution of numerical, categorical and semantic features into a global function.

Thus, a compatibility measure that combines numerical, categorical and semantic features to generate a hierarchical classification will be presented (Batet, Gibert et al. 2011). Chapter 2 presents several ways of comparing semantic features on the basis of similarities/distances taken from the literature as well as measures based on the assessment of the distance with some background knowledge, such as ontologies (Studer, Benjamins et al. 1998b) or the web. It also introduces our own proposals, SC_{Eu} and SC_{log} to improve results in front of one or more (MSC_{log}, chapter 3) background ontologies. In those chapters was seen that our proposals perform better than other proposals in the literature and MSC_{log} is a generalization of SCl_{og} when more than one ontology are available. That's why the comparison function used in our clustering method will be $SC_{log.}$

After this study, we will base our clustering method on the knowledge available in ontologies, because:

- The ontology-based semantic similarity measure proposed in section 2.2.1 is used in order to compare the values of semantic feature (i.e. terms). Its main advantages are: it has achieved good results in the tests when evaluated using different standard benchmarks, its low computational cost, the fact that it not depends on tuning parameters, the availability of complete general and domain ontologies, and due to ontology-based measures are not affected by the language labels used to refer to concepts because the knowledge represented in an ontology is language-independent (see the full justification in section 2.4).
- A method for integrating the knowledge of multiple ontologies that avoids both the dependence on the completeness of a single ontology and most of the limitations and drawbacks of related works has been developed (section 3.2), allowing to consider the semantics given in different ontologies all together.
- Ontologies are a powerful tool for semantic knowledge management. Moreover, nowadays we can find large ontologies in different domains, as well as general-purpose resources.
- The processing of the ontologies is easy since most of them provide application interfaces.
- Specific corpus or the Web are not always available. The former needs a manual pre-processing in order to achieve good results and about the latter, queries to web search engines have a high computationally.

The clustering approach proposed in this chapter can be used in any domain where the comparison between objects described with these three types of features is required. For example, as it will shown in chapter 5, we have worked in collaboration with the Technological Park for Tourism and Leisure making an analysis to identify profiles of tourists that visit a certain area, which involves numerical (e.g. age), categorical e.g. (sex) or semantic (e.g. motivations) relevant features that describe the tourist itself.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

4.2.1 Generalizing compatibility measures to introduce semantic features

Literature is abundant on references on compatibility measures for numerical and categorical features values can be found in the literature (Ichino and Yaguchi 1994; Ralambondrainy 1995; Gibert and Cortés 1997; Diday 2000; Doring, Borgelt et al. 2004; Ahmad and Dey 2007).

Several metrics for mixed numerical and categorical values can be found in the literature (Gibert and Cortés 1997; Diday 2000; Doring, Borgelt et al. 2004; Ahmad and Dey 2007).

In this section, we propose a generalization of Gibert's mixed metrics (Gibert and Cortés 1997) including semantic features. In (Gibert and Cortés 1997) Gibert introduces the mixed metrics as a weighting between the normalized Euclidean metrics for numerical variables and the Chi-squared metrics, χ^2.

$$d^2_{(\alpha,\beta)}(i,i') = \alpha d^2_{\zeta}(i,i') + \beta d^2_{Q}(i,i') \quad \text{with } (\alpha,\beta) \in [0,1]^3, \alpha+\beta=1 \tag{60}$$

The decision on the Euclidean and Chi-squared metrics are based on the works of (Annichiarico, Gibert et al. 2004; Gibert, Nonell et al. 2005), which propose that combination.

In (Gibert, Nonell et al. 2005) it is compared with other proposals in the literature and it is seen that, for the particular context of clustering methods, this proposal overcomes the performance of other proposals, in the sense, they recognize better the underlying structure of the target data sets.

Many real applications show the validity of this proposal to extract robust clusters in from real data (Gibert and Sonicki 1999; Annichiarico, Gibert et al. 2004; Gibert, Martorell et al. 2007; Gibert, García-Rudolph et al. 2008; Gibert, Garcia Alonso et al. 2010; Gibert, Rodríguez-Silva et al. 2010; Gibert, Rojo et al. 2010) from many different domains.

With respect to the semantic features, the measures developed in chapters 2 (SC_{Eu} and SC_{log}) and 3 (MSC_{log}) will be used.

According to the different types of features (i.e. numerical, categorical and semantic features), the distance between a pair of objects *i* and *i'*, is calculated as the combination of applying a specific distance for each type of feature X_k. Then, the distance is defined in eq. 59 (Batet, Valls et al. 2010b; Batet, Valls et al. 2010c; Batet, Gibert et al. 2011):

$$d^2_{(\alpha,\beta,\gamma)}(i,i') = \alpha d^2_{\zeta}(i,i') + \beta d^2_{Q}(i,i') + \gamma d^2_{S}(i,i') \quad \text{with } (\alpha,\beta,\gamma) \in [0,1]^3, \alpha+\beta+\gamma=1 \tag{61}$$

In this definition we have three main components:

ζ = {k : X_k is a numerical feature, k =1:n_ζ }
Q = {k : X_k is a categorical feature, k =1: n_Q }
S = {k : X_k is a semantic feature, k =1:n_S, linked Θ }

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Such that K= $n_{\zeta+}\, n_{Q+}\, n_{S.}$

The distances for each of type of features are the following:

$d^2_{\zeta}(i,i')$ is the normalized Euclidean distance for numerical features,

$d^2_Q(i,i')$ is a rewriting of the χ^2 metrics for categorical values that can be directly computed on the symbolic representation of categorical data, without requiring split to dummy variables,

$d^2_S(i,i')$ is the MSC_{log} similarity measure.

In (59) each component has an associated weight. The weighting constants (α,β,γ) are taken as functions of the features' characteristics. In particular, they depend on the range of distances of each type of feature and how many variables refer (this will be explained in section 4.2.1.1). So the final expression for the compatibility measure is:

$$d^2_{(\alpha,\beta,\gamma)}(i,i') = \alpha \sum_{k\in\varsigma} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \frac{\beta}{n_Q^2} \sum_{k\in Q} d_k^2(i,i') + \frac{\gamma}{n_S^2} \sum_{k\in S} ds_k^2(i,i') \qquad (62)$$

where s_k^2 is the variance of the numerical feature X_k, n_Q=*card(Q)*, n_S=*card(S)*.

Under analogous philosophy of Gibert's mixed metrics, and 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.

The first component corresponds to the Euclidean distance used for numerical data, the second component $d_k^2(i,i')$ is the contribution of categorical feature X_k according to the Chi-squared metrics. In particular, the Gibert's proposal is based on the idea that χ^2 metrics (Gibert, Nonell et al. 2005) upon categorical variables is directly related with the quantity of information provided by the variable itself (Benzécri 1980). To improve the efficiency, we will use a rewriting of the χ^2 metrics for categorical ones avoiding the splitting to a complete incidence table classically required for χ^2 computation. It is defined as follows (eq 61):

$$d_k^2(i,i') = \begin{cases} 0 & if \quad x_{ik} = x_{i'k} \\ \frac{1}{I_{k^i}} + \frac{1}{I_{k^{i'}}} & if \quad x_{ik} \neq x_{i'k} \\ \frac{(f_i^{k_s} - 1)^2}{I^{k_s}} + \sum_{j\neq s}^{n_k} \frac{(f_i^{k_j})^2}{I^{k_j}} & if \quad x_{ik} = c_s^k, \text{and } i' \text{ is a class} \\ \sum_{j=1}^{n_k} \frac{(f_i^{k_j} - f_{i'}^{k_j})^2}{I^{k_j}} & i \text{ and } i' \text{ are both classes} \end{cases} \qquad (63)$$

In expression (61), I^{k_j} I^{k_j} is the number of objects *i* such that x_{ik}=c_j^k, being c_j^k the *j*-th categorical value of X_k (*j*=1:n_k); I_k^i=card{1:1, $i \in I \& x_{1k} = x_{ik}$}. As *i* represents a

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

subclass of objects and X_k is not constant inside the class, $f_i^{k_j}$ is the proportion of objects from the i^{th} class with value c_j^k. In fact

$$f_i^{k_j} = \frac{I_i^{k_j}}{\sum_{j=1}^{n_k} I_i^{k_j}} \tag{64}$$

Finally, $ds_k^2(i,i')$ is the contribution of semantic feature X_k which is calculated on the basis of our own proposal for computing semantic similarity from multiple ontologies (MSC_{log}) presented in section 3.2 and using as basis the SC_{log} , semantic similarity measure, proposed in section 2.2.1 (when card(Θ) =1 only SC_{log} is used). Since we proposed similarity measures and clustering algorithms work with distances or dissimiliarities the sign of this measure is changed to have dissimilarity rather than a similarity.

4.2.1.1 Weighting indices α, β, γ

The computation of the global distance must balance the influence of each group of features, because each of the terms of the expression (61) is representing the contribution of a pack of variables of a common type, which can be very different, both in magnitude and in the number of variables represented. For this reason, each component of the distance is weighted accordingly. With a similar argument as the one provided in (Gibert and Cortés 1997) and being α, β and γ the weight for numerical, categorical and semantic features, it is enough to index the expression (59) with $(\alpha,\beta,\gamma) \in [0,1]^3$ with $\alpha+\beta+\gamma$ =1 since bounding the addition of weights defines an equivalent relationship on the set of achievable hierarchies, where the same hierarchies available from $(\alpha,\beta,\gamma) \in \mathcal{R}^3$ are found.

In (Gibert and Cortés 1997; Gibert, Sonicki et al. 2002) some heuristic criteria are introduced to find acceptable values for the weighting constants α and β. Several real applications (Annichiarico, Gibert et al. 2004; Gibert, Nonell et al. 2005) shown a successful performance of the original proposal in front of other values, for the particular case of recognizing underlying classes on a given domain.

Here, the same criteria are used to extend the proposal to a third type of features (Batet, Valls et al. 2010b; Batet, Gibert et al. 2011). So, the proposal is to calculate the values for α, β and γ in the following way:

$$\alpha = \frac{\frac{n_\zeta}{d_{\zeta\ \max^*}^2}}{\frac{n_\zeta}{d_{\zeta\ \max^*}^2} + \frac{n_Q}{d_{Q\ \max^*}^2} + \frac{n_S}{d_{S\ \max^*}^2}} \tag{65}$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$\beta = \frac{\frac{n_Q}{d^2_{Q\,\max^*}}}{\frac{n_\zeta}{d^2_{\zeta\,\max}*} + \frac{n_Q}{d^2_{Q\,\max^*}} + \frac{n_S}{d^2_{S\,\max^*}}} \tag{66}$$

$$\gamma = \frac{\frac{n_S}{d^2_{S\,\max^*}}}{\frac{n_\zeta}{d^2_{\zeta\,\max}*} + \frac{n_Q}{d^2_{Q\,\max^*}} + \frac{n_S}{d^2_{S\,\max^*}}} \tag{67}$$

We have $(\alpha,\beta,\gamma) \in [0,1]^3$ with $\alpha+\beta+\gamma=1$, and $n_\zeta=card(\zeta)$, $n_Q=card(Q)$, $n_S=card(S)$ and $d^2_{\zeta\,\max^*}$, $d^2_{Q\,\max^*}$, and $d^2_{S\,\max^*}$ are the truncated maximums of the different subdistances. Using the truncated maximums instead of the absolute maximums will provide robustness in front of multivariate outliers. Maximums are usually truncated to 95% but other possibilities could be considered as well.

This expression has the following properties:

- The proposal represents an equilibrium among the different components of the final distance, since they are referred to a common interval. Dividing every term by the maximum value they can present, the three components will have equal influence on $d^2_{(\alpha,\beta,\gamma)}(i, i')$. So,

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

- The proposal is robust to the presence of outliers, because it is considering truncated maximums. As outliers produce big distances with respect to the other objects, they will not be taken as reference points. Doing this, the other distances will not be concentrated in a subinterval $[0,c_0]$, $c_0 << 1$, avoiding even numerical instability. Moreover, when outliers are not present, the eliminated distances will be almost of the same range as $d^2_{\zeta\,\max^*}$, $d^2_{Q\max^*}$ and $d^2_{S\,\max^*}$ respectively, and the real working interval will be $[0,c_0]$, $c_0 \approx 1$, what does not imply a major change.
- The proposal gives to every subdistance an importance proportional to the number of features it represents. So,

$$\alpha \propto n_\zeta \quad \& \quad \beta \propto n_Q \quad \& \quad \gamma \propto n_S$$

Consequently, we have defined a set of weights (α,β,γ) that depends on the importance of each type of feature and the magnitude of each term.

4.2.2 Clustering algorithm

In this work a hierarchical reciprocal neighbours clustering with Ward criteria (Ward 1963) is proposed.

The algorithm is formalized as follows:

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Input:

- a set $I=\{1,..,n\}$ of objects represented in a data matrix χ
- a data matrix χ (rows can be considered as vectors $x_1,...,x_n$ each one describing one object) in I,
- and a set of ontologies $\Theta=\{o_1,..,o_n\}$ as knowledge source.

Output: a set C of $2n$-1 clusters ordered hierarchically as a binary tree (C,E) with $2(n-1)$ edges and n leaves.

$\forall\, i\ 1\leq i\leq n : c_i:=\{x_i\}$
$C:=C':=\{c_1,\ldots,c_n\}$
$E:=\varnothing$
$j:=n+1$
while $(|C'|>1)$ **do**
$(c_{u'},c_{v'}):=\text{argmin}_{(cu,cv)\in C'xC'}\, Proximity(c_u \cup c_v)$
$c_j = c_{u'} \cup c_{v'}$
$C'=C' \cup \{ c_j \}-\{ c_{u'},c_{v'}\}$
$C=C \cup \{ c_j \}$
$E=E \cup \{(c_{u'}, c_j),(c_{v'}, c_j)\}$
end while
return (C,E)

Proximity(c_u,c_v) is calculated according the aggregation criteria of the reciprocal neighbours. Here, those pairs of reciprocal neighbours are joined in order to construct the aggregation tree.

Definition 10: A pair of objects i and i' are called *reciprocal nearest neighbours (RNN's)* if i is the nearest neighbour for i' ($i=NN(i')$) and vice-versa ($i'=NN(i)$). That is, if $d(i,i'') \leq d(i, i'')$ and $d(i, i') \leq d(i', i'')$, for all i'', $1 \leq i'' \leq n$, $i''\neq i$ and $i'' \neq i'$.

In this work, the *chained reciprocal neighbours algorithm* (Rham 1980) which is based on the concept of RNN is used. At every step, a pair of *RNN* is aggregated in a new class. The chained version is a quick algorithm of $O(n\log n)$ worst case complexity. So, this algorithm requires the definition of a measure in order to compute the distance between two objects, and, thus, to identify the pair of nearest elements (see section 4.2.1).

A combination of Reciprocal Neighbours with Ward's criterion is used. So nearest neighbour is found by minimal inertia rather than simple distance between objects.

In our case, the representative of the inner classes generated during the process is built using the arithmetic mean for numerical features and the mode for non-numerical ones, either semantic or categorical.

As usual in hierarchical clustering, the output can be represented as a dendrogram.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

4.3 Summary

The exploitation of data from a semantic point of view establishes a new setting for data mining methods.

On a hand, in this chapter has been presented a survey of different clustering algorithms. In particular there is a clear division between partitional and hierarchical approaches. Pros and cons have been analyzed.

On the other hand, it has been proposed a method to include semantic features into an unsupervised clustering algorithm. It has been identified that the key step to be modified is the calculation of the similarity between the objects. To achieve the integration of the three types of features, the similarity must be able to deal with this heterogeneous information. Thus, a combination function that combines numerical, categorical and semantic features has been introduced. Special attention has been devoted to analyze the contribution of semantic features, proposing a formulation that permits to estimate the semantic similarity of textual values from a conceptual point of view exploiting ontologies. In particular, the management of the semantic knowledge associated to the values of the data matrix is done by relating the values with concepts in one or more ontologies.

We have also studied the problem of weighting the contribution of the different types of information. When there is no information about the weights of the variables, some automatic process must be defined. We have proposed a formulation that balances the amount of information provided by each type of feature, giving an equilibrated influence of each variable in the calculation of the global distance.

Finally, the most important contribution is the specification of the ontology-based clustering methodology. The Ward method has been taken. It is a well-known clustering method, available in many statistical and data mining packages (see Annex B). Moreover, the Ward clustering method has some interesting properties for the dendograms generated, as explained in this section.

All these facts make us to expect that the results obtained with this new clustering method will be more adequate to identify profiles or typologies of objects.

The next chapter will analyze this hypothesis, presenting the evaluation of this methodology. The results will 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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Chapter V

5 Clustering evaluation

Up to this point, we have presented the main contributions of our work, describing a clustering method that is able to exploit the semantics of words, when at least one ontology is available. Multiple ontologies can be used to assess the semantic similarity of this type of data, which is then combined with the traditional Euclidean and Chi-Squared distances for numerical and categorical data.

The clustering method proposed has been integrated into the KLASS software system (Gibert and Cortés 1998; Gibert, Nonell et al. 2005) 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* allows clustering with numerical or/and categorical data matrices using the metrics introduced before Gibert's mixed metrics, among other 7 different metrical proposals (Gibert, Nonell et al. 2005). Moreover, different clustering algorithms can be used, as the reciprocal neighbours hierarchical algorithm with the Ward's criterion. It may graphically represent the resulting dendrogram and it can recommend the final number of classes using a heuristic based on Calinski-Harabaz criterion. This tool has been modified including the contributions developed in this work.

In this chapter, we will use the semantic clustering method in different datasets in order to evaluate its behaviour. In particular, we first present two preliminary studies: on one hand, we illustrate how the introduction of semantic similarity measures improves the clustering results with respect of considering the qualitative values as simple categorical; on the other hand, we present a comparative study with different ontology-based semantic similarity functions available in the literature (see section 2.1). Our hypothesis is that those semantic similarity measures that provide best results comparing pairs of terms when evaluated in a standard benchmark will also provide more accurate clusters when they are used to compute similarities inside a clustering method.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

After these preliminaries, we will show different tests made on a real data set of visitors of a National Park in Catalonia (Spain).

There are different ways to evaluate clusters in hierarchical clustering. Clustering, as any unsupervised technique must evaluate how well the results model the domain without reference of external information. However, because of its very nature, cluster evaluation or validation is a difficult task. To evaluate the quality of the results obtained with our method with respect to other clustering approaches, we have used a measure of comparison between partitions.

The chapter is organized as follows. Section 5.1 details how differences between cluster partitions have been quantified. Section 5.2, presents the results of the semantic clustering in a preliminary study using a reduced set of cities. In Section 5.3, some tests on real data have been presented. In particular, in Section 5.3.1, a dataset of visitors of the Ebre Delta Natural Park is introduced, and an analysis of the results of a previous study is done. In section 5.3.2, the same data set is tested using the presented clustering approach, and the results are studied. In section 5.3.3, the methodology for assessing semantic similarity from multiple ontologies is evaluated by using three different ontologies. The final section summarizes this chapter.

5.1 Distance between partitions

In the tests that we have performed, the best cut in the dendogram generated by the clustering method has been selected. The criterion for this selection is based on the Calinski-Harabasz index that optimizes the ratio between the inertia inter and intra clusters (Calinski and Harabasz 1974). This cut generates a partition of the individuals. This partition maximizes the Calinski-Harabaz index indicating the best proportion between the maximum homogeneity inside the classes (their elements are quite similar) and the maximum heterogeneity between the clusters, indicating that they are distinctive among them (the elements of each cluster are different from the elements in other clusters). This is, in fact, the goal of the clustering methods.

