Developing a Generic Framework for Constructing Visual Data Mining Tools

Adrian Hargreaves
School of Information Technology
Whitireia NZ
64-4-237 3103
adrian.hargreaves@whitireia.ac.nz

ABSTRACT

A recent study has estimated that the "digital universe" has grown by a factor of nine since 2005 to 1.8 zettabytes (1000 million terrabytes) of data in 2011. Despite this improving ability to store data, it has been suggested that there has been little corresponding improvement in the availability of knowledge. This predicament arises since our capacity to generate and store data by far exceeds our current ability to explore and analyse these vast volumes. Visual approaches to data mining have attempted to address this problem by exploiting human perception in the discovery of interesting and meaningful patterns in abstract data. Through visualisation data may be explored directly and intuitively to gain new insights and to draw cogent conclusions. Visualisation also helps to promote confidence and trust in data mining models and to justify actions taken in light of those models. Although visualisation has proved an effective technique in this regard, as yet little other than guidelines exist to support the development of visual data mining tools. In view of this shortcoming, this paper explores the rationale for constructing a generic framework for the development of visual data mining tools that would attempt to tightly integrate visualisation with data mining techniques. This paper also identifies a programme of research that will lead to the construction of such a framework.

Categories and Subject Descriptors

D.3.3 [**Database Applications**]: Data mining

General Terms

Theory

Keywords

Visual data mining, Conceptual framework, Knowledge discovery.

1. INTRODUCTION

Traditional statistical methods have been available as tools for data analysis long before the advent of computers. However, the efficacy of these tools when applied to datasets stored within machines is limited because of their primary focus on the testing of hypotheses. This focus reflects the traditional concern of statisticians about confirming or rejecting an identified hypothesis, rather than discovering new knowledge within data.

Statistical methods are also focused on problems that typically involve much fewer variables and cases than maybe encountered in many databases [8]. Even a simple database recording an organisation's day-to-day transactions will now typically involve datasets of large volume and dimensionality, reflecting the huge growth in stored data as documented in a recent study by Gantz and Reinsel [12]. As a consequence of the limitations of statistical methods when applied to such datasets, new techniques for extracting patterns from data have emerged. Collectively, these techniques have become known as data mining methods.

Data mining has been defined as the "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [9]. Data mining is therefore concerned with the discovery of new knowledge within data rather than using data to verify existing knowledge, as in statistics. Patterns within datasets are discovered by applying machine learning methods which include classification [27], associations [2], clustering [11] and sequential patterns [23]. However, a central problem in the knowledge discovery process is the identification of patterns that reflect the data analyst's interests and purposes. Indeed, the success of the data mining process is largely dependent upon the availability of the data analyst's insights and biases, even though the process may use unsupervised learning algorithms [16]. Through interaction with the data mining process, data analysts are able to filter out irrelevant patterns and direct the selection of appropriate learning algorithms in an iterative process of knowledge discovery.

One approach for human-computer interaction of particular relevance to data mining is visual analytics [16]. Thomas and Cook [33] define visual analytics as "the science of analytical reasoning facilitated by interactive visual interfaces". Over the last 21 years, a variety of visualisation techniques have emerged that incorporate a visual interface that facilitates interactive browsing and analysis of data. Examples of such applications include Tree Maps [14], Cone Trees [28], Hyperbolic 3D [25] and VisAware [20]. The basic premise upon which all of these applications are based is that patterns, trends and relationships are often identified more easily if data is graphically presented. Visualisation is the first and foremost act of cognition that allows us to create patterns or impose order [21] and is widely considered the most effective approach for transforming complex data into information [1]. A classic example of the benefits of shifting the cognitive load to faster, visual processes is provided by Snow's

This quality assured paper appeared at the 3rd annual conference of Computing and Information Technology Research and Education New Zealand (CITRENZ2012) incorporating the *25th Annual Conference of the National Advisory Committee on Computing Qualifications,* Christchurch, New Zealand, October 8-10, 2012. Mike Lopez and Michael Verhaart, (Eds)..

[31] dot map. Using a visual metaphor rather than a textual or tabular representation allowed Snow to clearly demonstrate a link between an outbreak of cholera and a single water pump in a London street.

