Poster: Visualizing Schema Clusters for Agile Information Sharing

Beth Yost, Craig Bonaceto, Michael Morse, Chris Wolf, Ken Smith

The MITRE Corporation

ABSTRACT

Very large enterprises present distinct information management challenges: it is difficult for their Chief Information Officers (CIOs) and information architects to obtain a clear overview of enterprise information assets and it is hard to quickly recognize opportunities for productive data sharing. To support users in these endeavors, we have created Affinity, a tool for visualizing clusters of related schemas. The contributions of this work are: 1) a new application of clustering, schema clustering, to enable rapid and agile information sharing, 2) a visualization technique combining hierarchical clustering and 2D point placement and 3) an open source tool for visualizing schema clusters.

KEYWORDS: Information visualization, data integration, schema matching, multidimensional visualization, hierarchical clustering.

1 MOTIVATION

Very large enterprises, such as governments, present CIOs and information architects with distinctive challenges. For example, the United States Department of Defense Metadata Registry (MDR) is a repository meant to support the reuse of information assets that contains *thousands* of schemas (i.e., the data models of specific assets). At this scale, it is cognitively challenging to rapidly: a) discover relevant information assets, b) recognize groups of related information assets reflecting "communities of interest" among their owning organizations, and c) recognize promising opportunities for data sharing and integration.

At the heart of these challenges is the need to recognize relationships within large groups of schemas, and the information assets and organizations they represent. As stated in [1] "the ability to identify clusters of related schemas is vital, providing CIOs with a big picture view of enterprise data sources and revealing to integration planners the most promising (i.e., tightly clustered) candidates for integration". The goal of *schema cluster visualization* is to provide insight into important relationships among enterprise information assets and to make the most promising opportunities for data sharing clearly visible.

To address these challenges, we created Affinity, a tool for visualizing clusters of schemas. Affinity takes multiple schemas as input and calculates the similarity between schemas using Jaccard's distance (i.e., the overlap in terms between each pair of schemas). Using these similarity values, agglomerative hierarchical clustering is used to group schemas. Decision makers then use interactive visualizations to understand and explore the results. The use of Affinity is a decision maker's first step in the process of information sharing and integration using a larger open source framework [2]. In this paper we describe Affinity, the visualization techniques used in Affinity, and future directions.

202 Burlington Road, Bedford, MA 01730
{bethyost, cbonaceto, mdmorse, cwolf, kps}@mitre.org

2 AFFINITY

Affinity is written in Java and uses the Standard Widget Toolkit (SWT) [3]. Affinity can accept schemas from a variety of sources; we currently select schemas within the SchemaStore repository that is part of a larger open source framework for information integration [2]. After a user selects a subset of schemas from a list of all available schemas in the repository, the similarity matrix generation and clustering are performed automatically.

2.1 Dendrogram View

The results of the schema clustering are visualized using two linked views as shown in Figure 1. The view on the right is a standard hierarchical clustering visualization, the dendrogram, which displays the results of the hierarchical clustering. Similar to the Hierarchical Clustering Explorer [4], Affinity includes an adjustable minimum similarity bar in the dendrogram (represented using the vertical dotted line) to show all clusters at a specified level of similarity. As this bar is moved to the right, more inclusive clusters are highlighted. As the bar is moved to the left, more exclusive clusters are highlighted. The setting a user chooses for this bar will depend on their desire to identify larger more loosely related groups, or smaller more similar groups of information assets for data integration or communities of interest.

To assist in studying the formation (or dissolution) of groups, forward and back buttons are provided to enable users to easily step through the clustering algorithm. Affinity also uses alternating colors to help users visually separate adjacent clusters; without these, users must follow the horizontal lines across the display to determine which schemas are in the same cluster.

2.2 2D View with Superimposed Clustering

In addition to the dendrogram view, a more familiar representation of similarity is provided (i.e., the left view in Figure 1). This view represents each schema as a point in a 2D space where the proximity of two points represents the similarity between those schemas. This means that similar information assets for data integration or communities of interest will generally be located in close physical proximity. We currently use the Force Directed Placement algorithm provided by Prefuse to calculate point placement [5]. The 2D view has the advantage of being a more familiar representation for conveying clustering and similarity than the dendrogram. The point placement is based on the same similarity data, but is independent from and therefore complementary to the hierarchical clustering algorithm used by the dendrogram.

To show the correspondence between the 2D layout and the hierarchical clustering, we superimpose the hierarchical clusters onto the 2D view, as illustrated in the left of Figure 1. Our approach is similar in spirit to Shepard's hand-drawn graphics, where contours representing hierarchical clustering are embedded in the two-dimensional spatial representation [7]. Paulovich and Minghim have also combined 2D placement with hierarchical clustering information in HiPP [6]. Our approach differs from HiPP's in that we use hierarchical clustering information as a

means of dynamically linking two related views of the same data (the similarity matrix), without obscuring the features of either.

When the results of the 2D placement and hierarchical clustering are similar, the superimposed 2D view in Affinity resembles a topographic map in which "peaks" correspond to the most exclusive clusters. Where point placement did not correspond well to the hierarchical clustering, a user can see that the contours are not hierarchical. For example, the *yana* and *igor* schemas form a cluster in the dendrogram but are not closest in proximity to each other in the 2D view shown in Figure 1.

