A Framework for the Expansion of Spatial Features Based on Semantic Footprints
Technical Report

Raimundo F. Dos Santos Jr
Spatial Data Management Lab
Virginia Tech
Falls Church, VA – USA
rdossant@vt.edu

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

Chang-Tien Lu
Spatial Data Management Lab
Virginia Tech
Falls Church, VA – USA
ctlu@vt.edu

ABSTRACT

Geographic feature expansion is a common task in Geographic Information Systems (GIS). Identifying and integrating geographic features is a challenging task since many of their spatial and non-spatial properties are described in different sources. We tackle this expansion problem by defining semantic footprints as a measure of similarity among features. Furthermore, we propose three quantifiers of semantic similarity: spatial, dimensional, and ontological affinity. We show how these measures dilute, concentrate, harden, or concede the feature space, and provide useful insights into the semantic relationships of the spatial entities. Experiments demonstrate the effectiveness of our approach in semantically associating the most appropriate spatial features.

1. INTRODUCTION

Geospatial web services as well as Geographic Information Systems (GIS) commonly exchange data for a multitude of application domains from real estate to marketing. For these systems, one major challenge has been interoperability: the capacity for understanding different data sources in spite of syntactic and semantic differences in language. Several organizations have attempted to mitigate this problem with standardized specifications. The Open Geospatial Consortium (OGC), for instance, has proposed a set of frameworks in an attempt to bring uniformity to spatial data processing [8]. In general, these frameworks use standard grammars such as *Extensible Markup Language* (XML) for data transport. Google and Yahoo! often use KML (Keyhole Markup Language) in their mapping APIs. Government agencies often use Geography Markup Language (GML) for data exchange [12]. One advantage of XML is its hierarchical structure which helps define relationships among entities. As a consequence, it also lends itself well to object orientation that is so prevalent in modern computing.

Consider the two GML examples depicted in Figure 1: *Data Source 1* describes a *geometryProperty* named *Leon Dept of Housing*, whereas *Data Source 2* describes another geometric object called *Hope Apartments*. What is the relationship between these two geographic features/objects? A quick look at their attributes provides some hints: they are within close proximity of each other (lines 1-3), both are urban structures (line 6), and one object occupies similar but less area than the other (lines 7-9). Based on these observations, the following possibilities arise: (1) *Hope Apartments* is part of the *Leon Dept of Housing*; (2) They are indeed the same since *Leon Dept of Housing* was renamed *Hope Apartments* and moved across the street from its original location into a smaller facility; (3) They are two independent facilities that are coincidentally co-located. Without further contextual considerations, only domain experts can make a complete and necessary determination of the nature of relationship between these two geographic features.

	Data Source 1	Data Source 2	
1	<gml:coordinates>	<gml:coordinates>	1
2	-56.3159,	-56.3101,	2
3	52.5168	52.5199	3
4	</gml:coordinates>	</gml:coordinates>	4
5	</gml:Point>	</gml:Point>	5
6	</ogr:geometryProperty>	</ogr:geometryProperty>	6
7	<ogr:building>	<ogr:building>	7
8	<ogr:AREA>	<ogr:AREA>	8
9	5.000	3.932	9
10	</ogr:AREA>	</ogr:AREA>	10
11	<ogr:PERIMETER>	<ogr:PERIMETER>	11
12	25.010	22.882	12
13	</ogr:PERIMETER>	</ogr:PERIMETER>	13
14	<ogr:NAME>	<ogr:NAME>	14
15	Leon Dept of Housing	Hope Apartments	15
16	</ogr:NAME>	</ogr:NAME>	16
17	<ont:living space/>	<ont:apartment/>	17
18	<ogr:LAT>	<ogr:LAT>	18
19	543831	523300	19
20	</ogr:LAT>	</ogr:LAT>	20
21	<ogr:LONG>	<ogr:LONG>	21
22	56100	52449	22

Figure 1 – Example GML Data Sources

The discussion above illustrates the challenges in reasoning on disparate data sets. Work in this field of research proposes a wide variety of approaches to handle data disparity: value comparisons, word distances, disambiguation, look-ups on gazetteers, and others [24,25]. While some of these approaches have been successful to some extent, they often introduce a high level of complexity in semantic processing. Our work aims to reduce this complexity by proposing a semantic framework which exploits spatial relationships built into the geographic features. The framework will help elicit hidden and useful semantic information about the geographic features and their neighbors. Our goal is not only to determine possible matches, but also to determine whether geographic features can be deemed complementary (or irrelevant) to one another. We would like to determine if *Leon Dept of Housing* and *Hope Apartments* are the same building or just similar facilities. We are also interested in measuring their physical proximity and then combine their associated descriptions so that a higher authority (i.e., the domain expert) may make a final decision based on his/her own constraints.

We propose a method of semantic footprints based on the three relational concepts: the spatial affinity within the data space; the dimensional affinity within the XML hierarchy; and the ontological similarity based on the feature's class label. In addition, we describe an approach that utilizes the above measures to associate and link disparate geographic features. Because the

number of geographic features is potentially large, we devise the concepts of dilution, hardness, concentration, and concession as a means to efficiently and effectively perform semantic analysis on the data. These concepts provide criteria to evaluate the ongoing progress of our analysis and help answer the following questions: are geographic features/objects being found in close proximity to the initial geographic feature query? If so, do these geographic features add sufficient relevant information to the initial geographic feature query? If the user is initially seeking only *k* number of features, then are the current ones sufficiently relevant or should the process continue to search for others that may be more relevant? Our motivation relates to tools and technologies that rely on hierarchically semi-structured data (e.g., XML, GML, and KML), have strong syntactic capabilities, but lack semantic support for data processing, and can exploit semantic footprints as an auxiliary tool to enhance semantic alignment.

This paper is organized as follows: In Section 2, we give related approaches to feature reconciliation and object matching. Section 3 gives the general problem statement, expands on our theoretical approach to *Semantic Footprints*, and elaborates on a semantic analysis approach. Experiments are described in Section 4 and conclusion is provided in Section 5.

