EuroVis Workshop on Visual Analytics (2014)
M. Pohl and J. Roberts (Editors)

Guided Sketching for Visual Search and Exploration in Large Scatter Plot Spaces

Lin Shao[1], Michael Behrisch[1], Tobias Schreck[1], Tatiana v. Landesberger[2], Maximilian Scherer[2], Sebastian Bremm[2] and Daniel Keim[1]

[1]Universität Konstanz, Germany
[2]Technische Universität Darmstadt, Germany

Abstract

Recently, there has been an interest in methods for filtering large scatter plot spaces for interesting patterns. However, user interaction remains crucial in starting an explorative analysis in a large scatter plot space. We introduce an approach for explorative search and navigation in large sets of scatter plot diagrams. By means of a sketch-based query interface, users can start the exploration process by providing a visual example of the pattern they are interested in. A shadow-drawing approach provides suggestions for possibly relevant patterns while query drawing takes place, supporting the visual search process. We apply the approach on a large real-world data set, demonstrating the principal functionality and usefulness of our technique.

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Search process

1. Introduction

Data is created at incredible pace in many applications. For example, data is made increasingly available in so-called open data repositories, containing large amounts of scientific or government public data. While more data becomes available, it has become increasingly difficult for users to specify a starting point for data exploration. For example, one challenge is that the contents of open data repositories may be heterogeneous in nature, or not include consistent metadata annotations which would allow prioritized or catalog-based access. Example-based search can help as a starting point to explore the data space. But then, a challenge is that users are not aware of the patterns in the data such that an exact query formulation is not always obvious.

We introduce an approach for exploration in large scatter plots spaces. It is based on the idea of allowing the user to provide a visual example of data patterns to search for. This is useful for example, if an analyst wants to check whether a given target data distribution is present in the data set. A sketch for the target distribution can be matched against the data and the results can be visualized to reject or confirm hypotheses. In other cases, users may only have a vague idea about which data patters to expect, or they do not know what will be of interest among this data. To this end, we enhance the sketch interface by a so-called *shadow drawing* component. This component continuously executes a background search while the user is sketching, and overlays potentially matching or complementary data patterns in the editor. Thereby, already while searching, users may get an overview of data contents and can even adapt their query sketch in real time. Our search approach is based on an adapted image similarity search method which provides useful invariance properties for matching data patterns at different scale and location, making the approach applicable to find local patterns of interest. Our approach also supports navigation in scatter plot spaces by cluster analysis, allowing to compare sets of data. As for heterogeneous data, scatter plots may comprise different measurements, we also support exploring the relationships between scatter plot patterns and the distribution of measurement labels.

2. Related Work

Our approach relates to several areas, including analysis of scatter plot data and content-based image retrieval. It can

