Data Integration

Mission Need

With the increasing number of heterogeneous and independent information sources that the different DHS applications need to access to achieve their mission, it is becoming more difficult to integrate information from these different sources and turn it into actionable knowledge. PURVAC is taking a multipronged approach to solve challenges inherent to the critical problem of data integration. The issues we are addressing relate to:

- **Schema matching** to find correspondences or matches between the different attributes present in the data representations or schemas found in different information sources. These attributes are usually represented differently from one system to another due to design choices made in the different systems, language differences, or use of different standards.

- **Entity resolution** to detect and eliminate duplicates improves the consistency of the data significantly. Often, the same real world entity, e.g., a given person, may be represented differently in two systems and even in the same system due to manual entry, lack of uniform standards for content, parallel data entry, software bugs, data exchange, or wrong assumptions towards default values.

- **Privacy-preserving data sharing** between different information systems to enable collaboration between different entities that do not necessarily trust each other. We focus on the peer-to-peer model where, to acquire specific information, a given peer’s request would need to hop through several other peers. While the peers collaborate in answering the request, the requester’s peer does not want to have its request and the response to this request be exposed to any other peer except the one which hold the response.

- **Benchmarking** of data integration techniques, especially schema matching, to enable uniform evaluation of these techniques and give decision makers a tool to choose the right technique based on their needs.

Benefit: Data Integration is critical for homeland security applications to obtain actionable knowledge from the large number of heterogeneous and independent information sources. PURVAC has embarked on a multipronged approach to solve several critical issues in data integration that will enable actionable decision making environments for homeland security applications.

Collaborators:
- Regenstrief Institute
- Purdue University Student Health Clinic (PUSH)

Funded by:
- US Department of Homeland Security
- National Science Foundation
- Lilly Endowment

Early Development | Lab Prototype | Commercial Product
January/2008

Data Integration for Public Health Surveillance in Support of Visual Analytics
**Solutions**

Our multipronged approach to data integration has already produced outstanding results with respect to schema matching and privacy preserving data sharing.

For schema matching, we proposed a novel approach which we call *usage-based schema matching*. In contrast to relying on the schema information or the data instances, as most existing techniques do, our technique exploits the usage information of the attributes in the query logs to find matches. Our technique is suitable for matching schemas even when their attribute names are opaque, use totally different languages, or when they have different layouts.

Our work on addressing the security issues in peer to peer systems where queries and answers need to be mapped across the different peers is the first to propose a complete solution to limit the disclosure of queries, query results, and mappings needed to translate the query between peers. We proposed a *privacy-preserving* query answering protocol based on *noise insertion* and *commutative encryption* methods that guarantees the non-disclosure of query results and mappings.

**Impact**

We are working on applying our different techniques in the realm of public health surveillance and entity behavior extraction.

Public health surveillance collects, analyzes and interprets data about biological agents, diseases, risk factors and other health events from several information sources. The goal is to provide timely dissemination of collected information to decision makers. Providing systematic database support and data integration will enable efficient and effective access to the different data sources needed for surveillance. PURVAC has established access to anonymized data to support our different projects in health surveillance from the Indiana State Department of Health and the Purdue University Student Health center (PUSH).

For the entity resolution challenge, we are working on a new technique based on analyzing similarities between entities behavior. The proposed technique studies the transaction log which records the actions performed by a given entity with a timestamp. These actions can be used to provide information about the entity behavior over time. A time series is then built to extract information about the actions pattern and periodicity. Afterwards, the behavior information is used to improve the detection of duplicate data.

The impact of our data integration efforts will be far beyond healthcare to touch on every single mission of DHS where there is a need to access and integrate data from heterogeneous and independent sources.

---

*The modified Hyperion architecture after adding the privacy-preserving components (Left). A process diagram representing behavior extraction (Right).*

For more information, contact:
Dr. Mourad Ouzzani, mourad@cs.purdue.edu
http://www.purvac.org