AFRL-SR-AR-TR-10-0308

REPORT DOCUMENTATION PAGE

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
18-06-2010	FINAL	4/01/07 - 3/31/10

4. TITLE AND SUBTITLE
InfoFuse: Interleaved Information Gathering and Reasoning for Information Fusion

5a. CONTRACT NUMBER
FA9550-07-1-0416

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Craig A. Knoblock

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Southern California / Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 North Randolph Street, Suite 325
Arlington VA 22203-1768

RSL

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The vast amount of geospatial data now available covering the entire world presents new and exciting opportunities to derive new information through information fusion. These data sources include mapping services (Google Maps, Yahoo Maps, etc.), Web 2.0 based collaborative projects (WikiMapia and OpenStreetMap), traditional geospatial data sources (raster maps, KML vector layers), and non-traditional geospatial data sources (phone books, property records, etc.). This large amount of diverse data increases the probability of encountering missing or inconsistent data and requires efficient reasoning algorithms to scale to large problem instances during information fusion. To address these issues, we have developed a geospatial fusion framework that integrates the various types of geospatial data available within a region. Our approach builds on our past work on constraint satisfaction reasoning and data access. This framework supports the ability to gather and fuse information, and uses conflict resolution strategies to disambiguate data inconsistencies. We implemented our approach into a system called InfoFuse and successfully demonstrate this approach on the real-world data for Belgrade.

15. SUBJECT TERMS
Geospatial data integration, geospatial reasoning, satellite imagery, information integration.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Craig A. Knoblock
Unclassified	Unclassified	Unclassified	Unlimited	10	19b. TELEPHONE NUMBER *(Include area code)* (310) 448-8786

Final Technical Report

InfoFuse: Interleaved Information Gathering and Reasoning for Information Fusion

USAF, Air Force Office of Scientific Research
Award Number: FA9550-07-1-0416
Period of Performance: 4/01/07 – 3/31/10

Craig A. Knoblock (PI)
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
Phone: 310-448-8786
Fax: 310-822-0751
knoblock@isi.edu

June 23, 2010

Executive Summary

The vast amount of geospatial data now available covering the entire world presents new and exciting opportunities to derive new information through information fusion. These data sources include mapping services (Google Maps, Yahoo Maps, etc.), Web 2.0 based collaborative projects (WikiMapia and OpenStreetMap), traditional geospatial data sources (raster maps, KML vector layers), and non-traditional geospatial data sources (phone books, property records, etc.). This large amount of diverse data increases the probability of encountering missing or inconsistent data and requires efficient reasoning algorithms to scale to large problem instances during information fusion. To address these issues, we have developed a geospatial fusion framework that integrates the various types of geospatial data available within a region. Our approach builds on our past work on constraint satisfaction reasoning and data access. This framework supports the ability to gather and fuse information, and uses conflict resolution strategies to disambiguate data inconsistencies. We implemented our approach into a system called InfoFuse and successfully demonstrate this approach on the real-world data for Belgrade.

Objectives of the Research Effort:

In a previous AFSOR funded project, we presented a constraint satisfaction approach to identifying the buildings shown in a satellite image by fusing imagery, road vector data, and online telephone books. The resulting fused information can then be used to augment and update a geospatial database such as a gazetteer. This previous work demonstrated that combining traditional and non-traditional data is a means to deriving information from multiple sources that is not available in any single source. The work also highlighted the benefits of fusing diverse data sources. In general, the vast amount of data now available covering the entire world provides new and exciting opportunities to derive new information through information fusion.

However, there are several key challenges that need to be addressed in order to fully realize the benefits of fusing these diverse types of sources. These data sources include mapping services (Google Maps, Yahoo Maps, etc.), Web 2.0 based collaborative projects (WikiMapia and OpenStreetMap), traditional geospatial data sources (raster maps, KML vector layers, gazetteers), and non-traditional geospatial data sources (phone books, property records, etc.). The large amount of data that can be exploited also increases the probability of encountering source failures or data inconsistencies. Therefore, it is imperative that any systems that deal with real-world data sources have the ability to deal with these potential issues. By introducing more data, we are also presented with a scaling problem. Any reasoning algorithms and conflict resolution strategies need to scale to larger problem instances and support larger numbers of data sources.

