Towards Ontological Similarity for Spatial Hierarchies

Raimundo F. Dos Santos
Virginia Tech
7054 Haycock Rd
Falls Church, VA 22043 USA
rdossant@vt.edu

Arnold P. Boedihardjo
US Army Corps of Engineers
Topographic Engineering Center
7701 Telegraph Rd
Alexandria, VA 22315 USA
arnold.p.boedihardjo@usace.army.mil

Chang-Tien Lu
Virginia Tech
7054 Haycock Rd
Falls Church, VA 22043 USA
ctlu@vt.edu

ABSTRACT

Ontological structures provide a rich hierarchy of concepts and relationships that are helpful in exploratory analysis. Ontologies, however, are often categorical, which introduces ambiguity, and makes numerical analysis difficult. Adding to the problem is the fact that as the number of ontological concepts increases so does computational complexity for a variety of analytical tasks. In this paper, we propose both spatial and ontological co-occurrence as a means to derive similarity among categorical values. More specifically, we devise a method that combines entity location as well as categorical frequency into a numerical measure of similarity for any pair of categorical values. In addition, we show how different ontological levels can hide or uncover information content while influencing the number of processed categorical values. We provide experiments that demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks—*data mining, spatial information retrieval*

General Terms

Algorithms, Theory

Keywords

Categorical data, ontologies, spatial hierarchies

1. INTRODUCTION

An ontology is commonly defined as a graph structure that models connectivity among concepts and relationships in a real-world domain [13]. One of its strengths is knowledge sharing, in which field experts agree a priori on standard concepts and linkage among them. The *International Classification of Diseases* (*ICD*), is such an example [2]. Provided by the *World Health Organization* (WHO), *ICD* is a hierarchical structure that assigns codes to diseases and differentiates them at nested levels. In Figure 1, Codes [I00 - I99.9], for instance, cover *Diseases of the circulatory system*, which are further reclassified into subcategories: *Acute rheumatic fever (I00-I02.9)*, *Cerebrovascular diseases*(I60-I69.99), and others.

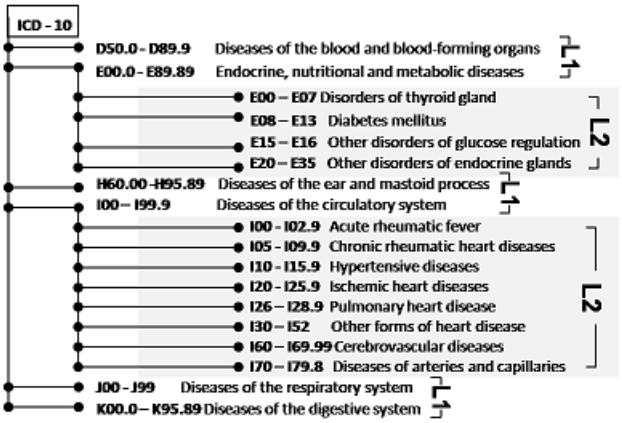

Figure 1: International Classification of Diseases (ICD) - partial view.

Ontologies are designed such that, at shallow levels, concepts are general in purpose. As traversal moves to deeper levels, concepts become increasingly specific. Thus in Figure 1, while Level 1 (L1) has a general node for *Diseases of the circulatory system*, Level 2 (L2) decomposes it into several possibilities, such as *Ischemic heart diseases*. Other levels exist, though not shown.

Ontological concepts are quite often categorical or nominal by nature, making it difficult to establish a similarity measure. Take for instance Figure 2, which depicts a few occurrences of heart disease in the region around Washington, D.C. In between *Reston* and *Tyson's Corner*, there is one occurrence of *hypertensive* ($\otimes$) heart disease along with one occurrence of *ischemic* ($\triangle$) heart disease. Spatially speaking, their distance is just a few miles apart. Ontologically, however, it is neither apparent nor intuitive whether *hypertensive* is closer to *ischemic* than to other diseases, such as *pulmonary* ($\square$), or vice-versa. The hierarchy of Figure 1 shows that these elements reside in the same level (L2), but their positioning relative to one another is at best arbitrary. Without a proper numerical similarity, the ontological space becomes challenging to reason over, which is a requirement

ACM SIGSPATIAL QUeST'12, Nov. 6, 2012, Redondo Beach, CA, USA

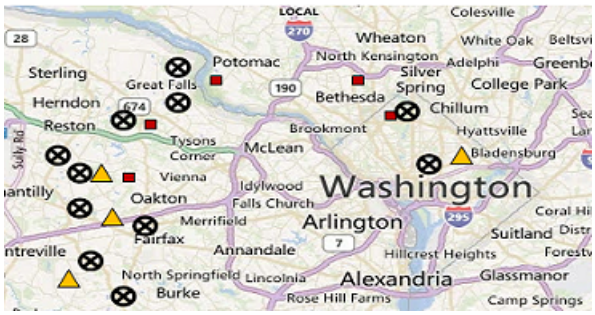

Figure 2: Hypothetical heart disease occurrences in Northern Virginia.

in many analytical techniques. Certain classification approaches rely on numerical similarity to work properly [15], while other data transformation techniques strictly require numerical values, as is the case in *PCA*-based multidimensional reduction [10] and certain clustering techniques [3]. Determining similarity as a condition of identity and relatedness among spatial entities is normally approached from a *Euclidean*[1] perspective. Distance and other quantitative measures (e.g., a person's age or water temperature) are ideal as comparative norms since they provide an inherent basis for differentiation and ordering.

