A Flexible Approach for Visual Data Mining

Matthias Kreuseler and Heidrun Schumann

Abstract—The exploration of heterogenous information spaces requires suitable mining methods as well as effective visual interfaces. Most of the existing systems concentrate either on mining algorithms or on visualization techniques. This paper describes a flexible framework for Visual Data Mining which combines analytical and visual methods to achieve a better understanding of the information space. We provide several preprocessing methods for unstructured information spaces such as a flexible hierarchy generation with user controlled refinement. Moreover, we develop new visualization techniques including an intuitive Focus+Context technique to visualize complex hierarchical graphs. A special feature of our system is a new paradigm for visualizing information structures within their frame of reference.

Index Terms—Information visualization, multidimenisional information modeling, hierarchies, focus+context techniques, clustering, maps, information analysis.

1 Introduction

Exploration of complex information spaces has become one of the "hot topics" in many research fields, including computer graphics, data mining, pattern recognition, and learning, and other areas of statistics, as well as data bases and data warehousing. A variety of novel mining techniques, visualization paradigms, and frameworks have been developed in recent years. Nevertheless, extracting useful knowledge or models from observed data is still a complicated nontrivial process.

In this context, visualization offers a powerful means of analysis that can help to uncover patterns and trends hidden in unknown data. Additionally, visualization provides a natural method of integrating multiple data sets and has been proven to be reliable and effective across a number of application domains. Still, visual methods cannot entirely replace analytic nonvisual mining algorithms. Rather, it is useful to combine multiple methods during data exploration processes [31].

The new area of visual data mining focuses on this combination of visual and nonvisual techniques as well as on integrating the user in the exploration process. Integrating visual and nonvisual methods in order to support a variety of exploration tasks, such as identifying patterns in large unstructured heterogeneous information or displaying information context (e.g., frame of spatial or domain references), requires sophisticated mining, visualization and interaction techniques. This carries over entirely new qualities of problems. Some of the most important ones can be summarized as follows:

- *Extracting patterns and controlling the mining*: The exploration of large unstructured information spaces requires information preprocessing. In this regard "filtering out uninteresting items" and merging similar objects into groups are necessary in order to reveal hidden patterns. Suitable metrics have to be applied for obtaining similarities and structures in high-dimensional feature space. Furthermore, the degree of abstraction has to be controlled interactively in order to supervise and steer the search for patterns during the mining process. This interaction is of outstanding importance to support explorations at arbitrary levels of detail.
- *Visualizing information sets*: The success of visual data analysis depends very much on its ability to support a variety of exploration tasks such as overview, zoom in on items of interest or details on demand. Different visualization methods are required for revealing information structure and information contents such as attribute values. Furthermore, novel interaction techniques are needed for controlling the degree of abstraction within visual representations and for providing navigational aids in information space.
- *Visualizing the frame of reference*: Effective explorations of spatially referenced information (e.g., health data in certain areas) require the combination of an adequate display of the spatial frame of reference with the visualization of complex information structures. It is necessary to find an appropriate mapping between information and frame of reference. In particular, we address the problem of displaying complex graphs over geographical maps, a problem that has not been widely studied.

Ankerst [4] classifies current visual data mining approaches into three categories. Methods of the first group apply visualization techniques independent of data mining algorithms. The second group uses visualization in order to represent patterns and results from mining algorithms graphically. The third category tightly integrates both mining and visualization algorithms in such a way that intermediate steps of the mining algorithms can be visualized. Furthermore, this tight integration allows users

• *The authors are with the Universität Rostock, FB Informatik, PostBox 999, 18051 Rostock, Germany.*
E-mail: {mkreusel, schumann}@informatik.uni-rostock.de.

Manuscript received 12 apr. 2001; accepted 10 July 2001.
For information on obtaining reprints of this article, please send e-mail to: tocg@computer.org, and reference IEEECS Log Number 114504.

to control and steer the mining process directly based on the given visual feedback.

