Data Alignment and Integration

An information theory-based technology measures the importance of observations and leverages them to quantify the similarity between entities, improving accuracy and reducing the time required to find related entities in a population.

Patrick Pantel
Andrew Philpot
Eduard Hovy
Digital Government Research Center, USC Information Sciences Institute

To handle the wide range of geographic scales and complex tasks that it must administer, the government splits its data in many different ways, collecting it at different times and through different agencies. The resulting massive data heterogeneity makes it impossible to effectively locate, share, or compare data across sources, let alone achieve computational data interoperability.

Many settings urgently need some form of data alignment or merging. For example, an air-quality scientist at a state environmental agency such as the California Air Resources Board (CARB) reconciles air emissions data from local regions to monitor overall patterns and support air-policy regulation. In a homeland security scenario, analysts identify and track threat groups using separately collected and stored individual behaviors such as phone calls, e-mail messages, financial transactions, and travel itineraries.

Addressing these issues requires finding similarities between entities within or across heterogeneous data sources. To date, most approaches for integrating data collections, or even for creating mappings across comparable data sets, require manual effort. Despite some promising recent work, the automated creation of such mappings is still in its infancy because equivalences and differences manifest themselves at all levels, from individual data values through metadata to the explanatory text surrounding the data collection as a whole.

Some data sources contain auxiliary information such as relational structure or metadata, which have proven useful in interrelating entities. However, such auxiliary data can be outdated, irrelevant, overly domain-specific, or simply nonexistent. Therefore, a general-purpose solution can't rely on such auxiliary data. All we can count on is the data itself: a set of observations describing the entities.

Applying this purely data-driven paradigm, we've built two systems: Guspin for automatically identifying equivalence classes or aliases, and Sift for automatically aligning data across databases. The key to our underlying technology is identifying the most informative observations and then matching entities that share them.

We've used our systems to align US Environmental Protection Agency (EPA) data between the Santa Barbara County Air Pollution Control District (SBCAPCD) and Ventura County Air Pollution Control District (VCAPCD) emissions inventory databases and the CARB statewide inventory database, as well as to identify equivalence classes in the EPA's Facilities Registry System (FRS). This work can significantly reduce the amount of human effort involved in creating single-point access to multiple heterogeneous databases.

GOVERNMENT PARTNERS

The staff at CARB annually integrates the emissions inventory databases belonging to California's 35 air quality management districts (AQMDs) to create a state inventory. They submit this inventory annually to the US EPA, which performs quality-assurance tests on state inventories and integrates them into a national emissions inventory for use in tracking national air-quality policies' effects.

To deliver their annual emissions data to CARB, air districts must manually reformat the data according to the specifications of CARB's California Emission Inventory Development and Reporting System. Every time CARB revises the CEIDARS

Comparing all data in a large collection housed in one or more databases can be an overwhelming task.

data dictionary (as it did in 2002 and several other times recently), AQMD staff must translate emissions data into the new format. Likewise, when CARB provides emissions data to the US EPA's National Emission Inventory (NEI), the CARB staff must expend significant effort translating data into the required format. Our goal with this data set is to automatically integrate the AQMD databases with the CARB database.

The FRS is a centrally managed database recording information about US facilities such as refineries, gas stations, and manufacturing sites that are subject to environmental regulations. Because various sources provide the FRS entries, the database contains many duplicates. Our goal for this data set is to automatically discover the duplicate entries.

INFORMATION MODEL

Comparing all data in a large collection housed in one or more databases can be an overwhelming task. But not all data is equally useful for comparison. Some observations are more informative and important than others.

When assessing the similarity between entities, important observations should be weighted higher than less important ones. The "Leveraging Important Data" sidebar provides an intuitive example of this concept.

Claude Shannon's classic 1948 article provides a way of measuring the information content of events.[1] This theory of information provides a pointwise mutual information metric quantifying the association strength between two events by measuring the amount of information one event tells us about the other. By applying this theory to our problem, we can identify the most important observations for each entity in a population.

