ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
CONSIGLIO NAZIONALE DELLE RICERCHE

A. Formica, E. Pourabbas

CONTENT BASED SIMILARITY OF GEOGRAPHIC CLASSES ORGANIZED AS PARTITION HIERARCHIES

R. 624 Dicembre 2004

Anna Formica – Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR, Viale Manzoni 30 - 00185 Roma, Italy. Email: anna.formica@iasi.cnr.it.

Elaheh Pourabbas – Istituto di Analisi dei Sistemi ed Informatica del CNR, Viale Manzoni 30 - 00185 Roma, Italy. Email: elaheh.pourabbas@iasi.cnr.it

To appear in Knowledge and Information Systems (KAIS), 2009.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica, CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.rm.cnr.it
URL: http://www.iasi.rm.cnr.it

Abstract

In this paper we propose a method to measure the semantic similarity of geographic classes organized as partition hierarchies within *Naive Geography*. The contribution of this work consists in extending and integrating the *information content* approach, and the method for comparing concept attributes in the ontology management system *SymOntos* developed at IASI. As a result, this proposal allows us to address both the concept similarity within the partition hierarchy, and the attribute similarity of geographic classes and, therefore, to reduce the gap among the different similarity approaches defined in the literature.

Keywords: *Geographic class, Naive Geography, similarity reasoning, partition hierarchy, information content.*

1. Introduction

Semantic similarity (or semantic proximity) consists of a field of study whereby a set of terms/concepts is assigned a metric based on the commonalities and differences of their semantic content. It covers wide areas and application domains in artificial intelligence, cognitive science, databases, and software engineering. Recently, this study is growing in importance in different settings, such as digital libraries, heterogeneous databases, and the World Wide Web.

The research on concept similarity is particularly relevant for the retrieval and integration of data within *Geographic Information Systems* (*GIS*s) [42]. In fact, users of geographic data have different backgrounds and a very rich domain of concepts, due to the varieties in human languages for expressing and communicating geographic information. In the literature, a number of semantic similarity methods have been proposed. They refer essentially to *map* similarity [2], similarity *search* and *retrieval* [44, 47, 55], similarity of *geographic events* and *processes* [29], *spatial relationships* [20], *geospatial entity-classes* [42], etc.

In this paper, the problem of semantic similarity of *geographic classes* is addressed. Geographic classes are related by the inclusion relationship and organized as partition hierarchies [35, 45]. A geographic class is defined by a name, a tuple of typed attributes, and a set of geometric types. Partitions are an important spatial concept, and are widely used in the generation of many scale-dependent maps from single databases [40, 18]. Examples of partitions in the real world are the subdivision of a given territory into administrative boundaries such as countries, states and counties, the classification of a land according to soil type, etc. In this paper, we refer to territorial administrative partitions.

The proposed method, called *GSim*, has been defined by revisiting, extending, and integrating two approaches: (i) the *information content* approach as defined in [27], that is inspired by [38], and (ii) a method for attribute similarity which is based on the *maximum weighted matching* problem in bipartite graphs, that has been proposed within the ontology management system *SymOntos* [15]. The latter is based on axiomatic similarity degrees for concept names (*axiomatic similarity*), that are established according to a consensus system by a panel of experts in the application domain. In this paper, attribute similarity has been extended and modified as explained below.

The contribution of this paper is manyfold. First, we define the notion of *information content similarity* (*ics*) by introducing the notion of *SynSet*, i.e., a set of synonyms for a geographic knowledge base, and by integrating the notion of similarity given in [27]. Second, attribute similarity is evaluated according to the notion of *ics* (rather than the axiomatic similarity defined in *SymOntos*). Third, we address the similarity of typed attributes which was not defined in [15]. Fourth, in our approach we also include a similarity measure for geometric types, which is a first preliminary proposal to be refined in future work.

Our approach falls under the general area of *Naive Geography* [11] or common-sense geographic reasoning that is a field of study concerned to capture people's concepts about space and time. The main task is to derive common sense definitions of geographic entities or classes (entities modeled in GIS) such that the classes are described by their essential properties. These properties cover three components of geographic information, which are spatial or geometric, thematic and temporal. The spatial or geometric component refers to the geographic coordinates of entities on Earth's surface. The thematic component refers to the characteristics of the entities (e.g., density of population, temperature, territorial subdivision, geomorphology), and temporal component refers to the variation of the spatial and thematic components over time. Research on spatial similarity is focused on the matching of the geometric properties of geographic objects. Some examples are content-based image retrieval [6, 22], topological relations [3], metric details of spatial relations [12]. The studies on semantic similarity are addressed to capture the thematic properties of geographic classes leaving out the similarity of the geometric properties (see for instance [42]). Our goal goes beyond the spatial similarity and it is mainly oriented to achieve the common-sense definition of the GISs users through the semantic similarity in order to bridge the gap between user needs and the concepts provided from GIS.

Note that in the literature most of similarity models defined by computer scientists are based on the hierarchical structure of concepts, in contrast with the work of psychologists who typically focus on concept features (or attributes) [42]. Our proposal represents an effort to capture both the concept similarity within the hierarchy, and the attribute similarity (here referred to as *tuple* similarity) of geographic classes. In this paper we focus on partition hierarchies, however our approach can also be applied to ISA hierarchies [1, 16]. Our work allows an enhanced treatment of different class attributes since they are assigned a similarity score rather than being treated simply as different features. For implementation issues, we referred to $WordNet$ taxonomy of concepts [52], however a more detailed lexical databases for the English language can be adopted as well.

As a final remark, and as also mentioned by the authors in [42], we emphasize that the evaluation of the semantic similarity of geospatial entity classes according to the information content approach is an interesting research topic to investigate for which there are no proposals in the literature. The contribution of this paper is just in this direction.

Our proposal can also be used to provide an approximate answer when the user expresses a query in terms of geographic classes that have no match or are missing in the database. For instance, if the user asks the list of hotels in the counties of U.S.A., and the concept of county is not provided in the database, then state, which is the concept at the coarse level in the hierarchy, is proposed. Furthermore, suppose the user erroneously asks the list of hotels in the counties of Italy. Our method can be used to detect the most similar concept to county within the Italian administrative subdivisions, that is municipality.

The paper is organized as follows. In the next section, we analyze different approaches to similarity measuring that have been proposed in the literature. In Section 3, a preliminary informal presentation of the main characteristics of the geographic classes addressed in this paper is given and, successively, the geographic data model is formally defined. In Section 4, we describe the information content and tuple similarity methods. They will be then integrated in our proposal. A running example is also given throughout the paper. Finally, in Section 5 the experiment we have performed in order to evaluate our method is presented. We selected three among the most representative proposals in the literature, and the experiment shows that $GSim$ has the highest correlation with human judgement. Section 6 concludes.

2. Related Work

In the domain of GISs, similarity is mainly discussed from the *spatial* point of view. This study originated from the increased availability of geographic data collected from satellite imagery and other remote sensors. Therefore, spatial similarity has been defined as matching and ranking according to a certain context (function, goal), scale (coarse o finer level), and technology (for searching, retrieving, and recognizing data). In particular, the research on spatial similarity ranges from *data retrieval* [33, 44], *problem solving* [53], *conflation* [7], *inter-operability* [21], etc.

Similarity of concepts and their relationships, that is the focus of our paper, has been largely discussed in the literature. Concerning the relationships, the following different types have been considered: *associative* (e.g., cause-effect), *equivalence* (or synonymy), *hierarchical* (e.g., *ISA* or hyperonym-hyponym, *Part_of*, etc.). The hierarchical relationship, particularly, has attracted a lot of interest in the research community since it has been considered very suitable for mapping the human cognitive view of classification (i.e., taxonomy).

The traditional approach to evaluate semantic similarity in a taxonomy is the so-called *edge-counting* approach [26, 36], which corresponds to the semantic distance approach. The semantic distance is equal to the minimal path length between concepts. The main limitations of this approach are represented by the underlying assumptions that, often, are not explicitly defined and, in many realistic cases, it is difficult to define a taxonomy where the distances between any two adjacent nodes are equivalent.

In order to overcome the drawback of the edge-counting approach, another methodology has been introduced, namely the *information content* (or *node-based* approach) [38, 39], which has been successively refined by Lin in [27]. This relies on the association of probabilities with the nodes of the taxonomy. The similarity between concepts is measured by the ratio between the amount of information shared by the concepts and the sum of the amounts of information of concepts. This approach is recalled in Section 4 and is the basis of the method for evaluating semantic similarity of geographic classes proposed in this paper. As also mentioned in [23], it shows a higher correlation with human judgement with respect to other existing proposals.

