Visualisation and dynamic querying of large multivariate data sets

James Witter

University of Durham

Department of Computer Science

M.Sc. Thesis

Visualisation and Dynamic Querying of Large Multivariate Data Sets

James Witter

2003

Abstract

The legitimacy and effectiveness of current methods and theories that guide the construction of visualisations is in question and there is a lack of any scientific support for many of these methods. A review of existing visualisation techniques demonstrates some of the innate strengths and weaknesses within the approaches used. By focusing on the more specific task of developing visualisations for large sets of multivariate data, the lack of any kind of guidance in this development process is acknowledged. A prototype visualisation tool based on the well-documented techniques of Parallel Coordinates and Dynamic Queries has been developed taking into account these findings. Incorporating new and novel ideas addressing identified weaknesses in current visualisations, this prototype also provides the basis for demonstrating, testing and evaluating these concepts.

Acknowledgement

I would like to thank the Computer Science Department for their friendliness and encouragement throughout my time at Durham.

I would especially like to thank my supervisors, Prof. Malcolm Munro and Dr. Nigel Thomas for their time, support and guidance throughout my time of study without which this thesis may never have been completed.

I would also like to thank the members of the Visualisation Research Group (VRG) for their valuable feedback and insights into my work.

Copyright

Declaration

No part of the material offered has previously been submitted by the author for a degree in the University of Durham or in any other University. All the work presented here is the sole work of the author and no one else.

Contents

List Of Figures

List Of Tables

Chapter 1 Introduction

1.1 Introduction

The aim of this chapter is to provide an introduction to the work contained within this thesis. It provides an overview of the general problem that the work in this thesis looks to address, and details the approach taken by the research in resolving the problem. The chapter also outlines the criteria for success for this work and outlines the structure of the remainder of the thesis.

1.2 Overview

1.2.1 The problem

In modern times use of the term 'information' has become more widespread; this growth in popularity of the term can be linked to its strong association with new technologies and their related activities.

There is little doubt that information is central to our lives. We are information-seeking organisms [Marc95, March99]. We naturally look for information that can further our knowledge, and the many books that line the shelves of libraries and bookshops all over the world are testament to this. However, in more recent times, information has grown in significance. It has become a commodity. One that is valued almost above all others. Organisations now expend vast amounts of time and other resources on the tasks of collecting and analysing data in the hope that it could provide information that benefits their business processes in some way. Today information is sold to the highest bidder; it's big business. Gathered at all possible junctures, companies use every possible means to collect information, bringing the boundaries of privacy and personal information into question.

Rapid advances in computer and communication technologies are the main driving force behind the increased significance assigned to 'information'. They provide the means of creating, storing, and accessing much greater volumes of data from any number of sources. The development of the Internet and associated technologies has made these large and potentially distributed sources of data available to the masses, and now more than ever people everywhere are confronted with vast amounts of data in their everyday lives.

1.2.2 Research aims and Focus

Data sets are growing in both size and complexity, and it is becoming increasingly difficult to derive useful information from these sources. However, the importance assigned to data continues to grow, and vast resources are employed in its collection and storage. For this reason there is a growing interest from the research community in developing techniques to deal with such data sets, supporting investigation and analysis, and the derivation of information. This interest from the research community has given rise to the creation and development of a number of research fields to include: data mining, knowledge discovery in databases (KDD) and visualisation. The work in this thesis follows the path of visualisation research; an approach that looks to support the human role in deriving information from vast data repositories, instead of replacing them with complex algorithms and computation.

The work first looks to clarify the problem under investigation and the terms key to this and later discussions. Based on these foundations the visualisation research area is introduced, providing background on how the topic developed and details of the current state of research. Focusing on more specialised research into the development of visualisations for large sets of abstract multivariate data, this work aims to identify problems inherent, not only in the development of visualisations for such data sets, but visualisations in general. Based on these findings this work aims to:

- Develop a prototype visualisation capable of representing large sets of abstract multivariate data, taking into account further points identified in the research undertaken.

- Demonstrate the usefulness of this prototype, outlining how it could be used to fulfil different user goals.

- Provide a review of the work to help measure the success of the work undertaken and the progress made.

1.3 Criteria for Success

This section outlines the set of criteria for success of this work. These criteria represent a set of objectives that the work on this project tries to address. It is hoped that by later reviewing these criteria they will provide some insight into the progress made in attaining these objectives, and some indication as to the success of the work.

The criteria for success:

1. To investigate visualisations capable of presenting large amounts of multivariate data.
2. To develop a prototype tool to demonstrate the visualisations developed.
3. To explore the use of these visualisations in a component brokerage system.
4. To develop visualisation mechanisms to handle relationships within data sets.

1.4 Structure of Thesis

This section outlines the contents of each of the remaining six chapters of this thesis.

Chapter 2 provides an introduction to the topic of information and discusses its growing importance in today's society. It begins by outlining the interpretations used in the thesis for terms central to work detailed later in this chapter and throughout the remaining chapters. Based on these foundations, the chapter considers the growing role information plays in society, and the new opportunities and challenges brought about by these changes. The problems individuals and organisations face with the growing volume and complexity of data sets being produced is of central concern, and the chapter presents examples of the problems faced and approaches to solving these issues.

Chapter 3 presents a review of the visualisation research area. It provides background details on how the field of research has developed and outlines how it relates to other research aiming to resolve similar problems. The chapter also contains a review of the current state of visualisation research, focusing on the more specialised problems faced when developing visualisations for large sets of abstract multivariate data.

Chapter 4 presents a prototype visualisation tool. It begins by reiterating the problems that the prototype attempts to resolve, and outlines the approach taken in

its development with details of the rationale behind this approach. The chapter provides an outline of the tool and the functionality it offers, demonstrating how its different features work. Based on this implementation the chapter also identifies problems inherent in the approach taken and in the tool itself, proposing areas that could be subject to further development.

Chapter 5 presents a Case study that demonstrates how the prototype tool could function within a real world situation. A Component Brokerage scenario is outlined, detailing the different users, their objectives, and how the tool would fit into such an environment. The chapter discusses how the tool could be used to support each of the roles identified and help them achieve their objectives, making use of examples based on theoretical data sets to support the discussions.

Chapter 6 contains a criteria based evaluation of the prototype tool. The criteria look to provoke discussion as to the extent of the tool's support for visualising large abstract multivariate data sets, and the extent to which external considerations such as the user and data are taken into account. These discussions provide some indication as to the progress made and the success of the work undertaken.

Chapter 7 contains a conclusion formed on the basis of the work carried out in this thesis. Reviewing the criteria for success outlined in this chapter, the conclusion looks to measure the progress of the work in relation to these points. Based on the findings from this work, the chapter also identifies and outlines opportunities to extend this work in the future.

Chapter 2 Data, Information and Knowledge

"We are drowning in information but starved for knowledge"

- John Naisbett

2.1 Introduction

Information, as we have come to perceive it today, is synonymous with everyday life. Within our social and working lives we are bombarded by information from all areas: from the media through television, radio, magazines, etc., but more significantly from the technology that we surround ourselves with and find ourselves becoming increasingly reliant upon. This, coupled with mounting pressures on people to process, assimilate and react to these growing volumes of information forms the core of the problem approached by this thesis.

By looking at information and its associated problems, this chapter aims to provide support for the exploration and discussion of more refined problems later in this thesis. Addressing these aims, this chapter looks first to develop a consistent view of what is meant by 'information' and related terms. On these foundations, the pivotal role of information and its associated problems in today's society are discussed, and finally the chapter introduces different approaches to resolving some of the issues it identifies.

2.1.1 Data, Information and Knowledge

Historically, the word 'information' has little broad definitional power [Sholl99]. However, in more recent times 'information' is being used more frequently. Influenced greatly by the influx and advancement of technology, especially in data storage and the Internet, use of the term 'information' has become fashionable, associated with many aspects of technology and associated research areas [Sholl99, Buck91, Belk76].

