Eurographics Conference on Visualization (EuroVis), Posters Track (2016)
T. Isenberg and F. Sadlo (Guest Editors)

Query by Visual Words: Visual Search for Scatter Plot Visualizations

Lin Shao[1], Timo Schleicher[2] and Tobias Schreck[1]

[1]Graz University of Technology, Austria
[2]University of Konstanz, Germany

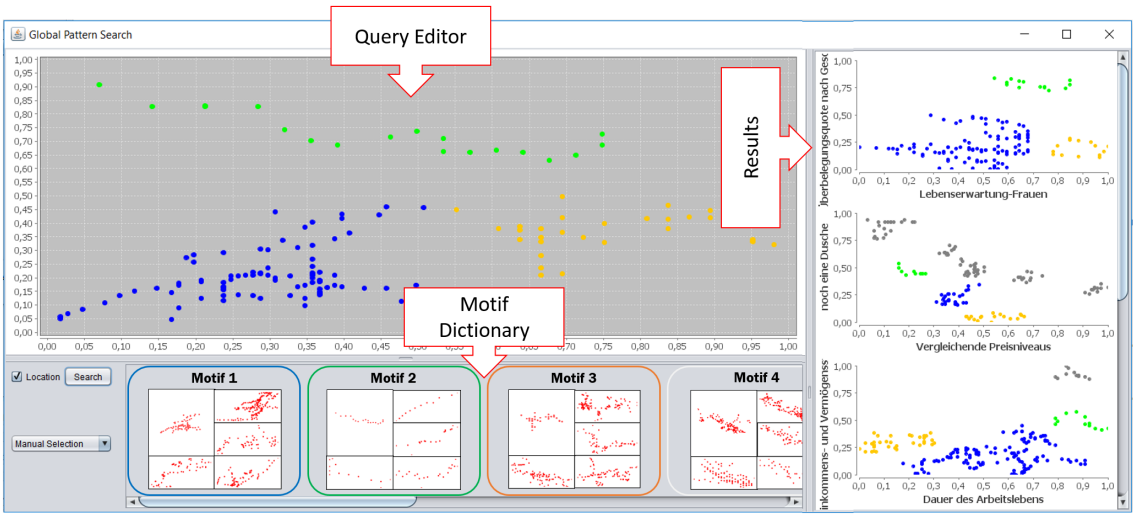

Figure 1: *Scatter plot retrieval by visual words. A motif-based dictionary is used to compose scatter plot queries and enables the search-based discovery of novel scatter plot compositions. To create a query, users can select an interesting motif and freely position the motif prototype (enlarged plot in the glyph representation) on the query editor.*

Abstract

Finding interesting views in large collections of data visualizations, e.g., scatter plots, is challenging. Recently, ranking views based on heuristic quality measures has been proposed. However, quality measures may fail to reflect given user interest, since interestingness is strongly dependent on the application domain and user context. As an alternative, interactive exploration in combination with example based user queries can be used to find patterns of interest. We introduce a novel approach for searching in large sets of scatter plot views based on a dictionary of frequent local scatter plot patterns. The dictionary is used for interactive construction of scatter plot queries, taking into account similarity of local scatter plot patterns as well as their approximate location in the plot. We introduce the overall approach, present a glyph design for visualization of dictionary entries, and illustrate the applicability of our implementation.

1. Introduction

Exploration of large data sets is a challenging task particularly if the dimensionality grows. A problem of high-dimensional visualization techniques, such as scatter plot matrices or parallel coordinates, is that the complexity of analyzing patterns and the amount of relevant projections will increase by the number of dimensions. Prior research has focused on quality metrics to filter and rank large projection views as a starting point for exploration. In [WAG05, TAE*11, SSB*15], global and local properties of scatter plot patterns were used to heuristically estimate the interestingness of a plot.

DOI: 10.2312/eurp.20161137

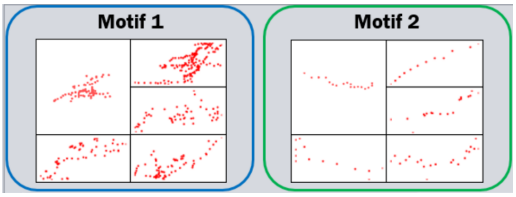

Figure 2: *Two entries of the dictionary of frequent scatter plot patterns. The enlarged plot depicts the motif cluster prototype and is used as motif template.*

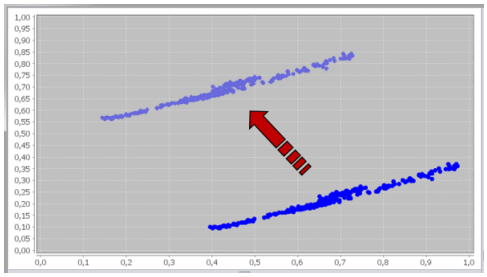

Figure 3: *Illustration of creating a query. Users can freely position selected motif prototypes on a template coordinate system.*

However, these quality measures may fail to cover the users' interest, since interestingness is strongly domain- and user-dependent. An alternative for exploring large visualization spaces would be to use sketch-based search techniques, which allow users to express their interest in view patterns (or motifs) [SBS*14]. Recently, several visual analytics systems based on interactive whiteboards have been presented for data exploration [LSR*15, BLC*11].