Accomplishments/New Findings

We developed a general fusion framework that integrates the various types of geospatial data available within a region. Our approach builds on our past work on constraint satisfaction reasoning and data access. This framework supports the ability to gather and fuse information, using conflict resolution strategies to disambiguate data inconsistencies. This framework supports both the integration and reasoning of heterogeneous geospatial data. The data integration tasks involve gathering the available geospatial data from a wide variety of sources, such as those listed above. The geospatial reasoning processes can infer new and useful knowledge about a region by applying various reasoning methods over the integrated data. An example of geospatial reasoning process is identifying streets and street names from raster maps. Figure 1 shows an example screenshot where a variety of data sources and reasoning capabilities have been integrated into a single integrated framework. In this figure, the fusion of the datasets and reasoning processing make it possible to identify the locations of the buildings, the names of the streets, and the businesses associated with each of the buildings.

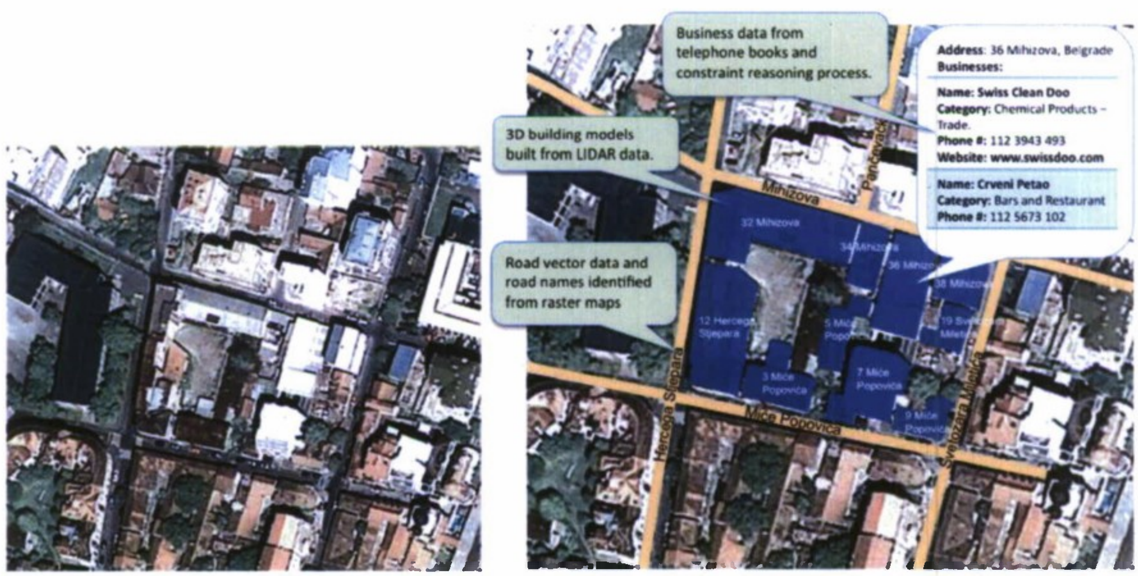

Figure 1: Area in Belgrade before and after the geospatial fusion process

The integrated framework provides a common platform for geospatial data integration and reasoning tasks. It allows the user to interactively fuse different kinds of geospatial data sources and exploit the integrated data to carry out various geospatial reasoning processes. We have also developed constraint satisfaction techniques that enable the framework to automatically infer constraint models from problem instance data and improve problem-solving performance. We now describe the accomplishments in detail.

Inferring Constraint Models from Problem Instance Data

We have shown that Constraint Programming (CP) is an effective paradigm for modeling and solving the building identification problem. However, the modeling of this problem remains an art, requiring a CP expert to specify the variables, their domains, and the set of constraints that govern a particular Constraint Satisfaction Problem (CSP). Further complicating the modeling process is the need to specialize a given constraint model for all cities throughout the world exhibiting some addressing variations. To automate the modeling process and alleviate the load placed on the human user, we developed a framework to enrich the generic constraint model by adding to it the addressing constraints that apply to a given problem instance (Michalowski et al. 2007a). These additional constraints are inferred from the input data of a problem instance.