A variety of visualization methods which have been developed in different domains can be classified into the first group referring to the classification given above. Among these are techniques for visualizing multidimensional information. These methods try to map correlations of objects in high-dimensional information space to spatial correlations in a 2D or 3D presentation space. Among these are approaches like IVORY [10], VR-VIBE [6], and Narcissus [12], which exploit spring models to place objects according to their similarities, whereby similar objects are placed spatially close together. Other systems, like Lyberworld [11] and SPIRE [33], use different visual metaphors like Relevance Spaces [11], Information Galaxies, or Themescapes [33] in order to visualize document collections or results from data base retrieval. FOCUS [25] is an interactive table viewer which supports the exploration of complex object-attribute tables by a combination of a focus+context technique, a hierarchical outliner for large attribute sets and a general easy-to-use dynamic query mechanism.

Other visual interfaces have been developed for visualizing and interacting with hierarchies, like Cone Trees [8] or Disc Trees [16], which use horizontal and vertical cones or discs to layout hierarchies. FSN [28] and Information Pyramids [2] exploit the metaphor of 3D information landscapes to depict large hierarchical information spaces. Other approaches, such as Treemaps [17] and CHEOPS [7], are well-known 2D techniques which use available screen space very effectively.

The visualization of mining models (category 2 of the classification of visual data mining approaches) can be found in [26], where hierarchical cluster structures are discovered and visualized based on implicit surfaces. Other examples are WebSOM [1], which applies color coded planes to visualize results of a Self-Organizing Map algorithm, or OPTICS [4], which displays hierarchical clusterings.

Systems like Descartes [3] or Devise [9] provide solutions for visualizing geographically related information. Different types of icons, diagrams, colored faces, and maps are used for depicting data within their spatial frame of reference. These systems, however, do not support the visualization of rather complex information structures, as, for instance, abstract node link graphs or hierarchies.

Most of the systems mentioned above solve, each in its own manner, some of the single problems introduced earlier in this section. Up to now, there are still open questions of how to provide a flexible framework for solving those problems in a more general way.

The work reported in this paper was inspired by the research stated above. In Section 2, we briefly sketch our approach for modeling information space. We suggest a scalable visualization framework (cf. Section 3) in order to address the introduced problems. Basically, our framework integrates a *scalable preprocessing pipeline* for organizing large unstructured high-dimensional information spaces (see Section 4) with several new *scalable visualization techniques* (cf. Section 5) for visualizing information structure along with information contents, as well as displaying and interacting with mining results. We propose a new paradigm for integrating the visualization of information structures and their spatial frame of reference in Section 6. Future work and conclusions are covered in Section 7.

2 Information Model

The design of a scalable visualization framework requires a formal and easily adaptable information model for describing information units and the general characteristics of the information space. It's our goal to define a general model which is suitable for different domains and a variety of visualization applications.

References [30] and [31] use objects to represent information. In order to formalize this information representation, we introduces the concept of *information objects* IO_i to describe the *information space*. The term "information object" denotes a necessary abstraction of the data which represent the information. Information objects are concrete objects (e.g., documents, files, or real world objects like cars, houses, or cities) which may contain other information objects.

The information set IM is a discrete set of information objects.

$$\mathbf{IM} = \{IO_1, \ldots, IO_n\} \tag{1}$$

$$with \quad IO_i = IO_j \quad \Leftrightarrow \quad i = j \quad i, j, n \in \mathbb{N}.$$

Information objects are characterized by a set of attributes. Those attributes can have arbitrary continuous or categorical ranges of values in order to describe object properties and the characteristics of the information. The function *attr* provides all attributes of a set of information objects.

$$attr(\{IO_1, IO_2, \ldots IO_n\}) = \{A_1, A_2, \ldots, A_k\} \tag{2}$$

$$with \quad A_i = A_j \quad \Leftrightarrow \quad i = j \quad i, j, k, n \in \mathbb{N}.$$

The attribute set AM is the set of all attributes A_i of the set of information objects.

$$\mathbf{AM} = attr(\{IO_1, \ldots, IO_n\}) \qquad n \in \mathbf{N}. \tag{3}$$

Those attributes define dimensions and span the *information space* IR, whereas the ranges of attribute values define the scaling of the related axes of the information space.

The dimensionality of the information space IR is defined as the cardinality of the attribute set AM.

$$dim(\mathbf{IR}) = |\mathbf{AM}|. \tag{4}$$

In other words, the attributes and their ranges of values represent the dimensions of the information space in our model. Thus, information objects IO_i can be understood as points within multidimensional information space.

In order to model arbitrary relations between IO_i which might either be given explicitly or obtained implicitly, we introduce the information structure IS.

