IMT: A Mixed-Initiative Data Mapping and Search Toolkit

Michael Zang[1], Adam Gray[2], Joe Kriege[1], Kalyan Moy Gupta[3], David W. Aha[4]

[1]CDM Technologies, Inc.; San Luis Obispo, CA 93401
[2]Collaborative Agent Design Research Center (CADRC);
California Polytechnic State University; San Luis Obispo, CA 93405
[3]Knexus Research Corp.; Springfield, VA 22153
[4]Navy Center for Applied Research in Artificial Intelligence;
Naval Research Laboratory (Code 5514); Washington, DC 20375
{mzang,adgray,jkriege}@cdmtech.com kalyan.gupta@knexusresearch.com david.aha@nrl.navy.mil

Abstract

Interoperability requires the resolution of syntactic and semantic variations among system data models. To address this problem, we developed the Intelligent Mapping Toolkit (IMT), which employs a distributed multi-agent architecture to enable the mixed-initiative mapping of metadata and instances. This architecture includes a novel federation of service-encapsulated matching agents that leverage case-based reasoning methods. We recently used the IMT matching service to develop several domain-specific search applications in addition to the IMT mapping application.

The Motivation for Developing IMT

Interoperability among information systems constitutes a primary concern when integrating processes both within and across organizations. As the distribution process owner (DPO) for the U.S. Military, this is particularly true for the United States Transportation Command (USTRANSCOM), which integrates distribution processes (e.g., supply requisition, inventory management, and transportation) across the individual military services, suppliers, shippers, and host nation support systems. To facilitate the requisite levels of interoperability among system-specific information models, USTRANSCOM has developed the Distribution Process Information Exchange Data Model (DPIEDM) and initiated an effort to map existing system-to-system interfaces to this logical data model. DPIEDM's goal is to provide a much improved semantic and contextual specification to information exchanges, thus improving current and future process integration across the extended enterprise.

The essential operation in data mapping is *Match*, which takes two schemas (or table extensions) as input and produces a mapping between elements of them that correspond semantically (Rahm and Bernstein 2001). For two schemas with n and m elements respectively, the number of possible matches is $n*m$, implying a manually prohibitive effort when mapping to schemas containing thousands of elements, such as the DPIEDM. This implication prompted USTRANSCOM to automate aspects of their mapping task to significantly decrease the requisite level of effort (i.e., time and expertise) while reducing errors. No usefully applicable, commercial products for semantic mapping automation exist. Thus, USTRANSCOM sponsored the development of the IMT operational prototype, which applies artificial intelligence techniques to this compelling problem. The IMT project was a collaborative endeavor involving CDM, CADRC, Knexus, NRL, and USTRANSCOM's semantic mapping community.

The IMT Prototype Description

We introduced IMT in our IAAI-08 paper *Enabling the Interoperability of Large-Scale Legacy Systems* (Gupta, et al. 2008). Here we summarize it only briefly. IMT is novel in several ways. It maps large-scale schema (i.e., metadata) and instance data. It employs a distributed multi-agent architecture that includes a federation of matching agents for case-based similarity assessment and learning. IMT semi-automatically acquires domain-specific lexicons and thesauri to improve its mapping performance. It also provides an explanation capability for mixed-initiative mapping. ***IMT's primary task is to suggest mappings to users for final verification and acceptance***. Its architecture includes the three layers of components shown in Figure 1 and described below.

The ***GUI Layer*** comprises a graphical user interface that allows users to perform actions such as: importing, selecting, and visualizing problem elements; acquiring auxiliary resources; invoking matching agents; consulting the agent explanation facility; and exporting mapping solutions for use in other applications.

The ***Agent Layer*** provides *Matching* agents that compute the similarity between problem elements (i.e., tables and fields) by employing similarity assessment procedures typically used in case-based reasoning (CBR). Each agent uses a different feature representation to address a variety of syntactic and semantic variations. For example, the N-gram Matcher converts element names and descriptions into n-grams, each of which becomes a feature, to address the morphological variations in the text pertaining to verbs and nouns (e.g., *description* vs. *describe*). Likewise, the Word Matcher tokenizes multi-word descriptions into words that will be used as features for linguistic matching. Unlike the N-gram Matcher, the