Formally, pointwise mutual information quantifies the association strength between two events. It essentially measures the amount of information one event x gives about another event y, where $P(x)$ denotes the probability that x occurs, and $P(x, y)$ the probability that both events occur jointly:

$$mi(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

Pointwise mutual information compares two models (using Kullback-Leibler, or KL, divergence[2]) for predicting the co-occurrence of x and y: One model is the maximum-likelihood estimate (MLE) of the joint probability of x and y, and the other is the MLE of x and y occurring independently. Pointwise mutual information is high when x and y occur together more often than they would by chance, which is computed by the probability of x and y occurring independently.

In the example from Figure A2 in the sidebar, assuming that the total frequency count of all phone calls from all people is 1.32×10^{12}, then the pointwise mutual information between events *John* and *calls-to-DC* is:

$$mi(John, calls\text{-}to\text{-}D.C.) = \log \frac{\frac{336}{1.32 \times 10^{12}}}{\frac{1{,}300{,}281}{1.32 \times 10^{12}} \times \frac{1606}{1.32 \times 10^{12}}} = 5.33$$

and between *John* and *calls-to-Bogota* is

$$mi(John, calls\text{-}to\text{-}Bogota) = \log \frac{\frac{21}{1.32 \times 10^{12}}}{\frac{227}{1.32 \times 10^{12}} \times \frac{1606}{1.32 \times 10^{12}}} = 7.88$$

COMPUTING SIMILARITY

Given a method of ranking observations by their relative importance, we still need a comparison metric for determining the similarity between two entities. The metric must not be too sensitive to unseen observations—that is, the absence of a matching observation is not as strong an indicator of dissimilarity as the presence of one is an indicator of similarity. (Some metrics, such as the Euclidean distance, don't make this distinction.)

Many metrics could apply here. We chose one of the more common ones: the cosine coefficient metric. The similarity between each pair of entities e_i and e_j, using the cosine coefficient metric,[3] is given by:

$$sim(e_i, e_j) = \frac{\sum_o mi(e_i, o) \times mi(e_j, o)}{\sqrt{\sum_o mi(e_i, o)^2 \times \sum_o mi(e_j, o)^2}}$$

where o ranges through all possible observations (for example, phone calls). This measures the cosine of the angle between two pointwise mutual information vectors. A similarity of 0 indicates orthogonal vectors (that is, unrelated entities), whereas a similarity of 1 indicates identical vectors. Two very similar elements will have vectors that are very close, and their angle's cosine will approach 1.

Leveraging Important Data

Claude Shannon's classic 1948 article gives us a way to measure the information content of events. Shannon's theory of information provides a pointwise mutual information metric that quantifies the association between two events by measuring the amount of information one event tells us about the other.

The following scenario illustrates the power of pointwise mutual information.

Assume you're a narcotics officer charged with tracking two individuals—John Doe and Alex Forrest—from a population of Southern California residents. Would knowing that last year both John and Alex called Hollywood about 21 times a month increase your confidence that John and Alex are the same person or from the same social group? Possibly.

Now, suppose we also told you that John and Alex both called Bogota about 21 times a month. Intuitively, this observation yields considerably more evidence that John and Alex are similar because not many Southern California residents call Bogota so frequently. Measuring the relative importance of such observations—calling Hollywood and calling Bogota—and leveraging them to compute similarities is the key to our approach.

Table A lists John's most frequently called cities along with the call frequencies. It's not surprising that a Californian would call Los Angeles, Culver City, Anaheim, and even Washington, D.C. If Alex had similar calling patterns to these four cities, it would somewhat increase our confidence that he and John are similar, but obviously our confidence would increase much more if Alex also called the more unusual cities, Bogota and Medellin.

Looking only at the call frequencies in Figure A1, we would place more importance on matching calls to Los Angeles than to Bogota. But mutual information provides a framework for reranking calls by their relative importance (information content). Figure A1 shows the frequencies of John calling Washington, D.C., John calling any city, and anyone calling Washington, D.C.; Figure A2 illustrates the same for Bogota. Although John calls Washington, D.C., more frequently than Bogota, many more people in the population call Washington, D.C., than Bogota.

