Argos: Dynamic Composition of Web Services for Goods Movement Analysis and Planning

José Luis Ambite, Genevieve Giuliano, Peter Gordon, Qisheng Pan[*],
Naqeeb Abbasi, LanLan Wang, Matthew Weathers
University of Southern California [*]Texas Southern University

1. INTRODUCTION

This Project Highlight describes Year 2 activities of our Argos research. The purpose of the research is to develop a flexible data query and analysis system based on the web services paradigm. Our application domain is metropolitan goods movement, including modeling of goods movement flows and effects of goods movement activities on urban spatial structure. The project began in August 2003. This research seeks to blend computer science and social science approaches by developing new data integration tools and applying them to social science research problems. The research has three objectives: 1) to advance computer science research by developing an expressive web services description language and techniques for dynamically composing web services, 2) to develop and conduct test applications of an intra-metropolitan goods movement flow model using web services in cooperation with government partners, and 3) to use the model to conduct social science research on intra-metropolitan economic linkages and spatial structure. Although the focus is on the specific topic of urban goods movement, the approach to web service composition is general and can be applied to other scientific data gathering and analysis tasks.

The Computer Science Research Problem Significant advances have been made in both data integration and web services, but limitations still exist. Query evaluation systems in data integration systems are composed of classical relational algebra operations, but do not include arbitrary computations as required in scientific workflows. Current web services tools require manual composition. Our approach is to combine the benefits of data integration and web services by developing an architecture for automated web service composition.

The Domain Research Problem Goods movement is becoming an increasingly important phenomenon in metropolitan areas, yet our ability to obtain information, model, and plan for urban freight flows is limited. Many limitations are related to data: 1) firm specific information is proprietary and therefore not available, 2) available data sources are incomplete, gathered in different units or time periods or geographic scale, 3) detailed information on flows on the highway network are costly to obtain and therefore scarce, 4) traditional methods for modeling freight flows depend on highly disaggregate data which can only be obtained via field survey. Our approach is to use data from widely available secondary sources to develop goods movement monitoring and modeling tools. Such an approach requires efficient computational tools that can integrate data from heterogeneous sources and process it to generate new results.

dg.o2005 Atlanta, GA, USA

2. YEAR 2 ACTIVITIES

Over the past year we have focused on three tasks: 1) developing an ontology for describing the various data elements and sources that compose the goods movement domain; 2) identifying new data sources to improve our previous goods movement workflow and model; 3) developing a method for automatic composition of computational workflows.

2.1 Ontology Development

Integrating and analyzing data from multiple sources requires assigning formal semantics. We define an ontology for this purpose [2]. The ontology describes the data items in the different sources and the data produced by the operations uniformly as multi-dimensional objects. That is, each object is defined as having values along a set of attributes or dimensions. These values are themselves objects that are organized in hierarchies, most of which have part-of semantics. The data to be described include production, consumption and flows of goods by quantity, value, type, location, time period. Our ontology has the following core dimensions: geographic entity (e.g. Los Angeles CMSA), temporal extent of measurement, commodity type according to various classifications, measurement unit, and flow type.

The ontology is parsimonious by design: we want sufficient dimensions to describe any given data element within the domain so that data requests can be answered. We used the Resource Description Framework (RDF) [4] as description language, augmented with the rule language Triple [5].

Ontology development was challenging due to the complexity of the data. For example, product data are reported in a variety of classifications, including SCTG (Standard Classification of Transported Goods), SIC (Standard Industrial Classification), NAICS (North American Industry Classification System). Each product category must be carefully described so that data from different sources can be reconciled. Location of production and consumption is reported at different levels of geography, from census tracts to "rest of the world." Spatial units are therefore described in terms of containment, e.g. Los Angeles County is part of the Los Angeles CMSA (Consolidated Metropolitan Statistical Area), which is part of California, and so on.