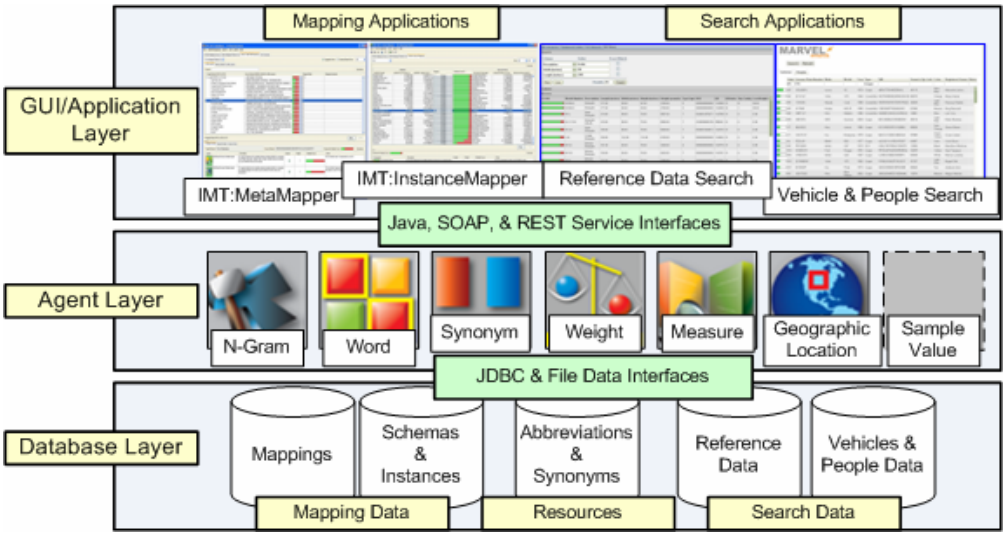

Figure 1: The IMT Architecture

Word Matcher uses information from the Synonym Matcher to process semantic variations. The Synonym Matcher computes the similarity of two features by using the Abbreviations and Synonyms Libraries. The Word Matcher then incorporates these results into the overall similarity assessment.

The ***Database Layer*** includes JDBC-compliant repositories for persisting the mapping problem and solution representation—supporting mapping among schemas, tables, fields, and instances—and the resources for storing the abbreviations and synonyms—supporting the strength of association among synonyms for use by matching agents Additionally, schema and instance data may be imported directly from mapping problem sources.

The New Capabilities and Applications

Since completion of the initial IMT prototype for USTRANSCOM, the underlying similarity assessment framework and agents have been re-factored and cleanly partitioned into a Similarity Assessment Service supporting a number of interfaces (e.g., Java, SOAP, and REST) and a new IMT semantic data mapping toolkit revision. This approach has generalized the original GUI Layer into an Application Layer supporting other problem domains.

In addition to the IMT application, the Similarity Assessment Service now supports domain-specific search tools including: (1) an application that identifies desired records in military reference data and (2) an application that identifies vehicles or people of interest. Capabilities currently under development for the IMT mapping application include an agent for providing schema match scores from corresponding sample data values and the generation of data transformation code from the semantic mappings produced by IMT.

Demonstration Description

Our demonstration employs a combination of display posters, self-running slide shows, hands-on software interaction by attendees, and narrated software presentations to showcase IMT's ability to specify, import, and refine a metadata or instance data mapping problem. The demonstration further illustrates the practical decision support assistance provided by IMT towards resolving these problems. Additionally, our demonstration will incorporate one or more intuitive IMT-technology-derived search applications developed as Cal Poly student senior projects under the auspices of the CADRC. These applications will show the weighted combination of multiple *Match* methods—including N-Gram, Word with Synonym replacement, Measured Quantity, Geographic Location, and Sample Value comparison—to assess similarity between distinct data elements such as the schemas, tables, and fields of two databases to be mapped.

References

Gupta K.M., Aha D.W., & Moore P.G. (2006). Rough set feature selection algorithms for textual case-based classification. *Proceedings of the Eighth European Conference on Case-Based Reasoning* (pp. 166-181). Ölündeniz, Turkey: Springer.

Gupta, K.M., Zang, M.A., Gray, A., Aha D.W., Kriege J. (2008). Enabling the Interoperability of Large-Scale Legacy Systems. To appear in *Proceedings of the Twentieth Innovative Applications of Artificial Intelligence Conference*. Chicago, IL: AAAI Press.

Rahm, E., & Bernstein, P.A. (2001). A survey of approaches to automatic schema matching. *International Journal on Very Large Databases*, 10, 334-350.