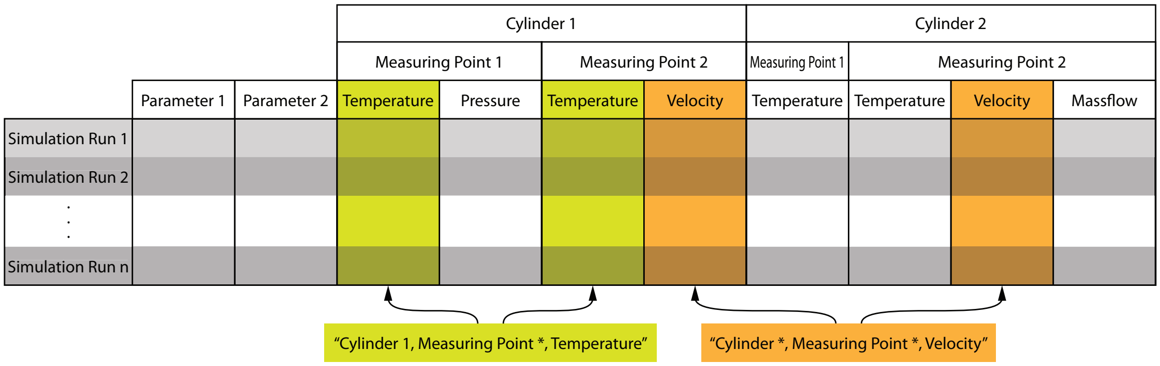
Dimension Sets: Visual Analysis of Structured High-Dimensional Data

Harald Piringer*
VRVis Research Center, Vienna, Austria

Wolfgang Berger†
VRVis Research Center, Vienna, Austria



			Cylinder 1				Cylinder 2			
			Measuring Point 1		Measuring Point 2		Measuring Point 1	Measuring Point 2		
	Parameter 1	Parameter 2	Temperature	Pressure	Temperature	Velocity	Temperature	Temperature	Velocity	Massflow
Simulation Run 1										
Simulation Run 2										
⋮										
Simulation Run n										

"Cylinder 1, Measuring Point *, Temperature"
"Cylinder *, Measuring Point *, Velocity"

Figure 1: A simplified example of a data model from multi-run simulations in combustion-engine design. Dimensions are structured hierarchically and dimension sets serve as a mechanism to define groups of semantically related subsets.

Abstract

The high dimensionality of a dataset can be a result of the chosen data model. In such cases, the challenge of an analysis is to consider the structure and the semantics of the dimensions. This poster proposes the concept of dimension sets, which are user-defined sets of related dimensions for parameterizing views and interactive selections. We describe options for visual layouts and implications when relating multiple dimension sets, and we outline extensions to interactive queries. Examples from the analysis of multi-run simulations in combustion-engine design illustrate the work. This domain also provides an application area where first evidence for the usefulness has been collected.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction Styles

1 Introduction

Datasets with hundreds of attributes are becoming increasingly common. When analyzing such high-dimensional data, it is important to consider the source of the high dimensionality. In many cases, all dimensions are considered unrelated a priori and it is a goal of the analysis to reduce the dimensionality by determining relations. Two fundamental strategies for dimension-reduction are to project data to low-dimensional space and to structure the dimensionality for selecting a subset thereof. Examples of projections include linear methods like Principle Component Analysis [3] and non-linear methods like Multi-Dimensional Scaling [5]. Examples of dimension-based structuring include semi-automatic approaches for grouping similar dimensions [7] or ranking dimensions with respect to different notions of relevance [6]. Recently, interactive workflows have been proposed [2, 1] where more comprehensive overviews of related work can be found.

In some cases, however, the high dimensionality is a result of the data model and the structure of the dimensions is known in advance. Our application background, for example, is the analysis of multi-run simulations in combustion-engine design [4]. Such data involves different simulation results (e.g., temperature, pressure, etc.) at a – typically small – number of measuring points within a conceptualized block model of an engine. As illustrated in Fig. 1, the high dimensionality results from discriminating different simulation results per measuring point as distinct dimensions. All dimensions are structured hierarchically to reflect different levels of the block model (e.g., measuring points, cylinders, etc.). While the data could in principle be laid out differently, the domain experts decided for this data model for two reasons: First, each row represents an entire simulation run. This matches their mental model when analyzing the data. Second, the number and the type of simulation results varies across different measuring points (see Fig. 1). Alternative data models (e.g., encoding the measuring point as categorical dimension) would thus be inefficient due to many missing values, or make queries complex due to splitting the data across multiple relational tables.