In some of the tests that we will explain, we have a reference partition to which compare the results of our method. To perform this comparison and obtain a numerical value for the distance between partitions, we have used the distance defined in (López de Mántaras 1991). It is worth to note that this distance takes two partitions of the same data set P_A and P_B.

Being P_A a partition whose clusters are denoted as A_i and P_B a partition whose clusters are denoted as B_j, the distance is defined as:

$$d_{Part}(P_A, P_B) = \frac{2 * I(P_A \cap P_B) - I(P_A) - I(P_B)}{I(P_A \cap P_B)} \quad (68)$$

where $I(P_A)$ is the average information of P_A which measures the randomness of the distribution of elements over the set of classes of the partition (similarly for and $I(P_B)$), and $I(P_A \cap P_B)$ is the mutual average information of the intersection of two partitions. They are computed as

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$I(P_A) = -\sum_{i=1}^{n} P_i \log_2 P_i \quad (69)$$

$$I(P_B) = -\sum_{i=1}^{m} P_j \log_2 2P_j \quad (70)$$

$$I(P_A \cap P_B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log_2 P_{ij} \quad (71)$$

where the probabilities of belonging to the clusters are $P_i=P(A_i)$, $P_j=P(B_j)$, and $P_{ij}=P(A_i \cap B_i)$ computed as observational frequencies.

Notice that the distance values obtained are normalized in the [0..1] interval, where 0 indicates identical clusters and 1 maximally different ones.

This distance has been already used to compare partitions by many authors, for instance (Singh and Provan 1996; Armengol and Plaza 2003; Valls 2003; Kumar, Ramakrishnan et al. 2008) among others.

This is useful for experimental essays or synthetic cases in which a reference partition exists and the theoretical properties of the method are evaluated. However, in real application, often a reference partition lacks, as getting it is exactly the target problem to be solved through the clustering. In these cases, the structural goodness of the classification together with the *interpretability* of the final classes is the most common criterion used to validate the quality of the results. The first criterion is guaranteed by the application of Ward's criterion for clustering and Calinski-Harabaz index for cutting the dendrogram.

5.2 Effect of semantics in clustering results

In this section two preliminary studies of the clustering method are presented. First, it is analyzed the improvement of introducing semantic features (whose values are compared using semantic similarity measures) in the clustering with respect to manage values as categorical. Then, the performance of ontology-based semantic similarity functions available in the literature is studied when applied to clustering.

5.2.1 Consequences of including semantic features in clustering

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

(country, language, geographical situation, major city interest and geographical interest).

Table 16. 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

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 version of Gibert's mixed metrics 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.

5.2.1.1 Clustering without considering semantic information

In this case, semantic features are treated as ordinary categorical features, that is, their semantics is not considered and the original Gibert's mixed metrics (which uses Chi-squared distance for categorical variables) is applied. 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 6 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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Figure 8 (up) shows the Class Panel Graph (Gibert, Nonell et al. 2005) for this partition. The following descriptions of the clusters are 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).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

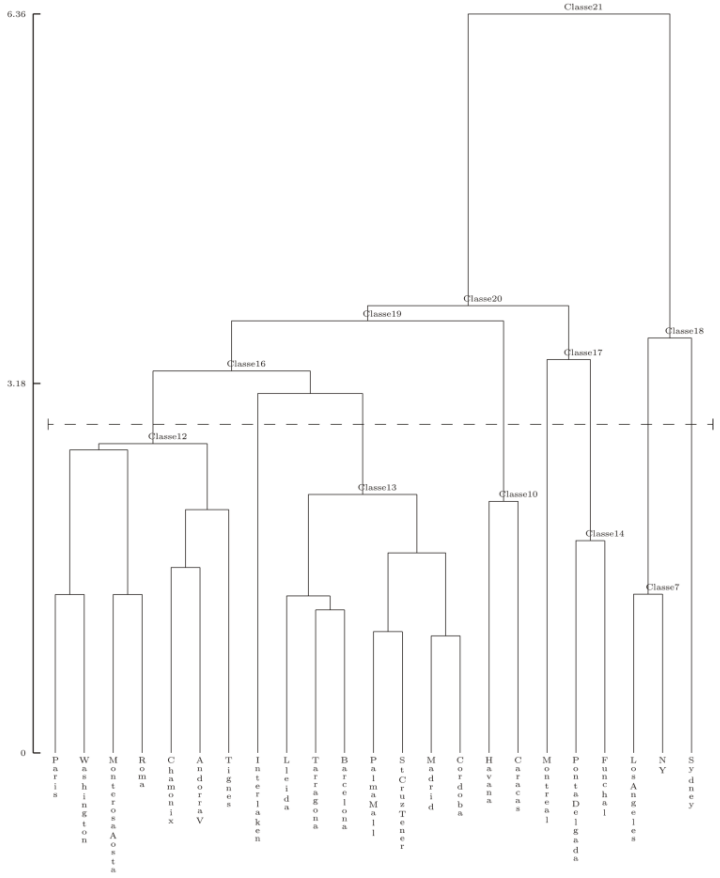


Figure 6. Dendogram without considering ontologies.

5.2.1.2 Clustering with semantic information

In this case, 4 variables were treated as semantic using the WordNet ontology in the similarity assessment: country, language, geographical situation and major interest. Continent and city ranking are treated as categorical. The proposal presented in section 3.2 (MSC_{log}) using SC_{log} measure (section 2.2.1) has been used to compare de values of semantic features. Figure 7 shows the resulting dendrogram, quite different from Figure 6 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, made from the class panel graph (see Figure 8 down), looks more coherent:

- Class10 has country capitals from Latin cultures (Cuba, Venezuela, Italy) speaking Romance languages with religious architecture as main interest.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- 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.

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.

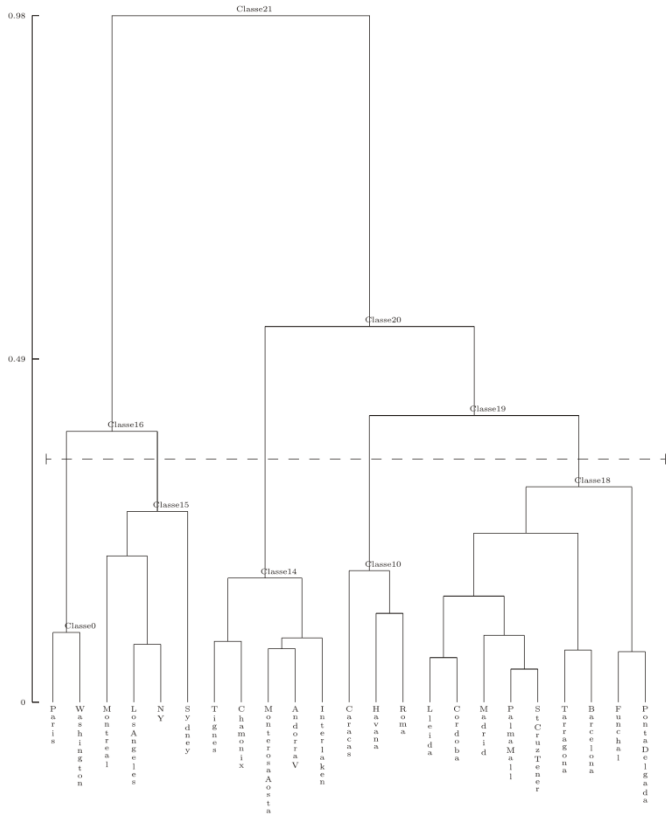


Figure 7. Dendrogram using ontologies.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

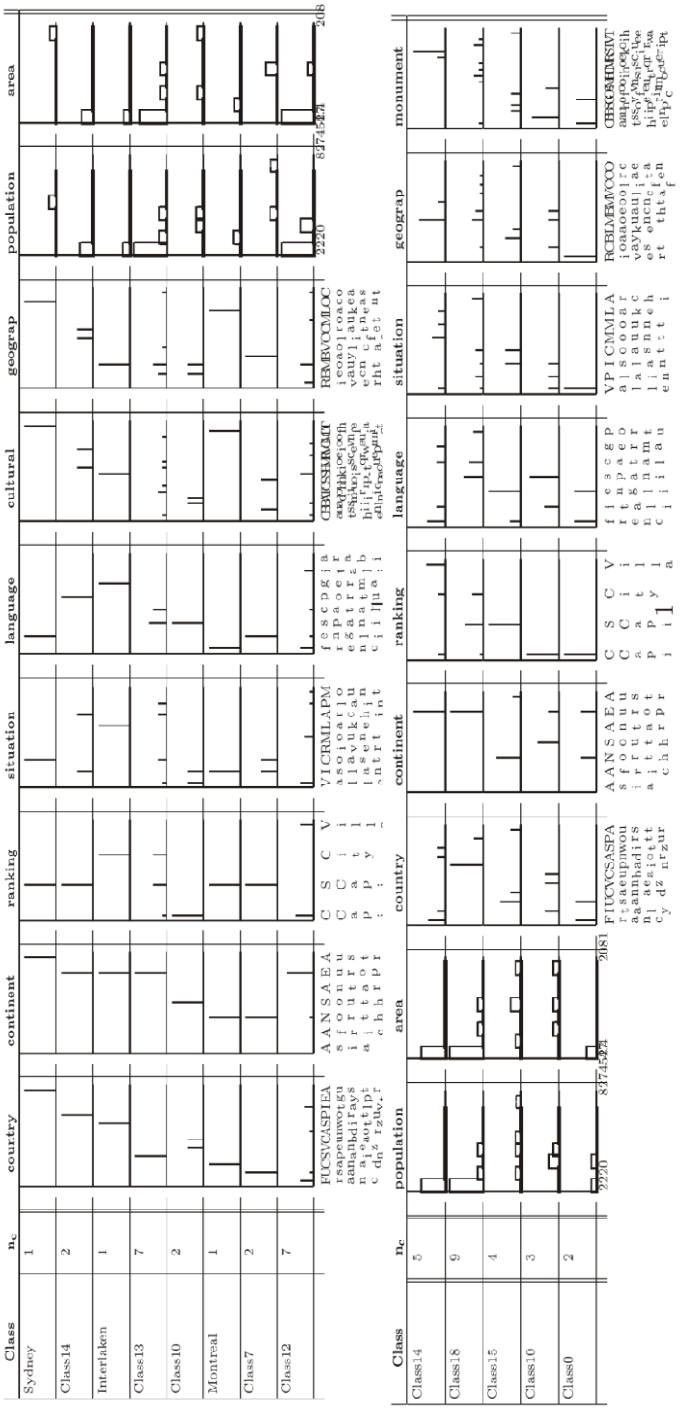


Figure 8. Class Panel Graph without semantic features (up) and with semantic features (down)

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

5.2.2 Performance of semantic similarity functions in clustering

In the previous experiment, we have used the proposal of semantic clustering presented in section 4.2 where the comparison between objects is performed by the generalized Gibert's mixed metrics. We presume that the fact of using our method for similarity assessment MSC_{log} (section 3.2) is a good alternative as it has obtained accurate results. However, the values of a semantic feature (i.e. terms) could be compared using other semantic measures.

In this section, the performance of different ontology-based semantic similarity measures when used in a clustering process is studied (Batet, Valls et al. 2010b).

Thus, the clustering of the touristic cities used in the previous section has been performed using three well-known semantic similarity functions (Section 2.1.1.1) and they will be compared with our proposal: Rada et. al. (Path Length), Wu and Palmer (WP), Leacock and Chodorow (LC) and our ontology-based measure SC_{log} (see section 2.2.1). For this experiment we will study the clustering results, using the WordNet ontology.

The clusters obtained with each of the different similarity measures are displayed in Table 17. In the case of our measure, in parenthesis, there are also the class labels of the same test as shown in dendogram of Figure 7.

From the point of view of interpretability of the clusters given in Table 17, Path Length provides some classes difficult to understand, putting together Chamonix and Santa Cruz the Tenerife despite they are not the same type of tourist destination at all. The experiment done using the WP similarity provides more interpretable classes. However, Tarragona and Lleida are included in a cluster with cities with ski resorts, although skiing is not possible at all in those cities. The experiment done using LC is able to detect correctly the class of skiing cities. However, ClassD in LC is quite heterogeneous. Finally, using our function the interpretation of clusters looks more coherent. Cities are grouped as country capitals; state capitals from North America or Australia placed in islands or near the coast where the spoken language is English; European cities placed at mountains where the main attraction is ski; country capitals from Latin cultures with religious architecture; and Spanish and Portuguese cities, not located in mountains, speaking romance languages, with tourist attractions (see the pervious section for the full description of these classes).

As explained in section 2.3, typically, the performance of semantic similarity measures is evaluated through a benchmark that consists of a list of word pairs ranked by a group of people. The agreement of the computational results and human judgments is assessed by looking at how well the computed ratings correlate with human ratings for the same data set.

The correlation between the results obtained with each of those similarity measure and the experts is given in Table 4. The ranking from the best to the worst is: our ontology-based measure (SC_{log}), Wu and Palmer and Leacock and Chodorow with the same correlation, and Rada et. al.

Then this dataset of cities was evaluated by a group of 4 subjects who partitioned them in different clusters of similar tourist destinations. The clusters obtained were compared against the human classification using the distance between partitions

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

presented in section 5.1. Since the human classifications were not identical, an average distance of each semantic similarity with respect to the 4 subjects is given.

Table 17. Results of ontology based semantic clustering using different semantic similarity functions.

Semantic Similarity	Partition	Cities in each cluster
Rada	ClassA	Paris, Montreal, Sydney, Los Angeles
	ClassB	NY, Washington
	ClassC	Caracas, Lleida,Tarragona, Córdoba, Palma de Mallorca, Habana, Roma
	ClassD	Interlaken, Tignes, Monterosa-Aosta, Andorra la Vella
	ClassE	Chamonix, Madrid, Santa Cruz de Tenerife
	ClassF	Funchal, Ponta Delgada, Barcelona
WP	ClassA	Paris, Washington
	ClassB	Montreal, Los Angeles, NY ,Sydney
	ClassC	Roma, Córdoba, Habana, Caracas
	ClassD	Santa Cruz de Tenerife, Madrid, Palma de Mallorca, Funchal, Ponta Delgada, Barcelona
	ClassE	Lleida, Tarragona, Andorra la Vella, Interlaken, Chamonix, Tignes, Monterosa-Aosta
LC	ClassA	Paris, Washington, Montreal, Los Angeles, NY, Sydney
	ClassB	Andorra la Vella, Interlaken, Chamonix, Tignes, Monterosa-Aosta
	ClassC	Tarragona, Barcelona, Funchal, Ponta Delgada
	ClassD	Roma,Habana, Palma de Mallorca, Caracas, Lleida, Córdoba. Madrid, Santa Cruz de Tenerife
SC	ClassA (Class0)	Paris, Washington
	ClassB (Class15)	Montreal, Los Angeles, NY, Sydney
	ClassC (Class14)	Andorra la Vella, Interlaken, Chamonix, Tignes, Monterosa-Aosta
	ClassD (Class10)	Habana, Caracas, Roma
	ClassE (Class18)	Lleida, Córdoba, Madrid, Palma de Mallorca, Santa Cruz de Tenerife, Madrid, Tarragona, Barcelona, Funchal, Ponta Delgada

Table 18 summarizes the distances between the human classifications and the partitions obtained applying the clustering method with different semantic similarity functions. Table 18 shows that the partition obtained using the Rada et. al. measure is the least accurate with respect to the human evaluations with a distance of 0.56. The WP and LC measures clearly outperform the poor results obtained with Rada et. al. measure. In fact, LC offer the best results of those measures based on the path length (0.44). The best results are obtained by our measure, with a distance of 0.32. In conclusion, the results obtained using those measures that have higher correlations using a standard benchmark also have the best performance in the clustering process.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 18. Distance between human partitions and computerized partitions.

Similarity method	Average d_{Part}
Path	0.560
WP	0.456
LC	0.441
SC	*0.32*

5.3 Evaluation with real data

In this section, we test our semantic clustering approach from tourism data obtained from a questionnaire done to visitors of the Ebre Delta Natural Park (in (Batet, Valls et al. 2010c; Batet, Gibert et al. 2011)).

Natural parks have increased their importance as a tourism destination in the recent decades. In fact, the recreational and tourist use of natural areas has a growing importance in relation to economic development (Epler Wood 2002). For that reason getting any kind of knowledge about the characteristics of the visitors is of great importance for planning, improving facilities and increasing the economic or biologic potential of an area.

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) (Figure 9), 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.

Figure 9. Ebre delta

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

5.3.1 Area of study and related work

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 questions are about demographic and socio-economical aspects of the visitor (e.g. origin, age, sex or level of studies), aspects of the trip organization (e.g. previous information, material), and characteristics of the visit (e.g. means of transport or activities done in the park) and, finally, the interests and satisfaction degrees on different features of the park.

From this set of variables, two groups of interest have been defined (Anton-Clavé, Nel-lo et al. 2007): 4 variables that define the tourist profile (origin, age group, accompanying persons and social class) and 6 that model the trip profile (previous planning, reasons for trip, accommodation, length of stay and loyalty). We performed a proper descriptive analysis for data cleaning. Table 19 shows the frequency values of features reporting the first and second reasons to come to Ebre Delta.

In (Anton-Clavé, Nel-lo et al. 2007), techniques of dimensionality reduction were used to find visitor's profiles, in order to improve the management of the area according to a better knowledge of the kind of people that visits the park and their main interests.

In particular, a multivariate homogeneity analysis was carried out. Two dimensions were selected for the analysis, keeping a 30% and 26% of variance respectively. In the interpretation phase, it was seen that Dimension 1 can discriminate among the variables relating to type of accommodation, length of stay and reason for the trip. It shows the degree of involvement of the tourist with the nature. The second dimension is determined by the type of group and by age and shows the degree of involvement with the services, such as accommodation. It is important to note that the reasons for visiting the park play a role in both dimensions, being, at the end, the major factor used to distinguish the two main big groups of tourists.

From that, five clusters (Table 20) of visitors were identified, from which the two first groups include a total of 83.9 % of the individuals. In (Anton-Clavé, Nel-lo et al. 2007) it was concluded that the rest of groups were really small and targeted to a very reduced group of visitors. For this reason, only the two main groups, corresponding to *EcoTourism* and *BeachTourism*, were characterised and discussed and a χ^2 (Chi-square) independence test was performed to show the significant difference between those two profiles regarding different variables.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 19. Frequencies of the reported values of features reporting the first and second reasons.

Linguistic Value	**First Reason**		**Second Reason**	
	Freq	**%**	**Freq**	**%**
Nature	339	17,4	211	11,4
Relaxation	146	7,5	222	11,4
Beach	125	6,4	45	2,3
Wildlife	61	3,1	88	4,5
Landscape	49	2,5	31	1,6
Culture	46	2,4	39	2,0
Second residence	45	2,3	9	0,5
Visit	40	2,1	20	1,0
Sightseeing	20	1,0	6	0,3
Holidays	19	1,0	6	0,3
Sports	13	0,7	19	1,0
Tranquillity	10	0,5	13	0,7
Others	10	0,5	6	0,3
Gastronomy	9	0,5	3	0,2
Loyalty	8	0,4	12	0,6
Business	6	0,3	1	0,1
Education	5	0,3	3	0,2
Familiar tourism	5	0,3	5	0,3
Walking	5	0,3	3	0,2
By chance	4	0,2	1	0,1
Fishing	3	0,2	2	0,1
Photography	2	0,1	1	0,1
Recommendation	2	0,1	3	0,2
Before disappearance	1	0,1	2	0,1
Bicycling	1	0,1	2	0,1
Clime	1	0,1	2	0,1
Ecotourism - Birds	0	0,0	2	0,1
Missing value	0	0,0	218	11,2
Total	975	100,0	975	100,0

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 20. Typology of visitors to the Ebre Delta Natural Park (Anton-Clavé, Nel-lo et al. 2007).

