Measuring Similarity in Ontologies by Means of Boolean Matrices

Montserrat Batet[a], Aïda Valls[a], Karina Gibert[b]
[a] *ITAKA, Intelligent Technologies for Advanced Knowledge Acquisition, Dept. of Computer Engineering and Maths, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain.*
[b] *Dept. Statistics and Operational Research, Universitat Politècnica de Catalunya, Ed. C5, c/ Jordi Girona 1-3, Barcelona, Catalonia, Spain.*

Abstract Ontologies are a suitable way for representing complex domains. Our work is focused on using ontologies as background knowledge to improve the results of the traditional clustering methods. In this report, a new proposal to measuring similarity in ontologies is presented. This approach is based on projecting the ontology into a numerical data matrix and then applying Euclidean distances. Different ways of projecting the knowledge are described and compared. The results, obtained with two different medical ontologies, show that the results are quite similar in some cases. Finally, the advantages and drawbacks of each approach are highlightened.

1. Introduction

Nowadays, ontologies have emerged as an important research area. Ontologies are used in applications related to knowledge management, natural language processing, e-commerce, intelligent integration information, information retrieval, bio-informatics, education, etc. The main benefit of using ontologies is the possibility of reusing and sharing knowledge.

Our recent research has been focused in apply ontologies as background knowledge during the clustering process, in particular they are used to compare linguistic attributes in order to improve the clustering results reducing the arbitrariness of traditional clustering algorithms [Xu, 2005].

Many approaches of measures to compute the similarity and the relatedness between concepts in an ontology have been defined [Pedersen, 2004, Nguyen, 2006, Zhang, 2007]. The semantic similarity measures can be classified into two classes: (1) to measure the semantic similarity is by using the is-a properties of the ontology that define taxonomic relations [Wu & Palmer 1994, Leacock & Chodorow 1998, or Nguyen, 2006], and (2) to use training corpora and information content (IC), which is a corpus–based measure of the specificity a concept, to estimate the semantic similarity and relatedness between two concepts[Nguyen, 2006].

This report is a first step to familiarize with ontologies before stating our work with semantic similarity measures and relatedness measures, which are introduced in section 2. The background knowledge provided by the ontologies and the similarity results are used in order to group a set of elements in clusters. Incorporating background knowledge in the clustering process the arbitrariness of the clustering algorithm is expected to be reduced and the clusters obtained will be more suitable and interpretable in a particular domain. The purpose of this report is to present and compare three ways for calculate the similarity between concepts in an ontology transforming the representation space. In this approach we only consider the taxonomical relations among the concepts in the ontology. These relations are represented into a Boolean matrix. In particular, this report proposes three ways of building that matrix (section 4):

(1) the number of different adjacent concepts (superclasses of a concept and subclasses of a concept),

(2) the number of superclasses or fathers and

(3) the paths between a particular concept and the root node of the taxonomy. That is, the number of different ancestors (fathers, grandfathers, etc) in the ontology hierarchy tree.

The medical domain is one of the more active in defining ontologies [Pisanelly, 2004a]. For example, the web-based application called BioPortal[1] allows to access to the Open Biomedical Ontologies (OBO) library. This library contains a large collection of ontologies in biomedicine as well as model organism communities, biology, chemistry, anatomy, radiology, and medicine. In section 5, we compare the three approaches to similarity measurement with two medical ontologies: Galen ontology and an ontology named APO (Actor Profile Ontology) that has been build in the EU Research Project K4CARE. This kind of ontologies are developed to solve problems such as the demand for reusing and sharing patient data and medical knowledge. They are introduced in more detail in section 3.

2. Ontologies

Ontologies help to build better and more interoperable information systems [Pisanelli, 2004b]. There are many definitions about what an ontology is. One of the most known was given by Studer et al. [Studer et al., 1998]: "An ontology is a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group". Sowa [Sowa, 1999] explains the subject of an ontology as: "The subject of an ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalogue of types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D". Neches et al .[Neches 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.

We can classify ontologies 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, 1990].
- *Domain-ontologies*: describe the vocabulary related to a generic domain by specializing the concepts introduced in the top-level ontology. They are a set of concepts within a domain and their relationships. There are a lot of examples of this type of ontologies in e-commerce UNSPSC[2], NAICS[3]. Or in medicine GALEN[4]; UMLS[5]; ON9, etc.;
- *Task ontologies*: describe the vocabulary related to a generic task or activity by specializing the top-level ontologies. For instance, the Scheduling Task Ontology [Mizogucgui et al. 1995] includes generic names, generic verbs, generic adjectives, etc.
- *Application ontologies*: In application ontologies, concepts often correspond to roles played by domain entities. They contain all the definitions needed to model the knowledge required for a particular application.