2. RELATED WORK

Early research on spatial entities is related to the works of GIS. With the support of organizations such as the OGC, standards have been established for the management of geographic features [8] using common communication protocols (e.g., HTTP) and XML-based encodings (e.g., GML). With the advent of geospatial portals (e.g., Google Maps, Yahoo! Maps), geographic features have taken on increased popularity. Traditionally, geographic feature matching and expansion have been primarily utilized in spatial indexing methods for database systems. The use of spatial indices is abundant in this area as exemplified in [1, 5, 10]. However, our work does not focus on spatial indices but rather emphasize on the development of an approach that will enhance the extraction, processing, and analysis of semantic information in spatial data. Other aspects such as data quality and composability of grammars are described in [16, 17]. Current literature in semantic information processing can be classified into one of the following categories:

Schema Matching: Rahm *et al.* proposed the decomposition of complex schemas into simpler sets [2,14]. Doan *et al.* used a set of semantic mappings to learn other mappings using machine learning techniques [7]. Islam *et al.* proposed a method to determine the semantic similarity of words and another for word segmentation [4]. Schema matching becomes challenging when many schemas are involved. In addition, it often only works with textual elements which makes spatial processing inefficient and/or impractical. We depart from the above works by considering the spatial characteristics of objects, which is not in the scope of any of the aforementioned works.

Object Consolidation: The difficulty of combining objects described in different sources is addressed by Beeri *et al* [11]. They extend the one-sided nearest neighbor join into mutually nearest neighbors. As described by Bleiholder *et al.*, data fusion can also be performed at a query language level [13]. Instead of relying on schema information, objects are considered for their attribute values rather than attribute types. Seghal *et al.* proposed entity resolution primarily as a function of locations [15]. The spatial component is deemed similar when their distance meets a certain threshold. We differ from these approaches by extending our work beyond object fusion and propose methods to evaluate semantic relationships within the attribute and ontological spaces. An example output of our method includes determination of geographic features that are complementary within an application domain.

Ensemble Reasoning: This class of techniques combines characteristics of both schema matching and object consolidation to provide semantic analysis. They tend to be more effective in applications in which prior knowledge of the schemas is available. Fazzinga *et al.* proposed a query language to combine partial answers from different sources on the basis of limited knowledge about the local schemas in XML documents [3]. Leitao *et al.* proposed a method to detect duplicate objects in XML data using Bayesian networks [6]. A schema matching approach, Protoplasm, is an aggregation of several existing methods to reconcile named entities [9]. Unlike our proposed framework, these studies do not consider the spatial component of an object and rely primarily on non-spatial textual content.

Class	Name	Primary Focus	Goal	General Spatial Applicability
Schema Matching	Rahm [2][14] Doan [7]	Logical Structure	Feature Matching	Low
Object Consolidation	Beeri [11] Bleiholder[13]	Attribute Values	Feature Matching	Medium
Ensemble Reasoning	Fazzinga [3] Leitao [6]	Structure, Attributes, Types	Feature Matching & Likeness	Medium
Ensemble Reasoning	Semantic Footprints	Spatial Structure	Feature Matching, Likeness & Complement	High

Table 1 – Summary of Semantic Information Processing Approaches

Table 1 provides a summarized view of the literature in semantic feature analysis. The last row gives a snapshot of how our work differs from existing approaches. Our proposed framework is unique in several ways. **First**, we take a qualitative view of feature expansion by avoiding explicit comparisons on data values. **Second**, we extend the notion of spatial co-location to include the most semantically relevant nearby features which are not necessarily the closest in geographic space. For example, if a source describes several buildings and water bodies, nearby houses are possibly more relevant to a query originating from a house than a water body. **Third**, our framework is oriented towards data sources of similar application domains. As an illustration, consider the marketing realm. In its context, nearby stores and malls would most likely provide more relevant information than, for instance, weather data. We propose spatial proximity, dimensional affinity, and ontological similarity to improve the efficiency of our semantic analysis by limiting the number of geographic features or objects under consideration.

3. PROBLEM DEFINITION OF SPATIAL FEATURE EXPANSION

The nomenclature below formalizes the spatial feature expansion problem.

Given:

- Set $D = \{d_1,\ldots,d_i,\ldots,d_n\}$ *where* d_i *is a semi-structured hierarchical data source (e.g., GML file).*
- Geographic feature set $f_{geo}(d_i) = \{g_1,\ldots,g_j,\ldots,g_m\}$ *where the* g_j*'s are all the geographic features or objects of data source* d_i *and* $m = |d_i|$ *is the number of geographic features in* d_i.
- Set $G = U_{i=1\ldots n} f_{geo}(d_i)$**.** The set *G is the union of all geographic features in all data sources* $d_1 \ldots d_n$.
- Attribute set $f_{att}(g_j) = \{a_1,\ldots,a_k,\ldots,a_q\}$ *where the* a_k*'s are all element/attribute types of the geographic feature* g_j.

Objectives:

I. From a starting geographic feature g_s (initial query), find the set $G_{close}(g_s) = \{g_j \mid g_j \in G$ and $dualÅff(g_s, g_j) \geq \xi_{close}\}$ where *dualÅff* is a measure of the degree of spatial closeness and ξ_{close} is a user-defined threshold.

II. From a starting geographic feature g_s, find the set $G_{dim}(g_s) = \{g_j \mid g_j \in G_{close}(g_s)$ and $dimÅff(g_s, g_j) \geq \xi_{dim}\}$ where G_{dim} is a measure of attribute similarity and ξ_{dim} is a threshold based on the ranking order of $dimÅff(g_s, g_j)$.

III. From a starting geographic feature g_s, find the set $G_{ont}(g_s) = \{g_j \mid g_j \in G_{close}(g_s)$ and $ontÅff(g_s, g_j) \geq \xi_{ont}\}$ where G_{ont} is a measure of ontological similarity and ξ_{ont} is a threshold based on the ranking order of $ontÅff(g_s, g_j)$.

IV. From a starting geographic feature g_s, find an ordered set $G_{final}(g_s) = \{g_j \mid g_j \in G_{close}(g_s)$ and $(i < j \rightarrow Sem\varphi(g_s, g_i) \geq Sem\varphi(g_s, g_j)\}$ where $Sem\varphi$ is a measure of similarity based on $dimÅff$ and $ontÅff$.