The embedded information that we exploit is a set of instantiated variables (i.e., variable-value pairs) which we call landmark data points (i.e. buildings with known addresses). Our framework tests the features of these data points in order to select, from a library of constraints, those addressing constraints that should be added to the generic constraint model of the problem. The creation, storage, and maintenance of individual constraint models, for all cities, that account for all of the applicable addressing constraints is an unrealistic and formidable endeavor. However, the work required of the expert to define constraints that capture all of the characteristics of addressing seen to date is easier and more manageable. Moreover, combining this expert knowledge with known building addresses provided by public sources such as gazetteers allows our framework to dynamically build a constraint model of an area of interest. This constraint model plays a vital role in determining the precision of the returned solutions.

Improving Problem-Solving Performance

The benefits of an accurate model are only fully realized when the solving mechanism takes advantage of the structure and characteristics of a problem instance. To generate a precise solution, the solving component must be flexible in supporting varying problem models. To improve the performance of problem solving, the solver should exploit the structure of the problem and incorporate appropriate heuristics to reduce the explored search space. Therefore, we extended the solver we developed under a previous AFSOR grant to support the constraint models we infer. By developing a standardized representation for all problem instances, which includes the inferred constraint model, we developed an end-to-end building identification application that can identify buildings in areas larger then previously possible (Michalowski et al. 2007b). This application's architecture is shown in Figure 2. Our empirical evaluations show that the solution quality and runtime performance is greatly improved when using such an end-to-end system when compared to our previous approach.

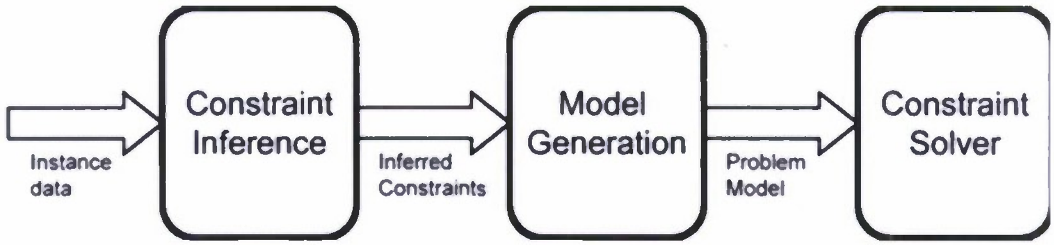

Figure 2. Building Identification Application Architecture

Framework for Geospatial Data Integration and Reasoning

Our goal is to develop a framework for integrating and reasoning about geospatial data. The various geospatial layers are integrated on top of a base layer, such as the satellite imagery for a given area. The system imports other data into the system and converts them into a uniform KML format. The reasoning processes then operate on the data layers that are available and either generate new layers or associates the results of the reasoning with the existing layers. This uniform approach to representing and reasoning about the data hides the heterogeneity present in the input data formats from the geospatial reasoning processes, thus allowing them to focus on the logic. This heterogeneity has been a major hurdle for achieving semantic interoperability of geospatial data sources. The reasoning methods are able to exploit the integrated data and present the results on a map or image using this framework.

Figure 3 shows the interface for InfoFuse, which is implemented on top of GoogleMaps. The right column shows the various operations available to import data and reason about the existing data. The system operates entirely on real-world data for the city of Belgrade. This figure shows the streets and buildings for a given region in Belgrade. At this point in the processing, the system has imported the data for each of the streets shown from the white pages and yellow pages for Belgrade. Thus, it simply has a list of the residents and business that are on a given street. The next task is to apply an information reasoning process to determine which address is associated with which building. In order to determine how to map the telephone book data to the individual buildings, the system turns the problem into a constraint satisfaction problem (CSP) [Bayer et al., 2007; Michalowski & Knoblock, 2005]. The CSP formulation (Figure 3) integrates the vector data that defines the layout of the streets in a city, the building locations along the street, the addresses obtained from online phonebooks, and the addressing patterns used in the given city.

