GIVA: A Semantic Framework for Geospatial and Temporal Data Integration, Visualization, and Analytics

Isabel F. Cruz, Venkat R. Ganesh, Claudio Caletti, Pavan Reddy
ADVIS Lab
University of Illinois at Chicago
{ifc, vsekar, ccaletti, preddy}@cs.uic.edu

ABSTRACT

The availability of a wide variety of geospatial datasets demands new mechanisms to perform their integrated analysis and visualization. In this demo paper, we describe our semantic framework, *GIVA*, for *Geospatial and temporal data Integration, Visualization, and Analytics*. Given a geographic region and a time interval, GIVA addresses the problem of accessing simultaneously several datasets and of establishing mappings between the underlying concepts and instances, using automatic methods. These methods must consider several challenges, such as those that arise from heterogeneous formats, lack of metadata, and multiple spatial and temporal data resolutions. A web interface lets users interact with a map and select datasets to be integrated, displaying as a result reports where values pertaining to different datasets are compared, analyzed, and visualized.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation*

Keywords

Spatial databases, GIS, Data integration, User interfaces

1. INTRODUCTION AND MOTIVATION

Spatio-temporal data are a fundamental resource for a variety of applications including those in public administration, transportation networks, and environmental studies. Within environmental studies, a possible scenario entails the study of two indicators: flu and precipitation to detect if they are correlated or if, for example, precipitation is a predictor of flu occurrences. Other scenarios may compare dependencies between these two indicators in two different cities. This scenario is depicted in Figure 1.

Several indicators can be studied at the same time and multiple dependencies considered in the emerging urban metabolism field [9]. To conduct these studies, vast amounts of geospatial information must be accessed and integrated using automatic methods, so that environmental scientists do not have to manually establish connections among highly heterogeneous data. We have been considering several scenarios motivated by two projects in which we collaborate, namely BURST (Building Urban Resilience and Sustainability)[1] and TerraFly [17]. Both projects are intended for experts in a variety of domains including urban metabolism and public health (BURST), hydrology and disaster mitigation (TerraFly), and transportation (BURST and TerraFly).

In this demo, we describe a semantic framework, *GIVA*, for *Geospatial and temporal data Integration, Visualization, and Analytics*. Using this framework, users can select regions in a map, specify time intervals, and select datasets to produce reports where values pertaining to different datasets are compared, analyzed, and visualized.

At the core of GIVA is its capability to deal with data, metadata, and their heterogeneity, by addressing the following issues: (1) *wide variety of formats*, both standardized (e.g., GML, KML, Shapefile, MapInfo TAB) and non-standardized (e.g., HTML tables and flat files); (2) *lack of metadata*, which stems in great part from non-standardized formats; (3) *multiple spatial and temporal resolutions*, due to different data acquisition techniques (e.g., surveys for census data and sensing methods for precipitation); (4) *different vocabularies and schemas*, which are created by diverse organizations (an example in public administration is that of land use codes [18]) and is illustrated for the two cities of Figure 1. In addition, there are overarching issues when dealing with geospatial data, namely that of uncertainty [15, 19].

2. FRAMEWORK

This section introduces our semantic framework (Figure 2) and describes briefly its components.

2.1 Data Extraction

Data of interest to geospatial information appears in a variety of formats, which we represent in the hierarchy of Figure 3. We refer to the formats approved by OGC[2] and that implement its standards as *standardized* and the rest as *non-standardized* data formats. A *geographic component* in these data formats uses geodetic systems such as WGS84 and geometric objects (e.g., polygon, polyline).

However, GIS data that are represented in web tables or text need special processing. Web tables are primarily con-

SIGSPATIAL'13, Nov 05-08 2013, Orlando, FL, USA
ACM 978-1-4503-2521-9/13/11. http://dx.doi.org/10.1145/2525314.2525324
.

