Exploring Schema Similarity At Multiple Resolutions

Ken Smith, Craig Bonaceto, Chris Wolf, Beth Yost, Michael Morse, Peter Mork, Doug Burdick
The MITRE Corporation
{kps,cbonaceto,cwolf,bethyost,mdmorse,pmork,dburdick}@mitre.org

ABSTRACT

Large, dynamic, and ad-hoc organizations must frequently initiate data integration and sharing efforts with insufficient awareness of how organizational data sources are related. Decision makers need to reason about data model interactions much as they do about data instance interactions in OLAP: at multiple levels of granularity. We demonstrate an integrated environment for exploring schema similarity across multiple resolutions. Users visualize and interact with clusters of related schemas using a tool named *Affinity*. Within any cluster, users may drill-down to examine the extent and content of schema overlap. Further drill down enables users to explore fine-grained element-level correspondences between between two selected schemas.

Categories and Subject Descriptors: H.5 Information Interfaces and Presentation. I.5.3 Clustering.

General Terms: Human Factors, Design, Algorithms

Keywords. Schema Similarity Exploration

1. INTRODUCTION

In large and changing organizations, chief information officers (CIOs) typically lack detailed awareness of how information assets relate to each other: "Do they address similar topics? Do they share common data elements? To what extent?" Clarity about data source relatedness is vital to making informed decisions about data source integration and retirement projects (e.g., "Can the replacement for System A replace System B as well?"). *Ad-hoc* organizations provide similar challenges. Consider a coalition of providers unexpectedly thrown together due to a hurricane, such as hospitals, churches, the Coast Guard, and local police. In this situation, it is vital to quickly identify communities of interest (COIs) - groups of providers which can pool resources (and exchange their data) to provide a service like food or medical supplies.

Schema similarity is the extent to which a set of schemas share common features, providing a strong indication of data similarity within the systems. Underlying both scenarios above is the need to explore schema similarities in support of the decision making process. While numerous options provide decision support about data in systems, there is a paucity of systems providing decision support about the data systems themselves. This demonstration presents a system addressing this gap.

Schema similarity can be examined at multiple resolutions, analogous to examining data at different levels of aggregation in OLAP systems. We argue that schema relationships at three resolutions are *particularly* useful: a) clustering relationships within a schema repository, b) content overlap relationships within a schema cluster, c) attribute correspondence relationships for a schema pair. Each resolution provides insights which are valuable in different ways. Determining whether two agencies could share a substantial range of information requires "rolled up" schema-level information, while "drilling down" to details about attribute-level similarity reveals schema content, structure, and design hidden at coarser levels. Insights at multiple levels can be synergistic: two clustered health schemas may, upon detailed examination, include a provider of veterinary medicines and a provider of cosmetics; unlikely data sharing partners in a hurricane!

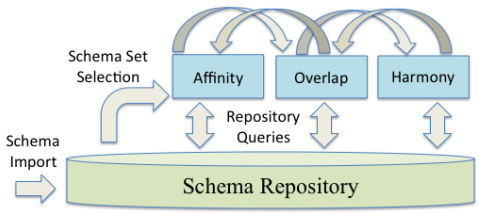

Figure 1: ***Repository-based system architecture***

Contributions

1. *Exploration environment.* We demonstrate a single integrated environment enabling users to explore schema similarity relationships across multiple resolutions.

2. *Schema cluster visualization tool.* Affinity computes and displays schema cluster topology in two interactively linked views: as a hierarchical dendrogram and as a 2D point placement map.

3. *Open source project.* This environment is part of the OpenII [4] information integration framework, which any organization may use and extend for free.