Inspired by aforementioned techniques, we propose a mixed approach for search in scatter plot data by combining interest measures based on local patterns [SSB*15] with a query-by-example search technique from image retrieval [ELK14]. The basic idea is to decompose a visualization into regions of interest. A dictionary of *visual words* represents frequently occurring regions of interest (or local scatter plot patterns). We use the dictionary to compute a similarity function between scatter plots, and allow the composition of queries by interactive selection from the dictionary.

2. Our Approach

Based on [SSB*15], we build a dictionary of scatter plot patterns by the following steps:

1. *Segmentation of Local Scatter Plot Patterns.* We perform a segmentation of each scatter plot into regions of interest. To this end, a Minimum Spanning Tree is constructed and the longest links are removed, hence segmenting the data into a number of dense clusters.
2. *Visual Feature Extraction.* For each scatter plot segment, a feature vector based on gradient orientation and density histogram is computed [PJW00].
3. *Dictionary Generation.* The dictionary is formed by clustering the set of scatter plot segments, using the visual feature vector as a basis. The dictionary consist of the clusters of scatter plot segments together with the size of the cluster members and possibly cluster quality statistics.

As shown in [SSB*15], we can describe each scatter plot by a $tf \times idf$-type (weighted) histogram of occurring patterns from the dictionary. In this work, we visualize the obtained dictionary entries by a overview glyph representation (see Figure 2) and use it for designing new queries. The glyph representation depicts the visual appearance of a motif by showing the motif cluster prototype (medoid segment of a dictionary entry) and a small multiple view of various associated motifs. Thus, users can quickly explore the local pattern space and choose interesting motifs for search. A query can be composed by selecting motif prototypes from the dictionary. Moreover, users can choose interesting motifs as visual words and relocate them on a template coordinate system to discover novel scatter plot compositions of motifs (see Figure 3). The best matching results are shown on the right hand side of Figure 1.

3. Queries Based on Motif Dictionary

A user query Q consists of a number of selected patterns from the dictionary, $Q = \{q_1, q_2, ..., q_n\}$, as well as approximate spatial positions of each of the local patterns. We want to find those scatter plots $S = \{s_1, s_2, ..., s_n\}$ which show the highest similarity to Q. We take into account the similarity of scatter plot patterns as well as their spatial positions.

We use a 1:1 alignment to match a pattern of S to a motif of Q. Since the search includes the desired location of scatter plot patterns, we use the spatial position of patters with respect to their affiliation (i.e., dictionary entry) as a constraint. To determine the spatial distance between a pattern of S to a motif of Q, we compute the center of mass on the normalized scatter plot axes of the patterns and motifs respectively. A 1:N alignment may result in scatter plots that contain one query motif several times whereas other query motifs may not be contained at all, although they have a potential matching partner. In this case, we compute the nearest neighbor pattern on the 2D plane that can be matched in a 1:1 fashion. This is especially important if the set of scatter plot patterns and the set of query motifs have a varying size. If for instance a scatter plot consists of two patterns but the user specified more query motifs, then only the two best matching patterns are taken into account.

A second important search constraint is the number of motifs. Since the user also defines the number of desired motifs, this needs to be considered in addition. For this reason, we use an optional weighting factor depending on the number of patterns that have not been matched from the query due to a too small number of motifs in a respective scatter plot. Another matching case that needs a specific treatment is the N:M alignment for identical motifs. For this scenario, we minimize the overall matching distance of scatter plot patterns S and query motifs Q.

4. Conclusion

The basic idea of this work is to transfer the well-known Bag-of-Words concept from image retrieval to data visualization techniques. In the scope of this poster, we introduce our approach and pipeline for scatter plot retrieval using a motif-based dictionary as described in [SSB*15]. By using this approach, we support the exploration process for locally interesting areas in scatter plots and bridge the gap between motif exploration and global scatter plot search.

References

[BLC*11] BROWNE J., LEE B., CARPENDALE S., RICHE N., SHERWOOD T.: Data analysis on interactive whiteboards through sketch-based interaction. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (New York, NY, USA, 2011), ITS '11, ACM, pp. 154–157. 2

[ELK14] ELLIOTT D., LAVRENKO V., KELLER F.: Query-by-example image retrieval using visual dependency representations. In *Proc. Int. Conf. on Computational Linguistics* (Aug 2014), pp. 109–120. 2

[LSR*15] LEE B., SMITH G., RICHE N. H., KARLSON A., CARPENDALE S.: Sketchinsight: Natural data exploration on interactive whiteboards leveraging pen and touch interaction. In *2015 IEEE Pacific Visualization Symposium (PacificVis)* (April 2015), pp. 199–206. 2

[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* (2000), ACM, pp. 51–54. 2

[SBS*14] SHAO L., BEHRISCH M., SCHRECK T., VON LANDESBERGER T., SCHERER M., BREMM S., KEIM D. A.: Guided Sketching for Visual Search and Exploration in Large Scatter Plot Spaces. In *Proc. Int. Workshop on Visual Analytics* (2014), EG. 2

[SSB*15] SHAO L., SCHLEICHER T., BEHRISCH M., SCHRECK T., SIPIRAN I., KEIM D. A.: Guiding the exploration of scatter plot data using motif-based interest measures. In *Proc. Int. Symp. on Big Data Visual Analytics* (Sept 2015), pp. 1–8. 1, 2

[TAE*11] TATU A., ALBUQUERQUE G., EISEMANN M., BAK P., THEISEL H., MAGNOR M., KEIM D.: Automated analytical methods to support visual exploration of high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 17*, 5 (2011), 584–597. 1

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. In *In Proceedings of the IEEE Symposium on Information Visualization* (2005), pp. 157–164. 1