Towards Semantic Integration of Legacy Databases for Homeland Security

Terry L. Janssen, Ph.D.

Lockheed Martin
12355 Sunrise Valley Dr.
Reston, VA 20191
terry.janssen@lmco.com

Abstract

Sharing information between diverse heterogeneous databases and users has become one of the common goals of ontology-based systems (Musen, 1992; Gruber, 1993). Several commercial endeavors have been successful, such as Yahoo! with taxonomies for categorizing websites and Amazon.com for categorizing products, respectively. However, large legacy databases raise many challenges. The data base schemas are often poorly designed and the metadata is poorly documented. Rarely is there a standard vocabulary for describing entities. Ontology can be developed for databases, but mappings between legacy databases remains one of the grand challenges. This paper addresses some of these issues in context of homeland security.

Introduction

Geospatial and other types of intelligence data is being collected at unprecedented rates and stored in a variety of databases. Geospatial-intelligence databases primarily contain images and maps. Together with other types of intelligence data and text they provide a wealth of information. However, these databases are too voluminous, and too many, to be fully exploited without an efficient way to find, understand and mine the data. Furthermore there are many data mining and discovery tools and techniques (Fayyad, et al 1996; Westphal and Blaxton, 1998). It is often difficult to know which data are relevant and appropriate to mining and discovery for the objectives at hand. For example, data mining and discovery tools make it possible to identify recurring patterns, predict events and generate potentially interesting hypotheses.

Traditionally data warehouses have been developed to support data mining and discovery functions, and they typically use entity-relationship (E-R) models. However, these models do not include most of the semantic information beyond the data element definitions and relationships. Database schemas that represent those E-R models tend to be brittle, i.e., do not evolve automatically for data that do not fit into the defined schema. Furthermore, data warehouses have primarily supported statistical approaches that do not fully exploit the information contained within the data. This is evident for one-of-a-kind events, as is often the case with terrorist activities, possibly induced using ontologies and logic but not by statistic approaches.

Applications like homeland security require a variety of mining and discovery techniques, and a mining and discovery environment that can readily accommodate them using a variety of data such as human intelligence reports (text) and geospatial-intelligence contained in images and maps. Homeland security requires discovery of deep relationships between people, places, things and events, over time and space. This challenge requires new innovations in data mining and discovery, and environments that are more conducive to discovery. A large number of heterogeneous databases need to come together more effectively, with many data mining and discovery tools, so that users can focus on mining and discovery rather than lower level technical hurdles, while at the same time protecting privacy and security of the data.

Approach

Ontologies have been emerging for the last couple of decades as a knowledge representation (Staab, 2004), and the adoption of the Web Ontology Language (OWL) by the W3C is making them more widely accepted and used (see www.w3c.org/2004/owl). For deep discovery, ontologies with logic engines appear to be shining stars of the future. This is especially the case if knowledge layers on top of legacy databases can be achieved. Evidence of this can be found in proof-of-concept demonstrations, such as produced by Cyc Corp (www.cyc.com).

A knowledge layer is a means of semantically linking diverse heterogeneous databases and users. A knowledge layer is a means of representing entities, facts and events contained in databases and text, and can be overlaid on top of multiple databases. An ontology-based knowledge layer represents entities, facts, events and their relationships, and can provide capabilities such as semantic query. The purpose is not to replace databases and data warehouses. Clearly data mining applications like OLAP and statistical analysis software work well on relational databases, i.e., data stored in rows and columns (tabular data). But even here, a knowledge layer implemented with ontology has the potential to benefit the user in use of these tools. For example, semantic search of metadata might raise a user's awareness of potential data sources and their characteristics.

Knowledge layers on multiple, large scale databases and text is generally hard to achieve. Recent research is making progress, such as the integration of images, buildings and related information using semantic web technology (Michalowski, et al 2004). However poorly designed data models, and poor documentation, makes semantic integration efforts costly and high risk. The labor required for establishing common data element definitions and relating them to well structured ontology has tended to be beyond an organization's resources and political will. Partial alleviation has been achieved by semi-automatic extraction of facts from text and representation in ontology. Commercially available software products provide parts of the solution. Products integrated together provide a larger, but still limited, capability. For example, AeroText™ can quite accurately extract facts and events from text (www.aerotext.com) and those facts can be stored in a scalable knowledge server like Tucana™ (www.tucanatech.com). These integration efforts are starting to make headway toward knowledge layers for large scale, operational databases, but may questions remain unanswered. For example, how can geospatial-intelligence data be included within a knowledge layer, since it mostly contains images and maps? There is research on automated extraction, but with today's technology this primarily requires semantic markup of images and maps, e.g., RDF and OWL for semantic integration. These and many other issues need to be explored.

An ontology with logic engine provides several capabilities, such as:

- Semantic browsing and search, e.g. search by concepts;
- Semantic linking, e.g., inference of deep complex relationships;
- Learning and inferences that inform other inferences.

Semantic linking facilitates link analysis and visualizations of data and relationships, to many degrees of separation. For example, entities and relationships can visualized from multiple perspectives based on views from the ontology. The commercial product VisualLinks™ (www.visualanalytics.com) is a data mining tool that assists the user in "walking" databases to find relationships based on associated values . The semantic integration of data across multiple, heterogeneous databases will reduce the level of effort required to performing link analyses, and it will most likely improve accuracy. It is not uncommon for analysts to have to access tens or hundreds of different databases with unique data models and in many cases poor documentation. The bottom line is that data is hard to access and understand, and this in general hinders the fluent use of data mining and discovery tools.

