ClusTR: Exploring Multivariate Cluster Correlations and Topic Trends

Luigi Di Caro[1] and Alejandro Jaimes[2]

[1] Universita' di Torino, Torino, Italy
[2] Telefonica Research, Madrid, Spain
dicaro@di.unito.it, ajaimes@tid.es

Abstract. We present a demonstration of ClusTR, a highly interactive system for exploring relationships between different clusterings of a dataset and for viewing the evolution in time of topics (e.g., tags associated with objects in the dataset) within and across such clusters. In particular, ClusTR allows exploration of generic multi-dimensional, text labeled and time sensitive data.

1 Introduction

In many applications (e.g., bioinformatics, telecommunications, marketing, etc.) the focus of Data Mining is often on Knowledge Discovery (KD), largely because of the complexity of the data and the heterogeneity in terms of attributes and possible groupings of objects. The same reasons that create the needs for a focus on Knowledge Discovery, however, make it a very challenging process. On one hand, it is an activity that cannot be fully automated—by definition, the discovery process aims to gain insight from the data and therefore implies a user-in-the-loop approach. On the other hand, except in very simple cases, insight cannot be gained without often significant automatic processing intertwined in the human analysis loop. The combination of automatic techniques and human analysis is thus central to the KD process and the two could offer new potentiality in both areas: presenting information in a way that can exploit human perceptual skills, and user input to improve automatic results. Thus, a large part of the KD process focuses on discovering and understanding of not only similarities between objects, but also on how different groupings of objects, based on different attributes relate to each other. Furthermore, in cases in which time is one of the attributes and objects have textual attributes, it is also often important to discover the evolution of such textual attributes (e.g., topics or categories) over time with respect to the possible groupings of the objects.

In this paper, we address these problems with ClusTR, a highly interactive system for exploring relationships between different clusterings of a dataset and for viewing the evolution in time of topics (e.g., tags associated with objects in the dataset) within and across such clusters. In particular, ClusTR allows exploration of generic multi-dimensional, text labeled and time-sensitive data. Our system follows the goals of visual analysis tasks [9], as well as user interface design principles (e.g., overview first, zoom-and-filter, then details-on-demand and others [3]), placing strong emphasis on a highly interactive environment that combines automatic techniques with human analysis. In particular, we apply the concept of dynamic queries [3],

W. Buntine et al. (Eds.): ECML PKDD 2009, Part II, LNAI 5782, pp. 722–725, 2009.

which continuously update the data that is visualized: user actions work instantly, triggering clustering, filtering, and allowing other operations in the database.

Related work. There is a large body of work on interactive KD, the most relevant being in the areas of interactive exploration and queries, data visualization and temporal data exploration. Related systems and approaches include [2][1][10][11][4] [6][7][8], and [14].

2 ClusTR

Our system works with any kind of multi-dimensional data, but some of the functionality was implemented specifically for data that includes time and textual features. For the demonstration we use a freely available Flickr image dataset [13] containing 25,000 images where each image has the following features: textual tags, geo-tag coordinates, and camera settings (aperture, exposure, ISO, and a Flash use flag). In the current implementation we use EM for clustering, which assigns a probability distribution to each instance that indicates the probability of it belonging to each of the clusters. The value of correlation between two clusters is calculated using the Jaccard coefficient (the size of the intersection divided by the size of the union).

Interaction with the system takes place as follows. The first step is selection of the data for analysis. After data is loaded into ClusTR, exploration consists of two stages: (1) selection of features to create multiple cluster groups (Figure 1); and (2) exploration of correlations between resulting clusters and analysis of time-evolution of topics within clusters (Figure 2). We describe the two stages below.

Given k features (figure 1) in the dataset, in the first interaction screen the system creates a correlation panel of k x k squares. The user selects features by drawing sets of dots (D) and lines (L), whose combination determine the features to use in the creating cluster groups. In figure 1, for example, two cluster groups are created, one using flash, ISO and one using ISO, aperture, and exposure. The two cluster groups are created on-line when the user presses the "next" button (each cluster group contains the same data). Each dot and each line drawn by the user defines one combination of features to create a cluster group.

The drawing style of interaction used is meant to intuitively allow quick selection of features instead of cumbersome interaction with boxes or lists. ClusTR plots the value distributions for the dataset for each feature (Figure 1, right) to give the user quick

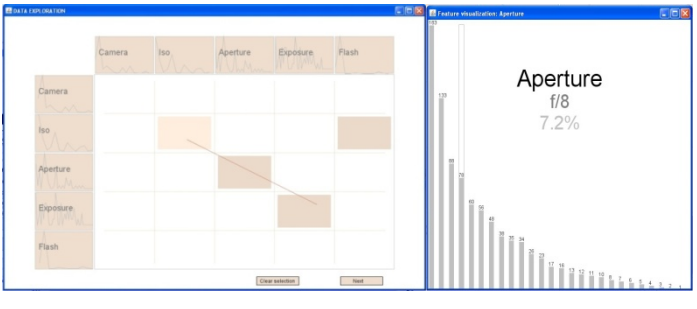

Fig. 1. The panel for selecting the features to create cluster groups

insight into the variability of the data in each feature space. The correlation layout has the advantage of allowing us to graphically represent correlations between features.