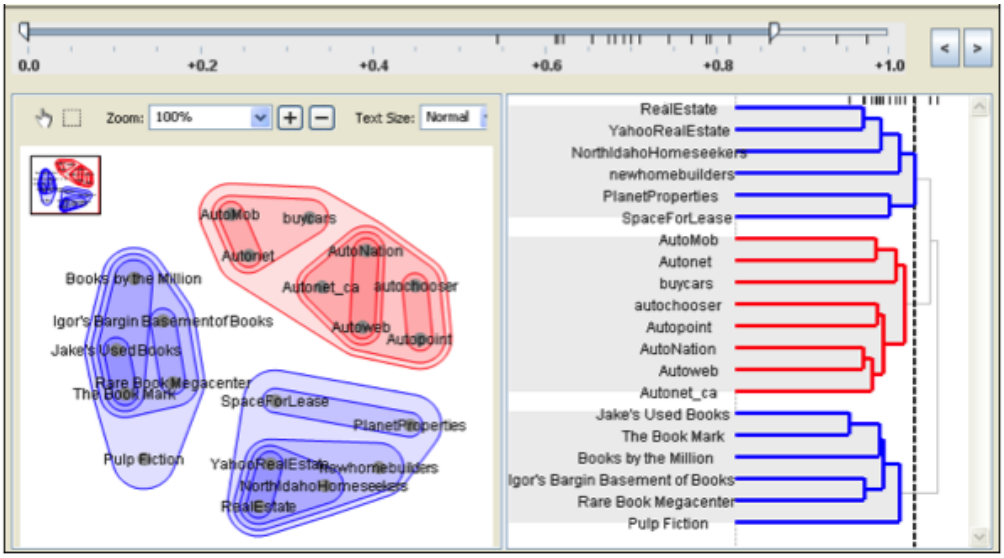

Figure 2: ***A dual-pane Affinity visualization for a set of 22 schemas. The three subtrees selected in the dendrogram (on the right) correspond to three clusters on the left involving books, automotive sales, and real estate.***

Demonstration Scenario

We will demonstrate this environment using a schema repository (Figure 1) loaded with schemas of comparable size, complexity, and quantity typical of government disaster response agencies. Users will be able to explore the contents of the entire repository, and drill down into various schema clusters, and schema pairs, of interest.

Exploration Workflow

Although the actual demo will utilize a larger number of more complex disaster response schemas, we illustrate the workflow with a "toy" e-commerce scenario. AllStuff.com has recently acquired a collection of smaller e-commerce sites which now require consolidation, and the CIO has tasked AllStuff's integration engineer with the task of identifying the best consolidation candidates. After a panicked search, the integration engineer installs this environment, and gets to work.

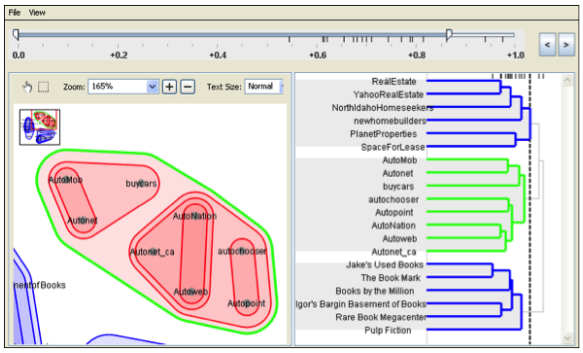

Figure 3: ***Zooming in on a cluster involving books.***

Initially, the engineer imports the schemas into a schema repository, and uses Affinity to cluster them into groups which may reveal potential content overlap. The Affinity visualization for the schemas in the AllStuff.com repository is shown in Figure 2. He can then interactively locate a cluster of interest by zooming into regions in the left pane, or exploring the *dendrogram* in the right pane.

A cluster of interest is selected by clicking on it, as shown by the green highlight of the book cluster in Figure 3. The engineer can drill down into that cluster to further explore the pairwise content overlap for each schema pair in the cluster. Figure 5 illustrates the *overlap view* for the schemas in the book cluster. *Similarly*, a schema pair of interest in the overlap view (e.g., one displaying view high overlap) can be selected for further drill down and exploration of similarity at the schema element level. In this case, the Harmony [6, 3] schema matching tool (Figure 4) provides the needed exploration functionality.

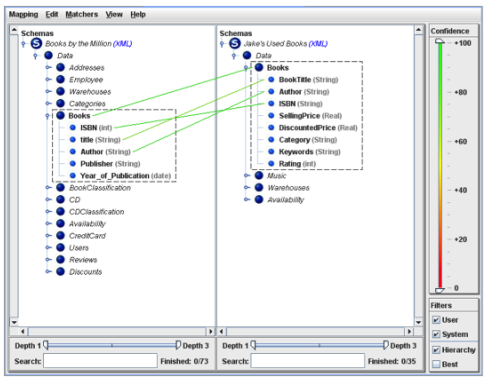

Figure 4: ***Examining element-level similarity for a schema pair using a schema matcher.***

Note that each drill down action opens a new pane in our GUI (the Eclipse framework). Open panes represent the current exploration path, drilling down into increasingly detailed schema representations. Once exploration down one path (e.g., cluster) is exhausted, the user can close those panes and open new ones exploring a different path.

2. SYSTEM DESIGN

As illustrated in Figure 1, all enterprise schemas are first imported into a schema repository. Importers exist for DDL and XSD, as well as custom schema formats; internally, schemas are represented in a general metamodel. The repository, importer, and *M3* metamodel used are from the OpenII project [4]. Leveraging such a framework facilitates interoperability among applications which operate over schemas. Each of the three applications in the demonstrated environment: the Affinity clustering tool, the overlap visualization, and the Harmony schema matcher are designed to use the OpenII interfaces. They thus can readily exchange schematic information, and query the repository for more detailed schema information (e.g., schema element names and documentation) as needed during the exploration of schema similarity.