With regard to feature (or tuple) similarity, one of the most commonly used similarity measure is the *Dice*'s function [28, 37, 4]. It provides the coefficient of correlation between feature vectors, and it is given by the ratio between the number of features that are common to two vectors and the sum of the numbers of the features of each vector. Note that the Dice's function does not allow tuple similarity to be computed by explicitly considering the similarity degrees of components. Whereas, we address the similarity degrees of components through the notion of *information content similarity* (see Section 4, and Section 5).

Recently, few attempts have been performed to model geographic class similarity. For instance, in [41, 42] the authors define a computational method for assessing semantic similarity among geospatial entities that is based on the Tversky's model [50]. According to such a model, the similarity is a linear combination of weighted shared features and weighted distinct features. This method, called *Matching-Distance Similarity Measure* (MDSM), determines the similarity by using a matching process over synonym sets, semantic neighborhoods, and distinguishing features (such as parts, functions, and attributes). In particular, in [42], two different approaches, namely the *variability* and *commonality*, for determining features' relevance have been presented. Both these approaches will be used in our experiment presented in Section 5. It is important to emphasize that in [41, 42] the authors concentrated on the cognitive properties of the semantic similarity assessment, and they do not address the similarity of the geometric properties of geographic entities. The model of Tversky has also been used in the context of Ontologies [34] in [25] to determine semantic similarity of geographic categories (or entities). In particular, the problem of cross-mapping of geographic ontologies has been addressed and a systematic methodology for comparing categories has been presented.

To illustrate our proposal, we have selected MDSM, Dice, and Lin which are among the most representative methods defined in the literature (see Section 5).

Finally, we remark that our work is also a first attempt to integrate the geometric types of geographic classes in the similarity evaluation, which is omitted in all the proposals discussed above.

3. The Geographic Data Model

We consider a simple object-oriented geographic data model. This is characterized by the notion of *geographic class*, which is informally defined below. Such a notion will be formally defined in the next subsection.

A *geographic class* describes a set of geographic objects having the same set of *attributes* (or *properties*). It is specified by a *name* and a class *expression*. The class expression contains a *tuple* of typed attributes, and one or more *geometric* types from {*point*, *polyline*, *polygon*}. Each attribute is associated with an atomic type (e.g., *integer*, *string*, etc.). We assume a set of geometric types is associated with a geographic class because in multiple-representation databases, or depending on the scale, this type may change (e.g., municipality as a *polygon* or municipality as a *point*). For each geometric type a set of operations is defined. These operations, that can be either unary or binary, concern the topological relationships among objects. The definition and the formalization of these relationships, and their semantics, have been extensively discussed in [9, 10] and go beyond the scope of this paper.

A *geographic knowledge base* is essentially a set of geographic classes where each class is identified by a name, and there are no dangling class names, i.e., every name is associated with a class expression.

A geographic knowledge base can be organized according to a *partition* hierarchy. Such a hierarchy captures the well known *is-in* relationship or *inclusion* property. In the literature, basically, three different kinds of semantics for inclusion have been identified: *class*, *meronymic*, and *spatial* [45, 51]. Class inclusion is the standard subtype/supertype relationship which has been widely discussed in the database literature, and it is indicated by ISA [8, 49, 48]. Concerning meronymic (*Part_whole*, or *Part_of*) inclusion, many studies have been carried out. One of the various semantics of meronymic relationships discussed in [45] is the *place-area.* It concerns parts which are similar to whole, and they cannot be separated (for instance, the reception area is part of an office). Finally, the semantics of spatial inclusion differs from place-area in that it represents objects that are surrounded by others but they are not part of them (as for instance, car is-in city).

In this paper, we consider the place-area semantics for the inclusion relationship that is more suitable for capturing the meaning of inclusion in the geographic context and, in particular, of geographic classes organized as partition hierarchies [45]. In the rest of the paper, for the sake of simplicity, we will use the term *Part_of* hierarchies to mean this kind of hierarchies. Note that the place-area semantics gives us the possibility of applying the information content approach of Resnik, conceived for ISA hierarchies, to partition hierarchies.

Example 3.1. The following is a geographic class of name *Country*:

$$\begin{aligned}&Country = \langle [name : string,\\ &\quad countryCode : integer, president : string, flag : string], \{polygon\}\rangle\end{aligned}$$

having *name*, *countryCode*, *president*, and *flag* as tuple of typed attributes, and *polygon* as the singleton containing the geometric type of the class. In the next subsection we will see that such a class can be disjunctively partitioned according to different administrative organizations, e.g., the geographic classes *Region*, *State*, or *Department*. For instance, the countries Italy, U.S.A., and France are subdivided into regions, states, and departments, respectively.

3.1. Geographic Knowledge Base

In the following we assume that countable sets $\mathcal{N}, \mathcal{A}$ of class names and attribute names, respectively, are given. Let T and G be the sets of atomic and geometric types defined, respectively, as follows:
$T = \{string, integer, real, boolean\}$,
$G = \{point, polyline, polygon\}$.
The notion of *geographic class* is defined below.

Definition 3.1. [Geographic class] *A* geographic class *(*class *for short) is specified by a name and an expression as follows:*

$$n = < [a_1 : \tau_1, ..., a_h : \tau_h], G_n >, \qquad h \geq 1$$

where n is the name *of the class from $\mathcal{N}$, the a_j's are attribute names from $\mathcal{A}$, the τ_j's are types from T, and $G_n \subseteq G$ is the set of geometric types of the class n.*

Note that in the sequel, the term *concept* will be used to denote a class, an attribute, or a type name. Before introducing the notion of *geographic knowledge base*, the notion of *partial order on partitions* is given below.

Definition 3.2. [Partial order on partitions] *Let E be a subset of the plane, and P a partition of E, i.e., $\bigcup_P = E$, and $\forall$ p,p' $\in$ P, p$\cap$p' $= \emptyset$. Let $\mathcal{P}$(E) be a set of partitions of E and $\sqsubseteq$ be the following partial order on $\mathcal{P}$(E) (indicated as ($\mathcal{P}$(E),$\sqsubseteq$)):*

$P \sqsubseteq P'$ *iff* $\forall$ $p \in P$, $p' \in P'$, $p \cap p' \in \{p, \emptyset\}$.

Example 3.2. Let E be a subset of the plane called *Country*, and consider a set of partitions defined as follows:

$\mathcal{P}$(*Country*) = {*Region, State, Department, Province,*
County, Municipality}

and the partial order $\sqsubseteq$ on the above partitions:

$Municipality \sqsubseteq Province$
$Province \sqsubseteq Region$
$Region \sqsubseteq Country$
$County \sqsubseteq State$
$State \sqsubseteq Country$
$Department \sqsubseteq Country$,

as shown in Figure 1. As we mentioned before, *Country* is partitioned into three disjunctive partitions, called *Region*, *State*, and *Department*. The first two are partitioned again into *Province* and *County*, respectively, and are related to *Region*, and *State* by a partial order relationship. Similarly, *Province* into *Municiplaity*. In the following, the partial order on partition allows us to define the notion of *Geographic knowledge base*.

