A General Framework for Conflation

Benjamin Adams, Linna Li, Martin Raubal, Michael F. Goodchild

University of California, Santa Barbara, CA, USA
Email: badams@cs.ucsb.edu, linna@geog.ucsb.edu, raubal@geog.ucsb.edu, good@geog.ucsb.edu

1. Introduction

To date GIS-based conflation research has primarily been concerned with specific algorithms and tools for performing conflation on specific types of datasets. Most of these techniques focus on matching the geometry of geospatial features represented as points, polylines, and polygons. More recently the matching of the semantics of geospatial features has been identified as a key component of the conflation problem. In addition, conflation was conceived in terms of traditional GIS, but with the advent of neogeography, and in particular web-based maps with community-generated content, the representation of geospatial features on the earth has become increasingly heterogeneous in terms of geometry, semantics, data provenance, and error (Goodchild, 2007). In light of these complexities we present a general framework for conflation that aims to handle this diversity of representation.

To further motivate the need for a general framework we propose that the conflation operation is an important part of answering the question "Where am I?" The "Where am I?" question is a perennial favorite of spatial cognition, navigation, and location services researchers. For our purposes we take a very specific interpretation of this question. Namely, the answer to this question is the identification of one's location as a point on the Earth. Any location on the Earth can be represented by a set of properties. We can identify our location if and only if this set of properties used to represent our location is unique. In the case that we have the latitude/longitude property or a unique identifier of a discrete object at our location, such as an address, we can establish uniqueness of our location. In the absence of these singleton sets we can only answer this question by assembling a set of properties that are unique to our location. As these properties will often come from separate information sources a conflation operation is required. Given that conflation merges representations based on similarity thresholds and that errors exist in all geographic data sets, there exists a degree of uncertainty as to the uniqueness of our location identified from conflated data. We make no claims as to sources of the representations being conflated and in fact as to whether they are internal mental representations or external representations stored in an information system.

For an example of the "Where am I?" question take the following natural language description of a person's location that may have been sourced from a number of different places (see figure 1 for a corresponding map):

> *I am in Southern California standing on a two-lane road facing north. Directly behind me is a parking lot about 100 meters square and further back are a number of buildings. Directly in front of me I see a small airport and behind that a town and a mountain range rising above it. A highway intersects the town. I know the ocean is south of me.*

Here we have a number of geospatial feature types (e.g., buildings, airport, mountains), topological and spatial relationships (e.g., intersects, south of), attributes (e.g., two-lane), and geometry (100 m^2), which can all be used to uniquely identify the location of the questioner.

Although conflation research has focused on the geometric representation of the features, this scenario illustrates that for unstructured and heterogeneous data both geometry and semantics must be considered. Such unstructured geographic data is ubiquitous. For example, many neogeographic data sources, such as Geonames[1], define the semantics of feature type terms using natural language descriptions. Furthermore, an analogy can be drawn between the distortion in a person's (or ontology's) representation of space and location (including both geometry and semantics) and the geometric distortion in GIS data resulting from different measurement techniques.