When combined with spatial properties, a categorical similarity measure may be devised quantitatively according to data point frequency. In this manner, an "ontological similarity between diseases" can be achieved. This idea makes two assumptions: distances between each pair of entities are available (or can be computed) and each entity is annotated with a categorical label. Indeed, this is an ideal situation for ontologies that reside both on the spatial domain (i.e., has location) and on the non-metric space (i.e., has categorical data). It is also considered a *local* approach since it relies on the spatial distribution of data points within a region. In addition, it lends itself well to specific contexts(e.g., *blood diseases*), but can be easily extended as a general purpose approach (e.g., *diseases*).

The number of levels in an ontology (i.e., depth) is countably infinite. While there's no limit on depth, there are certainly trade-offs between information accuracy and computational complexity at different levels. Figure 1, for instance, has $n = 6$ diseases at L1. Making a hypothetical calculation between one disease and all remaining others at that level would require $(n-1) = 5$ operations. Comparing all to all would require $\frac{n(n-1)}{2} = 15$ operations, assuming symmetry among diseases (i.e, AB is equal to BA). At L2, with $n = 12$, 1-to-all would need 11, while all-to-all would require 66 operations.

This example underscores the fact that, at deeper ontological levels, computation costs tend to rise. The good news, on the other hand, is that at deeper levels, information is richer: extracting L1 data only yields general types of diseases, while L2 provides more specific knowledge. For a given application, then, what level strikes the right balance between cost and information content? While certain queries may just ask for *metabolic diseases*(L1), others may demand the same, but related to *Diabetes mellitus*(L2). We use this trade-off as a motivation when calculating an ontological similarity. The major contributions of this paper are as follows:

- **A method to generate a quantitave similarity measure for categorical data points**. We extend the concept of *Pair Correlation Function*(PCF) [14] as *ontological similarities* which measure the frequency of co-occurrence of a pair of categories at specific spatial distances. In our set up, we are interested in any pairs of entities whose distance fall within a given range, and use their category values and frequencies to determine similarity.
- **Reasoning over ontological levels**. Determine when working at deeper levels impacts the tradeoff between information content and the number of categories to be processed.

In section 2, we compare existing literature to our work and note where they deviate or target different goals. Section 3 provides background information referenced throughout the remainder of the paper. The foundation for the categorical similarity measure, which includes identification of spatially-proximal entities, segmentation, and merging are given in Section 4. Section 5 binds our proposed approach in an algorithm, explains its phases, and provides a computational complexity analysis. We discuss our experiments and provide insight into the results in Section 6, and conclude in Section 7.

2. RELATED WORK

The notion of categorical similarity was applied on taxonomies in the early biological sciences (Gregor Mendel, 1822-1884 [12]). In modern times, similarity based on categorical data can be seen as one of two general types: *entity level*, where the entities themselves are compared based on the number of common characteristics; and *attribute level*, where divergence is established among attribute values, but not necessarily among their entities. Therefore, if two persons share a few disease symptons, one could say the two entities are similar. At the *attribute level*, however, the similarity is simply between symptons, and no assumption is extended to the similarity between the two persons.

Entity Level: These approaches are more related to the idea of concept merging. Consider Table 1, where (v_m, v_n) denotes two attribute values. Bouquet *et al.* propose a similar approach to *single hopping*, where the distance between two entities is determined by their shortest path to each other [5], as a function of their distance to the root over the distance of their *Least Common Ancestor* (LCA) to the root. A variation is given by Leacock *et al.* [9], where the shortest path length is scaled by the depth D of the ontological tree. Haase *et al.* considers the length of the shortest path between two concepts and the depth of the tree, adjusting them with parameters α and β respectively for differentiated weighing. In practical terms, they ignore correlation among entities since they do not consider what groups of categories are prevalent. Rather, only paths among individual entities are examined. As will be shown in Section 3, we depart from these approaches by looking not only where they occur, but also how commonly they appear in the dataset.

[1]Other distances are also applicable, but are outside the scope of this document.

Attribute Level: One of the simplest approaches is the *overlap* measure, in which a value of 1 is assigned when two attributes match, and 0 otherwise. This approach gives every match and mismatch the same importance, and thus may not work well for many applications due to its oversimplified measure. There have been other more robust approaches to the problem.

Inverse Occurrence Frequency (IOF) is related to *term-frequency* and *inverse document frequency* (*TF-IDF*) [6]. However, it operates on categorical data, not documents. The highest similarity occurs when v_m and v_n appear only once. When v_m and v_n are the only two values, and each occurs half the time, then the lowest similarity is observed. This approach also favors small categorical sets. Our approach favors entities that co-occur frequently.

Goodall examines the probability that a particular value is observed for a pair of objects in a random sample [8]. Infrequent values are given more importance, which is more applicable to outlier detection, and less to entity relevance. Our approach also investigates frequencies, but rather, we claim higher frequency as a better indicator of importance. Under Lin [11], the similarity value is rewarded on frequent matches, and punished on infrequent mismatches. A drawback is that in many datasets mismatches are dominant, and thus, may have an adverse effect on the overall similarity. For this reason, we do not take mismatches into consideration.

Eskin *et al.* take a somewhat different direction by looking at the number of values a particular category can take [7]. When the values match, the similarity is maximal. Otherwise, it decreases based on the number of possible values. This approach has better usage in small category sets, but is less than ideal in wide categories. While real-world applications do limit the number of categories to a certain extent, our approach leaves that restriction to the domain expert, not the algorithm.

An additional aspect that differentiates our approach is the use of spatial co-location. In the above methods, the relation between entities is independent of physical proximity. Our motivation is that, even though similarity is influenced by frequency, nearby entities should be more influential than distant ones. We address this problem by calculating both spatial and ontological similarities, while imposing on each a weighing scheme, as one of our contributions. Our approach is the last item in Table 1.

