Visalix: A Web Application for Visual Data Analysis and Clustering

Loïc Lecerf
Xerox Research Centre Europe
6, chemin de Maupertuis
38240, Meylan, France
loic.lecerf@xrce.xerox.com

Boris Chidlovskii
Xerox Research Centre Europe
6, chemin de Maupertuis
38240, Meylan, France
boris.chidlovskii@xrce.xerox.com

ABSTRACT

This paper presents Visalix, a Web-based interface aimed at facilitating human-computer cooperation in complex data analysis tasks. It implements an *interactive visualization* paradigm which assists users in matching their domain knowledge with the algorithmic power of data analysis and mining techniques. Visalix integrates a number of Visual Interactive Learning components for better understanding, easier interpreting complex datasets, and training prediction models.

Keywords

Data visualization, star coordinates, visual clustering, interactive learning

1. INTRODUCTION

The growing use of information visualization tools and data mining algorithms stems from two parallel lines of research. Researchers in information visualization believe in the importance of offering users various views and insights about the data distributions, while data mining researchers believe that statistical algorithms and machine learning can be relied on to find the interesting patterns [5].

We propose a new framework that integrates the efficiency of automatic machine learning algorithm and the flexibility of information visualization tools. We design Visalix, a Web application for visual interactive data analysis, clustering and annotation.

The core of Visalix is a visual clustering component which offers a 3D interactive projection of data [4] and possibilities to manipulate it. Beyond clustering, Visalix is designed to help users in data annotation tasks. Users can analyze a dataset, cluster items manually or by automatic optimization, annotate items, manage the label set, make label predictions for (yet) unannotated items and visualize them, choose the most relevant item to improve current prediction models, etc.

KDD'09, June 29-July 1, 2009, Paris, France.

Visalix is designed to be domain independent. The system copes with datasets where items are described by sets of characteristics (features) and may come with their textual or visual representation. The Visalix is available at `http://visalix.xrce.xerox.com`. Five datasets and two videos are available in the Visalix site for the demonstration purposes; they have been created for different tasks in document analysis and annotation, image categorization and medical diagnostics. Likewise, users can upload their datasets through the Web interface, visualize and analyze them.

1.1 Data Visualization Component

Semi-Supervised Star Coordinates [2] is an extension of star coordinates [3] to the spherical view and for semi-supervised learning cases, when a small part of labeled data coexists with a large part on unlabeled items. Star coordinates is an original approach to visualize and manipulate multidimensional data in a 2D or 3D space. Beyond the conventional manual settings, it enriches visual clustering with automatic settings where the projection distance metric is learned from the available set of user feedback in the form of either item similarities or direct item labels. It combines the advanced data analysis of automatic clustering with the flexibility of interactive visual clustering.

The *spherical coordinates* (SC) visualization model consists of *max-min* normalization followed by α-mapping. Max-min normalization addresses large-valued features and scale them in $[-1, 1]$ range. Then, α-mapping projects d-dimensional points into 3D space for the convenience of visual parameter tuning.

Let a point $Q(x, y, z)$ represent the image of a d-dimensional max-min normalized data point, $P(v_1, \ldots, v_d), v_i \in [-1, 1]$, in 3D space. $Q(x, y, z)$ is determined by the average of the vector sum of the d vectors $sc_i \cdot v_i$, where $sc_i = (cos(\theta_i), sin(\theta_i)sin(\phi_i), sin(\theta_i)cos(\theta_i))$, $i = 1, \ldots, d$, and $\theta_i \in [0, 2\pi]$ are the spherical coordinates that represent d dimensions in 3D visual space. According to α-mapping, a 3D projection point $Q(x, y, z)$ is determined as follows

$$Q(x,y,z) = \frac{1}{d}\begin{pmatrix} \sum_{i=1}^{d} \alpha_i v_i cos(\theta_i) - x_0 \\ \sum_{i=1}^{d} \alpha_i v_i sin(\theta_i) sin(\phi_i) - y_0 \\ \sum_{i=1}^{d} \alpha_i v_i sin(\theta_i) sin(\phi_i) - z_0 \end{pmatrix} \quad (1)$$

Here, the vector $\alpha = [\alpha_1, \ldots, \alpha_d], \alpha_i \in [-1, 1]$, provides the *scaling* adjustable parameters, one for each of the d dimensions; *rotation* parameters θ_i and ϕ_i are initially set to $2i\pi/d$ and can be adjusted afterwards. Point $o = (x_0, y_0, z_0)$ refers to the center of the display area. α-mapping is a linear mapping with a fixed set of α, θ, ϕ values. Figure 1 presents a projection of a dataset with $d = 8$ in the 3D spherical coordinates space. Arrows indicate feature axes (8 axes in total) and date items from 5 different classes; different colours are associated with different classes.

