Significance Information for Translation: Air Quality Data Integration

Andrew Philpot, Patrick Pantel and Eduard Hovy

Digital Government Research Center
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

{philpot,pantel,hovy}@isi.edu

ABSTRACT

The management of air quality involves local, state, regional, national, and international organizations. At each level, data are collected and used for analysis, assessment, and regulatory enforcement. Effective air quality management requires coordination among multiple organizations and, therefore, requires integration among their respective data sets. This integration remains a complex IT challenge due to the variety of collection, storage, format, and dissemination methods employed by each organization. In most cases today when organizations need to share data, specialized arrangements and significant manual effort are required to create usable mappings between the data sources. More general methods are required to bridge the gap between representations and organization schemes. We present an interactive web-based demo of our preliminary work adapting the statistical alignment and clustering methods from cross-language statistical machine translation. Using the demo, users can discover new relations and test likely candidate relations between two similar data sources from local and California state air quality management agencies.

Categories and Subject Descriptors

H.2.2 [**Database Management**]: Heterogeneous Databases.

General Terms

Algorithms, Experimentation.

Keywords

Database Integration, Mutual Information.

1. APPROACH

Aligning comparable databases is a challenging and important problem. Despite the wealth of information contained in these information sources, lack of metadata, differences in formats, representational emphases, and encoding schemes all combine to render the integration effort human-centered, time-consuming and error-prone. This problem is faced by thousands of large enterprises with numerous data collections, from Government agencies at all levels to the chemical and automotive industries to startup companies that link together and integrate websites.

Recently, automatic alignment approaches have shifted focus to data-driven techniques which can discover relationships inherent in data sets, without use of any particular form of metadata. Inspired by the analogous problem of cross-language Machine Translation (MT), we are investigating the use of an information theoretic model to perform data-driven alignments.

The key to our approach is to first identify, using an information-theoretic model, the most informative data elements and then match data sources that share these informative elements. For example, in a data set modeling Santa Barbara, California facilities required by statute to submit emissions data to local regulatory agencies, terms such as "lease" (as in oil rig lease), "Vandenberg AFB," "auto body", and especially "Santa Barbara" are very common; while terms like "Ford" or "Oregon" or "Wingerden" occur only rarely. Therefore, a pair of columns both of which contain "Vandenberg AFB" are intuitively not as similar as two columns both of which contain "Wingerden." Our system, which we call Similarity Information for Translation (*SIfT*) automatically detects and exploits situations corresponding to this intuitive case, assigning higher similarity relationships to those columns (or rows, or tables, with some modifications) which contain less common and thus more significant similarities.

In information retrieval, clustering, and related areas, objects (typically documents) are often represented in a feature vector space. Each word or aspect of the domain corresponds to one dimension of a multi-dimensional feature space. A document is then a vector trajectory through this hyperplane, with the value in a given dimension being a statistic (e.g., frequency of that occurrence) of the associated feature. Similarly, in our task, we model each database column using its data elements; the feature statistics are the pointwise mutual-information between a column and a data element (Church and Hanks, 1989). We also represent columns using expanded feature sets suggested by the data domains of the columns. For example, a column of mostly English text phrases can be modeled by extracting the words or three-character "trigrams" from each text field; a column of 10 digit phone numbers can be modeled by the area code, exchange, and suffix, etc. Expanded feature sets can increase the likelihood of a match given two different representations; they can also introduce spurious matches.

Given two columns represented by such feature vectors, their similarity is computed using a vector distance or alignment metric. We use cosine similarity (Salton and McGill, 1983) which has the very nice property of not being very sensitive to 0-frequency features. In other words, the absence of a matching feature does not indicate dissimilarity as strongly as the presence of a matching feature indicates in similarity.