Pointwise mutual information leverages this observation by adding importance for a city to which John calls frequently and deducting importance if many people in the general population call the same city. Table B shows the results of reranking the cities by the pointwise mutual information measure.

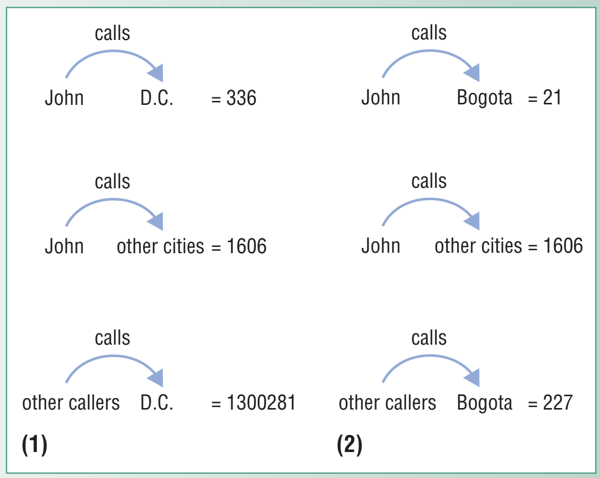

Figure A. Identifying important observations in our homeland security scenario of phone calls placed by Southern California residents: (1) frequency of calls that John and others placed to Washington, D.C., and other cities; and (2) frequency of calls that John and others placed to Bogota and other cities.

Table A. Frequency of phone calls that John Doe placed monthly.

City	Call frequency	City	Call frequency
Los Angeles	571	Boston	34
Washington, D.C.	336	Ventura	33
Hamburg	234	St. Louis	31
Culver City	199	Bogota	21
Anaheim	103	Hollywood	21
Leipzig	59	Covina	20
Medellin	51	Long Beach	16
Toronto	38	Carson	16

Table B. Reranking results of phone calls by pointwise information.

City	Reranked score	City	Reranked score
Bogota	7.88	Ventura	4.38
Medellin	7.05	Toronto	4.36
Leipzig	5.78	Boston	4.31
Hamburg	5.58	Covina	2.91
Culver City	5.48	St. Louis	2.40
Washington, D.C.	5.33	Long Beach	2.03
Los Angeles	4.77	Carson	1.62
Anaheim	4.46	Hollywood	1.43

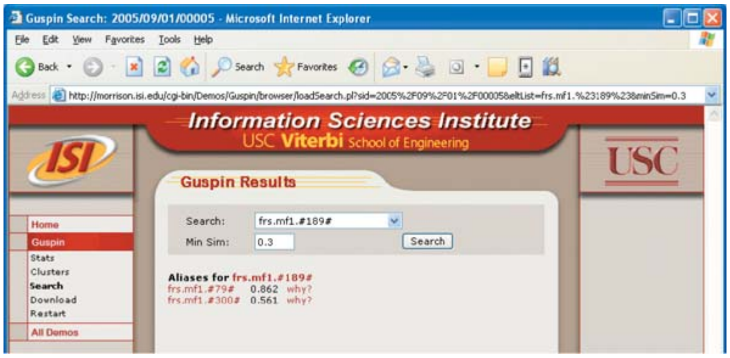

Figure 1. Guspin's search interface. In this example, facility 189 from the EPA's Facilities Registry System is most similar to facilities 79 and 300. A user can click on a facility name to view its observation data or on "why?" for a comparison of observation data from facilities 189, 79, and 300.

SYSTEMS

We've applied the technology to several problems, including automatically building a word thesaurus, discovering concepts, inducing paraphrases, and identifying aliases in a homeland security scenario (www.isi.edu/~pantel).

In the digital government context, we've built two Web tools—Guspin and Sift—and applied them to problems the EPA faces. At their core, both systems employ pointwise mutual information and similarity models.