Visual data mining (VDM) tools have been developed as a means of analysing large datasets to identify hidden patterns and relationships. VDM adopts a human-computer collaborative approach and combines visualisation techniques with data mining's analytical, mathematical and statistical methods to facilitate the intuitive extraction of hidden knowledge from data [17]. It aims to integrate the user by applying their flexibility, creativity and domain knowledge in the data exploration process. VDM concerns not only the presentation of the results of data analysis, but also the exploration of the underlying datasets and the mining models employed. However, as data mining becomes more extensive in industry and as the number of automated techniques employed increases, there has been a tendency for these models to become more complex [32]. In order to prevent data mining models from becoming unfathomable diviners of hidden knowledge, it is essential to develop techniques that consolidate automatically generated data mining knowledge with experts' domain expertise [4, 30]. In the absence of such techniques, there is a danger that important strategic decisions may be made without fully understanding the reasoning behind them. Thus the motivation behind the visualisation of data mining models can be summarised as issues concerning understanding and trust [32]. If data mining models can be appropriately visualised, users may be more able to reason about the underlying assumptions upon which these models are based and as a consequence, assess the efficacy of such models within a problem domain.

Cognitive research suggests that the design of visual representations and the mechanisms that facilitate interaction with them can influence the data analyst's conceptualisation of the phenomenon represented [22, 24]. A key research problem in visualisation is to discover "expressive and effective visual metaphors for mapping abstract data to visual forms" [7]. Providing appropriate tools that enable data analysts to understand multi-dimensional, multivariate datasets and which facilitate the extraction of useful patterns from such data is a problem of on-going research [18].

2. CHALLENGES AND RESEARCH OPPORTUNITIES IN VISUAL DATA MINING

A preliminary review of current literature concerning visual data mining has revealed the following opportunities for future research.

2.1 Promoting Understanding and Trust

Data mining applications are often deployed amongst a diverse and potentially non-technical end-user community. In contrast, data mining algorithms and the models they generate have tended to increase in complexity at a pace often not matched by visualisation tools. However, unless appropriate visual metaphors are integrated into data mining applications that promote understanding and trust, the models and the results they generate must be accepted on faith. A key challenge then is to exploit the end-user's innate visual processing capabilities to improve the transparency of data mining models. Promoting understanding and trust will also require mechanisms that enable the user to interact with visualisations dynamically, allowing users to iteratively explore, test and modify the mining models they develop [18]. The iterative nature of model development suggests that users also require a history of their exploration and interaction with models [26]. However, developing features that allow users to compare results between models is an area in need of further research and evaluation [16]. By providing mechanisms that promote understanding and trust and that support the exploratory and iterative nature of model development, the predictive quality of these models may be similarly enhanced.

2.2 Providing Re-useable Components

To maximise the potential that visualisation techniques may afford it will be necessary to tailor visual representations to the specific needs of users and their problem domain [13]. In contrast, researchers often seek more generic solutions in the hope that they may be more widely applicable [26]. While successful visualisations often re-use established techniques, they also require customisation to specific problem domains [19]. This suggests that a toolbox approach may be advantageous, in which common re-useable visualisation solutions may be applied to a variety of customised applications. Although attempts have been made to fill this gap [5, 10], very few visualisation tools provide more than a library of existing visualisations from which the user selects. To more fully address this issue, data analysts require a set of re-useable components for building novel customised visual designs [18].

2.3 Evaluation of VDM Tools

To promote a more wide-spread adoption of visual approaches to data mining, it has been suggested that researchers tackle real problems in partnership with business and government organisations [26]. This approach also provides more realistic opportunities to evaluate the potential and limitations of VDM tools. Case studies involving VDM tools in realistic settings have in particular been a neglected area of research [15].

2.4 Principles and Methodology

Although there are indications that the field of visual analytics is maturing [29], to date little other than guidelines exist to support the development of software solutions to problems concerning visualisation [3]. This shortcoming suggests that principles for generating visualisations that are applicable to a diverse range of end-users and a methodology for solving visualisation problems are required. A set of principles and a methodology would also provide a framework for a more complete understanding of the field and assist in the development of novel visualisation techniques [18].