[1] http://www.burst.uic.edu

[2] http://www.opengeospatial.org/standards/is

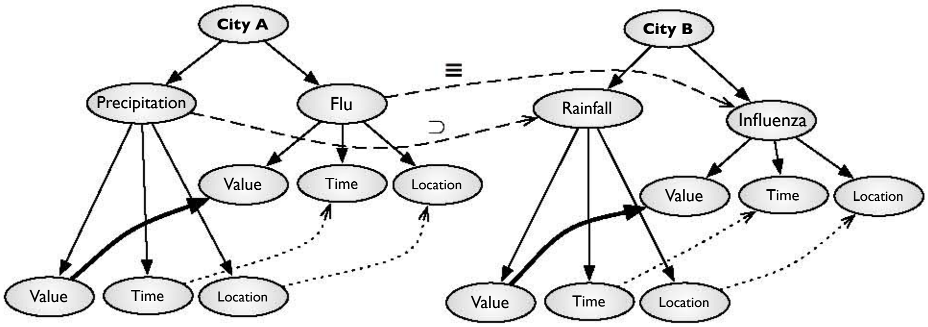


Figure 1: Comparison between two cities. Dashed edges represent concept similarity, dotted edges represent time or location similarity, and solid thick edges represent correlation between values of different concepts.

structured using the <*table*> tags for a variety of purposes such as, HTML forms, calendars, page layout, and relational data. However, in many cases web tables (even if they originate from relational databases) are not feature-rich because they do not contain clearly represented headers. The extraction of the corresponding feature-rich tables entails the identification of the headers (which are sometimes nested) and the storage of the table to produce a feature-rich table, which is stored in a structured file. For this kind of extraction we use a machine learning approach that encompasses a decision tree classifier model (C4.5) [16] using 20 different heuristics (including number of columns, rows, font size, and color) and trained it on 100 web tables with GIS data.

2.2 Data Translation

Data translation is the process of translating data from one format to another. Clean abstraction of data formats and methods to perform data translation are required for a sound solution to data integration [1]. Thus, before we attempt to create geospatial mappings between these data, they are translated into a common spatial data format. One issue is that *non-standardized* formats require semantic processing to identify the appropriate column headers that contain information about spatial coordinates and time stamps. We use string matching on the column headers and perform random sampling on the values to find pattern similarities. For instance, this ensures that an unclearly named column header (e.g., *Pos*) that contains geospatial coordinates (e.g., -85.46, 42.32) will be identified as indeed containing spatial coordinates and its name associated with a correct meaning. Further, data in *non-standardized* formats may contain implicit geographic components (e.g., Illinois). Special processing and techniques are required to identify these implicit geographic components as described in Section 2.4.2.

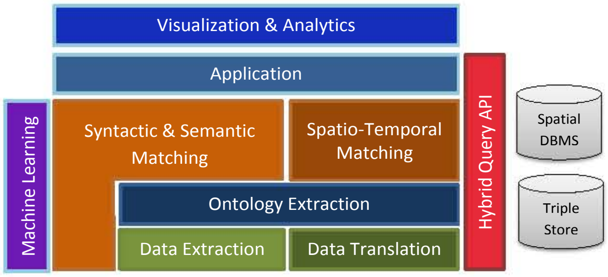


Figure 2: GIVA framework.

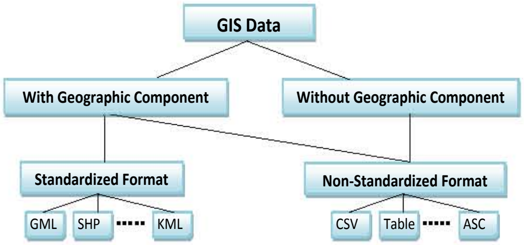


Figure 3: Hierarchy of spatial data types.

2.3 Ontology Extraction

The hierarchical characteristics of geospatial classification schemes can be modeled using a *part-of* or *is-a* relationship [4]. We have also devised methods to extract ontologies from a variety of formats, including from relational tables, XML, and RDF documents and to merge ontologies using matching and a data exchange approach by considering a global ontology [5]. This merging method is further described in Section 2.4.1 but we mention it here because it is related to recent ontology extraction approaches that use data exchange, machine learning, and user interaction [11].

2.4 Matching

The semantic integration of geospatial data requires the identification of correspondences among ontology concepts, properties, and instances, using syntactic and semantic characteristics of the ontologies, a process called *ontology matching* or *alignment*. The output of this process is a set of *mappings*. For spatial and temporal data, the spatial and temporal attributes of the data will also be considered.