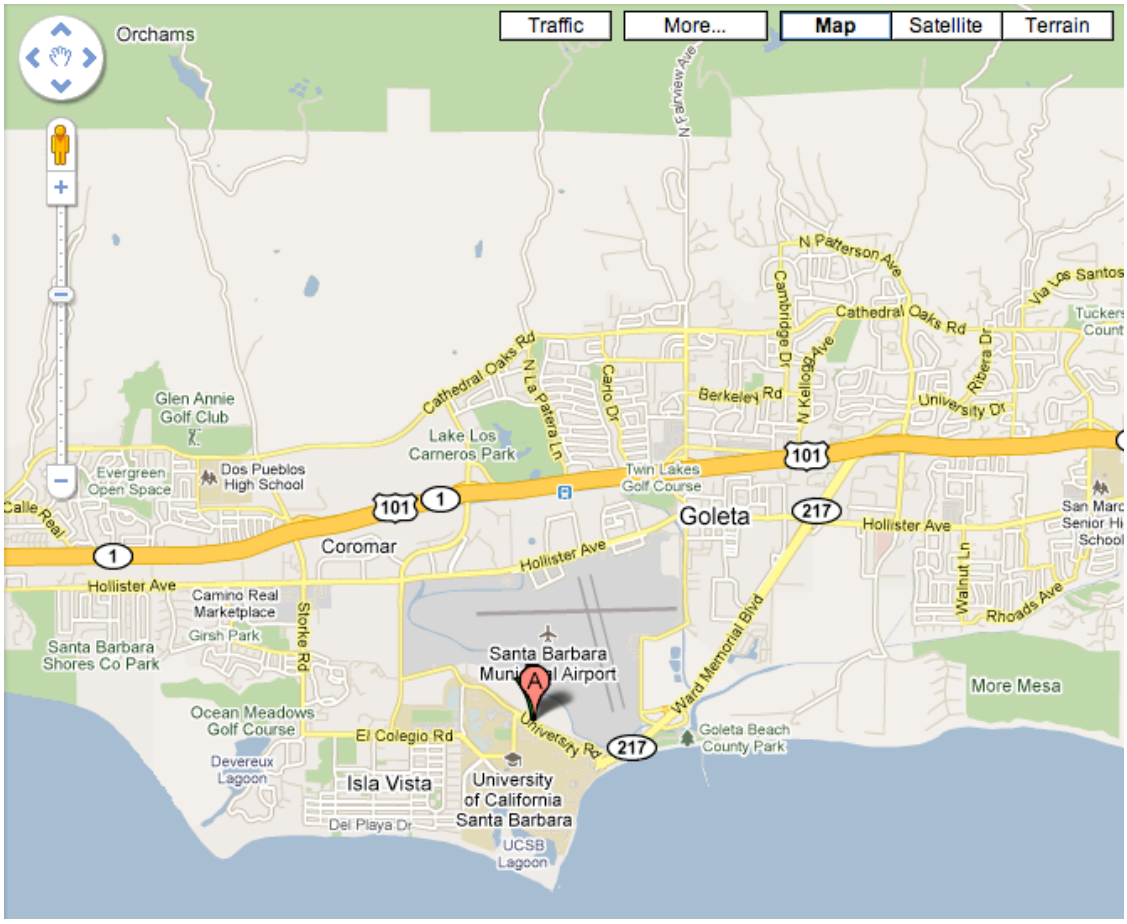

Figure 1. Map of location in scenario.

2. Geospatial Feature

Every geospatial feature on Earth can be represented by a 4-tuple $\langle G, F, A, T \rangle$ where:

G: Geometry
F: Feature type
A: Set of attributes

[1] http://www.geonames.org/

T: Set of topological relationships to other features

Arguably, F, A, and T could all be subsumed under the label semantics (and perhaps even G) (Kuhn 2003) but we find it helpful to make this separation because 1) feature types are especially salient in our conceptualizations of geospatial features (Frank 2003) and 2) topology is unique from other attributes in that it encodes relationships with other features. One might be tempted to also include a unique identifier for each feature along the lines of Ordnance Survey's Topographic Identifiers (TOIDs)[2], but epistemologically this kind of assignment is suspect. For example, is Lake Tahoe the same geospatial entity that it was before European settlers named it as such? This question also points to the fact that all geospatial features are dynamic on some time scale. Thus, we can designate different representations for the same feature at different points or intervals in time. For example, the geometric representation of the Aral Sea today looks very different from what it did in 1960.[3]

Every location can be generalized as an area containing some collection of the above tuples. A characteristic of the conflation operation is that it is invariant based on the extent of the area examined. In other words our proposed framework is designed to be universally applicable regardless of whether one is conflating feature sets with large or small footprints on the Earth.

3. Conflation is Similarity + Error Estimation

Given two collections of geospatial features, the process of conflating them generates a new collection of features along with estimations of the error propagated. The key component of any conflation algorithm is measurement of the similarity between features (Saalfeld 1988). We define feature similarity as a linear weighted sum of semantic and geometric similarities:

$$\omega_{ij} = \alpha S_{ij} + \beta D_{ij} \quad (1)$$

where S is semantic distance and D is geometric distance (distance and similarity are inversely related), and i and j are feature indices in the source datasets. The weights α and β provide a way to describe a context for conflation. We make the claim that in most cases geometry trumps semantics, because regardless of semantic similarity if two features are far apart in space they should never be conflated. Although semantic and geometric similarity measures are incommensurable, when they are both normalized to a [0..1] scale this will tend to result in $\beta >> \alpha$. However, there may be exceptions, especially when conflating highly unstructured data such as found in the "Where am I?" scenario described in the introduction.