3. PROPOSED RESEARCH AND AGENDA

3.1 Objectives

Based on the current research opportunities presented above, it is proposed that a generic framework for the development of VDM tools should be constructed. This framework would attempt to tightly integrate visualisation and data mining techniques with the objective that VDM tools developed from within this framework would better support data analysts in the following activities:

- The specification of data mining models
- The analysis of data mining models
- The evaluation of data mining models

3.2 Research Questions

In order to construct this framework it is suggested that the following research questions need to be addressed:

- How can visual metaphors be designed so that they promote understanding and trust in data mining models and provide a basis for set of re-useable components for building novel customised visual designs?
- What mechanisms will support data analysts in the comparison and evaluation of the data mining models they develop and the results they generate?
- How can these mechanisms and visual metaphors be incorporated into a conceptual framework that will support the development of effective VDM tools?

3.3 General Approach

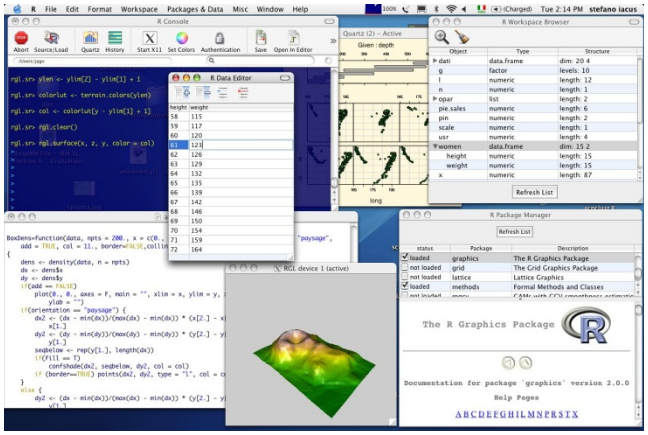
The design of the framework will be initially guided by research from several fields including the seminal work of Bertin [6] and MacEachren [21] in their studies on semiotics and graphical representation. Other significant fields of inquiry will include data mining, visual analytics and human computer interaction. The development of VDM tools for several problem domains within a design science research context will help to further refine the content of the framework. In this way, the knowledge derived from the instantiation of the VDM tools themselves will contribute to the development of the framework. Case studies set in real problem domains will provide a useful vehicle for evaluating the efficacy of a framework for the development of appropriate visualisation tools.

3.4 Technical Considerations

From a technical perspective, these tools will be constructed by extending the functionality of a chosen data mining application (Oracle Data Mining) with advanced and novel visualisation techniques, created and presented from within R; an open source programming language and software environment for statistical computing and graphics.

Oracle Data Mining (ODM) provides a rich variety of data mining algorithms for classification, regression, clustering, attribute importance, association and feature extraction. It also allows developers to fully expose mining models, parameter settings and generated results.

The R statistical programming environment enables the creation of sophisticated graphics and statistical analyses, as shown below. It contains a large set of built-in functions to build custom applications.



```
rgl.sr> ylen <- ylim[2] - ylim[1] + 1

rgl.sr> colorlut <- terrain.colors(ylen)

rgl.sr> col <- colorlut[y - ylim[1] + 1]

rgl.sr> rgl.clear()

rgl.sr> rgl.surface(x, z, y, color = col)
>
>
>
>
>
>
>
>
```

height	weight
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164

```
BoxDens=function(data, npts = 200., x = c(0.,                    "paysage",
    add = TRUE, col = 11., border=FALSE,collin
{
    dens <- density(data, n = npts)
    dx <- dens$x
    dy <- dens$y
    if(add == FALSE)
        plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
            ylab = "")
    if(orientation == "paysage") {
        dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x
            x[1.]
        dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y
            y[1.]
        seqbelow <- rep(y[1.], length(dx))
        if(Fill == T)
            confshade(dx2, seqbelow, dy2, col = col)
        if (border==TRUE) points(dx2, dy2, type = "l", col = c
    }
    else {
        dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * (y[2.] - y
            y[1.]
```

Object	Type	Structure
▶ dati	data.frame	dim: 20 4
g	factor	levels: 10
l	numeric	length: 12
n	numeric	length: 1
▶ opar	list	length: 2
pie.sales	numeric	length: 6
pin	numeric	length: 2
scale	numeric	length: 1
usr	numeric	length: 4
▼ women	data.frame	dim: 15 2
height	numeric	length: 15
weight	numeric	length: 15
x	numeric	length: 87

status	Package	Description
☑ loaded	graphics	The R Graphics Package
☐ not loaded	grid	The Grid Graphics Package
☐ not loaded	lattice	Lattice Graphics
☑ loaded	methods	Formal Methods and Classes

Documentation for package `graphics' version 2.0.0

Figure 1. The R Statistics and Graphics Environment.

