Computers & Graphics 62 (2017) 28–40

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on Graphics Interface 2016

Hierarchy-based projection of high-dimensional labeled data to reduce visual clutter

Dominik Herr [a,b,*], Qi Han [a], Steffen Lohmann [c], Thomas Ertl [a]

[a] *Institute for Visualization and Interactive Systems (VIS), University of Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany*
[b] *Graduate School of Excellence advanced Manufacturing Engineering (GSaME), University of Stuttgart, Nobelstraße 12 70569 Stuttgart, Germany*
[c] *Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Schloss Birlinghoven, 53757 Sankt Augustin, Germany*

A R T I C L E   I N F O

Article history:
Received 1 September 2016
Received in revised form
24 November 2016
Accepted 3 December 2016
Available online 13 December 2016

Keywords:
Projection visualization
Information search and retrieval
Clustering
Visual analytics

A B S T R A C T

Visualizing high-dimensional labeled data on a two-dimensional plane can quickly result in visual clutter and information overload. To address this problem, the data usually needs to be structured, so that only parts of it are displayed at a time. We present a hierarchy-based approach that projects labeled data on different levels of detail on a two-dimensional plane, whilst keeping the user's cognitive load between the level changes as low as possible. The approach consists of three steps: First, the data is hierarchically clustered; second, the user can determine levels of detail; third, the levels of detail are visualized one at a time on a two-dimensional plane. Animations make transitions between the levels of detail traceable, while the exploration on each level is supported by several interaction techniques, including halos, a darts view, and a magic lens. We demonstrate the applicability and usefulness of the approach with use cases from the patent domain and a question-and-answer website. In addition, we conducted a qualitative evaluation to assess the usefulness and comprehensibility of our approach.

1. Introduction

The abundance of data produced nowadays is often too much to show all at once and its structure regarding content and relation is often either unavailable or unknown. A common practice to label the data is using keywords or tags that originate from predefined metadata, a classification structure, or from extracted content. In addition, the overview of the relation between data is often simplified by clustering the data and then showing the cluster relations. The relations of the labels can span a high-dimensional space that is based on their co-occurrence. To visually analyze the labels' coherence, the labels may be shown on a two-dimensional plane, for example through projection. Yet, the two-dimensional visualization of many labels leads to the problem of visual clutter. This visual clutter combined with a possibly complex relationship measure can cause heavy cognitive load for an analyzing user and hinders the efficiency of the analysis. One common approach to ease this problem is to introduce a hierarchy to the labels, wherein a child node specifies the description of its parent node.

However, the visualization of such hierarchies usually focuses on the relationship between the parent and its children. Apart from the intuition that siblings with the same parent are closer related to each other than to other data objects on the same hierarchy level, the child–child relationship remains hidden in such visualization approaches.

We present an approach that aims to reduce the visual clutter of two-dimensional projections of a large number of labels through a smooth transition between overview and detail [1]. This is achieved by hierarchizing the labels based on a given similarity matrix. We show first highly aggregated clusters and then provide more details as the clusters are split apart. Also, our approach represents the child-child relationship of the generated hierarchy at a given level of detail through the projection of the relevant labels onto a two-dimensional plane. To do so, the approach is composed of three steps:

1. A hierarchy is generated from a given similarity matrix.
2. Based on the hierarchy, varying levels of detail are determined for the information shown to the user; the user can change those levels at will.
3. The information contained in the topmost level of detail is projected onto a two-dimensional plane, and the user can explore the shown data or increase the level of detail.

This article is an extension of a work we presented at the International Conference on Computer Graphics and Interactive Techniques

* Corresponding author at: Institute for Visualization and Interactive Systems (VIS), University of Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany.
E-mail addresses: dominik.herr@vis.uni-stuttgart.de (D. Herr), qi.han@vis.uni-stuttgart.de (Q. Han), steffen.lohmann@iais.fraunhofer.de (S. Lohmann), thomas.ertl@vis.uni-stuttgart.de (T. Ertl).

http://dx.doi.org/10.1016/j.cag.2016.12.004
0097-8493/

(GI'16) [2]. We extended the approach by a magic lens to alleviate a possible overlap of labels and an extension of the visual representation of the clusters by showing multiple contained elements to give a better overview of the clusters' contents. Furthermore, we conducted a user study to evaluate the usefulness and comprehensibility of introducing a hierarchy to a projection-based dimension reduction approach. We also updated and extended the discussion of related work and applied our approach to scholarly data to create example visualizations that were evaluated.

2. Related work

The following discussion of related work is divided into five parts. First, we examine works from the field of information visualization that use dimension reduction to visualize data spatially. Afterwards, we review visualization methods for hierarchical datasets and related interaction techniques. Then, we show visualization approaches for hierarchy-based data. Subsequently, we report on recent works in the area of word cloud visualization that are related to our approach. Finally, we summarize related work in the field of magic lenses.