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

5.3.2 Using Ontology-based clustering to find tourist profiles in Ebre Delta

In this section, a n ontology based semantic clustering as proposed in section 4.2 is done on the same dataset of visitors to the Ebre Delta (Batet, Gibert et al. 2011).

In order to be able to compare our study with the previous one, we have taken into consideration the same subset of attributes, formed by 4 variables that define the tourist profile (origin, age group, accompanying persons and social class) and 6 that model the trip profile (previous planning, first reason to come, second reason to come, accommodation, length of stay and loyalty) (Table 21).

Since the previous study did not make use of intelligent data analysis, we first have performed the clustering using traditional treatment of categorical features. The classic Gibert's mixed metrics (Gibert and Cortés 1997) has been used as the compatibility measure for the clustering. In this experiment, age group, length of stage and loyalty are taken as numerical features, while origin, accompanying persons, social class, previous planning, accommodation, and reason 1 and 2 for the trip are taken as categorical. The dendrogram for this experiment with 8 classes is shown in Figure 10.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 21. Features values of Ebre Delta dataset.

Feature	Linguistic Value
accommodation	Hotel, cottage, apartment, camping, home, house
planning	Reservation, improvisation, excursion
origin	Catalan, Spanish, foreigner
accompanying	Family, Friends, partner, classmates, solitude, Seniors, Companies
social class	Low, low-middle, middle, high-middle, high
reason	Nature, Relaxation, Beach, Wildlife, Landscape, Culture, Second residence, Visit, Sightseeing, Holidays, Sports, Tranquillity, Others, Gastronomy, Loyalty, Business, Education, Familiar tourism, Walking, By chance, Fishing, Photography, Recommendation, Before disappearance, Bicycling, Clime, Ecotourism - Birds
age	1..100
loyalty	0…90
length of stay	0…3000

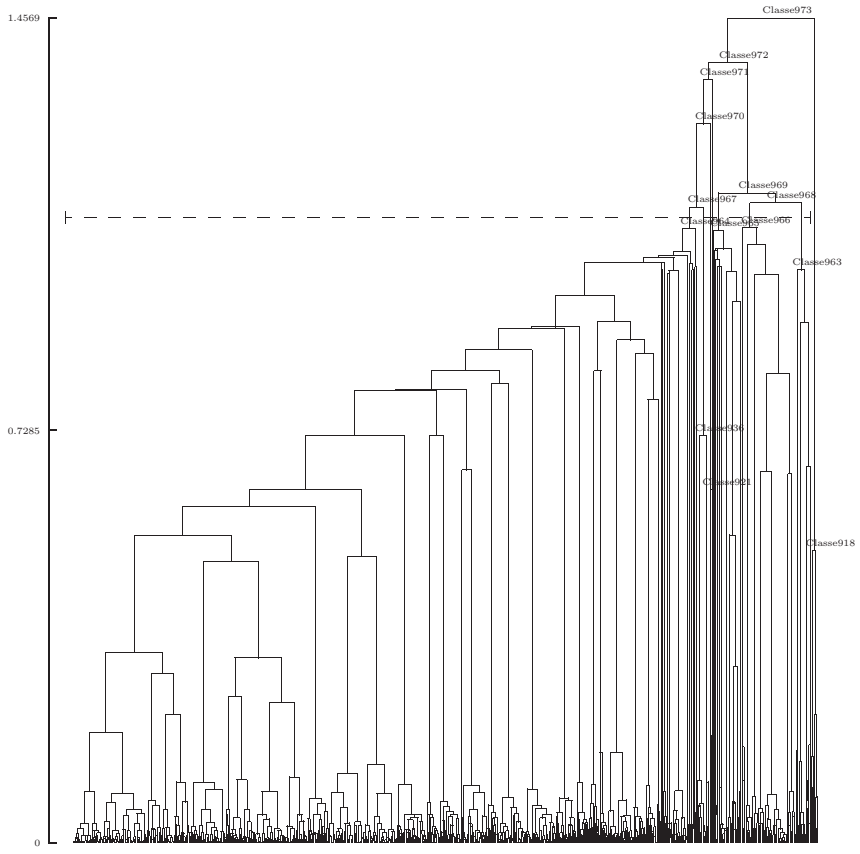


Figure 10. Dendogram with categorical features (8 classes).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

In this case, as usual in real complex domains, there is a very big class quite heterogeneous which seems to share all type of visitors (Table 22).

Table 22. Typology of visitors to the Ebre Delta Natural Park with categorical features.

Class	#	Description
864	1	Single outlier visitor
C963	20	Long stage, between 35-early 40s years, 52% stays at second home, 75% are Catalan people, it is not clear the first reason to come (some of them come for walking).
C966	72	Long stage, between 35-early 40s, 68% is at home, half Spanish, half Catalan, have a second residence near the park
C965	37	Long stage, higher fidelity, around 46 years, 65% home, 78% Catalan, their main interest is gastronomy
C921	4	Long stage, more fidelity, between 35-early 40s, 50% goes to the hotel, an important part makes reservation, part of the foreigners concentrated in this group 25% of the class are foreigners, they come for recommendation of other people, 50% Catalan, main interests: relaxation or landscape
C936	16	Shorter stage, between 35-early 40s, almost 60% home, 80% Catalan, main interests: nature or business
C918	8	Young, under 30s, 50% stay in camping, 75% makes reservation, 50% Spanish, education tends to be first reason
C964	817	Shorter stage, occasional visit, between 35-early 40ss, mainly hotel, 63% Catalan, main reasons: nature, landscape and sightseeing

Neither the second reason nor the social class discriminates at all among the classes. The main problem here is that *class 964* concentrates 817 visitors, which represents the 83.8% of the total sample size and is quite heterogeneous. Thus, although some differences can be seen in other profiles, these are just referring marginal groups of the population and the information provided by this clustering is not very useful. This is a common effect of clustering when many variables are used. Our hypothesis is that this effect could be minimized by introducing the semantics of the terms in the clustering process.

Thus, the ontology based semantic clustering has been applied by considering all textual variables (origin, accompanying persons, social class, previous planning, accommodation, first reason to come, and second reason) as semantic variables and using the metrics proposed before. WordNet ontology is used again to estimate the semantic similarity.

From the results of this experiment, a cut in 8 classes is recommended for its interpretability. The dendogram for this experiment is shown in Figure 11.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

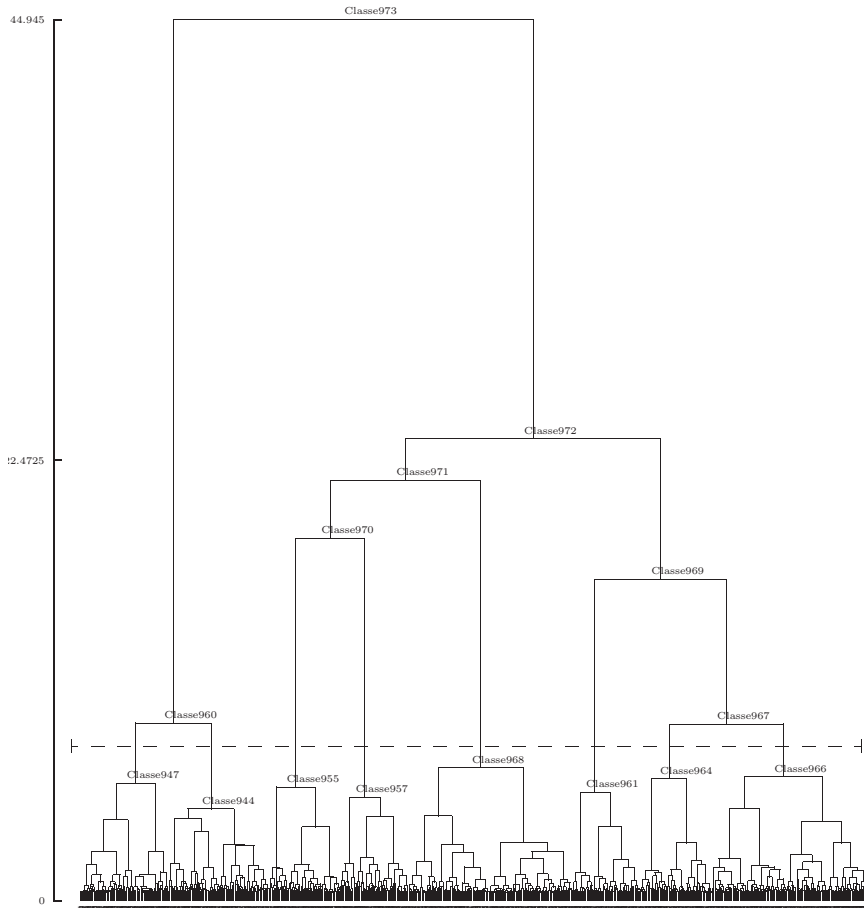


Figure 11. Dendogram with semantic features (8 classes).

This time, we obtain more equilibrated classes (Table 23) than in the previous experiment classes are more equilibrated that in the previous experiment (Table 22).

The interpretation has been made from their class panel graph (Figure 12).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Table 23. Typology of visitors using semantics.

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

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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

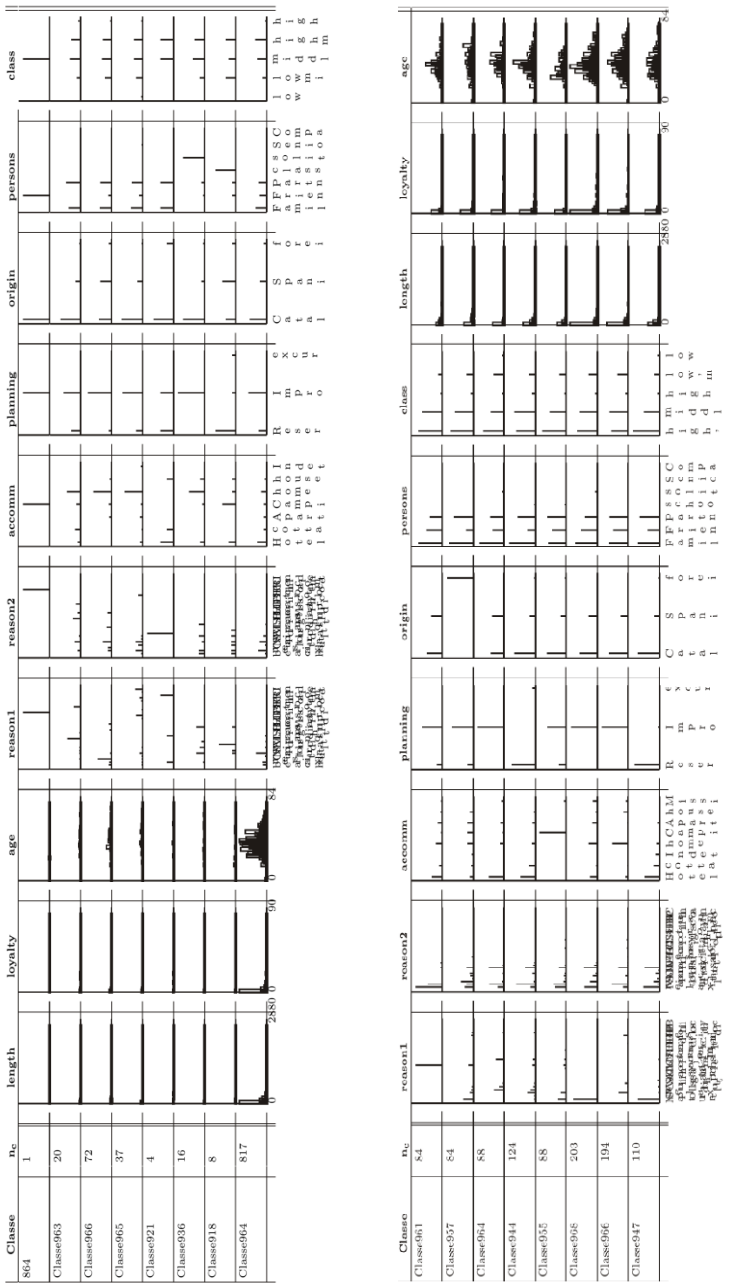


Figure 12. Class Panel Graph with categorical features(up) and with semantic features (down).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

The data analysis made on the visitors to the Ebre Delta National Park in (Anton-Clavé, Nel-lo et al. 2007) was a pure statistical multivariate approach, consisting in projecting the data in a new artificial space of factorial components which preserves as much information of the complete data set as possible. The success in the results when using these techniques depends, in general, on the experience of the data analyst, who must be able to find a proper interpretation of the selected dimensions. In this case, an interpretation for the first two dimensions was found and used to define the profiles.

Moreover, for this particular application, the data analysis provided a rather non-equilibrated partition with two very big groups and some very small ones. Experts focused on these two big groups, disregarding some interesting information contained in the other ones. However, the total variance represented by the two first dimensions is 56%, which means that 44% of the information contained in the data set is missed. This is a rather good result when qualitative variables are used with this technique in real applications. However, the reduction of data is so important that can seriously affect the correspondence with reality.

Since most of the variables were categorical, the standard techniques used worked as usual under a pure syntactic approach, where simple binary comparisons between modalities were performed, only distinguishing equal or different responses to each question of the survey, leading to a very poor estimation of the real similarity between different responses.

Data mining techniques, on the contrary, permit to make a qualitative analysis of the structural relations between the concepts expressed by data and the analysis is performed directly on the original variables space, guaranteeing the direct interpretability of the results. In particular, clustering algorithms are adequate to study the relationships among objects and to define a grouping with high intra-cluster homogeneity and high inter-cluster distinguishability. However, when they are applied to categorical variables, the same restrictions mentioned before appear, making difficult the establishment of differences between the objects.

In the cluster performed disregarding the semantic information, there are already significant improvements regarding classical data mining approaches, since here, the Gibert's mixed metrics is already enabling to take into account the quantity of information provided by qualitative variables. But, still, the qualitative variables are managed from the syntactic point of view and the ordering relationships between semantic concepts is not taken into account.

When semantics is disregarded, the clustering generates a big class (with 83.8% of the tourists) and other 7 small classes. The goal of the study is to determine a typology of tourists with certain specificities that permit to the managers of the park to make a better planning of the recreational uses. Therefore, although the interpretation of the partition in few classes is possible (see Table 22), from the point of view of the manager, this partition is useless because the majority of visitors belongs to the same profile and no clear between-class differences can be identified. Notice that, in the dendogram given in Figure 10 clustering has been performed in steps of the same importance, that is, there is not a big difference in inter and intra cluster distances.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

In the ontology-based approach proposed in this work, the clustering method is able to manage the meaning of values, relating them to concepts in a given ontology.

The partition obtained with this semantic-based clustering generates 8 clusters of more homogeneous dimension. This is an important fact, since now we can identify typologies of visitors that represent a significant proportion of the total number of visitors.

From the dendogram in Figure 11, it can also be seen that we have obtained clusters with high cohesion, which means that the distances between the members in the cluster are quite small in comparison with their distances with objects outside the cluster. Moreover, if the level of partition is increased, then the cohesion of the clusters decreases quickly, which also indicates that the clusters are well defined.

This clustering is coherent with the grouping made by (Anton-Clavé, Nel-lo et al. 2007) using multivariate analysis, because the variables about the reasons for visiting the park have a great influence in the formation of the groups. Interests on nature, beach and relax are present in different classes. However, thanks to the semantic interpretation of the concrete textual values provided by the respondents, we have been able to identify that visitors interested in nature are similar to those interested in wildlife. The system has been also able to identify the similarity between hotels and cottages and between second homes and familiar houses. This proofs that the estimation of the relative similarities among objects in terms of the meaning of the values improves the final grouping.

In this way, the two types of visitors identified in statistical analysis as Ecotourism and Beach Tourism (mainly guided by the variable *First Reason to visit the park*) have now been refined as follows:

- Ecotourism. Here different subtypes can be disguised:
  - visitors that stay in hotels and apartments for relax (C947),
  - visitors with familiars or a second residence (C966),
  - Catalan and Spanish visitors interested in wildlife (C968) and
  - tourists interested in culture (C955).
- Beach tourism. Here different subtypes can be disguised:
  - older people staying in hotels or apartments looking for relax (C944) and
  - people that live near the park (C961) or
  - second home (C957) and go to the beach quite frequently.

Notice that this is a more rich classification that establishes clear profiles of visitors with different needs. This is according with the hypothesis of the experts that suggested that the park attracts highly different types of visitors. After this study, the manager may study different actions according to the different types of demand.

5.3.3 Evaluation using multiple ontologies

In this section we present an analysis focused on the method for multiple ontologies (section 3.2) (Batet, Valls et al. 2010c), during the semantic clustering.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

We have centred the evaluation on testing the influence of different ontologies in semantic clustering results when applied to real data. In this case the same dataset of a questionnaire made to 975 visitors to Ebre Delta Natural Park is used. The questionnaire was designed in order to determine the main characteristics of the tourism demand and the recreational uses of this natural area.

Three ontologies have been selected to make this analysis. This selection has been done to study the influence of different types of ontologies in the results of the clustering:

1. WordNet (Fellbaum 1998): that is a very large and general purpose ontology, so it covers all the values that appear in the different variables (i.e. nature, relax, camping, etc.) (See section 2.3.2.1 for more information) .
2. Reasons: that is a domain ontology developed by an expert specifically for the purpose of analysing this particular data set. It includes a biased modelling of the terms regarding to reasons for visiting a natural park. Concepts are structured as environmental issues (e.g. wildlife, nature), hobbies (e.g. sport, gastronomy) and personal interests (e.g. visit family, culture) (Figure 13).
3. Space: a domain ontology about spatial information retrieved from the Swoogle ontology search engine. It covers aspects regarding to geographic data (e.g. mountain, valley) and types of buildings (e.g. governmental buildings, schools). In particular it includes information about different types of accommodations (e.g. camping, hotel).

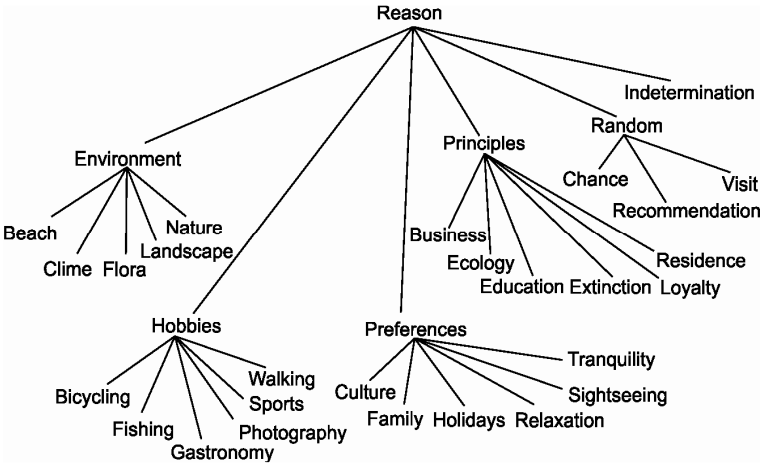


Figure 13. Reasons Ontology

In order to study the impact of adding new ontologies in clustering process, we made a first test with a single semantic feature. This variable refers to the visitors' reason to visit the Ebre Delta Park. WordNet and Reasons ontology are considered. So, the experiments are performed over two different cases: (1a) using only WordNet, (1b) and using both WordNet and the Reasons.owl ontology. From the results of this experiment, a cut in 4 classes is recommended for its interpretability. The dendograms for this experiment are shown in Figure 14 and Figure 15.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- (1a) The partition obtained with only WordNet is quite difficult to interpret. It can be seen that the results obtained generate a disassociated class, which corresponds to visiting the park for nature reasons C337, and other classes with the rest of reasons. One can distinguish a class C968 where the reason is going to beach, but the rest of classes (C970 and C925) are heterogeneous.

