Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

DETERMINATION, VISUALIZATION AND INTERPRETATION OF SEMANTIC SIMILARITY AMONG GEOGRAPHIC ONTOLOGIES

Marinos Kavouras, Margarita Kokla, Eleni Tomai

School of Rural and Surveying Engineering, National Technical University of Athens, 15780 Zografos Campus, Athens, Greece

1. INTRODUCTION

A close inspection of existent ontologies shows that although they refer to the same concepts, they use different semantics due to different contexts. This "Babel Tower" makes the association process and the establishment of an integrated ontology very problematic.

In the framework of our ontological research (http://ontogeo.ntua.gr/), previous work ([1], [2]) has introduced a methodology for the integration of heterogeneous ontologies in order to achieve semantic interoperability. In this endeavor, we further need to identify and resolve unambiguously existing heterogeneities. The purpose of the present research is to extract semantic information from definitions and to enrich the representation of categories with semantic relations [3] in order to disambiguate geographic categories. The ability to represent and visualize the degree of semantic similarity with concept mapping tools [4] can greatly facilitate the entire process.

For tackling these semantic heterogeneities, we explore the similarities/ dissimilarities of three well-known ontologies for geographic categories – CORINE LC, MEGRIN and WordNet. This paper shows to what extent differentiation of the given ontologies, with regard to the definitions used, influence the association of similar concepts.

2. APPROACHES ON DETERMINING SEMANTIC SIMILARITY

Similarity plays an important role in the categorization process. Furthermore, the determination of semantic similarity helps deal with semantic heterogeneities and therefore, facilitates interoperability. For this reason, several approaches have been trying to model the notion of similarity. As far as, similarity of geographic entity classes is concerned, [5] and [6] have proposed a computational method for assessing similarity between two ontologies by a similarity function that compares the distinguishing features of the entities such as parts, functions and attributes.

This paper goes into comparing the entities' definitions found in ontologies by Natural Language Processing and extract semantic relations for them. Also, it goes a step further into visualizing these ontologies as to provide metric information for relations among entities like inclusion, junction, and disjunction and therefore, provide a framework for the integration process of these specific ontologies and its result.

3. THE THREE ONTOLOGIES

In this approach we extract semantic relations from two European categorization schemata and WordNet. The problems of association and integration mentioned before are met when trying to compare distinct repositories of geographic information such as the following.

- CORINE LC [7] is a land cover categorization schema intended to provide consistent localized geographical information on the land cover of the member states of the European Community, by using satellite data. The CORINE Land

Cover has three hierarchies of categories. The upper level consists of 5 categories, the middle level of 15, and the lowest one of 44 categories;

- GDDD-Geographical Data Description Directory, MEGRIN's GDDD [8] contains information on the digital geographic information available from Europe's National Mapping Agencies. (NMAs). Layer names, feature type names and feature attribute types names correspond to the nomenclature used in the DIGEST Feature and Attribute Coding Catalogue (FACC);
- WordNet [9] on the other hand is a lexical database for the English language that has been designed, as its founders say, based on current psycholinguistic theories.

In this approach, as far as CORINE is concerned, we examined categories of the lowest level of the hierarchies. Also, for reasons of simplicity, we decided to restrict the study only to a small number of categories from the three ontologies, but still very representative ones, because they account for the heterogeneities we are trying to deal with. So the categories in question are taken form:

- CORINE LC's categories 4 (wetlands) and 5 (water bodies),
- MEGRIN's category hydrography, and then from
- WordNet we examined the definitions the database provides for the corresponding categories.

We ended up with the definitions of 17 distinct "category_types" (table 1). The term category_type refers to categories that can be found in different ontologies under the same terminology (name of the category) but exhibit differentiation in their definitions or the contexts they are used under.

Ontology	Category_type
CORINE Land Cover	Peat bog
	Water course
	Water body
MEGRIN	Bog
	Canal
	Lake/ pond
	Salt marsh
	Salt pan
	Watercourse
WordNet	Body of water
	Bog
	Canal
	Lake
	Pond
	Salt pan
	Watercourse
	Watercourse

Table 1 The category_types used in our approach

4. EXTRACTION OF SEMANTIC RELATIONS

The purpose of the present research is to extract semantic information from definitions and to enrich the representation of categories with semantic relations in order to exhibit similarities and heterogeneities between them. The field of Natural Language Processing develops methodologies for the automatic extraction of semantic information from definitions. According to [10], definitions include wealth of knowledge expressed in natural language, which can be analysed by Natural Language Processing Systems.

