Lecture Notes in Computer Science 4404

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Simeon J. Simoff Michael H. Böhlen
Arturas Mazeika (Eds.)

Visual Data Mining

Theory, Techniques and Tools for Visual Analytics

Springer

Volume Editors

Simeon J. Simoff
University of Western Sydney
School of Computing and Mathematics
NSW 1797, Australia
E-mail: s.simoff@uws.edu.au

Michael H. Böhlen
Arturas Mazeika
Free University of Bozen-Bolzano
Faculty of Computer Science
Dominikanerplatz 3, 39100 Bozen-Bolzano, Italy
E-mail: {boehlen,arturas}@inf.unibz.it

Library of Congress Control Number: 2008931578

CR Subject Classification (1998): H.2.8, I.3, H.5

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-71079-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-71079-0 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12280612 06/3180 5 4 3 2 1 0

Foreword

Visual Data Mining—Opening the Black Box

Knowledge discovery holds the promise of insight into large, otherwise opaque datasets. The nature of what makes a rule interesting to a user has been discussed widely[1] but most agree that it is a subjective quality based on the practical usefulness of the information. Being subjective, the user needs to provide feedback to the system and, as is the case for all systems, the sooner the feedback is given the quicker it can influence the behavior of the system.

There have been some impressive research activities over the past few years but the question to be asked is why is visual data mining only now being investigated commercially? Certainly, there have been arguments for visual data mining for a number of years – Ankerst and others[2] argued in 2002 that current (autonomous and opaque) analysis techniques are inefficient, as they fail to directly embed the user in dataset exploration and that a better solution involves the user and algorithm being more tightly coupled. Grinstein stated that the "*current state of the art data mining tools are automated, but the perfect data mining tool is interactive and highly participatory*," while Han has suggested that the "*data selection and viewing of mining results should be fully interactive, the mining process should be more interactive than the current state of the art and embedded applications should be fairly automated*[2]." A good survey on techniques until 2003 was published by de Oliveira and Levkowitz[3].

However, the deployment of visual data mining (VDM) techniques in commercial products remains low. There are, perhaps, four reasons for this. First, VDM, as a strong sub-discipline of data mining only really started around 2001. Certainly there was important research before then but as an identifiable subcommunity of data mining, the area coalesced around 2001. Second, while things move fast in IT, VDM represents a shift in thinking away from a discipline that itself has yet to settle down commercially. Third, to fully contribute to VDM a researcher/systems developer must be proficient in both data mining and visualization. Since both of these are still developing themselves, the pool from which to find competent VDM researchers and developers is small. Finally, if the embedding is to be done properly, the overarching architecture of the knowledge

[1] Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38 (2006)

[2] Ankerst, M.: The perfect data mining tool: Automated or interactive? In: Panel at ACM SIGKDD 2002, Edmonton, Canada. ACM, New York (2002)

[3] de Oliveira, M.C.F., Levkowitz, H.: From visual data exploration to visual data mining: a survey. IEEE Transactions on Visualization and Computer Graphics 9, 378–394 (2003)

discovery process must be changed. The iterative paradigm of *mine and visualize* must be replaced with the data mining equivalent of direct manipulation[4].

Embedding the user within the discovery process, by, for example, enabling the user to change the mining constraints, results in a finer-grained framework as the interaction between user and system now occurs *during* analysis instead of *between* analysis runs. This overcomes the computer's inability to incorporate evolving knowledge regarding the problem domain and user objectives, not only facilitating the production of a higher quality model, but also reducing analysis time for two reasons. First, the guidance reduces the search space at an earlier stage by discarding areas that are not of interest. Second, it reduces the number of iterations required. It also, through the Hawthorn Effect, has the effect of improving the user's confidence in, and ownership of, the results that are produced[5].

While so-called *guided* data mining methods have been produced for a number of data mining areas including clustering[6], association mining[4,7], and classification[8], there is an architectural aspect to guided data mining, and to VDM in general, that has not been adequately explored and which represents an area for future work.