Guspin

Guspin (http://guspin.isi.edu) is a general-purpose tool for finding equivalence classes within a population. It provides a simple user interface that lets a user upload one or multiple data files containing observations for a population. The system identifies duplicate (or near-duplicate) entities and presents the results to the user for browsing or download.

We used Guspin to identify duplicates in our two test sets (the CARB and AQMD emissions inventories and the FRS). For the CARB and SBCAPCD 2001 emissions inventories, Guspin extracted

- 50 percent of the matching facilities with 100 percent accuracy,
- 75 percent of the matching facilities with 90 percent accuracy, and
- 89 percent of the matching facilities with 92 percent accuracy for a given facility and its top-5 mappings.

Our second test used a sample of the FRS. Each FRS record includes a particular facility's name, address, state, and zip code. We upload the FRS data through Guspin's Web interface. Guspin measures the mutual information between entities and observations, computes the similarity between each entity pair, clusters entities into equivalence classes, and provides a browsing tool that an analyst can use to find equivalence classes and navigate a population's similarity space. The analyst can also download the resulting Guspin analysis for further examination.

With Guspin's search feature, users can search for individual entities. For example, Guspin found that facility 189 is grouped with facilities 300 and 79. Figure 1 shows the results of a search for facility 189's most similar entities. For each similar entity, Guspin shows the cosine similarity score and a "why?" link, which lets the user compare observations of the two facilities (recall that we use important observations to compute the similarity between entities).

Figure 2 shows comparisons between observations for facilities 189 and 79 and between observations for facilities 189 and 300. Blue and green observations apply to only one of the two facilities, whereas both facilities share red observations. The figure lists observations in descending order of mutual information scores.

For very similar entities, we expect that the most important observations (those at the top of the list) will be red. In fact, even though Figure 2 shows that facilities 189 and 79 share fewer common observations than facilities 79 and 300, the similarity between facilities 189 and 79 is greater because they share more important features (that is, more red features are at the top of the list).

Guspin is useful for other tasks. For example, it can identify occurrences of plagiarism in essays by representing essays with the words they contain. Or, it can help researchers find coregulated genes by representing genes with their expressions in a series of microarray experiments.

Sift

Sift (http://sift.isi.edu) is a Web-based application portal for cross-database alignment.[4] Given two relational data sources, Sift helps answer the question, Which rows, columns, or tables from source 1 have high correspondence with (all or part of) some parallel constructs from source 2? Most previous attempts at intersource alignment rely heavily on metadata, such as table and column names and data types.[5] Yet, as noted earlier, metadata is often unreliable or unavailable.

Sift provides most of the same functionality as Guspin, but adds control over the definition and use of observations in the data sources. Whereas

Figure 2. Guspin comparison of two entities' observations. (a) Comparison of observations for facilities 189 and 79 (similarity = 0.862) and (b) comparison of observations for facilities 189 and 300 (similarity = 0.561).Observations are sorted in decreasing order of pointwise mutual information scores. Blue and green observations apply to only one of the facilities; red observations are shared by both.

Guspin takes a population description as input, Sift more narrowly draws input from a pair of relational databases. The user controls which database elements to include in the alignment (for example, columns, rows, and tables).

Table 1 shows database column fragments taken from two databases, A and B. Because none of the observations in the data fields overlap exactly, Guspin wouldn't be able to find any match. Sift overcomes this problem by preprocessing observations to identify known observation types, such as phone numbers, zip codes, time, and date. Sift can then reformulate the observations into their atomic parts. For example, a phone number's atomic representation could be the area code and local phone number, whereas a date's atomic representation would be its month, day, and year components.

After this preprocessing, Sift reformulates the first field of the example in Table 1, A.T1.phone_number, to 310 and 555-6789. Sift then matches these observations to those in B.T2.area_code and B.T2.local_phone. The user controls which reformulations to apply.

Consider CARB's task of creating an annual emissions inventory for California—that is, a catalog of the emitting facilities, processes, and devices in the state's 35 local AQMDs and the measurement or estimated toxic and criteria pollutants they produce.

