

**Interface Techniques Supporting Knowledge Discovery in
Databases**

by

Mark Hollands

`mlhol1@student.monash.edu.au`

Literature Review

Bachelor of Software Engineering with Honours (2150)

Supervisor: Associate Professor Trevor Dix

`Trevor.Dix@infotech.monash.edu.au`

**School of Computer Science and Software Engineering
Monash University**

July, 2003

Contents

Abstract	iii
1 Introduction	1
2 KDD Process	1
3 User-centered Design	2
4 Human Computer Interaction	3
5 Interface Techniques	4
6 Applying Interface Techniques to KDD	4
7 Conclusion	5
References	6

Interface Techniques Supporting Knowledge Discovery in Databases

Mark Hollands, BSE(Hons)
Monash University, 2003

Supervisor: Associate Professor Trevor Dix

Abstract

Data mining is only one step in the Knowledge Discovery in Databases (KDD) process. Applying interface techniques, such as conceptual models, the KDD process can be used to assist the usability and knowledge discovery of data mining systems. First describing data mining and how it fits within the KDD process, this review then moves onto explain how user-centered design techniques can assist in the development of data mining systems. This paper then examines the human-computer interface present in data mining applications and how interface design can assist the user in knowledge discovery, with specific reference to this study.

1 Introduction

Data mining is defined as "the exploration and analysis by automatic or semi-automatic means of large quantities of data in order to discover meaningful patterns or rules" (Berry and LinHoff, 1997). It first received recognition as a significant field of research during the early 1990s. Considerable commercial interest in extracting patterns from large real world data sets has seen it become one of the most active research fields in modern computing. Data mining has combined and built upon, research borrowed from fields such as pattern recognition and machine learning, towards its aim of knowledge discovery.

The term *knowledge discovery* implies more than just mathematical techniques applied to real world data. Without successfully imparting this discovered knowledge to the user, data mining is a pointless exercise. Towards this end, the Knowledge Discovery in Databases (KDD) process has endeavored to build a multidisciplinary framework that aims to support the user's interactions with the underlying mathematical data mining system.

"KDD aims to provide tools to automate (to the degree possible) the entire process of data analysis and the statistician's *art* of hypothesis selection" (Fayyad, Piatetsky-Shapiro and Smyth, 1998). Some of these tools solve specific problems such as: how the data is initially accessed from the source; how algorithms can be scaled to massive data sets and still run efficiently; how results can be interpreted and visualized and a number of other aspects for generally assisting man-machine interactions to make the data mining environment conducive to user learning.

2 KDD Process

The process is iterative in nature; the system begins with low-level real world data sets, combined with user interaction; then refines this to become high-level knowledge and hopefully understanding on the part of the user. The KDD process can be generally described as (Robinson and Shapcott, 2002):

1. Determining what problem is to be solved, i.e. what domains are to be explored. These constitute hypotheses that the data mining process is being conducted to test.
2. Creating the relevant data set for mining from the source database.
3. Pre-processing of the data set, dealing with missing values and/or invalid data.
4. Performing data reduction upon the data set, to simplify or change the domain of the model to be generated. This step is particularly useful after the first iteration in refining the data set to produce a more accurate model.

5. data mining, using the system to build the required model from the data.
6. Analysing the model with potential use of visualisation against the hypotheses. This may be coupled with returning to any of the previous steps for further iteration. Refinement of the domain of the model being processed based upon the information provided by the DM is part of assisting the user to gain a better understanding from a more comprehensive model.
7. Acting upon the discovered knowledge, in using the knowledge directly, in conjunction with another system, or merely documentation and reporting the knowledge.

Different researchers have classified KDD in various ways, dividing or combining certain steps, yet the underlying process is essentially the same. For example, Fayyad et al. (1998) gives a process where the data mining (Step 5) is expanded out to a series of sub-steps covering matching the goals of data mining (Step 1) to a particular data mining method and selecting specific algorithms to use in the data mining step.

3 User-centered Design

The goal of the KDD process is to assist the user's learning towards knowledge discovery. To achieve this goal it is essential to understand the ways that users interact and perceive data mining systems. These principles can be practically applied under the model of software design known as User-Centered Design (UCD). Historically, computer aided learning systems have used certain aspects of UCD to assist the usability of their systems (Lewis, Brand, Cherry and Rader, 1998).

