Rapid Integration of Online and Geospatial Data Sources for Knowledge Discovery

Article

3 authors:

Craig A. Knoblock
University of Southern California
351 PUBLICATIONS **10,746** CITATIONS
SEE PROFILE

Cyrus Shahabi
University of Southern California
361 PUBLICATIONS **6,028** CITATIONS
SEE PROFILE

John P Wilson
University of Southern California
148 PUBLICATIONS **2,710** CITATIONS
SEE PROFILE

Rapid Integration of Online and Geospatial Data Sources for Knowledge Discovery

Craig Knoblock, Cyrus Shahabi and John Wilson
University of Southern California
[knoblock,shahabi,jwilson]@usc.edu

Much of the work on information integration has focused on the dynamic integration of structured data sources, such as databases or XML data. With the more complex geospatial data types, such as imagery, maps, and vector data, researchers have focused on the integration of specific types of information, such as placing points or vectors on maps, but much of this integration is only partially automated. With the huge amount of geospatial data now available and the enormous amount of data available on the Web, there is a terrific opportunity to exploit the integration of online sources with geospatial sources for knowledge discovery. The challenge is that the dynamic integration of online data and geospatial data is beyond the state of the art of existing integration systems.

There are two general challenges that must be addressed in order to fully exploit the combination of these different types of sources. First, automated techniques are needed to integrate the diverse source types. For example, integrating maps with imagery or online schedules with road or rail vectors are needed in order to mine the information available by integrating these source types. Second, given the ability to integrate these diverse types of sources, general integration and visualization frameworks are needed to rapidly assemble these sources to support knowledge discovery. For example, one might want a mediator that can support ad hoc queries that require dynamically integrating geospatial and online data sources. Or one might want a more specialized integration framework that supports the integration of specific types of sources to support a given knowledge discovery task.

To illustrate the importance of the integration of online and geospatial data sources, consider the inadvertent bombing of the Chinese Embassy in Belgrade. On 7 May 1999, B-2 bombers dropped 5 GPS-guided bombs on what had been incorrectly identified as the headquarters of the Yugoslav Federal Directorate for Supply and Procurement (FDSP). An intelligence analyst had correctly determined that the address of the FDSP headquarters was Bulevar Umetnosti 2, but the analyst then used a flawed procedure to identify the geographic coordinates of that address. The results were tragic, especially in light of the fact that the data was available in the telephone book to determine that the target was in fact the Chinese Embassy and not the FDSP headquarters (Pickering 1999). Using sources available today, the telephone book for Belgrade, which is available online, could be superimposed on an image of Belgrade to determine the likely identity of the buildings in an image. Unfortunately, a system to automate this task does not exist today.

Consider some other examples of how the integration of these various types of sources can be exploited for knowledge discovery. Online news reports of terrorist events could be superimposed on a map and organized by time to look for patterns in activities. Online schedules can be integrated with transportation vector data to make predictions about the locations of trains, buses, or ferries. Detailed maps can be integrated with high-resolution satellite imagery to automatically determine the names of the roads in an image, which are typically not available in the road vector data available from NIMA. There are many other ways the integration of these different types of sources could be exploited. But the point is that there is no way we could even anticipate all the possible ways that this information could be combined. Thus, what is needed are tools that support an analyst in the rapid, dynamic, and accurate integration of these various types of sources in order to mine the available data.

In the remainder of this paper we describe some of our initial efforts on geospatial data integration, which illustrates the types of integration that are possible.

Mediators, Wrappers, and Geospatial Data

In previous research, we have developed two large-scale mediator systems, SIMS (Arens et al. 1996) and Ariadne (Knoblock et al. 2001). SIMS, our original system, integrates traditional databases and programs. Ariadne extends the SIMS architecture so that web-based sources can be accessed as well. The general mediator framework enables multiple, heterogeneous information sources — including

databases, programs, and web sites — to be linked together and queried as if they were a single, virtual database.

An important part of this work developed machine learning techniques for rapidly converting online Web sources into databases (Knoblock et al. 2000). These techniques greatly simplify the problem of turning web pages into structured data. The user provides a few examples of the information to be extracted and the system learns a set of extraction rules that can be used in either real-time or offline to extract the required information. This work has been patented and licensed to a USC spinoff company, Fetch Technologies (www.fetch.com).