In the second stage, each of the clusters is represented by a solid circle and a label on the outside of a circle (Figure 2). The circular interface aims at showing correlations between clusters, while the bottom panel (bottom of Figure 2) shows topic trends. Each cluster group is represented by a different color. Horizontal bars on the outside of the circle associated with each cluster indicate the size of the cluster relative to its corresponding group. Lines between clusters represent correlations (thicker edge implies higher correlation). User interaction takes place in the following ways:

- Clicking anywhere on the screen, holding the button and moving the mouse up or down: the threshold for correlation is modified and shown. Lines dynamically (immediately) appear, disappear, increase or decrease thickness.
- Clicking on a cluster (solid circle) shows details of the cluster (topic evolution described next, and a tag cloud with their most frequent values).

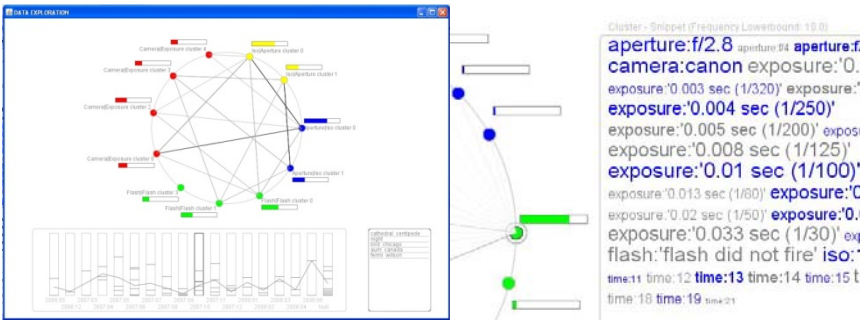

Fig. 2. Visualization of correlation strengths and evolution of topics for the clustered data. When the user selects a specific cluster, the system shows a dynamic tag cloud with the most frequent contents in the selected cluster.

Note that for clusters with the same color there's no edge/correlation given that they belong to the same clustering group so they contain distinct instances. For data that contains a textual feature (e.g., tag, free-text, etc.), it is possible to consider such values to be topics, or to extract them using a variety of techniques. For this demonstration we focus on the tag features available in the data (each image has a number of tags associated with it) and refer to them as *topics*.

The topic evolution component of the interface complements the cluster correlation exploration. The goal of this component is to show the trends in time of the topics within clusters and to compare topic trends across clusters. Topics are analyzed according to the level of co-occurrence in different periods of time using Latent Semantic Analysis [15], mapping them into a one-dimensional space **T** as in [12], in a way that their relative latent semantic distances are preserved. The bottom panel of the interface shows the general evolution of the topics over time (bottom of Figure 2). Each vertical bar represents a single period of time (e.g., a day, week, season, etc.). The horizontal lines within each bar represent the topics of items in the cluster and the vertical location of each line represents its value in **T**. The user can click over one

single period of time to have an insight of the topics/labels contained in it. In addition, the system allows the user to explore the differences in topic trends for different clusters. When the user clicks on a cluster in the circle, its topics trace is projected in the topics evolution panel. In this way, the user can visually relate clusters to time, seeing how their topic evolution overlaps.

3 Conclusions and Future Work

We have presented a demonstration of ClusTR, a highly interactive system for exploring relationships between different clusterings of a dataset and for viewing the evolution of topics within and across clusters. The main features of the system are its interactive-query mechanisms, the functionality to select features for clusterings using a 2D correlation space, and allowing the dynamic exploration of topics over time for multiple clusters. Future work includes further addition to the functionality (e.g., adding several clustering methods, more visualization options, etc.), a user study, case studies with other datasets, etc.

References

1. Ahn, J.-w., Brusilovsky, P., Grady, J., He, D., Syn, S.Y.: Open user profiles for adaptive news systems: help or harm? Int. conf. WWW 2007 (2007)
2. Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B., Williams, J.G.: Visualisation of a document collection: The VIBE system. In: Inf. Proc. and Managem. (1993)
3. Shneiderman, B.: Designing the User Interface. Addison Wesley, Reading (1997)
4. Zytkow, J.M., Rauch, J.: Circle Graphs: New Visualization Tools for Text-Mining. In: Żytkow, J.M., Rauch, J. (eds.) PKDD 1999. LNCS (LNAI), vol. 1704, pp. 277–282. Springer, Heidelberg (1999)
5. Kliger, J.: Methods for Visualizing User Models. MIT Media Lab, Cambridge
6. Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., Plaisant, C.: Discovering interesting usage patterns in text collections: Integrating text mining with visualization. HCIL Technical report 2007-08
7. Seo, J., Shneiderman, B.: A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections. In: Infovis 2004 (2004)
8. Inselberg, A.: Visual Data Mining with Parallel Coordinates. Comp. Statistics (1998)
9. Nocke, T., Schumann, H.: Goals of Analysis for Visualization and Visual Data Mining Tasks. In: CODATA Workshop Information, Presentation and Design (2004)
10. Catarci, T., Costabile, M.F., Levialdi, S., Batini, C.: Visual Query Systems for Databases: A Survey. Journal of Visual Languages and Computing 8
11. Derthick, M., Kolojejchick, J., Roth, S.F.: An Interactive Visual Query Environment for Exploring Data
12. Qi, Y., Candan, K.S.: CUTS: CUrvature-Based Development Pattern Analysis and Segmentation for Blogs and other Text Streams. In: Hypertext 2006 (2006)
13. Hu., M.J., Lew, M.S.: The MIR Flickr Retrieval Evaluation
14. Berchtold, S., Jagadish, H.V., Ross, K.A.: Independence Diagrams: A Technique for Visual Data Mining. AT&T Laboratories (1998)
15. Deerwester, S., Dumais, S., Furnas, G., Harshman, R., Landauer, T., Lochbaum, K., Streeter, L.: Computer Information Retrieval using Latent Semantic Structure