Examples:

Information Systems
Information Processing
Information Science

'Information' has become commonplace in the English language. However, even though the word's use is seen as trivial, the meaning derived from it is not. 'Information' is being used to represent a growing number of concepts within varying contexts, and it is largely accepted that the term has become overloaded with meaning [Sholl99,Buck91,Flor02]. It is becoming more and more difficult to look at 'information' in isolation, without a context: "with the kind of detachment that scientific enquiry demands" [Broo80]. Taking a much-simplified view of 'information', Buckland [Buck91] still manages to identify 3 separate concepts it represents:

1. Information-as-Thing
2. Information-as-Process
3. Information-as-Knowledge

The broad definitional power of 'information' could be seen as one of the reasons for such a broad range of interest in its associated research areas. Nevertheless, it is acknowledged that the different meanings associated with the term are detrimental to its use, and can lead to a lack of understanding.

Information-related research recognises that understanding and defining 'information' is a key component of their research, and that its ambiguous use remains a problem. Attempting to overcome problems connected with the term's use, papers on related topics often look to establish the authors' own interpretation of 'information' before presenting the main topic. By presenting their own interpretation, the author hopes to aid the readers' understanding, but at the same time they are also contributing to the continued overloading of the term.

In an effort to form a foundation for later work, this thesis develops the interpretation of 'information' as it is used in the remainder of the thesis. It is hoped that by taking this approach, the problem of ambiguity cited earlier in this section will be resolved. However, it is impossible to develop a full notion of 'information' without introducing the terms 'data' and 'knowledge'. The meaning of these terms is tightly coupled to that of 'information', and it has been said that a definition of any one of these terms is useless without the definitions of the remaining two [Lehn97, Sten01]. This close relationship is well recognised and is clearly demonstrated in the way many of their definitions rely so heavily on the meanings of the others. The

dictionary definitions of these terms are no exception to this, and the next section presents these, demonstrating this close link.

2.1.2 Dictionary Definitions

(Oxford English Dictionary 2nd Edition - 1991)

Information

i. *"Knowledge communicated concerning some particular fact, subject or event; that of which one is apprised or told; intelligence, news"*

ii. *"An item of information or intelligence; a fact or a circumstance of which one is told"*

iii. *"Separated form, or without implication of reference to a person informed; that which inheres in one of two or more alternative sequences, arrangements, etc., that produce different responses in something, and which is capable of being stored in, transferred by, and communicated by inanimate things."*

Knowledge

i. *"The fact of knowing a thing, state, etc. or a person, acquaintance; familiarity gained from experience."*

ii. *"Acquaintance with a fact; perception, or certain information, a fact or matter; state of being aware or informed; consciousness (of anything) "*

Data (plural of datum)

i. *"A thing given or granted; something known or assumed as fact, and made the basis of reasoning or calculation; an assumption or premises from which inferences are drawn."*

ii. *"The quantities, characters, or symbols on which operations are performed by computers and other automatic equipment, and which*

may be stored or transmitted in the form of electrical signals, records on magnetic tape or punched cards, etc."

Discussion of Definitions

The link between information and knowledge is most clearly demonstrated within the first definition of information. Here, the definition makes explicit use of the term knowledge, describing information as a specialised case of knowledge: *"communicated concerning some particular fact, subject or event"*. The second definition of information relates closely to the first definition for knowledge. Where information is described as *"a fact or a circumstance of which one is told"*, the definition of knowledge suggests that it is the storage of these facts: *"The fact of knowing a thing"*. The second definition of knowledge also makes explicit use of the term information.

The definitions of data and information also demonstrate the close relationship between these terms. Although not making explicit use of the term 'data', the second definition of information describes it as *"a fact or a circumstance of which one is told"*, this relates closely to the first definition of data, *"A thing given or granted; something known or assumed as fact"*.

Evidence of a direct link between data and knowledge is much harder to find in these definitions. Instead they suggest a more indirect relationship through information.

This discussion lends support to presenting a relationship as shown in Figure 2.1, where data is linked to information and information to knowledge.

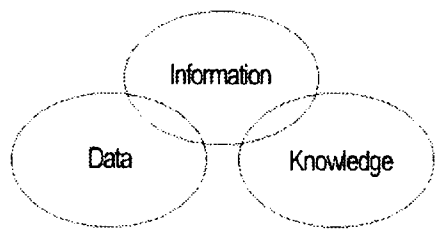


Figure 2.1 - Relationship between Data, Information and Knowledge

Although the definitions in this section demonstrate, to some extent, the existence of links between information-knowledge and information-data, it is not clear from these definitions how they are related. Other definitions emphasise the connection between the terms and provide more evidence for relationship as shown in Figure 2.1.

Example [Flor02, Chec90]:

Information = Data + Meaning

Knowledge = Information + Processing

At a more fundamental level, Bateson defines information as a 'difference':

"A 'bit' of information is definable as a difference which makes a difference." [Bate79].

Breaking down this definition: "a difference" can be described as a discrete state and thus could be defined as data. To "make a difference" the difference must be meaningful in someway, thus Bateson's view of information as a difference can be directly referenced to that in the example given [Flor02].

Confusion over the true meaning and the differences between these terms can lead to one being used in place of another. This is especially the case with data and information [Mach83]. It is not uncommon to find one being used to represent the meaning of the other and sometimes they are used interchangeably to convey the same meaning.

2.2 Definition of Terms

This section aims to outline how information, data and knowledge will be used within the context of this thesis. By interpreting and describing the relationships between these, as well as their attached meaning, it is hoped that the ambiguity in their later use will be greatly reduced.

The section considers these terms from a relatively high level of abstraction, focusing to a greater extent on the concepts they represent. The approach taken, in the main, is from a human perspective and uses ideas based on a number of key resources that focus on this subject [Buck91, Broo80, Quig99, Popp72].

Presenting a model encompassing the three concepts, information, data and knowledge, the section looks to differentiate one term from the others, detailing how the concepts relate to one another and thus how they differ. The section then presents descriptions for each of the terms derived from this study. Each of these descriptions references relevant texts, providing the means to demonstrate similarities or disparities in their interpretations and approaches. However, it must be stated that this section does not in any way represent a definitive guide to these words or their usage. It simply represents an interpretation derived from and, compared and contrasted against, core texts on and around the topic of 'information'.

2.2.1 The Relationship between Data, Information and Knowledge

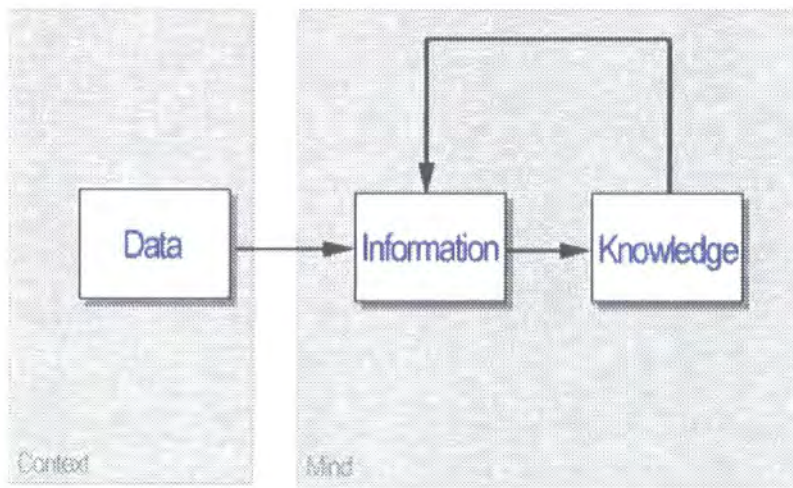


Figure 2.2 - Relationship between Data, Knowledge and Information

Explanation

The model displayed in Figure 2.2 provides an overview of how data, information and knowledge relate to each other. It shows the three concepts within context and, from a relatively high level of abstraction, details of how they are related. This figure does not aim to show the depths of these links, only the fact that they exist.

Overview

Looking at the main entities and the interconnecting, directional lines in Figure 2.2, a process flowing from data through to knowledge is clearly visible. This 'hierarchical style' approach to specifying the relationships between data, information and knowledge is one described by a number of authors [Ack97, Bell97]. Flowing from data to knowledge, the process indicates a derivation path, where the concepts represented by one of the entities can be derived, possibly via some transformation, from another. Evidence for this kind of link between the concepts of data, information and knowledge is also supported in the following definitions:

> Information = Data + Meaning
> Knowledge = Information + Processing

Data

Taking this approach, data could be defined as something from which information can be derived. A perspective supported by the interpretation presented in this section.