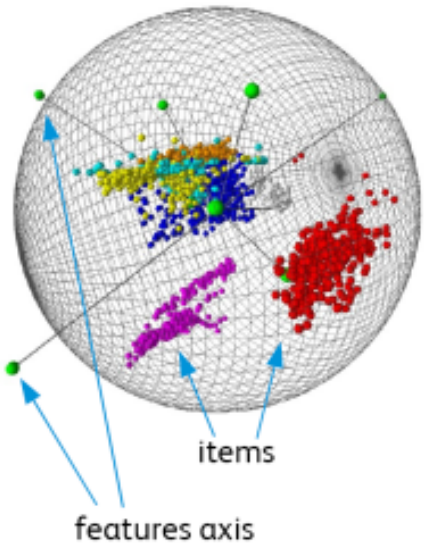

Figure 1: 3D Star Coordinates model.

The visual clustering visualization component provides a number of interaction features, which users can utilize to improve their understanding of the datasets. The basic features of the visual clustering in the following:

Scaling: Scaling transformation allows users to change the length of an axis, thus increasing or decreasing the contribution of a particular data attribute on the resultant visualization.

Rotations: Rotation transformation modifies the direction of the unit vector of an axis, thus making a particular data attribute more or less correlated with other attributes. When multiple axes are rotated to point in about the same direction, their contributions are effectively aggregated in the visualization. As a result, these transformations provide a restructuring of the data points in the visualization based on the criteria chosen.

View point and zoom: Users can rotate the view point and zoom in/out to better observe and analyze the current data scene. Mouse manipulations to change the view point and to zoom are intuitive and straightforward. Once positioned on the display area, left mouse drags rotate the view, while the right mouse drags move the scene nearer or farther from the viewer. 3D visualization model gives an important advantage over 2D models used so far and facilitate the neater dataset understanding. This additional dimension allows the 3D "fly-over" effect similar to the 3D Map View used in advanced Web applications.

Dimension selection: If the initial dataset contains too many dimensions (which are entity features), users may want to preliminary discard the less relevant ones. The number of dimensions may influence both the easy of manual cluster tuning and, to less extend, the complexity of algorithms. The Visalix provides a function to reduce the number of dimensions of your dataset, using the Principal Component Analysis (PCA) method.

Learning an optimal distance metric: Semi-supervised clustering assumes that we dispose the user feedback in the form of either direct item labels or item (dis-)similarities [1]. Using this feedback, the better clustering can be achieved by adjusting the underlying *distance metric*, thus allowing to capture the user's view of which items should be put together or apart. The original data representation may not be embedded in a space where clusters are clearly separated. Modifying the distance metric transforms this representation in such a way that the distance between same-cluster items is minimized, while distances between different-cluster items are maximized. In Visalix, the optimal projection distance metric M for the spherical coordinates is obtained by using canonical variates which generalize the Fisher Discriminant Analysis (FDA) to C classes and 3 projected dimensions. The projection in Figure 1 is actually obtained by learning the optimal distance metric from the dataset and item labels. The metric provides the optimal values of α, θ, ϕ for the best separation of classes in the projections space.

1.2 Model Uncertainty Visualization

In the semi-supervised learning, a particular place is devoted to the active learning. This approach plays a particular role when the data labeling is expensive and one needs to minimize the labeling effort. Different methods have been developed to guide the user through the labeling process. In Visalix, we propose a *Visual Active Learning* component which combines the active learning with the visualization. It represents unlabeled data in an uncertainty space and allows the user to choose the next element to label. The model is very intuitive in representing the uncertainty of a current model with respect to a dataset. This approach is complementary to the conventional active learning, since the user can combine the model predictions with her domain knowledge and insights.

The Model Uncertainty Visualization copes with *class points* and *data points* and presents in the 3D uncertainty space. Class points are shown as small spheres; each class $c \in C$ has its point $Y^c = (x^c, y^c, z^c)$, with its unique color; C denotes the set of possible classes. Data points are the representation of unlabeled items in the uncertainty space. The position and the color of a data point are computed in the function of the class predictions for the corresponding item.

Data point positions. The position (x_i, y_i, z_i) of a data point X_i in the uncertainty space is defined as a function of the model uncertainty and (current) positioned of class points Y^c, as follows:

$$(x_i, y_i, z_i) = \left(\sum_c W^c P_i^c x^c, \sum_c W^c P_i^c y^c, \sum_c W^c P_i^c z^c\right),$$

where $P_i^c = P(c|X_i)$ are class probabilities for X_i according to the current prediction model, and W^c are weights of class

points $Y^c = (x^c, y^c, z^c)$. By default, weights W^c are set to 1 for all classes, but the user may manually adjust them according to her particular interest.

