Fast Interactive Visualization for Multivariate Data Exploration

Changhyun Lee
School of Computational
Science and Engineering
Georgia Tech
Atlanta, GA 30332, USA
clee407@gatech.edu

Wei Zhuo
College of Computing
Georgia Tech
Atlanta, GA 30332, USA
wzhuo3@cc.gatech.edu

Jaegul Choo
School of Computational
Science and Engineering
Georgia Tech
Atlanta, GA 30332, USA
jaegul.choo@cc.gatech.edu

Duen Horng (Polo) Chau
School of Computational
Science and Engineering
Georgia Tech
Atlanta, GA 30332, USA
polo@gatech.edu

Haesun Park
School of Computational
Science and Engineering
Georgia Tech
Atlanta, GA 30332, USA
hpark@cc.gatech.edu

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract

We are investigating a fast layout method for visualizing and exploring relationships between multivariate data items. We improve on existing works that use the force-directed layout, which has high running time and cannot scale up for large-scale visual analysis. Our method, based on *Mean Value Coordinates*, has a closed-form solution that can determine items' locations in a single iteration. In addition, it has a fast running time that is linear in the number of items. We are also exploring multiple interactive visualization techniques to help users make sense of the data, such as blending multiple heat maps to simultaneously express multiple types of data distributions; and techniques to create topics, and to merge or split topics in real time.

Author Keywords

Multivariate data visualization; visual analytics; interacting data exploration

ACM Classification Keywords

H.4 [Information Systems Applications]: User Interfaces; I.3.6 [Computer Graphics]: Interaction techniques

General Terms

Algorithms, Design, Human Factors

Introduction

One of the widely-used methods to visualize the relationships between data items in a screen space is the *force-directed graph layout*, which places items on the screen such that the (Euclidean) distance between each pair of items roughly corresponds to their given distance.

However, as the number of items that we want to visualize increases, the force-directed graph layout faces several challenges. First, the layout can no longer accurately preserve all pairwise relationships, which can contain maximally n^2 relationships for n data items. Second, even if all relationships are correctly mapped in the layout, the resulting visualization can be visually overwhelming, which greatly impede human understanding.

To tackle these problems, a notable line of research focuses on incorporating some additional meaningful objects or concepts as new nodes, called *anchors*, and let users manipulate these anchor nodes' positions in the layout. Accordingly, the layout algorithm then updates the positions of data items based on their relationships with respect to the anchor nodes. For example, suppose we have *documents* as our data items, and we add *topics* as our anchor nodes; then in the layout, documents that are highly related to a particular topic would be drawn closer to that topic's anchor node (see Figure 1 for an example).

Representative works that use similar ideas include the *Dust & Magnet* system (DnM) [8] and the TopicViz system [3], where the anchor nodes correspond to individual variables and semantically meaningful topics in document data, respectively. Parallel Coordinates and Graph-based tools are also used in analyzing multivariate data [2]. A major benefit of the proposed approach is that, instead of n^2 relationships, users can focus their analyses on much fewer relationships, only $n \times k$ of them, where k is the number of anchor nodes.

However, a major problem still remains, which is that the force-directed layout algorithm does not scale computationally, when we have more data items, due to its quadratic running time (in the number of data items) and its iterative nature (may need many iterations until stable layout results are attained). As a result, force-directed layout may be less favorable for real-time large-scale visual analysis.

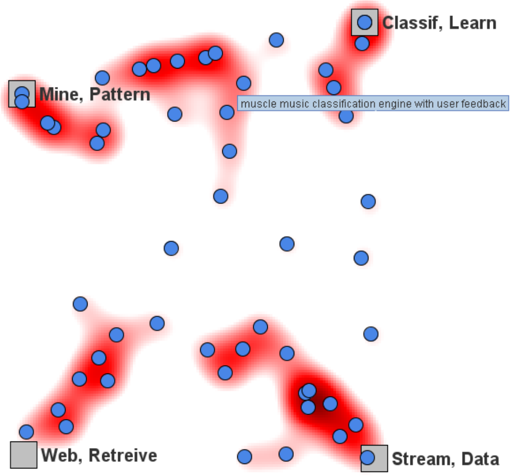


Figure 1: Main view of our proposed visualization. The squares represent anchor nodes and the circles are data items. The data items are placed by the *closed-form layout* presented in this work. The heat map represents the density of the data items. The tooltip shows a title of the data item relevant to the topics shown as '*Mine, Pattern*' and '*Classif, Learn*'.

As an alternative way to improve these problems, we propose an efficient layout algorithm using interpolation.

To be specific, unlike the force-directed layout, where the algorithm goes through many iterations, our approach requires only a single iteration, by having a closed-form solution and a linear time complexity (in the number of items). Owing to such significantly efficient computation, our approach achieves the above-described interactions almost immediately in real time for large-scale data.

As will be discussed later, a potential drawback of our proposed approach is that nontrivial overlapping between data items can occur in visualization since it does not consider any repulsive forces as in the force-directed layout. As a matter of fact, this problem arises severely throughout most approaches including even the force-directed layout when the number of data items, which are to be visualized in a limited screen space, becomes large.