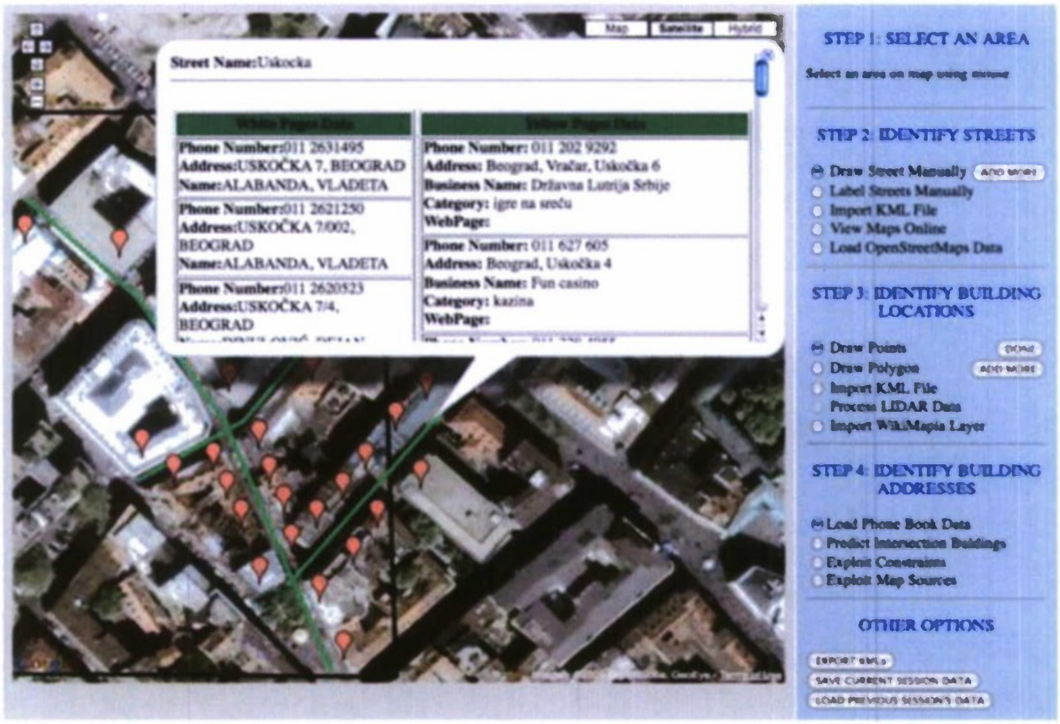

Figure 3: InfoFuse: A system for integrating and reasoning about geospatial data.

The reasoning task consists of various integration and geospatial reasoning steps. In InfoFuse, we focused on a specific integration task to solve the real-world problem of identifying buildings in imagery. The main steps involved in this task are identifying streets in an image, identifying building locations, identifying building addresses and linking business data. InfoFuse provides several alternative operations through the interface to carry out these tasks. The user can identify streets by automatically loading the OpenStreetMap data using a software wrapper, import an existing road vector layer or interactively drawing the road line. Figure 4 shows an example of streets identification for a selected area.

Figure 4: Streets identified (green lines) with OpenStreetMap data.

The user can identify the building locations using similar operations available in InfoFuse. For example, it can be done by manually drawing points or polygons (Figure 5) to represent the buildings, loading in an existing KML layers for the building locations, or extracting data from another source, such as WikiMapia.

Figure 5: Building locations manually identified as polygons.

InfoFuse gathers current information about people and businesses for a region by executing the wrappers over Yellow Pages and White Pages website. InfoFuse then links the extracted data with the road vector data and makes it available for viewing. Figure 6 shows the businesses listing and phonebook data in the popup for the street Uskočka of Belgrade City. The CSP reasoner combines the road vector data, the building location data, and the phone book data in a reasoning process to map the addresses to the individual buildings. This reasoning processs takes the phone book data associated with the roads vectors, performs the reasoning over data, and links the resulting data to the individual buildings. [Bayer et al., 2007; Michalowski & Knoblock, 2005]. Instances of building variables that are mapped to a single address are depicted with green placemarks and instances mapped to multiple addresses are depicted with red placemarks. The ambiguity of multiple possible addresses mapped to a single location is due to the uncertainty that may be present in the input data, such as missing addresses in the phonebook.

List of Personnel Associated with the Research Effort

Craig Knoblock, PI
Maria Muslea, Research Scientist

Figure 6: InfoFuse displays the resulting mapping.

Martin Michalowski, Graduate Research Assistant
Matthew Michelson, Graduate Research Assistant
Shubham Gupta, Graduate Research Assistant
Pedro Szekely, Research Assistant Professor
Rattapoom Tuchinda, Graduate Research Asistant
Berthe Choueiry, Associate Professor, University of Nebraska
Ken Bayer, Research Assistant, University of Nebraska

Publications

Greg Barish and Craig A. Knoblock, Speculative plan execution for information gathering, Artificial Intelligence, 172(4-5):413--453, 2008. http://dx.doi.org/10.1016/j.artint.2007.08.002.

