VDM@ECML/PKDD2001: The International Workshop on Visual Data Mining at ECML/PKDD 2001

Simeon J. Simoff
University of Technology Sydney
Sydney, NSW 2007
Australia
simeon@it.uts.edu.au

Abstract
This brief report presents an overview of the International Workshop on Visual Data Mining, conducted on 4 September 2001 in conjunction with the 12th European Conference on Machine Learning (ECML'01) and 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01). It includes summary of the presentations and discussions, and provides pointers to relevant resources in the area.

Keywords
Visual data exploration, information visualization, visual reasoning, data mining.

Introduction
The International Workshop on Visual Data Mining VDM@ECML/PKDD2001, took place in Freiburg, Germany after the Visual Data Mining Workshop at ACM KDD 2001 in San Francisco. The presentations at this workshop, however, did not overlap with the ones presented in San Francisco. The workshop started after Antony Unwin's conference keynote address on "Statistification or Mystification, the need for statistical thought in Visual Data Mining", which setup the context for the workshop presentations. The workshop was followed later in the week by a tutorial on Visual Data Mining and Exploration of Large Databases, conducted by Daniel A. Keim and Mihael Ankerst. This 'package' of events on visual data mining provided ECML/PKDD2001 participants with comprehensive and state-of-art coverage of the methods, techniques, hot issues and applications in the area.

John W. Tukey emphasized that seeing may be believing or disbelieving, but above all, data analysis involves visual, as well as statistical, understanding. Perhaps the most famous and certainly one of the oldest visual explanations in mathematics is the visual proof of the Pythagorean theorem. This proof is unusual in its brevity and its complete appropriateness to the problem. Pictures and diagrams are also used in non-geometrical parts of mathematics, mostly for psychological reasons: harnessing our ability to reason "visually" with the elements of a diagram in order to assist our more purely logical or analytical though processes. Thus, a visual reasoning approach to the area of data mining and machine learning promises to overcome some of the difficulties experienced in the comprehension of the information encoded in data sets and the models derived by other quantitative data mining methods.

Visual data mining is a collection of interactive reflective methods that support exploration of data sets by dynamically adjusting parameters to see how they affect the information being presented. This emerging area in explorative and intelligent data analysis and mining is based on the integration of concepts from computer graphics, visualization metaphors and methods, information and scientific data visualization, visual perception, cognitive psychology, diagrammatic reasoning, visual data formatting and 3D collaborative virtual environments for information visualization. Hence, the diverse program committee, which included James L. Alty (Loughborough University, UK), Heinz-Dieter Boecker (GMD National Research Center for Information Technology, Germany), Chaomei Chen (Brunel University, UK), Di Cook (Iowa State University, USA), John Debenham (University of Technology Sydney, Australia), Alberto Del Bimbo (Universitá degli Studi di Firenze, Italy), Edwin Diday (Université Paris IX - Dauphine, France), Chitra Dorai (IBM Thomas J. Watson Research, USA), Alex Duffy (University of Strathclyde, UK), Erik Granum (Aalborg University, Denmark), Georges Hebrail (EDF R&D, France), Maolin Huang (University of Technology Sydney, Australia), Alfred Inselberg (Multidimensional Graph Ltd and Tel Aviv University, Israel, and San Diego Super-computing Center, USA), Daniel A. Keim (University of Konstanz, Germany), Carlo Lauro (University of Naples, Italy), Torsten Möller (Simon Fraser University, Canada), Bruce Thomas (University of South Australia, Australia), Carl H. Smith (University of Maryland, USA), Masaki Suwa (Chukyo University, Japan), Osmar R. Zaïane (University of Alberta, Canada), and Ahmed Zighed (Universite Lumiere Lyon, France) and the workshop organizers - Simeon J. Simoff (University of Technology Sydney, Australia), Monique Noirhomme-Fraiture (Institut d'Informatique, FUNDP, Namur, Belgium) and Michael H. Böhlen (Aalborg University, Denmark). All papers were extensively reviewed by three referees drawn from the program committee.

Scope of the workshop
The co-occurrence of PKDD and ECML offered unique and perhaps the best opportunity for discussing the latest developments in visual data mining and how machine learning and knowledge discovery methodologies can benefit from it and build on it. The workshop was extensively supported by the 3DVDM group from Aalborg University, Denmark. From the eight submissions selected for presentation, two came out of this group.

