Enhancing the Visual Clustering of Query-dependent Database Visualization Techniques using Screen-Filling Curves

Extended Abstract

Daniel A. Keim

Institute for Computer Science, University of Munich
Leopoldstr. 11 B, D-80802 Munich, Germany, Phone (+49) 89 2180-6267
keim@informatik.uni-muenchen.de

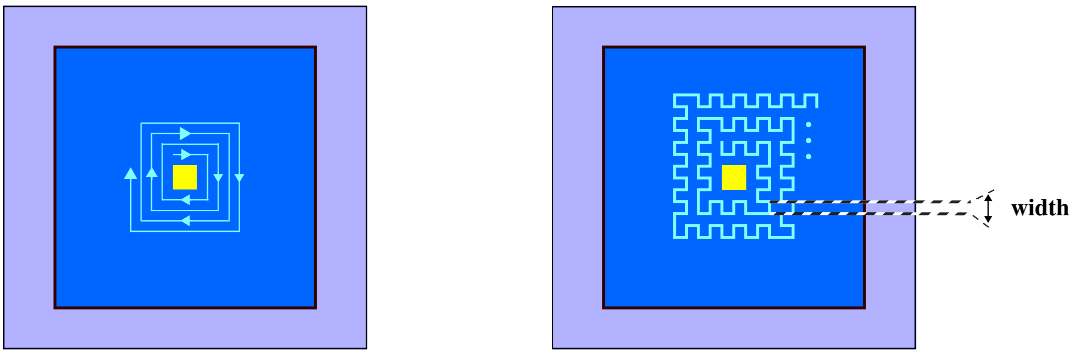
1. Introduction

An important goal of visualization technology is to support the exploration and analysis of very large amounts of data which are usually stored in databases. Since number and size of the databases is growing rapidly, there is a need for novel visualization techniques which allow a visualization of larger amounts of data. Most of today's databases store typical transaction-generated multi-attribute data which does not have any inherent two- or three-dimensional semantics and therefore does not lend itself to some two- or three-dimensional visualization.

In general, databases can be seen as multidimensional data sets with the attributes of the database corresponding to the dimensions of the multidimensional data set. There are a variety of well known techniques for visualizing arbitrary multidimensional data sets: scatterplot matrices and coplots [And 72, Cle 93], geometric projection techniques (e.g., [Hub 85, ID 90]), iconic display techniques (e.g., [Che 73, SGB 91]), pixel-oriented techniques (e.g., [KKS 93, KKA 95]), hierarchical techniques (e.g., [BF 90, Shn 92]), dynamic techniques (e.g., [BMMS 91, AWS 92, Eic 94, ADLP 95]), and combinations hereof (e.g. [Asi 85, AS 94]). Most techniques, however, are only suitable for rather small data sets consisting of at most a few thousand data items. Only dynamic and pixel-oriented techniques are able to handle larger data sets with hundreds of thousands of data items. Dynamic techniques reveal interesting properties of the data by generating series of visualizations. In general, however, only a limited portion of the data can be visualized at one point of time. In contrast, pixel-oriented techniques are able to visualize the largest amounts of data possible at one point of time. On current displays, they allow a visualization of up-to 1,000,000 data values.