Affinity

Affinity accepts a set of schemas as input, clusters them, and then visually renders these schemas in the context of their cluster relationships as illustrated in Figure 2.

Clustering. Affinity initially computes a matrix of interschema distances. Jaccard's distance is computed over each schema pair, based on key words extracted from the schema (with stemming and stopword removal). The resulting distances correspond well to the intuition of domain experts. Schemas are then merged into successively larger groups via hierarchical agglomerative clustering (complete linkage) [2] until the algorithm terminates, producing a schema cluster tree whose root is a single cluster containing all schemas.

Visualization. Affinity utilizes a novel combination of visualizations [7]. As shown in the right pane of Figure 2, Affinity renders the schema cluster tree as a *dendrogram*, a representation commonly used to visualize gene sequence clusters. As in [5], users can interactively drill down into the dendrogram. Note, however, that the order of the vertical listing is not always meaningful. For example, the two most *unrelated* items may be adjacent in vertical sequence. Thus, although dendrograms accurately reflect the result of a hierarchical clustering algorithm, important information about the proximity of schemas is not always visually obvious.

To complement this view, Affinity also lays schemas out on a 2D plane using a force-directed point placement algorithm. This 2D view sacrifices accuracy due to dimensionality reduction, but renders clustering relationships more clearly than the dendrogram. By adjusting the two slider bars, cluster information from the dendrogram is interactively superimposed onto the 2D view, giving the sense of exploring a topographic map until exactly the "strata" (granularity) of clustering of interest to the user becomes visible.

Overlap View

There are numerous ways schemas can be related with respect to overlapping content. A small schema may replicate a component of a larger schema, or two schemas of the same size may have strong element-to-element matches. Because such similarity relationships are not revealed by Affinity's clustering, we created a view to compute and display pairwise content overlap, as shown in Figure 5. Each schema is represented as a labeled circle with area proportional to its size in elements. Binary Venn diagrams are constructed with overlap proportional to the number of elements having matches, as determined by schema matching algorithms

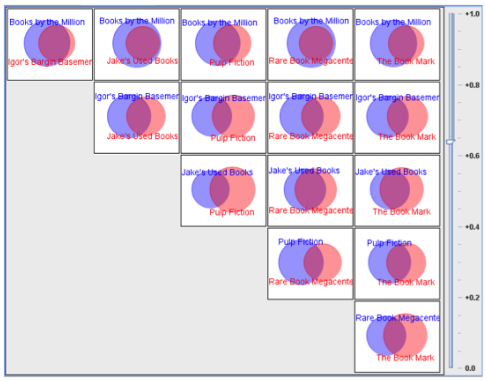

Figure 5: ***Pairwise content overlap within a cluster of 6 book schemas.***

(e.g., name similarity, synonymy). Match strength is modulated by the slider bar on the right. Mousing over any region reveals the underlying schema elements in a textbox. For example, mousing over a blue area reveals the blue schema elements having no match in the red schema at the currently selected match strength.

3. CONCLUSIONS

We demonstrate a single integrated environment in which users explore schema similarity relationships at three complementary resolutions. Schema clusters are visually rendered by a novel tool, Affinity, in an interactive linked dual-pane display. The impact for decision makers is analogous to that of OLAP systems. Instead of instance data, decision makers reason about the schemes of information systems through roll up and drill down, gaining insights into potential integration projects and data sharing partnerships.

Participating in the OpenII consortium provides an integrating framework for our tools, and enables interoperation with related contributions (e.g., Schemr [1]). We are currently providing this environment to MITRE customers with enterprises on the scale of thousands of schemas. In the near future, we plan to compare interschema distances to a manually determined gold standard.

We wish to thank Len Seligman and and Catherine Macheret.

4. REFERENCES

[1] K. Chen, J. Madhavan, and A. Halevy. Exploring schema repositories with schemr. In *SIGMOD '09*, 2009.

[2] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* John Wiley and Sons, Inc., New York, 2001.

[3] P. Mork, A. Rosenthal, L. Seligman, J. Korb, and K. Samuel. Integration workbench: Integrating schema integration tools. In *InterDB*, 2006.

[4] OpenII. Project webpage. openintegration.org, 2009.

[5] J. Seo and B. Schneiderman. Interactively exploring hierarchical clustering results. *IEEE Computer*, 35(7):80–86, July 2002.

[6] K. Smith, P. Mork, L. Seligman, A. Rosenthal, M. Morse, D. Allen, and M. Li. The role of schema matching in large enterprises. In *CIDR 09*, 2009.

[7] B. Yost, C. Bonaceto, M. Morse, C. Wolf, and K. Smith. Visualizing schema clusters for agile information sharing. In *The 2009 IEEE Information Visualization Conference (Poster)*, October 2009.