3.1 Concept of Semantic Footprints

Hierarchical structures encapsulate a rich set of relationships not always visible to the naked eye. Names do not always match, locations are ambiguous, and characteristics may range wildly. These differences arise because data is affected by many factors, such as external noise, human subjectivity, and un-calibrated measuring tools. While some systems attempt to match features by introspecting their properties [18], we avoid exhaustive attribute comparisons as they tend to increase computational complexity when many geographic features are present. To establish an efficient and effective representation of semantic relationships, we define semantic footprints and their components in the subsections below.

3.2 Spatial Affinity within the Data Space

Geographic features are commonly described in terms of their locations and hence, we give our first definition for describing spatial closeness:

Definition 1: *Geographic feature* g_i *is said to be locally-fit (LF) in data source* d_i *if its minimum bounding rectangle (MBR) is explicitly provided in the data source.*

For example, given five locally-fit geographic features $g_1 \ldots g_5$ residing in data sources $d_1 \ldots d_5$, respectively, we investigate whether g_1, the starting query feature, has any spatial significance to $g_2 \ldots g_5$. We give the spatial significance, namely *dual affinity*, by:

$$DualÅff(g_i, g_j) = 1 - \frac{Dist(g_i, g_j) - MinDist(g_i, g_j)}{MaxDist(g_i, g_j) - MinDist(g_i, g_j)} \quad \text{(Eq. 1)}$$

Assuming that the geographic features g_i and g_j share a common coordinate system, *Equation 1* defines dual affinity as the degree of spatial closeness between the features. The *Dist* function can be generalized to any appropriate spatial distance, for example, we often consider the geodesic distance for latitudinal and longitudinal coordinates. Other distances such as Euclidean or Manhattan distances can also be used. Furthermore, the choice of locations of spatial extents can be approximated by its centroid, which is an acceptable approach in many types of application. For example, $Dist(g_i, g_j)$ may use the centroids of g_i's and g_j's MBRs as their representative locations. The functions *MinDist* and *MaxDist* represent the shortest and longest possible distances between two geographic features respectively.

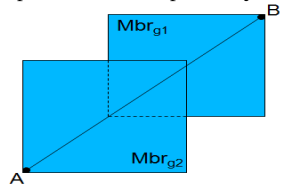

Figure 2 – MinDist and MaxDist for Two MBRs

For example, in Figure 2 the geographic features are described by their MBRs, therefore the *MaxDist* between any two objects is the length of the segment AB and *MinDist* is zero since the MBRs overlap. From a spatial point of view, two features have maximal affinity when their locations are the same, i.e., $dualÅff=1$. Hence, to achieve *Objective I*, $G_{close}(g_s)$ can be determined by collecting all features whose *dualÅff* is higher than a given ξ_{close}.

We build upon *DualÅff* to define the spatial footprint of a geographic feature:

Definition 2: The *footprint* φ of a geographic feature g_s is given by the set of all attributes of all geographic features in $G_{close}(g_s)$.

$$\varphi(g_s) = \bigcup_{i=1..|G_{close}(g_s)|} (f_{att}(g_i)) \text{ where } g_i \in G_{close}(g_s) \quad \text{(Eq. 2)}$$

The footprint represents the maximal collection of attributes types within the set of $G_{close}(g_s)$. This maximal set will impose a bound on the computational complexity of the proceeding semantic operations.

3.3 Dimensional Affinity in the Data Space

One attractive aspect of XML is its ability to define class relation in a hierarchical fashion. This idea gives rise to *dimensional affinity* and applies to all geographic features, whether they are locally-fit or do not have an explicit location. In these cases, we observe the dimensions of the feature (its attributes/elements), while relying on the location of its parent. In Figures 3 and 4, the five features (the circles) are within some MBR not of their own, indicated by the encompassing squares covering an area larger than the features themselves. In Figure 3, only the location of the parent is available (locally-displaced feature), and Figure 4 has no location but the bounds of the data set (globally-displaced). While these two cases do not have an explicit location, they can still be useful to establish a semantic footprint. Dimensional affinity gives the ability to measure how similar two geographic features are in relation to their elements and attributes.

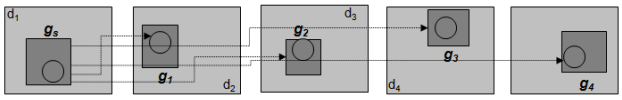

Figure 3 – A set of 5 locally-displaced features in 5 data sets

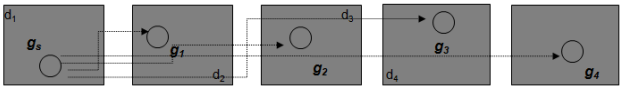

Figure 4 – A set of 5 globally-displaced features in 5 data sets

We define dimensional affinity as follows:

$$DimÅff\ (g_s, g_k) = \frac{|(f_{att}(g_s) \cap f_{att}(g_k))|}{|\varphi(g_s)|} \quad \text{(Eq. 3)}$$

where $g_s, g_k \in G_{close}(g_s)$.

DimÅff gives the ratio of common attributes between two geographic features, g_s and g_k, in relation to its total number of attributes, i.e., its footprint. Hence, the dimensional affinity is dependent upon the spatial proximity of features in $G_{close}(g_s)$ and what attribute types they share in common. If *Leon and Stellar* together have 22 attributes, but only 5 in common, then *DimÅff(Leon,Stellar) = 5/22= 0.23* and if the ξ_{dim}, is met, the geographic features can later be utilized in the analysis of the complete semantic footprint. *Objective II* is then achieved by forming $G_{dim}(g_s)$ as the sorted set of all geographic features with dimensional affinity $\geq \xi_{dim}$.

3.4 Ontological Class Affinity

Ontologies represent a classification scheme to group similar objects and are commonly used in a wide range of fields, from medicine to the data sciences [19,20]. Given this as a motivation, we show a method to compute the hierarchical ontological distance among features as the third component of our semantic footprint. We define the class distance between two nodes in a common hierarchical ontology as follows [23]:

$$Class\_d(g_s,g_k) = d(LCA(g_s, g_k), g_s) + d(LCA(g_s, g_k), g_s) \quad \text{(Eq. 4)}$$

where $d(g_i, g_j)$ is the edge length between the classes of g_i and g_j and $LCA(g_i, g_j)$ is the Lowest Common Ancestor defined as the farthest node from the root that is the most immediate ancestor of both g_i and g_j.