UCD is a software development methodology comprising of three main principles (Gould and Lewis, 1985). The first of which is the way that UCD is primarily concerned with the user's perspective of the system. It involves controlled collection of data from the users regarding their goals and subsequent actions within the system. This involvement should begin early in the software development lifecycle. The second principal is gathering empirical measurement of usage of the system. This can be carried out in the form of usability testing which, combined with feedback from the users, forms an important source of information for further direction of the software. Iterative design is the third principal of UCD software development. The usability goals gathered from interactions in the previous two points are repeatedly assessed through iterations in the development process of design, implementation and testing (Alexander, 2003).

The principal of iterative development can be seen to be very similar to the rapid prototyping model of Software Engineering. Gulliksen, Lantz and Boivie (1999) have stressed

while this can provide invaluable insight in the UCD process it can also lead to *functionality creep* if it is not successfully managed.

4 Human Computer Interaction

The process of human-centered system design subdivides a task between a machine and a human (de Figueiredo and Lai, 2000) :

- The machine acts on computationally intensive sub-tasks.
- The human acts on sub-tasks that require complex perception.
- The human-machine interface reconciles the state of the machine with the state of the human.

This division of tasks relates very well to the interactions between a user and a data mining system. Implementation of a data mining system modeled upon the KDD process seems to lend itself naturally to the principals of UCD. The need for enhanced user interactions in an essentially task-(or job-) orientated system matches well with the structured usability testing focused upon users and their tasks. To successfully *reconcile* the gap between the state of the machine and the state of the human, the language in which the two communicate must be considered.

Basden (1996) has examined data mining systems from the point of view of knowledge base creation. Previously two forms of knowledge have been predominantly used in knowledge base construction. In the first, knowledge is drawn from a domain expert, conceptualized and then assembled into a knowledge base through the use of knowledge representation language and software. The second technique, commonly found in neural networks and case based reasoning, elicits knowledge through rule based induction. He explores that the type of creative thinking used in the KDD process is instantiating a third type of knowledge base creation. It differentiates itself from the two main techniques because the source of much of the content is created by the design process itself. He hypothesizes from this idea that the existing user interface conventions, commonly found in API libraries present in "Windows" computing, are stifling the creative design process by limiting the extent to which a user can be immersed in their task. The interface applications of this hypothesis are discussed further in section 4.

This fits with the thoughts of interface expert, Norman (1988), who speculates that "The best computer programs are the ones in which the computer itself *disappears*, in which you work directly on the problem without having to be aware of the computer."

This concept of making the computer *disappear* is attempting to bridge the gap between the user and computer, or the human - machine interface. When a user is immersed within the environment that the computer provides, the practical limitations of the computer are minimized.

5 Interface Techniques

This can be achieved through the use of one of the most powerful user interface techniques, the conceptual model. A conceptual model assists a user in understanding the way a system works by associating functionality with an existing metaphor they would be familiar with. This familiarity if used well can provide the effect of immersing the user in their task. Due to the iterative and somewhat complex nature of the KDD process, finding a suitable conceptual model is by no means a simple task.

Chattratchat, Guo and Syed (1999) have applied an iconic visual programming technique with graphical visualization techniques to create an internet-based framework for convenient and effective data mining. They model the concept of work flow using the established convention of arranging and linking graphical icons. Together with the use of threading and multi-tasking, they have explored ways to maintain a level of interactivity with the user despite the constraints placed upon the system by deploying over the internet.

Robinson and Shapcott (2002) have expanded upon the concept of work flow, experimenting with a highly graphical representation of data flow attempting to move data visualisation past the classical *charts and graphs*. Using 3d graphics techniques common in computer gaming, they model the concept of water flowing through pipes to describe data *flowing* through the data mining system. Their model's inability to scale to datasets with large numbers of attributes is one of the limitations of an otherwise innovative approach.