The organizers have grouped the papers in three streams: *Methodologies for Visual Data Mining*; *Applications of Visual Data Mining*; and *Support for Visual Data Mining*. The papers in the proceedings are of particular interest to the broad community of researchers in cognitive technologies in data mining and

analysis. The workshop included two additional presentations – one from Erik Granum, the head of the 3DVDM group, who presented an overview of this unique interdisciplinary group, current projects and research opportunities there; and one from Monique Noirhomme-Fraiture, who demonstrated unique visualization tools for visual data mining of symbolic data. Due to overlap with the conference keynote address on visual data mining, the first session on methodologies was moved to the afternoon, hence, the workshop started with the application session. The order of the presentations below follows the actual workshop, which differs from the workshop program in the proceedings.

Applications for Visual Data Mining

The "Applications" session started with the presentation on *Visual Data Mining Using a Constellation Graph* by Tokihiko Niwa (Kwansei Gakuin High School, Nishinomiya, Japan), Kenji Fujikawa, Kazuyoshi Tanaka (Hitachi Systems and Services Ltd., Japan), and Mayumi Oyama (Kwansei Gakuin University, Nishinomiya, Japan). Constellation graph, although not a new visualization model, offers an effective way to display multivariate data on to a two dimensional plane. Authors take this metaphor further, creating an interactive visual data mining tool. The stellar position is determined by an angle that is calculated from the variable value and by a vector that is dependent on the weight of each variable. Once the display is generated, the method allows the researcher to interactively change the weighting factors of the variables in consideration, which affects the constellation graph and the operator can immediately see how changes affect the clustering of the data. When a cluster of stars becomes very cluttered, it is possible to select the cluttered segment for further investigation. The authors demonstrate the use of this technique in two examples - the analysis of data sets with the works of three well-known Japanese writers (Yasushi Inoue, Yukio Mishima and Atsushi Nakajima) and the identification of distinct characteristics of each author, and in the analysis of diabetes diagnosis data. Authors have submitted parts of their work to the Japanese Patent Office. In *Mining Travel Data with a Visualiser*, Colin Ho (BT Asia-Pacific, Hong Kong) and Ben Azvine (BTexact Technologies) presented an unusual visual data mining application in travel pattern analysis. Proposed approach combines the Geographic Information Systems technology of mapping and visualizing geographic data, visual queries techniques and path traversing for visualizing travel patterns. The "Travel Visualizer" system identifies the hot spots depending on the time of the day and provides alternative travel patterns, based on the existing travel data. The travel models are built visually, however, when it comes to estimates and selecting of the feasible travel patterns, the system also provides some explanatory facilities (basically showing the rules that have governed the selection of the the travel patterns). This integrated application can be classified as visual mining by querying. In *Customer Data Mining and Visualization by Generative Topographic Mapping Methods*, Jinsan Yang and Byoung-Tak Zhang (Seoul National University, Seoul, Korea) addressed the application of generative topographic mapping methods to the analysis of web customer data and compared their relative merits on the clustering and visualization with self-organizing maps and the results of principle component analysis. The KDD-CUP 2000 data was used for the investigation. According to presenter, visualizations based on generative topographic mapping methods showed clear underlying structures of clusters. In the analysis missing values were treated as taking another value (= 0). From the knowledge discovery point of view, visualizations of the results of principal component analysis did not show clearly the structure of the data since the complexity of data was beyond linearity. In SOM, the clusters were visualized, but the ambiguity remained to some extent since the expression was limited up to the node set. Generative topographic mapping offered much greater flexibility since data points can be expressed over the whole latent plane, showing each cluster more compactly. The paper will be interesting to web business analysts, that try to understand various characteristics of potential customers.

The session was concluded with Erik Granum's presentation about the 3DVDM group at Aalborg University, Denmark. The large-scale project initiated by the group deals with novel interdisciplinary approaches to data mining by combining expertise in statistics, database systems, visualization, and perceptual and cognitive psychology with facilities for advanced Virtual Reality (VR) 3D visualizations. Through the merging of different disciplines, the collaboration of mature researchers, and the education of young scientists, the project creates a new type of expertise. National companies and institutions can directly exploit the research results and the expertise offered by the group, as well as the special facilities and supporting software. Details about the projects and the opportunities can be obtained at the group web site at http://www.cs.auc.dk/3DVDM/.