2.4.1 Semantic Matching

Ontologies exhibit structural and conceptual heterogeneity, which we attribute to data creation by different organizations. The alignment of these ontologies require the sophisticated combination of various mechanisms geared to the identification of various classes of similarities. We use AgreementMaker [3], which is a proven system for ontology matching. AgreementMaker is also used for the mapping of the ontologies that are extracted from relational, XML, and RDF sources, enabling the mapping of similar concepts independently of where they appear (e.g., titles of relational tables, names of properties, or values). Data integration is achieved by rewriting a query expressed in terms of an ontology to another ontology using the established mappings [5].

AgreementMaker uses machine learning techniques to automatically change its configuration to maximize precision and recall [2].

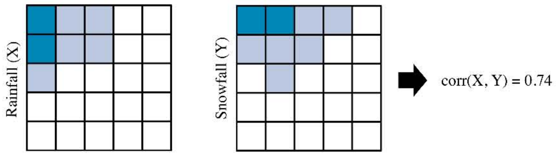


Figure 4: Comparison of values across data sets.

2.4.2 *Spatio-Temporal Matching*

Two important tasks to be addressed by this component are described below:

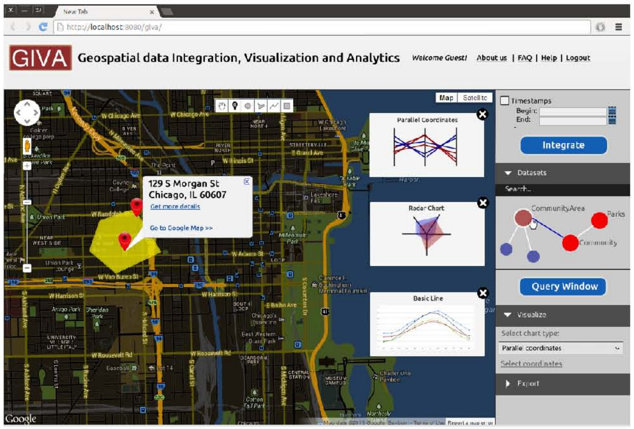
Resolving implicit geographic component. The process of assigning apposite geographic coordinates is referred to as *geocoding*, and that of identifying a geographic context is referred to as *geoparsing* [13]. For instance, geocoding helps in identifying the word *Illinois* and assigning its respective geographic component (e.g., state boundary of Illinois), if available. However, geospatial ambiguities often exist. For instance, the *Illinois river* may refer to the river in the state of Illinois or to the river of the same name in the state of Oregon. We implement geoparsing using a Named Entity Recognition (NER) technique and use semantic mappings as discussed in Section 2.4.1 for geocoding.

Managing spatial and temporal resolution. Heterogeneities in spatial and temporal resolution are introduced when data are published using different data acquisition techniques. For instance, precipitation data may be published associated with different areas depending on the density of the placement of the gages or the assumed coverage of each of them (e.g., a rectangle in a grid or a circle). We deal with this integration problem by introducing a new spatial resolution method that establishes a grid. The integration is performed by partitioning the space and computing a weighted average of the values in each of the original datasets, as illustrated in Figure 4. This produces a new dataset at a new resolution. Uncertainty increases when the dimensions of the grid are small in comparison with the measurement resolution, hence the grid dimensions can be defined depending on the dataset and the desired level of uncertainty. Temporal resolution can be resolved similarly.

This technique can be used when considering datasets about the same concept, for example *rainfall* or about different concepts, for example if the user wants to build a dataset about *precipitation* starting from two datasets about *rainfall* and *snowfall*. In this case, we can merge the datasets by adding the values of the two original datasets and by introducing an appropriate uncertainty value associated with this merging. Correlation between the datasets (instances) (see Figure 4) can assist the semantic matching process.

2.5 Storage Systems and Application

Our framework includes two different types of storage systems. A *Spatial DBMS* is used for storing and indexing geographic data and a *Triple Store* is used for handling semantic data and also to store the final alignments. A *Hybrid Query API* combines the query functionality of these two systems. An *Application* (web or stand-alone) is necessary to communicate with the other components of the framework and for the user interaction. This application also acts as Web Feature Service (WFS) interface to publish the integrated data to the domain stakeholders. For this demonstration, we develop a web application. The implementation details are described in Section 3.


http://localhost:8080/giva/

129 S Morgan St
Chicago, IL 60607
Get more details
Go to Google Map >>

Figure 5: GIVA web interface design.