The interpretation of data used here, is largely based on work developed by Michael Buckland. Within his work: "Information-As-Thing" [Buck91], Buckland identifies and describes 3 alternative ways in which the term 'information' is used today. One of these alternatives, information-as-thing, forms the basis of the interpretation of data within this thesis.

> Information-as-thing: *The term "information" is also used attributively for objects, such as data and documents, that are referred to as "information" because they are regarded as being informative, as "having the quality of imparting knowledge or communicating information; instructive."* (Oxford English Dictionary, *1989, vol. 7, p. 946).*
> *[Buck91]*

Buckland further supports this notion of data as, informative, 'real world' objects within his framework of information-as-thing when he states: *"we are unable to say confidently of anything that it could not be information"* [Buck91]; a view also taken

by Brookes who describes the physical environment to be as much a source of information as *"marks on a document"* [Broo74]. To demonstrate the dependency of 'information-as-thing' on its context Buckland associates it with the concept of 'evidence', examining how its informative nature is highly dependent upon both the physical and mental contexts within which it is found.

Examples:

- Physical Context:
    - A knife in a kitchen and a knife at a murder scene present us with quite different views of the same object.
- Mental Context:
    - A book written in a language you don't understand is just a collection of pages with marks. You do not have the mental knowledge to decipher its potentially informative content.

Figure 2.2 represents the physical context of data as the encapsulating 'Context' entity around it. The all-encompassing 'Information Context' entity represents the combination of both the mental and physical contexts within which data resides.

This interpretation of data goes much further than approaches taken by Machlup and Belkin & Robertson who focus specifically on interaction between two individuals instead of between an individual and everything else [Mach83, Belk76]. Despite work done elsewhere, sub classing this environment of objects: e.g. the document movement and 'natural sign' (things that are informative without them being intended for communication), this thesis leaves the interpretation of data in this more generalised form.

Following the link from data to information shown in Figure 2.2, the context moves from the physical domain of data to the abstract domain of information and knowledge, representing a movement from the physical world to the abstract world in our minds.

Information

'Information', abstract in nature, is notoriously difficult to define succinctly [Lehn97, Sten01]. The interpretation of 'information' presented here focuses on how 'information' is gained and how it is used. It sees 'information' as the facts and meanings that can be derived from data: *"what is learned by direct observation of the physical environment"* [Broo74]. However, this approach does not reflect the full interpretation of the term presented in Figure 2.2. This shows knowledge as another potential source of information; an approach that is largely based on the work of Wersig [Wers79], who defines 3 sources of information:

1) Generated internally by mental effort
2) Acquired by sheer perception of phenomena
3) Acquired by communication

Sources 2 and 3 fit within this thesis' interpretation of data. However, source one represents the belief that by asserting a certain amount of mental effort new information can be derived from existing knowledge. This link from knowledge to information is included within Figure 2.2.

Knowledge

Knowledge, as with information, represents an abstract concept. Questions similar to "what is knowledge?" have been proposed by many over the years and have been at the centre of much philosophical thought, posing as many questions as they offered answers. However, it is clear that knowledge is somehow stored in the mind. For simplicity, and to avoid a deeply philosophical discussion, this thesis takes knowledge to be a mental state, a state that can be interacted with and transformed both from information generated internally, through some kind of mental effort, and through the incorporation of new information [Kidd94]. This interaction is shown in Figure 2.2 by the arrow moving from information into knowledge.

So far in presenting the interpretation of data, information and knowledge, this section has provided descriptions of the terms and a model defining their interrelationships (see Figure 2.2). The model defines a process flowing from data

through information to knowledge. This section now looks to define the parts of this process in greater depth, emphasising not only how the concepts relate, but also how they differ.

2.2.2 The Relation between Data and Information

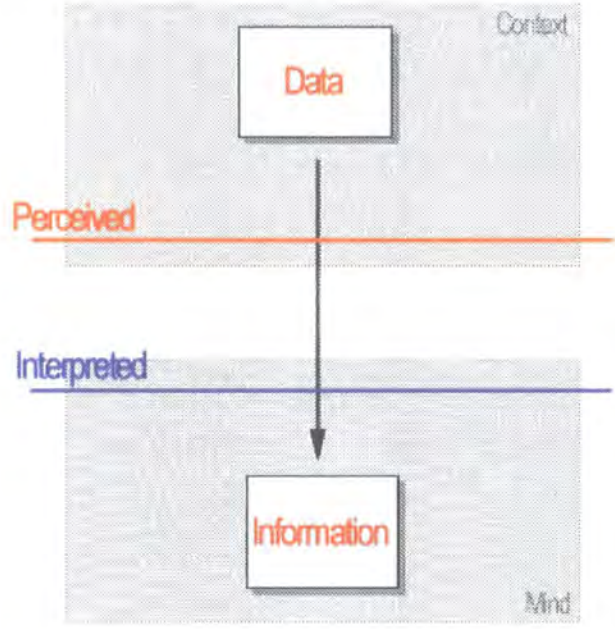


Figure 2.3 - Relationship between Data and Information

Explanation

Figure 2.3 shows the relationship between data and information in greater depth. The model presented aims to show that the derivation of information from data is largely dependent on how the data is perceived and interpreted by an individual. An individual perceives informative features from the physical environment and interprets these using their current knowledge. As Lehner and Maier point out, the physical view of the world is little use without the ability to extract information from the scenes using knowledge built up within the mind:

> *"Information can be generated by perceiving, selecting, and interpreting data according to the person's knowledge" [Lehn97]*

However, this model introduces an interesting discussion as to whether data exists within the 'real world' or within the world that the individual perceives. This is not

a problem looked at by the work in this thesis and so the model described above is assumed.

2.2.3 The Relation between Information and Knowledge

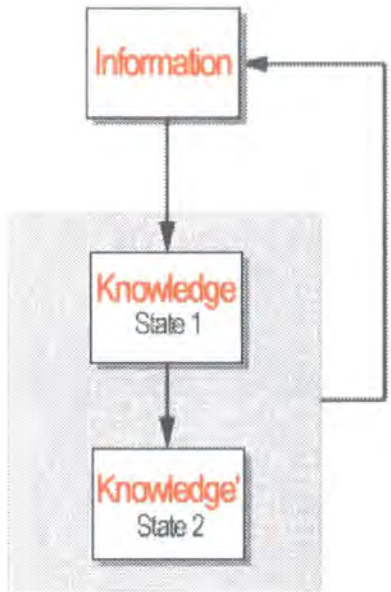


Figure 2.4 - Relationship between Information and Knowledge

Explanation

Figure 2.4 looks in greater depth at the relationships between information and knowledge. The boundary between the concepts of information and knowledge is a great deal fuzzier than that between data and information. Both are amorphous and exist only in the mind, and they are extremely hard to define and therefore to differentiate from each other. However, for the purpose of describing this model, it is only necessary to clarify how information and knowledge interact. In this model knowledge is not directly derived from information; in the words of Kidd *"it does not become knowledge but it alters the existing knowledge by increasing (or shifting) the individual's knowledge state"* [Kidd94]. Thus the arrow pointing from information to knowledge can be interpreted not as a process of derivation but as a process of interaction on an individual's knowledge state. The interaction of information with a knowledge state causes the state to change; this is shown in Figure 2.4 by knowledge state 1 moving to knowledge state 2.

The final arrow moving from knowledge to information is as described in the overview of the model. Thought processes applied on the current knowledge state can give way to the discovery of new information, this in turn may interact with and transform the knowledge state further.

2.2.4 Review of the Model Presented

Although the model presented indicates a process flowing from data to knowledge this does not mean to imply that the process can only flow in one direction. The model represents a view showing how individuals process data, and the impact this can have on their knowledge. However, the model could equally have shown the reverse process, where knowledge flows through to data: this would be the case if an individual wanted to communicate part of their knowledge. The flow of the model presented also does not imply that data is any less important than knowledge. Data can be understood and processed (possibly by computer) and so could be considered to be of much greater value than information and knowledge. Toumi presents the hierarchy in reverse for this purpose [Toum99].

The model described in this section maps quite closely to the work of philosopher Karl Popper. Although Popper's work, for the most part, looked at a quite different problem, his development of 3 worlds is similar to some of the concepts portrayed in this section [Popp72].