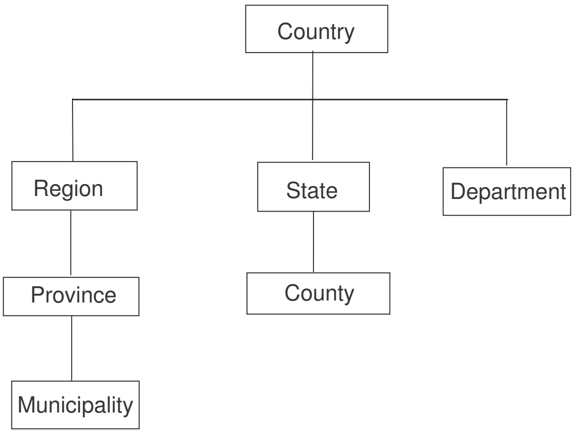

Figure 1: Partial order on the *Country* partitions

Definition 3.3. [Geographic knowledge base] *Let E be a subset of the plane and $(\mathcal{P}(E),\sqsubseteq)$ be a partial order on the set of partitions of E. A* geographic knowledge base $\Sigma_E = (C, A, Cls)$ *consists of finite sets $C \subset \mathcal{N}$, $A \subset \mathcal{A}$ and a finite set Cls of geographic classes such that, for each $n \in C$, n is the name of precisely one geographic class in Cls, and Cls also contains a class expression of name E. Furthermore, $C = \mathcal{P}(E)$. Then, $(\mathcal{P}(E),\sqsubseteq)$ is referred to as Part_of hierarchy of the geographic knowledge base Σ_E.*

Example 3.3. For instance, consider the set of partitions of *Country* of the previous example. A geographic knowledge base, called *GeoKB*, can be defined by the following set of geographic classes:

$$Country = \langle [name : string,\\ countryCode : integer, president : string, flag : string], \{polygon\}\rangle$$

$$Region = \langle [label : string, healthCenter : string,\\ president : string], \{polygon\}\rangle$$

$$State = \langle [tag : string, countryCode : integer,\\ governor : string], \{polygon\}\rangle$$

$$Department = \langle [name : string, area : integer,\\ wineZone : integer, dweller : integer], \{polygon\}\rangle$$

$$Province = \langle [name : string,\\ prefecture : string, localEducationOffice : string], \{polygon\}\rangle$$

$$Municipality = \langle [identity : string, head\_of\_council : string,\\ countryCode : integer, inhabitant : integer, area : integer], \{point, polygon\}\rangle$$

$County = \langle [countyID : integer, population : integer,$
$surface : integer], \{polygon\} \rangle$

Then, the hierarchy shown in Figure 1 is the $Part\_of$ *hierarchy of the* $GeoKB$ *example.*

Note that the partition hierarchies on a single geographic class, e.g. *Country*, are used to generate and visualize 2D multi-scale representation. Hence, *polygon* is the common geometric type used to model these classes. However, some geographic entities, like *Municipality*, can be represented by *point* at a small scale map representation. In this paper, we focus on 2D representation of geographic classes.

3.2. Weighted Part_of Hierarchy

Following the approach proposed in [38], below the notion of *weighted* $Part\_of$ *hierarchy of a geographic knowledge base* is introduced. It is essentially given by the $Part\_of$ hierarchy of the geographic knowledge base enriched with a function which associates *probabilities* with the concepts (class, attribute or type names) defined in the hierarchy.

Definition 3.4. [Weighted *Part_of* hierarchy of a geographic knowledge base] *Given a geographic knowledge base* Σ_E*, consider the* $Part\_of$ *hierarchy of* Σ_E *and a function* p *associating each concept (class, attribute, or type name)* c*, in the* $Part\_of$ *hierarchy of* Σ_E*, with a value in the interval [0,1] such that* $p(c)$ *is the probability that an instance belongs to the concept* c*. Furthermore, assume that the* $Part\_of$ *hierarchy has a unique* Top *node - the most abstract concept - such that for each* c*,* $partOf(Top, c)$ *holds, and* $p(Top) = 1$*. Then the* $Part\_of$ *hierarchy is referred to as* a weighted Part_of hierarchy of the geographic knowledge base Σ_E*, and it is indicated as* $\mathcal{H}_p^{\Sigma}$*.*

In the following, we refer to a hierarchy (taxonomy) of concepts represented by nouns in English. The *frequencies* of concepts are estimated using noun frequencies from large text corpora, as for instance the Brown Corpus of American English [17].
The *probability* $p(c)$ of the concept c is defined as follows:

$$p(c) = \frac{freq(c)}{M}$$

where $freq(c)$ is the frequency of the concept c, and M is the total number of observed instances of nouns in the corpus.

Example 3.4. In this paper, probabilities have been estimated according to $SemCor$ project [14], which labels subsections of Brown Corpus to senses in the WordNet lexicon [52]. According to $SemCor$, the total number of observed instances of nouns is $88,312$.

The definitions of the classes of our running example, and the related frequencies (the number in parenthesis) according to WordNet are given below:

(276) Country – the territory occupied by a nation;

(67) Region – the extended spatial location of something;

(272) State – the territory occupied by one of the constituent administrative districts of a nation.

(34) County – an area created by territorial division for the purpose of local government.

(34) County – the territory [1] *created by administrative division for the purpose of local government.*

(1) Department – the territorial and administrative division of some countries (such as France);

(4) Province – ...;

(1) Municipality – ... .

In Figure 2 the weighted $Part\_of$ hierarchy of the $GeoKB$ example is shown. Note that, in $WordNet$ the frequencies of the concepts *Municipality* and *Department* are not indicated, therefore, we assume they are equal to 1.

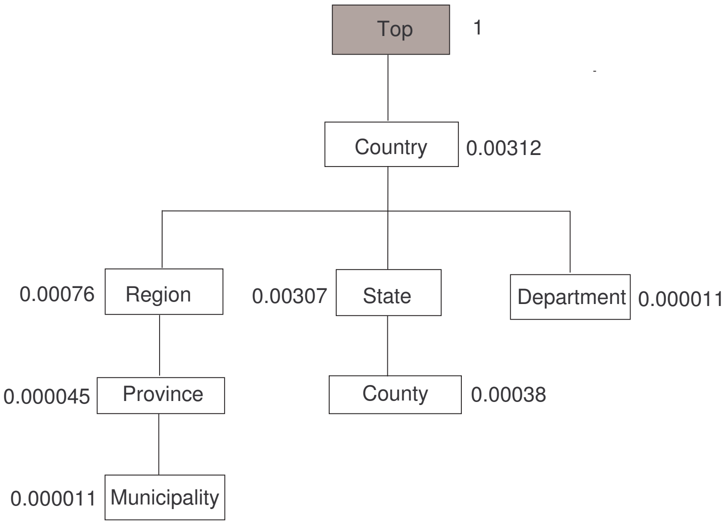

Figure 2: Weighted $Part\_of$ hierarchy of the $GeoKB$

Before concluding this section, we need to introduce the notion of *SynSet for a geographic knowledge base* Σ_E. It deals with synonyms, which have a fundamental importance in evaluating concept similarity since they are concepts differing in the names while having the same meaning.

Definition 3.5. **[Set of synonyms for a geographic knowledge base**] *Given a geographic knowledge base* $\Sigma_E = (C, A, Cls)$, *let* $B_i \subset \mathcal{N} \cup \mathcal{A}$, i = 1...n, *be sets of synonyms according to a lexical database for the English language. Then,* $SynSet_\Sigma$ *is a (possibly empty)* set of synonyms for the geographic knowledge base Σ_E *if:*

$SynSet_\Sigma = \{B_1, ..., B_n\}$, n ≥ 0,

and $B_i \cap$ *(C∪A)* $\neq \emptyset$, *for* i = 1...n.

In this paper, for homogeneity, we again refer to the set of synonyms provided by $WordNet$.

Example 3.5. The following set, namely $SynSet_{GeoKB}$, is a set of synonyms for the geographic knowledge base $GeoKB$, which has been defined according to WordNet (we focus on the following set for lack of space):

[1] Note that here the term *territory* has been used in place of *region* (as defined in WordNet) in order to avoid confusion.

$SynSet_{GeoKB} = \{$
$\{name, label, tag, identity, mark\},$
$\{surface, area\}$
$\{population, inhabitant, dweller, denizen, indweller\}$
$\}$

4. Class Similarity

In this section, the similarity of geographic classes ($GSim$) organized according to $Part\_of$ hierarchies is addressed. The proposed method is based on three notions, namely the *information content similarity* (ics), the *tuple similarity* ($tSim$), and the *geometric type similarity* (γSim). In particular, ics is devoted to compute the similarity between class names, attribute names, and atomic types, $tSim$ is conceived to measure the similarity of tuples of typed attributes, and γSim allows us to evaluate the similarity of sets of geometric types. Such notions are presented below, and in Subsection 4.3 they are combined to obtain the notion of $GSim$.

Note that, the formal properties of concept similarity depend on the selected data model, e.g., the *geometric model*, the *featural model*, the *network model*, etc. [32]. Among these formal properties, it is worth mentioning *symmetry* and *transitivity.* For instance, the featural model of Tversky [50] is neither dimensional nor metric and, consequently, symmetry and transitivity do not necessarily hold. Since our approach focuses on geographic classes organized as partition hierarchies, similarity is symmetric and transitive. This is in line with the network model and other frameworks for measuring similarity between and within ontologies (see for instance [13]).