Sometimes, the information regarding a particular domain can be represented by a taxonomy (an is-a hierarchy). In that case, taxonomies are also considered ontologies [Studer et al, 1998] if they provide a consensual conceptualization of the domain [Lassila and McGuineness, 2001]. For instance, UNSPSC, Gene Ontology, proposals for standards on the e-commerce domain, or a taxonomy for searching the Web. In relation to these considerations, two types of ontologies are distinguished [Gómez-Pérez, 2004]:

- *Lightweight* ontologies, which usually are taxonomies, these ontologies include concepts, concept taxonomies, relationships between concepts, and properties that describe concepts.

[1] The website is maintained by The National Center for Biomedical Ontology that is part of the National Centers for Biomedical Computing supported by the NIH Roadmap, USA. URL: http://www.bioontology.org/tools/portal/bioportal.html

[2] www.unspsc.org

[3] http://www.naics.com/

[4] http://www.opengalen.org/

[5] http://www.nlm.nih.gov/research/umls/

- *Heavyweight* ontologies, which model the domain in a deeper way and provide more restrictions on domain semantics, these ontologies add axioms and constraints to lightweight ontologies.

3. Case Study

In this section we present the ontologies that will be used to study the similarity approach that we are going to present in this report. As it has been said, we have chosen two ontologies in the medical domain: APO ontology and GALEN ontology. Although they are ontologies for quite specific domains, they are quite large. In the case of the APO, we have 399 concepts, and in the Galen there are about 23142 concepts. For this reason, only a portion of these ontologies will be used to illustrate the results obtained in this study.

3.1. Galen ontology

The Galen ontology is an ontology developed by the non-profit organization OpenGalen[6]. Galen is a representative ontology in the domain of medicine. This ontology has a primary taxonomy that is divided in four primary conceptual categories (with are subclasses of the DomainCategory). This taxonomy is represented with dotted lines in **Figure 1**:

- Structures (GeneralisedStructure): abstract or physical things with parts independent of time
- Substances(generalizedSubstance): continuous abstract or physical things independent of time
- Processes(GeneralizedProcess): changes which occur over time
- Modifiers(ModifierConcept): properties, or characteristics of structures, substances or processes

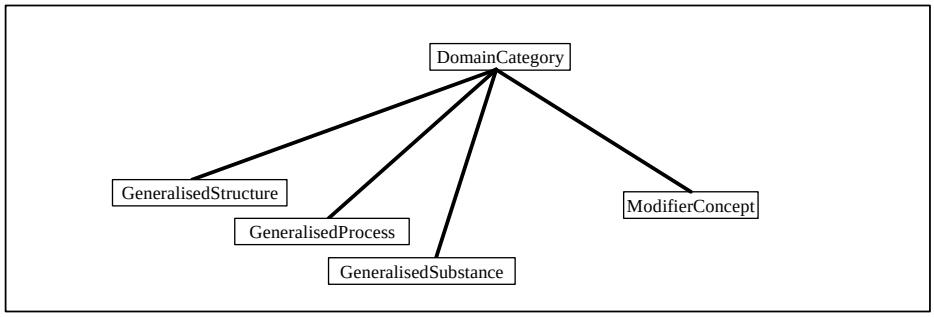

Figure 1 Overview of Galen ontology

A secondary asserted taxonomy is superimposed over the primary, in order to capture the medical intuition of 'disease' or 'disorder'. The category *Phenomenon* represents the disjunction of domain categories, more specifically with structures, processes, and substances, and with the modifiers of feature, state, and collection. (the category *Phenomenon* and its subclasses are represented in bold lines in Figure 2):

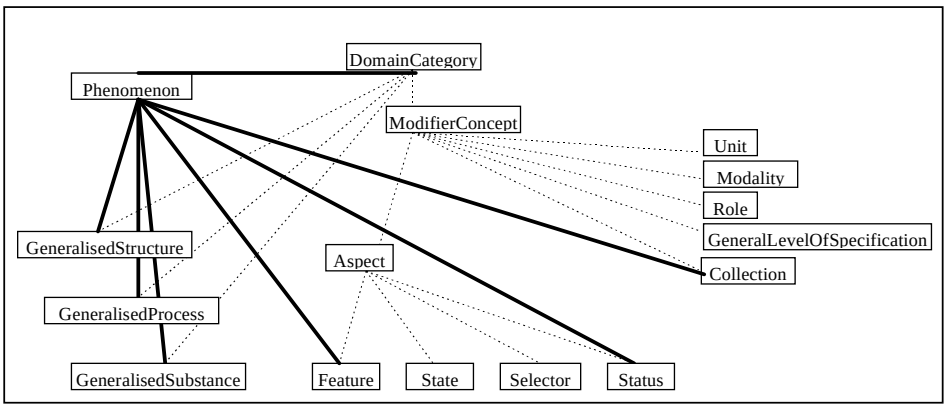

Figure 2. Secondary Hight Level Taxonomy

[6] http://www.opengalen.org/

3.2. Actor Profile Ontology ontology

The Actor Profile Ontology (APO) is an ontology that has been developed in the EU K4Care project[7]. The European project K4Care: "Knowledge-Based Homecare eServices for an Ageing Europe" is an ehealth project that tries to manage the care of senior citizens and chronically disabled persons developing a distributed platform to ensure the home-care assistance of the increasing number of senior population.