Table 1: Similarity Measures

Measure on Entity ⓔ or Attribute ⓐ	spatial
ⓐ $overlap = \begin{cases} 1 \; if \; v_m = v_n \\ 0 \; otherwise \end{cases}$	✘
ⓐ $Goodall = \begin{cases} 1 - \sum_{m=1}^{\lvert V \rvert} Fr(v_m)Fr(v_m - 1) \; if \; v_m = v_n \\ 0 \; otherwise \end{cases}$	✘
ⓐ $Lin = \begin{cases} 2 \times log(Fr(v_m)Fr(v_m - 1)) \; if \; v_m = v_n \\ 2 \times log(Fr(v_m)Fr(v_m - 1) + Fr(v_n)Fr(v_n - 1)) \\ otherwise \end{cases}$	✘
ⓐ $Eskin = \begin{cases} 1 \; if \; v_m = v_n \\ \frac{Fr(v_m)}{Fr(v_m)+2} - \frac{Fr(v_n)}{Fr(v_n)+2} \;\; otherwise \end{cases}$	✘
ⓐ $IOF = \begin{cases} 1 \; if \; v_m = v_n \\ \frac{1}{1+log(Fr(v_m))\times log(Fr(v_n))} \;\; otherwise \end{cases}$	✘
ⓔ $Haase = \begin{cases} e^{-\alpha l \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}} \; if \; v_m \neq v_n \\ 1 \;\; otherwise \end{cases}$	✘
ⓔ $Bouquet = \begin{cases} \frac{d(v_m,root)+d(v_n,root)}{2\times d(LCA(v_m,v_n),root)} \end{cases}$	✘
ⓔ $Leacock = \begin{cases} max[-log(\frac{d(v_m,v_n)}{2D})] \end{cases}$	✘
ⓐ $O\sigma^*_{(v_m,v_n)} \longrightarrow$ see Section 4.3	✔

v_m, v_n: categorical values $Fr(v_m)$: frequency of v_m
$d(v_m, root)$: distance from v_m to tree root
$LCA(v_m, v_n)$: Least Common Ancenstor of v_m and v_n

3. PRELIMINARY CONCEPTS

Entities in a high-dimensional space are often qualified by attributes of different natures: discrete, continuous, categorical (or nominal), interval, ordinal, among others. In our approach, a similarity measure δ defined by a shared attribute a_k between entities e_p and e_q must meet the following requirements:

1. $\delta^{a_k}(e_p,e_q) \geq 0$ iff $e_p \neq e_q$ (*positive definiteness*)
2. $\delta^{a_k}(e_p,e_p) = 0$ (*equality*)
3. $\delta^{a_k}(e_p,e_q) = \delta^{a_k}(e_q,e_p)$ (*symmetry*)

Positive definiteness maintains that for any given distance[2] function, two entities can only lie apart from each other if they are separate objects. Further, two separate objects can have zero distance (i.e., same location). Item #2 complements that thought by stating that an object can never be distant from itself. Symmetry establishes that the distances are the same regardless of origin and destination (i.e., going from A to B is the same as going from B to A).

These definitions lend themselves to practical use under certain assumptions: values are numerical and allow ordering to be established. In addition, data points (entities, elements, objects) are well defined in unambiguous format: two objects are either the same exact thing or completely separate elements. The effect of the above rules have significant impact on data reasoning as it simplifies the computation of similarity measures via numerical analysis. In our earlier example, we can easily determine that a heart-disease patient at 239 mg/dL cholesterol is more similar to a 220 mg/dL person than to a 189 mg/dL person.

Commonly, the 3 above requirements are not applicable to categorical analysis. In the same manner that the *curse of dimensionality* causes metric data to become sparse in high dimensions [4], categorical data dilute the meaning of similarity in a complex ontological space. Take for instance the attributes *side-effects* = {*anxiety, nausea, tiredness, swelling, coughing*} and *risk-factors* = {*gender, heredity, smoking, nutrition*}. If persons pe_1 and pe_2 suffer anxiety, while persons pe_2 and pe_3 have the same nutritional diets, which 2 persons are closer, the ones that share *side-*

[2]Distance and similarity have an inverse relationship, their converse used interchangeably in this paper.

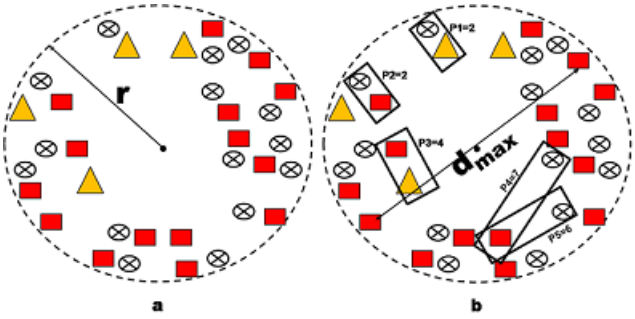

Figure 3: (a) A hypothetical region of radius r. (b) 5 pairs of entities segmented by distance.

effects or the ones that share *risk-factors*? For these 2 categorical attributes, items #1 and #2 (see the requirements above) do not hold since there's no distance within or between *side-effects* and *risk-factors*. However, we do know they are separate concepts, and should be treated likewise. As such, this paper constrains the discussion to a single categorical attribute with many possible values. Item #3 is important because it allows bidirectional processing of attributes, and lowers overall processing when comparing elements. The following definitions will be used going forward:

I) D is a multi-dimensional dataset.

II) $E = \{e_1,e_1,...,e_j\}$ is a set of entities.

III) δ:E x $E \Longrightarrow \mathbb{R}$ is a spatial distance function.

IV) The region R w.r.t. radius r, denoted R_r, constrains the set of entities E such that $\forall k, z \in \{1,...,n\}$, z≠k, the spatial distance $\delta(e_k,e_z) \leq 2r$.