2.2 New Data Sources

In previous work we developed a model for estimating intra-regional freight flows on the highway network [3]. Our objective was to develop a model that was behaviorally robust, transferable across metropolitan areas, and used data from widely available sources. Tests of the model gave encouraging results [1]. However, the model included many assumptions due to lack of data, and relied on some unique data sets (one-time surveys). Models relying on such sources are not easily updated (because one-time surveys are infrequent) or transferable (because different metro areas collect different data). As part of the Argos project, we are developing a freight flow model based almost entirely on commonly available, inexpensive secondary data sources. The major research steps are: 1) utilize a regional input/output transactions table to estimate zone level[1] intra-regional commodity-specific trip attractions and trip productions; 2) estimate commodity-specific interregional and international trip attractions and productions for zones with import/export nodes (airports, seaports, etc.); 3) create a regional origin-destination matrix using estimates from the previous two steps; and 4) load the O-D matrix on a regional highway network with known passenger flows.

We use as our case study the Los Angeles CMSA. The Los Angeles CMSA includes the ports of Los Angeles and Long Beach, together the largest container port in the US. Los Angeles is also the nation's second large air cargo hub. Consequently inter-regional and international commodity flows represent a large portion of goods movement within the region. Much of our work this year has focused on the estimation of inter-regional and international trip attractions and productions (supplies and demands for commodities). Using various data sources required a method for reconciling different product classification systems. We have developed conversion matrices for the classifications used in our key data sources: 1) IMPLAN input-output transactions table (509 sector codes), 2) Commodity Flow Survey (SCTG system), 3) Waterborne Commerce of the United States (WCUS), 4) WISERTrade.

Producing the O-D matrix requires combining IMPLAN and CFS data to estimate all productions that are consumed in the region, outside the region but inside the US, and outside the US. Similarly, we must estimate all goods consumed in the region that are produced in the region, outside the region but inside the US, and outside the US. These productions and attractions must be located within the region to the zone level, must be expressed in tonnage units, and must be allocated to specific modes (truck, rail, air water). We are currently refining our work on inter-regional and international flows.

2.3 Automatic Workflow Composition

A major goal of this research is to develop methods for automatically compose computational workflows. We are developing an architecture based on expressive web service descriptions that enables services compositions to be automatically derived similarly to the way query plans are generated in data integration systems. Because the goods movement domain contains many instances of data items with hierarchical values for each dimension, we have focused on the problem of automatically generating workflows that include aggregation operations. Our initial results appear in [2].

Our composition approach includes five steps. First, we model the application domain as an RDF/S ontology. Second, we formally describe the contents of each data source by defining the objects they provide according to our ontology. Third, we model the possible computational operations. Data transformations are modeled as black boxes, defined only by input/output behavior. Some inputs and outputs can be described directly from our ontology; others require a Triple logic program, in particular aggregations operations. Fourth, we combine the ontology, sources, and operations into a Triple program that can automatically generate computational workflows in response to user requests. Finally, our system executes the workflow by translating the generated operation graph to the XML workflow language BPEL4WS. Preliminary experiments with this approach are promising [2].

3. CONCLUSIONS

We are now about half-way through this research project. We have accomplished much of what we had intended, and the collaboration between computer science and social science colleagues has been both challenging and productive. The complexities of secondary data sources required more time and effort than anticipated to understand and resolve; however we feel that the final products of this research will be enriched by the effort.

4. REFERENCES

[1] J. L. Ambite, G. Giuliano, P. Gordon, Q. Pan, and S. Bharracharjee. Integrating heterogeneous sources for better freight flow analysis and planning. In *Proceedings of the Second National Conference on Digital Government Research (dg.o 2002)*, Redondo Beach, California, 2002.

[2] J. L. Ambite and M. Weathers. Automatic composition of aggregation workflows for transportation modeling. In *Proceedings of the 6th Annual National Conference on Digital Government Research (dg.o2005)*, Atlanta, Georgia, 2005.

[3] P. Gordon and Q. Pan. Assembling and processing freight shipment data: developing a gis-based origin-destination matrix for southern california freight flows. Final report of METRANS research project 99-25, University of Southern California, 2001. http://www.metrans.org/Research/Final_Report/99-25_Final.pdf.

[4] F. Manola, E. Miller, and B. McBride. RDF primer, W3C recommendation, 10 february 2004. http://www.w3.org/TR/rdf-primer/, 2004.

[5] M. Sintek and S. Decker. TRIPLE: A query, inference, and transformation language for the semantic web. In *International Semantic Web Conference (ISWC)*, Sardinia, June 2002.

[1] The spatial unit is TAZ (Traffic Analysis Zone), similar in size to census tracts.