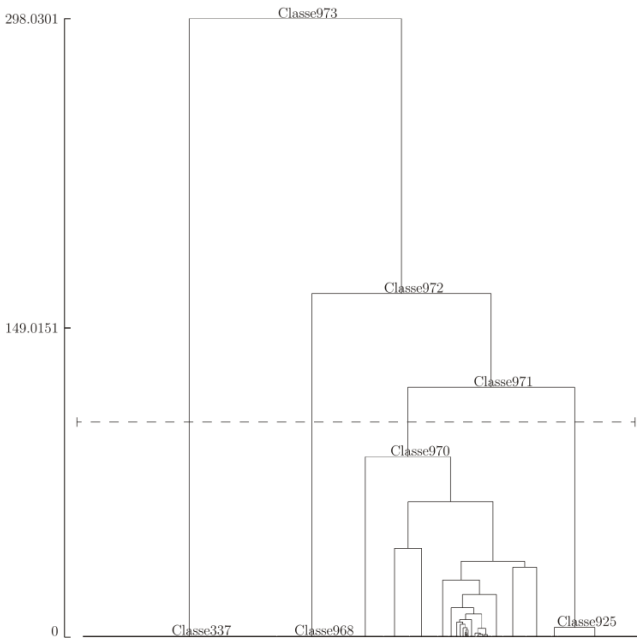


Figure 14. Dendogram of one semantic feature using WordNet

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

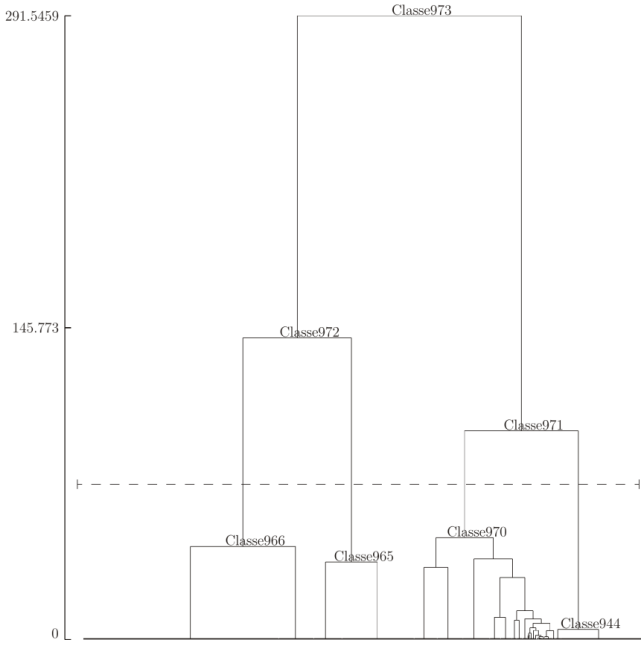


Figure 15. Dendogram of one semantic feature using both WordNet+Reasons ontology.

- (1b) In the partition obtained using the Reasons ontology in addition to WordNet the classes are more equilibrated that in the previous experiment, having clusters with more homogeneous dimension. C966 contain people mainly interested in nature (84%). In class 970 people are interested in culture. sightseeing and visit the family (43%). In class C944 the reason for visiting the delta is relaxation. And finally, C965 contains people interested in the beach (72%) and also in the landscape (28%).

In both cases, the partition has been done optimizing the Calinski-Harabasz as usual (Calinski and Harabasz 1974). We can see that the partition obtained in (1b) has been done at a lower level than the one in (1a), which means that the cohesion of the clusters in the latter is much higher than in the former.

A bivariate cross table of the partitions obtained between these two experiments is given in Table 24. We can see that objects have been distributed in a different way:

- In the partition including the Reasons ontology, C966 has 61 objects that belong to C970 in the other partition, these correspond to individuals whose main reason was "*wildlife*". In the domain ontology developed by the expert, "*wildlife*" is associated to "*nature*", whereas in WordNet "*wildlife*" can be associated also to "beach" and other concepts.
- The same happens with class C965, which is related with personal preferences about rest, relaxation and tranquillity. The term "*tranquillity*" has a more general meaning in WordNet, so the people that answered with this term were grouped in a more heterogeneous class.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- Class C970 is related to personal preferences of the tourist while the class with the same name using only WordNet is more heterogeneous.
- Finally, two classes are identical in both partitions (C944) for the one including Reasons ontology and C925for the one using only WordNet). Old people staying in hotels and apartments looking for relax.

Table 24. Bivariate table comparing the partition obtained using only WordNet and using both Wordnet+Reasons

WN \ WN+Reasons	C966	C970	C944	C965
C337	339			
C970	61	245		49
C925			156	
C968				125

Now a second experiment is presented. In the second experiment we evaluate the influence of considering multiple ontologies when combining numerical and semantic features. The study includes two semantic features referring to the *main reason* to visit the park and the *type of accommodation* (i.e. hotel, camping, etc.), and a numerical feature *age*. The three ontologies explained before have been used with three different settings: first we used the general purpose ontology, WordNet (2a), then, we add the Reasons ontology (2b), and finally we add the Space domain ontology that was not designed specifically for this dataset (2c). The results of the tests are the following:

- (2a) Using only WordNet: the best partition is in 5 classes. With 5 classes we can find:
  - In class C967 that contains visitors whose main reason to come to de park is nature (96%), and most of them are in a hotel (32%) but not in a camping.
  - In class C969, people come for relax, visit the family or because they have a second residence (65%). However this class includes other reasons. Visitors stay at home, at a relative's house, apartment or cottage.
  - In class C962 the reasons are nature (37%), relax (20%) or beach (17%), and visitors stay in a camping (90%).
  - Class C968 has visitors quite heterogeneous, interested in wildlife, relax, and landscape among other, some of them stay at a hotel (64%).
  - Finally, in class C965 the main reason is going to the beach, a 33% stay in a hotel.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

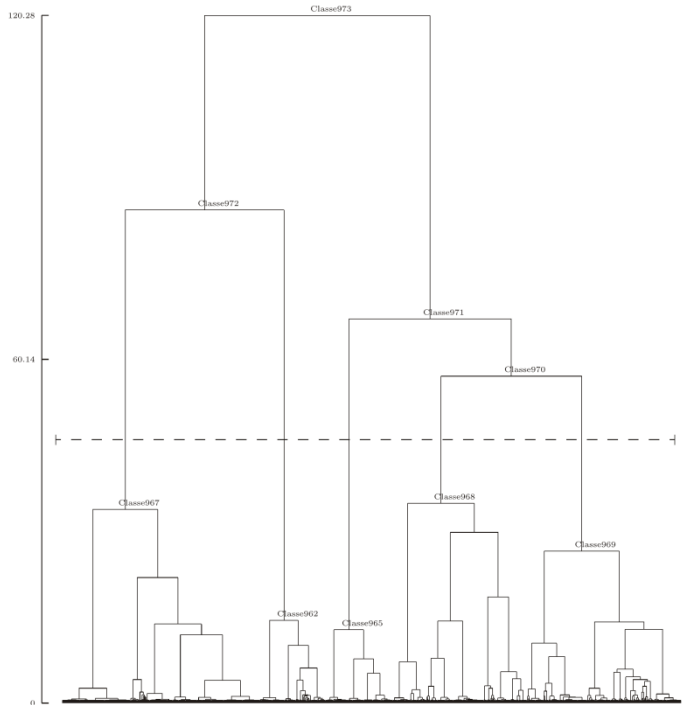


Figure 16. Dendogram generated using WordNet ontology

- (2b) Using WordNet and the Reasons ontology: the recommended partition is in 5 classes. Here, the interpretation of the clusters with regards to the reasons for visiting this kind of park is done in a more clear way.
  - Class C966 where visitors come from environmental related reasons (nature (82%) and wildlife (15%)) and stays at a hotel (34%).
  - Two classes of beach tourism:
    - one group C961 come for nature (34%) relaxation (18%) and beach (15%) but stay at home,
    - and another group C959 of visitors come for beach (72%) and landscape (27%) but stay in a camping.
  - Finally, two other classes are related to personal preferences (relaxation, visit family, culture or second residence):
    - the former C965 contains visitors who stays in a hotel,
    - in the latter C969 users stay in an apartment or cottage.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

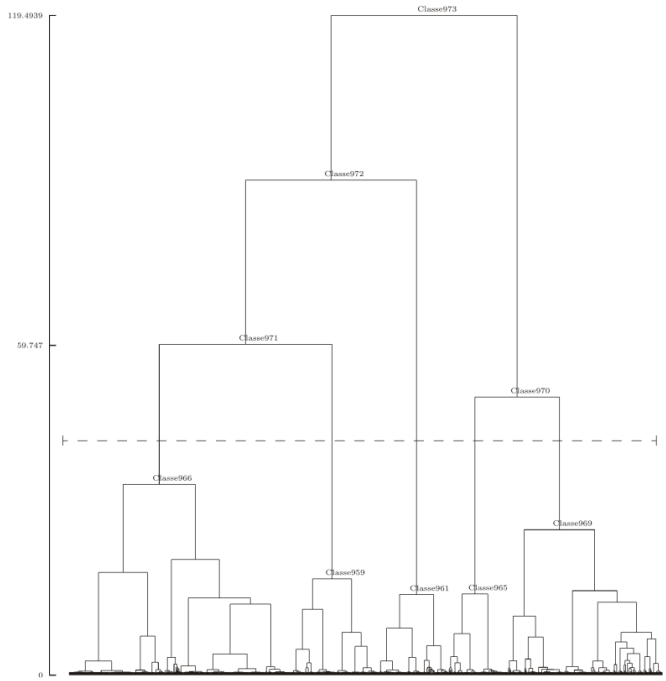


Figure 17. Dendogram generated using WordNet, and Reasons ontologies

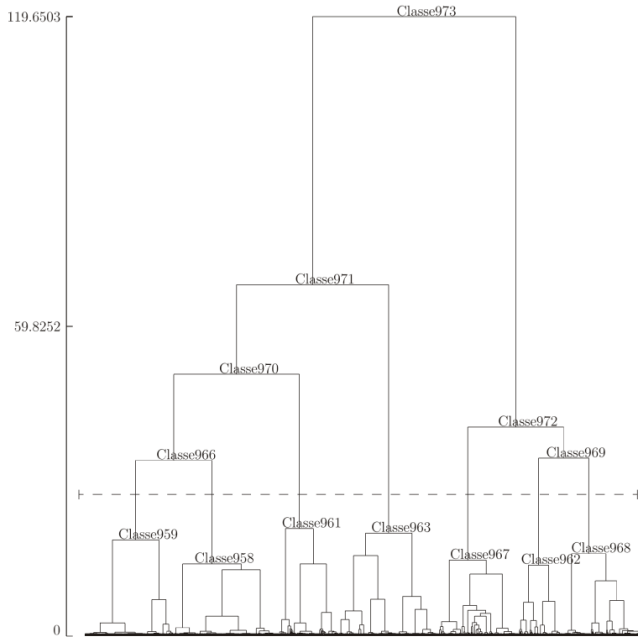


Figure 18. Dendogram generated using WordNet, Reasons and Space ontologies

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- (2c) Using WordNet, Reasons and Space ontologies: the best partition here is in 7 classes, which permits to make a finest classification of the visitors. As in the previous experiment, we find two classes C959 and C958 where visitors come for environmental reasons (in the former the reasons are nature and wildlife and stay at a hotel (however also includes people interested in bicycling, the clime etc.), and in the latter the reason is nature (98%) and stay in a cottage, apartment or at home). Another group, C961, with users interested both in nature (50%) and relaxation (23%), staying in a camping. Class C968 represent people who come only for relax. C962 represent people who come for culture, sightseeing and photography and they mainly stay in a hotel. Class C963 has visitors interested in the beach. Finally, C967 contains people who stay at home or in a relative's house, and interested in diverse things, such as visit family, gastronomy or taking photos.

In order to perform a more analytical comparison between the results given by the three tests, we have done a cut in 7 clusters in the three dendograms. From the point of view of the manager of the park, the third classification (2c) in 7 groups is better than the others, because it is able to find more profiles of visitors, so that he can initiate different specialized actions for each particular target group.

To evaluate the difference between the partitions, the results have been compared using the Mantaras' distance between partitions, defined in (López de Mántaras 1991). The distance between partition of test (2a) and 2(b) is 0.3, between test (2b) and (2c) is 0.27, and between (2a) and (2c) is 0.43 (see Table 25). This shows that effectively the inclusion of more ontologies as background knowledge produces different clustering results. So, the beneficial effect of having more ontologies proved for the behaviour of the methodology for assessing similarity proposed in section 3.2 is also propagated to the benefits on getting more accurate profiles when clustering algorithms introduce this measure. The proposed methodology seems to able to consider the knowledge provided by each individual ontology in an integrated way.

Table 25. Distance between partitions

Partition	**2a**	**2b**	**2c**
2a	0	0.3	0.43
2b		0	0.27
2c			0

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

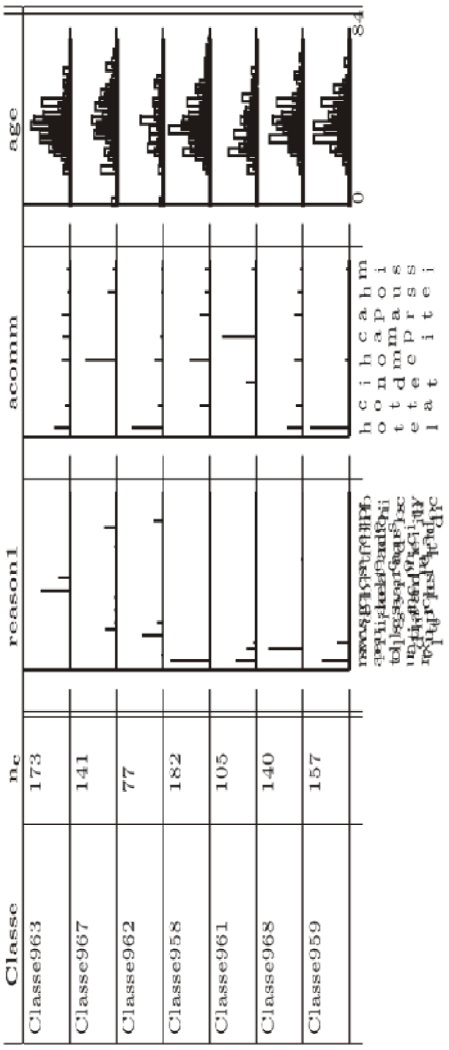


Figure 19. Class Panel Graph using WordNet, Reasons and Space ontologies

5.4 Summary

In this chapter we have evaluated our approach of semantic clustering.

In a preliminary study, we have analysed the influence of including semantic information in clustering and the influence of using different semantic similarity

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

measures in the results of clustering. Considering 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.

These semantic similarity measures can be integrated into the compatibility measure presented in section 4.2.1 that is capable of combining the contribution of a set of numerical, categorical and semantic features to compute the distance between individuals in a clustering process. 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. The evaluation has shown 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.

Then, we test our approach with real data using a dataset obtained from a survey done to the visitors of a Natural Protected Park. 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 which enriches the previous results and provides useful knowledge about the characteristics of the visitors.

In addition, the case in which multiple ontologies have been evaluated over the same data set is also considered. This last aspect is especially interesting as many overlapping ontologies are typically available for a domain, each one modeling the same knowledge in a different way. By heuristically combining the knowledge modelled in each one, we are able to exploit the benefits of a more accurate knowledge representation that an ontology may provide for a given pair of concepts.

Results obtained with real data and several general purpose and domain ontologies have sustained this hypothesis. The combination of general purpose ontologies with domain ontologies seems to generate more interpretable clusters, because each ontology provides part of the knowledge required in the data mining process. We have also seen that including ontologies especially designed for analyzing a particular dataset may introduce a user's desired bias on the clustering.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Chapter VI

6 Applications

Semantic clustering has many interesting applications. The support for textual data opens the possibility to deal with many data mining tasks in which a semantic interpretation is needed.

In this chapter, we present two applications of the semantic clustering method proposed in this thesis in two very different fields.

The chapter is organized as follows:

Section 6.1 presents the application of semantic clustering within the scope of the Spanish project DAMASK.

Section 6.2 presents the application semantic clustering as a measure of validation of data utility in the context of privacy protection of textual attributes.

The last section summarizes the chapter and presents some conclusions.

6.1 DAMASK project

The work presented in this document opened one of the research lines proposed in the Spanish research project *Data-Mining Algorithms with Semantic Knowledge* (DAMASK TIN2009-11005)[9]. This project proposes the use of semantic domain knowledge, represented in the form of ontologies, to define new methods for extracting and integrating information from heterogeneous Web resources with varying degrees of structure, performing an automatic classification and description of the clusters identified from a semantic point of view. Finally, a recommender system based on these technologies will be developed.

The project will test the practical applicability of the proposed methods in the strategic area of Tourism. As it is argued in the project, different studies indicate that the Web is a great mechanism for tourist destination promotion at all levels (including local one). Consumers all around the world use Internet in order to obtain information about travelling, and 54% of them use Internet to initiate the search for travel agencies in the Web. Thus, information about destinations is of crucial importance in decision making in the Tourism field. In consequence, the difficulty of accessing those data

[9] http://deim.urv.cat/~itaka/CMS2/

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

may introduce a bias between different destinations. In this sense, one of the most important topics in e-Tourism is to avoid the saturation of common destinations, by means of the promotion of new ones, achieving a sustainable tourism. In order to achieve this goal, it is very important that the user receives all the possible tourist offers that match with his preferences. As information is usually available in a textual form, proper extraction, interpretation and classification as well as personal recommendation tools are required. So, this is the focus of the DAMASK project.

The main goal of this project is to develop new data mining methods driven by the semantic domain knowledge. In order to accomplish this goal a system able to group semantically related Web contents will be developed. The main goals of the project are the following:

1) Pre-processing of input data, focusing on their acquisition from freely and massively available resources such as Web resources, their integration and their transformation in a format which may be directly processed.
2) Methods of automatic classification of data, considering any type of heterogeneous information, including numerical, categorical and conceptual data. The work presented in this thesis fits in this goal.
3) Methods for interpreting the classes obtained in the previous step and give the proper recommendations to the user.

In order to achieve these goals, the project has been divided into 3 main tasks:

- *Task 1*: *Ontology-based information extraction and integration from heterogeneous Web resources.* The output of this task are the data values the studied domain, including numerical, categorical and textual data, represented in an object × attribute matrix (data matrix) which will be the input of the task 2. This task relies on several scalable knowledge acquisition techniques like linguistics patterns or statistical analyses (Sánchez and Moreno 2008a; Sánchez and Moreno 2008b) in order to 1) extract relevant information describing an entity (e.g. a touristic destination), 2) associate these data to ontological concepts. More details can be found in (Sánchez, Isern et al. 2010).
- *Task 2*: *Automatic clustering of entities based on the semantics of the concepts and attributes obtained from the Web resources.* This task starts when the different data types to be considered in this project have been acquired (numerical, categorical, conceptual, textual, etc.). Based on the data matrix and an ontology, it will be designed a method for automatically building clusters with the help of the contextual semantic knowledge provided by the domain ontology. Moreover, an automatic interpretation process of the clusters will also be studied, in order to obtain a semantic description of the clusters that can help the user in his/her decision making tasks.
- *Task 3*: *Application of the developed methods to a Tourism test case.* In this task the goal is to evaluate the deployment of the methods designed in the previous tasks in a particular case study: a personalized recommendation system of touristic destinations. A Web application will be designed to offer this kind of recommendations to any user. The tool will be focused on searching touristic destinations in the different types of touristic resources available in Internet using the tools developed in task 1. The clustering methods defined in task 2 will then be

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

applied to obtain a classification of touristic destinations based on the domain knowledge and the user preferences, in order to be able to recommend the set of places that match better with the user's interests.

