Matching and Integration Across Heterogeneous Data Sources

Patrick Pantel, Andrew Philpot and Eduard Hovy
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
{pantel,philpot,hovy}@isi.edu

ABSTRACT

A sea of undifferentiated information is forming from the body of data that is collected by people and organizations, across government, for different purposes, at different times, and using different methodologies. The resulting massive data heterogeneity requires automatic methods for data alignment, matching and/or merging. In this poster, we describe two systems, *Guspin*™ and *Sift*™, for automatically identifying equivalence classes and for aligning data across databases. Our technology, based on principles of information theory, measures the relative importance of data, leveraging them to quantify the similarity between entities. These systems have been applied to solve real problems faced by the Environmental Protection Agency and its counterparts at the state and local government level.

Categories and Subject Descriptors

H.2.5 [**Database Management**]: Heterogeneous Databases.

General Terms

Algorithms, Experimentation.

Keywords

Information theory, mutual information, database alignment, equivalence class detection.

1. INTRODUCTION

In face of the growing mass of often undifferentiated data being collected in at an unprecedented pace by government agencies, data users need automated assistance to make sense of their data sets by interrelating, clustering, grouping, contained elements etc. The general class of problems is that of finding similarities between entities within or across heterogeneous data sources. To date, most approaches to entity consolidation and cross-source data integration have relied heavily on manual effort and on auxiliary information such as relational structure or metadata. In the work described in this poster, we address the increasingly common case where metadata is outdated, irrelevant, overly domain specific, or simply non-existent. A general-purpose solution to this problem cannot therefore rely on such auxiliary data. Unfortunately, all one can count on is the data itself: a set of observations describing the entities.

In this "data only" paradigm, we have developed an information theoretic model for matching and integration of data sources. The key to our underlying technology is to identify the most informative observations and then match entities that share them. Applying this model, we have built two systems, *Guspin* for automatically identifying equivalence classes or aliases, and *Sift* for automatically aligning data across databases.

2. INFORMATION MODEL

When interrelating entities based on observational data (e.g., matching people based on their financial transactions and communication patterns), certain observations are much more informative and important and thus indicative of similarity than others. When assessing the similarity between entities, important observations should be weighed higher than less important ones.

Shannon's classic 1948 paper [4] provides us with a way of measuring the information content of observed events. This theory of information provides a metric, called pointwise mutual information, which quantifies the association between two events by measuring the amount of information one event tells us about the other.

Consider the following scenario, illustrating the power of pointwise mutual information, in which you are a drug trafficking officer charged with tracking two particular individuals *John Doe* and *Alex Forrest* from a population of Southern California residents. If you were told that last year both *John* and *Alex* called Hollywood about 21 times a month, then would this increase your confidence that *John* and *Alex* are the same person or from the same social group? Yes, possibly. Now, suppose we also told you that *John* and *Alex* each called Bogotá about 21 times a month. Intuitively, this observation yields considerably more evidence that *John* and *Alex* are similar, since not many Southern California residents call Bogotá with such frequency. Measuring the relative importance of two such observations—calling Hollywood and calling Bogotá—and leveraging the measurements to compute similarities between entities is the key to our approach.

In our formulation, we use pointwise mutual information to measure 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 they both occur:

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

Given this method of ranking observations by relative importance, we use the cosine coefficient metric [1] to determine the similarity between two entities. In comparison to other candidate metrics, such as Euclidean distance, cosine is less 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. The similarity between each pair of entities e_i and e_j, using the cosine coefficient metric, 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 (e.g., phone calls). A similarity of 0 indicates orthogonal vectors (i.e., unrelated entities) whereas a similarity of 1 indicates identical vectors. For two very similar elements, their vectors will be very close and the cosine of their angle will approach 1.

3. SYSTEMS

We have applied this mutual information model to several problems, including automatically building a word thesaurus, discovering concepts, inducing paraphrases, and identifying aliases in a homeland security scenario. In the context of digital government, we have built two web tools, *Guspin* and *Sift*, and applied them to problems faced by the Environmental Protection Agency (EPA). At the core, both systems employ the pointwise mutual information and similarity models described in the previous section.

3.1 Guspin™[1]

Guspin is a general purpose tool for finding equivalence classes within a population. It provides a simple user interface where aan analyst user uploads one or multiple data files containing observations for a population. The system then identifies and clusters duplicate (or near-duplicate). *Guspin* provides an analyst with a browsing tool for finding equivalence classes and navigating the similarity space of the supplied population. The analyst may also download the resulting *Guspin* analysis for further examination.In an experiment identifying duplicates facilities given between national, state, and local facility catalogs (described in greater detail in our poster), *Guspin* (i) with 100% accuracy, extracted 50% of the matching facilities; (ii) with 90% accuracy, extracted 75% of the matching facilities; (iii) for a given facility and the top-5 mappings returned by the system, with 92% accuracy, extracted 89% of the matching facilities.

3.2 Sift™[2]

Sift is a web-based application portal for cross-database alignment [2] [3]. Given two relational data sources, *Sift* helps answer the question "which rows, columns, or tables from data source S_1 have high correspondence with (all or part of) some parallel construct(s) from S_2?" Using domain-independent and domain-dependent probabilistic knowledge-based and syntactic data recognizers (e.g., for phone numbers, CAS registry numbers, SIC/NAICS codes, date/time formats), Sift can fortify the as-received observation space with computed observation types. This additional type knowledge is brought to bear during the normal similarity vector space match, allowing for instance that a phone number in S_1 with attached area code might match a phone number in S_2 whose area code is stored in a different column, etc. In an experiment aligning California state and local emissions inventory databases (again described in greater detail on the poster proper) *Sift* discovered 295 alignments, of which 75% were correct. There were 306 true alignments, of which *Sift* identified 221 or 72%. Interestingly, when *Sift* found a correct alignment for a given column, then the alignment appears in the first two returned candidate alignments.

4. CONCLUSIONS

A general-purpose solution to the problem of matching entities within or across heterogeneous data sources cannot rely 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 principles of information theory, measures the importance of observations and then leverages these to quantify the similarity between entities. Though the technology is applicable to a wide range of applications, we have built two web solutions, called *Guspin*™ and *Sift*™, addressing the general problems of building equivalence classes or aliases for a population and of aligning heterogeneous databases. These systems have been applied to solve real problems faced by the Environmental Protection Agency and allied state and local environmental quality agencies with remarkable accuracy. Our systems can dramatically reduce the time an analyst needs to find related entities in a population. However, the power of the technology is critically dependent on gathering the right observations that entities might share, which in itself is a very interesting avenue of future work. Our model has the potential to address several serious and urgent problems faced by the government such as terrorist detection, identity theft, and data integration.

5. REFERENCES

[1] Baeza-Yates, R. and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Wokingham: Addison-Wesley.

[2] Pantel, P.; Philpot, A.; and Hovy, E.H. 2005. Aligning Database Columns using Mutual Information. In Proceedings of *Conference on Digital Government Research (DG.O-05)*. pp. 205-210. Atlanta, GA.

[3] Pantel, P.; Philpot, A.; and Hovy, E.H. 2005. An Information Theoretic Model for Database Alignment. In *Proceedings of Conference on Scientific and Statistical Database Management (SSDBM-05)*. pp. 14-23. Santa Barbara, CA.

[4] Shannon, C.E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 20:50–64.

[1] *Guspin* is available from http://guspin.isi.edu.

[2] Sift is available from http://sift.isi.edu.