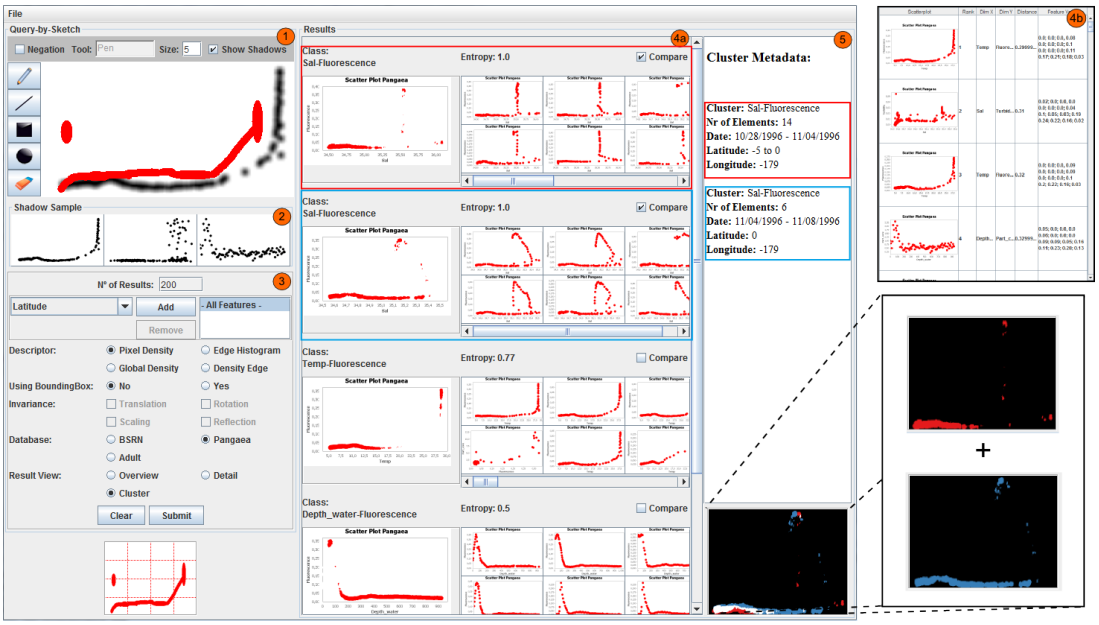


Figure 1: *Our approach for sketch-based scatter plot search and exploration. (1) user sketch interface; (2) shadow-draw templates for user guidance; (3) query settings; (4) result views ((4a) one-to-many comparison and (4b) one-to-one); (5) many-to-many comparison based on meta data and cluster representatives.*

be applied to applications for high-dimensional data via dimensionality reduction or the scatter plot matrix [WGK10] representation. The scatter plot diagram can be a common denominator for many different types of data. The Scagnostics approach [WAG06] provides a set of graph-based measures which characterize a given scatter plot, and can be used for ranking and clustering. Based on Scagnostics features, in [BCBL13] a metaheuristic optimization algorithm to predict interesting scatter plot patterns was proposed. In [LAE*12] a multi-step approach to analyze large scatter plot matrix spaces based on visual quality measures, matrix reordering, and visual abstraction was introduced. A recent study added an overview of perceptional separation factors useful for linking scatter plot properties with perceived interestingness or interpretability [STMT12]. In [EDF08], animated navigation through scatter plot matrix space by extrusion-based transitions between views was proposed. The detection of clusters and correlation in scatter plots highly depends on the appropriate aspect ratio of a scatter plot. In [FHSW13], the Delaunay triangulation is used to determine appropriate aspect ratios. In [TMF*12], we combined subspace search with projection-based scatter plot overviews to navigate high-dimensional data spaces.

In content-based image retrieval, the goal is to design functions to compare and rank images for similarity of content. Typically, various low-level image features including color histograms, edge histograms, or texture measures can be used [DKN08]. A problem in content-based search is often how to define a query if no example search object is available. Sketch-based approaches allow to match a user-provided sketch against image content [EHBA10]. Shadow-drawing was originally introduced to help untrained users execute appropriate sketches [LZC11]. While that work operated on real-word images, we apply the concept to the task of retrieval in scatter plot visualizations. Previous works have considered specific search methods for navigating in visualization spaces. To search for interesting local patterns in time series data, a sliding-window approach together with interactive query selection was introduced in [HS04]. In previous work, we considered search systems for time series data [BBF*11] and graph data [vLBBS10]. In [SvLS12], we compared several distance functions for scatter plot retrieval, based on an appropriately defined benchmark which is also used in this work. The contribution in this work is the shadow-draw enabled sketch interface for explorative search in scatter plot data, providing effective comparative views.

3. Sketch-Based Scatter Plot Retrieval and Exploration

We next describe our approach for shadow-draw and sketch-based search and exploration in scatter plot spaces.

3.1. Sketch-Based Scatter Plot Retrieval

The basis of our approach is a sketching interface (see Figure 1 (1)) to compare user sketches with the search scatter plot space. The interface allows to compose a query sketch by a

free-hand drawing tool with undo and erasing functionality. We compare a given sketch with a given scatter plot using a feature-based similarity function. Specifically, we subdivide the sketch and scatter plot diagram areas into a regular grids of size 8×8. For each grid cell, we compute the density of points contained, and a histogram of edge orientations of the cell points based on Laplace image filtering [PJW00]. Both density and edge features are concatenated to form a global feature vector which is used for search. We also employ a sliding-window based variant of this descriptor, to match a given sketch against a target pattern invariant with respect to different positions, rotations and scales.