From the class distance measure above, we define the ontological class affinity *OntÅff* as follows:

Definition 3: The *ontological class affinity* $OntÅff(g_s, g_k)$ is the degree of similarity between the classes of g_s and g_k from a common hierarchical ontology:

$$OntÅff(g_s, g_k) = \frac{1}{1+Class\_d(g_s,g_k)} \quad \text{(Eq. 5)}$$

Hence, if geographic features g_s and g_k are of the same class, $OntÅff(g_s,g_k) = 1$. For example, if Leon is classified as an "*apartment*" and *Stellar* is a "*house*", assuming these two classes are two hops apart in the ontology, then their $OntÅff = \frac{1}{1+2} = 0.333$. *Objective III* can then be achieved by creating $G_{ont}(g_s)$ as the sorted set of all geographic features with ontological class affinity $\geq \xi_{ont}$.

One goal of this study is to maintain the total number of threshold parameters to a minimum under the assumption that spatial, dimensional, and ontological affinities are jointly independent. Our framework minimally maintains only one threshold for each of the components of the semantic footprint (*DualÅff, DimÅff,* and *OntÅff*). Although we assume joint independence amongst these components, existence of correlations does not affect the effectiveness of our semantic measures. In fact, potential correlations between these components can be discovered and further explored via our proposed semantic analysis process discussed in the proceeding Section 3.5.

Fusing Dual Affinity, Dimensional Affinity, and Ontological Class Affinity

Combining the measures of *OntÅff* and *DimÅff*, we propose *semantic footprint Semφ* as a total measure of the semantic similarity between two geographic features of $G_{close}(g_s)$. Formally, semantic footprint *Semφ* is defined as follows:

Definition 4: The *semantic footprint between two geographic features* g_s *and* g_k *is given by:*

$$Sem\varphi(g_s, g_k) = \frac{DimÅff(g_s,g_k)+OntÅff(g_s,g_k)}{2} \quad \text{(Eq. 6)}$$

Because *OntÅff* and *DimÅff* apply to elements of G_{close}, *Semφ* inherits the spatial similarity constraint (via *DualÅff*) of the geographic features. Hence, *Semφ* provides a similarity measure between geographic features based on spatial, dimensional, and ontological affinities.

From our example in Figure 1, the semantic footprint between Leon and Stellar is *Semφ(Leon,Stellar)= (0.23 + .33)/2 = 0.28. Equation 6* helps us achieve *Objective IV* by establishing a ranking criterion for $G_{final}(g_s)$ as the set of all geographic features starting from g_s.

3.5 Complexity Analysis

This section provides an analysis of the costs for computing the terminal set of geographic features in $G_{final}(g_s)$ for a given geographic feature query g_s. The total cost for generating the set $G_{final}(g_s)$ is:

(Eq. 7)

$$Cost\left(G_{final}(g_s)\right) = Cost\left(G_{close}(g_s)\right) + Cost\left(G_{dim}(g_s)\right) + Cost\left(G_{ont}(g_s)\right)$$

Assuming that no spatial indexing has been applied to the geographic feature set *G*, the cost for generating $G_{close}(g_s)$ is:

(Eq. 8)

$$Cost\left(G_{close}(g_s)\right) = |G| * DistCalc\_Cost = O(|G|)$$

where *DistCalc_Cost* is the cost of calculating the distance between two features. The distance calculation is a constant time operation.

To obtain $G_{dim}(g_s)$, the footprint is generated and the set intersect operation is performed between g_s and all other geographic features in $G_{close}(g_s)$. The set intersect operation is implemented using a hash table which gives a linear time cost. The total cost for computing the set $G_{dim}(g_s)$ is thus:

(Eq. 9)

$$Cost\left(G_{dim}(g_s)\right) = \sum_{i=1..|G_{close}(g_s)|}(|f_{att}(g_s)| + |f_{att}(g_i)|) = O\left(|G_{close}(g_s)| * Max_{i=1..|G_{close}(g_s)|}(|f_{att}(g_i)|)\right) = O(|G_{close}(g_s)| * |\varphi(g_s)|)$$

where $\varphi(g_s)$ is the footprint.

The set $G_{ont}(g_s)$ is obtained by performing ontological class distance calculations between g_s and all other geographic features in $G_{close}(g_s)$. A lookup table of the class IDs which link to the class

nodes in the ontology allows for *O(1)* search time for a given geographic feature class. Once the pair of nodes is found in the ontology graph, the *Lowest Common Ancestor* (LCA) can be determined in time linear to the ontology level size by traversing to the root node and obtaining the longest common node sequence between the two geographic feature classes. The following provides the total cost of generating $G_{ont}(g_s)$:

(Eq. 10)

$$Cost\big(G_{ont}(g_s)\big) = O(Ont_{ls})$$

where Ont_{ls} is the level size of the ontology.

Hence, the total cost of generating $G_{final}(g_s)$ is:

(Eq. 11)

$$Cost\left(G_{final}(g_s)\right) = O(|G|) + O(|G_{close}(g_s)| * |\varphi(g_s)|) + O(Ont_{ls})$$

3.6 Progressive Dilution, Hardness, Concentration, and Concession

Traversing data sources in search of related features is an ongoing process for which no halting point is clearly defined. Using the concepts of our approach, we present a systematic method to evaluate the progression of the relevant features from a starting geographic feature g_s as more geographic features $g_{1...}\ g_m$ become available for processing. The goal is to observe the changes in semantic footprint as more geographic features are analyzed, and determine to which extent *DimÅff* and *OntÅff* are contributing to the semantic footprint *Semφ*. For this purpose, we present four definitions also referred to as *density sets*:

Definition 5: The set $G_{dilution}(g_s) = \{g_j \mid g_j \in G_{close}(g_s)$ and $DimÅff(g_s,g_j) \leq t_{dim}$ *and* $Sem\varphi(g_s,g_j) \geq \xi_{sem}\}$, *where* ξ_{sem} *is a user-defined threshold for high semantic footprint and* t_{dim} *is a user-defined threshold that establishes a low level for dimensional affinity.*

Dilution is the set of features with high semantic footprint, but low dimensional affinity. It is indicative of features that do not share many attributes in common. In such cases, a high *Semφ* is mostly dependent on *OntÅff*, the second component of the semantic measure.