To compensate this issue, we offer two techniques that represent how densely data items are placed in a particular region. The first is to utilize heat map to encode the data density as the background color. The temperature encoded as color saturation conveys the probability distributions of data items. The second is to preserve the screen real-estate of the information representation by applying offset distances to each circular node [9]. Additionally, based on the proposed approach, we provide various interactions: multiple heat maps, document-induced topic creation, merging/splitting anchor nodes.

Implementation

In this paper, we use a real-world document dataset, called *Four Area* to demonstrate our techniques. The dataset consists of papers published in four computer science areas: machine learning (ML), database (DB), data mining (DM), and information retrieval (IR), from 2001 to 2008.

From this data set, we construct an author-by-author matrix, which counts the number of papers that two different authors work together. To identify topic-wise representations of documents, we apply LDA [5] and receive the soft-clustering results representing the tendency of each author. With this result, we build relationships between authors and topics. We pick a topic's two most frequent words the topic's label (see Figure 1).

Our visualization is currently implemented in both Java (version 1.6; we also use the JUNG Java library [1]) and Javascript (with scalable vector graphics).

Method details

Scenario

We will use a scenario to explain the visualization and interaction in the context of documents/topics exploration. That is, data nodes represent *documents* and anchor nodes represent *topics*. The techniques can be applied to general multivariate data visual analysis.

1. Closed-form Layout: Mean Value Coordinates

Design of information layout presents a continuing challenge in the visualization and interaction design, motivating the development of novel techniques and approaches for multivariate data exploration. Closed-form geometric techniques are widely used for interactive visual analysis due to their simplicities and available utilities for efficient rendering). We propose to use the closed-form layout that employs linear interpolations. The inverse problem of generating this layout (i.e. using a node's position to solve for its coefficients related to polygonal control node positions) is known as the *mean value* *coordinates* (MVC) [4]. In this work, we use MVC to refer to the layouts given by linear interpolations.

2. MVC + Packing offsets
Directly using MVC for data layout may introduce significant overlapping, especially if many documents have the same coefficients. To solve this problem, we add a packing offset for each document node, and this prevents overlapping while keeping semantically similar document nodes gathered as groups. (Fig. 2 shows applying packing to one group of nodes.)

Mathematically, the layout is expressed as

$$D_i = \sum_{j=1}^{n_t} w_{ij} T_j + O_i$$

where D_i is the position of the i-th document and w_{ij} is the normalized weight coefficient of D_i to the j-th topic, T_j. O_i is the packing offset from the mean value position solved in realtime.

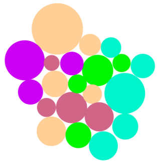


Figure 2: Packing a group of circular nodes by keeping them gathered without overlapping

In comparison, we show two layouts generated by MVC only and by the combination of MVC and packing offsets (Figure 3). An interactive comparison is available at www.cc.gatech.edu/~wzhuo3/svg/SoftClusterVis.svg.

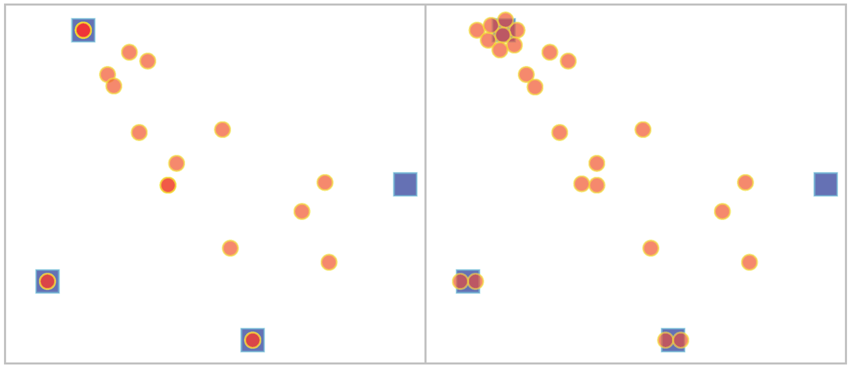


Figure 3: Packing multiple groups of nodes by applying packing offsets prevents overlapping and hence preserve the screen real estate of all the document nodes.

Dragging a topic node T_j mobilizes all related document nodes while irrelevant document nodes remain fixed.

3. MVC + Heat mapping
As the number of documents increases, the screen space eventually becomes a bottleneck with large collections of documents. Here we propose to utilize blurring and tone mapping to encode the data density as the background color, or the 'heat'. The temperature is high for regions where many documents are mapped into it by the mean value layout. Specifically, the temperature $T(P)$ at pixel location P is computed as

$$T(P) = \sum_{i=1}^{n_d} f(||D_i - P||)$$

where f is the Gaussian function with its input as the Euclidean distance between D_i and P. By precomputing the blurring kernel, the heat map is generated in real-time with arbitrary kernel size.

4. Multiple Heat Maps

Sometimes ambiguity arises if we provide only one heat map: It informs us the document distribution, but nothing about the relationships between different groups of documents. For example, different groups of documents are characterized by different coefficients but mapped to the same locations due to the projection from high dimensional weight space to 2D screen space. In fact, this is a problem with most 2D layout techniques.