4.1. Information Content Similarity

The notion of *information content similarity* allows us to compute the similarity of class, attribute, and atomic type names. It is based on the definition of *semantic similarity*, previously introduced by Resnik in [38], and successively refined by Lin in [27]. The starting assumption of the approach is that the *information content* of a concept (class, attribute, or type name) c is defined as - log $p(c)$, that is, as the probability of a concept increases, the informativeness decreases, therefore the more abstract a concept, the lower its information content [43]. According to the Resnik's approach, the similarity of hierarchically organized concepts, Sim_R, is given by the maximum information content shared by the concepts, that is, the more information two concepts share, the more similar they are. Intuitively, given two concepts c_1,c_2, $Sim_R(c_1, c_2)$ is defined by the information content of the concept that is the least upper bound (lub) of c_1,c_2 in the taxonomy, i.e., - log $p(lub(c_1, c_2))$. Successively, concept similarity has been defined according to Lin as the maximum information content shared by the concepts (Sim_R) divided by the information contents of the comparing concepts. This is formally defined by point 2 in the definition of *information content similarity* (ics) below. Furthermore, point 1 states that the ics of two concept names is equal to 1 if they coincide or are synonyms, and in all the other cases, their similarity is equal to zero (point 3).

Definition 4.1. [Information content similarity (ics)] *Given a geographic knowledge base* $\Sigma_E = (C, A, Cls)$, *a weighted* $Part\_of$ *hierarchy* $\mathcal{H}_p^{\Sigma}$ *and a* $SynSet_{\Sigma} = \{B_1, ..., B_n\}$, $n \geq 0$ *for* Σ_E, *consider the sets* T *of atomic types, as defined above. Given* n_1,$n_2 \in C \cup A \cup T$, *the* information content similarity *of* n_1,n_2, *indicated as* $ics(n_1, n_2)$, *is defined as follows:*

1. *if* $n_1 = n_2$ *or* n_1,$n_2 \in B_k \in SynSet_\Sigma$*, for some* k*,* $1 \leq k \leq n$*:*

$$ics(n_1, n_2) = 1$$

2. *if* n_1,$n_2 \in C \cup A$ *and* n_1,$n_2 \notin B_k \in SynSet_\Sigma$*, for any* k*:*

$$ics(n_1, n_2) = \frac{2 \log p(c')}{\log p(n_1) + \log p(n_2)}$$

where c' *is a concept providing the maximum information content shared by* n_1, n_2*, i.e.:*

$$-\log p(c') = \max_{c \in \mathcal{S}(n_1,n_2)} [-\log p(c)]$$

and $\mathcal{S}(n_1, n_2)$ *is the set of concepts that are upper bounds of both* n_1,n_2 *in the Part_of hierarchy;*

3. *otherwise:*

$$ics(n_1, n_2) = 0.$$

Example 4.1. In our running example consider the *Province* and *County* classes. The following holds:

$$\begin{aligned} ics(Province, County) &= \frac{2 \log p(Country)}{\log p(Province) + \log p(County)} \\ &= \frac{2 * 8.32180}{14.43032 + 11.34286} = 0.64577 \end{aligned}$$

Note that, as already mentioned, according to Lin's approach the maximum information content shared by the classes and the information contents of the classes to be compared are both considered. This is not true if we adopt Resnik's approach for which, for instance, the similarity (Sim_R) between *Province* and *County* and the similarity between *Municipality* and *County* coincide. In fact, both pairs share the same maximum information content, that is provided by *Country*, i.e.:

$Sim_R(Province, County) = Sim_R(Municipality, County) = -\log p(Country)$,

as it is easy to see in Figure 2.

4.2. Tuple Similarity

The tuple similarity (*tSim*) is inspired by a method based on the *maximum weighted matching* problem in bipartite graphs [19, 15]. Below, we start by presenting the notion of *candidate sets of pairs*. Essentially, given two class names n_1, n_2, it allows the identification of the sets of pairs of attributes, each pair formed by one attribute from n_1 and the other from n_2, which have to be considered in order to maximize the sum of the *ics* of the pairs. This method is in line with other feature-based approaches, as for instance, the Dice's function [4]. As better illustrated in Section 5, Dice's function does not allow tuple similarity to be computed by explicitly considering the similarity degrees of components whereas, in our proposal, they are addressed by integrating the notion of *ics* in the *tSim*.

Definition 4.2. **[Candidate sets of pairs]** *Let n_1,n_2 be the names of two classes of a geographic knowledge base defined as follows:*

$$n_1 = < [a_1 : \tau_1, ..., a_h : \tau_h], G_1 >, \qquad h \geq 1$$
$$n_2 = < [b_1 : \delta_1, ..., b_k : \delta_k], G_2 >, \qquad k \geq 1$$

where $h \leq k$ and:

$\mathcal{A}(n_1) = \bigcup_i a_i$ *and* $\mathcal{A}(n_2) = \bigcup_j b_j$.

Then, the set of candidate sets of pairs *of n_1, n_2, indicated as $\mathcal{P}(n_1, n_2)$, is defined by all possible sets of h pairs of attribute names, as follows:*

$$\mathcal{P}(n_1, n_2) = \{ \ \{\langle a_i, b_j \rangle_v\}_{v=1...h} \mid a_i \in \mathcal{A}(n_1),\ b_j \in \mathcal{A}(n_2),\ and\ \forall\ \langle a_s, b_r \rangle_q,$$
$$a_s \in \mathcal{A}(n_1),\ b_r \in \mathcal{A}(n_2),\ 1 \leq q \leq h,\ (i = s\ or\ j = r) \Rightarrow (a_i = a_s\ and$$
$$b_j = b_r)\}.$$

In other words, given two geographic classes, namely n_1,n_2, $\mathcal{P}(n_1, n_2)$ consists of all the sets of pairs of attributes (each pair formed by attributes not belonging to the same class), such that there are no two pairs of attributes sharing an element. For instance, assume that $\mathcal{A}(n_1)$ and $\mathcal{A}(n_2)$ represent a set of boys and a set of girls, respectively. A candidate set of pairs defines a possible set of marriages (when polygamy is not allowed) [19].

The notion of *tuple similarity* of two geographic classes n_1,n_2 is based on the set of candidate sets of pairs $\mathcal{P}(n_1, n_2)$, as defined below.

Definition 4.3. **[Tuple similarity (tSim)]** *Let n_1,n_2 be the names of two geographic classes as in the previous definition, such that $h \leq k$. Then the* tuple similarity *of n_1,n_2, indicated as $tSim(n_1, n_2)$, is defined as follows:*

$$tSim(n_1, n_2) = \frac{1}{k} \left(\max_{P \in \mathcal{P}(n_1,n_2)} \sum_{\langle a,b \rangle \in P} ics(a, b) * ics\,(typeOf(n_1, a), typeOf(n_2, b)) \right)$$

where ics is the information content similarity of a, b, standing for attributes of n_1, n_2, respectively, $\mathcal{P}(n_1, n_2)$ is the set of candidate sets of pairs of n_1, n_2, and $typeOf(n_1, a)$, $typeOf(n_2, b)$ are the types associated with a,b, in the classes n_1,n_2, respectively.

As mentioned in the Introduction, *tSim* has been inspired by the definition of attribute similarity given in *SymOntos*. The latter is based on the notion of *axiomatic similarity* (*as*) that is established according to a consensus system by a panel of experts in the application domain. In *tSim*, *as* has been replaced with *ics*. In addition, in *SymOntos* the similarity of types is not addressed. (Therefore, in place of the product of the *ics* above, in *SymOntos* we simply have $as(a, b)$.)

Example 4.2. Consider the geographic classes *Municipality* and *Department*. In this case, k=5 and h=4. One possible set of four pairs of attributes that maximizes the above formula is the following:

$$\{\langle identity, name\rangle, \langle head\_of\_council, wineZone\rangle, \langle inhabitant, dweller\rangle, \langle area, area\rangle\}$$

In fact, according to Definition 4.1, we have:

$ics(identity,name) = ics(inhabitant,dweller) = 1$

as defined in the $SynSet_{GeoKB}$ of Example 3.5, and, of course:

$ics(area,area) = 1,$

and the same holds for the related types, since $ics(string,string) = ics(integer, integer) = 1$, whereas:

$ics(head\_of\_council,wineZone) = 0.$

Therefore (recall that k=5):

$$tSim(Municipality,Department) = \frac{1}{5} * 3 = 0.60.$$

Note that, another possible set that maximizes the sum above contains the pair $\langle countryCode, wineZone\rangle$ in place of $\langle head\_of\_council, wineZone\rangle$ (in fact $ics(countryCode,\ wineZone) = 0$).