There are many possibilities with semantic integration via ontologies, e.g.:

- Ontologies are less brittle than data models rigidly contained within databases, and ontologies evolve more naturally as new entities and relationships are added;
- Semantic integration facilitates *semantic search*, e.g., users can locate data sources more readily and accurately than traditional search engines like Google;
- Semantic integration provides more capability than the data warehouse tool suites have traditionally offered (e.g. on-line analytical tools, OLAP, and tools limited to multidimensional summary views of data);
- Data represented within an ontology is *semantically linked* which facilitates data mining approaches such as link analysis and inference;
- Ontologies, especially those based on first or higher order logic, are a more expressive than the entity-relationship representations within data models;
- Ontologies with logic engines facilitate knowledge discovery from deep inferencing, logical induction of hypotheses, evidential reasoning, and other capabilities missing from traditional data warehouses (Barwise, et al 2002).

Issues

Several issues need to be addressed, e.g.:

- Large legacy databases tend to contain "noisy" data, such as typos or different spelling of the same address, or phonetically similar spellings for the same name; means of reducing this error, such fuzzy matching and inferring "sameness" from well developed ontologies, needs to be advanced before semantic integration efforts can be more fully realized;
- What user-interface designs will facilitate acceptance and use by end users? Users, like intelligence analysts, need easy to understand and use systems, such as the capability to adding,

retrieving, changing and deleting facts (<predicate, subject, object>) in knowledge servers with huge numbers of facts and many contexts;
- Do users find existing means of visualizing knowledge and data to be adequate, or are new visualization techniques needed before users can effectively maintain and use huge knowledge servers? Various kinds of visualizations used with knowledge need to be tested with actual end-users.

Semantic integration in huge legacy database systems is non-trivial and appears to be extremely difficult to achieve without semantic enablement. Some of the semantic integration process can be partially automated using commercially available software products. For example, entities can be extracted from text (e.g. AeroText™ and MetaCarta™) and organized into concept lattices and taxonomies (e.g. Stratify™ and Entrieva™). Knowledge servers of various capabilities are also now commercially available, e.g., Cerebra Server™ (networkinference.com), Semantic EII (ibm.com) and Tucana Knowledge Server™ (tucanatech.com). Each respective vendor makes claims of high scalability and performance. Generally facts are represented as *triples* in the form of <predicate, subject, object> and most provide description logic engines (OWL; see w3c.org/2004/OWL/).

Technology Transfer to Large Scale System?

The question remains as to how well the above knowledge servers can meet the challenge, e.g. implementation of a knowledge layer that facilitates an environment where the intelligence analyst can focus on data mining and discovery rather than the many lower level and mundane tasks. Can commercially available technology effectively bridge the gap between ontology development and domain modeling using logic, to knowledge layers on top of large legacy databases?

As part of our exploration we have considered commercial knowledge server products, e.g., Cerebra Server™ (networkinference.com), Semantic EII (ibm.com) and Tucana Knowledge Server™ (tucanatech.com), and the less commercialized CYC-L™ that provides a higher order logic engine and an ontology consisting of several micro theories (cyc.com).

To learn how well these knowledge servers scale it is necessary to test them with large populations of data and text, and we are using metrics and tests to identify their respective strengths and weaknesses. The primary consideration is how they perform in fielded deployments.

Another major issue is user adoption. Most of the knowledge servers provide ontology engineering environments that are useful to technicians and programmers, e.g., graphical interfaces for developing models with logic, but intelligence analysts will most likely require very different user-interfaces. Clearly end-users in general need to spend their time doing their jobs, enabled by tools and a mining and discovery environment, and not hindered with requirements to learn technical details of ontologies and formal logic. What interface designs are a good match to the end-users, and will they work with commercially available knowledge servers, or are different products needed?

Summary

In summary, the accomplishment of semantic integration and development a knowledge layer on top of multiple, huge legacy database systems remains a major challenge. We are (a) defining the problem and some of the issues, (b) exploring the use of ontologies and logic for semantic integration and development of a knowledge layer, (c) investigating commercially available knowledge servers and techniques, and (d) considering next steps in preparation for potential deployment of a partial knowledge layer on select, large databases. This should benefit mining and discovery applications related to homeland security.

References

Barwise, J. and J. Etchemendy, 2002, *Language, Proof and Logic*, CSLI Publications, Leland Stanford Junior University, USA.

Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, 1996, *Advances in Knowledge Discovery and Mining*, AAAI Press, Menlo Park, CA.

Gruber, T.R., 1993, A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition* **5**: 199-220.

Hendler, J., and D.L McGuinness, The DARPA Markup Language. *IEEE Intelligent Systems* **16**(5): 67-73.

Michalowski, M., J. Ambite, S. Thakkar, and R. Tuchinda, 2004, “Retrieving and Semantically Integrating Heterogenous Data from the Web,” *IEEE Intelligent Systems* **19**(3): 72-79.

Musen, M.A., 1992, Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research* **25**: 435-467.

Staab, S. and R. Studer (eds.), 2004, *Handbook on Ontologies*, Springer, New York.

Westphal, C, and T. Blaxton, 1998, *Data Mining Solutions,* John Wiley and Sons, New York.