We have also developed an integration framework called Heracles (Knoblock et al. 2001; Knoblock et al. 2001), that extracts and organizes a wide variety of information into a single, easy-to-access package. Heracles combines information from disparate data sources and displays it in a single integrated framework. We built an application of Heracles called the WorldInfo Assistant (Knoblock et al. 2001), that combines a variety of online data about countries: news, weather, airports, economic, and political information, imagery, and maps. This application exploits a wide variety of geospatial data, including image, map, vector, point, and elevation data. On our own servers we maintain roughly 2 terabytes of geospatial data that covers most of the world. Figure 1 shows example map, vector, and elevation data. The focus of the WorldInfo Assistant is on access to different types of information, but not specifically on the integration of the geospatial data.

Figure 1: Image, Map, Vector, and Elevation Data Displayed in Heracles

Integrating Vector Data and Imagery

In a recent study (Chen et al. 2003), we focused on the problem of accurate integration of geospatial vector data with (satellite or aerial) images. One application for such integration could be for the purpose of automatic recognition and annotation of spatial objects in imagery. We utilized a wide variety of geospatial and textual data available on the Internet in order to efficiently and accurately identify objects in the satellite imagery. To demonstrate the utility of our technique, we built an application that utilizes the satellite imagery from the Microsoft TerraService and the Tigerline vector files from US Census Bureau (as well as some online sources) to annotate buildings on the imagery.

Our main challenge is that geospatial data (specifically, vector and image data) obtained from various data sources may have different projections, different accuracy levels, and different inconsistencies. The applications that integrate information from various geospatial data sources must be able to overcome these inconsistencies accurately, in real-time and for large regions. Traditionally, this problem has been in the domain of the image processing and GIS systems. However, the *conflation* approach (Saalfeld 1993) used in various GIS systems to manually or semi-automatically align two geo-spatial data sets does not scale up to large regions. Image processing techniques to identify objects in the image in order to resolve vector-image inconsistencies require significant CPU time and may result in inaccurate results.

To explain our approach, we first need to explain the conflation process. The conflation process divides into following tasks: (1) find a set of conjugate point pairs, termed “control point pairs”, in both vector and image datasets, (2) filter control point pairs, and (3) utilize algorithms, such as triangulation and rubber-sheeting, to align the rest of the points and lines in two datasets using the control point pairs. Traditionally, human input has been essential to find control point pairs and/or filter control points. Instead, we developed completely automatic techniques to find control point pairs in both datasets and

designed novel filtering techniques to remove inaccurate control points. We developed two different techniques to find accurate control point pairs. Our first technique generates control points using localized image processing. The second technique finds control points by querying information from online web sources. Due to lack of space, we only briefly describe the first technique, which relies only on the imagery and vector data for accurate integration. We find feature points, such as the road intersection points, from the vector dataset. For each intersection point, we perform image processing in a small area around the intersection point to find the corresponding point in the satellite image. The running time for this approach is dramatically lower than traditional image processing techniques due to localized image processing. Furthermore, the road directions information makes detecting edges in the image much easier problem, thus reducing the running time even more.

a) Before Conflation

b) Using control point pairs to for image processing

c) After Conflation

Figure 2: Automatic conflation of vector data with imagery

Integrating Maps and Imagery

In a follow-up study (Chen et al. 2003b) to our automatic vector to image conflation, we developed efficient techniques to the even more challenging problem of automatically conflating maps with satellite imagery. There is a wide variety of geo-spatial data available on the Internet that provides satellite imagery and maps of various regions. The National Map, MapQuest, University of Texas Map Library, Microsoft TerraService, and Space Imaging are good examples of map or satellite imagery repositories. In addition, a wide variety of maps are available from various government agencies, such as property survey maps and maps of oil and natural gas fields. Satellite imagery and aerial photography have been utilized to enhance real estate listings, various military targeting applications, and other applications. Again, by integrating these spatial datasets, one can support a rich set of knowledge discovery queries that could not have been answered given any of these datasets in isolation. For example, when you are looking for a park in a neighborhood, the satellite imagery may provide you better view of the park, while the map is essential to see the surrounding streets and how to get to the park. However, accurately integrating maps and imagery from different data sources remains a challenging task. This is because spatial data obtained from various data sources may have different projections and different accuracy levels. If the geographic projections of these datasets are known, then they can be converted to the same geographic projections. However, the geographic projection for a wide variety of geo-spatial data available on the Internet is not known. The fact that many of the online maps sources do not provide the geo-coordinates of the maps makes the integration even more complicated.