4.3. Geographic Class Similarity

Before introducing the notion of *geographic class similarity* (*GSim*), we need to present the notion of *geometric type similarity* (γSim). The latter allows us to measure the similarity of the sets of the geometric types of two geographic classes. We remark that within *Naive Geography* the similarity of geometric types is not addressed, and this is a first preliminary proposal to be extended in future work.

Definition 4.4. [Geometric type similarity (γSim)] *Consider two sets of geometric types, G_1 and G_2. The* geometric type similarity *of G_1,G_2, indicated as $\gamma Sim(G_1, G_2)$, is defined as follows:*

$$\gamma Sim(G_1, G_2) = \begin{cases} 1 & \textit{if } G_1 \cap G_2 \neq \emptyset \\ 0 & \textit{otherwise} \end{cases}$$

In our running example, consider the classes *County* and *Municipality*. Since the intersection of the sets of the geometric types of these classes (which are {*polygon*} and {*point*, *polygon*}, respectively) is non-empty, then $\gamma Sim(County,\ Municipality)$=1.

Now, the notion of *GSim* presented below can be defined on the basis of the information content similarity and the tuple similarity presented in the previous subsections, and the γSim above.

Definition 4.5. [Geographic class similarity (GSim)] *Let $\Sigma_E = (C, A, Cls)$ be a geographic knowledge base, consider a weighted $Part\_of$ hierarchy $\mathcal{H}_p^{\Sigma}$, a $SynSet_\Sigma = \{B_1, ..., B_n\}$, n ≥ 0, for Σ_E, and two geographic classes of Σ_E, of names n_1, n_2, respectively defined as follows:*

$n_1 = < [a_1 : \tau_1, ..., a_h : \tau_h], G_1 >, \quad h \geq 1$
$n_2 = < [b_1 : \delta_1, ..., b_k : \delta_k], G_2 >, \quad k \geq 1$

and suppose that $h \leq k$. Then, the geographic class similarity *of n_1,n_2, indicated as $GSim(n_1, n_2)$, is defined as follows:*

$$GSim(n_1, n_2) = \begin{cases} ics(n_1, n_2) * w_p + tSim(n_1, n_2) * w_a & \textit{if } \gamma Sim(G_1, G_2) = 1 \\ 0 & \textit{if } \gamma Sim(G_1, G_2) = 0 \end{cases}$$

where ics is the information content similarity of class names n_1, n_2, tSim is the tuple similarity, w_p and w_a are weights, such that $w_p + w_a = 1$, expressing the relevance to be given to the *partition hierarchy (ics), and attributes (tSim), respectively, and γSim is the geometric type similarity defined above.*

In the definition above, if w_a is equal to 0 then $GSim$ is given by the shared information content between n_1, and n_2, as derived from the partition hierarchy of the geographic classes. In contrast to the previous case, if w_p is equal to zero then $GSim$ corresponds to the tuple similarity. Note that, although the definition of $tSim$ is based on ics, the parameter w_p affects the ics of class names n_1, n_2, without influencing the ics of the related class attributes and types in $tSim$, and viceversa for w_a.

It is important to note that the evaluation of the tuning parameters is a complex problem which goes beyond the scope of this paper. In general, in the literature the relative definition is left to the domain expert, and one of the challenging topics in this field is to define (semi-)automatic criteria to evaluate such parameters rather than relying on human expertise [5, 54, 31, 46, 24]. For instance, in [42] a proposal for determining features' relevance is presented, where features stand for attributes, parts, or functions. Such a proposal is based on two different approaches, namely *variability* and *commonality.* This problem is related to that of defining w_p and w_a in our paper since it concerns the evaluation of the relevance of both the hierarchical (parts) and attribute components of a knowledge base within a given domain. For this reason, in this paper we determine w_p and w_a in line with both the variability and commonality approaches defined in [42]. In the following these approaches are briefly recalled, taking into account that the term "feature" used in [42] below stands for attributes or parts.

According to the variability approach, the relevance of a feature is related to the feature's informativeness, therefore the feature's relevance decreases if it is shared by most of geographic classes in the knowledge base. Viceversa, according to the commonality approach, high frequency of a feature corresponds to high relevance. Given a geographic knowledge base, assume t is a type of feature, i.e. t stands for attributes (a), or parts (p). Then, let P_t^v and P_t^c (where v stands for *variability* and c for *commonality*, respectively) be defined as follows:

$$P_t^v = 1 - \sum_{i=1}^{m} \frac{o_i}{nm}$$

$$P_t^c = \sum_{i=1}^{m} \frac{o_i}{nm} = 1 - P_t^v$$

where o_i is the number of occurrences of a feature in the geographic knowledge base, n is the number of geographic classes and m is the number of features in the geographic knowledge base. According to the variability approach, the weights are determined as follows:

$$w_p^v = \frac{P_p^v}{P_a^v + P_p^v}$$

$$w_a^v = \frac{P_a^v}{P_a^v + P_p^v}$$

whereas, according to the commonality approach:

$$w_p^c = \frac{P_p^c}{P_a^c + P_p^c}$$

$$w_a^c = \frac{P_a^c}{P_a^c + P_p^c}$$

By focusing on our running example, the number of geographic classes is $n = 7$, in the case of parts $m = 7$, and in the case of attributes $m = 13$. Therefore, the following holds:
$w_p^v = 0.46$ and $w_a^v = 0.54$
$w_p^c = 0.59$ and $w_a^c = 0.41$.

The semantic similarity measure proposed in this paper can be used to perform query approximation in GISs. For example, when querying a geographic database, the user can express a query in terms of geographic classes that have no match in the database. Therefore, approximate queries can be formulated by replacing the missing concepts with similar ones, which are then proposed to the user as acceptable.

Example 4.3. Suppose that the user wants to know "all the hotels of the counties in Italy", but he/she is not aware about the Italian administrative subdivisions. In order to provide an approximate answer, the database can be coupled with the partition hierarchy of Figure 2. Hence, three geographic classes *Region*, *Province* and *Municipality* are proposed to the user, with the related similarity scores with *County*. In the following, let $GSim^v$ and $GSim^c$ be the $GSim$ measure computed by assuming $w_p = w_p^v = 0.46$ and $w_a = w_a^v = 0.54$ (the variability approach) and $w_p = w_p^c = 0.59$ and $w_a = w_a^c = 0.41$ (the commonality approach). The following holds:
$GSim^v(Municipality, County) = (0.5993 * 0.46 + 0.40 * 0.54) * 1 = 0.4917$
$GSim^c(Municipality, County) = (0.5993 * 0.59 + 0.40 * 0.41) * 1 = 0.5176$
where:
$ics(Municipality, County) = \dfrac{2 * 8.32180}{11.34286 + 16.43032} = 0.5993$
$tSim(Municipality, County) = \dfrac{1}{5} * 2 = 0.40$
and, as shown previously:

$\gamma Sim(Municipality, County)$=1.

Consider now the *Province* and *County* classes. In this case:
$GSim^v(Province, County) = (0.6458 * 0.46 + 0 * 0.54) * 1 = 0.2971$
$GSim^c(Province, County) = (0.6458 * 0.59 + 0 * 0.41) * 1 = 0.3810$

where:
$ics(Province, County) = \dfrac{2 * 8.32180}{11.34286 + 14.43032} = 0.6458$
$tSim(Province, County) = \dfrac{1}{3} * 0 = 0$
$\gamma Sim(Municipality, County)$=1.

If we consider the *Region* and *County* classes, then we have:
$GSim^v(Region, County) = (0.7667 * 0.46 + 0 * 0.54) * 1 = 0.3527$
$GSim^c(Region, County) = (0.7667 * 0.59 + 0 * 0.41) * 1 = 0.4524$
where:

$ics(Region, County) = \frac{2 * 8.32180}{11.34286 + 10.36423} = 0.7667$
$tSim(Region, County) = \frac{1}{3} * 0 = 0$
$\gamma Sim(Municipality, County)$=1.

As we can see, according to both the variability and the commonality approaches, within the Italian administrative subdivisions the most similar concept to $County$ is $Municipality$. However, the user can decide to obtain the list of hotels from the Italian regions or provinces if, according to his/her opinion, they are more appropriate.