2.6 Visualization and Analytics

We consider two components: one for visualization and the other one to support analytic methods.

The visualization component is fundamental to develop information processing in the context of different stages of scientific research and decision making. A use-based approach has long been proven to be an effective way to reinforce human understanding of abstract data [12]. We implement both an interactive map and plots for multidimensional visualizations, such as star plots and parallel coordinates graphs, where users will be able to display one or multiple variables simultaneously as shown in Figure 5.

The analytics component aims at providing the scientists with a suite of statistical models for spatial data exploration and multivariate analysis. We offer libraries for spatial autocorrelation and spatial regression as well as for factor analysis. In particular, we implement measures of spatial autocorrelation, such as Moran's I and Geary's C, and libraries to run OLS regression and spatial lag models. However, the analytics tool is meant to be an extensible part of the framework according to the needs of the scientists.

3. IMPLEMENTATION

We use PostGIS, a well-known spatial extension of the PostgreSQL database system, as our *Spatial DBMS* and OWLIM [10], an RDF database management system implemented in Java, as our *Triple store*. We develop a *Hybrid Query API* in Java to interlink PostGIS and SPARQL queries. The *Data Extraction* is developed using WEKA's [8] implementation of C4.5 algorithm to train the model and to extract the feature-rich web tables. The extracted tables are converted to a tab delimited file. *Data translation* is implemented in an XML framework that extends GDAL [7] to extract geospatial data with proper handling of geodetic systems. This module also implements the semantic processing techniques described in Section 2.2 to handle flat files (CSV and TSV). For *Ontology Extraction*, we use Apache OpenNLP[3] as an NLP toolkit and DBpedia[4] to receive suggestions for class names during the ontology construction. Automatic ontology extraction is a complex task and its performance depends on the organization of the schemas.

[3] http://opennlp.apache.org/

[4] http://dbpedia.org/

To overcome this issue, we allow users to optionally review the extracted ontology. The resulting RDF-Schema is used to generate triples. *Semantic Matching* is performed using AgreementMaker [3] and *Spatio-temporal Matching* uses the *Hybrid Query API* and an implementation of a matching mechanism as described in Section 2.4.2. A web interface is developed using the latest web technologies, namely AJAX and jQuery. For visualization and analytics, we use the interactive JavaScript visualization library—D3.js.[5]

4. RELATED WORK

A mobile application for an urban environment is presented by Della Valle et al. [6] to answer semantic queries such as finding the nearest tourist spots. Their data preparation module handles *Point* data from several ESRI Shapefiles, which are then manually processed and converted into an RDF format using PostGIS. These data are used along with an earlier platform that they developed, which provides SPARQL end points and a semantic framework with a reasoner to answer queries.

Urbmet[6] is an interactive map application to analyze urban data. Datasets about energy, material, and population are processed manually to provide reports for the very specific purpose of displaying potential spatial patterns that exist among them. Many similar applications can be found in OpenCityApps.[7] However, each of these applications is limited to providing visualizations or reports for pre-defined purposes and does not support data integration.

Middel presents an integrated framework for visualizing multivariate geodata [14]. The framework stores the spatial data mapped to uniform grids that cannot be changed and uses multinomial logistic regression to estimate characteristics of two different attributes for visualization. The drawbacks with this method are: (1) the possibility of a large amount of generated gridded data that could drastically reduce the performance of the system; (2) the potentially large addition of uncertainty in the partitioned grids that can impact the quality of the visualization.

In all of the systems we reviewed, there is no process that automatically integrates heterogeneous datasets. Also, the heterogeneity that is present in the data formats or metadata is either not resolved or is resolved manually.

5. CONCLUSIONS

We have introduced GIVA, a semantic framework that assists domain experts in integrating highly heterogeneous datasets and in analyzing and visualizing dependencies among them. The system supports three types of users: *administrator*, *domain expert*, and *casual user*, with different types of access. Given the complexity of the overall framework—in fact, we could not find any framework whose overall functionality can be compared in breadth with the one we propose—it is the case that every component of the framework offers opportunities for expansion and for improvement.