Here, we consider the column alignment between CARB and the SBCAPCD. The CARB and SBCAPCD emissions inventory databases used in our experiments each contain approximately 300 columns, thus a completely naïve human must consider approximately 90,000 alignment decisions in the worst case.

After selecting reformulation parameters, Sift measures the mutual information between columns and observations (data fields), computes the similarity between each pair of columns, and presents the user with an interface for browsing the alignment.

Figure 3 shows a correct alignment discovered by Sift for the columns containing process descriptions. Sift also displays the most important observations contributing to the alignment (including the pointwise mutual information scores and frequency).

Like Guspin, Sift provides a "why did these match?" link for comparing the observations of the two aligned columns. Figure 4 illustrates why Sift aligns the process description columns. It shows a complete view of the observations for both columns. The observations are in descending order of pointwise mutual information. Observations in red belong in both columns, whereas observations in blue are from the SBCAPCD column only, and observations in green are from the CARB column only. Sift aligns the two columns because they share many high mutual information observations.

Sift discovered 295 alignments, of which 75 percent were correct. Of the 306 true alignments, Sift identified 221, or 72 percent. Interestingly, when Sift finds a correct alignment for a given column, it finds it in the first two returned candidate alignments.

Considering only two candidate alignments for each possible column will reduce the number of possible decisions a human expert makes. Assuming that we must consider each of the 90,000 candidate alignments (in practice, human experts easily reject many alignments) and that for each column Sift outputs at most k alignments, a human expert would have to inspect only $k \times 300$ alignments. For $k = 2$, a human must inspect only 0.67 percent of the possible alignment decisions, an enormous time savings.

Table 1. Database column fragments.

A.T1. phone_number	B.T2. area_code	B.T2. local_phone
310-555-6789	310	555-6789
310-666-0987	310	666-0987
213-777-9393	310	777-9393

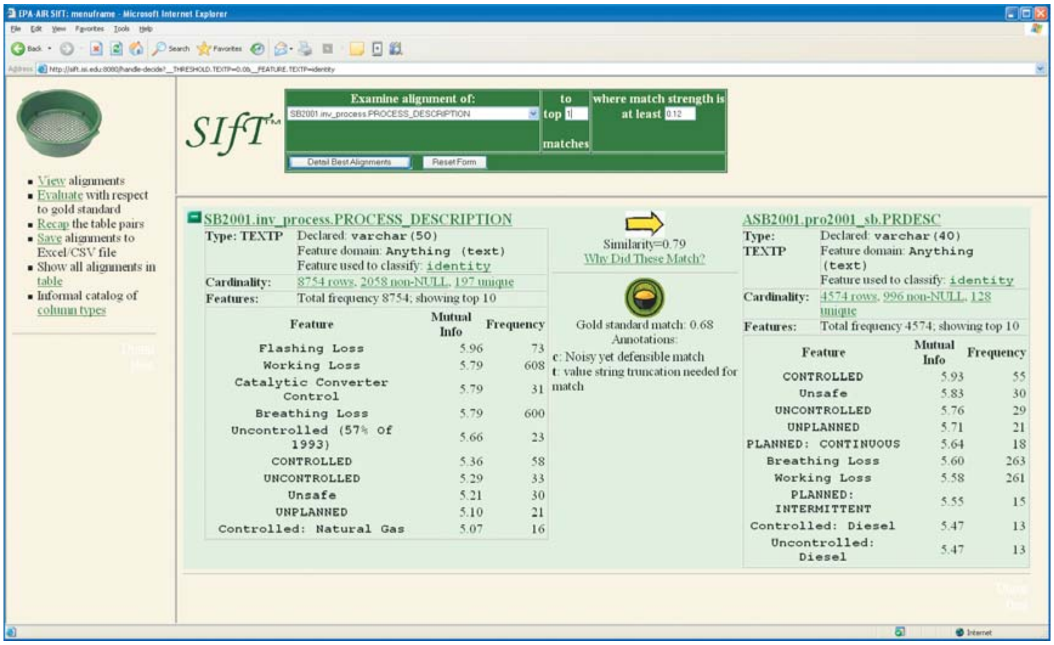

Figure 3. Sift display. Sift discovered a correct alignment between the process description columns in the Santa Barbara County Air Pollution Control District and California Air Resources Board databases.