5. Experimental Results

In this section, we present the experiment we have performed to evaluate our proposal. We have compared the similarity values obtained with $GSim$ against that obtained according to three representative methods selected in the literature: (i) the proposal of Lin [27], (ii) the Dice's function [4], and (iii) the Matching-Distance Similarity Measure (MDSM) [42]. Before illustrating our experiment, the Dice's function and the MDSM's method are briefly recalled below.

5.1. Dice and MDSM methods

According to Dice's function [4], given two concepts, say c_1, and c_2, each described by a set of features (or attributes), say $F(c_1)$, $F(c_2)$, respectively, their similarity ($Dice(c_1, c_2)$) is defined as follows:

$$Dice(c_1, c_2) = \frac{2 \mid A(c_1, c_2) \mid}{\mid F(c_1) \mid + \mid F(c_2) \mid}$$

where:

$$A(c_1, c_2) = \{(a, b) \mid a \in F(c_1), b \in F(c_2), a, b \in D \subseteq Aff\}$$

Aff is the set of sets of attributes showing affinity that is computed similar to the maximum weighted matching problem in bipartite graphs, and, for any set S, $|S|$ is the cardinality of S. For instance, consider $Municipality$ and $State$. The number of pairs of attributes showing affinity are (*identity*,*tag*), and (*countryCode*,*countryCode*). Therefore:

$$Dice(Municipality, State) = \frac{2 * 2}{5 + 3} = 0.50$$

Regarding MDSM, the similarity measure is a weighted sum of the similarity measures for $parts$, $functions$ and $attributes$, that are the distinguishing features of spatial entity classes in [42]. Given two classes a, b, their similarity according to MDSM, indicated as $MDSM(a, b)$, is defined as follows:

$$MDSM(a, b) = w_p * S_p(a, b) + w_f * S_f(a, b) + w_a * S_a(a, b)$$

where $S_t(a, b)$ - t standing for parts (p), functions (f), and attributes (a) - is defined below, and w_p, w_f, and w_a are weights defining the relative importance of parts, functions, and attributes, such that $w_p + w_f + w_a$=1, which are defined according to the variability and commonality approaches. The definitions of w_p and w_a have been recalled in Subsection 4.3, and w_f can be defined analogously. Let A, B be two description sets for the classes a and b, respectively. Then, the following holds:

$$S_t(a,b) = \frac{\mid A \cap B \mid}{\mid A \cap B \mid + \alpha(a,b) \mid A - B \mid + (1 - \alpha(a,b)) \mid B - A \mid} \tag{1}$$

where $A \cap B$ and $A - B$ are the set-theory intersection and difference of the sets A, B, respectively, and $0 \leq \alpha(a,b) \leq 1$ is a function defining the "relative importance" of the non-common characteristics of the classes. Such a function is essentially defined in terms of the distance among the classes a, b and the class representing their least upper bound in the hierarchy, where the distance is given by the number of arcs along the shortest path between classes (for details, see [42]).

In our experimental results, since functions are not present, we focus on parts and attributes. For instance, in the case of *Municipality* and *State*, $\alpha(Municipality, State)$=0.25 because their least upper bound is *Country*. Furthermore, assume that A and B are the sets of attributes of *Municipality* and *State*, respectively. Then:

$$\mid A \cap B \mid = 2$$

due to the presence of the pairs of attributes (*identity*,*tag*) (that, similar to the case of Dice's function, are synonyms), and (*countryCode*,*countryCode*). Furthermore, the set differences are respectively:

$$\begin{array}{ll} A - B = & \{head\_of\_council, inhabitant, area\} \\ B - A = & \{governor\} \end{array}$$

therefore:

$$S_a(Municipality, State) = \frac{\mid A \cap B \mid}{\mid A \cap B \mid + 0.25 \mid A - B \mid + 0.75 \mid B - A \mid} = 0.5714$$

For parts we have:

$$S_p(Municipality, State) = 0$$

By indicating with MDSM^v and MDSM^c the MDSM measure calculated according to the variability and commonality approaches respectively, since $w_p^v = 0.46$, $w_a^v = 0.54$, and $w_p^c = 0.59$, $w_a^c = 0.41$ (see Subsection 4.3), the following holds:

$$MDSM^v(Municipality, State) = 0.3086$$

$$MDSM^c(Municipality, State) = 0.2343.$$

5.2. The Experiment

In our experiment we computed the similarity values for all pairs of geographic classes of our running example according to Lin, Dice, *tSim*, MDSM and *GSim* (for the last two by following both the variability and commonality approaches defined in [42]). In order to evaluate the

Table 1: Comparison among some representative similarity measures of geographic classes with $GSim^v$ ($w_p = w_p^v = 0.46$) and $GSim^c$ ($w_p = w_p^c = 0.59$)

Pairs of classes	*HJ*	*Lin*	*Dice*	*$MDSM^v$*	*$MDSM^c$*	*tSim*	*$GSim^v$*	*$GSim^c$*
(Municipality,Province)	**0.9200**	**0.9352**	0.2500	0.1800	0.1367	0.2000	0.5382	**0.6338**
(Municipality,Region)	0.7500	0.7736	0.2500	0.1800	0.1367	0.2000	0.4639	0.5384
(Municipality,Country)	0.7300	0.6724	0.4444	0.2700	0.2050	0.4000	0.5253	0.5607
(Municipality,State)	0.6500	0.6718	0.5000	0.3086	0.2343	0.4000	0.5250	0.5604
(Municipality,County)	0.6300	0.5993	0.5000	0.2842	0.2158	0.4000	0.4917	0.5176
(Municipality,Department)	0.7100	0.5065	**0.6667**	**0.3812**	**0.2894**	**0.6000**	**0.5570**	0.5448
(Province,Municipality)	**0.9100**	**0.9352**	0.2500	**0.5680**	**0.6720**	0.2000	0.5382	**0.6338**
(Province,Region)	0.8900	0.8360	**0.3333**	0.1800	0.1367	**0.3333**	**0.5646**	0.6299
(Province,Country)	0.7600	0.7315	0.2857	0.1350	0.1025	0.2500	0.4715	0.5341
(Province,State)	0.6900	0.7308	**0.3333**	0.1800	0.1367	**0.3333**	0.5162	0.5679
(Province,County)	0.6000	0.6458	0.0000	0.0000	0.0000	0.0000	0.2971	0.3810
(Province,Department)	0.5700	0.5393	0.2857	0.1473	0.1118	0.2500	0.3831	0.4207
(Region,Municipality)	0.7100	0.7736	0.2500	0.5680	0.6720	0.2000	0.4639	0.5384
(Region,Province)	0.8500	0.8360	0.3333	**0.6400**	**0.7267**	0.3333	0.5646	0.6299
(Region,Country)	**0.9200**	**0.8907**	**0.5714**	0.2700	0.2050	**0.5000**	**0.6797**	**0.7305**
(Region,State)	0.8200	0.8897	0.3333	0.1800	0.1367	0.3333	0.5893	0.6616
(Region,County)	0.6600	0.7667	0.0000	0.0000	0.0000	0.0000	0.3527	0.4524
(Region,Department)	0.7100	0.6212	0.2857	0.1543	0.1171	0.2500	0.4207	0.4690
(Country,Municipality)	0.7200	0.6724	0.4444	0.6760	0.7540	0.4000	0.5253	0.5607
(Country,Province)	0.6900	0.7315	0.2857	0.6400	0.7267	0.2500	0.4715	0.5341
(Country,Region)	**0.9200**	0.8907	**0.5714**	**0.8200**	**0.8633**	**0.5000**	0.6797	0.7305
(Country,State)	0.8700	**0.9987**	**0.5714**	**0.8200**	**0.8633**	**0.5000**	**0.7294**	**0.7943**
(Country,County)	0.6300	0.8464	0.0000	0.4600	0.5900	0.0000	0.3893	0.4994
(Country,Department)	0.7900	0.6724	0.2500	0.5950	0.6925	0.2500	0.4443	0.4992
(State,Municipality)	0.6500	0.6718	0.5000	0.2400	0.1822	0.4000	0.5250	0.5604
(State,Province)	0.6900	0.7308	0.3333	0.1800	0.1367	0.3333	0.5162	0.5679
(State,Region)	0.8200	0.8897	0.3333	0.1800	0.1367	0.3333	0.5893	0.6616
(State,Country)	**0.9200**	**0.9987**	**0.5714**	0.2700	0.2050	**0.5000**	**0.7294**	**0.7943**
(State,County)	0.7300	0.8476	0.0000	**0.4600**	**0.5900**	0.0000	0.3899	0.5001
(State,Department)	0.7300	0.6718	0.2857	0.1543	0.1171	0.2500	0.4440	0.4989
(County,Municipality)	0.6300	0.5993	0.5000	0.2571	0.1952	0.4000	0.4917	0.5176
(County,Province)	0.6000	0.6458	0.0000	0.0000	0.0000	0.0000	0.2971	0.3810
(County,Region)	0.6600	0.7667	0.0000	0.0000	0.0000	0.0000	0.3527	0.4524
(County,Country)	0.6700	0.8464	0.0000	0.0000	0.0000	0.0000	0.3893	0.4994
(County,State)	**0.7600**	**0.8476**	0.0000	0.0000	0.0000	0.0000	0.3899	0.5001
(County,Department)	0.7200	0.5993	**0.5714**	**0.2945**	**0.2236**	**0.5000**	**0.5457**	**0.5586**
(Department,Municipality)	0.7100	0.5065	**0.6667**	**0.3411**	**0.2589**	**0.6000**	**0.5570**	0.5448
(Department,Province)	0.5700	0.5393	0.2857	0.1620	0.1230	0.2500	0.3831	0.4207
(Department,Region)	0.7100	0.6212	0.2857	0.1543	0.1171	0.2500	0.4207	0.4690
(Department,Country)	**0.8200**	**0.6724**	0.2500	0.1350	0.1025	0.2500	0.4443	0.4992
(Department,State)	0.7300	0.6718	0.2857	0.1543	0.1171	0.2500	0.4440	0.4989
(Department,County)	0.7200	0.5993	0.5714	0.3240	0.2460	0.5000	0.5457	**0.5586**
Correlation	**1.0000**	**0.7127**	**0.2972**	**0.3774**	**0.3419**	**0.3318**	**0.7233**	**0.8224**