The project started in January 2010. Tasks 1 and 2 are being developed in parallel, corresponding the second one to the work presented in this document.

In broad lines, task 2 is subdivided in different subtasks:

- State of the art about the techniques for automatic clustering of data and about the existing methods for similarity measurement for semantic concepts. This study will focus on the drawbacks of the traditional clustering methods, which do not use contextual semantic knowledge to guide the process of classification. It will be studied which similarity measures are being used for comparing a pair of concepts from a domain ontology or lists of words using semantic structures like WordNet. Finally, the applicability of those semantic similarity measures into a clustering algorithm will be analyzed.
- Identification of the clustering steps in which contextual domain knowledge could be included to improve the results.
- Adaptation of the traditional clustering algorithms to permit the use of some of the existing semantic similarity measures based on ontologies and linguistic terms. Analysis of the improvement in the quality of the results, with respect to the interpretability of the clusters in a particular domain. Identification of the drawbacks or weak points of those measures. Use this information to design and implement 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. Afterwards, those new methods must be included into the clustering algorithms, obtaining a software prototype that will be tested in different case studies. This will permit to evaluate the results of the new domain-centred approach in relation to the previous ones.

The work presented in this thesis corresponds to these tasks. It has extensively analyzed the similarity paradigm and studied their behaviour both from theoretical and practical points of view. As a result, new approaches for semantic similarity estimation have been proposed which have proved to improve related works, including the exploitation of several knowledge sources, while fitting into the requirements of the project. Moreover, a semantic clustering method has been proposed being able to integrate in a uniform manner all the data compiled in the first stage (numerical, categorical or semantic). The evaluation showed that a classification process that takes into consideration the semantics of input data is able to provide better tailored classes than a method focused on numerical or categorical data.

Due to task 1 is still under development, the semantic clustering methodology has not been integrated with the extraction tools yet. However, some evaluations have been performed with pre-compiled domain data (i.e., information about touristic destinations). As shown in section 5.2 and 5.3, both synthetic and real data (from the Delta de l'Ebre Natural Park) have been used to test the accuracy of our method. The results of these tests show the validity of our method and that our semantic clustering approach is easily applicable to this project and recommender or expert systems dealing with textual data.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

6.2 Anonymization of textual data

In this section we analyze the connections between data mining tools and privacy preserving techniques. Privacy Preserving Data Mining is an emerging research field that has appeared with the increase of the ability to store data from individuals(Domingo-Ferrer 2008). After a study of the field, we have identified that some of the outputs of this thesis can be applied in the context of privacy preserving of textual data. In fact, some initial contributions have already been made.

Statistical agencies generally provide summarised data generated from a collection of responses given by a set of individuals. Therefore, because responses are not directly published, an individual's privacy can be easily guaranteed. However, this summarized information may be not useful enough if a detailed analysis of the responses is needed. Because of the potential benefits of exploiting the original data (i.e. micro-data), new masking techniques are being developed to minimise the risk of re-identification when this information is made available.

To satisfy a certain anonymity degree, several masking methods have been designed to modify the original values, building a new masked version of the dataset (Domingo-Ferrer and Torra 2004). This data transformation results in a loss of information, which is a function of the differences between the original and masked datasets. These differences may compromise the utility of the anonymized data from the data mining point of view. Ideally, the masking or anonymization method should minimize the information loss and maximize data utility.

In the past, many masking methods were designed to build groups of continuous-scale numerical data. Numbers are easy to manage and compare, so the quality of the resulting dataset from the utility point of view can be optimised by retaining a set of statistical characteristics. However, extending these methods to categorical attributes is not straightforward because of the limitations on defining appropriate aggregation operators for textual values, which have a restricted set of possible operations. Moreover, textual attributes may have a large and rich set of modalities if the individuals are allowed to give responses in textual form. Because of the characteristics of this kind of values and the ambiguity of human language, defining suitable masking techniques is even more difficult. Semantics play a crucial role in properly interpreting these data but this dimension is often ignored in the literature. The quality of masked data is typically considered by preserving the distribution of input data. Although data distribution is a dimension of data utility, we agree with other authors (Xu, Wang et al. 2006) that retaining the semantics of the dataset is more important if the aim is to draw substantive conclusions from data analyses.

In (Martínez, Sánchez et al. 2010a), a new method for anonymization of textual attributes using ontologies is proposed. In this method, semantic technologies based on ontologies have been exploited to develop an algorithm that is able to interpret textual data, associating words to ontological concepts and perform a masking process that considers the semantics of data. Compared with previous approaches that consider semantic features in a categorical fashion, the incorporation of the semantics into the algorithm has shown to preserve the amount of information in the dataset,

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

while guaranteeing a certain degree of anonymization (Martínez, Sánchez et al. 2010a).

Most of the premises about textual data processing and the role of ontologies and semantic similarity are shared to those presented in this work, showing that, as stated in the introduction, semantic technologies have a predominant role in many areas of application.

Two direct synergies have been found between both works. On one hand, the semantic similarity measures developed in this thesis are being applied to assist the process of anonymization of textual data. The method is based on the substitution of sensible values for other ones that increase the level of anonymity. This implies that the values of a record that could be de-identified are substituted for new ones, which are semantically "near" to them. The algorithm assumes that the most appropriate record to substitute a non-anonymous one is the record that minimizes the semantic distance with respect to the original. Due to the fact that anonymization methods typically deal with huge amounts of data, we believe that the measures presented in chapter 2, will aid to improve the process while maintaining the necessary scalability and efficiency of the algorithms.

On the other hand, the clustering method developed in this work has been directly applied to the evaluation of the anonymization results. A critical aspect of any anonymization method is the preservation of the data utility. As data anonymization necessarily implies a transformation of data which is in function on the desired degree of privacy, one should carefully consider that this transformation preserves, as much as possible, the utility of the original data. The anonymization method presented in (Martínez, Sánchez et al. 2010a) has paid especial care in the preservation of data utility, which, in the case of textual information necessarily relies in the preservation of data semantics.

To evaluate the data utility, traditional approaches uses measures focused on maintaining the distribution of the values, without considering any semantic information. In (Martínez, Sánchez et al. 2010b) we propose to evaluate the accuracy of semantic anonymization methods applying a data mining method and comparing if the results obtained with the masked dataset are similar enough to the ones that would be obtained with the original dataset. For this purpose, semantic data mining methods are required. The use of the clustering method presented in Section 4.2 is crucial for achieving this task. By quantifying the differences between the clusters obtained from original data against those obtained from the anonymized data, we are able to quantify the degree of preservation of data utility. The partitions obtained can be compared by means of the distance presented in section 5.1).

The analysis performed in (Martínez, Sánchez et al. 2010b) has shown that an anonymization method incorporating a proper semantic analysis was able to better retain the data utility than classical methods focused only on numerical or categorical information. Again, this conclusion is very similar to what it has been observed in the present work and shows the importance of a proper understanding of data semantics in order to interpret, process or classify them.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Chapter VII

7 Conclusions and future work

In this chapter, a final summary of the work, highlighting the contributions is provided. Then, the general conclusions of the work are given, focusing on both the semantic clustering method proposed and the new semantic similarity measures developed. The chapter also enumerates the publications derived from this work. Finally, in the last section, we suggest several lines of future work and give some ideas on how they can be tackled.

7.1 Summary and contributions

The main aim of this work has been the development of a methodology able to exploit the semantic content of terms when used in clustering methods, called *ontology based semantic clustering*. This work is a contribution in the field of Domain-Driven Data Mining, in which we have studied how domain knowledge can been exploited during the clustering process. In particular, we have studied two domain knowledge sources to infer the semantic content of terms: ontologies and the Web.

In the past, classical clustering methods dealt with numerical features. Later proposals introduce heterogeneous data matrices, but they are often limited to numerical and qualitative data. However, these approaches do not attempt to interpret the conceptual meaning of textual terms, which is crucial in many applications related to textual data. In this work, we have focused our efforts on extending these clustering methods by considering the meaning of the terms. A key aspect in the clustering process is the way in which semantic features are evaluated and compared. For doing so, semantic similarity measures have been studied.

Following the goals introduced in section 1.5, the contributions of this work are the following ones:

1. The concept of *semantic feature* is introduced to formalize those categorical features for which additional semantic information is available. Although a general frame is purposed admitting a variety of knowledge sources, the use of ontologies has been exploited as more suitable for clustering purposes.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2. Several measures to assess the similarity from ontologies and textual corpora have been studied in detail. First, two similarity measure SC_{log} and SC_{Eu} are proposed in section 2.2.1 that compiles as much taxonomic ontological knowledge as available in a input ontology. This measure is based on exploiting the full taxonomic knowledge of an ontology by considering the amount of different and equal superconcepts of a pair of concepts. Secondly, a contextualized version for the computation of the IC from the Web (section 2.2.2) has been designed named CIC_T_IR. It is based on constructing suitable queries using information available in ontologies, in particular the least common subsumer of the terms. The advantages and limitations of both contributions and of related works have been carefully analysed. After this analysis SC_{log} is selected to be used during the clustering process for its accuracy, lacking of corpora-dependency and parameter-tuning, and its low computational cost.
3. Ontology-based semantic similarity measures are affected by the degree of completeness of the ontology. In chapter 3, we proposed MSC_{log}, a new method for computing the semantic similarity from multiple ontologies extending the similarity SC_{log} measure defined in section 2.2.1, which avoids most of the limitations of previous approaches (e.g. the necessity to scale the results in order to compare the results obtained from different ontologies and the complexity of the casuistic during the integration of partial results).
4. A compatibility measure to compare objects simultaneously described by numerical, categorical, and semantic features is proposed. A generalization of Gibert's mixed metrics is presented in 4.2.1. We have proposed a weighting policy that compensates the contribution of each type of features in the global measurement.
5. An ontology based semantic clustering approach able to exploit the semantics of data is presented. Concretely, the hierarchical reciprocal neighbours algorithm with Ward criterion has been extended to include also semantic features. The integration of the different measures is done by means of the compatibility measure proposed.
6. Several tests to evaluate the accuracy of the proposed semantic similarity measures and of the semantic clustering method have been performed. Standard benchmarks have been used to compare the new semantic similarity measures against the state of the art. For the case of evaluating the behaviour of the semantic clustering, we have made some analyses comparing the quality of the clusters, considering their interpretability and distribution of the objects in the different clusters.
7. A real application of the proposal to data from tourists visiting the National Park Delta de l'Ebre (a dataset obtained in an opinion poll made to 975 visitors) has been made. Successful results were obtained and a proposal of a new profiling of the visitors of the Ebre Delta which identified more refined profiles than the ones obtained using traditional multivariate techniques.

We expect to have contributed in the semantic processing of textual data in those tasks in which clustering is part of the process. In fact, considering that classical clustering does not allow a semantic interpretation of the meaning of terms, and semantics improve results, we believe that semantic clustering can be useful. The

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

semantic processing of textual data is of great importance in different fields. In that sense, the proposed semantic clustering approach can be used in several tasks and domains that require a semantic interpretation of terms, such as electronic commerce (e.g. grouping similar products or to obtain a characterization of users), medicine (e.g. clustering of electronic health records), tourism (recommending tourist destinations to users), even in privacy preserving.

7.2 Conclusions

From the research conducted in this work, including the developed similarity measures, the clustering method and the results obtained from their evaluation, 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. Related works consider different premises for computing concept similarity from ontologies: some measures are based on the computation of the length of the path between a pair of concepts (Rada, Mili et al. 1989; Wu and Palmer 1994), other measures consider different kinds of ontological features (Tversky 1977; Petrakis, Varelas et al. 2006) and, finally, there are approaches that rely on an intrinsic computation of IC (based on the number of concept hyponyms)(Seco, Veale et al. 2004; Zhou, Wang et al. 2008). The results obtained with the ontology-based similarity measure presented in section 2.2.1 show that our approach improves most of related works because the measure shows a great accuracy against related works when evaluated using a standard benchmark.
- As other authors enounced in the past (Brill 2003; Cilibrasi and Vitányi 2004; Etzioni, Cafarella et al. 2004), 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 2.2.2) 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 low computational cost and the fact that even a single ontology is required as background in the case of the semantic similarity, and the use of publicly available enough search engines to extract statistics in the case of the computation of the IC from a resource like the Web, which contains information of almost every possible domain of knowledge, make them suitable for being applied in many domains.
- Different paradigms to compute semantic similarity can be found. Considering the advantages and disadvantages of each measure reviewed in chapter 2 and their accuracy, the ontology-based similarity measure SC_{log} presented in section 2.2.1 is clearly appropriated to be used in clustering methods. The reasons are: its high accuracy, stability, its low computational cost, the fact that it does not depend on

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

tuning parameters and the availability of large and detailed general-purpose and domain-specific ontologies.

- The similarity assessment depends on the completeness and homogeneity of the ontology. However, there are few authors that are worked in computing semantic similarity from multiple ontologies (Rodríguez and Egenhofer 2003; Al-Mubaid and Nguyen 2009). We have proposed a method that overcomes both the dependence on the completeness of a single ontology and most of the limitations and drawbacks of related works. When multiple ontologies are considered by heuristically combining the knowledge modelled in each one, we are able to exploit the benefits of a more complete and accurate knowledge representation than when using a single ontology. The results obtained show two important principles:
  - The more ontologies considered the higher the accuracy of the similarity assessment.
  - The improvements of similarity assessment due to an additional ontology increase as mach new knowledge is provided by this new ontology (redundant new ontologies provide smaller improvement).
- With respect to the ontologies studied and used in the tests, we have seen that there exist large, detailed and well-structured domain ontologies from which extract semantic evidences to support similarity assessments, such as the case of biomedicine. WordNet, in particular, has shown accurate results in many domains, being an appropriated resource for domains without large specific ontologies.
- 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. The combination of numerical, categorical and semantic features is done by using a compatibility measure. This allows treating each value that describes an object in the most appropriated way. The definition of a compatibility measure takes advantage of semantic similarity measures. Thus, improvements on the semantic feature comparisons, using this compatibility measure, bring improvements on the methods.
- 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. Moreover, the clusters obtained with several ontologies have shown that the combination of general purpose ontologies with domain ontologies seems to generate more interpretable clusters, because each ontology provides a partial and complementary view of the knowledge required in the data mining process.
- The semantic similarity functions which perform better in non-specific tests also show more goodness in clustering processes.
- Both the developed semantic similarity and the semantic clustering method 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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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.

7.3 Publications

The results of the work done in this thesis have been published in several international conferences and journals. In this section we give the list of publications, divided into three parts: first, publications related to semantic clustering and its applications; second, publications about the contributions in semantic similarity measures and their evaluation; finally, some complementary papers that deal with aspects related with this work.

In short, the work presented in this document has been published in 6 international peer-reviewed journals, 5 of them are indexed in the ISI Web of Knowledge. Contributions have also been presented in several conferences in the area, most of them being international conferences. A total of 12 papers have been published in peer-reviewed conferences.

Moreover, at the moment of writing this document, we have 3 papers submitted in ISI journals, covering the last contributions of the thesis.

1. *Publications concerning the design of the semantic clustering methods and the evaluation of the proposal.*
   1.1. Batet, M., Gibert, K., Valls., A. (2011). Semantic Clustering Based On Ontologies: An Application to the Study of Visitors in a Natural Reserve. 3th International Conference on Agents and Artificial Intelligence, Rome, Italy. In press.
   1.2. Batet, M., Valls, A., Gibert, K., Sánchez, D. (2010). 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, A. Moreno and J. Aguilar. Amsterdam, IOS Press: 207-216.
   1.3. Martínez, S., Sánchez, D., Valls, A., Batet, M. (2010). The Role of Ontologies in the Anonymization of Textual Variables. Artificial intelligence research and development. Proceedings of the 13th International Conference on the Catalan Association for Artificial Intelligence. R. Alquézar, A. Moreno and J. Aguilar. Amsterdam, IOS Press: 153-162.
   1.4. Batet, M., Valls, A., Gibert, K. (2010). 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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

1.5. Batet, M., Valls, A., Gibert, K. (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.
1.6. *Submitted*: Martínez, S., Sánchez, D., Valls, A., Batet, M. Privacy protection of textual attributes through a semantic-based masking method. Submitted to Special issue Information fusion in the context of data privacy (Information Fusion An International Journal on Multi-Sensor, Multi-Source Information Fusion)
1.7. *Submitted*: Gibert, K., Batet, M., Valls, A. Introducing semantic variables in mixed distance measures. Submitted to Data Mining and Knowledge Discovery Journal.

2. Publications concerning the survey, design and evaluation of semantic similarity measures.

2.1. Batet, M., Sánchez, D., Valls, A. (2010). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*. In press. *Imp. Fact: 2.432 (1st quartile).*
2.2. Sánchez, D., Batet, M., Valls, A., Gibert, K. (2010). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems* 35(3): 383-413. *Imp. Fact: 0.98 (3rd quartile).*
2.3. Sánchez, D., Batet, M., Valls, A. (2010). Web-Based Semantic Similarity: An Evaluation in the Biomedical Domain. *International Journal of Software and Informatics* 4(1): 39-52
2.4. Batet, M., Sánchez, D., Valls, A., Gibert, K. (2010). 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.
2.5. Sánchez, D., Batet, M., Valls, A. (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.
2.6. Batet, M., Sánchez, D., Valls, A., Gibert, K. (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.
2.7. Batet, M., Valls, A., Gibert, K. (2010). 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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

2.8. *Submitted*: Batet, M., Sánchez, D., Valls, A., Gibert, K. Computing semantic similarity from multiple ontologies: an application to the biomedical domain. Submitted to the Artificial Intelligence in Medicine Journal.
2.9. *Submitted*: Batet, M., Valls, A., Gibert, K. Measuring the Distance between Words with the Taxonomic Relations of an Ontology. Submitted to Fuzzy Sets and Systems Journal.

3. Publications addressing ccomplementary issues (study, definition application and tailoring of ontologies, as well as some other approaches to include domain knowledge in clustering)

3.1. Batet, M., Isern, D., Marín, L., Martínez, S., Moreno, A., Sánchez, D., Valls, A., Gibert, A. (2010). Knowledge-driven delivery of home care services. *Journal of Intelligent Information Systems* (in press). *Imp. Fact: 0.98 (3rd quartile)*.
3.2. Valls, A., Gibert, K., Sánchez, D., Batet, M. (2010). Using ontologies for structuring organizational knowledge in Home Care assistance. *International Journal of Medical Informatics* 79(5): 370-387. *Imp. Fact: 3.126 (1st quartile)*.
3.3. Valls, A., Batet, M., López, E.M. (2009). Using expert's rules as background knowledge in the ClusDM methodology. *European Journal of Operational Research* 195(3): 864–875. *Imp. Fact: 2.093 (2nd quartile)*.
3.4. Batet, M., Valls, A., Gibert, K., Martínez, S., Morales. E. (2009). Automatic Tailoring of an Actor Profile Ontology. Knowledge Management for Health Care Procedures, ECAI 2008 Workshop, K4HelP 2008, Patras, Greece, July 21, 2008, Revised Selected Papers. D. Riaño, Springer. LNAI 5626:104-122.
3.5. Batet, M., Martínez, S., Valls, A., Gibert, K. (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.
3.6. Batet, M., Gibert, K., Valls, A. (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.

7.4 Future work

In this section, we describe several lines for future research and present some ideas on how they can be tackled.