Definition 6: The set $G_{hardness}(g_s) = \{g_j \mid g_j \in G_{close}(g_s)$ and $OntÅff(g_s,g_j) \leq t_{ont}$ *and* $Sem\varphi(g_s,g_j) \geq \xi_{sem}\}$, *where* ξ_{sem} *is a user-defined threshold for high semantic footprint and* t_{dim} *is a user-defined threshold that establishes a low level for ontological affinity.*

Hardness defines a set of features with high semantic footprint, but low ontological affinity. When the features are not similarly-typed (i.e., far in the ontological classification), a high *Semφ* must rely primarily on *DimÅff*.

Definition 7: The set $G_{concentration}(g_s) = \{g_j \mid g_j \in G_{close}(g_s)$ and $DimÅff(g_s,g_j) > t_{dim}$ and $OntÅff(g_s,g_j) > t_{ont}$ *and* $Sem\varphi(g_s,g_j) \geq \xi_{sem}\}$, *where* ξ_{sem} *is a user-defined threshold for high semantic footprint and* t_{dim}, t_{ont} *are thresholds for minimum values of for dimensional and ontological affinities respectively.*

Concentration *is the set of features that yield a high semantic footprint from both a high number of shared attributes and close ontological proximity. It* balances a mix of geographic features that are not only similar in attribute commonality, but also similar in attribute types.

Definition 8: The set $G_{concession}(g_s) = \{g_j \mid g_j \in G_{close}(g_s)$ and $g_j \notin (G_{concentration}(g_s) \cup G_{dilution}(g_s) \cup G_{hardness}(g_s))$

Concession *is the set of features that cannot be classified as any of the types in Definitions 5-7. Practically, they represent geographic features with low affinity in general, both dimensional, ontological, and as a consequence, have a low semantic footprint.*

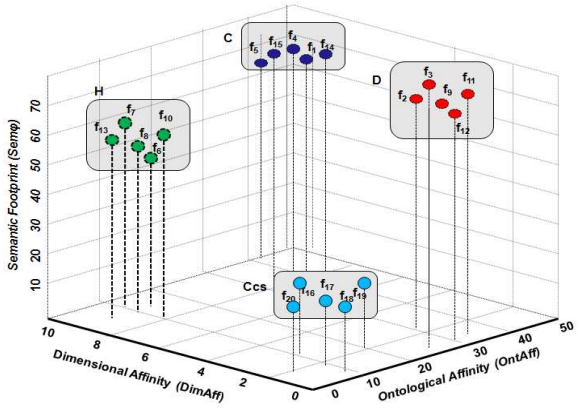

Figure 5 – A Hypothetical Snapshot of Dilution, Hardness, Concentration, and Concession

Figure 5 illustrates the progression graph of a hypothetical geographic feature traversal. The **H**-set shows an area of hardness composed of five features with high semantic footprint, but low ontological affinity. Dilution can be seen at the **D**-set where dimensional affinity is low. In this case, the high semantic footprint can be explained from the high ontological affinity. The concentration set ***C*** shows features with both high dimensional and ontological affinity, whereas all other cases fall under the concession ***Ccs***-set. A concentration set (***C***) is possibly a richer source of information that can enhance the starting geographic feature more so than *D* or *H*.

Thresholds t_{ont} , t_{dim}, and ξ_{sem} can be manipulated to accommodate the application requirements. For instance, if dimensional affinity (i.e., common attributes) is more desirable than type matching (i.e., ontological proximity), the application should explore a hardness set (and vice-versa for a dilution set). When both factors are important, a concentration set provides a more suitable mix. It is also possible to provide an initial and automatic determination of t_{ont} , t_{dim}, and ξ_{sem} by using the centroid of the semantic footprints of the geographic features in G_{final}. The automatically generated thresholds can serve as the starting point for which further adjustments can be made as the analysis progresses. The thresholds t_{ont} , t_{dim}, and ξ_{sem} can be obtained as follows for a given starting geographic feature query g_s:

(Eqs. 12)

$$t_{ont} = \frac{\sum_{i=1..|G_{final}(g_s)|}(OntÅff(g_s,g_i))}{|G_{final}(g_s)|}$$

$$t_{dim} = \frac{\sum_{i=1..|G_{final}(g_s)|}(DimÅff(g_s,g_i))}{|G_{final}(g_s)|}$$

$$\xi_{sem} = \frac{\sum_{i=1..|G_{final}(g_s)|}(Sem\varphi(g_s,g_i))}{|G_{final}(g_s)|}$$

Similarly, the medoid of the semantic footprints can also be used in lieu of the centroid. Employing the medoid can provide a more robust threshold set as it less sensitive to any outliers that may exist in G_{final}.

Algorithm 1 - Identifying Dilution, Hardness, Concentration, and Concession Sets

Inputs: g_s, G_{close}, ξ_{sem}, t_{dim}, t_{ont}

Outputs: $G_{dilution}(g_s)$, $G_{hardness}(g_s)$, $G_{concentration}(g_s)$, $G_{concession}(gs)$

1: **using** g_s **and** g_i **in** G_{close} **where i in {1..n}**
2: **for each** g_i
3: calculate $DimÅff(g_s,g_i)$ (Eq. 4) ;
4: calculate $OntÅff(g_s,g_i)$ (Eq. 6) ;
5: $Semϕ(g_s,g_i) = DimÅff(g_s,g_i) + OntÅff(g_s,g_i)$;
6: **If** $(DimÅff(g_s,g_i) \le t_{dim}$ && $Semϕ(g_s,g_i) > \xi_{sem})$
7: add $g_i \rightarrow G_{dilution}(g_s)$;
8: **Else If** $(OntÅff(g_s,g_i) \le t_{ont}$ && $Semϕ(g_s,g_i) \ge \xi_{sem})$
9: add $g_i \rightarrow G_{hardness}(g_s)$;
10: **Else If** $(DimÅff(g_s,g_i) > t_{dim}$ && $OntÅff(g_s,g_i) > t_{ont}$ && $Semϕ(g_s,g_i) > \xi_{sem})$
11: add $g_i \rightarrow G_{concentration}(g_s)$;
12: **Else** add $g_i \rightarrow G_{concession}(g_s)$;
13: **end for**
14: **output** $G_{dilution}$, $G_{hardness}$, $G_{concentration}$, $G_{concession}$

Algorithm 1 shows a method that uses Definitions 5,6,7, and 8. First, the semantic components are calculated in Lines 3 and 4, and combined as the total semantic footprint in Line 5. Lines 6-12 apply simple logic to determine if the current geographic feature falls under dilution, hardness, concentration, or concession. Each feature is stored into its appropriate set for later examination.