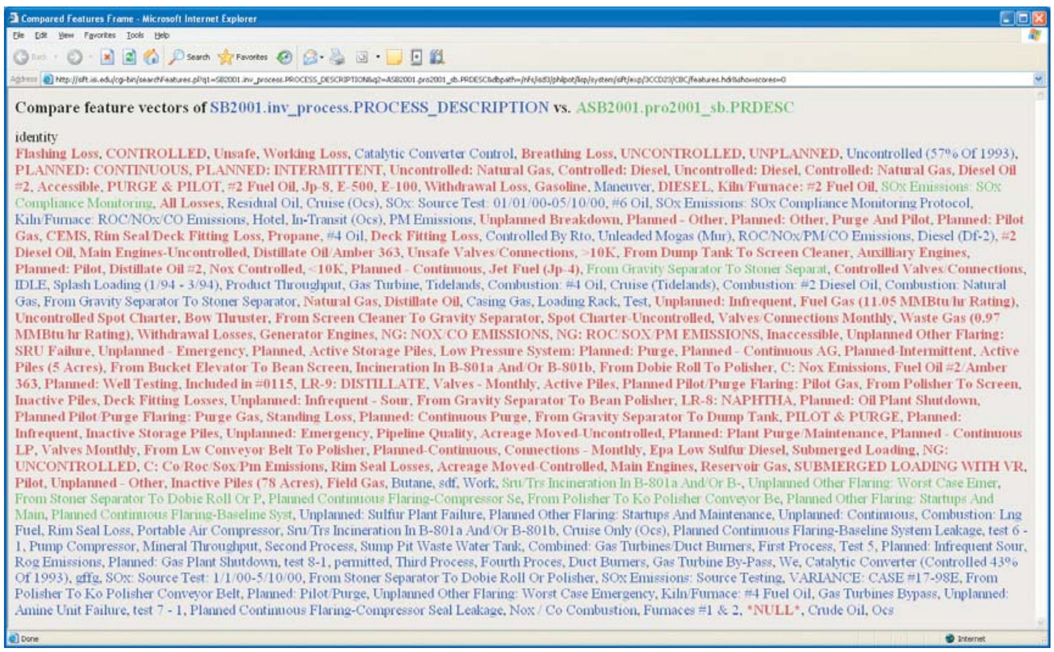

Figure 4. Observations for the process description columns in the SBCAPCD and CARB databases, in descending order of mutual information. Observations in red belong to both columns, observations in blue are only from the SBCAPCD column, and observations in green are only from the CARB column. The predisposition of red observations to appear at the head of the list demonstrates Sift's effectiveness.

To be of practical use to government, a system must effectively analyze data transfers between district and state and integrate new data as it becomes available each year. This is challenging because data formats can change on both the collectors' and the integrators' sides. However, because year-to-year changes are rarely significant, we can reconcile the possibly divergent evolutions automatically, thereby closing the loop by automatically generating the data integration.

We used this process to integrate the 2002 databases of both the Ventura County Air Pollution Control District (VCAPCD) and SBCAPCD into CARB.[4] We randomly sampled 50 columns in the automatically integrated CARB 2002 databases, then we asked a human judge to classify each aligned column according to the following guidelines:

- *Correct*. The column is aligned correctly according to the gold standard.
- *Partially correct*. The aligned column is a subset or superset of the gold standard alignment. This situation arises when we transfer only part of the column to CARB or when we must perform a join on the district tables to match the CARB schema. Solving these problems requires more than simple column alignments and is beyond this article's scope.
- *Incorrect*. The column is not aligned correctly according to the gold standard.

Table 2. Results for automatically generating a CARB 2002 database from VCAPCD and SBCAPCD 2002 databases.