Regarding the clustering process, some issues can be addressed:

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

- *To study to what extend the results depend on the coverage and homogeneity of the ontologies.* This is an important issue in order to define some process for the selection of the most appropriate ontologies to be considered when executing the semantic clustering. The tests shown in this work indicate that including new ontologies does not decrease the quality of the semantic estimation, but further analyses should be done. For example, tests with reduced domain ontologies might show the degree of dependency of the quality of the results with respect to the ontology coverage.
- *Extension to multi-valued semantic features.* At the moment, semantic features only have a single value. In some applications, it is needed to allow that the values of these features could be a set of terms. For example, in touristic destinations, the value of the feature "religious buildings" in a big city could contain {chapel, cathedral, mosque}. This extension implies to have a data matrix that contains multiple terms for the values of semantic features. In addition, it is necessary to define how the relative similarities between terms pairs are managed in order to compute the global similarity. So, *n* terms (the value of a semantic feature for a particular object) will be compared with *m* terms (the value of the same feature for another object). Some aggregation of partial similarities must be defined. The use of different aggregation policies could be studied, from simple approaches where the similarity could be the average similarity from each pair of terms to complex heuristics based on the knowledge in the ontology.
- *Tools for interpretation of the clusters using the semantic features.* A new technique for generating descriptions of the clusters guided by the domain ontology could be defined. It should be studied how to use the semantic knowledge represented in the ontology in order to help the user to distinguish and correctly interpret the main characteristics that define each of the clusters. This will contribute to facilitate the interpretability of clusters to the user. One possible line of work is related to the construction of prototypes of each cluster that summarize, subsume or represent the semantics of the objects contained in a particular cluster.
- *Integration of the presented clustering approach in the DAMASK project.* First, it must be studied the requirements for the connection between the ontology-based information extraction task from heterogeneous Web resources with the clustering method developed in this work. The information extraction method provides a data matrix where the value of a semantic feature is a concept represented in the domain ontology. In this way, the presented clustering approach could work with data extracted automatically from textual resources (i.e. the Web). This integration will avoid a manually processing of the textual data in order to format them to be used with the semantic clustering. Second, the clustering will be integrated into a recommender system that suggests tourist destinations according to the personal preferences, using the tools of semantic data mining. These tasks could allow the possibility of making a wider evaluation of the semantic clustering approach presented in this work in multiple and heterogeneous domains.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Regarding to the study of semantic similarity some issues can be addressed:

- *Extending the presented similarity approach to the computation of semantic relatedness between concepts.* As stated in chapter 2 relatedness computation is a broader concept than similarity. It considers, in addition to taxonomic relations, other inter-concept non-taxonomic relationships. Even though non-taxonomic knowledge is rarer than taxonomical, large and rich ontologies such as WordNet partially models it. For example, non-taxonomic relationships contained in ontologies such as WordNet (e.g. meronyms, holonomy, antonymy or related terms) (Hirst and St-Onge 1998) or SNOMED CT (e.g. attributes or other relations that represent other characteristics of the concept) should also be studied since they could provide additional evidences about concept alikeness. In addition, for the proposed approach of IC computation from the Web, due to absolute term co-occurrence in the Web covers any kind of semantic relation between concepts, our presented approach could be easily adapted by including, as context, other types of common ontological ancestors, exploiting non-taxonomic relationships such as meronymy, holonomy, antonymy, etc. However, the exploitation of these relationships should not compromise the generality of the defined approach.
- *An study of intrinsic IC computation.* Due to related works proposing intrinsic IC computation models provided good results exploiting only an ontology, new intrinsic computation measures could be defined by exploiting as most as possible the taxonomical knowledge of an ontology.
- *Application of the ontology-based semantic similarity in privacy preserving.* Continuing with the research done in anonymization of databases with textual values, a promising area of interest for us is the use of ad-hoc ontologies for improving anonymization techniques. Ontology-based masking methods could be developed thanks to the tools of semantic similarity assessment developed in this work. For example, it could be interesting to work on the anonymization consisting in microaggregation (Domingo-Ferrer and Torra 2001). This method builds groups of similar objects taking into account some restrictions on the nature of the groups (*f.i.* preserving the diversity and ensuring some level of indistinguishability). The semantic similarity measures developed in this work could be used to compare the objects in during the microaggregation, as it has been done in clustering.
- *An study of the effect of using domain ontologies in the semantic similarity assessment.* Semantic similarities should be evaluated in specific domains in which textual comprehension and ontologies play an important role when developing new intelligent services (Prantner, Ding et al. 2007).

Considering all the research issues proposed in this section, we feel that the contributions presented in this work can be the first stone in many research works.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

References

Ahmad, A. and L. Dey (2007). "A k-mean clustering algorithm for mixed numeric and categorical data." Data Knowledge Engineering **63**(2): 503-527.

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.

Al-Mubaid, H. and H. A. Nguyen (2009). "Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies." IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **39**(4): 389-398.

Allampalli-Nagaraj, G. and I. Bichindaritz (2009). "Automatic semantic indexing of medical images using a web ontology language for case-based image retrieval." Engineering Applications of Artificial Intelligence **22**(1): 18-25.

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

Annichiarico, R., K. Gibert, U. Corrtés, F. Campana and C. Caltagirone (2004). "Qualitative profiles of disability." Journal of Rehabilitation Research and Development **41**(6A): 835-845.

Anton-Clavé, S., M.-G. Nel-lo and A. Orellana (2007). "Coastal tourism in Natural Parks. An analysis of demand profiles and recreational uses in coastal protected natural areas." Revista Turismo & Desenvolvimento **7-8**: 69-81.

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

Armengol, E. and E. Plaza (2003). Remembering Similitude Terms in CBR. Machine Learning and Data Mining in Pattern Recognition, Third International Conference, MLDM 2003, Leipzig, Germany, July 5-7, 2003, Proceedings. P. Perner and A. Rosenfeld, Springer. **LNAI 2734:** 121-130.

Aseervatham, S. and Y. Bennani (2009). "Semi-structured document categorization with a semantic kernel." Pattern Recognition **42**(9): 2067-2076.

Baeza-Yates, R. A. (1992). Introduction to data structures and algorithms related to information retrieval. Information Retrieval: Data Structures and Algorithms. W. B. Frakes and R. Baeza-Yates. Upper Saddle River, NJ, Prentice-Hall, Inc.**:** 13–27.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Ball, G. H. and D. J. Hall (1965). ISODATA, a novel method of data analysis and classification.

Banerjee, S. and T. Pedersen (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. 18th International Joint Conference on Artificial Intelligence, IJCAI 2003, Acapulco, Mexico, Morgan Kaufmann, 805-810.

Basu, S., I. Davidson and K. L. Wagstaff, Eds. (2008). Constrained Clustering: Advances in Algorithms, Theory, and Applications. Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC.

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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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, S. Martínez and E. Morales (2009). Automatic Tailoring of an Actor Profile Ontology. Knowledge Management for Health Care Procedures, ECAI 2008 Workshop, K4HelP 2008, Patras, Greece, July 21, 2008, Revised Selected Papers. D. Riaño, Springer. **LNAI 5626:** 104-122.

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

Bechhofer, S., F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinees, D. Patel-Schneider and L. Stein. (2009). "OWL Web Ontology Language Reference [http://www.w3.org/TR/owl-re] (Accessed May 2009)."

Beliakov, G., H. Bustince and J. Fernández (2010). On the median and its extensions. Computational Intelligence for Knowledge-based Systems Design, LNAI 6178, Springer**:** 435-444.

Benzécri, J. (1980). Pratique de l'analyse des données, Paris: Dunod.

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

Bergamaschi, B., S. Castano, S. D. C. d. Vermercati, S. Montanari and M. Vicini (1998). An Intelligent Approach to Information Integration. Proceedings of the First International Conference Formal Ontology in Information Systems, 253-268.

Berners-Lee, T., J. Hendler and O. Lassila (2001). "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." Scientific American **284**(5): 34-43.

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. New York, Plenum Press.

Bichindaritz, I. and S. Akkineni (2006). "Concept mining for indexing medical literature." Engineering Applications of Artificial Intelligence **19**(4): 411-417.

Blank, A. (2003). Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology. Words and Concepts in Time: towards Diachronic Cognitive

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Onomasiology. R. Eckardt, K. von Heusinger and C. Schwarze. Berlin, Germany, Mouton de Gruyter**:** 37-66.

Bollegala, D., Y. Matsuo and M. Ishizuka (2007). Measuring Semantic Similarity between Words Using Web Search Engines. 16th international conference on World Wide Web, WWW 2007, Banff, Alberta, Canada ACM, 757-766.

Bollegala, D., Y. Matsuo and M. Ishizuka (2009). A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web. Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, Singapore, Republic of Singapore, ACL and AFNLP, 803–812.

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 (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA, 10-15.

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

Cadez, I., D. Heckerman, C. Meek, P. Smyth and S. White (2003). "Model-Based Clustering and Visualization of Navigation Patterns on a Web Site." Data Mining and Knowledge Discovery **7**(4): 399-424.

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

Cao, L., P. S. Yu, C. Zhang and Y. Zhao (2010). Domain Driven Data Mining, Springer.

Carpineto, C., S. Osinski, G. Romano and D. Weiss (2009). "A survey of Web clustering engines." ACM Computing Surveys **41**(3): 1-38.

Caviedes, J. E. and J. J. Cimino (2004). "Towards the development of a conceptual distance metric for the UMLS." Journal of Biomedical Informatics **37**(2): 77-85.

Ciampi, A., Y. Lechevallier, M. Castejón and A. Gonzalez (2008). "Hierarchical clustering of subpopulations with a dissimilarity based on the likelihood ratio satistic: application to clustering massive datasets." Pattern Analysis and Applications **11**(2): 199-220.

Cilibrasi, R. L. and P. M. B. Vitányi (2004). "Automatic meaning discovery using Google." Available at: http://xxx.lanl.gov/abs/cs.CL/0412098.

Cilibrasi, R. L. and P. M. B. Vitányi (2006). "The Google Similarity Distance." IEEE Transactions on Knowledge and Data Engineering **19**(3): 370-383.

Cimiano, P. (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer-Verlag.

Cimiano, P., S. Handschuh and S. Staab (2004). Towards the self-annotating web. 13th international conference on World Wide Web, WWW 2004, New York, USA, ACM, 462 - 471.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Corchado, J. M. and C. Fyfe (2000). "A comparison of kernel methods for instantiating case based reasoning systems." Computing and Information Systems **7**: 29-42.

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

Curran, J. R. (2002). Ensemble Methods for Automatic Thesaurus Extraction. Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Philadelphia, PA, USA, Association for Computational Linguistics, 222–229.

Chavent, M., Y. Lechevallier and O. Briant (2007). "DIVCLUS-T: A monothetic divisive hierarchical clustering method " Computational Statistics & Data Analysis **52**(2): 687-701.

Chen, H.-H., M.-S. Lin and Y.-C. Wei (2006). Novel association measures using web search with double checking. 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, COLING-ACL 2006, Sydney, Australia, ACL, 1009-1016.

Chen, J. Y. and S. Lonardi, Eds. (2009). Biological Data Mining. Mining and Knowledge Discovery Series, Chapman & Hall/CRC.

Chen, Y., E. K. Garcia, M. R. Gupta, A. Rahimi and L. Cazzanti (2009). "Similarity-based classification: Concepts and Algorithms." Journal of Machine Learning Research **10**(Mar): 747-776.

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.

Deerwester, S., S. T. Dumais and R. Harshman (2000). "Indexing by latent semantic analysis." Journal of the American Society for Information Science **41**(6): 391-407.

Diday, E. (2000). Analysis of symbolic data: exploratori methods for extracting statistical information from complex data. , Springer-Verlag.

Dillon, W. R. and M. Goldstein (1984). Multivariate Analysis: Methods and Applications, Wiley.

Ding, L., T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi and J. Sachs (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. thirteenth ACM international conference on Information and knowledge management, CIKM 2004, Washington, D.C., USA, ACM Press, 652-659.

Domingo-Ferrer, J. (2008). A survey of inference control methods for privacy preserving data mining. Privacy preserving data mining: models and algorithms. **55-80**.

Domingo-Ferrer, J. and V. Torra (2001). A quantitative comparison of disclosure control methods for microdata. Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam, Elsevier**:** 11-134.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Domingo-Ferrer, J. and V. Torra (2003). "Median based aggregation operators for prototype construction in ordinal scales." International Journal of Intelligent Systems **18**(6): 633-655.

Domingo-Ferrer, J. and V. Torra, Eds. (2004). Privacy in Statistical Databases.

Doring, C., C. Borgelt and R. Kruse (2004). Fuzzy clustering of quantitative and qualitative data. Conf. North American Fuzzy Information Processing Society (NAFIPS 2004), Banff, Alberta, Canada, IEEE, 84-89.

Dujmovic, J. and H. Bai (2006). "Evaluation and Comparison of Search Engines Using the LSP Method." Computer Science and Information Systems **3**(2): 711-722.

Dunn, J. (1974). "Well separated clusters and optimal fuzzy partitions." Journal of Cybernetics **5**: 95-104.

Epler Wood, M. (2002). Ecotourism: Principles, Practices and Policies for Sustainability, United Nations Publications.

Ester, M., H.-P. Kriegel, J. Sander and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, AAAI Press, 226-231.

Etzioni, O., M. Cafarella, D. Downey, K. Kok, A. Popescu, T. Shaked, S. Soderland and D. S. Weld (2004). WebScale Information Extraction in KnowItAll. WWW2004, New York, USA.

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.

Fan, B. (2009). "A hybrid spatial data clustering method for site selection: The data driven approach of GIS mining " Expert Systems with Applications **36**(2): 3923-3936.

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

Fensel, D. (2000). "The Semantic Web and Its Languages." IEEE Intelligent Systems **15**(6): 67-63.

Fisher, D. (1987). "Knowledge acquisition via incremental conceptual clustering." Maching Learning **2**: 139–172.

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

Gangemi, A., D. Pisanelli and G. Steve (1998). Ontology Integration: Experiences with Medical Terminologies. Formal Ontology in Information Systems, IOS Press, 163-178.

Gibert, K. (1996). "The use of symbolic information in automation of statistical treatment for ill-structured domains." AI Communications **9**(1): 36-37.

Gibert, K. and U. Cortés (1997). "Weighing quantitative and qualitative variables in clustering methods." Mathware and Soft Computing **4**(3): 251-266.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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. Garcia-Rudolph, A. Garcia-Molina, T. Roig-Rovira, M. Bernabeu and J. Tormos (2008). "Response to TBI-neurorehabilitation through an AI & Stats hybrid KDD methodology." Medical Archives **62**: 132-135.

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., C. Garcia Alonso and L. Salvador Carulla (2010). "Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support." Health Research Policy and Systems **8**(28).

Gibert, K., X. Martorell, R. Massanet, M. Sánchez-Marrè, J. Martín-Sánchez and A. Martorell (2007). "A Knowledge Discovery methodology for identifying vulnerability factors of mental 44 disorder in an intellectually disabled population." Frontiers in Artificial Intelligence and Application **163**: 426-435.

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.

Gibert, K., G. Rodríguez-Silva and I. Rodríguez-Roda (2010). "Knowledge discovery with clustering based on rules by states: A water treatment application." Environmental Modelling and Software **25**(6): 712-723.

Gibert, K., E. Rojo and J. Rodas (2010). "Effects of using mean curve in the analysis of repeated time series." Acta Informatica Medica **18**(3): 140-146.

Gibert, K. and Z. Sonicki (1999). "Clustering based on rules and medical research." Journal on Applied Stochastic Models in Business and Industry **15**(4): 319-324

Gibert, K., Z. Sonicki and J. C. Mart´ın (2002). "Impact of data encoding and thyroids dysfunctions." Studies in Health Technology and Informatics **90**: 494–498.

Godo, L. and V. Torra (2000). "On aggregation operators for ordinal qualitative information." IEEE Transactions on Fuzzy Systems **8**(2): 143-154.

Goldstone, R. L. (1994). "Similarity, interactive activation, and mapping." Journal of Experimental Psychology: Learning, Memory, and Cognition **20**(1): 3-28.

Gómez-Pérez, A., M. Fernández-López and O. Corcho (2004). Ontological Engineering, 2nd printing. Springer-Verlag. ISBN: 1-85233-551-3.

Gowda, K. C. and E. Diday (1991). "Symbolic clustering using a new similarity measure." IEEE Transactions on Systems, Man and Cybernetics **22**.

Gower, J. C. (1967). "A comparison of some methods of cluster analysis." Biometrics **23**: 623-628.

Gower, J. C. (1971). "A general coefficient for similarity." Biometrics **27**: 857 – 872.

Gruber, T. R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Workshop on Formal Ontology in

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Conceptual Analysis and Knowledge Representation. N. Guarino and R. Poli. Padova, Italy, Kluwer Academic Publishers.

Guarino, N. (1998). Formal Ontology in Information Systems. 1st International Conference on Formal Ontology in Information Systems, FOIS 1998, Trento, Italy, IOS Press, 3-15.

Guha, S., R. Rastogi and K. Shim (2000). "ROCK: A robust clustering algorithm for categorical attributes." Information Systems **25**(5): 345–366.

Guha, S., R. Rastogi and K. Shim (2001). "CURE: An efficient clustering algorithm for large databases." Information Systems **26**(1): 35-58.

Gupata, S., K. Rao and V. Bhatnagar (1999). K-means clustering algorithm for categorical attributes. 1st Int. Conf. Data Warehousing and Knowledge Discovery (DaWaK'99), Florence, Italy, 203–208.

Hamasuna, Y., Y. Endo and S. Miyamoto (2010). "On tolerant fuzzy c-means clustering and tolerant possibilistic clustering." Soft Computing **14**(5): 487-494.

Han, J. and M. Kamber (2000). Data Mining: Concepts and Techniques, Morgan Kaufmann.

Han, J., M. Kamber and A. Tung (2001). Spatial Clustering Methods in Data Mining: A Survey. Geographic Data Mining and Knowledge Discovery. H. J. Miller and J. Han. London, Taylor & Francis**:** 201-231.

Hansen, P. and B. Jaumard (1997). "Cluster analysis and mathematical programming." Mathematical Programming **79**(1-3): 191–215.

Harris, Z. (1985). Distributional structure. The Philosophy of Linguistics. J. J. Katz. New York, Oxford University Press**:** 26–47.

Herrera, F. and L. Martínez (2000). "A 2-tupple fuzzy linguistic representation model for computing with words." IEEE Transactions On Fuzzy Systems **8**(6): 746-752.

Hirst, G. and D. St-Onge (1998). Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An Electronic Lexical Database. C. Fellbaum, MIT Press**:** 305–332.

Hliaoutakis, A. (2005). Semantic Similarity Measures in the MESH Ontology and their Application to Information Retrieval on Medline. Technical Report, Technical Univ. of Crete (TUC), Dept. of Electronic and Computer Engineering.

Hliaoutakis, A., G. Varelas, E. Voutsakis, E. G. M. Petrakis and E. E. Milios (2006). "Information Retrieval by Semantic Similarity." International Journal on Semantic Web and Information Systems **2**(3): 55-73.

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

Huang, H., Y. Cheng and R. Zhao (2008). A Semi-supervised Clustering Algorithm Based on Must-Link Set. 4th International Conference on Advanced Data Mining and Applications, ADMA 2008. LNAI 5139, Chengdu, China, Springer, 492-499.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Huang, Z. (1998). "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values." Data Mining and Knowledge Discovery **2**: 283–304.

Ichino, M. and H. Yaguchi (1994). "Generalized Minkowski Metrics for Mixed feature-type data analysis." IEEE Transaction on Systems, Man and Cybernetics **22**(2): 146–153.

Jain, A. K. and R. C. Dubes (1988). Algorithms for clustering data. Michigan, USA, Prentice-Hall.