V) $A = \{a_1,a_2,...,a_j\}$ is a set of attributes of e_k in E.

VI) $V = \{v_1,v_2,...,v_j\}$ is a set of categorical values of a_k in A.

VII) $Fr(v_k)$ is the frequency of value v_k in the entire dataset D.

VIII) The spatial segment SS^q w.r.t. R_r is comprised of all pairs of entities whose distance is equal to or greater than $q \times c$ and less than $(q+1) \times c$. q is the segment index and c is a user-defined length of each segment.

IX) The ontological segment $OS_{(v_m,v_n)}$ w.r.t. R_r, is comprised of all pairs of entities whose attribute values are (v_m, v_n) or (v_n, v_m) (i.e., symmetrical).

Objective:

1. Devise an *Ontological Similarity* $O\sigma^*_{(v_m,v_n)}$ between pairs of categories based on co-occurrence frequency and spatial distance.

4. ESTIMATING A LOCAL CATEGORICAL SIMILARITY

In this section, we describe a method to estimate the similarity between two categorical labels based on co-occurrence frequency and the Euclidean distance between pairs of entities. Other distance types can be substituted.

Given an entity e_k in a dataset of |E| entities, a_i is a single-valued random attribute of e_k assuming one of the values in V. The probability of observing $a_i = v_j$ is the frequency of v_j in the dataset:

$$Fr\,(v_j) = P\,[a_i = v_j] = \frac{|E|^{v_j}}{|E|} \quad (1)$$

where $|E|^{v_j}$ is the number of entities with a v_j attribute and $|E|$ is the total number of entities in $\mathbb{R}_r$. Figure 3(a) has a total of 34 entities within radius r. Each entity is represented by its disease category, i.e., □ or ⊗ or △. There are 15 □, 15 ⊗, and 4 △. According to Equation (1), $Fr\,(\square) = \frac{15}{34}$, $Fr\,(\otimes) = \frac{15}{34}$, and $Fr\,(\triangle) = \frac{4}{34}$.

In order to group pairs by distance, we must define two values: the number of segments we would like to work with; and a uniform length of each segment. These values are application-specific. For instance, we can create 5 segments, each one being 2.5 miles wide. Each segment has an index $q \in \{1..n\}$ whose greatest value denotes the total number of segments created. The length of each segment corresponds to parameter c in Section 3, *definition VIII*. We use each segment as follows: within region R_r, all pairs of entities located less than 2.5 miles apart are allocated to Segment 1. Segment 2 encompasses all pairs of entities between 2.5 miles and less than 5 miles, and so forth. Figure 3(b) shows 5 pairs of entities (p_1 through p_5) at different distances. Pair p_1 is composed of two entities with categorical values (⊗△), which are located 2 distance units from each other. Pair p_2 has the same distance, but with different categories. Both would go to the same segment since they have the same distance. Pair $p3$ would go to segment 2, and pairs $p4$ and $p5$ would go to segment 3.

The number of segments can be set randomly or according to application need, since no single approach can be shown always appropriate under different spatial distributions and varying frequencies. A feasible method, however, is to find it based on d_{max}, the distance between the two farthest entities in R_r, as shown in Figure 3(b). Dividing it by a user-defined value λ, produces segment length c:

$$c = \frac{d_{max}}{\lambda} \quad (2)$$

Therefore, if the 2 farthest entities are 7.5 miles apart from each other, and λ=3, we obtain 3 segments, which can be set for distance ranges [0-2.5),[2.5,5.0), and [5.0,7.5). Adjusting λ from lower to higher values influences the granularity from coarser to more fine-grained.

4.1 Spatial Pair Segmentation

The steps explained earlier create a separation of entities in space. Our motivation is that nearby entities tend to be more similar than distant ones, as is commonly accepted in geographic systems. Based on spatial distance, each pair of entities is allocated to a unique spatial segment:

$$SS^q = \bigcup \{e_j, e_k\} \quad (3)$$

such that: $q \times c \leq \delta(e_j, e_k) < (q+1) \times c$ as in Definition VIII. In Figure 4, there are 3 segments: SS^1 and SS^3 have 2 pairs of entities each, while SS^2 has 1 pair, based on Figure 3(b). For simplicity, this example only shows 5 pairs, though all others would need to be accounted for, as well.

4.2 Segment Merging

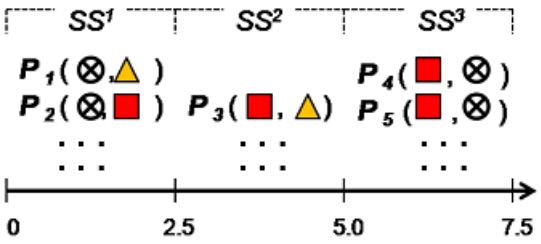

Figure 4: Segmentation based on spatial distance

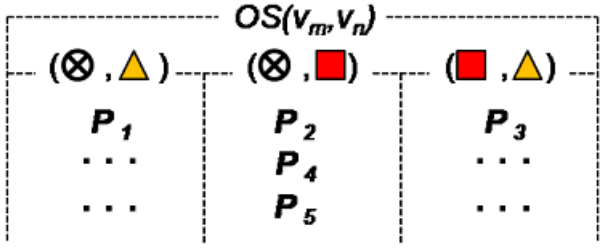

Figure 5: Ontological segmentation

Having in hand all spatial segments, we must now create separate ontological segments, as described in *Defintion IX*. This is simply a matter of allocating all pairs of entities with the same categories to the same segment. In Figure 3(b), for example, we can allocate $p1$ to Segment 1 and $p3$ to Segment 3. As for $p2$, $p4$ and $p5$, they go into segment 2 since they have the same pairs of attributes. The partial results are shown in Figure 5.