Expanding the semantic part of equation 1 to differentiate feature type, attributes, and topology we get:

$$\omega_{ij} = \alpha_f Sf_{ij} + \alpha_a Sa_{ij} + \alpha_t St_{ij} + \beta D_{ij} \quad (2)$$

---

[2] http://www.ordnancesurvey.co.uk/oswebsite/freefun/geofacts/geo1201.html

[3] In fact, it might be represented as three different features today: the North Aral Sea and western and eastern South Aral Sea.

Note, in most cases the distance measures of the semantic components will be strongly correlated, which should be considered when assigning the α weights.

3.1 Geometric and Semantic Distortion

Uncertainty in GIS data may come from many sources. The real world is infinitely complex with lots of information; therefore generalization or simplification introduces uncertainty in the creation of a GIS database. Different people may adopt different conceptualization schemes when they create datasets about the same features. In addition, uncertainty also depends on the measurement method and the scale of measurement. Finally, uncertainty may be propagated in data processing and analysis after data are captured. Therefore, one of the key research questions in conflation is how to identify the same features in spite of inherent uncertainty in various geospatial databases. Analogously, when answering the question "Where am I?" humans conflate information presented on a map with incomplete and imperfect information in their internal representations along with implicit knowledge of how well they trust their own information to generate a better understanding of their location.

Semantic similarity measurement is an active research topic in GIScience, and several techniques have been developed (Schwering 2008). Every dataset either commits to an implicit ontology that exists in the head(s) of the schema designer(s) or an explicit ontology defined in some degree of formalization. Just as two representations of a feature will be geometrically distorted, they will also be *semantically distorted* by the ontology. In some cases there is no ground-truth to identify which of the features is more distorted from reality; there is simply a difference in definition. An example of this is the Río de la Plata in South America, which may or may not be classified as a river. However, there are cases where something close to ground-truth can be identified. For example, if the land-use of an area is represented as farmland in one database, but a raster image of the same area with the same timestamp shows concrete everywhere then it is a fair assessment to say that the first representation is highly semantically distorted from reality. In general, two representations will be semantically distorted from each other, quantified in terms of the results of semantic similarity measurement.

3.2 Temporal Effects on Similarity Measurement

The role that time plays in conflation is a largely unexamined problem. Even under "perfect" measurement conditions, all of the components, $\langle G,F,A,T \rangle$, of a representation will change over time. The rate of change, which we designate the *temporal malleability,* of a geospatial representation varies considerably depending on what is being represented. However, the similarity threshold required to identify two representations as referring to the same feature will always be indirectly related to the magnitude of their temporal difference (i.e., representations of the same feature will become less similar the further they are spread over time). Historical GIS is one application where temporal malleability is of particular importance. Furthermore, there is the issue that features can come into existence, cease to exist, join together, or separate into multiple features at different moments in time (Hornsby and Egenhofer 2000).

4. Conclusion

Today geographic data is not only found in GIS databases but also on the Web in a variety of formats. A framework for conflation must be sufficiently general to be applicable to these heterogeneous data. Here we presented a framework of conflation

that defines geospatial features as 4-tuples and the process of conflation as a linear combination of geometric and semantic similarity measures with corresponding error estimation. This lays the groundwork for developing models that incorporate both geometric and semantic distortion for matching heterogeneous data. We also examined issues related to temporality in conflation. An important next step is the formalization of semantic distortion and how it relates to existing models of geometric distortion. In addition, more work needs to be done developing adequate models for error estimation and recording for both geometric and semantic matching.

Acknowledgements

This work is supported by NGA-NURI grant HM1582-10-1-0007.

References

Frank A, 2003, Ontology for spatio-temporal Databases. In: M. Koubarakis and e. al. (Ed.), *Spatiotemporal Databases: The Chorochronos Approach*. Lecture Notes in Computer Science 2520, pp. 9-77, Springer, Berlin.

Goodchild M, 2007, Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69: 211-221.

Hornsby K and Egenhofer M, 2000, Identity-based change: a foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science,* 14(3): 207-224.

Kuhn W, 2003, Semantic Reference Systems. *International Journal of Geographical Information Science,* 17(5): 405-409.

Saalfeld, A. 1988, Conflation Automated map compilation. *International Journal of Geographical Information Systems, 2*(3), 217 - 228.

Schwering A, 2008 Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS,* 12(1): 5-29.