various proposals, we had to address the problem related to the so-called "ideal" (or "right") values. In general, ideal values are established according to human judgement and, in the literature, often the similarity scores assigned by human subjects in the Miller&Charles experiments are addressed [30]. It is important to observe that our proposal originates by relying on the experimental results already defined in [27], and [23]. In particular, in the mentioned papers it has already been shown that Lin's approach shows a higher correlation with human judgement, and therefore provides better results, than other "hierarchy-centered" methods, including the traditional edge-counting approach [36].

In the experiment we performed, ideal values have been established by asking 24 students to asses the similarity among all pairs of geographic classes of our running example. Students were asked to assign a similarity score, on a scale of 0 to 1, to each of the 42 pairs of classes that can be obtained from the $GeoKB$ of Example 3.3. For each pair, the average of these values is shown in the second column of Table 1 (Human Judgement - HJ). Besides the similarity scores obtained according to Lin, Dice, and *tSim*, Table 1 also shows MDSM^v, MDSM^c, and *GSim*^{v}, *GSim*^{c} values by setting $w_p^v = 0.46$ ($w_a^v = 0.54$), and $w_p^c = 0.59$ ($w_a^c = 0.41$), as determined in the previous section.

For each similarity method, the highest similarity scores among a class and the remaining six classes in the geographic knowledge base are shown in bold. Note that similarity in MDSM is asymmetric, see for instance the values of MDSM^v for the pairs (*Municipality*,*Province*) (0.18) and (*Province*,*Municipality*) (0.57). The same also holds for human judgement, see for instance the slightly difference between the values of (*Municipality*,*Province*) (0.92) and (*Province*,*Municipality*) (0.91), that is more relevant between, for instance, (*Country*,*State*) (0.87) and (*State*,*Country*) (0.92). Looking at the matches of highest scores with human judgement, Lin provides the best result (6 correct matches), against *GSim*^{c} (4 correct matches), Dice and *tSim* (3 correct matches), and MDSM^v, MDSM^c, and *GSim*^{v} (2 correct matches). However, as shown in the last row of Table 1, the highest correlation with human answers is provided by *GSim*^{c} (0.82). In the case of *GSim*^{v}, we have a slightly higher correlation with human judgement (0.72) with respect to Lin (0.71).

Note that, in most cases, the values obtained according to *GSim* are greater than the values computed according to Dice and MDSM, and they are less than the measures calculated according to the information content approach by Lin. The reason is that, on one hand, our method captures the informativeness of geographic classes organized as a hierarchy, while Dice and MDSM are mainly based on the common and non-common characteristics of classes. For instance, in some cases the values obtained by Dice and MDSM are equal to zero. This indicates that there are no common features between the considered pairs of classes. On the other hand, similar to Dice and MDSM, our approach considers the structures of classes, and the heterogeneity of the attributes, in some cases, significantly impacts on the similarity values that are considerably less than the ones obtained according to Lin.

Table 1 shows how the similarity scores significantly differ by following the hierarchy-centered approach of Lin or the feature-centered approaches of Dice or MDSM. The aim of *GSim* is just to reduce the gap between the hierarchy-centered and the feature-centered methods proposed in the literature. In fact it has been conceived for knowledge bases where concepts are hierarchically organized and, at the same time, are associated with sets of attributes. For instance, consider the pairs of classes (*Municipality*,*Region*) and (*Municipality*,*County*) and analyze their values according to Table 1. Since *Region* is partitioned (although indirectly) into *Municipality* (see Figure 1), their similarity according to the hierarchy-centered approach of Lin is greater than the similarity between *Municipality* and *County*, i.e., 0.77 and 0.60, respectively. On the other hand, *Municipality* share with *County* one more attribute with respect to *Region* (taking into account also synonyms) then, the similarity of (*Municipality*,*County*) and (*Municipality*,*Region*) according to the feature-centered approach of Dice is 0.50 vs 0.25, respectively. Similar results are obtained according to MDSM^v, i.e., 0.28 vs 0.18, and MDSM^c, i.e., 0.22 vs 0.14. Accordingly, *tSim* of (*Municipality*,*County*) is greater than *tSim* of the pair (*Municipality*,*Region*) (0.40 vs 0.20), whereas in the case of $GSim^v$ and $GSim^c$ we have closer results, i.e., 0.49 vs 0.46, and 0.52 vs 0.54, respectively.

6. Conclusion and Future Work

In this paper, we focused our attention on the semantic similarity of geographic classes organized as *Part_of* hierarchies. We have proposed a method for similarity measuring which takes into account both the concept similarity within the *Part_of* hierarchy (through the information content approach) and the tuple similarity (through the sets of typed attributes). The proposed method takes advantage of the information-theoretic notion of similarity which overcomes the drawbacks of the traditional edge-counting approach and, accordingly, provides more reliable measures for comparing geographic classes. The information content approach relies on taxonomies of concepts that can be acquired from *WordNet* or more detailed lexical databases for the English language.

A direction for future research is to extend the proposed method to the dynamic component of the geographic data model, i.e., to the set of operations that can be performed on geographic data.

References

[1] C. Beeri (1990) A formal approach to object-oriented databases, Data & Knowledge Engineering 5, North-Holland 353-382.

[2] M. W. Berry, T. A. Letsche (1997) Large-Scale Information Retrieval with Latent Semantic Indexing, Information Sciences 100(1-4) 105-137.

[3] T. Bruns, M. J. Egenhofer (1996) Similarity of Spatial Scenes, Seventh International Symposium on Spatial Data Handling, Delft, The Netherlands, M.-J. Kraak and M. Molenaar (eds.) 31-42.

[4] S.Castano, V.De Antonellis, M.G.Fugini, B.Pernici (1998) Conceptual Schema Analysis: Techniques and Applications, ACM Trans. on Database Systems 23(3) 286-332.

[5] S. Castano, A. Ferrara, G.N. Hess (2006) Discovery-Driven Ontology Evolution, Semantic Web Applications and Perspectives (SWAP), 18-20 December, Pisa, Italy.

[6] T. Chiang, T. Tsai (2008) Querying color images using user-specified wavelet features, Knowledge and Information Systems 15(1) 109-129.

[7] M. Cobb, M. Chung, H. Foley, F. Petry, K. Shaw (1998) A Rule-Based Approach for the Conflation of Attributed Vector Data, GeoInformatica: An International Journal on Advances of Computer Science for Geographical Information Systems 2(1), 7-37.

[8] E. F. Codd (1979) Extending the database relational model to capture more meaning, ACM Transactions on Database Systems 4(4), 397-434.