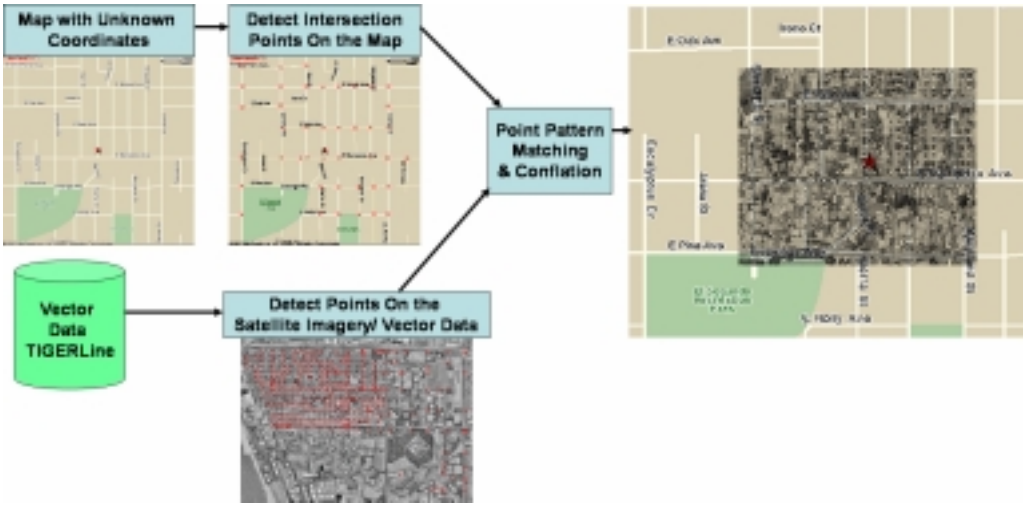

Figure3. Align imagery with maps

An overview of our approach is illustrated in Figure 3. Basically, we use common vector datasets as "glue" to integrate imagery with maps. In particular, our approach utilizes the road intersection points automatically identified on imagery and maps (whose geo-coordinate is unknown in advance), and applies a specialized point matching algorithm to compute the alignment between the two point sets. Experimental results on the city of El Segundo demonstrate that our approach lead to remarkably accurate alignments of maps and satellite imagery. The aligned map and satellite imagery can then be used to make inferences that could not have been made from the map or imagery alone. Figure 4 show an example of our results on the city of El Segundo.

Figure4 MapQuest map to imagery conflation (semi-transparent image)

Integrating Online Schedules (Moving Objects) with Vectors

In our prior work (Shahabi et al. 2001; Shahabi et al. 2002), we investigated challenges in efficient support of queries on moving objects (e.g., trains and cars). Here, we discuss one of the query types, which focuses on the integration of online schedules with vectors. This is an example of the integration of temporal data (train schedules) with spatial data (train tracks and stations) and how the combination allows answering questions which could have not been answered on the sources individually.

In (Shahabi et al. 2001), we show how a temporal source, e.g., a website providing train schedule information, can be integrated efficiently with a spatial source that contains railroad vector data. In particular, we study efficient execution of spatio-temporal range queries on the integrated sources. A spatio-temporal range query imposes bounds on spatial and temporal attributes and asks for all tuples satisfying the constraints. For example, given a geographical area (e.g., bounded by a rectangle) and a time interval, we would like to find all the trains that would be in that area in the given time-interval. The GUI of this application is shown in Figure 5.