Data point colors. The higher the probability P_i^c of labeling item X_i with class c is, the closer the data point of X_i will be to the class point Y_c. When the number of classes is high, using the positions only may be insufficient to get a right insight on the model uncertainty. Thus we complete the point positions with colors in order to ease the interpretation of uncertainty.

Class point colors are expressed as color values in the RBG color space (R for red, G for green and B for blue). Colors for class points are set by default and may be adjusted by the user. The color values for the data points are computed from the colors of class points Y^c to reflect the probability assigned by the model that the respective item is labeled with class c. Any color value is expressed in the RGB color space; it indicates how much of red (R), green (G), and blue (B) color is included.

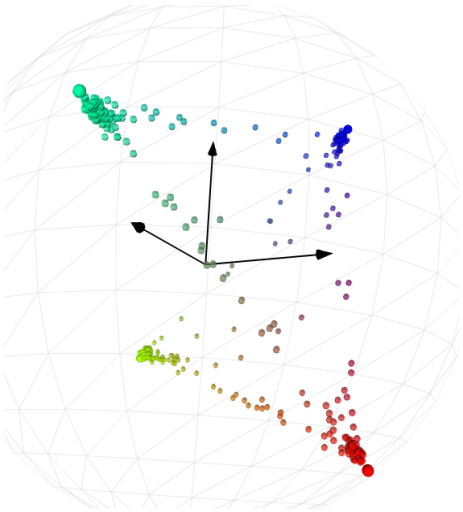

Figure 2: Model Uncertainty Visualization example.

The color of class point Y_c is defined by its three components (R_c, G_c, B_c). The color (R_i, G_i, B_i) of a data point X_i is a mixture of class colors weighted by the class probabilities, as defined below:

$$(R_i, G_i, B_i) = \left(\sum_c W^c P_i^c R^c, \sum_c W^c P_i^c G^c, \sum_c W^c P_i^c B^c\right).$$

Figure 2 shows an example of uncertainty space with the model predictions for about 300 unlabeled items and 4 classes.

This visualization provides a simple insight on the model "quality" and a deeper knowledge of model (un)certainty. In the labeling/annotating process, the annotator may either follow the system suggestion or choose the most relevant item to label herself. After each annotation (or a group of annotations), the prediction model is augmented with the newly labeled elements, and the uncertainty representation is updated accordingly. Visalix additionally offers the user a possibility to move class points and colors as well as change their weights.

1.3 Item Interpretation

Each item in the dataset can be associated with a textual or visual representation. This considerably simplify the data interpretation during the analysis and labeling process. In the current version, Visalix supports four different types of association:

Data: This default option supports the association of an item with the set of its descriptors.

Text: Another standard option is to associate an item to a text, such a word, phrase or a plain text document. This support allows to address various text clustering and classification problems, like spam detection or e-mail categorization.

Images: Yet another useful support stems from the image domain. Images assicated to items might be generic or special one, like country flags, face photos, a business logo, etc. Unlike data and text supports, Visalix counts on the user or the external routines for the feature extraction from images.

Object zones: A particular attention has been paid in Visalix to the task of inner-document annotation. This includes metadata annotation, as well as annotating graphical objects such as pictures, diagrams and tables. Document fragments are addressed the zone definition mechanism. In Visalix, these zones are defined as a set of polygons.

Figure 3 shows the Visalix Web interface and its application to the document annotation task. The interface is composed of the two views: the left window shows the 3D visual space; the right window associates data items to document elements, where an association includes a document zone and a textual fragment. The two windows get synchronized, any action or item labeling in one window is promptly propagated tn another one. Alternative cases of text, image and zone associations can be found on the Visalix site.

1.4 Item annotation

While each class is represented by its proper color, the unlabeled items remain either shaded in 3D projection space or colored in the 3D uncertainty space. The annotation of items can be done in the projection space, in the uncertainty space or using the item representation. To choose a class for the item, user can activate a context menu associated with each item. The annotation of item groups is yet implemented in the Web version of Visalix. The local version offers a mechanism allowing the user to select a group of data points using a 3D convex hull envelope.

Visalix is designed to easily manage the annotation classes. The user can define classes through in the interface; and change them during the annotation process, by adding a new class or by deleting an old one. From the available set of labeled items, a prediction model can be learned, using the logistic regression or SVM algorithm.