4. EXPERIMENTS

Given a starting geographic feature, our goal is to find other related features within one or more data sources. Our datasets are composed of features of the cities of Frankfurt, Leverkusen, and Konigswinter [21]. For the ontology, we used NASA's SWEET [22], which we extended with urban structure concepts of *home, apartment, hotel, building, warehouse, and construction.*

Our first step is to extract features from the first available data source and calculate their semantic footprint (*DualÅff, DimÅff, OntÅff*). Subsequently, regions of dilution, hardness, concentration, and concession can be identified, allowing their respective sets to be populated according to *Algorithm 1*.

In terms of measurement, we are interested in: (a) obtaining $G_{final}(g_s)$ when different parameters are considered; (b) identifying sets of dilution, hardness, concentration, and concession related to the starting geographic feature.

g_s = Geb537	$\|f_{att}(g_s)\|$ i.e., Attribute Count (g_s)	$\|f_{att}(g_s) \cap f_{att}(g_i)\|$ i.e, Shared Attribute Count Range (g_i)	*Class_d, i.e.* Ontological Variation (g_i)
Query I	30	min=5, max=24	min=0, max=25
Query II	30	10	min=1, max=29
Query III	30	18	min=10, max=38

Table 2 – Evaluation Queries

Table 2 summarizes three representative queries selected from the experiments. We desire to find features located within ξ_{close} *=100 km* of the starting geographic feature (g_s=*Geb537*) that are considered "most related" in terms of their semantic footprint. The features in this data set have anywhere from 12 to 40 attributes (or elements) and have a variation of labels in the ontology (e.g., house, apartment, construction, warehouse, etc…).

High Overall Semantic Footprint (*Semφ)*

Query I sets the starting geographic feature at *Geb537* with 30 total attributes, and labeled as a "house". For the target features, the number of shared attributes varies considerably from 5 to 30. The ontological distance varies from zero hops (i.e., *Class_d*) for one feature and all the way to 25 for others. Figure 6 gives a visual representation of the top 10 elements in G_{final}(*Geb537*) with arrows pointing in the direction of the 10 geographic features and labels for the semantic footprint values. Interestingly, the most related geographic features are not necessarily the closest ones. In fact, Figure 6 shows that even though *Geb537* is surrounded by nearby buildings, its footprint is composed of several farther away buildings. Figure 7 shows all geographic features as indicate by the id field of Table 3.

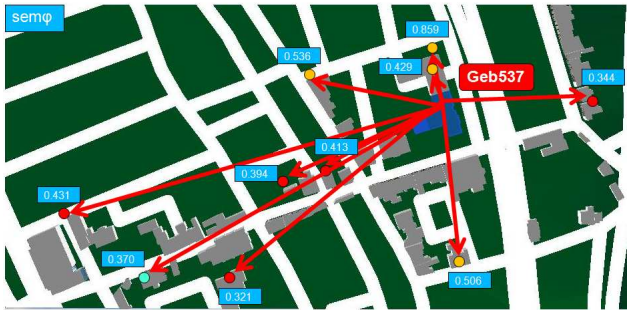

Figure 6 – Top 10 Highest Semantic Footprint Features related to *Geb537*

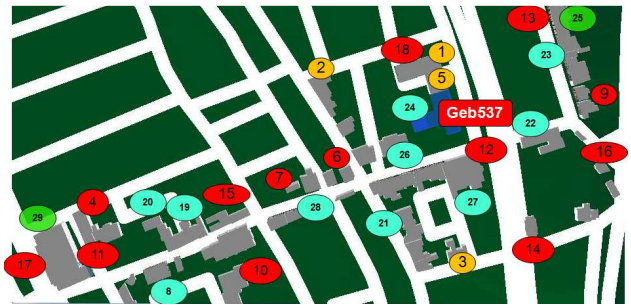

Figure 7 – Features Related to g_s=*Geb537* According to Table 3

High Dimensional Affinity (*DimÅff)*

Query II targets a more regular data set. We keep the same geographic starting point considering 20 total attributes. Of those, 10 are shared across all features. This configuration has the effect of setting an equal dimensional affinity across the data set (not shown). The ontological distance, however, can be fairly large. Elements are as close as one hop apart in the ontological hierarchy, and as far as 29 hops away. Figure 6 shows the top 10 most related elements, most of which have high dimensional affinity. In this scenario, the ontological affinity provides at best a low contribution to the semantic footprint.

High Ontological Affinity (*OntÅff)*

Still using Geb537 as g_s , *Query III* operates on features that share many attributes (i.e., high dimensional affinity on 18 shared attributes). The ontological distance, in addition, is low for most elements, varying from 10 to 38 hops. While ontological affinity is very low, the semantic footprint remains somewhat constant at ~ 0.6 since dimensional affinity is the same across the data set. Since all features are described with similar attributes, it can be inferred that such data set most likely originated from the same

provider using the same geographic standards. This is a real-world scenario, albeit possibly less common than *Query I*, where GIS often deal with a high variety of data descriptions from disparate sources.