The APO defines the data of the actors involved into the sanitary model of home-care defined in the K4Care project [Casals, 2008]. This model intends to be a standard in Europe. In the K4Care model people involved in the home-care assistance is organized in three groups of actors: patients, stable members, and additional care givers. Moreover, the APO also stores data about the services, procedures and actions. All these concepts are organised in a hierarchical structure.

In a simple approach, it could be said that the K4Care system provides a web-based platform to support the provision of home-care services to the patients. The APO contains the knowledge needed to automatically know the services in which actors can take part, the procedures of those services (a non-ordered list of actions that conform a service) and the particular actions that each actor can perform. In addition the ontology has a relation of all the documents that are created and read during the provision of services. Finally, the APO stores the permissions of the different kinds of actors to read and/or write each document, which is very important to prevent that non-authorised actors access to sensitive and private documents.

This ontology represents the minimum elements needed to provide a basic HomeCare assistance according to the HomeCare model proposed in the K4CARE project: HCNS (Home Care Nuclear Services). The model allows the definition of HC Accessory Services (HCAS) that extend the HCNS with specialized services such those coming from Oncology or Rehabilitation units. This packages of services are known as Care Units. So, a secondary asserted taxonomy is superimposed over the primary, in order to capture the Care Units. In **Figure 4** the secondary taxonomy is represented with dotted lines while the solid lines represent the primary hierarchy. The APO is represented by an ontology written in OWL[8].

The following list of classes summarises up the contents of the APO presented in this section.

- Entity: An Entity is somebody who can perform an action in the K4Care project. This class is divided in two subclasses:
  - Actor: The Actor class represents somebody in the K4Care. Actors are people interacting with the HC system. The class Actor is divided in subclasses such as Stable Members (composed of Family Doctors, Physicians in Charge, Head Nurses, Social Workers and Nurses), Additional Care Givers (Specialist Physician, Social Operator, Continuous Care Provider, Informal Care Giver), and Patients.
  - Evaluation Unit: this class represents a medical evaluation unit, a group that works together in the assessment and evaluation of patients. It is composed of a Family Doctor, a Physician in Charge, a Head Nurse, and a Social Worker.
- Services: The service class represents the home-care services. These services are classified into Access services, Patient Care services, and Information services.
- Procedure: this class represents a procedure associated to a service.
- Action: The class Action represents a set of simple actions that are executed inside one or more procedures.
- Document: This class represents the documents that are read or written during the execution of one action. These documents are stored in the EHR.

[7] K4CARE: www.k4care.net
[8] OWL: http://www.w3.org/tr/owl-features/. (2007)

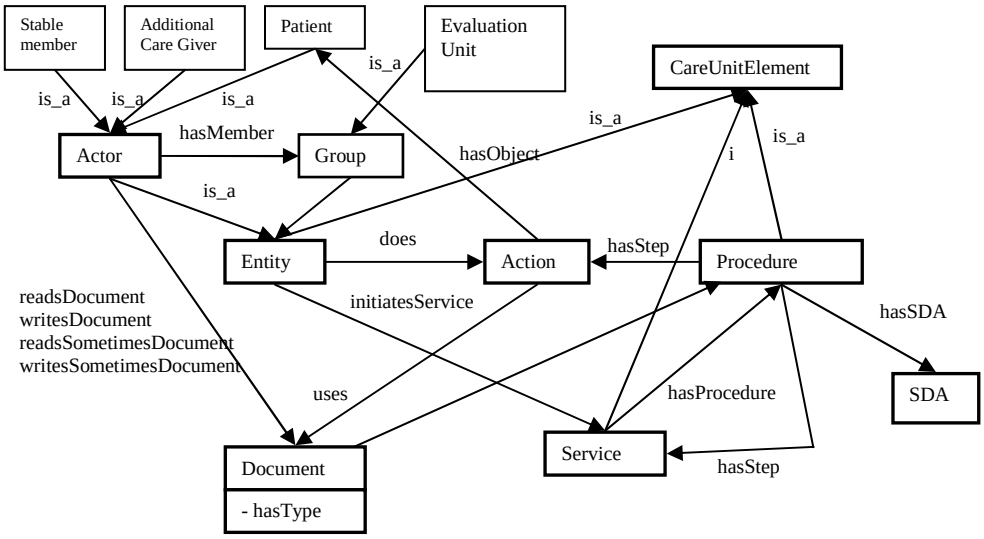

Figure 3. APO structure

4. Representing the ontology taxonomy with a Boolean matrix

In clustering, the input data is a set of objects described by means of a set of attributes. The information on the objects is usually given in a data matrix of valued pairs (object x attribute). To generate the clusters, there are well-known distances that determine the closeness between a pair of objects represented by numerical or Boolean values [Xu, 2005].