The next step is to merge the spatial and ontological segments. This is accomplished by comparing the q^{th} SS^q segment against each of the $OS_{(v_m,v_n)}$ segments, and combining the common pairs of attributes:

$$\overline{Seg}^{SS^q}_{OS_{(v_m,v_n)}} = SS^q \bigcap OS_{(v_m,v_n)} \quad (4)$$

We denote them as *merged segments*. To paraphrase, we break down each spatial segment based on ontological segment. For example, intersecting SS^1 of Figure 4 with $OS_{(\otimes,\triangle)}$ of Figure 5 generates only 1 common pair, which is $\overline{Seg}^{SS^1}_{OS_{(\otimes,\triangle)}} = \{P_1\}$. Likewise, $\overline{Seg}^{SS^1}_{OS_{(\otimes,\square)}} = \{P_2\}$.

4.3 Ontological Similarity Computation

By looking at each merged segment, we first define the Segmented Correlation Factor *(SCF)*:

$$SCF^q_{(v_m,v_n)} = \frac{|\overline{seg}^{SS^q}_{OS_{(v_m,v_n)}}|}{|\overline{seg}^q_*|} \quad (5)$$

SCF computes the probability of observing a given pair of categories among all different pairs of categories in all merged segments with the same index q (denoted by $\overline{seg}^q_*$). In practice, it implies the frequency of entities that share the same attributes within a certain spatial distance. From the previous section, where $\overline{Seg}^{SS^1}_{OS_{(\otimes,\triangle)}} = \{P_1\}$ and $\overline{Seg}^{SS^1}_{OS_{(\otimes,\square)}} = \{P_2\}$, then we can calculate $SCF^1_{(\otimes,\triangle)} = \frac{1}{2}$. Similarly, $SCF^1_{(\otimes,\square)} = \frac{1}{2}$. It is possible to observe pairs of categories that are very frequent, whereas others are very rare. This leads us to define *Ontological Similarity*:

$$O\sigma^q_{(v_m,v_n)} = \frac{SCF^q_{(v_m,v_n)}}{Fr(v_m).Fr(v_n)} \quad (6)$$

For all purposes, the *Ontological Similarity* is just the normalized version of the SCF calculation. By taking into account the frequencies of each category calculated earlier in this section, it prevents extremely common categories from dominating all others. Therefore, $O\sigma^1_{(\otimes,\triangle)} = \frac{\frac{1}{2}}{\frac{15}{34}\cdot\frac{4}{34}} = 9.70$. Because each segment has an ontological similarity per pair of categories, we end up with a matrix of values as shown in Figure 6. Since our goal is to obtain a single measure between categories, we consolidate the various ontological similarities as follows:

$$O\sigma^*_{(v_m,v_n)} = \sum_{i=1}^{\lambda} \frac{O\sigma^i_{(v_m,v_n)}}{i} \quad (7)$$

Equation (7) adds all ontological similarities of a particular pair into one value. Note that it also divides each one by its segment number i. Lower segment numbers have spatially closer entities, and therefore contribute more to the overall value. As the segment numbers increase, the overall contribution decreases. In Figure 6, we then have $O\sigma^*_{(\square,\triangle)} = \frac{6.20}{1} + \frac{2.19}{2} + \frac{2.77}{3} = 8.21$. Likewise, $O\sigma^*_{(\square,\otimes)} = 6.11$ and $O\sigma^*_{(\triangle,\otimes)} = 14.8$. Based on spatial co-location and frequency of attributes, these ontological similarities allows us to infer: $\triangle$ is more similar to $\otimes$ than to $\square$. In addition, $(\square,\otimes)$ is the least similar of the 3 pairs.

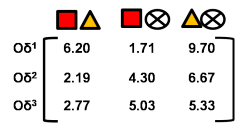

Figure 6: Hypothetical matrix of ontological distances.

5. COMPUTATIONAL COMPLEXITY

Algorithm 1 puts togegther our approach for ontological similarities in 6 phases. As inputs, it expects a set of entities for which one of its attributes is annotated with a categorical value in a set V, and λ, the maximum number of spatial segments to be considered.

Phase I generates several components needed throughout the algorithm. First, matrix *sp_d_m* stores the spatial distance between every pair of entities in the set E (Lines 1-5). Because our data is spatially symmetric (i.e., distance from A to B equals distance from B to A) only the upper half of the matrix needs to be populated. The frequencies of each category are gathered in Lines 6-8, each of which is stored in its own variable. The next pre-processing step is to obtain from the spatial distance matrix the largest distance between any two entities (Line 9). In combination with λ, it helps determine the length c of each segment that we will work with, and also the index q, to identify each generated segment.

Segmentation is done in **Phase II**, where the first step is to examine the spatial distance matrix. For spatial segmentation (Line 12), each pair of entities whose distance is in the same range is allocated to the same segment q. For ontological segmentation (Line 13), all pairs of entities that share the same categories are also stored in the same seg-

Algorithm 1: Computing Ontological Similarities

inputs : set of entities E, λ, set of attribute values V
output: a descending list of ontological distances