Support for Visual Data Mining

Starting an afternoon session, just after lunch is not a trivial exercise. Monique Noirhomme-Fraiture provided "live entertainment" for the workshop audience by demonstrating the Zoom Star visualization of the so-called symbolic variables. Symbolic variables may have values such as subsets of categories, intervals in the real line, or frequency distributions, which differs to the classical approach, where only one single number or category is allowed. There are a number of visualization methods to represent classical multivariate data, aiming at highlighting the correlation between variables. A common feature of all these visual representations is that they work only with variables of the same type, i.e. variables that are either all quantitative or all categorical. The demonstration of Zoom Star showed that it was possible to present variables with values in the form of intervals, sets, weighted values, logical dependencies and taxonomies of values. The demonstration included 2D and 3D editions of the Zoom Star representation. Different levels of details, animated features that allow to analyze the evolution of the symbolic variable – these are just few of the attractive features of this visualization paradigm. Readers, who are interested in the concepts and theory behind the Zoom Star, can find the details in [2]. In *Supporting Data Analysis Through Visualizations*, Paolo Buono, Maria Francesca Costabile, Francesca A. Lisi (Università di Bari, Bari, Italy) presented their work on information visualization in the context of the project FAIRWIS (Trade FAIR Web-based Information Services), funded by the European Union. The project aimed at offering on-line innovative services to support the business processes of trade fairs, both real and/or Web-based virtual fairs. FAIRWIS primarily addresses the needs of professional users in the fair context, namely fair organisers, exhibitors, and visitors attending the fair for business reasons. The module in consideration generated data visualizations on the WWW, which were exploited to facilitate human-computer

interaction, to allow easy access to the stored data, and to present the retrieved information in appropriate ways, thus helping users in their data analysis for marketing activities. The authors presented examples of developed visualizations that support various information queries. By exposing possible information gaps, such visualizations assist narrowing queries and improving information retrieval activities. In *Introducing Signature Exploration: a Means to Aid the Comprehension and Choice of Visualization Algorithms*, Penny Noy and Michael Schroeder (City University, London, UK) discussed two very important issues in visualization and visual data mining – the choice of visualization method and the comprehension of resultant visualizations. They looked at information visualization from a data mining perspective, when a complex data set had to be transformed to a level at which it could be displayed, as constrained by the medium. Hence – the result is usually a general loss of comprehensibility. Comprehensibility can influence significantly the validity of the results – an issue, that was pointed out by Antony Unwin in his keynote presentation. Authors have experimented with visualizations of data involving dimension reduction using their tool Space Explorer - a visualization application for multivariate and proximity data. An initial investigation showed how one data set could be displayed in a number of different ways, producing different clusters and outliers. This is a well known problem, but one for which general solutions are not evident. This work proposed a new concept, signature exploration, defined as the exploration of the behaviour of a visualization algorithm by means of the use of specially constructed data sets. Five types of constructed data were suggested; generic, constructed, query, landmark and feedback. The concept of signature exploration was illustrated with two examples- a feasibility test and a feedback application. The concept has the potential to become a general visualization design requirement.