We propose to address this by using multiple heat maps. For example, suppose the user is interested in two topics (Figure 4: top-left and top-right); a color is then assigned to each topic (red and blue). Documents are assigned the color of the topic that they most likely belong to (red or blue). This creates to two heat maps (document density distributions). We then blend these heat maps; visually, the user can then tell that the regions with blended colors (purple) contain documents related to both topics.

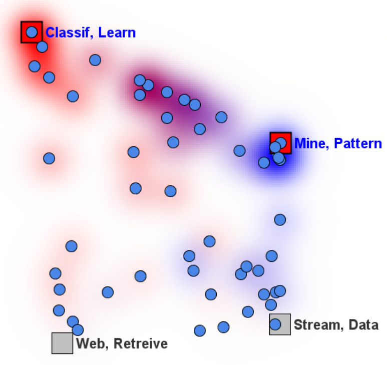
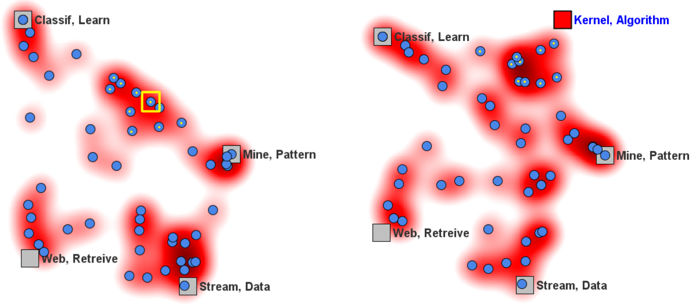


Figure 4: Two upper square nodes are selected and the heat map is generated based on these nodes. Some region has blended color, which indicates that they have relationships with both anchor nodes.

Interactions

Document-induced Topic Creation

Topics (anchor nodes) are typically identified by soft clustering methods or topic models (e.g., LDA [7]). But sometimes, these techniques may not be able to identify the right set of topics. We are exploring techniques to allow users to correct for these situations. One idea is to allow the user to create new topics based on *exemplar* documents. For example, in Figure 5a, the user sees that the high concentration of documents at the region between the two topics at the top indicates that those documents should form a new topic. The user can then induce a new topic by selecting a document (in yellow square) and create a topic based on it.



(a) Find exemplary document to become topic

(b) After the exemplary document is converted into the topic

Figure 5: A document which is indicated by yellow square (a) is converted into the topic, '*Kernel, Algorithm*' (b). Nearby documents marked with yellow dots are shown to be moved accordingly due to the new topic creation.

In Fig. 5 (a), to subdivide the topics, we select an exemplary document that is located in the middle of '*Mine, Pattern*' and '*Classif, Learn*' topics, and then convert it into the topic. Fig. 5 (b) shows that yellow marked documents that are placed around the new topic are shown to be moved depending on the new topic. This way, users can generate new topic from the set of documents which the user feel interesting.

Merging & splitting the topics
As we've said before, it is difficult to find suitable number of topics at first. To this end, our model offers merging and splitting interactions in order to adjust the number of topics.

Summary & Next Steps

We presented a significantly faster visualization method, as an alternative to the force directed layout/ We proposed to solve the overlapping problem of items by visualizing the data density in a form of shape packing or heat maps. We are exploring various interaction techniques to help the user make sense of the data, such as topic (anchor) creation, and merging and splitting.

We outline our next steps below:

- We are exploring how to enable fast heat map creation. Currently, we aggregate high-resolution kernels (one for each data item) to create heat maps. This approach's running time will hit a bottle neck when we have more than a few hundred items on the screen. We also explore alternative approachs, such as box blur, to create the heat map.
- We also plan to experiment with additional soft clustering algorithms such as NMF [6]. In addition, we will explore visualization and interaction techniques to support more topics and documents.

Acknowledgements

This work is supported in part by the project IC-Gen Rsch and the National Science Foundation grants CCF-0728812 and CCF-0808863; and the U.S. Army Research Office (ARO) and Defense Advanced Research Projects Agency (DARPA) under Contract Number W911NF-11-C-0088.

References

[1] Jung. http://jung.sourceforge.net/.

[2] K. L. Chung and W. Zhuo. Graph-based visual analytic tools for parallel coordinates. In *ISVC (2)*, pages 990–999, 2008.

[3] J. Eisenstein, D. H. P. Chau, and A. Kittur. TopicViz: Semantic navigation of document collections. Technical Report 1110.6200, ArXiv, 2011.

[4] M. S. Floater. Mean value coordinates. *Computer Aided Geometric Design*, 20:2003, 2003.

[5] P. Howland, J. Wang, and H. Park. Solving the small sample size problem in face recognition using generalized discriminant analysis. *Journal of Pattern Recogntion*, 39(2):277–287, 2006.

[6] H. J. Kim, Y.He. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*.

[7] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164, 2012.

[8] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4:239–256, 2005.

[9] W. Zhuo and J. Rossignac. Curvature-based offset distance: Implementations and applications. *Computers & Graphics*, 36(5):445–454, 2012.