Similar columns in *SIfT* are discovered using a clustering algorithm called CBC (Pantel and Lin, 2002). We also report

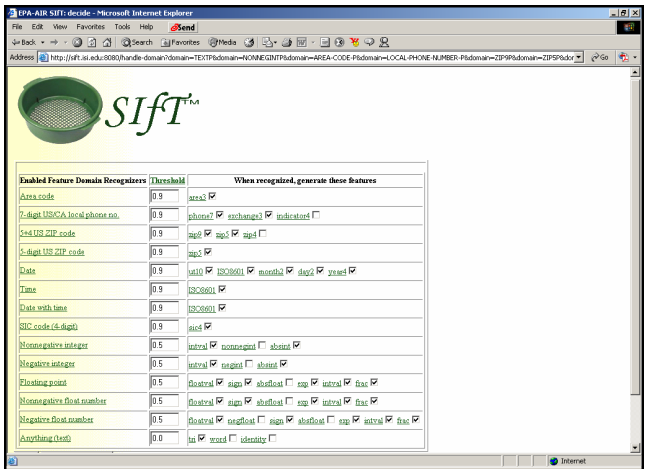

Figure 1. User specifies features to generate and thresholds.

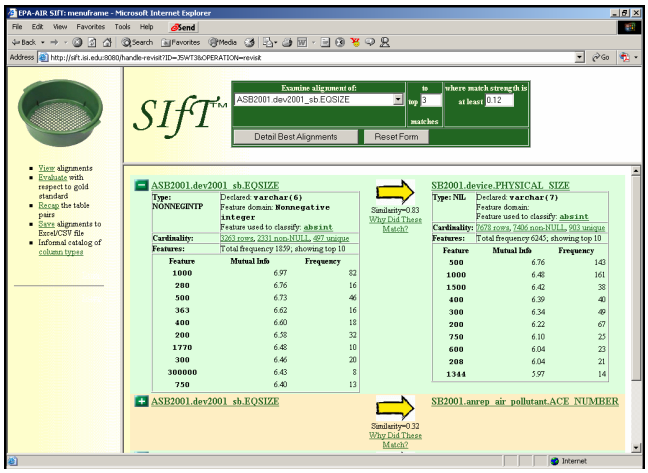

Figure 2. The user examines a proposed candidate column/ column match.

precision and recall measures relative to a human-derived "gold-standard" alignment, if available.

2. DEMO

In the interactive web-based *SIfT* demo, a user begins by choosing two or more information sources to align, from a set of roughly 20 emissions inventory databases maintained by air quality management districts within California and/or the California Air Resources Board statewide inventory, organized by model year. The user may thus attempt to find relationships between one or more local air quality inventories and the statewide inventory, or between one local district and another, or over different model years for the same district.

The user then chooses whether alignment should be conducted column to column or row-to-row. Given a set of data sources and an alignment, the user can then select which feature models are of interest (*e.g.,* are phone numbers, SSNs, or CAS abstract codes to be considered?) and given a feature mode, which features should be generated (*e.g.,* match phone numbers directly, or using area code and prefix separately, etc.). One can also choose feature model thresholds, (*e.g.,* what proportion of column values should look like SSNs before deciding to treat the column as a set of SSNs) (see Figure 1). The system makes the typical exploration of this parameter space straightforward.

The system then computes mutual information scores, generates clusters, presents a candidate set of alignments, and compares these alignments to those in the gold standard, when available. The user then can explore the produced alignment (Figure 2) or export it for later processing.

The demo is also available online at the URL http://sift.isi.edu/.

3. EVALUATION

SIfT has been evaluated using a gold standard which was constructed by the (non-domain expert) authors using a hand alignment of two emissions inventory databases both dealing with Santa Barbara County, California emissions reports from 2001. (No gold standard is currently available from the two supplying agencies.) *SIfT* found 75.0% of the mappings in that gold standard, and also proposed several highly significant yet previously undetected mappings. More details on the evaluation methodology and results are available in (Pantel et al, 2005).

4. REFERENCES

[1] Church, K. and Hanks, P. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of ACL-89*. pp. 76–83. Vancouver, Canada.

[2] Salton, G. and McGill, M.J. 1983. *Introduction to Modern Information Retrieval*. McGraw Hill.

[3] Pantel, P. and Lin, D. 2002. Discovering word senses from text. In *Proceedings of SIGKDD-02*. pp. 613–619. Edmonton, Canada.

[4] Pantel, P., Philpot, A., and Hovy, E. 2005. An Information Theoretic Model for Database Alignment. Submitted to *National Conference on Digital Government Research-05*. Atlanta, Georgia.