2.2.5 Summary of the Terms

Data

- Any physical entity within the 'real' world
- The information that can be derived from data is very much dependent upon the context within which you find it: both physically and mentally
- It needs to be both perceived and interpreted for information to be derived
- All data can hold information and thus has a certain amount of information potential

Information

- Derived from data
- Abstract entity that exists only within the mind, and not in the real world

- Information can be derived from data but it can also be gained through the application of mental effort on an individuals existing knowledge state.

Knowledge

- Again this represents an abstract entity that exists only in the mind
- It is a mental state
- Information interacts with and can alter an individuals knowledge

2.3 Information and Society

It is largely accepted that we are living in, or at least are moving towards becoming, an information society [Sholl99, Quig99, Gog99]. Society is evolving. The industrial revolution transformed the world from a culture of subsistence to one of mass production and choice. Now, it is said that we are in the midst of an 'information revolution', shifting society from these industrial foundations to a more information-centred culture.

Humans are essentially 'information seeking organisms' [Marc95, March99] and information as a route to extending our knowledge has always been a crucial part of our lives. However, in more recent times the role of information has gained in significance. Advancements in information and communication technologies, and the convergence of these two disciplines provide the ability to generate, store and offer access to increasingly large volumes of data. This data and the potential information it carries are essential to the decision-making processes of the individual and organisation, a vital component that governs the ability of both to survive and thrive in this 'information society'. Thus the need to derive useful information from these data collections is now more than ever central to the working and social lives of a growing number of individuals.

"Information is the life blood of modern society" [Gog99]

Data as a source of potential information is now recognised as a valuable commodity, and more business resources are being directed towards the need to gain, produce and manage this resource. The need to derive the information inherent in this data has become 'big business': the global market for information goods, services and technology has been estimated to be over $1.2 trillion [Guv98]. Roszak comments on the changing face of information and its new role as a valuable commodity:

> *"in the 1950's, information had come to be identified with the secret of life. By the 1970's, it had achieved an even more exalted status. It had become a commodity, the most valuable commodity in business" [Rosz86]*

As all sectors of the economy become more information intensive, more and more people are being employed to collect, process, or transmit data.

> *"The manual and unskilled worker class is shrinking in the society, while at the other end of the continuum the class of knowledge workers is becoming prominent."*
> *- Cambridge Reform Club, 1873, 'The future of the Working Classes' [Bell74]*

Computers are the primary tools in this 'information society'. As well as playing a role in the development of data sources, computers play an important mediating role between data and the individual. However, they are no longer tools that simply enable people to work with and handle large sets of data; we are becoming more *"intimate with our technology"* [Hill97] and computers are increasingly responsible for the shape of our working and social lives. The development of the Internet has played a major role in giving individuals access to the increasing volumes of data being produced. It could be said that computers provide a window into this world of potentially invaluable information, but the Internet provides a vehicle through which the data can be accessed. It provides *"a large market place where information demand meets information supply"* [Prop99], providing direct access to both data collections and people as sources of information covering an unimaginable range of topics.

There is little doubt that today's society is changing rapidly, so much so that this period has been likened to that of the industrial revolution. The industrial revolution altered the way we process, sort and rearrange, recombine and transport atoms and it has been suggested that this period is facilitating similar changes for data [Brow00].

2.3.1 The Promises of the Information Society

Our world continues to advance, and as technology becomes available to the masses it moves closer to becoming a global 'information society'. However, as we move towards this state, we are only beginning to understand and realise the potential benefits it has to offer. Potential that has been the subject of much discussion and hype:

Borgman comments that:

> *"The information society has the potential to improve the quality of life of European citizens, the efficiency of our social and economic organisation and to reinforce cohesion." [Borg00]*

Marchionini [March99] reports that the information society promises much, including:

- Increased productivity
- Increased collaboration and participatory democracy
- And improved health and quality of life

However, he also states that reaching this potential is dependent on having all of the following in place:

- Ability to rapidly access comprehensive amounts of information
- New storage technology along with new organisational tools and techniques
- Powerful analytical tools (e.g. statistical packages and data mining procedures)
- Global communications

Many of these promises remain simply potential benefits of the 'information society', yet to be realised, if they can be realised at all. Nonetheless, some of this potential is beginning to benefit individuals and organisations with access to the required technologies.

2.3.2 The Challenges of the Information Society

As well as providing great potential, it is clear that the information society also introduces new problems and challenges, some of which must be resolved in order for its potential to be fully realised. In response to investigating the promises of this new culture, Marchionini details some of the problems we already face in this increasingly information centred world:

- Information overload and multi-tasking stresses due to the increased volume and complexity of data along with increased expectancy on the individual.

- Various inequities involving opportunities to access and derive information from the growing pool of data.
- Disorientation, distraction and addiction within the large and complex data sets.
- Privacy and secrecy especially with sensitive data
- Social control

[March99]

These identified problems are predominantly related to issues humans encounter when handling and working with many, increasingly large and complex information sources.

As society continues to become more information focused, the requirements placed on its citizens are changing. Increased use of information and communications technologies, in conjunction with the need to cope with evermore massive and complex data sets, are changing the skills required to work and function effectively in today's society: people are expected to keep up with the latest technologies and the techniques associated with their use. This constant need to develop new skills to keep pace with the changing requirements of the information society has caused several people to propose the need for lifelong and self-directed learning. However, a process of lifelong learning must be supported by mechanisms allowing individuals to quickly learn and develop the required information skills. In the same way literacy helps us to learn new skills in the real world, the new requirements of the information society require us to be information literate.

Information Literacy

As the lives of the individual centre more and more around information, there is little doubt as to the benefit of having, and developing, information skills. However, the questions of what these skills are, and what is meant by being 'information literate" still remain.

There are a variety of propositions on what constitutes information literacy. Some see information literacy as an extension of computer literacy, whereas the library community see it as an extension to the bibliographic instruction. A broader view of information literacy is presented by Shapiro & Hughes. They list 7 components that aggregate to make up information literacy [Shap96]:

1. Tool literacy – relating to traditional computer literacy
2. Resource literacy – main part of bibliographic instruction
3. Socio-structural literacy – looking at the context for the data
4. Research literacy – methods and the tools involved
5. Publishing literacy – outputting the content
6. Emerging technology literacy – relates to the idea of lifelong learning
7. Critical literacy – evaluation of information sources and technologies

Marchionini links the concept of information literacy to the development of what he calls 'personal information infrastructures' [Marc95, March99]. Individuals develop and build up these structures - which include conscious and unconscious filtering and finding strategies - from birth to help deal with large amounts of information and reduce the risk of information overload (see section 2.4.1). These structures are composed of:

> *"Mental models for knowledge domains, search systems, past information-seeking events; general cognitive skills and specific information skills; attitudes and mental control mechanisms; and material resources." [March99]*

Information Seeking [March99, Mizz96] is a fundamental and high-level human process related to learning that is often embedded within much larger processes. Humans have always been "information seeking organisms" [Marc95, March99]; however, with the development of the information society, information seeking has become a fundamental skill for large proportions of the population. Thus, as a critical activity within the information society, information seeking can be seen as an important component of what is meant by being information literate.

This concept of information seeking is closely linked to that of information retrieval. Both look to search through information; however, information seeking in general is interpreted as a more human-centred activity where the element sought may not be known. Nonetheless, the concepts of information need and relevance are central to both these themes:

- Information Need – initiator for the information seeking process: knowledge gap or information problem. [Mizz96,Tayl86,Derv86]
- Relevance - important in identifying information sources of particular interest but also for looking at the performance of a given information-seeking strategy. [Marc95, Mizz96, Mizz98]

"Relevant information is discovered and then absorbed by the user to fill the knowledge gap" [Prop99]

This most human of activities is becoming increasingly coupled with technology. However, although support has been developed for some parts of the information seeking process, it is sadly lacking in others.

Emphasis remains on the individual to develop these 'information skills' with little or no support from technology. As society moves towards becoming even more information centred, little work is being done to educate its citizens of these changing requirements, or in developing technology to support the new role of the individual.

The impact of our changing society stretches further than the individual, manipulating the structure of our working and social lives. The effect on organisations is clear. New technology is becoming the focal point for a growing number of organisations and increasingly their structure is tailored to meet its associated needs. Nevertheless, many of these organisational issues originate from the core changes and problems faced by individuals.