Since the goal of our work is to use ontologies to improve the clustering process, it is interesting to use them in the similarity calculation. To take advantage of the existing similarity measures, different ways of representing an ontology by means of Boolean matrices are presented in this paper. In this approach, the data matrix has *n* objects in the rows and *n* attributes in the columns. Both the objects and the attributes are all the concepts in the ontology. Then, a value of 1 represents an *is-a* relation between the object in the row and the concept in the column, and a 0 means that there is no is-a relation between them.

4.1. Representation

The is-a relations in the taxonomy of the ontology can be represented in different ways. Three possibilities are presented, analysed and compared in this report:

(1) an adjacency matrix, which represents the superclasses and subclasses of the concept C,

(2) a matrix that only represents the superclasses of C, but not its descendents and

(3) a matrix that represents all the ancestors of C, that is, the path to the root concept.

According to each of these models, different ways of measuring proximity between concepts in an ontology are implemented and tested. All of them are based on the Euclidean distance as it is explained below.

4.1.1. Adjacency matrix

An adjacency matrix represents the adjacent concepts of each particular concept in the ontology. That is, the superclasses and subclasses in the is-a hierarchy of a particular concept. In this approach a Boolean matrix *X* of dimension *n*x*n* is built, where *n* is the number of concepts of the ontology. Each row x_i represents the adjacent concepts of a particular concept *i*, in which each column contains 0 if the column concept is not an adjacent concept of *i*, or 1 if it is an adjacent concept.

4.1.2. Fathers or superclasses

In this case, the matrix only represents the superclasses of a concept, but not its descendents. Each row x_i represents the superclasses of the concept *i*. A column contains 0 if the column concept is not a superclass of *i*, or 1 if it is a superclass. So, The entry x_{ik} is indicating if concept *k* is the superclasses of concept *i*.

4.1.3. Ancestors

The last proposal consists in taking into account all the ancestors that is the path from a particular concept to the root concept. In this approach, each row x_i represents the ancestors of the concept *i*, in which each column contains 0 if the column concept is not an ancestor of *i*, or 1 if it is an ancestor.

4.2. Distance

When the information of the ontology is represented in a Boolean data matrix, the relation or similarity between concepts can be calculated using standard distances. The advantage of this approach is that distances for numerical/Boolean values are well defined. In this report, the similarity between concepts is calculated using the Euclidean distance, which is defined as:

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

The distance δ gives values between $0 \leq \delta \leq 1$, where the values near 0 mean that the two concepts are similar, while values near 1 mean that the two concepts are very different.

To normalize this distance into the unit interval, the result of Eq. 1 is divided by the square of the maximum distance between concepts of the ontology, that is, the maximum different number of adjacent concepts, fathers or ancestors between two concepts in *X*.

$$\delta(x_i, x_j) = \frac{\sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}}{\sqrt{\max.\delta_X}} \qquad (2)$$

5. Results

We implemented the proposed approaches and conducted comparisons on their respective results. In all these approaches, ontologies are viewed as a graph limited to represent the taxonomic aspect of the ontology. As it has been explained, we have tested our approach with two different medical ontologies: the APO ontology and the GALEN ontology.