The information structure **IS** is defined as a relation on the information set **IS**:

$$\mathbf{IS} \subseteq \mathbf{IM} \times \mathbf{IM}. \tag{5}$$

The absolute value of **IS** may be 0, i.e., in some cases there may be no description of the relation between information objects.

Summarizing our model, the information space **IR** is defined by means of the information set **IM**, attributes which describe the information properties and represent the dimensions of **IR** and the information structure **IS**.

The information definition given above allows modeling of complex information spaces. Arbitrary visualization scenarios can be handled due to the use of attributes for characterizing information objects and the use of relations for describing connections between pieces of information. Spatially referenced information spaces can be described as well when treating the spatial frame of reference as a special attribute.

3 Basic Concept of a Scalable Framework

In order to solve the problems addressed in Section 1, we propose a framework which integrates a scalable preprocessing pipeline and different visualization modules. Basically, our preprocessing pipeline implements several algorithms, such as interactive filters, clustering, dynamic hierarchy computation, and neural networks for analyzing unstructured information spaces. Combining different techniques within a flexible framework helps to scale preprocessing with respect to the characteristics of the information space and users' exploration tasks. In order to display preprocessing results and to explore information space graphically, the framework offers several new visualization techniques as well.

3.1 Scalable Preprocessing

Preprocessing large information spaces often requires reducing the active data size to processible levels without losing relevant information. Other preprocessing tasks, such as gaining structure, identifying groups of related information objects, or forming meaningful subsets of the given data, are nontrivial because there is no general mathematical framework or paradigm on how to build those groups or subsets.

In order to address these problems and to achieve flexibility in the exploration process, we propose two major methods for preprocessing information spaces within our framework. Interactive user-driven approaches are used for selecting dimensions or subsets of the information space manually. Algorithmic computational procedures are applied for obtaining structures and patterns in the data automatically.

3.1.1 *Interactive Preprocessing*

The objective of interactive preprocessing is user-driven information structuring and reduction in order to determine the information which is relevant for the visualization. This is achieved by user controlled filtering out of nonrelevant information. Our framework provides several interaction methods, such as sliders, mouse-based visual selections, etc., in conjunction with different visual previews onto the data set in order to support information selection processes. These interactive procedures are useful because they allow direct considerations of users' domain knowledge and exploration tasks during the preprocessing. Basically our framework offers the following three interactive preprocessing methods:

- **Interactive reduction of the number of dimensions**—Visual previews on multidimensional information sets are used for supporting the selection of those dimensions which might be most relevant for the visualization. These previews are created with a technique which we called Data-Table-View. The Data-Table-View reveals ranges of values, value distributions and correlations between dimensions in order to support a qualified selection of dimensions (see Section 4.1).
- **User-driven filtering of data ranges**—In conjunction with the preview, interactive sliders can be utilized for specifying value intervals such that only those information objects which fulfill predetermined value conditions are shown in the visualization.
- **Interactive hierarchy specification**—Based on the user-defined hierarchy approach introduced in [32], users can impose arbitrary hierarchical organizations on a given information set even if it is not a natural hierarchy. Due to this interactive strategy, users can bring domain-specific and task-specific knowledge to the hierarchy specification that can be utilized for obtaining structures and revealing patterns in the data.

3.1.2 *Algorithmic Preprocessing*

The algorithmic-based preprocessing approach exploits similarities between information objects in high-dimensional information space. Therefore, we have to provide adequate measures $s_{ij} = s(IO_i, IO_j)$ for calculating similarities between information objects IO_i and IO_j.

As stated in [5], computing similarity measures is rather complicated because similarity can be defined in various ways and, often, domain specific expertise is required for determining appropriate measures. Furthermore, the decision if two objects are similar or not is specific to user goals. Let's consider an example. A number of firms are described by the volume of sales over a period of several years. As it is the objective to group those firms with similar sales rates within this time period, Euclidean Distance or some Minkowski Distances [18] are sufficient measures. In contrast to that, the Dot product or a Correlation coefficient [18] are appropriate if it is the intention to group firms with similar sales growth within that period of time. Thus, any of the different measures might be appropriate in certain cases.

Furthermore, the applicability of a specific similarity measure depends on the basic data types of the information object's attribute values. Thus, similarities might have to be computed from variables that are binary, nominal, ratio scaled, or a combination of these (cf. [18] for further information about these data types).