g_s=Geb537 $f_{att}(g_s)$ = 30								
g_i	id	$\|fatt(g_i)\|$	$\|fatt(g_s) \cap fatt(g_i)\|$	$dimÅff(g_s,g_i)$	$Class\_d(g_s,g_i)$	$ontÅff(g_s,g_i)$	$Sem\ \varphi(g_s,g_i)$	Dilution (D), Hardness (H), Concentration (C), Concession (Ccs)
Geb855	1	25	23	0.719	0	1.000	0.859	C
Geb521	2	25	20	0.571	1	0.500	0.536	C
Geb592	3	35	22	0.512	1	0.500	0.506	C
Geb600	4	40	30	0.750	8	0.111	0.431	H
Geb597	5	34	22	0.524	2	0.333	0.429	C
Geb579	6	40	30	0.750	12	0.077	0.413	H
Geb645	7	40	30	0.750	25	0.038	0.394	H
Geb653	8	27	11	0.239	1	0.500	0.370	D
Geb593	9	40	24	0.522	5	0.167	0.344	H
Geb545	10	33	21	0.500	6	0.143	0.321	H
Geb877	11	37	22	0.489	6	0.143	0.316	H
Geb557	12	30	18	0.429	4	0.200	0.314	H
Geb504	13	38	23	0.511	8	0.111	0.311	H
Geb559	14	32	20	0.476	6	0.143	0.310	H
Geb595	15	39	23	0.500	8	0.111	0.306	H
Geb874	16	29	17	0.405	4	0.200	0.302	H
Geb889	17	36	22	0.500	9	0.100	0.300	H
Geb589	18	31	19	0.452	6	0.143	0.298	H
Geb875	19	26	11	0.244	2	0.333	0.289	D
Geb560	20	23	10	0.233	2	0.333	0.283	D
Geb514	21	10	7	0.212	2	0.333	0.273	D
Geb562	22	20	8	0.190	2	0.333	0.262	D
Geb540	23	22	8	0.182	2	0.333	0.258	D
Geb516	24	14	6	0.158	2	0.333	0.246	D
Geb865	25	28	13	0.289	4	0.200	0.244	Ccs
Geb532	26	16	6	0.150	2	0.333	0.242	D
Geb550	27	18	6	0.143	2	0.333	0.238	D
Geb522	28	12	5	0.135	2	0.333	0.234	D
Geb561	29	24	11	0.256	4	0.200	0.228	Ccs

Table 3 – Data results for Query I

Dilution, Hardness, Concentration, and Concession Sets

Using *Algorithm 1*, we generate Table 4 to list how variations in *DimÅff* and *OntÅff* create sets of dilution, hardness, concentration, and concession. We set both t_{dim} and t_{ont} at 0.3 to designate our minimum cutoff requirements for dimensional and ontological affinity. If the domain expert has a strict demand for both attribute and type similarity, Table 4 identifies four features in $G_{concentration}(Geb537)$ that are comprised of those characteristics. The 10 features in $G_{dilution}(Geb537)$ group elements with high ontological/low dimensional affinity, whereas the 7 features in $G_{hardness}(Geb537)$ provide the converse. Figure 8 gives a plot of the geographic features in Table 3 (only a subset of the geographic features are shown). The three cases above underscore the importance of exploratory tasks in semantic data analysis. Understanding how features compare with and complement one another promote good information extraction and knowledge discovery.

$t_{dim=0.3}$, $t_{dim=0.3}$	$G_{concentration}(g_s)$	$G_{dilution}(g_s)$	$G_{hardness}(g_s)$	$G_{concession}(g_s)$
Query I	Geb855	Geb653,Geb875	Geb600,Geb579	Geb865
	Geb521	Geb560,Geb574	Geb645,Geb593	Geb561
	Geb592	Geb562,Geb540	Geb545,Geb877	
	Geb597	Geb516,Geb532	Geb857,Geb504	
		Geb550 Geb522	Geb559,Geb874	
			Geb889,Geb589	

Table 4 – Feature sets in $G_{dim}(g_s)$ and $G_{ont}(g_s)$

Discussion

From a mathematical perspective, semantic footprint is a measure of similarity between two geographic features. But in practice, we would like to understand its qualitative aspect, i.e., how similar the features are or how related they may be according to their natural characteristics. Looking closer at *Query I* and according to *Geb537*'s semantic footprint, its most related element is *Geb855:* they share many attributes (Table 3 row 1) in addition to being the same type of feature in the ontology ("houses"). For example, their shared attributes include *appearance, rgbTexture, image, ambientIntensity, and diffuseColor*, among others. Other geographic features in Table 3 lack some of those attributes, such as *image* and *texture*, which are not populated consistently. This scenario depicts an ideal case where semantic footprint is high from both a dimensional and an ontological perspective. As the number of shared elements decreases, so does the dimensional affinity values. Rows 2-5 still maintain a high semantic footprint due to the fairly high dimensional affinity. Row 7 (*Geb645*) finds a feature much farther in the ontological space (*Class_d*=25), causing the semantic footprint to drop as compared to the previous 5. These results force the semantic footprint to fluctuate as expected and demonstrate that semantic footprint is as an effective measure of relatedness.

For geographic features with far-apart types, the behavior of the semantic footprint can have a different connotation. For instance, looking into *Geb537* and *Geb645*, the ontology indicates they are 25 hops apart. The traversal path goes through "*house→private residence→living Space→…,…→ construction→ building→ private → warehouse"*. The framework punishes the relationship between these two elements as possibly "unrelated" due to the different nature between *house* and *warehouse*. In spite of that, the semantic footprint is still kept high to reward their high number of shared attributes. The implication of this behavior reflects possible real-life scenarios whether the domain expert is looking for a *house-house* or a *house-warehouse* correlation. The semantic footprint is flexible enough to allow these adjustments to occur without dismissing one or the other as unrelated.

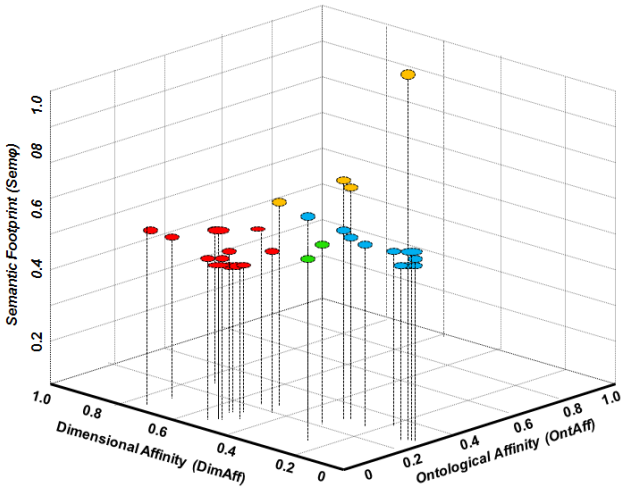

Figure 8 – Sets of Concentration, Dilution, Hardness, and Concession

In terms of density sets, the framework provides interesting insights. First, geographic features originating in the same data set tend to be highly concentrated, i.e., their semantic footprint is fairly balanced from both an attribute and ontology perspective. While this is not exactly surprising, variations in application domain often give rise to diluted and hardened sets even when the sources are the same or different, but from the same provider. We observed this behavior after processing geographic features (buildings in general) from Koenigswinter and Leverkusen. Some of the data sources come in different levels of detail which are hard to compare due to the differences in attributes, but are

common in CityGML format. In addition, attempts to relate applications of different domains (e.g., marketing and health) may easily yield concession sets, where the semantic footprint suffers significantly from a lack of common attributes and the fact that the same ontology may not always be the same for each source. In our study, we do not propose ontology merging or disambiguation, as it is outside of our scope. However, our framework still operates correctly by placing a lower premium on geographic features for which no common ontology is applied.