5.1. Study with the APO Ontology

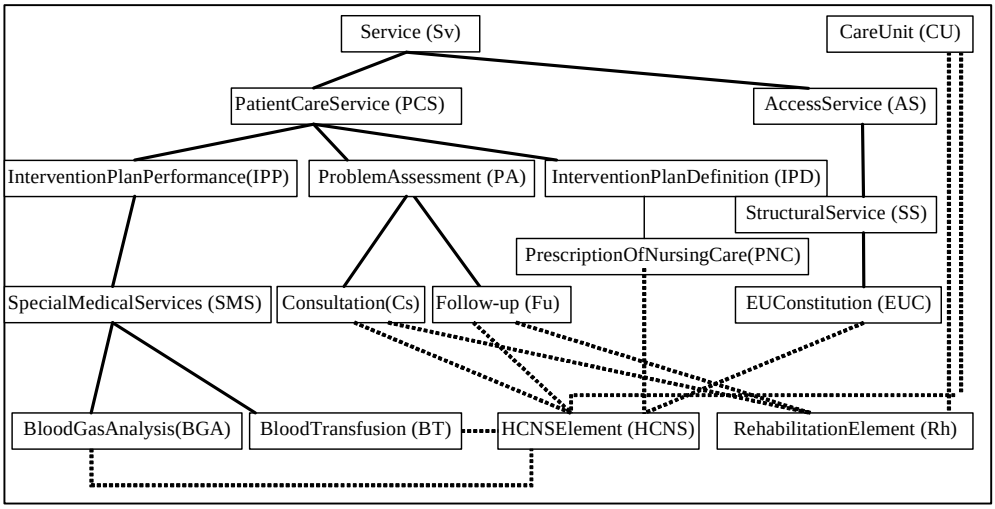

Figure 4. A subset of the APO ontology

Figure 4 shows the portion of the Actor Profile Ontology that has been used to make this study. The taxonomical information of the APO ontology has been projected into a Boolean matrix of ancestors to be able to test the behaviour of the Euclidean distance. Table 1 shows the matrix of superclasses of each concept. Table 3 shows the representation of the ancestors of 9 concepts of the APO. Columns correspond to the 17 nodes in this subset of the ontology. Finally, Table 2 is a subset of the table that represents the superclasses and the subclasses of each concept. In fact, there are concepts that have more subclasses. For example, the subclasses of ProblemAssessment are: ClinicalAssessment, ComprensiveAssessment, PhysicalExamination, Follow-up, RequestOfLaboratoryAnalysisService, FunctionalEvaluation, Multi-DimensionalEvaluation, SocialNeeds-NetworkAssessment, Consultation, EnvironmentalEvaluation, RequestOfDiagnosticExaminationService.

Table 1. Boolean matrix that represents the superclasses of APO concepts.

	PCS	CU	HCNS	Rh	IPP	SMS	Cs	FU	IPD	PNC	PA	BT	BGA	EUC	Sv	SS	AS
Cs			1	1			1				1						
FU			1	1				1			1						
IPD	1								1								
PNC			1						1	1							
PA	1										1						
BT			1			1						1					
BGA			1			1							1				
EUC			1											1		1	
Sv															1		

Table 2 Boolean ma matrix that represents the subclasses and superclasses of APO concepts.

	PCS	CU	HCNS	Rh	IPP	SMS	Cs	FU	IPD	PNC	PA	BT	BGA	EUC	Sv	SS	AS
Cs			1	1			1				1						
FU			1	1				1			1						
IPD	1								1	1							
PNC			1						1	1							
PA	1						1	1			1						
BT			1			1						1					
BGA			1			1							1				
EUC			1											1		1	
Sv	1														1		1

Table 3. Boolean matrix that represents the ancestors of APO concepts.

	PCS	CU	HCNS	Rh	IPP	SMS	Cs	FU	IPD	PNC	PA	BT	BGA	EUC	Sv	SS	AS
Cs	1	1	1	1			1				1				1		
FU	1	1	1	1				1			1				1		
IPD	1								1						1		
PNC	1	1	1						1	1					1		
PA	1										1				1		
BT	1	1	1		1	1						1			1		
BGA	1	1	1		1	1							1		1		
EUC		1	1											1	1	1	1
Sv															1		

Table 4, **Table 5**, and **Table 6** show the results of applying the Euclidean distance to these three different Boolean matrices. Each of those tables shows the distances between each pair of concepts on **Figure 4**.

Table 4. APO Euclidean distances of the superclasses matrix (F)

	Cs	FU	IPD	PNC	PA	BT	BGA	EUC	Sv
Cs	0.0	0.5	0.866	0.791	0.707	0.791	0.791	0.791	0.866
FU		0.0	0.866	0.791	0.707	0.791	0.791	0.791	0.866
IPD			0.0	0.612	0.5	0.791	0.791	0.791	0.707
PNC				0.0	0.791	0.707	0.707	0.707	0.791
PA					0.0	0.791	0.791	0.791	0.707
BT						0.0	0.5	0.707	0.791
BGA							0.0	0.707	0.791
EUC								0.0	0.791
Sv									0.0

Table 4 shows the obtained distances using the matrix of superclasses. The similarity distance values are in [0.5, 0.886] and there are 5 different possible values. Notice that in many cases the distances are equal (for example, in 18 cases the result is 0.791). In conclusion, this way of representing the ontology is not very informative.