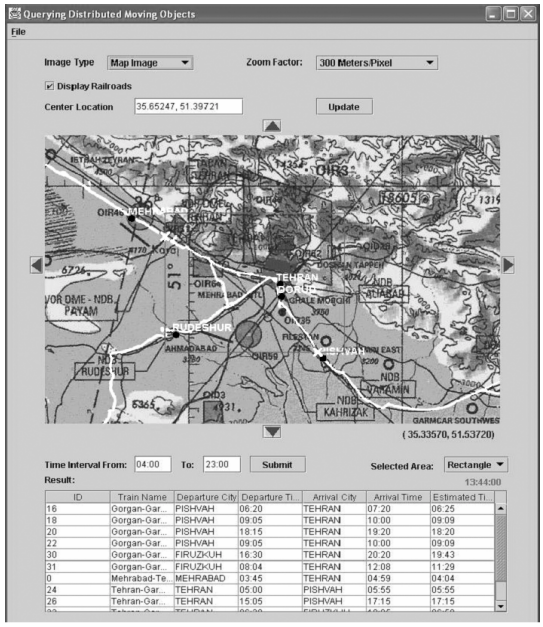

Figure 5 Integration of train schedules with vector data and maps

Evaluation of spatio-temporal range queries on distributed sources is time consuming because of the complex computational geometry functions (e.g., the shortest path function) that need to be executed on large volume of vector data as well as the temporal intersections that need to be applied among large sets of time intervals. One solution to reduce the query processing time of spatio-temporal range queries is to pre-compute the required information and materialize it using a moving object data model such as the 3D Trajectory model (Vazirgiannis et al. 2001). This is a feasible approach if we assume that different schedules, railroads, and stations information are all local and over which we have full control. However, with our assumed distributed environment, the sources of information that we would like to access are autonomous and dynamic.

Therefore, we investigated alternative distributed query plans to realize the integration of spatial and temporal information from distributed, heterogeneous web sources. We introduced a novel spatio-temporal filter (termed deviation filter), which can exploit the spatial and temporal characteristics of the data simultaneously to improve the selectivity.

References

Arens, Y., C. A. Knoblock and W.-M. Shen (1996). "Query Reformulation for Dynamic Information Integration." Journal of Intelligent Information Systems, Special Issue on Intelligent Information Integration **6**(2/3): 99--130.

Chen, C.-C., S. Thakkar, C. Knoblock and C. Shahabi (2003a). An Information Integration Approach to Automatically Annotate Spatial Objects in Satellite Imagery. The 8th International Symposium on Spatial and Temporal Databases (SSTD'03) , Santorini island, Greece, July, 2003.

Chen, C.-C., C. Knoblock , C. Shahabi and S. Thakkar (2003b). Automatically and Accurately Conflating Satellite Imagery and Maps, International Workshop on Next Generation Geospatial Information (NG2I'03), Cambridge (Boston), Massachusetts, USA, October, 2003

Knoblock, C. A., J. L. Ambite, S. Minton, C. Shahabi, M. Kolahdouzan, M. Muslea, J. Oh and S. Thakkar (2001). Integrating the World: {T}he {WorldInfo} Assistant. Proceedings of the 2001 International Conference on Artificial Intelligence {(IC}-{AI} 2001). Las Vegas, NV.

Knoblock, C. A., K. Lerman, S. Minton and I. Muslea. (2000). "Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach." IEEE Data Engineering Bulletin **23**(4).

Knoblock, C. A., S. Minton, J. L. Ambite, N. Ashish, I. Muslea, A. G. Philpot and S. Tejada (2001). "The ARIADNE Approach to Web-based Information Integration." International Journal of Cooperative Information Systems (IJCIS), Special Issue on Intelligent Information Agents: Theory and Applications **10**(1/2): 145-169.

Knoblock, C. A., S. Minton, J. L. Ambite, M. Muslea, J. Oh and M. Frank (2001). Mixed-Initiative, Multi-Source Information Assistants. The Tenth International World Wide Web Conference {(WWW10)}.

Pickering, T. (1999). Oral Presentation by Under Secretary of State Thomas Pickering on June 17, 1999 to the Chinese Government Regarding the Accidental Bombing of the PRC Embassy in Belgrade.

Saalfeld, A. (1993). Conflation: Automated Map Compilation. Computer Vision Laboratory, Center for Automation Research, University of Maryland.

Shahabi, C., M. Kolahdouzan and M. Sharifzadeh (2002). A Road Network Embedding Technique for K-Nearest Neighbor Search in Moving Object Databases. the 10th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS'02), McLean, VA.

Shahabi, C., M. Kolahdouzan, S. Thakkar, J. L. Ambite and C. A. Knoblock (2001). Efficiently Querying Moving Objects with Pre-defined Paths in a Distributed Environment. The Ninth ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS), Atlanta, Georgia, U.S.A.

Vazirgiannis, M. and O. Wolfson (2001). A Spatiotemporal Model and Language for Moving Objects on Road Networks. Proceedings of Symposium on Spatial and Temporal Databases (SSTD).