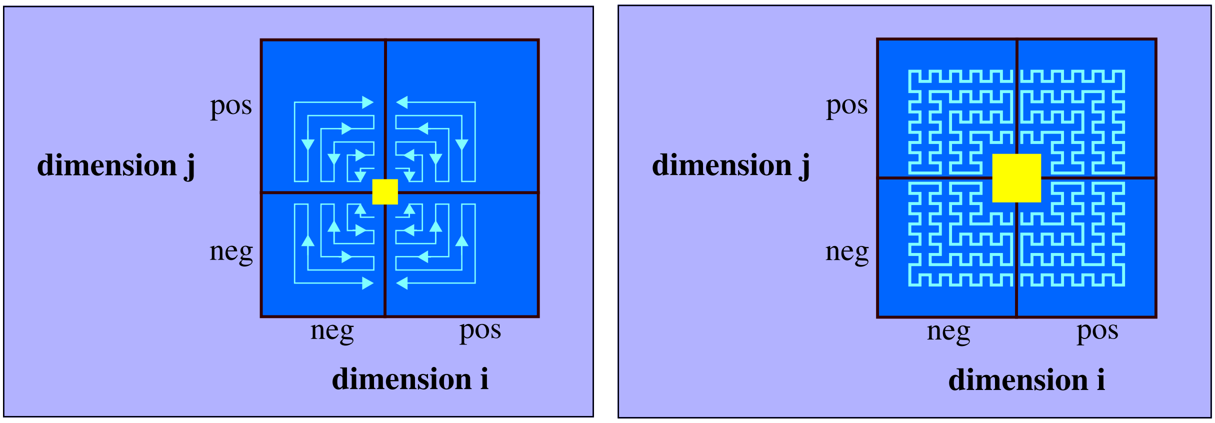
2. Pixel-oriented Visualization Techniques

The basic idea of pixel-oriented techniques is to map each data value to a colored pixel and present the data values belonging to different dimensions (attributes) in separate subwindows. Pixel-oriented visualization techniques can be divided into query-independent techniques which directly visualize the data (or a certain portion of it) and query-dependent techniques which visualize the data in the context of a specific query. Examples for the class of query-independent techniques are the screen-filling curve and recursive pattern techniques. The screen-filling curve techniques are based on the well-known Morton and Peano-Hilbert curve algorithms [Pea 90, Hil 91, Mor 66], and the recursive pattern technique is based on a generic recursive scheme which generalizes a wide range of pixel-oriented arrangements for displaying large data sets [KKA 95]. Examples for the class of query-dependent techniques are the Spiral- and 2D-techniques [KKS 93], which visualize the distances with respect to a database query. The arrangement of Spiral- and 2D-techniques centers the most relevant data items (data items fulfilling the query) in the middle of the window, and less relevant data items (data items approximately fulfilling the query) are arranged in a rectangular spiral-shape to the outside of the window (c.f. Figure 1a). In case of the 2D-technique, the data set is partitioned into four subsets according to the direction of the distance for two dimensions: for one dimension negative distances are arranged to the left, positive ones to the right and for the other dimension negative distances are arranged to the bottom, positive ones to the top (c.f. Figure 2a).



a. Spiral Technique **b. Snake Variant**

Figure 1: Spiral Technique and Snake Variant



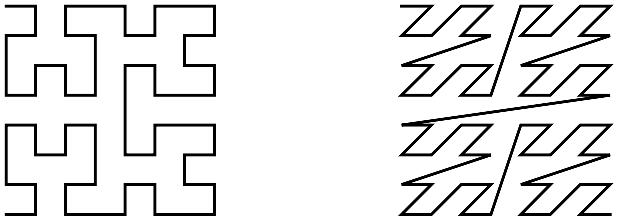
a. 2D-Technique **b. Snake Variant**

Figure 2: 2D-Technique and Snake Variant

3. Generalization of Query-dependent Visualization Techniques

A problem of the Spiral- and 2D-visualization techniques is that the local clustering properties of the Spiral arrangement are rather weak. Since the spiral is only one pixel wide, it is perceptually impossible to detect small clusters. The reason for this problem is that the mapping from the ordered sequence of data items to the position of the pixels on the two-dimensional display does not preserve locality. More specific, the probability that two pixels which are close together on the screen are also close together in the one-dimensional ordered sequence of data items is rather low. Arrangements which do provide a maximum of locality preservation are wave- or snake-like arrangements such as those used in the recursive pattern technique and screen-filling curves such as the Peano-Hilbert and the Morton curve. A problem of those techniques, however, is that the most relevant data items are placed in one corner of the visualization and that the ordering of data items according to their relevance does not become clear in the visualization.