Another area of future work for the VDM community is quantification. Although the benefits that VDM can provide are clear to us, due to its subjective nature, the benefits of this synergy are not easily quantified and thus may not be as obvious to others. VDM methods can be more time-consuming to develop and thus for VDM to be accepted more widely we must find methods of showing that VDM demonstrates either (or both of) a time improvement or a quality improvement over non-visual methods.

This book has been long awaited. The VDM community has come a long way in a short time. Due to its ability to merge the cognitive ability and contextual awareness of humans with the increasing computational power of data mining systems, VDM is undoubtedly not just a future trend but destined to be one of the main themes for data mining for many years to come.

April 2008 John F. Roddick

[4] Ceglar, A., Roddick, J.F.: GAM - a guidance enabled association mining environment. International Journal of Business Intelligence and Data Mining 2, 3–28 (2007)

[5] Ceglar, A.: Guided Association Mining through Dynamic Constraint Refinement. PhD thesis, Flinders University (2005)

[6] Anderson, D., Anderson, E., Lesh, N., Marks, J., Perlin, K., Ratajczak, D., Ryall, K.: Human guided simple search: combining information visualization and heuristic search. In: Workshop on new paradigms in information visualization and manipulation; In conjunction with the 8th ACM international conference on Information and Knowledge Management, Kansas City, MO, pp. 21–25. ACM Press, New York (2000)

[7] Ng, R., Lakshmanan, L., Han, J., Pang, A.: Exploratory mining and pruning optimizations of constrained association rules. In: 17th ACM SIGACT-SIGMOD-SIGART Symposium on the Principles of Database Systems, Seattle, WA, pp. 13–24. ACM Press, New York (1998)

[8] Ankerst, M., Ester, M., Kriegel, H.P.: Towards an effective cooperation of the user and the computer for classification. In: 6th International Conference on Knowledge Discovery and Data Mining (KDD 2000), Boston, MA, pp. 179–188 (2000)

Preface

John W. Tukey, who made unparalleled contributions to statistics and to science in general during his long career at Bell Labs and Princeton University, emphasized that seeing may be believing or disbelieving, but above all, data analysis involves visual, as well as statistical, understanding. Certainly one of the oldest visual explanations in mathematics is the visual proof of the Pythagorean theorem. The proof, impressive in its brevity and elegance, stresses the power of an interactive visual representation in facilitating our analytical thought processes. Thus, visual reasoning approaches to extracting and comprehension of the information encoded in data sets became the focus of what is called *visual data mining*. The field emerged from the integration of concepts from numerous fields, including computer graphics, visualization metaphors and methods, information and scientific data visualization, visual perception, cognitive psychology, diagrammatic reasoning, 3D virtual reality systems, multimedia and design computing, data mining and online analytical processing, very large databases last, and even collaborative virtual environments.

The importance of the field had already been recognized in the beginning of the decade. This was reflected in the series of visual data mining workshops, conducted at the major international conferences devoted to data mining. Later, the conferences and periodicals in information visualization paid substantial attention to some developments in the field. Commercial tools and the work in several advanced laboratories and research groups across the globe provided working environments for experimenting not only with different methods and techniques for facilitating the human visual system in examination and patterns discovery, and understanding of patterns among massive volumes of multi-dimensional and multi-source data, but also for testing techniques that provide robust and statistically valid visual patterns. It was not until a panel of more than two dozen internationally renowned individuals was assembled, in order to address the shortcomings and drawbacks of the current state of visual information processing, that the need for a systematic and methodological development of visual analytics was placed in the top priorities on the research and development agenda in 2005.

This book aims at addressing this need. Through a collection of 21 chapters selected from more than 46 submissions, it offers a systematic presentation of the state of the art in the field, presenting it in the context of visual analytics. Since visual analysis is such a different technique, it is an extremely significant topic for contemporary data mining and data analysis.

The editors would like to thank all the authors for their contribution to the volume and their patience in addressing reviewers' and editorial feedback. Without their contribution and support the creation of this volume would have been impossible. The editors would like to thank the reviewers for their thorough reviews and detailed comments.

Special thanks go to John Roddick, who, on short notice, kindly accepted the invitation to write the Foreword to the book.

April 2008

Simeon J. Simoff
Michael Böhlen
Arturas Mazeika

Table of Contents

Part 3 – Tools and Applications