5. CONCLUSION

In this study, we approach spatial data analysis from an exploratory perspective. Our work proposes semantic footprints as a framework for geographic feature expansion based on three concepts: spatial, dimensional, and ontological affinity. Together these concepts reason over attributes and types to uncover the most related geographic features to a starting point. In addition, they show the dilution, concentration, hardness, and concession of the feature space. Experiments on real data sets demonstrate how semantic footprints provide useful insight into data sources and the adequacy of ontological techniques for spatial applications.

6. REFERENCES

[1] Y. Chen, T. Suel and A. Markowetz. Efficient Query Processing in Geographic Web Search Engines. In *Proc. of the ACM Int'l Conf. on Management of Data (ICMD)*, pages 277-288, Chicago, IL, USA, 2006.

[2] E. Rahm, H. Do, and S. Massmann. Matching Large XML Schemas. In *SIGMOD Record*, Vol. 33, No 4, 2004.

[3] B. Fazzinga, S. Flesca and A. Pugliese. Retrieving XML Data from Heterogeneous Sources through Vague Querying. In *ACM Trans. on Internet Technology*, Vol. 9, No. 2, May 2009.

[4] A. Islam, D. Inkpen and I. Kiringa. Applications of Corpus-Based Semantic Similarity and Word Segmentation to Database Schema Matching. In VLDB Journal Vol 17 , No 5, pages 1293-1320, 2008.

[5] W. Bae, P. Vojtechovsky, S. Alkobaisi, S. Leutenegger and S. Kim. An Interactive Framework for Raster Data Spatial Joins. *In Proc of the 15th Int'l Symp. on Adv. in Geographic Information Systems (ACM GIS)*, Seattle, USA, 2007.

[6] L. Leitao, P. Calado and M. Weis . Structure-Based Inference of XML Similarity for Fuzzy Duplicate Detection. In *Proc. of the 16th ACM Conf. on Information and Knowledge Management (CIKM)*, pages 293-302, Lisbon, Portugal, 2007.

[7] A. Doan, P. Domingos and A. Halevy. "Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach." In SIGMOD Record, Vol 30, No 2, 2001.

[8] The Open Geospatial Consortium (OGC) Web Feature Service Specification. http://www.opengeospatial.org/standards/wfs#downloads last accessed on June 2010.

[9] P. Bernstein, S. Melnik, P. Michalis and C. Quix. Industrial-Strength Schema Matching. In *SIGMOD Record*, Vol 33, No 4, 2004.

[10] G. Cong, C. Jensen and D. Wu. Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects. In *Proc. of the Conf. on Very Large Databases (VLDB)*, pages 337-348, Lyon, France, 2009.

[11] C. Beeri, Y. Kanza, E. Safra and Y. Sagiv. Object Fusion in Geographic Information Systems. In *Proc. of the 30th Int'l Conf. on Very Large Databases (VLDB)*, pages 816-827, Toronto, Canada, 2004.

[12] R. Fonseca Dos Santos, C.T. Lu., L. Sripada and Y. Kou. Advances in GML for Geospatial Applications. In *Geoinformatica Journal*. Vol 11, pages 131-157, 2007.

[13] J. Bleiholder, S. Szott, M. Herschel, F. Kaufer and F. Naumann. Subsumption and Complementation as Data Fusion Operators. In *Proc. of the Conf. on Extending Database Technology (EDBT)*, pages 513-524, Lausanne, Switzerland, 2010.

[14] E. Rahm and P. Bernstein. A Survey of Approaches to Automatic Schema Matching. In *VLDB Journal*, Vol 10, pages 334-350, 2001.

[15] V. Seghal, L. Getoor and P. Viechnicki. Entity Resolution in Geospatial Data Integration. In *Proc. of the 14th Int'l Symp. on Adv. of Geographic Information Systems (ACM GIS)*, pages 83-90, Arlington, VA, USA, 2006.

[16] S. Thakkar, C. Knoblock and J. Ambite. Quality-Driven Geospatial Data Integration. In *In Proc. of the 15th Int'l Symp. on Adv. in Geographic Informatin Systems (ACM GIS)*, Seattle, USA, 2007.

[17] W. Fan, M. Garofalakis, M. Xiong and X. Jia. Composable XML Integration Grammars. In *Proc. of the Int'l Conf. on Information and Knowledge Management (CIKM)*, Washington, D.C, USA, 2004.

[18] J. Carvalho and A. Silva. Finding Similar Indentities Among Objects from Multiple Web Sources. In *Proc. of the Int'l Workshop on Web Information and Data Management (WIDM)*, pages 90-94, New Orleans, LA, USA, 2003.

[19] M. Lieberman, J. Sperling. Augmenting Spatio-Textual Search With an Infectious Disease Ontology. In *Workshop of the Int'l Conf. on Data Engineering (ICDE)*, pages 266-269, Cancun, Mexico, 2008.

[20] S. Hwang. Using Formal Ontology for Integrated Spatial Data Mining. In *Computational Sciences and Its Applications (LNCS)*. Vol 3044, pages 1026-1035, Springer-Verlag.

[21] http://citygml.org/, last accessed in June 2010.

[22] Semantic Web for Earth and Environmental Ontology. (SWEET). http://sweet.jpl.nasa.gov/ontology/, last accessed in June 2010.

[23] A. V. Aho, J. E. Hopcroft, J. D. Ullman, "On Finding Lowest Common Ancestors in Trees." In Proc of the 5th annual ACM symposium on Theory of computing (STOC), pages 253-265, 1973.

[24] S. Mardis and J. Burger. Design for an integrated gazetteer database: Technical description and user guide for a gazetteer to support natural language processing applications. Technical report, Mitre, 2005.

[25] E. Rauch, M. Bukatin, and K. Baker. "A confidence-based framework for disambiguating geographic terms." In Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References, 2003.