3.2. Shadow Draw Support

We provide a shadow draw functionality to guide users in sketching a query of interest. Specifically, we overlay contours of candidate scatter plot results in real-time in the background so that the user can trace the potentially matching patterns. This approach is inspired by [LZC11], which supports users in freehand drawing of real-world object shapes. Our approach works by executing a similarity search each time the user ends a given stroke in constructing the query sketch. We apply a k-means clustering on the N best-matching results from the data base. These k clusters form the candidate shapes which are drawn in the background in a semi-transparent way. Thereby, we show possibly matching groups of scatter plot patterns which can guide and/or inspire the subsequent sketching process. Consequently, the user may save time during the sketching process. Furthermore, the user may develop an understanding of the target search space and how it may relate to the given information need, already during query specification.

As the shadow template we pick the cluster representative that is closest to the cluster's centroid. This ensures that we receive k rather different types of patterns, which are however still related to the current sketch, as they are computed from the current N best matches. Figure 2 (b) and (c) demonstrates the *sketch-guidance* with the user chosen parameter setting $k = 3$ and $N = 200$. Another functionality of our query interface is that each of the actually retrieved scatter plots can be used as a query template or form the basis for subsequent editing. Thus, the exploration process is turned into a feedback loop that advances in every step.

3.3. Result Comparison Views

Our approach includes three different views to compare the retrieved results. The first view is a one-to-one comparison, shown in Figure 2 (f). This view compares all retrieved scatter plots against the user sketch. Then, since the result set may contain many identical or only slightly differing pattern variations, we provide an optional clustering step which aggregates the result list in a number of clusters. This one-to-many view is illustrated in Figure 1 (4a). There, the whole result set is clustered by the DBSCAN algorithm [EKSX96], and each cluster is visualized by a weighted average of the scatter plots members of the respective cluster.

When the considered scatter plot data stems from high-dimensional or heterogeneous data sets, it is important to take the dimension labels (or units of measurements) into account. We compute an Entropy-based measure for the distribution of dimension labels within clusters to rank the obtained clusters. An entropy of 1.0 represents clusters with identical label combinations among its members. By means of the Entropy score, one may discover clusters which show similar data point distributions, but contain largely similar (or conversely, divergent) measurements. A drill-down functionality enables the user to explore the dimension labels in case of divergent label distributions. Different clusters can also be compared by cluster overlays in a many-to-many view, as Figure 1 (5) illustrates.

4. Application Examples

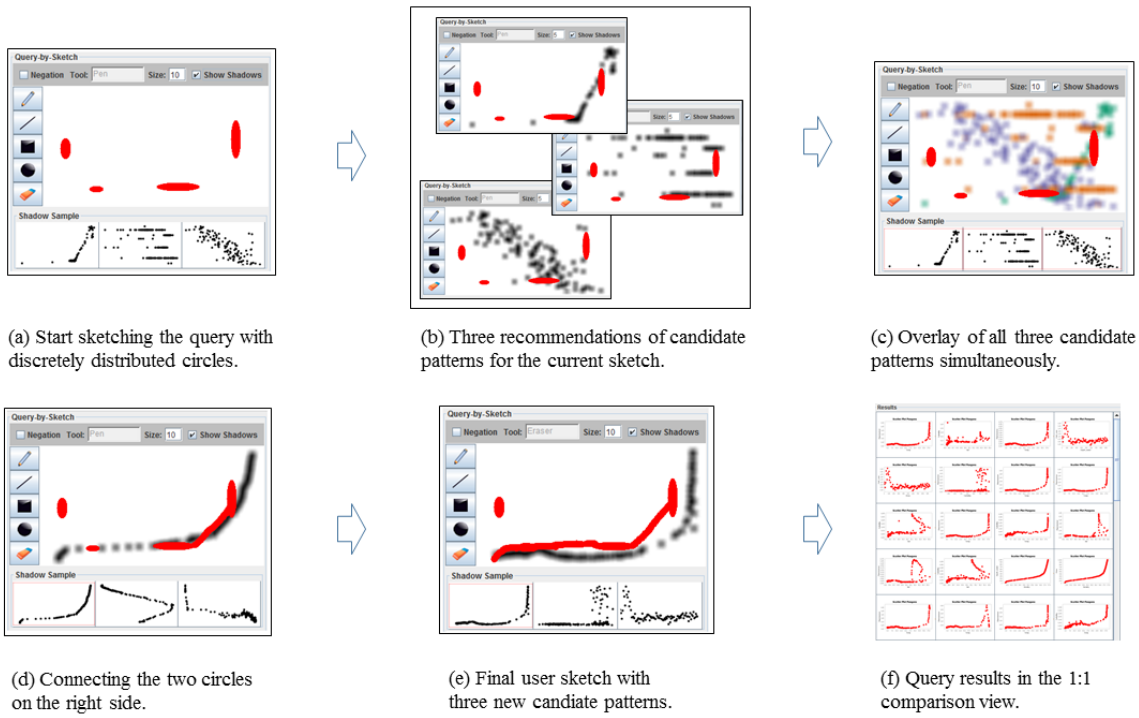
We apply our approach in an example use case to scientific data from earth observation research, provided by the PANGAEA repository [pan]. The repository hosts data on measurements of water, sediment, ice and atmosphere, among others. A collection of more than 24.000 scatter plots of many different earth-related measurements is considered (see [SvLS12] for details on the benchmark data).