Definitions are a kind of text with special structure and content. They are rich sources of knowledge and they reflect scientific knowledge of a domain. In geographic ontologies, definitions are the primary and usually the only descriptions of category terms, since other elements that could contribute to the semantic definition of geographic concepts (e.g.,

properties, functions, axioms) are either missing or are superficially described. Research on definitions is seeking ways to exploit the wealth of information latent in these special kinds of text.

Definitions are comprised of two parts: the genus and the differentiae. The genus or hypernym is the superordinate term of the defined word. For example, in the definition: "hotel: a building where travelers can pay for lodging and meals and other services", "building" is the genus of category "hotel".

The differentiae are other elements of the definition apart from the genus, which differentiate words with the same genus. Thus, in the definition: "skyscraper: a very tall building with many storeys", the word "skyscraper" has the same genus (i.e., "building") with the word "hotel", but the two words are distinguished by the differentiae (e.g., "where travelers can pay for lodging and meals and other services" and "tall", "with many storeys").

The methodology for analyzing definitions and extracting information in the form of semantic relations was introduced by [10] and further pursued by [11] and [12]. This approach consists in the syntactic analysis of definitions and in the application of rules, which examine the existence of certain syntactic and lexical patterns. Patterns take advantage of specific elements of definitions, in order to identify a set of semantic relations and their values based on the syntactic analysis. Patterns applied in the genus part of the definition extract the hypernym or superordinate relation. Patterns applied in the differentiae part extract other semantic relationships such as: purpose, location, material, time, part-of, size, etc.

The pattern for the extraction of the semantic relation PURPOSE [12] is: if the verb *used* (*created*, *intended*, *prepared*, *provided*, etc.) is post modified by a prepositional phrase with the preposition for, then create a PURPOSE relation with the head(s) of that prepositional phrase as the value. For example, a PURPOSE relation is extracted from the definition: "*canal: a manmade or improved natural waterway used for transportation"*, with value "transportation".

The methodology for extracting semantic information is used to analyze definitions of geographic categories into a set of semantic relations and their corresponding values. This formalized semantic information is further used to disambiguate similar categories by explicitly and objectively identifying similarities and heterogeneities between them.

More specifically, if the methodology for extracting semantic information is used for the analysis of category "lake" as defined by MEGRIN: "lake/pond: a body of water surrounded by land", three semantic relations are extracted: HYPERNYM with value "body", MATERIAL with value "water" and SURROUNDED BY with value "land". Respectively, from the analysis of same category as it is defined by WordNet: "lake: a body of (usually fresh) water surrounded by land", the same semantic relations and values are extracted. Therefore, it is evident that the two ontologies define equivalently the category "lake" (table 2).

	HYPERNYM	MATERIAL	SURROUNDED BY
Lake (MEGRIN)	body	water	land
Lake (WordNet)	body	water (usually fresh)	land

Table 2 Extraction of semantic relations for the category "lake"

If however, the above methodology is used for the analysis of category "ditch" as defined by the same ontologies, MEGRIN and WordNet, the resulting semantic relations and values identify heterogeneities between the definitions of the homonymous categories (table 3).

	HYPERNYM	PURPOSE	SIZE	NATURE
Ditch (MEGRIN)	channel	irrigation or drainage		
Ditch (WordNet)	waterway		small	natural

Table 3 Extraction of semantic relations for the category “ditch”

5. VISUALIZATION OF THE THREE ONTOLOGIES

To visualize the different ontologies we use the Multi-Dimensional Scaling method. The method uses a similarity/dissimilarity matrix to project the data into the projection space. In our case the projection space is a two dimensional space.

It is a dimensionality reduction method that represents multi-dimensional data sets by using a stress function [13] so that distances among data reflect the corresponding (dis)similarities. The value of the stress function is an indicator of the goodness-of –fit of the result. The higher its value the more the distortion imposed on the visualization of the entities; therefore the distances are greater than the corresponding dissimilarities. The output is a scatter plot of the data where similar entities are close in the representation space while dissimilar ones are far away.

The data that we have in this case are represented by a string of binary values (table 4). Value 0 denotes that the entity lacks the specific property while value 1 denotes that the entity possesses this property. When referring here to properties we take into account the semantic relations extracted from the Natural Language Processing as described earlier in the paper.

The similarity measure S is set by the ratio model in table 4, where C are the properties common in entities a and b, A are the properties in entity a but not in b, and B are the properties in entity b but not in a. as it can be understood the ratio is bounded between 0 and 1, the former denoting complete dissimilarity, the latter coincidence of the entities.