The need for individuals to constantly develop their information skills is clear, if only to keep pace with the complex computer systems we interact with. However, this approach focuses solely on taking advantage of human adaptability and learning. There is a clear case for looking into how technology and data can best take advantage of the existing and powerful skill sets individuals possess. The next section addresses some of these points by considering issues humans face with information.

2.4 Human problems with Information

The aim of this section is demonstrate some of the problems individuals face within the 'information society'. The problems detailed relate closely to the points raised

within the 'Challenges to the information society' section (section 2.3.2), and look to focus on those problems rooted in the abilities of humans, and the nature of their society and culture.

In the main, this section concentrates on the problem of 'information overload' that has been identified and documented by many. Furthermore, security issues with information are also examined, along with a brief outline of other ethical and organisational issues.

2.4.1 Information Overload

The concept of 'information overload' has been noted as one of the major problems people face as society becomes more information focused [March99, Mizz96]. Nevertheless, as with the term 'information', the interpretation of what is meant by 'information overload' is open to conjecture. The problem of 'information overload' has two main interpretations, based on two of the main uses of the term 'information':

1. Information as an abstract, mental entity – the interpretation used in this thesis (see section 2.2)
2. Information as data

'Information overload' may have a confused meaning, but there is little dispute over the cause of it. Advancements in technology and the development of the Internet have opened up growing numbers of large data pools to the masses. Where once there had been a lack of information sources on which to base sound decision-making processes, many sources now exist from which to derive relevant information. The two interpretations of 'information overload' look at the problems caused by this increased availability of information resources from different levels.

The first interpretation looks at the problems associated with the growing amount of information now available. This greater volume of information available provides support for educated decision making. However, as Roszak points out, having large amounts of information available is not always good:

"An excess of information may actually crowd ideas, leaving the mind (young minds especially) distracted by sterile, disconnected facts, lost among the shapeless heads of data" [Rosz86]

The amount of information available may be much greater, but not all this information is necessarily relevant to the task at hand. Excess information that is not relevant can distract from the task in hand, possibly resulting in the task taking much longer to complete. Marhionini describes this problem:

"we travel a narrow road towards our goals with a sea of seductive information to distract us on one side and a spiralling abyss of confusion and information overload on the other" [Marc95]

Technology is seen as a means of helping to resolve these problems, providing support for the identification and filtering of relevant information sources. However, Marchionini comments that although technology may help reduce the time to reach a stated goal, it also increases the possibility for distraction: the dangers on each side of the road. Technology may provide the means to improve our abilities and performance, but it can also distract and confuse. We are facing a paradox: as the availability of information increases, access to relevant information becomes increasingly difficult [Fab01].

"We are drowning in information but starved for knowledge"
- John Naisbett

The second interpretation of 'information overload' is that focused upon by work within this thesis. It is used to describe the problems associated with deriving potentially useful information from the large volumes of complex data now available. These vast volumes of data are seen as a commodity of great value, a source of potentially invaluable information that could be instrumental in the success of an individual or organisation. Thus there is increasing pressure on the growing masses of the 'information society' to gain, interpret and assimilate information from these swelling pools of data. Furthermore, as data grows both in volume and complexity it is becoming increasingly difficult to derive relevant information

through both natural and automated methods. Rephrasing John Naisbettt's quote it could be said that:

> *"We are drowning in data but starving for information"*

The major driving force being the growing volumes of data is the advances made in information technology that looks set of continue progressing as predicted in Moore's Law (computing power doubles every 18 months). Taking into consideration this and the expansion of the Internet, this kind of information overload has reached a new dimension. The volumes and complexity of data containing potentially useful information continues to increase, but the human abilities to take this input, derive relevant information and react remains relatively constant, and promises that the increased processing power of computers may help to resolve these issues seem unfounded. No longer do people have problems finding or accessing sources of potentially useful information. The problem now lies firmly with the abilities of humans to deal with the mass, scale and complexity of these sources.

Forming the source of this bottleneck, human cognitive and perceptual skills are key to our ability to harvest information from data. Human perceptual and cognitive abilities are remarkable and unique. We live to seek and build up knowledge of the world around us and develop frameworks that enable us to process more complex concepts. However, at the same time there is much that we cannot perceive [Pree94] and it is clear that our abilities need help and guidance to cope in today's information intensive society.

The way in which humans process information and the extent and limitations of these abilities has been the subject of much research over the years. A major preoccupation within the field of cognitive psychology, in more recent times, advances in technology has meant that interest in this area has spread to include research from more technology driven fields such as Human Computer Interface (HCI). Over the years various models and theories of these processes have been developed, assumed, and later tested through experiment:

> Information processing, Human information processing and more recently: Distributed cognition, Computation and Connectionist theories

[Pree94]

These processes have also been studied from a communication theory point of view following the lines of Shannon's theory of communication [Wea49]. Most of these studies focus on perception from a visual perspective, primarily because of the dominant nature of the visual cortex in the human sensory system. The models developed provide a foundation for further research into the areas that contribute to our ability to perceive and interpret data. Identifying potential strengths as well as weaknesses, these areas include:

- The eye and its abilities to discern colour and other graphical properties [Cleve85, Trei86]
- Human memory [Wick87, Atk68]
- Mental Modelling [Marc95, John83]
- Attentive and Pre-attentive processing of visual scenes [Trei80, Heal93]

[Pree94]

However, despite the volume of work undertaken, very little is known about how these human processes work. Research has helped to reveal the extent of human perception and cognition, and their limits have been recognised. It is these limits that remain the main bottleneck in today's information-intensive environments, and we continue to investigate how these abilities may be augmented through the use of technology.

Note: From this point onwards the Second interpretation of information overload is assumed.

2.4.2 Information Security

The development of new and more powerful technologies coupled with the growing mass, complexity and importance of data, and more importantly the potential information it holds, has raised many issues with respect to security.

- How do you protect this potentially invaluable commodity?
- Who is allowed access to this data?
- What can this data be used for?

- What will the information derived from the data be used for?

More and more people and organisations are storing and transferring increasingly sensitive data in electronic form. As consumer confidence in the Internet grows, people readily trust the security technology and protocols in place, providing personal and credit card details. Most of the world's money exists only in electrical form, stored as digital signatures on a disk or in memory; very little money has any physical form. Rather than armoured vehicles, electrical current is responsible for the vast majority of monetary transfers and payments. In this information society we place a huge amount of trust, almost blind faith, in the technology and protocols in place for the storage and transfer of data. However, all this data is open to abuse and exploitation through a number of different avenues.

- A company storing personal information may take advantage of this source of information for their own personal gain, or may be responsible for passing it on to a third party to the same effect, without the consent of the individuals involved.

- Any data or transfer of data that can be remotely accessed is open to attack by hackers. This hacking may give access to the data or alter the data in some way to its detriment.

Solutions to these problems lie mainly within legislation and technology:

- Legislation – used to protect the rights of the individual and organisation with respect to their personal information, e.g. the data protection act.

- Technology – used to develop new and elaborate security measures to prevent unauthorised access to the data and its contents. This includes developments such as the use of firewalls and the use of encryption.

As technology develops it enables us to develop more elaborate security systems to protect our data; nevertheless, it also provides increased power to the people who

may wish to break these systems. Thus the information society can be seen to have given rise to a new era of crime, an era where the criminal has no face, only an electronic signal that can be traced to their terminal.

2.4.3 Other ethical and organisational issues

The development of the Internet can be likened in many ways to the invention of the Gutenburg printing press in the 15th century. Although they have both contributed greatly to the proliferation of information and therefore knowledge, they can also be seen as a major contributor in the spread of propaganda and pornography. In essence the issues of propaganda and pornography do not represent the whole issue; it is the content and the availability of the material that represent the real problem.

Pornography represents one of the biggest and the most profitable industries on the World Wide Web. This in itself is not the problem. It is the possible content of these web sites and their accessibility that are the real issues.

- It is possible for the sites to have obscene and illegal content.
- People of all ages use the Internet and therefore have access to all it can offer: good and bad.