2. A WEB APPLICATION

The Web application of Visalix is composed of two major parts which communicate using Micromedia ActionScript Message Format (AMF) (see Figure 4). On one hand, Visalix uses Adobe Flash and Flex for the GUI component. On

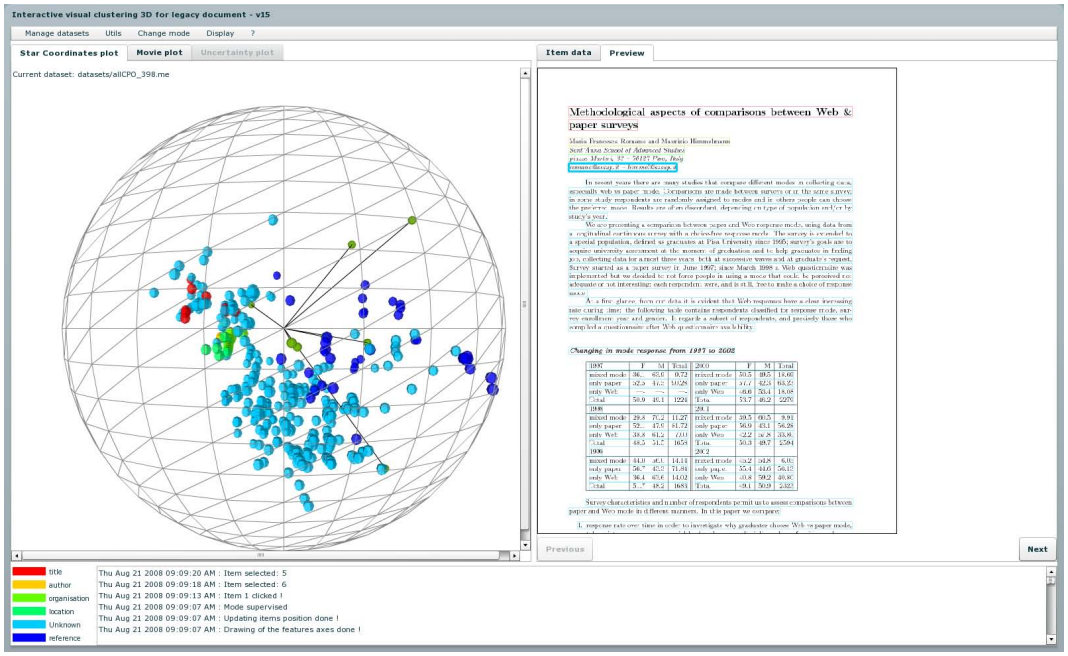

Figure 3: Document annotation with Visalix.

the other hand, a Python server manages datasets and calls Python packages to process data, to project them in the 3D visual space, to train prediction models, etc.

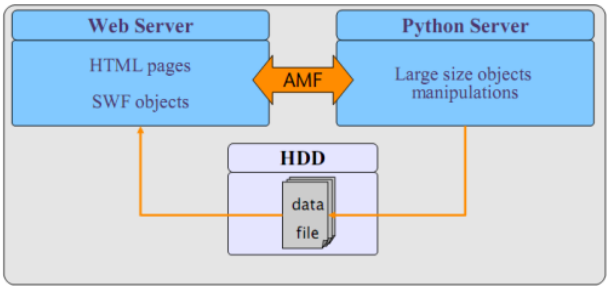

Figure 4: Visalix communication channel.

Python scientific libraries. The Python scripting language has recently seen an intensive extension toward scientific programming. Several Python packages such as Scientific Python (SciPy) allow to create a user friendly and application specific environment and to integrate an access to a range of efficient algorithms for matrix algebra and numerical calculus.

Flash and Flex technologies. Adobe Flash is commonly used to create animation, advertisements, and various web page components, to integrate video into web pages, and more recently, to develop rich Internet applications. Adobe Flex is a collection of technologies released by Adobe Systems for the development and deployment of cross-platform rich Internet applications based on the Adobe Flash platform. Flex applications provide a state-based client where significant changes require no page reloading. Similarly, Flex and Flash Player provide many useful ways to send and load data to and from server-side components without requiring the client to reload the view. Flash player plugin on 99% of browsers and it own rich 3D libraries.

3. CONCLUSIONS

We have presented Visalix, a Web application for the visual data analysis, clustering and annotation. The system is aimed at combining advanced data analysis algorithms and with the 3D visual interaction paradigm. Visalix is a work in progress. In the future, we intend to add more components, methods, support for other media like audio, video, Web content, and to extend to other application domains.

4. ACKNOWLEDGMENTS

This work has been partially funded by the French National Association for Research and Technology (ATASH project).

5. REFERENCES

[1] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD '04: Proc. 10th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, pages 59–68, 2004.

[2] B. Chidlovskii and L. Lecerf. Semi-supervised visual clustering for spherical coordinates systems. In *SAC '08: Proc. ACM Symposium on Applied computing*, pages 891–895, 2008.

[3] E. Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *KDD '01: Proc. 7th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, pages 107–116, 2001.

[4] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, Jan/Mar 2002.

[5] B. Shneiderman. Inventing discovery tools: combining information visualization with data mining? *Information Visualization*, 1(1):5–12, 2002.