Entity a (0,0,0,1,1,0,1,1,1,0) **Entity b** (1,0,0,0,0,0,1,1,1,1)	**S (a, b) =** $\frac{C}{A+B+C}$
So the similarity **S** for the given entities is S = 3/7 = **0.4286**	

Table 4 Examples of two entities’ property values and the measure of similarity.

6. INTERPRETATION OF RESULTS

As mentioned before, the output of the MDS is the set of coordinates for the entities in question. Then a clustering method is used to form homogeneous groups of entities that enjoy common properties. By this procedure we want to form a unified categorization schema out of the ones used in the analysis. We are then able to explore whether differences in naming denote the same entities, while sameness in naming but differentiation of the definitions provided for these entities denote distinct ones.

In the current approach we used a hierarchical clustering method to examine in what way the three distinct ontologies contribute in the formation of common upper-categories in a unified schema. The most significant finding of the clustering procedure is that we have a measure of the “consensus” for a specific category in the three ontologies, which is given to us by the distances of the projection of the same category_types into the projection space of the visualization.

7. CONCLUSIONS AND FURTHER WORK

The present research focuses on the extraction of semantic information from definitions of geographic concepts in order to identify and formalize similarities and heterogeneities between them. Visualization of semantic similarity proves to be a very useful tool for the association of similar concepts. Portraying similarities/ dissimilarities in a projection space gives us a concrete measure of the homogeneities/ heterogeneities of distinct schemata. We can then draw inferences:

- as to what extent different ontologies can be integrated;
- for the relations between category_types, like union, disjunction, inclusion;
- concerning the "consensus" in defining same categories in cross-ontological examination.

This work fits in the general puzzle of semantic interoperability. The next step involves the full implementation of this method and its incorporation in the development of a semantic translator for geodata collections.

8. BIBLIOGRAPHICAL REFERENCES

[1] M. Kokla & M. Kavouras, Fusion of top-level and geographical domain ontologies based on context formation and complementarity, *International Journal of Geographical Information Science* 15 (7): 679-687, 2001.

[2] M. Kavouras & M. Kokla "A method for the formalization and integration of geographical categorizations", *International Journal of Geographical Information Science*, 16 (5): 439 – 453, 2002

[3] M. Kokla & M. Kavouras, Extracting Latent Semantic Relations from Definitions to Disambiguate Geographic Ontologies, *GIScience 2002, 2nd International Conference on Geographic Information Science*, Boulder, Colorado, USA, 25-28 September 2002.

[4] E. Tomai & M. Kavouras, "Sharpening" Vagueness: Identifying, Measuring, and Portraying its Impact on Geographic Categories, *GIScience 2002, 2nd International Conference on Geographic Information Science*, Boulder, Colorado, USA, 25-28 September 2002.

[5] A. Rodríguez, M. Egenhofer, & R. Rugg, Assessing Semantic Similarities Among Geospatial Feature Class Definitions, *Interoperating Geographic Information Systems, Second International Conference, Interop '99*, Zurich, Switzerland, A. Vckovski, K. Brassel, and H.-J. Schek (eds.), *Lecture Notes in Computer Science*, Vol. 1580, Springer-Verlag, pp. 189-202, March 1999.

[6] A. Rodríguez & M. Egenhofer, Determining Semantic Similarity Among Entity Classes from Different Ontologies, *IEEE Transactions on Knowledge and Data Engineering*, 12 (2), 2003.

[7] European Environmental Agency: CORINE Land Cover Methodology and Nomenclature, http://reports.eea.eu.int/COR0-part1/en, http://reports.eea.eu.int/COR0-part2/en

[8] MEGRIN's PETIT project: http://www.eurogeographics.org/megrin/PROJECTS/PETIT/Prototyp_desc.html

[9] WORDNET 1.7.1 - a Lexical Database for the English Language, Cognitive Science Laboratory, Princeton University, http://www.cogsci.princeton.edu/~wn/

[10] K. Jensen & J.L. Binot, Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics* 13 (3-4): 251-60, 1987.

[11] Y. Ravin, Disambiguating and Interpreting Verb Definitions. *Natural Language Processing: The PLNLP Approach*, K. Jensen, G. E. Heidorn and S. D. Richardson (eds.), Kluwer Academic Publishers, Boston/Dordrecht/London, 1993.

[12] L. Vanderwende, The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries. Ph.D. thesis, Faculty of the Graduate School of Arts and Sciences, Georgetown University, Washington, D.C., 1995.

[13] J. B. Kruskal & M. Wish, *Multidimensional Scaling*. Sage University Paper series on Quantitative Applications in the Social Sciences, number 07-011. Sage Publications, Newbury Park, CA, 1978.