The Internet also holds the ability to act as a vehicle for making information available to the masses. Although this offers a powerful tool for the presentation of important and useful information, it also provides extremist groups with a channel that allows their influence into the homes of millions.

The faceless world of the computer and the Internet offer new opportunities that can, and have been exploited. Anyone with access to a computer connected to the Internet can produce a web page with the potential of communicating directly to millions of people across the world, without the need for identification. This raises issues such as censorship and the need to police the virtual world of the web. This may go against the global information access idealism of both the information society and the World Wide Web, but nevertheless some kind of policing is clearly required.

This is by no way an exhaustive list of the issues people face in the 'information society'; it hopes instead to demonstrate the extent to which humans are at the centre of foreseen problems with information and associated technologies.

2.5 Real World Examples

This section aims to demonstrate some real world examples of information overload and its associated problems. Information overload is one of the main issues society faces as information plays an increasingly pivotal role in our lives. By providing real world examples of information overload and its associated problems, this section aims to demonstrate their diverse and far reaching nature.

2.5.1 Software Development

It is widely acknowledge that the software industry is, and has been for some time, in a state of crisis. Commercial software projects are all too often delivered outside the requirements (time, cost, functionality, quality, etc) of the customer, and in extreme cases are never completed. The problems apparent in software engineering are the subject of much debate. Over the years there have been many false dawns in resolving these problems, but the search for Brookes' elusive silver bullet continues [Broo86].

One of the major problems faced by software developers is the growing size and complexity of systems they are being asked to produce. The programs produced, without considering any accompanying documentation, represent a large, potentially invaluable source of information about the developed system. Throughout the life of software there is a need to extract, understand and communicate information from these ever more complex and interrelated sources; processes that are critical to the success of the projects undertaken. It is no longer feasible for software to be developed by one person in a matter of weeks. Now, the majority of software developed is produced by teams, and can take months or even years to complete. The way the team organise themselves and cooperate is a key component contributing to the success of a software project. The ability to gain information from the available sources is critical to developing a common understanding of the development goals and progress, enabling the team to work together more effectively.

Software code and its accompanying documentation are of even greater importance after the project's delivery; when the delivered system must be maintained and updated as required. During this period it is unlikely that a team of maintainers will have any of the understanding gained by the original developers. Instead they are completely reliant upon the software itself and its accompanying documentation. All their understanding of the system and how it works has to be gleaned from these sources. However, given the low level complexity of the code and the lack of standardisation in accompanying documents, this is not an easy task, and has been the topic of much research into comprehending and maintaining software systems.

2.5.2 The World Wide Web

The World Wide Web is the biggest source of information available in the world today, and as everyone seeks to introduce his or her own presence in 'hyperspace' it looks set to expand even further. The World Wide Web has no fixed or predefined structure and finding sources of relevant information can be far from easy.

Search Engines such as Google remain the only realistic means of identifying sources of useful information. However, the results from using such tools can be far from effective, producing a whole array of irrelevant, often distracting information. Browsing and navigating the web can also prove problematic. Missing links, the turnover of URL's (Universal Resource Locators), and the complicated patterns of links between different web pages only add to these problems.

This expanding resource also offers access to other potentially invaluable data values based on it own structure, providing data that could be used to:

- Identify the existence of broken links on a web page
- Determine the structure of a web site or a group of connected sites
- Identify and reclaim web space no longer in use

2.6 Conclusions

Information is recognised as a powerful commodity and people are becoming increasingly aware of the information sources around them. However, as more try to 'mine' useful information from these sources, the problem of information overload is becoming increasingly prevalent.

The growing emphasis placed on deriving information from available sources has given rise to an influx of research looking to resolve some of the problems associated

with information overload. Looking at the problems from different perspectives, the research incorporates effort and resources from a wide range of fields. Approaches that include:

- The application of mathematical and statistical methods
- The application of algorithms
- The use of Artificial Intelligence and Machine Learning techniques
- The use visualisation techniques

The remainder of this thesis focuses on visualisation as an approach to resolving some of the issues associated with information overload. With the aim of focusing on the more specialised problem investigated by the work in this thesis, an overview of visualisation research and its relationship with other areas approaching similar issues is provided. This does not mean to advocate visualisation as the "silver bullet" solution for information overload. It instead introduces visualisation as a possible means of addressing some of the problems inherent in information overload.

Chapter 3 Visualisation

"I never waste memory on things that can easily be stored and retrieved from elsewhere."

- Albert Einstein (1879-1955)

3.1 Introduction

Chapter 2 introduces the topic of information: it outlines some of the problems faced by a society that is becoming more information orientated and focuses on the problem of 'information overload'. This chapter introduces visualisation research as an approach to addressing some of the problems associated with information overload: presenting an overview of the topic, including details of its origins and relationship with fields approaching the same or similar problems. This chapter also aims to focus attention on the more specialised problems connected with visualising large sets of potentially abstract, multivariate data targeted by the work in this thesis. For this reason the chapter's discussion of visualisation has been extended to include the topics of information visualisation and multivariate visualisations, presenting the more specialised problems faced by each. In presenting these problems and those present within visualisation research as a whole, as well as providing a review of current visualisation techniques, the chapter also looks at what the future may hold for visualisation, identifying points of weakness in current research practices and proposing how the research areas might develop.

3.2 Visualisation Background

Data is an important commodity in the modern world. A resource of potentially invaluable information, more and more companies believe that information inherent in the data they collect is, or at least could be, critical to their business processes and ultimately their success.

However, the human ability to comprehend these data sets that are both growing in size and complexity is severely limited, restricted by our own mental and perceptual competence. Norman points out that, *"The power of the unaided mind is highly overrated"* [Norm93]. We are easily overawed by the volume and complexity of data.

"But human intelligence is highly flexible and adaptive, superb at inventing procedures and objects that overcome its own limits. The real powers come from devising external aids that enhance our cognitive abilities. How have we increased memory thought and reasoning? By invention of external aids: It is things that make us smart." [Norm93]

It is this human capacity to utilise and transform our surroundings into tools complementing our own natural abilities that is our great strength. Making use of the undeniable link between what we see and what we think, these aids help us "use vision to think".

Developments in computer technology now enable highly detailed and interactive graphics to be produced. Machines capable of handling large data sets and producing such detailed graphics, combined with the possibility for direct interaction and manipulation offer a new and potentially powerful medium with which to further enhance our cognitive abilities. Visualisation research looks to take advantage of this new medium in the development of new visual and interactive cognitive aids. The resultant 'visualisations' provide a link between the two most powerful information-processing systems: the human mind and the modern computer, each with their comparative strengths and weaknesses. Card, Mackinlay et al. define visualisation as:

"The use of computer-supported, interactive, visual representations of data to amplify cognition" [Card99a]

Mapping data into a visual form, the goal of visualisation is to extend human cognition and provide useful insights into the data presented. Hamming identified that the *"purpose of computation is insight, not numbers"* [Hamm73] and in the same vein it was later suggested that the *"purpose of visualisation is insight not pictures"* [Card99a]. This need to gain insight into data has led to the development of visualisations supporting different objectives, such as: discovery, decision-making, and exploration. However, the common aim of all visualisations is to present the underlying data in an effective and expressive manor, hopefully leading

to a reduction in the cost associated with deriving the data's inherent information: the "cost-of-knowledge" [Card99b].

3.2.1 Origins of Visualisation

The use of graphical aids to support thought processes has ancient roots. However, it is only in much more recent times that studies in this area have been carried out in any great depth. Born within statistics and statistical analysis, the study of data graphics looks into the *"use of abstract, non-representational visual representations of data to amplify cognition"* [Card99a] and the foundations of visualisation research have developed from this. Seminal works in the area include those by Bertin [Bert81, Bert83] and Tufte [Tuft83,Tuft90,Tuft97].

However, modelling data using graphics in the traditional forms of graphs, charts and tables is for the most part no longer feasible when dealing with the massive data sets so common today. Developments in computer technology and their graphical capabilities provide a new medium to be explored. Computationally capable of dealing with large data sets, the use of computers to engineer graphical, external aids led to the birth of visualisation as a discipline; generally placed with the publication of the 1987 Report of the National Science Foundation's (NSF) Advisory Panel on Graphics, Image Processing, and Workstations [Def99]. From these foundations the area proceeded to develop and grow in stature, the first IEEE conference of visualisation was held in 1990. Today, visualisation research continues to grow and expand, encompassing a greater range and volume of resources.

