

# Information Integration using Information Theoretic Techniques

Eduard Hovy

Information Sciences Institute  
University of Southern California  
4676 Admiralty Way  
Marina del Rey, CA 90292  
hovy@isi.edu

## Introduction: Three Methods of Data Integration

The problem of data and information integration is widespread and growing in government and industry, and it is getting worse as legacy systems continue to appear. The absence of efficient large-scale practical solutions to the problem, and the promise of especially the information theoretic techniques on comparing sets of data values, make this an extremely rewarding and potentially high-payoff area for future research.

One of the principal problems when integrating nonhomogeneous data sets is terminology standardization. Defining *exactly* what data has been obtained, and making sure that the actual method of data capture employed did in fact accurately observe the specifications, is a task for specialists, and may require some very sophisticated analysis and description. It is not uncommon for specialists in different government agencies to spend weeks or even months understanding precisely what differences exist between their respective data collections, even when to the untrained eye the collections seem essentially identical. Not only technically difficult, the data integration process is fraught with potential unexpected legal and social ramifications.

Several IT researchers have performed research on data and information integration with Federal and State government data collections. Three principal approaches exist:

- Direct access, using information retrieval techniques (e.g., Ponte and Croft, 1998; Callan et al., 2001)
- Metadata reconciliation, using model/ontology alignment techniques (Arens et al., 1996; Hovy et al., 2001)
- Data mapping, using statistical information theoretic techniques (e.g., Pantel et al., 2005; Kang and Naughton, 2003)

The two older methods of direct access and metadata alignment approaches appear to be rather inaccurate and/or still require considerable human effort. In contrast, the third approach, of finding possible alignments across data collections by statistical measures on the actual data itself, holds great promise for the future.

## Why the Statistical Approach? Lessons from Our Past Work

Research group at USC's Information Sciences Institute have been working on problems in Information Integration since the early 1990s. Early work on the SIMS system (Arens et al., 1996) included a central domain model that was linked to the component databases and an AI-style planner that decomposed queries for efficient access. SIMS required the system designer to build a model of the application domain and to define the contents of each source (database, web

server, etc.) in terms of this model. The SIMS planner provided a single point of access for all the information: the user expressed queries without needing to know anything about the individual sources. SIMS translated the user's high-level request, expressed in a subset of SQL, into a query plan (Ambite and Knoblock, 2000), a series of operations including queries to sources of relevant data and manipulations of the data.

In later work, the EDC project (Hovy et al., 2002) took this a step further, addressing the problem of the semi-automated construction of the single central model and linking it to a large general-purpose term taxonomy or ontology Omega. The system provided dynamically planned access to data about petroleum products' prices and volumes, provided in a variety of forms and on a variety of media, by the Energy Information Administration, the Bureau of Labor Statistics, and the Census Bureau, and the California Energy Commission, in the form of over 50,000 data tables. In order to more rapidly construct the domain models, systems were developed for automatically identifying terminology glossary files from websites, extracting and formalizing the glossary definitions, clustering them appropriately, and automatically embedding them into the existing ontology and domain model.

The modeling bottleneck suggested using a novel data-centered approach: using statistical techniques from machine learning and natural language processing to suggest alignments between individual data items within and across data collections. The premise of our research was that it is possible to significantly reduce the amount of manual labor required in database wrapping and integration by automatically learning mappings in the data. In this research, we applied statistical algorithms to discover column correspondences across environmental databases. We have seen particular success in an information theoretic model, which we call Sift (Pantel et al., 2005), which performs data-driven column alignments. We have applied SIFT to mapping Santa Barbara and Ventura County Air Pollution Control Districts' 2001 and 2002 emissions inventory databases with the California Air Resources Board statewide inventory database. The application of SIFT yielded 75% precision and 72.2% recall on the column alignment task. On a task of integrating new district data with the statewide database, we achieved 55% accuracy for Ventura County and 59% accuracy for Santa Barbara County. We have subsequently applied this method to find duplicate entries within a single data collection. Our Guspian system has been used to identify duplicates within several databases of air quality measurements compiled by the US EPA, including the CARB and AQMDs emissions inventories as well as EPA's Facilities Registry System (FRS). In summary, Guspian's performance on the CARB and Santa Barbara County Air Pollution Control District 2001 emissions inventories was:

- with 100% accuracy, Guspian extracted 50% of the matching facilities;
- with 90% accuracy, Guspian extracted 75% of the matching facilities;
- for a given facility and the top-5 mappings returned by Guspian, with 92% accuracy, Guspian extracted 89% of the matching facilities.

We have created an online version of the SIFT and Guspian to which the user can upload suitably formatted data collections at <http://sift.isi.edu/>. Its results list potential alignments/overlaps with their likelihood scores, for human analysis.

## **The Future?**

I will outline the results from our work and then argue for a vision in which online services perform statistically-based data alignment on demand. Such data integration services would facilitate the Grid.

## References

- Arens, Y., C.A. Knoblock, and C.-N. Hsu. 1996. Query Processing in the SIMS Information Mediator. In A. Tate (ed), *Advanced Planning Technology*. Menlo Park: AAAI Press.
- Ambite, J.L. and C.A. Knoblock. 2000. Flexible and Scalable Cost-Based Query Planning in Mediators: A Transformational Approach. *Artificial Intelligence Journal*, 118 (1–2).
- Callan, J., W.B. Croft, and E.H. Hovy. 2001. Two Approaches toward Solving the Problem of Access to Distributed and Heterogeneous Data. In *DGnOnline*, online magazine for Digital Government. <http://www.dgrc.org/dg-online/>.
- Hovy, E.H., A. Philpot, J.L. Ambite, Y. Arens, J.L. Klavans, W. Bourne, and D. Sarioz. 2001. Data Acquisition and Integration in the DGRC's Energy Data Collection Project. *Proceedings of the NSF's National Conference on Digital Government dg.o 2001*.
- Hovy, E.H. 2003. Using an Ontology to Simplify Data Access. *Communications of the ACM*, Special Issue on Digital Government. January.
- Kang, J. and J.F. Naughton. 2003. On schema matching with opaque column names and data values. *Proceedings of SIGMOD-2003*. San Diego, CA.
- Pantel, P., A. Philpot, and E.H. Hovy. 2005. An Information Theoretic Model for Database Alignment. *Proceedings of the SSBDM conference*, pp. 14–23. San Diego, CA.
- Ponte, J. and W.B. Croft. 1998. A Language Modeling Approach to Information Retrieval. *Proceedings of the 21st International Conference on Research and Development in Information Retrieval*, pp. 275–281.