{**Phase I: pre-processing steps**}
1: **for** $i = 1$ **to** $|E|$ **do**
2: | **for** $j = i + 1$ **to** $|E|$ **do**
3: | | $sp\_d\_m(i, j) = \delta(e_i, e_j)$ /*build spatial distance matrix*/;
4: | **end**
5: **end**
6: **foreach** (v_i) *in* V **do**
7: | **set** $Fr(v_i) = \frac{|E^{v_i}|}{|E|}$ /*compute frequency of each category*/ ;
8: **end**
{define number and length of segments}
9: **set** $d\_max = max\{sp\_d\_m\}$;
10: **set** $c = d\_max/\lambda$; $q = 1$ **to** λ ;
{**Phase II: allocate entity pairs to spatial and ontological segments**}
11: **foreach** (e_i, e_j) *in* sp_d_m **do**
12: | $SS^q \longleftarrow sp\_d\_m(i, j)$ /*apply *definition VIII* in *Section 3*/;
13: | $OS_{(v_m,v_n)} \longleftarrow sp\_d\_m(i, j)$ /*apply *definition IX* in *Section 3*/;
14: **end**
{**Phase III: merge spatial and ontological segments**}
15: **foreach** (e_i, e_j) *in* SS^q **do**
16: | set $x = 1$ /*create a new index*/;
17: | **if** $(e_i, e_j) \subset OS_{(v_m,v_n)}$ **then**
18: | | $\overline{Seg}^{SS^x}_{OS_{(v_m,v_n)}} \longleftarrow (e_i, e_j)$;
19: | | x++;
20: | **end**
21: **end**
{**Phase IV: compute ontological similarities**}
22: **foreach** $\overline{Seg}^{SS^x}_{OS_{(v_m,v_n)}}$ **do**
23: | compute $SCF^x_{(v_m,v_n)}$ using Eq. (5) ;
24: | compute $O\delta^x_{(v_m,v_n)}$ using Eq. (6) ;
25: | $map((v_m, v_n, x), O\delta^x_{(v_m,v_n)})$ /*store values in a map*/;
26: **end**
{**Phase V: combine ontological similarities**}
27: **foreach** (v_m, v_n, x) || (v_n, v_m, x) *in map* **do**
28: | $O\delta^*_{(v_m,v_n)} =+ O\delta^x_{(v_m,v_n)}$ /*as in Eq. (7)*/;
29: **end**
30: **output** descending$(O\delta^*_{(v_m,v_n)})$;

ment, separate from the spatial segments.

Phase III is simply a matter of matching pairs of entities: when one pair in a spatial segment is also found in an ontological segment (Line 17), the algorithm creates a merged segment to store it (Line 18). A new index is created to keep track of the merged segments (Line 16).

With the merged segments of the previous step, **Phase IV** computes the *Segmented Correlation Factor* for each segment and category pair (Line 23). These values, along with the frequencies of each category from Phase I, are subsequently used in the calculation of the ontological similarities (Line 24). Since each pair of categories gets an ontological similarity value per segment, a map is used as a separate data structure to temporarily store those values (Line 25).

The map is used in **Phase V** as follows: for all pairs of segments with equal (or symmetrical) attribute values, their corresponding ontological similarities for all segments x are summed (Line 28). This allows the algorithm to finally sort and output, in descending order, a list of all ontological similarities (Line 30). The path of calculations leading to the *ontological similarities* goes through different stages. Initially, several pre-processing steps must take place to transform a raw dataset into usable data points. Since this work is based on n entities with an attribute having one of v possible values, their frequencies require $O(n).O(v)$ to be computed. Since $v \ll n$, the latter can be disregarded. Creating the *spatial distance matrix* requires visiting half of the entities in the region and comparing them to the remaining others at $O(\frac{n(n-1)}{2})$ (due to symmetry).

For spatial and ontological segmentation, SS^q and $OS_{(v_m,v_n)}$ both depend on pair values of the *spatial distance matrix*. Therefore each needs $O(|sp\_d\_matrix|)$. Merging the spatial and ontological segments runs at $O(|SS^q|.|OS_{(v_m,v_n)}|)$. Finally, computing the *ontological distances* depends on the number of merged segments at $O(|\overline{seg}^{SS^q}_{(v_m,v_n)}|)$ (individual frequency counts are disregarded since they have been precomputed).

In summary, the worst case for Algorithm 1 is $O(n) = O(n) + O(\frac{n(n-1)}{2}) + O(\frac{n(n-1)}{2}) + O(|SS^q|.|OS_{(v_m,v_n)}|) + O(|\overline{seg}^{SS^q}_{(v_m,v_n)}|)$. Through the manipulation of segment numbers and lengths, or the elimination of *merged segments*, the number of computations can be decreased to achieve shorter run times. Interestingly, the number of categories is less of a factor because they are treated in pairs and examined on a per-segment basis. In addition, many real-world domains limit categorical values per level, which is a reasonable assumption that makes the application more user-friendly.

6. EXPERIMENTS

To gauge the effectiveness of our proposed method, several evaluations have been performed. The process of converting a set of categorical values into a numerical similarity is influenced by different factors. Therefore, our goal is threefold: compare our proposed approach against some of the existing methods in the *Related Works*; observe how these factors impact our computations; and draw conclusions of what they mean in practical terms. Our dataset (*Dawn*)

Table 2: Dataset Characteristics

Total # of entities	380,126
Description	Each entity represents a visit to a hospital's emergency room due to a drug condition, such as allergic reaction or overdose, whether illegal or not.
# categories per level	(L1): 22 (L2): 195 (L3): 170 (L4): 990
Locations	cities of NY, Boston, Minneapolis, Chicago, and Detroit.
Spatial Segments in miles	① [0-240)→ (Bos-NY),(Chi-Det) ② [240-480)→ (Chi-Minn),(NY-Det) ③ [480-720)→ (Det-Minn),(Bos-Det),(NY-Chi) ④ [720-960)→ (Bos-Chi) ⑤ [960-1200)→ (NY-Minn),(Bos-Minn)

is provided by the *U.S Dept. of Health and Human Services* [1] and has the characteristics outlined in Table 2. The entire data is composed of 14 metropolitan areas throughout the United States. In our experiments, however, we limit our region of observation to the 5 locations of the table. The *Dawn* dataset records emergency room events in the 5 metro areas in the year 2009: those are individuals who suffered a drug reaction due to one of several factors, such as illicit drug use, accidental ingestion, suicide attempts, and others. Each event is annotated with its location, and has a *drug name* attribute from an ontological hierarchy of

four levels. Level 1 has 22 categories (e.g., *methenamine*, *dapsone*, *acyclovir*), Level 2 has 170 categories (e.g., *oxacyllin*, *carbamazepine*), Level 3 has 195 categories (e.g., *insulin*, *coumarin*), and Level 4 has 981 categories (e.g., *sulfonamides*, *penicillins*).