Database	Sample size	Correct	Partially correct	Incorrect	Accuracy
VCAPCD	50	25	5	20	55%
SBCAPCD	50	22	15	13	59%

Note: Alignments judged as partially correct count 1/2 point toward the accuracy.

Table 3. Accuracy of the top *K* alignments using the cosine similarity metric for 50 random samples from VCAPCD and SBCAPCD.

Database	Top 1	Top 5	Top 10	Top 25	Top 50
VCAPCD	100%	100%	60%	70%	55%
SBCAPCD	100%	100%	95%	76%	59%

Table 2 shows our evaluation's results. We compute the system's accuracy by adding one point for each correct alignment, half a point for each partially correct alignment, and no points for each incorrect alignment, and then dividing the total by the sample size. Sift doesn't align some district columns into the CARB database—that is, it doesn't find any alignment for these columns. In our 50 random VCAPCD samples, Sift left nine columns unaligned; of these, six were correct and three were incorrect.

Error analysis shows that Sift is particularly bad at aligning binary (yes/no or 0/1) columns. The pointwise mutual information model isn't useful here because many columns share binary values. A separate process should align such columns, which are easy to identify. For example, we might simply compare the ratio of 0s and 1s or the raw frequency of 0s and 1s. Likely, however, we'll need more complex table and row analysis.

Each alignment includes a similarity score (from the cosine similarity metric). We can view this similarity as Sift's confidence in each alignment. For both VCAPCD and SBCAPCD, we sorted the 50 randomly sampled alignments in descending order of Sift confidence and measured the accuracy for the top K alignments, for $K = \{1, 5, 10, 25, 50\}$. For binary columns, Sift disregards the similarity score and assigns a 0 confidence score. Table 3 illustrates the results. As expected, the higher the confidence Sift has in a particular alignment, the higher the chances that this alignment is correct.

A general-purpose solution to the problem of matching entities within or across heterogeneous data sources can't depend on the presence or reliability of auxiliary data such as structural information or metadata. Instead, it must leverage the available data (or observations) that describe the entities. Our technology, based on information theory principles, measures the importance of observations and then leverages them to quantify the similarity between entities.

At a minimum, our systems can dramatically reduce the time an analyst needs to find related entities in a population. However, the technology's power depends on gathering the right observations that entities might share, which in itself is an interesting avenue of future work. Given the right types of observations, our model can potentially solve several serious and urgent problems that governments face, such as terrorist detection, identity theft, and data integration. ■

References

1. C.E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical J.*, vol. 27, no. 3, 1948, pp. 379-423.
2. T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.
3. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.
4. P. Pantel, A. Philpot, and E.H. Hovy, "An Information Theoretic Model for Database Alignment," *Proc. Conf. Scientific and Statistical Database Management* (SSDBM 05), IEEE CS Press, 2005, pp. 14-23; http://2005.ssdbm.org/.
5. M. Tova and S. Zohar, "Using Schema Matching to Simplify Heterogeneous Data Translation," *Proc.*

24th Int'l Conf. Very Large Databases (VLDB 98), Morgan Kaufmann, 1998, pp. 122-133.

Patrick Pantel *is an assistant research professor and research scientist in the Natural Language Group at the University of Southern California Information Sciences Institute. His research interests include ontology learning and text mining, knowledge acquisition, and machine learning. Pantel received a PhD in computing science from the University of Alberta in Edmonton, Canada. Contact him at pantel@isi.edu.*

Andrew Philpot *is a research scientist in the Natural Language Group at the University of Southern California Information Sciences Institute. His research interests include ontologies and information integration. Philpot received an MS in computer science (artificial intelligence) from Stanford University. Contact him at philpot@isi.edu.*

Eduard Hovy *is a deputy division director at the University of Southern California Information Sciences Institute and the director for research of USC's Digital Government Research Center (www.dgrc.org). His research interests include topics in human language technology such as text summarization, question answering, machine translation, ontology learning, and text mining and other topics of interest to digital government, notably cross-database alignment. Hovy received a PhD in computer science from Yale University. Contact him at hovy@isi.edu.*