[9] M. J. Egenhofer, R. Franzosa (1991) Point-Set Topological Spatial Relations, International Journal of Geographical Information Systems 5(2), 161-174.

[10] M. J. Egenhofer, R. Franzosa (1995) On the Equivalence of Topological Relations, International Journal of Geographical Information Systems 9(2), 133-152.

[11] M. J. Egenhofer, D. M. Mark (1995) Naive Geography, Lecture Notes in Computer Science 988, Springer, 1-15.

[12] M. J. Egenhofer, A. Rashid, B. M. Shariff (1998) Metric Details for Natural-Language Spatial Relations. ACM Transanction on Information Systems 16(4), 295-321.

[13] M. Ehrig, P. Haase, M. Hefke, N. Stojanovic (2005) Similarity for Ontologies - A Comprehensive Framework, in: Proc. of European Conference on Information Systems (ECIS), Regensburg, Germany.

[14] C. Fellbaum (1998) A Semantic Network of English: the Mother of all WordNets, Computers and the Humanities 32, 209-220.

[15] A. Formica, M. Missikoff (2002) Concept Similarity in SymOntos: an Enterprise Ontology Management Tool, Computer Journal 45(6), 583-594.

[16] A. Formica, M. Missikoff (2004) Inheritance processing and conflicts in structural generalization hierarchies, ACM Computing Surveys 36(3), 263-290.

[17] W. N. Francis, H. Kucera (1982) Frequency Analysis of English Usage, Houghton Mifflin, Boston.

[18] A. U. Frank, G. S. Volta, M. McGranaghan (1997) Formalization of Families of Categorical Coverages, International Journal of Geographical Information Science 11(3), 215-231.

[19] Z. Galil (1986) Efficient algorithms for finding maximum matching in graphs, ACM Computing Surveys 18(1), 23-38.

[20] R. K. Goyal, M. J. Egenhofer (2001) Similarity of Cardinal Directions, in: Proc. of Seventh International Symposium on Spatial and Temporal Databases (SSTD), Los Angeles, 36-58.

[21] F. Harvey (1999) Designing for interoperability: Overcoming semantic differences, in: M. F. Goodchild, M. J. Egenhofer, R., Fegeas, C. A. Kottman, (Eds.), Interoperating Geographic Information Systems, Boston, Kluwer Academic Pub, 58-98.

[22] C.C. Hsu, W.W. Chu, R.K. Taira (1996) A Knowledge-Based Approach for Retrieving Images by Content, IEEE Transaction on Knowledge and Data Engineering, 8(4), 522-532.

[23] J. J. Jiang, D. W. Conrath (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, in: Proc. of the 10th International Conference on Research in Computational Linguistics (ROCLING), Taiwan, 1-15.

[24] I. Jurisica, J. Mylopoulos, E. Yu (2004) Ontologies for Knowledge Management: An Information Systems Perspective, Knowledge and Information Systems, 6(4), 380-401.

[25] M. Kavouras, M. Kokla, E. Tomai (2005) Comparing Categories among Geographic Ontologies, Computers & Geosciences, special issue, 31(2), 145-154.

[26] J. H. Lee, M. H. Kim, Y. J. Lee (1993) Information Retrieval Based on Conceptual Distance in IS-A Hierarchies, Journal of Documentation 49(2), 188-207.

[27] D. Lin (1998) An Information-Theoretic Definition of Similarity, in: J. W. Shavlik (Ed.), Proc. of 15th the International Conference on Machine Learning, Madison, Wisconsin, USA, Morgan Kaufmann, 296-304.

[28] Y. S. Maarek, D. M. Berry, G. E. Kaiser (1991) An Information Retrieval Approach For Automatically Constructing Software Libraries, IEEE Transactions on Software Engineering 17(8), 800-813.

[29] J. McIntosh, M. Yuan (2005) Assessing Similarity of Geographic Processes and Events, Transactions in GIS 9(2), 223-245.

[30] G. A. Miller, W.G. Charles (1991) Contextual Correlates of Semantic Similarity, Language and Cognitive Processes, 6(1), 1-28.

[31] R. Navigli, P. Velardi (2004) Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites, Computational Linguistics, 30(2), 151-179.

[32] K. A. Nedas, M. J. Egenhofer (2003) Spatial Similarity Queries with Logical Operators, in Int. Symposium on Spatial and Temporal Databases (SSTD), LNCS 2750, 430-448.

[33] D. Papadias, V. Delis (1997) Relation-based similarity, in: Proc. of the 5th ACM Int. workshop on Advances in Geographical Information Systems, Las Vegas, Nevada, ACM Press, 1-4.

[34] H.S. Pinto, J.P. Martins (2004) Ontologies: How can They be Built?, Knowledge and Information Systems, 6(4), 441-464.

[35] E. Pourabbas (2003) Cooperation with Geographic Databases, in: M. Rafanelli (Ed.), Multidimensional Databases: Problems and Solutions Idea Group Inc., IGP/INFOSCI/IRM Press, Hershey, PA - USA, 393-432.

[36] L. Rada, H. Mili, E. Bicknell, M. Bletter (1989) Development and Application of a Metric on Semantic Nets, IEEE Transactions on Systems Man and Cybernetics 19(1), 17-30.

[37] E. Rasmussen (1992) Clustering Algorithms, in: W. B. Frakes, R. Baeza-Yates (Eds.), Information Retrieval: Data Structures & Algorithms, Prentice Hall, 419-442.

[38] P. Resnik (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in: Proc. of the Int. Joint Conference on Artificial Intelligence (IJCAI), Montral, Qubec, Canada, August 20-25, Morgan Kaufmann, 448-453.

[39] P. Resnik (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, J. of Artificial Intelligence Reserach (JAIR) 11, 95-130.

[40] P. Rigaux, M. Scholl (1994) Multiple Representation Modelling and Querying, in: J. Nievergelt, T. Roos, H. J. Schek, P. Widmayer (Eds.), Lecture Notes in Computer Science 884, Springer Verlag, Berlin, 59-69.

[41] A. Rodriguez, M. Egenhofer (2003) Determining Semantic Similarity Among Entity Classes from Different Ontologies, IEEE Transactions on Knowledge and Data Engineering 15(2), 442-456.

[42] A. Rodriguez, M. Egenhofer (2004) Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure, International Journal of Geographical Information Science 18(3), 229-256.

[43] S. Ross (1976) A First Course in Probability, Macmillan.

[44] H. Samet (2004) Indexing Issues in Supporting Similarity Searching, in: Proc. of the Pacific Rim Conference on Multimedia, Tokyo, Japan, Lecture Notes in Computer Science 3332, Springer, 463-470.

[45] V. C. Storey (1993) Understanding Semantic Relationships, VLDB Journal 2(4), 455-488.

[46] H. Tan, P. Lambrix (2007) A method for recommending ontology alignment strategies, The 2nd International Semantic Web Conference - The 2nd Asian Semantic Web Conference, Busan, Korea, Nov. 11-15.

[47] A. Tombros, C.J. van Rijsbergen (2004) Query-sensitive similarity measures for information retrieval. Knowledge and Information Systems 6(5), 617-642.

[48] T. L. Teorey, D. Yang, J. P. Fry (1986) A Logical design methodology for relational databases using the extended entity-relational model, ACM Computing Surveys 18(2), 197-222.

[49] D. Tsichritzis, F. Lochovsky (1982) Data Models, New York, Prentice-Hall.

[50] A. Tversky (1977) Features of Similarity, Psychological Review 84(4), 327-352.

[51] M. E. Winston, R. Chaffin, D. Hermann (1987) A taxonomy of part-whole relations, Cognitive Science 11, 417-444.

[52] WordNet 2.0: A lexical database for the english language: http://www.cogsci.princeton.edu/cgi-bin/webwn.

[53] A. G. O. Yeh, X. Shi (2001) The Application of Case-based Reasoning in Development Control, in: S. Geertman, J. Stillwell (Eds.), Planning Support Systems in Practice, Berlin, Springer-Verlag, 223-248.

[54] K. Zhang, J. Tang, M. Hong, J. Li, W. Wei (2006) Weighted Ontology-Based Search Exploiting Semantic Similarity, Frontiers of WWW Research and Development, APWeb 2006, X. Zhou et al. (Eds.), LNCS 3841, pp.498-510.

[55] M. Zhou, M. Wong, K. Chu (2006) A geometrical solution to time series searching invariant to shifting and scaling, Knowledge and Information Systems 9(2), 202-229.