4.1. Content-based Search Using Guided-Sketching

We show an exemplified exploratory search for interesting patterns in the PANGAEA data by means of Figure 2. We start by sketching discretely distributed dense areas, shown in Figure 2 (a). Figure 2 (b) illustrates three retrieved candidate patterns. So far, the individual candidates are dissimilar to the sketch, but if we overlay all candidate patterns at once we see a common dense area of point distributions in Figure 2 (c). Hence, we gain a first insight of given scatter plot content after the first stroke. If we cover the depicted dense shadow area the probability to receive more often occurring patterns increases. Figure 2 (d) illustrates this adapted sketch with new and more precise suggestions for candidate patterns. All of these three candidates represent distinct patterns and can be retraced or modified to sketch a more focused query. Finally, we include one of the suggested candidates and trace the remaining parts of the shadow template. The final sketch and retrieval results are shown in Figure 2 (e) and (f).

4.2. Scatter Plot Comparison and Analysis

In the subsequent confirmatory search we analyze the results of the user sketch depicted in Figure 2 (f). The view indicates that most of the retrieved patterns contain axis combinations in conjunction with measurement of "Fluorescence". Thus,



(a) Start sketching the query with discretely distributed circles.

(b) Three recommendations of candidate patterns for the current sketch.

(c) Overlay of all three candidate patterns simultaneously.

(d) Connecting the two circles on the right side.

(e) Final user sketch with three new candiate patterns.

(f) Query results in the 1:1 comparison view.

Figure 2: *Our proposed user sketch interface (a) including sketch-guidance functionality ((b) - (e)) with three candidate pattern suggestions (bottom row in (c) - (e)). A final ranking of retrieved results is shown in (f).*

we filter for scatter plots with fluorescence parameters to reveal which dimensions are related with fluorescence. The clustered patterns of this filtered result set are shown in Figure 1 (4). The comparison of two selected clusters (highlighted in red and blue) shows that both clusters consist of the axis combination "Sal - Fluorescence", but differ visually. Another interesting aspect is that they have a parameter Entropy score of 1.0, meaning that all scatter plots share the same measurements. Figure 1 (5) shows the cluster comparison. Both representatives show only small overlap (white region). The meta data indicates that the only differences are the time period and geographical location. The red cluster was measured between 10/28/1996–11/4/1996 and the blue from 11/4/1996–11/08/1996. Both clusters were measured in the Pacific, but the red data was taken more to the south. This difference seems to be the reason for the dissimilar patterns, since the ocean salinity depends on the geographical location [Gre]. While these are first results obtained by us as non-experts, we want to further evaluate our approach together with domain scientists.

5. Conclusion and Future Work

We presented an approach for search-oriented visual exploration in large scatter plot spaces. It is based on a suitable distance function defined between scatter plots. A shadow overlay helps in specifying queries, and at the same time, provides context on the target data base. Appropriate result views allow to compare scatter plot patterns and distribution of labels by cluster-based aggregation. We demonstrated the approach by application to a large scatter plot data set.

We want to extend in different directions. Mixture model analysis may capture the notion of local patterns within scatter plots, and be a basis for visual query composition. We also want to investigate scalable visual representations for comparing sets of scatter plots. Glyph-based approaches could be interesting to this end. Methodologically, we are interested to develop the search tool to help scientists working with large experimental data in confirmatory or hypothesis-generation tasks. How to design appropriate user interfaces which allow integration of background knowledge in such serach processes is seen as a substantial challenge.