Our new techniques aim at combining the advantages of both, the query-dependent techniques and the screen-filling curves, while at the same time avoiding their disadvantages. Our new variants of the Spiral- and 2D-technique retain the overall arrangement of the Spiral- and 2D-technique, centering the most relevant data items in the middle of the screen, but enhance the clustering properties of the arrangement by using screen-filling curves locally. This means in case of the generalized spiral technique that the overall form of the arrangement remains a rectangular spiral shape. In contrast to the original technique, however, the spiral is composed of small Peano-Hilbert-, Morton-, or Snake-like curves. The width of the spiral is determined by the size of the ‘small’ curves. In case of the Snake variants, the width can be varied arbitrarily. The structure of



a. Peano-Hilbert **b. Morton**

Figure 3: Peano-Hilbert and Morton Arrangement of Order Three

the Peano-Hilbert and Morton curve does not allow arbitrary widths since the width is determined by the order of the Peano-Hilbert and Morton curves. For curves with order two, the resulting width is $2^2 = 4$, for curves with order three the resulting width is $2^3 = 8$, and for curves with order four the resulting width is $2^4 = 16$ (c.f. Figure 3). The Snake, Peano-Hilbert, and Morton variants of the Spiral- and 2D-techniques can be seen as generalization of the previously defined query-dependent pixel-oriented visualization techniques. The original Spiral- and 2D-techniques now become the special case of a Snake, Peano-Hilbert, and Morton curve of order one.

In the following, we show the advantage of improving the local clustering of query-dependent visualization techniques by screen-filling curves. In Figure 1, we provide an example visualization showing the effect of using the Snake variant of the Spiral-technique. The data set used consists of about 24,000 test data items with eight dimensions. Most of the data set (20,000 data items) is randomly generated in the range [-100, 100]. The remaining 4000 data items split up into two clusters which are only defined on the first five dimensions and are inserted at specific locations of the eight-dimensional space. The query used is [-20, 20] for each of the dimensions. Figure 1a shows the visualization generated by the Spiral technique and Figure 1b shows the visualization generated by the Snake-Spiral technique with a width of six pixels. In Figure 1a, almost no clustering is visible, while in Figure 1b, the clustering becomes quite obvious by the similar structure in the first five dimensions. The example clearly shows the advantage of the Snake-Spiral over the original Spiral technique. Figure 5 shows the visualizations resulting from different snake widths (snake width = 2, 4, 16).

Experiments with the Peano-Hilbert and the Morton variants of the Spiral- and the 2D-techniques show a similar performance compared to the Snake variant (cf. Figure 6). For larger widths, however, we the Peano-Hilbert and Morton variants to perform better since their local clustering is better. A disadvantage of the Peano-Hilbert and Morton variants is that the possible spiral widths are limited to 2^i depending on the order i of the local Peano-Hilbert and Morton arrangement.

4. Conclusions

The described techniques are all integrated into the *VisDB* database visualization and exploration environment. The *VisDB* system is implemented in C++/MOTIF and runs under X-Windows on HP 7xx machines. The system consists of an interactive interface which is divided into the visualization portion and the query specification portion. The query specification portion provides a slider-based direct-interaction interface which allows an intuitive specification of queries [Kei 94, KK 94]. The *VisDB* system has been specially designed to support the analysis and comparison of different visualization techniques. The goal of future work is to perform series of experiments which show the advantage of our new techniques on a wide range on data sets. The experiments are also important to determine the spiral widths which provide the best local clustering without destroying the global arrangement. Our experience in using the snake variants of the Spiral technique shows that snake widths between 6 and 10 provide good results. The optimal width, however, will probably depend on the properties of the data (e.g., size and number of clusters), and therefore, in our current implementation the user may switch arbitrarily between different widths.

References

[ADLP 95] Anupam V., Dar S., Leibfried T., Petajan E.: *'DataSpace: 3-D Visualization of Large Databases',* Proc. Int. Symposium on Information Visualization, Atlanta, GA, 1995.