Kenneth M. Bayer, Reformulating Constraint Satisfaction Problems with Application to Geospatial Reasoning, Master's Thesis, Department of Computer Science, University of Nebraska at Lincoln, August, 2007

Kenneth M. Bayer, Martin Michalowski, Berthe Y. Choueiry and Craig A. Knoblock. Reformulating CSPs for Scalability with Application to Geospatial Reasoning, In Proceedings of the 13th International Conference on Principles and Practice of Constraint Programming (CP-07), 2007

Kenneth M. Bayer, Martin Michalowski, Berthe Y. Choueiry, and Craig A. Knoblock. Reformulating Constraint Satisfaction Problems to Improve Scalability, In Proceedings of the 7th Symposium on Abstraction, Reformulation and Approximation (SARA-07), 2007

Jim Blythe, Dipsy Kapoor, Craig~A. Knoblock, Kristina Lerman, and Steven Minton, Information integration for the masses, Journal of Universal Computer Science, 14(11):1811--1837, 2008.

Shubham Gupta and Craig A. Knoblock, A Framework for Integrating and Reasoning about Geospatial Data, Extended Abstract, In proceedings of the 6th International Conference on GIScience, 2010.

Zachary G. Ives, Craig A. Knoblock, Steven Minton, Marie Jacob, Partha Pratim Talukdar, Rattapoom Tuchinda, Jose Luis Ambite, Maria Muslea, and Cenk Gazen., Interactive data integration through smart copy & paste, In Fourth Biennial Conference on Innovative Data Systems Research (CIDR), Pacific Grove, CA, January 2009

Martin Michalowski, Craig A. Knoblock, Berthe Y. Choueiry and Kenneth M. Bayer. Exploiting Automatically Inferred Constraint-Models for Building Identification in Satellite Imagery,
In Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems (ACM GIS-07), 2007

Martin Michalowski, Craig A. Knoblock and Berthe Y. Choueiry, Exploiting Problem Data to Enrich Models of Constraint Problems, In Proceedings of 6th International Workshop On Constraint Modeling and Reformulation (ModRef07), 2007

Martin Michalowski, Craig A. Knoblock, and Berthe Y. Choueiry, Reformulating Constraint Models Using Input Data, In Proceedings of the 7th Symposium on Abstraction, Reformulation and Approximation (SARA-07), Research Summary, 2007

Martin Michalowski, 2008, A General Approach to Using Problem Instance Data for Model Refinement in Constraint Satisfaction Problems, Ph.D. thesis, Department of Computer Science, University of Southern California, 2008

Matthew Michelson and Craig A. Knoblock, Unsupervised Information Extraction from Unstructured, Ungrammatical Data Sources on the World Wide Web, International Journal of Document Analysis and Recognition (IJDAR), Special Issue on Analytics for Noisy Unstructured Text Data, 10(3-4):211-226, 2007.

Matthew Michelson and Craig A. Knoblock, Creating Relational Data from Unstructured and Ungrammatical Data Sources, Journal of Artificial Intelligence Research (JAIR), 31:543-590, 2008.

Matthew Michelson and Craig A. Knoblock, Exploiting Background Knowledge to Build Reference Sets for Information Extraction, In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-2009), Pasadena, CA.

Matthew Michelson and Craig A. Knoblock, Mining the Heterogeneous Transformations Between Data Sources to Aid Record Linkage, In Proceedings of the International Conference on Artificial Intelligence (ICAI), 2009.

Matthew Michelson and Craig A. Knoblock, Constructing Reference Sets from Unstructured, Ungrammatical Text, Journal of Artificial Intelligence Research (JAIR), 38, p.189-221, 2010.

Rattapoom Tuchinda, Pedro Szekely, and Craig A. Knoblock, Building mashups by example, In Proceedings of the 2008 International Conference on Intelligent User Interface, January 2008.

Discoveries/Inventions/Patent Disclosures

Invention Disclosure: A Reference-Set Approach to Information Extraction from Unstructured, Ungrammatical Text

Invention Disclosure: Building mashups using the programming-by-demonstration approach