Using the creative knowledge discovery principal, a conceptual model focused upon a continual user process was created (Basden, 1996). The user is engaged in a consistent stream of thought that must not be interrupted. Through the use of interface concepts, the system allows the user to explore; using tentative action, to try things; discerning the change in system state on the fly. This is used to stimulate a deeper understanding for the model.

6 Applying Interface Techniques to KDD

Applying the UCD iterative design processes this project will develop a user interface for the existing data mining system, SNOB. Usability testing will measure the effectiveness of system development towards the goal of developing a successful conceptual model based upon the KDD process within the constraints of an internet-based architecture. Specifically, it will attempt to address the following aspects across the KDD process:

1. Determining the goals and hypotheses to be tested by the KDD process. Potentially outside the scope of a data mining system since it is essentially a purely human pursuit.
2. Creating the relevant data set for mining from the source database. Provide means for the user to easily enter the dataset and label the meaningful attributes.
3. The pre-processing phase found in KDD has given rise to the commercial trend of data warehousing, which has become a field in its own right (Fayyad et al., 1998). This refers to collating and cleaning data for ease of use in potential further projects. It has become increasingly popular from a business perspective to have data collected in an easily accessible format, to enable exportation to a range of potential data mining practical uses. By storing data in an easily accessible format, such as XML, this does not place undue constraints upon how the data can be used subsequently.
4. Limiting the domain of the system is most often carried out during the subsequent iterations of the KDD process. It is essential that after completing one iteration of the system, users do not become lost when they are forced to return to an earlier point and modify previous choices. Maintaining a consistent interface throughout the KDD environment can minimize the chances of the user becoming lost.
5. How efficiently data mining algorithms can be scaled, when used upon massive real world datasets, can assist the KDD process by cutting down execution time and giving accurate estimates as to how long the system will take to accomplish a task.
6. Visualisation and interpretation of results play a large role in assisting the user in determining the active state of the system. Based on this information they will be making decisions regarding further iterations through the KDD process. Hence it is essential that they are presented with an informative and up to date picture of the state of the model.
7. Providing facilities to allow the output of knowledge to other applications or documentation facilities to support later data mining exploration. This can support the user in acting upon the knowledge gained from the KDD process.

7 Conclusion

It is essential that data mining systems acknowledge that data mining is only one step in the knowledge discovery process. The user-centered design methodology is well suited to the development of data mining applications. Coupling this with interface techniques, tailored to the iterative nature of the KDD process, these factors can significantly enhance the usability of a data mining system, which ultimately makes the environment more conducive to user learning.

References

- Alexander, D. (2003). Redesign of the monash university web site: A case study in user-centred design methods, *AusWeb'03*. (last accessed July 28th, 2003).
URL: <http://ausweb.scu.edu.au/aw03/papers/alexander/paper.html>
- Baden, A. (1996). User interface for knowledge discovery, *Digest of Colloquium of Knowledge Discovery and Data Mining* .
- Berry, M. and LinHoff, G. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support*, Wiley Computer Publishing, New York.
- Chattratchat, J., Guo, Y. and Syed, J. (1999). A visual language for internet-based data mining and data visualization, *IEEE Symposium on Visual Languages*, IEEE, Tokyo, Japan.
- de Figueiredo, R. J. P. and Lai, G. C. (2000). A perspective of human-centered systems, *IEEE - Circuits and Systems* **11**(2): 3.
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1998). From data mining to knowledge discovery: an overview, *Advances in knowledge discovery in databases* .
- Gould, J. and Lewis, C. (1985). Designing for usability: Key principals and what designers think, *Communications of the ACM*, pp. 300–11.
- Gulliksen, J., Lantz, A. and Boivie, I. (1999). User centered design in practice - problems and possibilities, *User Centered Design in Practice - Problems and Possibilities*, Centre for User Orientated IT Design.
- Lewis, C., Brand, C., Cherry, G. and Rader, C. (1998). Adapting user interface design methods to the design of educational activities, *CHI'98 Proceedings*.
- Norman, D. A. (1988). *Design of Everyday Things*, Bantam Doubleday Dell Publishing Group.
- Robinson, N. and Shapcott, M. (2002). Data mining information visualisation - beyond charts and graphs, *Proceedings of the Sixth International Conference on Information Visualisation (IV'02)*, IEEE.