[And 72] Andrews D. F.: *'Plots of High-Dimensional Data',* Biometrics, Vol. 29, 1972, pp. 125-136.

[AS 94] Ahlberg C., Shneiderman B.: *'Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays'*, Proc. ACM CHI Int. Conf. on Human Factors in Computing (CHI94), Boston, MA, 1994, pp. 313-317.

[Asi 85] Asimov D.: *'The Grand Tour: A Tool For Viewing Multidimensional Data',* SIAM Journal of Science & Stat. Comp., Vol. 6, 1985, pp. 128-143.

[AWS 92] Ahlberg C., Williamson C., Shneiderman B.: *'Dynamic Queries for Information Exploration: An Implementation and Evaluation',* Proc. ACM CHI Int. Conf. on Human Factors in Computing (CHI92), Monterey, CA, 1992, pp. 619-626.

[BF 90] Beshers C., Feiner S.: *'Visualizing n-Dimensional Virtual Worlds with n-Vision'*, Computer Graphics, Vol. 24, No. 2, 1990, pp. 37-38.

[BMMS 91] Buja A., McDonald J. A., Michalak J., Stuetzle W.: *'Interactive Data Visualization Using Focusing and Linking'*, Visualization '91, San Diego, CA, 1991, pp. 156-163.

[Che 73] Chernoff H.: *'The Use of Faces to Represent Points in k-Dimensional Space Graphically',* Journal Amer. Statistical Association, Vol. 68, pp 361-368.

[Cle 93] Cleveland W. S.: *'Visualizing Data',* AT&T Bell Laboratories, Murray Hill, NJ, Hobart Press, Summit NJ, 1993.

[DE 82] Dunn G., Everitt B.: *'An Introduction to Mathematical Taxonomy',* Cambridge University Press, Cambridge, MA, 1982.

[Eic 94] Eick S.: *'Data Visualization Sliders'*, Proc. ACM UIST'94, 1994.

[Hil 91] Hilbert D.: *'Über stetige Abbildung einer Line auf ein Flächenstück'*, Math. Annalen, Vol. 38, 1891, pp. 459-460.

[Hub 85] Huber P. J.: *'Projection Pursuit',* The Annals of Statistics, Vol. 13, No. 2, 1985, pp. 435-474.

[ID 90] Inselberg A., Dimsdale B.: *'Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry'*, Visualization '90, San Francisco, CA., 1990, pp. 361-370.

[Kei 94] Keim D. A.: *'Visual Support for Query Specification and Data Mining',* Ph.D. Dissertation, University of Munich, July 1994, Shaker-Publishing Company, Aachen, Germany, 1995, ISBN 3-8265-0594-8.

[KK 94] Keim D. A., Kriegel H.-P.: *'VisDB: Database Exploration using Multidimensional Visualization',* Computer Graphics & Applications, Sept. 1994, pp. 40-49.

[KKA 95] Keim D. A., Kriegel H.-P., Ankerst M.: *'Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data',* Visualization '95, Atlanta, GA, 1995.

[KKS 93] Keim D. A., Kriegel H.-P., Seidl T.: *'Visual Feedback in Querying Large Databases',* Visualization '93, San Jose, CA,1993.

[Mor 66] Morton G. M.: *'A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing'*, IBM Ltd. Ottawa, Canada, 1966.

[Pea 90] Peano G.: *'Sur une courbe qui remplit toute une aire plaine'*, Math. Annalen, Vol. 36, 1890, pp. 157-160.

[SGB 91] Smith S., Grinstein G., Bergeron R. D.: *'Interactive Data Exploration with a Supercomputer',* Visualization '91, San Diego, CA, 1991, pp. 248-254.

[Shn 92] Shneiderman B.: *'Tree Visualization with Treemaps: A 2-D Space-filling Approach'*, ACM Trans. on Graphics, Vol. 11, No. 1, 1992, pp. 92-99.