Acknowledgments

We would like to thank Roberto Tamassia, Matteo Palmonari, Tom Theis, Ning Ai, Sybil Derrible, Sam Dorevitch, Naphtali Rishe, and Goce Trajcevski for useful discussions. Thanks are due to Matt Dumford and Anna Anderson for help with the software development. This work was supported in part by NSF Awards CCF–1331800, IIS–1213013, IIS–1143926, and IIS–0812258 and by a UIC-IPCE Civic Engagement Research Fund Award.

References

[1] S. Abiteboul, S. Cluet, T. Milo, P. Mogilevsky, J. Siméon, and S. Zohar. Tools for Data Translation and Integration. *IEEE Data Engineering Bulletin*, 22(1):3–8, 1999.

[2] I. F. Cruz, A. Fabiani, F. Caimi, C. Stroe, and M. Palmonari. Automatic Configuration Selection Using Ontology Matching Task Profiling. In *Extended Semantic Web Conference (ESWC)*, volume 7295 of *LNCS*, pages 179–194, 2012.

[3] I. F. Cruz, F. Palandri Antonelli, and C. Stroe. AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. *PVLDB*, 2(2):1586–1589, 2009.

[4] I. F. Cruz and W. Sunna. Structural Alignment Methods with Applications to Geospatial Ontologies. *Transactions in GIS, Special Issue on Semantic Similarity Measurement and Geospatial Applications*, 12(6):683–711, 2008.

[5] I. F. Cruz and H. Xiao. Ontology Driven Data Integration in Heterogeneous Networks. In A. Tolk and L. Jain, editors, *Complex Systems in Knowledge-based Environments*, pages 75–97. Springer, 2009.

[6] E. Della Valle, I. Celino, and D. Dell'Aglio. The Experience of Realizing a Semantic Web Urban Computing Application. *Transactions in GIS*, 14(2):163–181, 2010.

[7] GDAL Development Team. *GDAL - Geospatial Data Abstraction Library, Version 1.10.0.* Open Source Geospatial Foundation, 2013.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[9] C. Kennedy, S. Pincetl, and P. Bunje. The Study of Urban Metabolism and its Applications to Urban Planning and Design. *Environmental Pollution*, 159(8):1965–1973, 2011.

[10] A. Kiryakov, D. Ognyanov, and D. Manov. OWLIM–A Pragmatic Semantic Repository for OWL. In *Web Information Systems Engineering–WISE 2005 Workshops*, pages 182–192. Springer, 2005.

[11] C. A. Knoblock, P. Szekely, J. L. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan, and P. Mallick. Semi-automatically Mapping Structured Sources into the Semantic Web. In *International Semantic Web Conference (ISWC)*, pages 375–390. Springer, 2012.

[12] A. M. MacEachren and M.-J. Kraak. Exploratory Cartographic Visualization: Advancing the Agenda. *Computers & Geosciences*, 23(4):335–343, 1997.

[13] K. S. McCurley. Geospatial Mapping and Navigation of the Web. In *International World Wide Web Conference (WWW)*, pages 221–229. ACM, 2001.

[14] A. Middel. A Framework for Visualizing Multivariate Geodata. In *Visualization of Large and Unstructured Data Sets*, pages 13–22, 2007.

[15] D. Pfoser, N. Tryfona, and C. S. Jensen. Indeterminacy and Spatiotemporal Data: Basic Definitions and Case Study. *GeoInformatica*, 9(3):211–236, 2005.

[16] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[17] N. Rishe, B. Furht, M. Adjouadi, A. Barreto, E. Cheremisina, D. Davis, O. Wolfson, N. Adam, Y. Yesha, and Y. Yesha. Geospatial Data Management with TerraFly. In *Handbook of Data Intensive Computing*, pages 637–665. Springer, 2011.

[18] N. Wiegand, D. Patterson, N. Zhou, S. Ventura, and I. F. Cruz. Querying Heterogeneous Land Use Data: Problems and Potential. In *National Conference for Digital Government Research (dg.o)*, pages 115–121, 2002.

[19] M. Worboys. Computation with Imprecise Geospatial Data. *Computers, Environment and Urban Systems*, 22(2):85–106, 1998.

[5] http://d3js.org/

[6] http://urbmet.org/about/

[7] http://opencityapps.org/