ODM and R will be integrated into a single environment using an R package called RODM; R for Oracle Data Mining. The RODM interface allows R applications to mine data using ODM from within the R programming environment. It consists of a set of function wrappers written in R's source language that pass data and parameters from the R environment to an Oracle Database via an Open Database Connectivity (ODBC) interface.

The RODM interface will be used as an environment for the rapid development of VDM tools for selected problem domains and to attempt to address the research challenges identified above. Each VDM tool created in this integrated environment will be evaluated using a case study. An important evaluation criterion will be the ability of data analysts to discover new knowledge as a consequence of applying VDM tools developed from the framework.

4. RESEARCH METHODOLOGY

It is suggested that a mixed method approach be adopted for this research that combines two qualitative techniques; action research and exploratory case studies.

In the proposed research, data analysts within two target organisations will play a crucial contributory role in design of appropriate visual metaphors that will be incorporated into a customised VDM tool tailored to their specific needs. Thus action research could provide a viable framework as a basis for this collaboration as well as delivering a mutually satisfactory outcome.

An exploratory case study will be conducted within two target organisations. Each case study will draw upon primary, qualitative and quantitative data sources in the form of interviews with employees, observation of their business procedures and the analysis of datasets used in each target organisation. Action research will assist in promoting a collaborative setting in which to develop appropriate visual metaphors and VDM tools.

5. SUMMARY

The goal of the proposed research is to construct a generic framework for the development of visual data mining tools that will used to dynamically explore multi-dimensional, multivariate datasets and the mining models employed. To achieve this goal a number of research questions have been defined. These include determining appropriate mechanisms and visual metaphors that improve the transparency of and promote trust within data mining models. These mechanisms will attempt to support the iterative nature of model development by allowing results between mining models to be compared. In addition, the applicability of adopting a toolbox approach in which data analysts are able to tailor common re-useable visualisations for a variety of problem domains will be determined. These research questions will be addressed by conducting action research within a number of exploratory case studies. In addressing these research questions data analysts should possess more effective tools with which to more fully realise the goal of knowledge discovery.

6. REFERENCES

[1] Adams, M. 1995. Situation awareness and the cognitive management of complex systems. *Human Factors 37*, 85–104.Gantz, J. and Reinsel, D. 2010. The digital universe decade, Are you ready? IDC. EMC Corporation. http://www.emc.com/digital_universe.

[2] Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference* (Santiago, Chile), 487-499.Quinlan, J. 1986. Induction of decision trees. *Machine Learning 1*, 81-106.

[3] Aigner, W. 2005. Conference Report. In *Proceedings of the 3rd International Conference on Coordinated and Multiple Views in Exploratory Visualisation* (London, UK). Retrieved March 11, 2012 from http://publik.tuwien.ac.at/files/pub-inf_2994.pdf

[4] Baesens, B., Mues, C., Martens, D., and Vanthienen, J. 2009. 50 Years of data mining and OR: Upcoming trends and challenges. *Journal of the Operational Research Society 60*, S16-S23.

[5] Baumgartner, J., and Borner, K. 2002. Towards an XML toolkit for a software repository supporting information visualization education [Interactive Poster]. *IEEE Information visualization conference* (Boston, MA). Retrieved March 7, 2012 from http://ella.slis.indiana.edu/~katy/paper/ieee-iv02.pdf

[6] Bertin, J. 1981. *Graphics and graphic information processing*. Walter de Gruter, Berlin.

[7] Card, S., Mackinlay, J., and Shneidermann, B. 1999. *Readings in information visualization: Using vision to think.* Morgan Kaufmann, Los Altos, CA.