The key challenge is to enable an analysis of different levels of the dimension hierarchy without reflecting the potentially large number of involved dimensions by the complexity of visualization or interaction methods. For example, a scatterplot may visualize "temperature versus velocity" for a single measuring point or for multiple measuring points.

2 Visual Analysis Using Dimension Sets

The basis of the proposed approach is the concept of dimension sets. A dimension set is a group of semantically related dimensions with a comparable scaling, identical units, and a certain structural position within a given dimension hierarchy. For example, the dimensions representing "temperature" at all measuring points of cylinder 1 could be a meaningful dimension set in our application scenario. Conceptually, the structural position can be described as a path within the dimension hierarchy that may contain wildcards, e.g., "engine.cylinder1.measuringpoint*.temperature".

The key idea with respect to the visual analysis is to support a parameterization of views and interaction techniques like brushes in terms of dimension sets. For visualizations, the requirement of having a comparable scaling ensures that all dimensions of one set can reasonable be mapped to the same visual reference. Basic options for visual layouts include overlay and spatial separation.

Overlay draws the visual representation for all dimensions within the same visual space. As a consequence, a single data record typically has multiple visual representations - one for each dimension.

*e-mail: hp@vrvis.at
†e-mail:wberger@vrvis.at

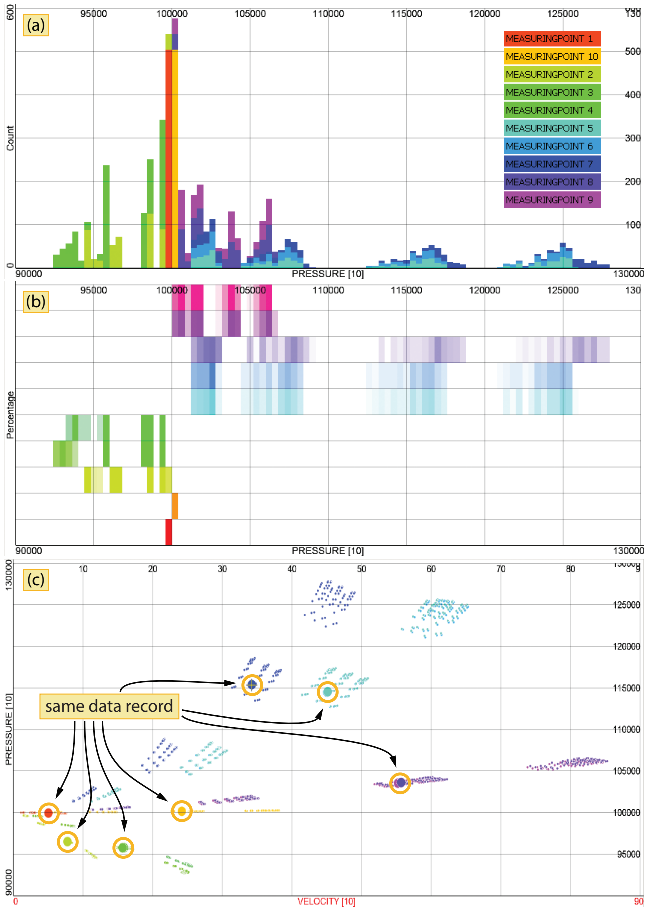


Figure 2: Visualization examples of dimension sets. (a) A histogram aggregating ten data dimensions; (b) a heatmap showing the same ten dimensions as adjacent rows; (c) a scatterplot relating two dimension sets when highlighting the visual items of one data record.