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

Jajuga, K., M. Walesiak and A. Bak (2003). On the general distance measure. Exploratory Data Analysis in Empirical Research. M. Schwaiger and O. Opitz.

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.

Karypis, G., E. Han and V. Kumar (1999). "Chameleon: Hierarchical clustering using dynamic modeling." EEE Computer **32**(8): 68–75.

Kaufman, L. and P. Rousseeuw (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York, John Wiley.

Keller, F. and M. Lapata (2003). "Using the web to obtain frequencies for unseen bigrams." Computational Linguistics **29**(3): 459-484.

Kimura, M., K. Saito, R. Nakano and H. Motoda (2010). "Extracting influential nodes on a social network for information diffusion." Data Mining and Knowledge Discovery **20**(1): 70-97.

Kohonen, T. (1997). Self-Organizating Maps. New York, Springer-Verlag.

Krishna, K. and M. Murty (1999). "Genetic K-means algorithm." IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics **29**(3): 433–439.

Krumhansl, C. (1978). "Concerning the Applicability of Geometric Models to Similarity Data: The Interrelationship between Similarity and Spatial Density." Psychological Review **85**: 445-463.

Kumar, D., N. Ramakrishnan, R. F. Helm and M. Potts (2008). "Algorithms for Storytelling." IEEE Transactions on Knowledge Data Engineering **20**(6): 736-751.

Lance, G. N. and W. T. Williams (1967). "A general theory of classificatory sorting strategies: II. Clustering algorithms." Computer Journal **10**: 271-277.

Landauer, T. and S. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge." Psychological Review **104**: 211-240.

Leacock, C. and M. Chodorow (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, MIT Press**:** 265-283.

Lee, J. H., M. H. Kim and Y. J. Lee (1993). "Information retrieval based on conceptual distance in is-a hierarchies." Journal of Documentation **49**(2): 188-207.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Lee, W.-N., N. Shah, K. Sundlass and M. Musen (2008). Comparison of Ontology-based Semantic-Similarity Measures. AMIA Annual Symposium, Washington DC, USA, AMIA, 384-388.

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.

Lenat, D. B. and R. V. Guha (1990). Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project. Boston, Massachusetts, Addison-Wesley.

Li, C. and G. Biswas (1999). Temporal pattern generation using hidden Markov model based unsupervised classification. Advances in Intelligent Data Analysis: Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis. LNCS 1642. D. Hand, K. Kok and M. Berthold, Springer Berlin**:** 245-256.

Li, C. and G. Biswas (2002). "Unsupervised learning with mixed numeric and nominal data." IEEE Transactions on Knowledge and Data Engineering **14**(4): 673–690.

Li, Y., Z. Bandar and D. McLean (2003). "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources." IEEE Transactions on Knowledge and Data Engineering **15**(4): 871-882.

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. (1998a). Automatic retreival and clustering of similar words. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL 1998, Montreal, Quebec, Canada, ACL / Morgan Kaufmann, 768-774.

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

Livesay, K. and C. Burgess (1998). "Mediated priming in high-dimensional semantic space: No effect of direct semantic relationships or co-occurrence." Brain and Cognition **37**: 102-105.

López de Mántaras, R. (1991). "A distance-based attribute selection measure for decision tree induction." Machine learning **6**: 81-92.

Lord, P., R. Stevens, A. Brass and C. Goble (2003). "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation." Bioinformatics **19**(10): 1275-1283.

Lund, K. and C. Burgess (1996). "Producing high-dimensional semantic spaces from lexical co-occurrence." Behavior Research, Methods, Instruments and Computers **28**(2): 203-208.

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

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Martínez, S., D. Sánchez and A. Valls (2010a). Ontology-Based Anonymization of Categorical Values. Modeling Decisions for Artificial Intelligence - 7th International Conference, MDAI 2010 Perpignan, France, Springer, 153-162.

Martínez, S., D. Sánchez, A. Valls and M. Batet (2010b). The Role of Ontologies in the Anonymization of Textual Variables. Artificial intelligence research and development. Proceedings of the 13th International Conference on the Catalan Association for Artificial Intelligence. R. Alquézar, A. Moreno and J. Aguilar. Amsterdam, IOS Press**:** 207-216.

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.

Melton, G. B., S. Parsons, F. P. Morrison, A. S. Rothschild, M. Markatou and G. Hripcsak (2006). "Inter-patient distance metrics using SNOMED CT defining relationships." Journal of Biomedical Informatics **39**(6): 697-705.

Mena, E., V. Kashyap and A. Sheth (1996). OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies. Proceedings of the International Conference of Cooperative Information Systems (CoopIS ’96).

Mika, P. (2007). "Ontologies are us: A unified model of social networks and semantics." Web Semantics: Science, Services and Agents on the World Wide Web **5**(1): 5-15.

Miller, G., C. Leacock, R. Tengi and R. T. Bunker (1993). A Semantic Concordance. Workshop on Human Language Technology, HLT 1993, Princeton, New Jersey, Association for Computational Linguistics, 303-308.

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

Mirkin, B. (2005). Clustering for data mining: a data recovery approach. London, Chapman & Hall/CRC.

Mollineda, R. and E. Vidal (2000). A relative approach to hierarchical
clustering. Pattern Recognition and Applications, Frontiers in Artificial Intelligence and Applications. M. Torres and A. Sanfeliu. Amsterdam, The Netherlands, IOS Press. **56**.

Monreale, A., G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo and S. Wrobel (2010). "Movement Data Anonymity through Generalization." Transactions on Data Privacy **3**(2): 91-121.

Morbach, J., A. Yang and W. Marquardt (2007). "OntoCAPE—A large-scale ontology for chemical process engineering." Engineering Applications of Artificial Intelligence **20**(2): 147-161.

Mori, J., M. Ishizuka and Y. Matsuo (2007). Extracting keyphrases to represent relations in social networks from web. 20th International Joint Conference on Artificial Intelligence, IJCAI 2007, Hyderabad, India, AAAI Press, 2820-2825.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

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

Nagy, G. (1968). "Proceedings of the IEEE." State of the art in pattern recognition **56**(5): 836–862.

Nakhaeizadeh, G. (1996). "Classification as a subtask of Data Mining experiences from some industrial projects." International Federation of Classification Societies **1**: 17-20.

Naphade, M. R. and T. S. Huang (2001). Semantic filtering of video content. Storage and Retrieval for Multimedia Databases, San Jose, CA, USA, SPIE, 270--279.

Neches, R., R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator and W. R. Swartout (1991). "Enabling Technology for Knowledge Sharing." AI Magazine **12**(3): 36-56.

Noy, N. F. and M. A. Musen (1999). SMART: Automated Support for Ontology Merging and alignment. Proceedings of the 12th Banff Workshop on Knowledge Acquisition, Modeling, and Management., Banff, Alberta, Canada, 1-20.

Patwardhan, S., S. Banerjee and T. Pedersen (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation. 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, CICLing 2003, Mexico City, Mexico, Springer Berlin / Heidelberg, 241-257.

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.

Pelleg, D. and A. Moore (2000). X-means: Extending K-means with efficient estimation of the number of clusters. 17th International Conference on Machine Learning (ICML 2000), Stanford, CA, USA, Morgan Kaufmann, 727–734.

Penz, J., S. Brown, J. Carter, P. Elkin, V. Nguyen, S. Sims and M. Lincoln (2004). Evaluation of SNOMED coverage of Veterans Health Administration terms. 11th World Congress on Medical Informatics, Medinfo 2004, IOS Press, 540-544.

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

Pirró, G. and N. Seco (2008). Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Content. OTM 2008 Confederated International Conferences CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, Springer Berlin / Heidelberg, 1271-1288.

Prantner, K., Y. Ding, M. Luger, Z. Yan and C. Herzog (2007). Tourism Ontology and Semantic Management System: State-of-the-arts Analysis. IADIS International Conference WWW/Internet 2007, IADIS 2007, Vila Real, Portugal, IADIS, 111-115.

Quevedo, J., C. Gibert and J. Aguilar-Martín (1993). Fuzzy semantics in expert process control IPMU '92—Advanced Methods in Artificial Intelligence, 4th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Palma de Mallorca, Spain, July 6–10, 1992 Proceedings. **LNCS 682:** 275-283.

Rada, R., H. Mili, E. Bichnell and M. Blettner (1989). "Development and application of a metric on semantic nets." IEEE Transactions on Systems, Man, and Cybernetics **9**(1): 17-30.

Ralambondrainy, H. (1995). "A conceptual version of the K-means algorithm." Pattern Recognition Letters **16**(11): 1147-1157.

Ratprasartporn, N., J. Po, A. Cakmak, S. Bani-Ahmad and G. Özsoyoglu (2009). "Context-based literature digital collection search." International Journal on Very Large Data Bases **18**(1): 277-301.

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.

Resnik, P. (1999). "Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language." Journal of Artificial Intelligence Research **11**: 95-130.

Rham, C. d. (1980). "La classification hierarchique ascendante selon la méthode des voisins réciproques." Les Cahiers de l'Analyse des Données **V**(2): 135-144.

Rodríguez, M. A. and M. J. Egenhofer (2003). "Determining semantic similarity among entity classes from different ontologies." IEEE Transactions on Knowledge and Data Engineering **15**(2): 442–456.

Romdhane, L. B., H. Shili and B. Ayeb (2010). "Mining microarray gene expression data with unsupervised possibilistic clustering and proximity graphs." Applied Intelligence **33**(2): 220-231.

Rousseeuw, P. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of Computational and Applied Mathematics **20**(1): 53-65.

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

Sahami, M. and T. D. Heilman (2006). A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. 15th International World Wide Web Conference, WWW 2006, Edinburgh, Scotland ACM Press, 377 - 386

Sánchez, D. (2008). Domain Ontology Learning from the Web, VDM Verlag.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Sánchez, D. (2009). "Domain Ontology Learning from the Web." knowledge Engineering Review **24**(4): 413.

Sánchez, D. (2010). "A methodology to learn ontological attributes from the Web." Data & Knowledge Engineering **69**(6): 573-597.

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.

Sánchez, D., D. Isern and M. Millán (2010). "Content Annotation for the Semantic Web: an Automatic Web-based Approach." Knowledge and Information Systems. DOI: doi:10.1007/s10115-010-0302-3 (in press).

Sánchez, D. and A. Moreno (2007). "Bringing taxonomic structure to large digital libraries." International Journal of Metadata, Semantics and Ontologies **2**(2): 112-122.

Sánchez, D. and A. Moreno (2008a). "Learning non-taxonomic relationships from web documents for domain ontology construction." Data & Knowledge Engineering **63**(3): 600-623.

Sánchez, D. and A. Moreno (2008b). " Pattern-based automatic taxonomy learning from the Web." AI Communcations **21**(1): 27-48.

Sato, M., Y. Sato and L. Jain (1997). Fuzzy Clustering Models and Applications.

Schallehn, E., K.-U. Sattler and G. Saake (2004). "Efficient similarity-based operations for data integration." Data & Knowledge Engineering **48**(3): 361-387.

Seco, N., T. Veale and J. Hayes (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. 16th Eureopean Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, IOS Press, 1089-1090.

Siler, W. and J. J. Buckley (2005). Fuzzy Expert Systems and Fuzzy Reasoning. New Jersey, Wiley.

Singh, M. and G. M. Provan (1996). Efficient Learning of Selective Bayesian Network Classifiers. ICML**:** 453-461.

Sneath, P. (1957). "The application of computers to taxonomy." Journal of General Microbiology **17**: 201–226.

Sneath, P. H. A. and R. R. Sokal (1973). Numerical Taxonomy. London,UK, Freeman.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Sokal, R. R. and C. D. Michener (1958). "A statistical method for evaluating systematic relationships." University Kansas Scientific Bulletin **38**: 1409-1438.

Sorensen, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons." Biologiske Skrifter **5**: 1–34.

Sowa, J. F. (1999). Knowledge Representation: Logical, Philosophical, and Computational Foundations. Pacific Grove, California, Brooks Cole Publishing Co.

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

Studer, R., V. R. Benjamins and D. Fensel (1998a). "Knowledge Engineering: Principles and Methods." Data and Knowledge Engineering **25(1-2)**(1-2): 161-197.

Studer, R., V. R. Benjamins and D. Fensel (1998b). "Knowledge Engineering: Principles and Methods." Data & Knowledge Engineering **25**(1-2): 161-197.

Stumme, G., M. Ehrig, S. Handschuh, S. Hotho, A. Madche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, Y. Sure, R. Volz and V. Zacharia (2003). The karlsruhe view on ontologies, Technical report, University of Karlsruhe, Institute AIFB, Germany.

Sugumaran, V. and V. C. Storey (2002). "Ontologies for conceptual modeling: their creation, use, and management." Data & Knowledge Engineering **42**(3): 251-271.

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.

Valls, A. (2003). ClusDM: A multiple criteria decision making method for heterogeneous data sets, Consejo Superior de Investigaciones Científicas.

Valls, A., M. Batet and E. M. López (2009). "Using expert's rules as background knowledge in the ClusDM methodology." European Journal of Operational Research **195**(3): 864–875.

Valls, A., K. Gibert, D. Sánchez and M. Batet (2010). "Using ontologies for structuring organizational knowledge in Home Care assistance." International Journal of Medical Informatics **79**(5): 370-387.

Walesiak, M. (1999). "Walesiak,." Argumenta Oeconomica **2**(8): 167-173.

Waltinger, U., I. Cramer and TonioWandmacher (2009). From Social Networks To Distributional Properties: A Comparative Study On Computing Semantic Relatedness. Thirty-First Annual meeting of the Cognitive Science Society, CogSci 2009, Amsterdam, Netherlands, Cognitive Science Society, 3016-3021.

Wan, S. and R. A. Angryk (2007). Measuring Semantic Similarity Using WordNet-based Context Vectors. IEEE International Conference on Systems, Man and Cybernetics, SMC 2007, Montreal, Quebec, Canada, IEEE Computer Society, 908-913.

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

Weinstein, P. and W. P. Birmingham (1999). Comparing Concepts in Differentiated Ontologies. Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management (KAW'99), Banff, Alberta, Canada.

Wilbu, W. and Y. Yang (1996). "An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts." Computers in Biology and Medicine **26**: 209-222.

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.

Xu, J., W. Wang, J. Pei, X. Wang, B. Shi and A. W.-C. Fu (2006). "Utility-based anonymization for privacy preservation with less information loss." SIGKDD Explor Newsl **8**(2): 21-30.

Xu, R. and D. Wunsch (2005). "Survey of clustering algorithms." IEEE Transactions on Neural Networks **16**(3): 645-678.

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

Zadeh, L. and J. Kacprzyk, Eds. (1999). Computing with Words in Information/Intelligent Systems. Series in Studies in Fuzziness and Soft Computing.

Zadeh, L. A. (1965). "Fuzzy Sets." Information and Control **8**(3): 338-353.

Zahn, C. T. (1971). "Graph-theoretical methods for detecting and describing gestalt clusters." IEEE Transactions on Computers: 68–86.

Zhang, T., R. Ramakrishnan and M. Linvy (1996). BIRCH: An efficient data clustering method for very large data sets. ACM SIGMOD Int'l Conf. on Management of Data, Montreal, Quebec, Canada, 103-114.

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.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Annex A

In this chapter, we demonstrate that the SC_{Eu} *measure* holds the properties required for a distance measure.

The properties for a distance are:

Identity: $d(c_i, c_j) = 0 \Leftrightarrow c_i = c_j \;\; \forall c_i, c_j \in O$

Symmetry: $d(c_i, c_j) = d(c_j, c_i) \; \forall c_i, c_j \in O$

Triangle inequality: $d(c_i, c_j) + d(c_j, c_k) \geq d(c_i, c_k) \; \forall c_i, c_j, c_k \in O$

The proofs of the Identity and Symmetry are straightforward. The Identity is fulfilled because iff two concepts are the same, the union of all the ancestors is equal to the set of common ancestors (i.e. intersection), which leads to a 0 at the numerator of eq. 37. Symmetry also holds because the union and intersection operators are symmetric.

So, in this section we concentrate on the proof of the triangle inequality property. First some notation is introduced to simplify the formulation of the problem.

Let:

$$\mathrm{A} = A(c_1), B = A(c_2), C = A(c_3)$$

The set of superconcepts of c_1, c_2 and c_3 respectively.

Then, the triangle inequality property is expressed as:

$$\sqrt{\frac{|\mathrm{A} \cup B| - |A \cap B|}{|A \cup B|}} + \sqrt{\frac{|B \cup C| - |B \cap C|}{|B \cup C|}} \geq \sqrt{\frac{|\mathrm{A} \cup C| - |A \cap C|}{|A \cup C|}}$$

Proposition 2. In a taxonomic relation without multiple inheritance, this property holds:

$$A \cap B = B \cap C \text{ or } A \cap B = A \cap C \text{ or } B \cap C = A \cap C$$

***Proof*.** Assuming that proposition 1 is not true, the following conditions must be true simultaneously:

C1:

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$A \cap B \neq B \cap C \Rightarrow (1.a)\ \exists cx \in A \cap B, cx \notin B \cap C \text{ or } (1.b)\ \exists cx \notin A \cap B, cx \in B \cap C$$

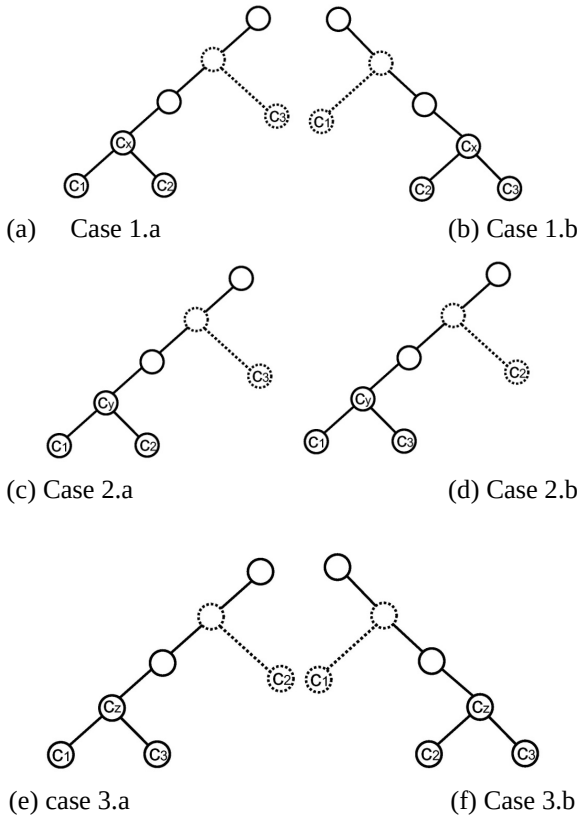
C2:

$$A \cap B \neq A \cap C \Rightarrow (2.a)\ \exists cy \in A \cap B, cy \notin A \cap C \text{ or } (2.b)\ \exists cy \notin A \cap B, cy \in A \cap C$$

C3:

$$B \cap C \neq A \cap C \Rightarrow (3.a)\ \exists cz \in A \cap C, cz \notin B \cap C \text{ or } (3.b)\ \exists cz \notin A \cap C, cz \in B \cap C$$

These six possible cases are graphically represented in figure 1.



(a) Case 1.a (b) Case 1.b

(c) Case 2.a (d) Case 2.b

(e) case 3.a (f) Case 3.b

Figure 20. Intersection cases.