[8] Elder, J. and Pregibon, D. 1996. A statistical perspective on knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Menlo Park, CA, 83-113.

[9] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, Eds. AAAI/MIT Press, Menlo Park, CA, 1-30.

[10] Fekete, J. 2004. The InfoVis Toolkit. In *Proceedings of the 10th IEEE Symposium on Information Visualization* (Austin, Texas), 167-174.

[11] Fisher, D. 1995. Optimization and simplification of hierarchical clusterings. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (Montreal, Quebec), 118-123.

[12] Gantz, J. and Reinsel, D. 2010. The digital universe decade, Are you ready? IDC. EMC Corporation. http://www.emc.com/digital_universe.

[13] Heer, J., Card, S. K., and Landay, J. A. 2005. Prefuse: A toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon), 421-430.

[14] Johnson, B. and Shneidermann, B. 1991. Tree-maps: A space filling approach to visualization of hierarchical information structures. In *Proceedings of the IEEE Visualization* (San Diego, CA), 284-291.

[15] Kang, Y., Gorg, C., and Stasko, J. 2009. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *Proceeding of the IEEE Symposium on Visual Analytics Science and Technology* (Atlantic City, New Jersey). IEEE, 139-146.

[16] Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. 2008. Visual analytics : Scope and challenges. In *Visual Data Mining* S. Simoff, M. Böhlen, and A. Mazeika, Eds. Springer, New York, 76-90.

[17] Keim, D., Muller, W., and Schumann, H. 2002. Information visualisation and visual data mining: State of the art report. In *Proceedings of the European Association for Computer Graphics* (Saarbrucken, Germany).

[18] Keim, D. and Zhang, L. 2011. Solving problems with visual analytics: challenges and applications. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* (Graz, Austria). ACM, New York, NY, 65, 1.

[19] Lee, B., Parr, C., Campbell, D., and Bederson, B. 2004. How users interact with biodiversity information using taxontree. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Gallipoli, Italy), 320-327.

[20] Livnat, Y., Agutter, J., Foresti, S., and Moon, S. 2005. Visual correlation for situational awareness. *IEEE Symposium on Information Visualization 1(5)*, 95-102.

[21] MacEachren, A. 1995. *How maps work*. Guilford, New York, NY.

[22] MacEachren, A. 1998. *Design and evaluation of a computerised dynamic mapping system interface.* National Centre for Health Statistics, Washington, WA.

[23] Mannila, H., Toivonen, H., and Verkamo, A. 1995. Discovering frequent episodes in sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (Montreal, Quebec), 210-215.

[24] Mark, D. 1989. Cognitive image-schemata for geographic information: Relations to user views and GIS interfaces. In *Proceedings of the GIS/LIS*, (Orlando, FL), 551-560.

[25] Munzner, T. 1998. Exploring large graphs in 3D hyperbolic space. *IEEE Computer Graphics and Applications 18(4)*, 18-23.

[26] Plaisant, C. 2004. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced visual interfaces* (Gallipoli, Italy), 109-116.

[27] Quinlan, J. 1986. Induction of decision trees. *Machine Learning 1*, 81-106.

[28] Robertson, G. and Mackinlay, J. 1991. Cone Trees: Animated 3D visualization of hierarchical information. In *Proceedings of the ACM SIGCHI conference on Human Factors in Computing Systems* (New Orleans, LA), 189-194.

[29] Shen, H., Lee, T. Chaudhuri, A. and Nouanesengsy, B. 2011. Visual analytics for enabling extreme scale scientific discovery. In *Proceedings of SciDac Conference 2011* (Denver, CO), 7-15.

[30] Sinha, A. and Zhao, H. 2008. Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems 46(1)*, 287–299.

[31] Snow, J. 1855. *On the mode of communication of cholera.* John Churchill, London.

[32] Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., and Sommerfield, D. 2001. Visualizing Data Mining Models. In *Information visualization in data mining and knowledge discovery*, U Fayyad, G. G. Grinstein and A. Wierse, Eds. Morgan Kauffman, San Francisco, 205-222.

[33] Thomas, J. and Cook, K. A. 2005. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press, Los Alametos, CA.