Early visualisations had to be developed within the restrictive boundaries of the computer's graphical capabilities. However, as technology continues to become more powerful and more affordable this is becoming less of a problem and we have now reached a point where, although still limited, support for extensive, real-time graphics is readily available, opening the topic up to the masses and widening the scope of possibilities.

Nevertheless, visualisation has never been the only approach to resolving the problems associated with the increasing mass and complexity of data being produced. The need to extract data's latent information has spawned new and significant research, not only under the banner of visualisation.

3.2.2 How visualisation relates to other research areas

Some traditional methods of statistical modelling and analysis are still in place today. However, many of these methods fail due to the increasingly large data sets considered. Even with the use of computationally powerful computers, issues of scale [Eick00] are not resolved; these techniques were developed to work on much smaller data sets and do not scale well to the massive sets so common today.

The need for new methods and approaches to help resolve the problems associated with large and complex data sets, and the continued need to 'mine' the data for information, has given rise to the areas of data mining and knowledge discovery in databases (KDD). These research topics both focus on the problem of 'information overload'. However, the meaning of each term and how they relate to each other, along with their relationships to other related areas such as visualisation, is often confused in literature; definitions of these terms, how they relate to each other and how they position themselves with respect to visualisation vary.

It is not the aim of this thesis to outline all interpretations of data mining and KDD. Instead it tries only to demonstrate the variety of these interpretations, with a view to outlining their relationship with visualisation.

Data Mining

There are many definitions of the term "data mining" taken from a number of different perspectives. Behind all these definitions lies the common goal of extracting information from large and often complex data sets. Cabena defines data mining as:

> "*the process of extracting previously unknown, valid and actionable information from large databases and then using the information to make crucial business decisions*"
> *[Cab97]*

There is some debate over what Cabena calls 'previously unknown' information [Pran98]; however, by focusing on the process of data mining from a relatively high level of abstraction, this definition steers clear of most contentious issues. Contradictions are commonly found in definitions that look to define the techniques

and processes which go to make up the activity of data mining. Unwin [Unw00] recognises two opposing views of how data mining is composed:

1) Computer Science approach views data mining as methods used to automatically search data sets for "interesting information"
2) Statisticians on the other hand use the term in a more general sense to describe searching large data sets for "anything of interest"

The restrictive computer science view presented by Unwin offers no place for visualisation, instead focusing more on the development of complex algorithms and the use of Artificial Intelligence and Machine Learning techniques. If not carefully managed, these approaches can distance the human 'miner' from the processes involved: the miner may often apply complex calculations with little knowledge of their meaning or how they may help identify useful information from the data. This could lead to a lack of understanding and the possibility of false interpretations being made about the data.

Visualisation as a tool for identifying interesting things (patterns, etc) within large data sets is very much a part of the 'statistician's' definition of data mining. The use of visualisation offers the potential to keep the user of a data mining system much closer to the processes involved, keeping the user informed not only about the results but also about the processes and data involved; 'the bigger picture'. This broader view of data mining presented by Unwin is the one taken in the Clementine User Guide, where the definition emphasises the role of data mining in extracting valuable information from available data sources:

> Data mining refers to: *"using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful" - Clementine User Guide, a data mining toolkit [Dill95]*

Knowledge Discovery in Databases (KDD)

KDD is a term often used interchangeably with data mining to represent the same meaning [John97, Goeb99]. However, others see KDD as a process that encapsulates that of data mining [Brac96, Fayy96].

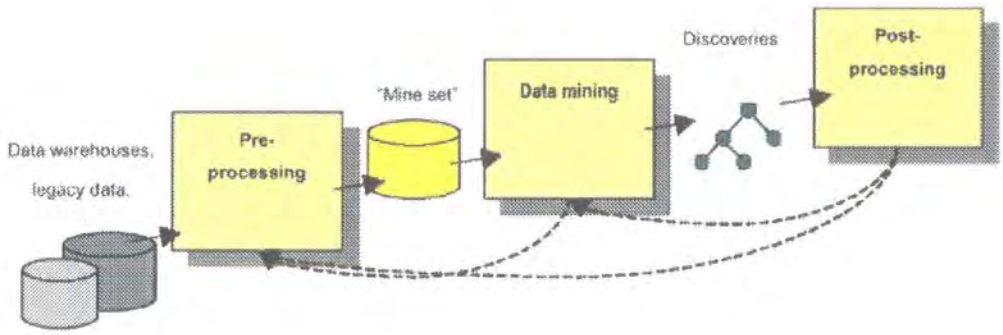


Figure 3.1 - Diagram of the KDD process [Feld98]

In this case, visualisation can be seen as a possible tool within the data mining sub-process shown in Figure 3.1. However, visualisation could also be part of the pre and post-processing stages:

- Pre-processing – visualisations could be used to identify the data (the 'mine set') the user wishes to explore.
- Post-processing – visualisations could be used to filter, sort and structure the data from the data mining processes to help present and identify useful portions of the data.

Results from both data mining and KDD are not definite but fuzzy, as with those gained through the application of visualisation. Like visualisation these processes centre on the need to derive snippets of information from the vast forests of data provided, and are heavily reliant on the users' proficiency with the tools and the techniques involved.

How research from other areas is used within visualisation

Visualisation takes forward many of the theories from Data Graphics as it looks to explore and take advantage of the flexible and powerful graphical medium afforded by the modern computer. Looking to help resolve the problem of 'information

overload', where the role of the human is key, visualisation incorporates interest from a number of different research communities, bringing together researchers and resources from computer science, user interfaces, psychology, perception and statistics.

Social sciences are proving to be an increasingly important component of visualisation research. Work continues to try and harness the potential of the human user, and studies in psychology and perception offer a unique insight into human behaviour, as well as our cognitive and perceptual abilities.

There is also a great deal of interest in visualisation from other areas of computer science research:

- The artificial intelligence community is looking closely at the mappings from data and task to visualisations, with the aim of automating the process of selecting appropriate techniques [Mack86].
- Using visualisations as interfaces to large sets of data has provoked interest from the user interface community as a possible component in a new generation of interface.
- The use of HCI and interface resources can also contribute to help visualisation development, supporting the creation of more usable and 'user friendly' tools.

Visualisation, in essence, hopes to present visual representations of data that enhance human cognitive abilities whilst taking best advantage of the human perceptual system.

3.2.3 Visual Representations

The statement: "A picture is worth a thousand words" has been the centre of much conjectured discussion in terms of both its meaning and application. This phrase could be interpreted with respect to a visualisation's ability to present large volumes of data in a small space. Many believe this saying to be an ancient Chinese proverb; however, instead of taking the term literally, the meaning it represents may be better derived by looking at its origin as traced by Mieder [Blac97]. Here the author traces the saying back to an advertising manager who uses the words to help sell

advertising space, and as Blackwell indicates, the claim is not so much about content but about effect and impact; pictures draw more attention than text.

This leads us to an interesting and complex problem posed by Petre and Blackwell:

> *"Does Visualisation Mean Pictures?" [Petr98]*

In discussing this issue, Blackwell looks to provoke thought into what should and should not be afforded to the term visualisation. Using examples he increasingly tests these beliefs with proposals involving varying amounts of graphical and textual display in combination. With specific attention to the relationships between graphical and textual programming he questions the divide between graphics and typography; a point that Blackwell investigates, in much greater depth, emphasising the role of diagrams and not visualisation (see Figure 3.2) [Blac98].