Table 5. APO Euclidean distances of the adjacency matrix (J)

	Cs	**FU**	**IPD**	**PNC**	**PA**	**BT**	**BGA**	**EUC**	**Sv**
Cs	0.0	0.277	0.65	0.439	0.707	0.439	0.439	0.439	0.588
FU		0.0	0.65	0.439	0.707	0.439	0.439	0.439	0.588
IPD			0.0	0.48	0.832	0.62	0.62	0.62	0.62
PNC				0.0	0.784	0.392	0.392	0.392	0.555
PA					0.0	0.784	0.784	0.784	0.784
BT						0.0	0.277	0.392	0.555
BGA							0.0	0.392	0.555
EUC								0.0	0.555
Sv									0.0

Table 5 shows the obtained distances using the adjacency matrix. The similarity distance values are between [0.277, 0.832]. In this case, the number of cases in which the distances are equal is lower because now not only the superclasses are used but also the subclasses.

The similarities obtained with the 9 concepts of the APO presented in Table 3 are shown in the following table. The similarity distance values are between [0.378, 0.707]. In addition, this range is wider than in **Table 4**.

Table 6. Euclidean in the numerical matrix representing ancestors (A)

	Cs	**FU**	**IPD**	**PNC**	**PA**	**BT**	**BGA**	**EUC**	**Sv**
Cs	0.0	0.378	0.655	0.598	0.535	0.655	0.655	0.707	0.655
FU		0.0	0.655	0.598	0.535	0.655	0.655	0.707	0.655
IPD			0.0	0.463	0.378	0.655	0.655	0.707	0.378
PNC				0.0	0.598	0.598	0.598	0.655	0.598
PA					0.0	0.655	0.655	0.707	0.378
BT						0.0	0.378	0.707	0.655
BGA							0.0	0.707	0.655
EUC								0.0	0.598
Sv									0.0

The results of the three measures, representing ancestors (A), superclasses (F) and adjacent concepts (J), show that the F matrix and J matrix don't provide, in general, accurate results when the pair of concepts to compare are far away from each other. This leads to different results. In that case, the results depend extremely of the number of superclasses and subclasses of the concepts to compare. For example, *similarity*(IPD, Sv) and *similarity*(PA, Sv) are different in **Table 5** although there are the same number of nodes between (IPD, Sv) and (PA, Sv) while they are equal in **Table 6**. These pair of concepts are brothers but in **Table 5** the *similarity*(PA, Sv) is lower than *similarity*(IPD, Sv) because PA has a greater number of subclasses. Another example can be found in *similarity*(BT, Sv) and *similarity*(Cs, Sv). Although they are brothers, in **Table 4** and Table 5 the first one pair is greater. This occurs because BT has a lower number of superclasses. Finally, PA and IPD are brothers but in **Table 5** the similarity is low because PA has a lot of subclasses. In fact, subclasses of a concept do not define its semantics as the ancestors do. In general, **Table 6** takes into account the number of ancestors and provides more accurate results than the F matrix. Moreover it considers all the possible category trees in the ontology at the same time, not only the shortest path in one of the categories.

5.2. Study with the GALEN Ontology

Figure 5 shows the portion of the Galen Ontology that has been used to make this study because it is quite large. In fact, it combines the primary and the secondary taxonomies represented in **Figure 1** and **Figure 2**.

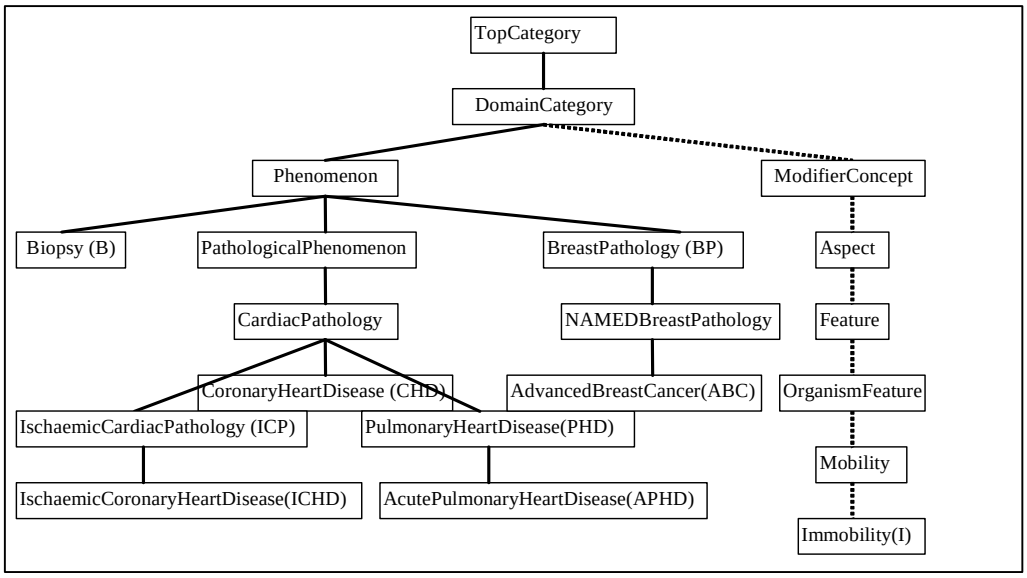

Figure 5. A subset of the GALEN ontology

The similarities obtained with the 9 concepts of the GALEN presented in **Figure 5** are shown in the following tables.