2.1. Dimension reduction and data projection

One common approach for visualizing the content in datasets, for example document collections, is to take extracted features, such as keywords or concepts, and compare the data based on those features. The natural language processing community developed many ways to extract keywords from documents [3]. The relatedness of such features is usually pairwise computed and therefore presents a high-dimensional reduction problem. Methods to solve this problem can be projection-based, for instance, by using Principal Component Analysis (PCA) [4], Multidimensional Scaling (MDS) [5], Least Squares Projection [6], or t-Distributed Stochastic Neighbor Embedding (t-SNE) [7]. Van der Maaten et al. [8] survey a number of techniques to project high-dimensional data onto a low-dimensional space. In such approaches, data is usually visualized as data points that take up almost no space. This representation suffers from visual clutter when the data is not clearly separable, which is even more problematic when the data is represented with labels.

Recently Sacha et al. [9] reviewed how dimension reduction techniques are used and integrated into interactive visualization techniques.

Due to the high complexity of projection techniques, other potentially faster and more intuitive approaches, such as force-based layouts [10,11], are often used when interactive visualizations are needed. However, García-Fernández et al. [12] concluded in their study that projection-based approaches are superior to force-based layouts, when a complete and large dataset needs to be visualized.

Also, there are other approaches based on neural networks, such as hierarchical self-organizing maps [13]. When this approach is used for each created area, as proposed in [14] or [15], this approach creates a visualization seemingly based on a hierarchy. However, when generating a visual hierarchy this way, only elements contained in an area are mapped relative to each other. In our approach, we use a hierarchy not only to visualize the relation between parents and their children, but also between the siblings across the clusters.

2.2. Hierarchical aggregation and visualization

In case that the data is already hierarchically structured, there are many approaches to visualize such data. If the user needs to inspect the data's distribution across the hierarchy tree, tree visualization techniques, such as dendrograms or icicle plots, are commonly used. However, dendrograms do not provide information about the relation between the elements across multiple clusters (even when they are at the same hierarchy level). Another prominent method for showing a hierarchy's structure are treemaps [16]. Due to a treemap's layout, its ability to provide information about the clusters' relations with different parent clusters is limited. It is almost impossible to indicate possible shifts of the clusters' similarities on different levels of detail as tree maps assume a fixed hierarchy without considering a possible relation between the clusters, which we want to provide.

The goal of most hierarchical visualization techniques is to show the structure of the hierarchy. Elmqvist and Fekete [17] propose guidelines how hierarchical data can be used to limit the amount of shown information. We adapt some of the proposed guidelines in our work, such as taking the most important element of a cluster as cluster representative.

2.3. Visualization approaches for hierarchy-based data

There are various approaches to provide information about more details about hierarchy-based data. For example, Dou et al. [18] generate topic models where the users can interactively modify the created hierarchical structure. Afterwards, the user can inspect the development of individual or groups of topics over time. In contrast to our approach, this attempt uses the hierarchies to aggregate topics for the analysis of changes over time. Instead, we use hierarchies to filter shown data and show the data's relation spatially. Like most hierarchical visualization approaches, they assume the availability of a hierarchical structure and use predefined hierarchy levels to show information. Our approach goes beyond that by enabling the users to set the shown hierarchy levels by themselves.

Liu et al. [19] developed an approach to build hierarchies based on topic graphs. They visualize the relations of extracted topics by using stacked trees in combination with force-based graph layouts. In contrast to our approach, they use hierarchies to distinguish between topics and not to set their content in a relation to other topics. Our approach aims not only to provide relational information across clusters, but also of their content when more information is shown.

Fried and Kobourov [20] presented a system, wherein they map the titles of papers in the DBLP database onto a two-dimensional landscape based on a hierarchy. The users can create a search profile whose results are highlighted on the landscape using a heat map visualization. They focus on temporal aspects of the data.

Wise et al. [21] proposed to show large document collections through a galaxy metaphor, in which every document is represented by a star. By doing so, the user gets a more intuitive understanding of the relations between documents. Similarly, SPIRE [22] and INSPIRE [23] use the same metaphor, but they combine it with a visual analytics approach to enable users to further analyze the data. The STREAMIT system [24] uses force-based layouts, clustering discovery techniques and topic modeling to visualize document streams in real-time. The clustering is based on the graph layout and does not create a hierarchy to examine the document streams on a semantical level. An early version of the recently published *Overview* system [25] projects documents onto a two-dimensional scatterplot, whilst showing a hierarchical tree structure of the documents in another view. The system uses brushing and linking to connect those two visualizations. However, the selected elements of the hierarchy view are not represented in the amount of data shown in the scatterplot.

Stahnke et al. [26] propose an interaction technique to interpret arrangements and errors in dimensionality reduction. They do so by enabling the user to "probe" the projected data and show more detailed information about the data and the relevant projection information.

Only a few of these landscape-based approaches support hierarchical data, and most that do assume the hierarchy to be given. Thom et al. [27] use hierarchical topic clustering combined with a