Fig. 2a shows a histogram that accumulates the frequencies for all dimensions of one set as an example of a frequency-based visualization technique. Item-based visualization techniques draw the visual elements for all dimensions (see Fig. 2c). Overlay is most useful to convey an overview of dimension sets in their entirety. For example, it indicates whether the values of any dimension have certain characteristics (e.g., exceed a certain threshold) without consuming much space. Sometimes, a distinction between individual dimensions is not even necessary for this task. Otherwise, color may indicate the contribution of each dimension for a small number of dimensions per set.

Spatial separation layouts multiple visualizations of the dimensions of one set, while a common scaling ensures the comparability within one visual context. Fig. 2b shows an example where the distribution of each dimension is shown as rows of a heatmap.

For visualization techniques that relate multiple variables (e.g., scatterplots or parallel coordinates), an important aspect is to combine dimensions from multiple sets. Meaningful combinations may be $1:n$ or $n:n$. $1:n$ refers to the combination of a set of n dimensions with the values of a single dimension. In this case, the single dimension is visually related to each dimension of the set. In our application, for example, the single dimension could be a parameter of the simulation. A respective scatterplot shows the distribution of a dimension set for all values of the parameter. In case of n:n combinations, the involved dimension sets are not only required to have an equal cardinality, but must also refer to common parents within the dimension hierarchy. For example, plotting temperature against pressure is typically only reasonable if pairs of dimensions refer to the same measuring point (as in Fig. 2c). The extension of this concept to combinations of more than two sets is straightforward, e.g., for 3D scatterplots.

Interactive selections also need to consider the concept of dimension sets. Representing single data records by multiple visual items introduces ambiguities that can be resolved in different ways. Typically, selections operate on entire data records, i.e., simulation runs in our application. In this case, two possible strategies may be *All* or *Any*, referring to logical AND or OR combinations, respectively, when applying a certain selection criterion to all dimensions of one or more sets. In our experience, a particularly useful application of the strategy *Any* is to highlight all visual representations that refer to a certain data record (see Fig. 2c). In combination with appropriate focus+context techniques, such a selection helps to identify structures across multiple dimensions for single data records.

Alternatively, selections may also operate on parts of data records, taking the structure of the dimension hierarchy into account. For example, selecting an item in a scatterplot could affect only those dimensions of the respective measuring point. While this approach is suitable to display details for specific visual representations, it is usually less suitable for linking visualizations of different parts of the dimension hierarchy.

3 Feedback and Limitations

First feedback from domain experts in combustion engine design confirmed several advantages of dimension sets. According to them, the most important advantage of dimension sets is to enable an analysis in terms of semantically meaningful and coherent groups despite potentially thousands of different simulation results. To them, the visualization of dimension sets provides useful overviews and enhances the comparability of multiple dimensions.

In some sense, dimension sets are an approach to transform a high-dimensional analysis problem into a visualization problem of many data items. However, as a key limitation, dimension sets are only applicable to certain types of high-dimensional data, i.e., data consisting of dimensions with corresponding semantics and a comparable scaling. Especially item-based visualizations may suffer from cluttering if the product of the number of dimensions times the number of records is large. Moreover, dimension sets arouse new challenging issues themselves, e.g., to automatically identify a concise name.

References

[1] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for Dimensional Analysis and Reduction. In *Proc. of the IEEE Conference on Visual Analytics Science and Technology (VAST 2010)*, pages 3–10. IEEE Computer Society, 2010.

[2] S. Johansson and J. Johansson. Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics. *IEEE Trans. on Visualization and Computer Graphics*, 15(6):993–1000, 2009.

[3] I. Jolliffe. *Principle Component Analysis*. Springer-Verlag, New York, 1986.

[4] K. Matkovic, M. Jelovic, J. Juric, and Z. Konyha. Interactive Visual Analysis and Exploration of Injection Systems Simulations. In *Proc. of the IEEE Conf. on Visualization 2005*, pages 391–398. IEEE Computer Society, 2005.

[5] A. Mead. Review of the Development of Multidimensional Scaling Methods. *The Statistician*, 33:27–35, 1992.

[6] J. Seo and B. Shneiderman. A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections. In *Proc. IEEE Symposium on Information Visualization 2004 (InfoVis 2004)*, pages 65–72, 2004.

[7] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets. In *Proc. IEEE Symposium on Information Visualization 2003 (InfoVis 2003)*, pages 105–112, 2003.