Table 7. GALEN Euclidean distances of the superclasses matrix (F)

	B	**ABC**	**BP**	**ICP**	**ICHD**	**CHD**	**PHD**	**APHD**	**I**
B	0.0	1.0	0.707	1.0	1.0	1.0	1.0	1.0	1.0
ABC		0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
BP			0.0	1.0	1.0	1.0	1.0	1.0	1.0
ICP				0.0	0.707	0.707	0.707	1.0	1.0
ICHD					0.0	1.0	1.0	1.0	1.0
CHD						0.0	0.707	1.0	1.0
PHD							0.0	0.707	1.0
APHD								0.0	1.0
I									0.0

Table 7 shows the obtained distances using the matrix of superclasses. The similarity distance values are in [0.707, 1.0]. Most of the results are equal (1.0) because the concepts do not share any superclasses. When the distance is 0.707 they share one superclass. In this case, the distance only shows if a pair of concepts are brothers or not.

Table 8. Galen Euclidean distances of the adjacency matrix (J)

	B	**ABC**	**BP**	**ICP**	**ICHD**	**CHD**	**PHD**	**APHD**	**I**
B	0.0	0.577	0.5	0.816	0.577	0.577	0.707	0.577	0.577
ABC		0.0	0.5	0.816	0.577	0.577	0.707	0.577	0.577
BP			0.0	0.866	0.645	0.645	0.764	0.645	0.645
ICP				0.0	0.577	0.707	0.816	0.816	0.816
ICHD					0.0	0.577	0.707	0.577	0.577
CHD						0.0	0.577	0.577	0.577
PHD							0.0	0.408	0.707
APHD								0.0	0.577
I									0.0

Table 8 shows the obtained distances using the adjacency matrix. The similarity distance values are between [0.5, 0.866].

Table 9. Euclidean in the numerical matrix representing ancestors (A)

	B	**ABC**	**BP**	**ICP**	**ICHD**	**CHD**	**PHD**	**APHD**	**I**
B	0.0	0.471	0.333	0.471	0.527	0.471	0.471	0.527	0.624
ABC		0.0	0.333	0.577	0.624	0.577	0.577	0.624	0.707
BP			0.0	0.471	0.527	0.471	0.471	0.527	0.624
ICP				0.0	0.236	0.333	0.333	0.408	0.707
ICHD					0.0	0.408	0.408	0.471	0.745
CHD						0.0	0.333	0.408	0.707
PHD							0.0	0.236	0.707
APHD								0.0	0.745
I									0.0

Finally, **Table 9** shows the obtained distances using the matrix of ancestors. The similarity distance values were in [0.236, 0.745]. In this case, the number of cases in which the distances are equal is lower and the results are more accurate.

The results applying our measurements to the GALEN ontology provide more information. J (**Table 7**) provides the poorest results. In that case, the measure only detects if a pair of concepts are brothers or if they have a father-son is-a relation. In the rest of cases the similarity is 1 because the number of different superclasses between two nodes and the maximum number of different superclasses is equal. This situation is really habitual when the concepts of an ontology have only one superclass.

In addition, we can also observe that the results depend extremely of the number of superclasses and subclasses of the pair of concepts to compare. For example, *similarity* (ICP, B) and *similarity* (ICP, BP) are equal in **Table 9** because B and BP are brothers while *similarity* (ICP, ABC) is greater. However, in **Table 8**, *similarity* (ICP, BP) and similarity (ICP, ABC) are equal although they aren't brothers. This occurs because BP and ABC have the same number of superclasses and subclasses. In addition, *similarity* (ICP, B) and *similarity* (ICP, BP) are different although B and BP are brothers. This occurs because they have different number of subclasses. Finally, we can highlight in **Table 8** that *similarity* (ICP, CHD) or *similarity* (ICP, PHD) is very small although they are brothers because the number of subclasses of each other.

6. Conclusions

This report has studied the behaviour of three approaches to measure the similarity between concepts in ontologies by means of Boolean data matrices. Three different matrices have been implemented to represent the concepts of the ontology: the adjacency matrix, a matrix that only takes into account the superclasses of a concept, and finally, a matrix that takes into account all the concepts between a particular concept and the root.

The Euclidean distance has been applied to each of these matrices to analyse their capacity to represent the semantics in the taxonomical relations of the ontology.