From these figures, it is easy to see that:

If expr. 1.a = true $\Rightarrow$ B ∩ C = A ∩ C (Fig. 1(a)), which violates condition C3.
If expr. 1.b = true $\Rightarrow$ A ∩ B = A ∩ C (Fig. 1(b)), which violates condition C2.
If expr. 2.a = true $\Rightarrow$ B ∩ C = A ∩ C (Fig. 1(c)), which violates condition C3.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

If expr. 2.b = true $\Rightarrow$ A $\cap$ B = B $\cap$ C (Fig. 1(d)), which violates condition C1.
If expr. 3.a = true $\Rightarrow$ A $\cap$ B = B $\cap$ C (Fig. 1(e)), which violates condition C1.
If expr. 3.b = true $\Rightarrow$ A $\cap$ B = A $\cap$ C (Fig. 1(f)), which violates condition C2.

This proves that the three conditions cannot be satisfied simultaneously. □

Proof. Triangle inequality.

For simplicity, let us rename the sets of union and intersection as follows:

$$x = |A \cup B| \; and \; p = |A \cap B| \; with \; 1 \leq p \leq x$$

$$y = |B \cup C| \; and \; q = |B \cap C| \; with \; 1 \leq q \leq y$$

$$z = |A \cup C| \; and \; r = |A \cap C| \; with \; 1 \leq r \leq z$$

Notice that p=1 iff $A \cap B$ = root node of the ontology, q=1 iff $B \cap C$ = root node of the ontology and r=1 iff $A \cap C$ = root node of the ontology. In addition, the union and intersection sets are equal only if the two concepts are exactly the same, because the sets of ancestors A, B and C include the concept in study, that is, $p=x$ iff $c_1=c_2$, $q=x$ iff $c_2=c_3$ and $r=x$ iff $c_1=c_3$.

So the triangle inequality can be expressed as:

$$\sqrt{\frac{x-p}{x}} + \sqrt{\frac{y-q}{y}} \geq \sqrt{\frac{z-r}{z}}$$

That is,

$$\frac{\sqrt{(x-p)yz}}{\sqrt{xyz}} + \frac{\sqrt{(y-q)xz}}{\sqrt{xyz}} \geq \frac{\sqrt{(z-r)xy}}{\sqrt{xyz}}$$

This is equivalent to demonstrate:

$$\sqrt{xyz - pyz} + \sqrt{xyz - xqz} \geq \sqrt{xyz - xyr}$$

According to Proposition 1, there are three different configurations: the first one is when $A \cap B = B \cap C$ (Fig. 1(d) and (e)), the second one has that $A \cap B = A \cap C$ (Fig. 1(b) and (f)) and the third one corresponds to $B \cap C = A \cap C$ (Fig. 1(a) and (c)). So, the demonstration of the triangle inequality property will be done for each of those cases separately.

CASE 1: $A \cap B = B \cap C$ (Fig. 1(d) and (e))

In this situation we have that $A \cap B = B \cap C \Rightarrow p = q \; and \; p \leq r$ and the triangle inequality (eq 6) can be written as follows:

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$\sqrt{xyz - pyz} + \sqrt{xyz - xpz} \geq \sqrt{xyz - xyr}$$

$$(\sqrt{xyz - pyz} + \sqrt{xyz - xpz})^2 \geq (\sqrt{xyz - xyr})^2$$

$$xyz + xyr - pz(x + y) + 2\sqrt{xyz - pyz}\sqrt{xyz - xpz} \geq 0$$

In this case we can rewrite *x*, *y* and *z* in terms of the cardinality of the sets of ancestors A, *B* and *C*, as follows: $a = |A| - p$, $b = |B| - p$ and $c = |C| - p$. So, *a*, *b*, *c* are the number of superconcepts of *A*,*B*,*C*, respectively, that not belong to the common ancestors, $A \cap B$.

$$x = a + b + p$$

$$y = b + c + p$$

$$z = a - (r - p) + c - (r - p) + r = 2 + 2p + c - r$$

The expression obtained is:

$$a^2b + 2abc + a^2c + 2acp + ac^2 + b^2a + 2b^2p + b^2c + 2bcp + bc^2 + 2abp - ap^2$$
$$- 2p^3 - cp^2 + apr + cpr + 2bpr + 2p^2r + 2\sqrt{xyz - pyz}\sqrt{xyz - xpz} \geq 0$$

Having that $p \leq r$ (the number of concepts of $A \cap B$ is less or equal than the number of concepts of $A \cap C$) and $p \leq x$ (the number of concepts of $A \cap B$ is less or equal than the number of concepts of $A \cup B$) and $p \leq y$ (the number of concepts of $B \cap C$ is less than number of concepts of $B \cup C$), we can determine the positivity of the pairs as indicated below:

$$a^2b + 2abc + a^2c + 2acp + ac^2 + b^2a + 2b^2p + b^2c + 2bcp + bc^2 + 2abp +$$

$$\underbrace{2p^2r - 2p^3}_{\geq 0} + \underbrace{apr - ap^2}_{\geq 0} + \underbrace{cpr - cp^2}_{\geq 0} + 2bpr + 2\sqrt{\underbrace{xyz - pyz}_{\geq 0}}\sqrt{\underbrace{xyz - xpz}_{\geq 0}} \geq 0$$

So, triangle inequality holds when $A \cap B = B \cap C$.

CASE 2: $A \cap B = A \cap C$ (Fig. 1(b) and (f))

In this situation we have that $A \cap B = A \cap C \Rightarrow p = r \text{ and } p \leq q$ and the triangle inequality (eq 6) can be written as follows:

$$\sqrt{xyz - pyz} + \sqrt{xyz - xqz} \geq \sqrt{xyz - xyp}$$

$$xyz + yp(x - z) - xqz + 2\sqrt{xyz - pyz}\sqrt{xyz - xqz} \geq 0$$

In this case we can rewrite *x*, *y* and *z* as:

$$x = a + b + p$$

$$y = b - (q - p) + c - (q - p) + q = b + 2p + c - q$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$z = a + c + p$$

,where $a=|A|-p$, $b=|B|-p$, and $c=|C|-p$. Then,

$$a^2b + 2a^2p + a^2c - 2a^2q + 2abc + 4acp + ac^2 - 2acq + 4abp + 4ap^2 - 4apq + ab^2$$

$$- 2abq + b^2c + 4bcp + bc^2 - 2bcq + 2b^2p + 5bp^2 - 3bpq - cpq + 2p^3 + cp^2 - 2p^2q +$$
$$+ 2bp^2 + 2\sqrt{xyz - pyz}\sqrt{xyz - xqz} \geq 0$$

In this case, the number of superconcepts of c_2 and c_3 are greater or equal than $|B \cap C|$, so we have that $q \leq c+p$ and $q \leq b+p$. In addition, we have that $p \leq x$ and $q \leq y$. So:

$$\underbrace{a^2p + a^2b - a^2q}_{\geq 0} + \underbrace{a^2p + a^2c - a^2q}_{\geq 0} + \underbrace{acp + ac^2 - acq}_{\geq 0} + \underbrace{acp + abc - acq}_{\geq 0} + \underbrace{2ap^2 + 2acp - 2apq}_{\geq 0} +$$
$$+ \underbrace{2ap^2 + 2abp - 2apq}_{\geq 0} + \underbrace{abp + ab^2 - abq}_{\geq 0} + \underbrace{abp + abc - abq}_{\geq 0} + \underbrace{bcp + b^2c - bcq}_{\geq 0} + \underbrace{bcp + bc^2 - bcq}_{\geq 0} +$$
$$+ \underbrace{2bp^2 + 2b^2p - 2bpq}_{\geq 0} + \underbrace{bp^2 + bcp - bpq}_{\geq 0} + \underbrace{2p3 + 2bp^2 - 2p^2q}_{\geq 0} + \underbrace{cp^2 + bcp^2 - cpq}_{\geq 0} +$$
$$+ 2\sqrt{\underbrace{xyz - pyz}_{\geq 0}}\sqrt{\underbrace{xyz - xqz}_{\geq 0}} \geq 0$$

So, the triangle inequality is demonstrated when $A \cap B = A \cap C$.

CASE 3: $B \cap C = A \cap C$ (Fig. 1(a) and (c))

In this situation we have that $B \cap C = A \cap C \Rightarrow q = r \; and \; p \leq q$ and the triangle inequality (eq 6) can be written as follows:

$$\sqrt{xyz - pyz} + \sqrt{xyz - xqz} \geq \sqrt{xyz - xyq}$$

$$xyz + yq(y - z) - pyz + 2\sqrt{xyz - pyz}\sqrt{xyz - xqz} \geq 0$$

In this case we can rewrite x, y and z as:

$$x = a - (p - q) + b - (p - q) + p = a + b + 2q - p$$

$$y = b + c + q$$

$$z = a + c + q$$

,where $a=|A|-p$, $b=|B|-p$, and $c=|C|-p$ and $q \leq p$. Then,

$$a^2b + ab^2 + 4abq - 2abp + 2abc + b^2c + 4bcq - 2bcp + 2b^2q + 5bq^2 - 3bqp + a^2c + 4acq -$$
$$- 2acp + ac^2 + bc^2 + 2c^2q - 2c^2p + 4cq^2 - 4cpq - apq + aq^2 + 2q^3 - 2pq^2 +$$

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

$$+ 2\sqrt{xyz - pyz}\sqrt{xyz - xqz} \geq 0$$

And having that $p \leq a+q$, $p \leq b+q$, $p \leq x$ and $q \leq y$:

$$\underbrace{a^2b + abq - abp}_{\geq 0} + \underbrace{ab^2 + abq - abp}_{\geq 0} + \underbrace{abc + bcq - bcq}_{\geq 0} + \underbrace{b^2c + bcq - bcp}_{\geq 0} + \underbrace{2b^2q + 2bq^2 - 2bqp}_{\geq 0} +$$
$$\underbrace{abq + bq^2 - bqp}_{\geq 0} + \underbrace{a^2c + acq - acp}_{\geq 0} + \underbrace{abc + acq - acp}_{\geq 0} + \underbrace{ac^2 + c^2q - c^2p}_{\geq 0} + \underbrace{bc^2 + c^2q - c^2p}_{\geq 0} +$$
$$\underbrace{2acq + 2cq^2 - 2cqp}_{\geq 0} + \underbrace{2bcq^2 + 2cq^2 - 2cqp}_{\geq 0} + \underbrace{abq + aq^2 - aqp}_{\geq 0} + \underbrace{2bq^2 + 2q^3 - 2q^2p}_{\geq 0} +$$
$$2\sqrt{\underbrace{xyz - pyz}_{\geq 0}}\sqrt{\underbrace{xyz - xqz}_{\geq 0}} \geq 0$$

Finally, the triangle inequality is demonstrated when $B \cap C = A \cap C$. □

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Annex B

In this chapter different software tools for clustering that are available, some of them developed by commercial companies, are listed:

Table 26. Data mining software tools

Name	SAS Enterprise Miner
Link	http://www.sas.com/technologies/analytics/datamining/miner/
Free	No
Description	Is a software tool able to perform data mining processes based on analysis of vast amounts of data with a broad set of tools. SAS provides a variety of clustering algorithms. It provides the different hierarchical agglomerative algorithms (single linkage, average linkage, complete linkage, centroid linkage or Ward's method). Also, it provides the K-Means algorithm, and the Self-Organizing Maps (SOM) algorithm (Kohonen Networks). It provides a wide range of graphic tools in order to study the results.
Name	SPSS
Link	http://www.spss.com/
Free	No
Description	SPSS is a powerful statistical tool and is one of the most widely used programs for statistical analysis in social science. It include descriptive statistics (Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics), bivariate statistics (Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests), Prediction for numerical outcomes (Linear regression), Prediction for identifying groups (Factor analysis, cluster analysis (two-step, K-means, hierarchical), Discriminant). It implements different clustering methods such as average linkage, single linkage, complete linkage, centroid method, median method, Ward's method and K-means.
Name	Clementine from SPSS
Link	http://www.spss.com/la/productos/clementine/clementine.htm
Free	No

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Description	SPSS Clementine provides two clustering algorithms, which are the K-Means algorithm, and the Self-Organizing Maps algorithm (Kohonen Networks). SPSS Clementine cannot cluster data hierarchically and it cannot cluster data set that has categorical variables. However, it can cluster a data set specifying the number of clusters before the process using K-Means algorithm. Also, it can cluster a data set without specifying the number of clusters before the process using the Kohonen Network algorithm.
Name	Intelligent Miner (IBM)
Link	http://www-306.ibm.com/software/data/iminer/
Free	No
Description	IBM Intelligent Miner is a set of "statistical, processing, and mining functions" to analyze data. It contains three main products: Intelligent Miner Modeling, Intelligent Miner Scoring, and Intelligent Miner Visualization. The first one develops analytic models such are Associations, Clustering, Decision trees, and Transform Regression PMML models via SQL API. The second one performs scoring operation for the models created by Intelligent Miner Modeling. The last one presents data modeling results using one of the following Visualizers: Associations Visualizer, Classification Visualizer, Clustering Visualizer, and Regression Visualizer. IBM's Intelligent Miner provides a variety of data mining techniques: Predictive modeling, Database segmentation or clustering, Link analysis (associations), Neural Classification, Neural Clustering, Sequential Patterns, Similar Sequences, Radial Basis Function (RBF)-Prediction, and Deviation detection (outliers). In particular, it provides only two clustering algorithms: the Demographic algorithm, and the Self-Organizing Maps algorithm (Kohonen Networks). So, IBM DB2 cannot cluster data set hierarchically and cannot cluster a data set based on a predefined number of clusters. However it can cluster a data set that has categorical variables using the Demographic algorithm.
Name	WEKA (Waikato Environment for Knowledge Analysis)
Link	http://www.cs.waikato.ac.nz/ml/weka/
Free	Yes
Description	Developed in the university of Waikato, New Zealand, Weka is a collection of machine learning algorithms for data mining tasks, implemented in Java. This software is one of the most complete ones of those free software packages. It can be executed from a command-line environment, or from a graphical interface, or it can be called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization, and is well-suited for developing new machine learning schemes. In particular, it contains different algorithms of clustering such as Cobweb, DBScan (Density-Based Spatial Clustering of Applications with Noise), EM (Expectation-Maximisation), FarthestFirst, FilteredClusterer, OPTICS (Ordering Points To Identify the Clustering Structure), x-means, MakeDensityBased-Clusterer algorithm, SimpleKMeans, CLOPE, SiB (sequential Information Bottleneck).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Name	Pentaho
Link	http://www.pentaho.com/
Free	No
Description	Pentaho Data Mining, provides a comprehensive set of machine learning algorithms from Weka. Its broad suite of classification, regression, association rules, segmentation, decision trees, random forests, neural networks, and clustering algorithms can be used to help an analyst understand the business better and to improve future performance through predictive analytics. So it have the same clustering algorithms than weka.
Name	RapidMiner
Link	http://rapid-i.com/
Free	Yes
Description	RapidMiner (formerly YALE (Yet Another Learning Environment)) is an open source toolkit for data mining. RapidMiner implements different clustering methods as, DBScan, EM, the Weka clustering schemes, Kernel K-Means, K-Means, K-Medoids, a Random Clustering, and an implementation of Support Vector Clustering.
Name	R language
Link	http://www.r-project.org/
Free	Yes
Description	R is a programming language and environment for statistical computing and graphics. Available for Windows, various Unix flavors (including Linux), and Mac. Provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. And it has interoperability with other languages as C, XML and Java. A number of different clustering methods are provided in this software. *Ward's* minimum variance method aims at finding compact, spherical clusters. The *complete linkage* method finds similar clusters. The *single linkage* method adopts a 'friends of friends' clustering strategy. The other methods can be regarded as aiming for clusters with characteristics somewhere between the single and complete link methods. Note however, that methods "median" and "centroid" are *not* leading to a *monotone distance* measure, or equivalently the resulting dendrograms can have so called *inversions* (which are hard to interpret).
Name	Rattle, Gnome Cross Platform GUI for Data Mining using R
Link	http://rattle.togaware.com/
Free	Yes
Description	Rattle(R Analytical Tool To Learn Easily) is a data mining toolkit used to analyze very large collections of data. Rattle presents statistical and visual summaries of data, transforms data into forms that can be readily modeled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets. It has a simple and logical graphical user interface based on Gnome. Rattle runs under GNU/Linux, Macintosh OS/X, and MS/Windows. In addition, Rattle can be used by itself to deliver data mining projects. Rattle also provides an entry into sophisticated data mining using the open source and free statistical language R.

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Name	Tanagra
Link	http://chirouble.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
Free	Yes
Description	It is a free (open-source) data-mining package that contains components for Data source (tab-delimited text), Visualization (grid, scatterplots), Descriptive statistics (cross-tab, ANOVA, correlation), Instance selection (sampling, stratified), Feature selection and construction, Regression (multiple linear), Factorial analysis (principal components, multiple correspondence), Clustering, Supervised learning (logistic regr., k-NN, multi-layer perceptron, prototype-NN, ID3, discriminant analysis, naive Bayes, radial basis function), Meta-spv learning (instance Spv, arcing, boosting, bagging), Learning assessment (train-test, cross-validation), and Association (Agrawal a-priori). It provides different clustering methods such as kMeans, Kohonen's Self Organization Map, LVQ (Kohonen's Learning Vector Quantizers), a "supervised" clustering algorithm, and HAC (Hierarchical agglomerative clustering).

Name	STATISTICA Data Miner
Link	http://www.statsoft.com/
Free	No
Description	*STATISTICA Data Miner* contains a selection of data mining solutions, with an easy-to-use user interface and deployment engine. STATISTICA Data Miner is highly customizable and can be tailored to meet very specific and demanding analysis requirements through its open architecture. Some characteristics are machine Learning (Bayesian, Support Vectors, Nearest Neighbour), General Classification/Regression tree models, General CHAID models,Boosted Tree Classifiers and Regression, Random Forests for Regression and Classification, MARSplines (Multivariate Adaptive Regression Splines), Cluster Analysis, Combining Groups (Classes) for Predictive Data Mining, Automatic Feature Selection, Ensembles of Neural Networks, etc. It provides different clustering methods such as kMeans and Generalized EM.

Name	*CLUTO*
Link	http://glaros.dtc.umn.edu/gkhome/views/cluto/
Free	Yes
Description	CLUTO is a family of data clustering and cluster analysis programs and libraries, that are well suited for low- and high-dimensional data sets. CLUTO is well-suited for clustering data sets arising in many diverse application areas including information retrieval, customer purchasing transactions, web, GIS, science, and biology. It has multiple classes of clustering algorithms (partitional, agglomerative (single-link, complete-link, UPGMA), and graph-partitioning based) and multiple similarity/distance functions (Euclidean distance, cosine, correlation coefficient, extended Jaccard, user-defined).

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC CLUSTERING
Montserrat Batet Sanroma
ISBN:9788469432327/DL:T. 1043-2011

Name	Oracle Data Mining (ODM)
Link	http://www.oracle.com/technology/products/bi/odm/index.html
Free	NO
Description	Oracle Data Mining is an option of Oracle Corporation's Relational Database Management System (RDBMS) Enterprise Edition (EE). It contains several data mining and data analysis algorithms for classification, prediction, regression, clustering, associations, feature selection, anomaly detection, feature extraction, and specialized analytics. ODM offers well known machine learning approaches such as Decision Trees, Naive Bayes, Support vector machines, Generalized linear model (GLM) for predictive mining, Association rules, K-means (Enhanced k-means (EKM)) and Orthogonal Partitioning Clustering (O-Cluster), and Non-negative matrix factorization for descriptive mining.
Name	DBMiner
Link	http://www.pentaho.com/
Free	No
Description	DBMiner implements a wide spectrum of data mining functions, including generalization, characterization, association, classification, and prediction. By incorporating several interesting data mining techniques, including attribute-oriented induction, statistical analysis, progressive deepening for mining multiple-level knowledge, and meta-rule guided mining, the system provides a user-friendly, interactive data mining environment with good performance.The underlying algorithm used in DBMoner is the *k*-means method.