Acknowledgements

This work was partially funded by the Juniorprofessor Program of the Landesstiftung Baden-Württemberg within the research project *Visual Search and Analysis Methods for Time-Oriented Annotated Data*. We thank the Alfred Wegener Institute, Bremerhaven and the PANGAEA portal for providing data which helped to develop this work.

References

[BBF*11] BERNARD J., BRASE J., FELLNER D., KOEPLER O., KOHLHAMMER J., RUPPERT T., SCHRECK T., SENS I.: A visual digital library approach for time-oriented scientific primary data. *Springer International Journal of Digital Libraries, ECDL 2010 Special Issue* (2011). 2

[BCBL13] BOUKHELIFA N., CANCINO W., BEZERIANOS A., LUTTON E.: Evolutionary Visual Exploration: Evaluation With Expert Users. *Computer Graphics Forum 32*, 3pt1 (2013), 31–40. 2

[DKN08] DESELAERS T., KEYSERS D., NEY H.: Features for image retrieval: an experimental comparison. *Inf. Retr. 11*, 2 (Apr. 2008), 77–107. 2

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis 2008) 14*, 6 (2008), 1141–1148. 2

[EHBA10] EITZ M., HILDEBRAND K., BOUBEKEUR T., ALEXA M.: An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics 34*, 5 (2010), 482–498. 2

[EKSX96] ESTER M., KRIEGEL H., SANDER J., XU X.: A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. of the Second International Conference on Knowledge Discovery and Data Mining* (1996). 3

[FHSW13] FINK M., HAUNERT J.-H., SPOERHASE J., WOLFF A.: Selecting the aspect ratio of a scatter plot based on its delaunay triangulation. *IEEE transactions on visualization and computer graphics 19*, 12 (Dec. 2013), 2326–35. 2

[Gre] GREICIUS T.: NASA's 'Salt of the Earth' Aquarius Reveals First Map. http://www.nasa.gov/mission_pages/aquarius/multimedia/gallery/pia14786.html, accessed 02/2014. 4

[HS04] HOCHHEISER H., SHNEIDERMAN B.: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization 3*, 1 (2004), 1–18. 2

[LAE*12] LEHMANN D. J., ALBUQUERQUE G., EISEMANN M., MAGNOR M., THEISEL H.: Selecting coherent and relevant plots in large scatterplot matrices. *Computer Graphics Forum 31*, 6 (Apr. 2012), 1895–1908. 2

[LZC11] LEE Y. J., ZITNICK C. L., COHEN M. F.: Shadowdraw: real-time user guidance for freehand drawing. *ACM Trans. Graph. 30*, 4 (2011), 27:1–27:10. 2, 3

[pan] PANGAEA Data Publisher for Earth & Environmental Science. http://www.pangaea.de/, accessed 02/2014. 3

[PJW00] PARK D. K., JEON Y. S., WON C. S.: Efficient use of local edge histogram descriptor. In *Proceedings of the 2000 ACM workshops on Multimedia* (New York, NY, USA, 2000), ACM, pp. 51–54. 3

[STMT12] SEDLMAIR M., TATU A., MUNZNER T., TORY M.: A taxonomy of visual cluster separation factors. *Computer Graphics Forum 31(3)* (2012), 1335–1344. 2

[SvLS12] SCHERER M., VON LANDESBERGER T., SCHRECK T.: A Benchmark for Content-Based Retrieval in Bivariate Data Collections. In *Proc. Int. Conference on Theory and Practice of Digital Libraries* (2012). 2, 3

[TMF*12] TATU A., MAASS F., FÄRBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D. A.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Procedings of IEEE Symposium on Visual Analytics Science and Technology* (2012), IEEE CS Press, pp. 63–72. 2

[vLBBS10] VON LANDESBERGER T., BREMM S., BERNARD J., SCHRECK T.: Smart query definition for content-based search in large sets of graphs. In *Proc. Int. Symposium on Visual Analytics Science and Technology* (2010), Eurographics Association, pp. 7–12. Peer-reviewed short paper. 2

[WAG06] WILKINSON L., ANAND A., GROSSMAN R. L.: High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Trans. Vis. Comput. Graph. 12*, 6 (2006), 1363–1372. 2

[WGK10] WARD M., GRINSTEIN G., KEIM D. A.: Interactive data visualization: Foundations, techniques, and application. A.K. Peters, Ltd, ISBN: 978-1-56881-473-5. 2