The results obtained with two different ontologies indicate that the adjacency matrix and the matrix of superclasses do not provide good results when the pair of concepts to compare is far away from each other. In these two cases, the similarity strongly depends of the number of superclasses and subclasses of the pair of concepts.

In addition, these two approaches do not take into account the depth of these concepts. However, in general, two concepts should be more similar if they are in a lower level of the ontology because this indicates that they are more specific, while concepts on the top of the ontology are more general. In conclusion, we can say that considering only the superclasses and subclasses is not enough to appropriately measure the similarity between concepts.

Using the matrix of ancestors we can elicit more semantic information of how similar two concepts are, because this approach is considering all the taxonomical relations from a particular concept to the root. In this case, the depth of the concepts in the ontology is also measured. Moreover, this measure takes into account if the concepts belong to more than one hierarchy inside the ontology. However, we have also noticed that sometimes considering all the hierarchies together can lead to non intuitive results. This should be studied in more detail.

The study concludes that the matrix of ancestors provides more relevant information about the similarity of two concepts in relation to all the taxonomy classification. The distances applied to the matrix of ancestors provided good results for measuring similarities. In general, the dimension of the

ontology also influences the results. As bigger the concepts' superclasses, subclasses or ancestors are, smaller the differences in the distance are. This effect deserves also a more detailed study.

In future work, our research will be focused on using the semantic similarities into a clustering process. We want to show that the knowledge in ontologies can improve the clustering results.

7. Acknowledgements

This work has been funded by the Student Research Grant of the University Rovira i Virgili. Authors would also like to acknowledge the support of the European K4CARE project (IST-2004-026968), ARES (CSD2007 -0004) and EAEGIS (TSI-2007-65406-C03).

8. References

[Casals, 2008]
Casals, J., Gibert, K. and Valls, A.: Enlarging a medical actor profile ontology with new care units. Knowledge Management for Health Care Procedures,From Knowledge to Global Care, LNAI 4924. 2008. 101-116.

[Gómez-Pérez, 2004]
Gómez-Pérez, A., Fernández-López, M. and Corcho, O.: Ontological Engineering, 2nd printing. Springer Verlag. ISBN: 1-85233-551-3. 2004.

[Guarino, 1998]
Guarino, N: Formal Ontology in Information Systems. In. Guarino N. (ed) 1st International Conference on Formal Ontology in Information Systems (FOIS'98). Trento, Italy, IOS Press. 1998. 3-15.

[Lassila and McGuineness,2001]
Lassila, O. and McGuineness, D.: The Role of Frame-Based Representation on the Semantic Web. Technical report KSL-01-02. Knowledge Systems Laboratory. Standford University. Satdford. California. 2001.

[Leacock,1998]
Leacock, C., and Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed., *WordNet: An electronic lexical database*. MIT Press. 1998. 265–283.

[Lenat, 1990]
Lenat, D.B. and Guha, R.V.: Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Boston, Massachusetts .1990.

[Neches, 1991]
Neches, R., Fickes, R.E., Finin, T. Gruber, T.R., Senator, T. and Swartout W.R.: Enabling technology for knowledge sharing. AI Magazine 12(3). 1991. 36-56

[Nguyen, 2006]
Al-Mubaid, H. and H. A. Nguyen: New Ontology-based Semantic Similarity Measure for the Biomedical Domain. In Proceedings of the IEEE Conference on Granular Computing, *GrC-2006*. Atlanta, GA, May 10-12, 2006. 623-28.

[Patwardhan , 2003]
Patwardhan, S. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, Univ. of Minnesota, Duluth. 2003.

[Pedersen, 2004]
Pedersen, T., Patwardhan S. and Michelizzi, J.: WordNet: Similarity –Measuring the Relatedness of Concepts. *AAAI*. 2004. 1024-1025.

[Pisanelly , 2004a]
Pisanelly, D.: Ontologies in Medicine. IOSPress. ISBN 1-58603-418-9. 2004.

[Pisanelly ,2004b]
Pisanelli, D.: If ontology is the solution, what is the problem? IOS Press. 2004.

[Studer , 1998]
Studer R, Benjamins VR, Fensel D Knowledge Engineering: Principles and Methods. IEEE Transactions on Data and Knowledge Engineering 25 (1-2). 1998. 161-197.

[Sowa, 1999]
Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, California . 1999.

[Wu,1994]
Wu, Z., and Palmer, M.: Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*. 1994. 133–138.

[Xu, 2005]
Xu, R. and Wunsch, D., I.: Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16 (3). May 2005. 645-678

[Zhang, 2007]
Zhang, X.; Jing, L.; Hu, X.; Ng, M. & Zhou, X.: A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering. *Advances in Databases: Concepts, Systems and Applications DASFAA . 4443*. 2007. 115-126.