In terms of spatial segmentation, we pre-processed the data in ranges of 240 miles (lat-lon distance, not driving distance). This gave us 5 segments that include, for instance, areas less than 240 miles apart, such as NY-Boston and Chicago-Detroit, in Segment 1. Metro areas farther than 240 miles, but less than 480 miles fall in Segment 2 (e.g., Chi-Minn and NY-Det) and so forth. We did vary that segmentation in some of our queries, but note when doing so. In the next subsections, we explain the results of applying Algorithm 1 to the above dataset using different parameters. Initially, we ran the pre-processing steps of **Phase I** of the algorithm ahead of time, to obtain the *spatial distance matrix* and the frequencies of each categorical value in the dataset.

6.1 Effect of Varying the Ontological Levels

The density of entities in each city is very high. New York alone has 58,645 entities while Boston accounts for 39,526. We take a gradual approach and use the data for each city in increments of 1% up to 5%, which yields the range of entities between 1,939 and 9,698. Spatial segmentation (*sp segs*) is kept at the original number of 5 and Levels 1-3 are used for the ontological segmentation (*ont segs*) which yields 231 segments (segments with only 1 occurrence are not considered). We compared our approach against those of *Leacock*

entities[min-max]	sp segs	ont segs	ontological level
1,939-9,698	5	231	1

(LE), *Goodall* (G), and *Lin* (LI). We are interested in how well our *Ontological Similarity* can differentiate categorical values in comparison to each approach. For this purpose, we plot the number of entities (n) against the ontological similarity (Oσ) of *merged segment* 1 and most popular category pair of each run. For the other approaches that does not use segments, we compute their similarity directly using their respective formulas. In the plot, we have also normalized the similarities from zero to one.

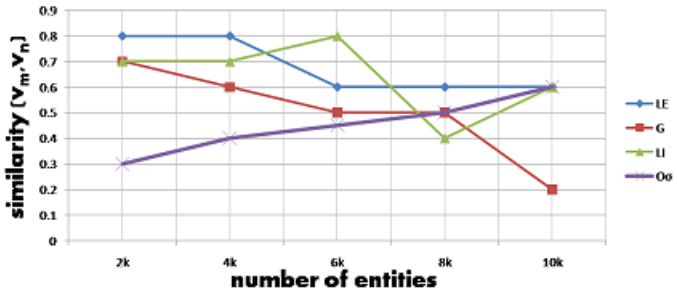

Figure 7: Number of Entities vs. Ontological Similarity - Level 1

Figure 7 shows the similarity between (*warfarin*,*ibuprofen*), which came out to be the most frequent drugs in Segment 1. The graph reveals an inverse trend between our approach

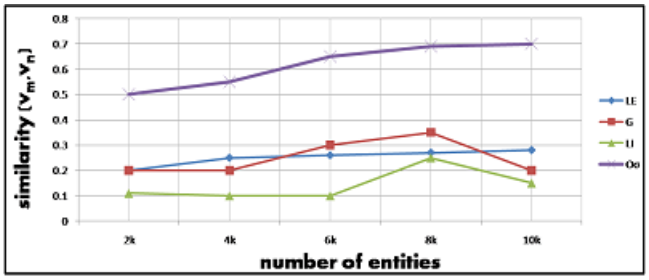

Figure 8: Number of Entities vs. Ontological Similarity - Level 2

($O\sigma$) and the others: while *LE*, *G*, and *LI* display a high initial similarity, they tend to fall as the number of entities increase. By contrast, ours start low and ends higher. The parameters for this run works at a shallow level of the ontology (L1), where the other approaches are fairly efficient. For the next run below, we move to ontological Level 2, where the number of categories increases to 195 and the parameters are as follows:

entities[min-max]	sp segs	ont segs	ontological level
1,939-9,698	5	2,313	2

Figure 8 shows a different result from the previous plot. Our approach tends to take advantage of a higher number of categories, and thus more ontological segments, which bumps up the values of the ontological similarity. LI is particularly susceptible to ups and downs as the number of entities change because it punishes the similarity when it finds mismatches, which we observe after the 8K mark. Our approach behaves slightly more consistently as more entities are added, displaying $O\sigma$ between 0.5 and 0.7. On the next run, we visit Level 3, where the number of categories substantially increases. After cleaning up the very infrequent pairs of categories, we end up with the following parameters:

entities[min-max]	sp segs	ont segs	ontological level
1,939-9,698	5	8,000	3

This setup is particularly interesting because Level 3 of the *Dawn* dataset is extremely large, as shown by the 8000 ontological segments it generates. Figure 9 shows the result. LE and LI behave more poorly than G and $O\sigma$. Particularly, G slightly outperforms our approach as the number of entities increases. The reason is that it has encountered a certain number of fairly infrequent values from which it benefits. For example, with 9K entities, we observed approximately 1,100 segments whose attribute pair occur no more than 5 times. In our approach, this fact helps decrease the $O\sigma$, while under *Goodall* it helps increase the similarity.