Affinity includes two different options for embedding the hierarchical clusters in the spatial layout. The approach shown in Figure 1 has no concavities. However, for larger datasets, we provide the option of viewing the contours in a way that minimizes area by tightly circling clusters. Figure 2 shows an example of tightly circling clusters by drawing contours around clusters that follow concavities.

Affinity includes a minimum similarity filter that can be used to remove information about "valleys" in the 2D view, which can help bring smaller tighter clusters into focus. The maximum similarity slider bar removes information about "peaks", bringing the larger clusters into focus. Modulating both sliders reveals clusters at the desired resolution.

3 Future Work

Affinity is an open source tool that is still being developed, and we are refining it through interaction with governmental agencies. We are adding options such as: the ability to click on a cluster to see a summary of the most distinctive terms shared by schemas in that group, b) the ability to "drill down" and visualize the overlap in matched terms (using a related tool, Harmony, also in [8]) between any two selected schemas, and eventually visualization of matching attributes across n-schemas. The code for Affinity and related resources is freely available from the Google Code repository [8].

References

[1] K. Smith, P. Mork, L. Seligman, A. Rosenthal, M. Morse, C. Wolf, D. Allen, and M. Li. The Role of Schema Matching in Large Enterprises. In *4th Biennial Conference on Innovative Data Systems Research (CIDR)* January 4-7, 2009, Asilomar, California, USA.

[2] http://sites.google.com/site/openinformationintegration/

[3] http://www.eclipse.org/swt/

[4] J. Seo and B. Shneiderman. Interactively Exploring Hierarchical Clustering Results. *IEEE Computer*, 35(7):80-86, July 2002.

[5] http://prefuse.org/doc/api/prefuse/action/layout/graph/ForceDirectedLayout.html

[6] F.V. Paulovich and R. Minghim. HiPP: A Novel Hierarchical Point Placement Strategy and its Application to the Exploration of Document Collections. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1229-1236, November/December 2008.

[7] R.N. Shepard. Representation of Structure in Similarity Data: Problems and Prospects. *Psychometrika*, 39(4):373-421, 1974.

[8] http://code.google.com/p/openii/

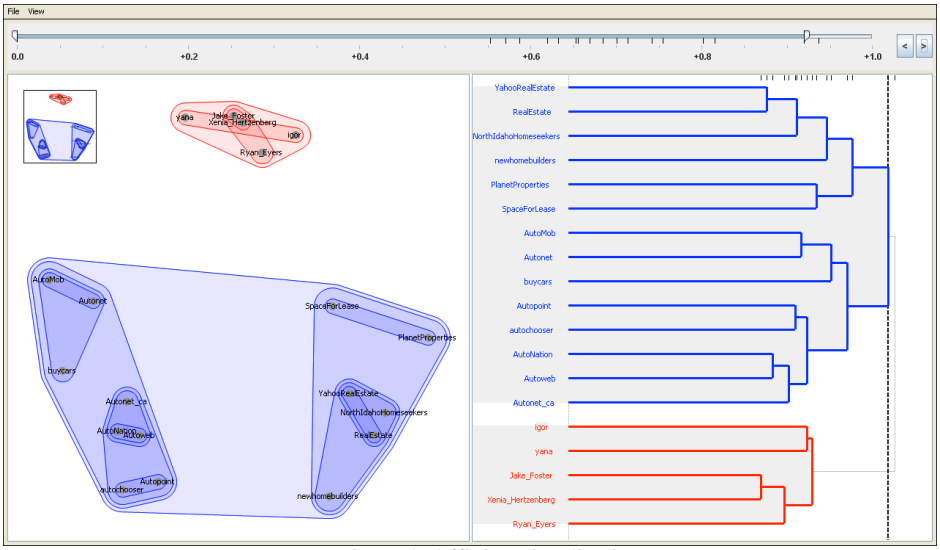

Figure 1. Affinity Visualization

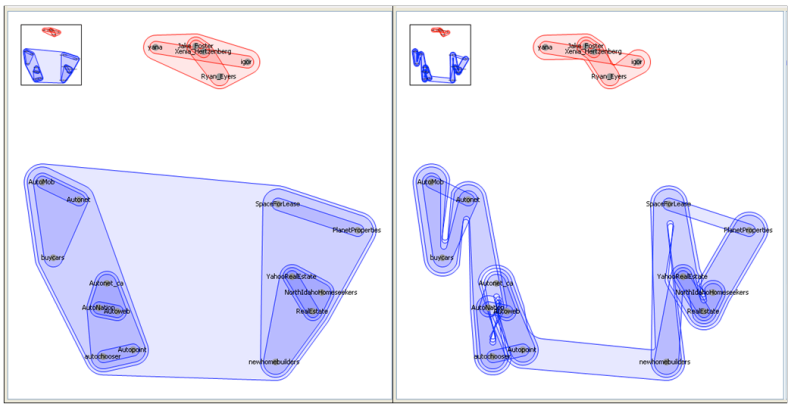

Figure 2. Superimposed Clustering without (left) and with (right) Concavity