Methodologies for Visual Data Mining

This session included two presentations of research works from the 3DVDM group at Aalborg University, Denmark. In *Using Nested Surfaces to Detect Structures in Databases*, Artüras Mažeika, Michael Böhlen and Peer Mylov presented the concept of nested surfaces and its computational model (following a topology-based definition of nested surfaces). Nested surfaces can be viewed as approximations that enclose the data at various density levels. The approach is based on the assumption that humans perceive surfaces much easier that individual observations, hence, the nested surface approach to mining data will assist the investigation of the structure of the data more easily than the scatter plot of the data itself. Nested surfaces are capable of smoothing the structures in the data. Presented method allowed identifying arbitrary shaped structures in a data set. The visual support provided for the nested surfaces facilitates the detection of multiple structures, which is important for data mining where the less obvious relationships are often the most interesting ones. Authors presented experimental results which demonstrated that surfaces were fairly robust with respect to the number of observations, and easy to perceive, and intuitive to interpret. The 'package deal' included several algorithms that compute surface grids and surface contours. In *Methods for Visual Mining of Data in Virtual Reality*, Henrik R. Nagel, Erik Granum and Peter Musaeus – another team from the 3DVDM group, presented modular system architecture with a series of tools for explorative analysis of large data sets in 3D virtual reality systems. The assumption behind this work in progress is that immersing the observer in the virtual space created by the 3D data visualization and the real-time navigation within this visualization will support the chances of finding interesting data structures and relationships. The key in the technique is that each observation in a data set is presented in a single geometric shape. The paper includes a very important section connected with the perception of visual cues and the features that facilitate visual data exploration in 3D virtual reality systems (including object and observer positions, pose, size, shape, color, texture, blinking and other dynamic cues). Several new methods for exploring data in 3D immersive environments were presented. Authors are also working on the development of design rules for constructing VR-visualizations for visual exploration by adopting experience from perceptual psychology. Designing visualizations was a central topic in *Towards the Development of Environments for Designing Visualization Support for Visual Data Mining*, where Simeon J. Simoff considered the issues connected with the development of a consistent approach to formal development, evaluation and comparison of visualization methods for visual data mining. The work introduced the Form-Semantics-Function framework as a formal approach towards constructing and evaluating visualization techniques, which was based on metaphor analysis, formalization and evaluation. The application of the framework was illustrated with two examples – one in construction a visualization scheme for identifying patterns in team collaboration and one in comparative evaluation of two visualization schemes used for visualization of text analysis results. The end of the presentation was setup to open the workshop discussion session.

Conclusion

The workshop ended with a discussion on the scope and methodologies of visual data mining. Since visual analysis is such a different technique, it is an extremely significant topic for the growing area of data mining and knowledge discovery research and development. Visualization has proven to be reliable, easy to learn, and extremely cost effective. Presentations at the workshop showed that visualization provides a natural method for integrating multiple data sets and has been used across a number of disciplines in academic and industry research. Participants noted that visual data mining expands visualization research with developing systematic methodologies for visual reasoning, visual data processing and discovery of patterns in visual data representations. During the visual analysis one can discover patterns of interest, based on boundary violations, frequency of occurrence and all sorts of data interdependencies. There was an agreement that the intimate involvement of the human in the twists and turns of the analysis leads to deeper understanding of the data set – sometimes much deeper, than the one that can be achieved with non-visual methods, like neural network training. A major problem in the development of visual data mining methods is that in absolute sense there is no one type of visualisation model that is better than another. The visualisation method of choice will usually be determined by such factors as appropriateness for the application domain, scope of the data sets, how data is mapped onto visualisation schemata, how the visualisation schemata utilises cognitive "strengths" of human information processing and how it overcomes its cognitive limitations. Another issue of concern was the statistical validity

and reliability of discovered patterns. This brought up the importance of the development of formal methods for evaluating the comprehensibility of a visualization model and guidance for the choice of the visualization algorithms. There was an agreement that the outcomes of visual data mining should be checked and validated by other data mining and statistical analysis methods. There was unanimous agreement to expand the research in the area of visual data mining and on the importance to organize another workshop on the topic.

Overall, the organizers evaluated the workshop as a successful one, where the diversity of the presentations sparked substantial discussion at the end of the workshop (the discussion lasted long past the scheduled end of the workshop). The workshop as part of the ECML/PKDD conference provided an excellent opportunity for exchange of insights and ideas, and establishing important contacts between the participants.

In order to support this emerging community, there has been created a forum to continue the explorations of visual data mining issues. The forum at this stage is supported by a mailing list at the following address: **vdm@it.uts.edu.au**

The proceedings of the workshop are available in electronic form at the:

- workshop web site:
http://www-staff.it.uts.edu.au/~simeon/vdm_pkdd2001/

or the ECML/PKDD 2001 conference website:

http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/workshop-proceedings.html

Acknowledgements

We are thankful to the members of the workshop program committee who helped us review the submissions. We are grateful to the ECML/PKDD workshop selection committee for providing us with the opportunity to stage this event. Special thanks go to Johannes Fürnkranz and Stefan Wrobel for the smooth and unobtrusive coordination of the organization of the workshops.

References

[1] Proceedings of the International Workshop on Visual Data Mining, 4 September, 2001, Freiburg, Germany, S. J. Simoff, M. Noirhomme-Fraiture and M. H. Böhlen (eds).

[2] Noirhomme-Fraiture, M. and Rouard, M., Visualizing and editing symbolic objects, in H. -H. Bock and E. Diday, (eds), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, Heidelberg, 2000.