6.2 Practical Implications

The previous section compared our approach with different methods and manipulated various factors to observe the behavior of our proposed approach. While that is all good, we also seek to understand the practical results of *Ontological Similarities*.

First and foremost, categorical co-occurrence is an integral part of our approach. One would then expect that when pairs of categories occur frequently, their ontological similarity would always be high. In fact, we observe this is not true

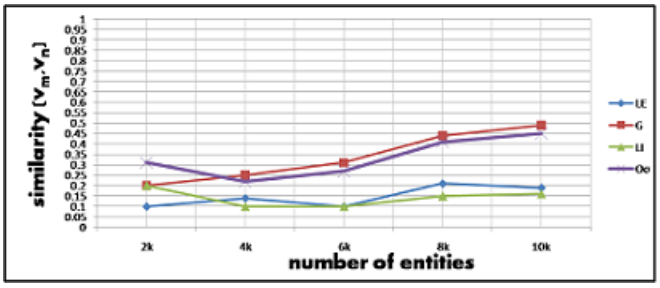

Figure 9: Number of Entities vs. Ontological Similarity - Level 3

quite often. In the Det-NY stretch, for example, the combination (*methenamine,glimepiride*) is a fairly common emergency room event. The former has 81 events and the latter has 112. Their *Ontolgical Similarity* $O\sigma = 42.20$, however, is fairly low as compared to other pairs with even lower frequencies, but that are located in closer spatial proximity. Two such pairs are (*topotecan,iodixanol*) whose $O\sigma$=51.0 and (*cidofovir,pilocarpine*) whose $O\sigma$=77.12. These two are located in the Bos-NY area, whose lesser distance helps increase their similarity measures. All *merged segments* have corresponding pairs of categories in at least one level of the ontology. However, not all *merged segments* have pairs that are present at all levels. As an example, we find that (*antihistamines*, *methylphenidate*) are two drugs present in Level 1. However, no drug sub-categories exist for them in Levels 2 through 4 (*Dawn* populates them with the value -7). This has a practical implication: working at different ontological levels may not be applicable in many domains whose ontologies provides low coverage at deeper levels.

Lastly, Table 3 presents the most similar pairs of drugs along with their frequencies and *Ontological Similiarities* $O\sigma$ for Level 3. It shows, for example, that *ethanol* is ontologically more similar to *heroin* than to any other drugs. Moreover, *marijuana* is highly similar to other drugs for which not data is reported (i.e., *drug unknown*). The high frequencies shown in the table also corroborate our approach that, along with co-occurence and spatial distance, these elements are able to devise a useful *Ontological Similarity* helpful in exploratory analysis.

Table 3: Most Similar Drug Occurences

v_m, v_n	$Fr(v_m)$	$Fr(v_n)$	$O\sigma$
ethanol, heroin	61,819	18,621	112.75
cocaine, miscellaneous agents	20,389	5,334	110.04
marijuana,drug unknown	12,875	8,057	81.01
antineoplastics,hydrocodone	5,615	5,237	77.76
inotropic agents,amoxycillin	4,527	4,513	71.55

7. CONCLUSION

Ontologies provide a wealth of information hidden in nested hierarchical levels. One of its limitations, however, is that ontological data is often categorical, making it inappropriate for many analytical tasks that require numerical values. In this paper, we proposed *Ontological Similarities*, a numerical measure of categorical values based on attribute frequencies and entity co-occurrences. Our approach considers spatial apects of the data and is able to determine application-specific similarity between any pair of categories. We also compare our work to existing methods, and show where our approach is more efficient. Further, we show how categorical pairs can be eliminated from the analysis to save computing cycles. Our work has been effective in uncovering insightful information hidden in different levels of the underlying data.

8. REFERENCES

[1] Dawn: Drug abuse warning network. US Dept. of Health and Human Services. http://www.samhsa.gov/data/DAWN.aspx - July 01, 2012.

[2] International classification of diseases. World Health Organization. http://www.who.int/classifications/icd/en/ , 25, 2012.

[3] M. R. Ackermann, J. Blömer, and C. Sohler. Clustering for metric and non-metric distance measures. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 799–808, 2008.

[4] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[5] P. Bouquet, G. Kuper, M. Scoz, and S. Zanobini. Asking and answering semantic queries. In *Proc. of Meaning Coordination and Negotiation Workshop (MCNW'04) in conjunction with ISWC '04*, 2004.

[6] T. H. Cao. *Conceptual Graphs and Fuzzy Logic: A Fusion for Representing and Reasoning with Linguistic Information*. Springer, Boston, Massachusetts, 2010.

[7] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Applications of Data Mining in Computer Security*. Kluwer, 2002.

[8] D. Goodall. A new similarity index based on probability. *Biometrics*, 22(4):882–907, 1966.

[9] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In C. Fellfaum, editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts, 1998.

[10] F. Li, J. Yang, and J. Wang. A transductive framework of distance metric learning by spectral dimensionality reduction. In *Proceedings of the 24th Intl. Conference on Machine Learning (ICML '07)*, pages 513–520, Corvallis, OR, December 2007.

[11] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th Intl. Conference on Machine Learning (ICML '08)*, pages 296–304, San Francisco, CA, 1998.

[12] G. Mendel. Experiments in plant hybridization. http://www.mendelweb.org - June 06, 2012.

[13] J. Partyka, N. Alipanah, L. Khan, B. Thuraisingham, and S. Shekhar. Content-based ontology matching for gis datasets. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 51:1–51:4, 2008.

[14] T. Reed and K. Gubbins. *Applied Statistical Mechanics: Thermodynamic and Transport Properties of Fluids*. Butterworth-Heinemann, Boston, Massachusetts, 1973.

[15] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, December 2009.