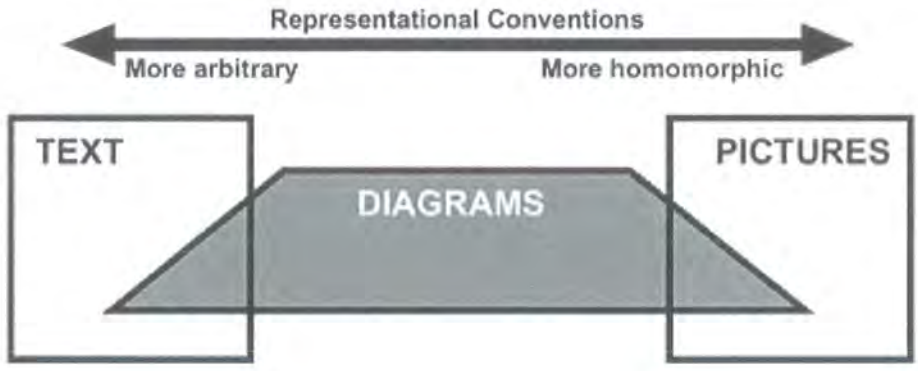


Figure 3.2 - Diagram showing relationship between text and pictures [Blac98]

In this work Blackwell questions the lines of division from text to diagrams and from diagrams to pictures, leading him and Petre to ask the question:

> *"Is there a dividing line after which we can confidently say that something is not a visualisation system?"[Petr98]*

However, instead of focusing on small details he encourages the development of a 'broad church' of research, utilising the potential offered through the use of both graphics and text.

This similarity in treatment of visualisations and diagrams presented in Blackwell's discussions is not unusual; visualisation can be seen as an extension of diagrams, taking these and other visual representations a step further, providing the possibilities of real-time interaction, and taking advantage of the computer's ability to cope with massive data sets. Thus many of the theories that now underpin developments in visualisation have their origins rooted in diagrams and research into their effectiveness.

3.2.4 Foundations of Visualisation

The process of visualising data is composed of a series of mappings, moving from the data to the visual form presented to the human perceiver:

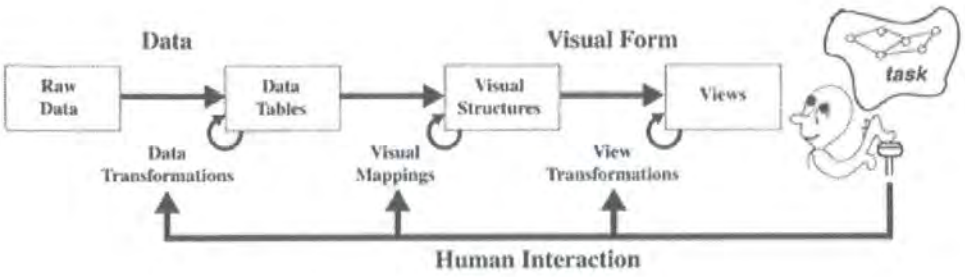


Figure 3.3 - Diagram showing the visualisation process [Card99a]

Figure 3.3 outlines this process as seen by Card et al. [Card99a], showing the potential transformations and interactions involved, and the role of human interaction in specifying these transformations. These transformations, as with other human-related activities, are prone to error, so each has the possibility of introducing false information and error into the process. The unrestricted nature of the mapping from data to visual form may provide the basis for developing many and varied visualisations, but it is also the point in the visualisation process where there is the greatest opportunity for the introduction of error. This introduction of error could be seen as the root of many failures within visualisation: Visualisations may offer the capability to obtain useful information from data, taking full advantage of human perceptual and cognitive abilities; however, it can also hide information and introduce the possibility that false information may be inferred. Tufte comments that:

"There are right and wrong ways to show data; there are displays that reveal the truth and displays that do not" [Tuft97]

Tufte uses the space shuttle Challenger incident as an example of where a visual display failed to tell the 'truth' about the data it represented [Tuft97].

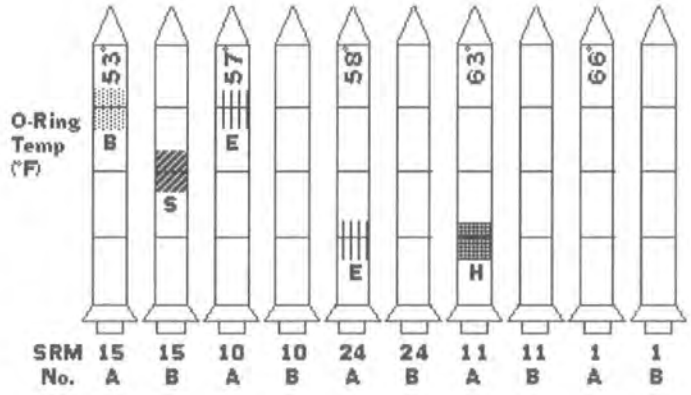


Figure 3.4 - Diagram similar to that used in the Challenger incident

In the time before 'lift-off' there was a question as to whether the shuttle launch should go ahead on the relatively cold day. This decision depended on the temperature and if it would cause the O-rings that sealed sections of the booster rockets to become unsafe. A diagram similar to that shown in Figure 3.4 was used to make the decision. This choice of diagram, although relatively elaborate in its presentation, obscured the main variables of importance: temperature and degree of damage. Temperature is shown textually and not graphically, while the degree of damage to the o-rings is presented but is not mapped onto any graphical scale. The pictures of the rockets serve very little purpose, cluttering the display and preventing the truth in the data from being identified. In his discussion of this case Tufte presents this same data in much-simplified scattergraph depicting the relationship between the two major variables. This presentation revealed the true dangers of launching on the day in question with its relatively low temperature, and demonstrates a valuable lesson: visualisations cannot simply be produced at random; they must be developed inline with available rules and guidelines as well as the goals of the visualisation.

The goal of visualisation is to use computer graphics to gain insight into large sets of data, providing support not only for the types of data - qualitative, ordinal and quantitative - but also for the user and the task they are trying to perform. Visualisations are essentially being used to communicate information, and they must do so with clarity, precision and efficiency [Tuft83]. To a finer degree, the goal of a visualisation is to reduce the costs associated with deriving the information inherent in data.

Researchers in visualisation have drawn on seminal works from the field of data graphics for guidance. Forming theories behind diagrams and their creation, these guidelines are now being used to help the development of visualisations and visualisation research.

Bertin

Bertin investigates the properties that make up all diagrams. Identifying this set of properties, he also outlines the potential of each property to support the display of specific types of data and their corresponding operations. Looking closely at how diagrams (or 'constructions' as he defines them) provide support for different questions that could be posed by a user, Bertin defines the most efficient diagrams as:

> *"those in which any question, whatever its type or level, can be answered in a single instant of perception, in a single image."[Bert83]*

Furthermore, he suggests that where it is not always possible to display a view to answer all questions, the graphic should be constructed to support the questions of preference. He comments on how the efficiency associated with each diagram is dependent on the tasks (or questions) under scrutiny, defining efficiency by the following proposition:

> *"If, in order to obtain a correct and complete answer to a given question, all other things being equal, one construction requires a shorter period of perception than another construction, we can say that it is more efficient for the question." [Bert83]*

Larkin and Simon

Although rooted in the realm of diagrams rather than visualisation, a study carried out by Larkin and Simon [Lark87] illustrates some of the reasons why visualisation can be so effective. In this study they compare the use of diagrams and textual representations for solving physics problems. Focusing on aspects of search, recognition and inference they found 3 main benefits from the use of diagrams:

1. Diagrams provide a way to group information that is to be used together, removing the need for large amounts of searching during problem solving.
2. Using location to group information about a single entity, diagrams remove much of the need to match symbolic labels.
3. Diagrams automatically support a large number of perceptual inferences

As commented by Card [Card99a], benefits one and two directly improve the costs associated with deriving information from the data (Cost-Of-Knowledge Characteristic), whereas the third reduces the cost of certain operations on the data. Visualisations should be developed to support and make use of these benefits.

Tufte

Work by Tufte [Tuft83,Tuft90,Tuft97] looks to develop notions of what general properties govern the effectiveness of a diagram, with reference to the display, the data represented, and the task at hand. Using these Tufte suggests guidelines and rules-of-thumb. Summarising some of Tufte's work and applying the described principles to visualisation, Globus [Glob94] defines 5 goals that visualisations should aim to achieve:

1. Content focus
2. Comparison rather than mere description
3. Integrity
4. High Resolution
5. Utilisation of classical designs and concepts proven by time

Tufte presents his guidelines with a view to specifying what a diagram should strive to be. Although these guides have little in the way of substantiation, they are still widely quoted and used within the field of visualisation.

Guideline Summary

The representation used should be carefully and deliberately designed, making the most of both textual and graphical features. The display should reflect a definite sense of scale with respect to the values represented and should only contain a comprehensible level of detail. Tufte also suggests that the representation should have a 'story telling' element to it, leading the user through a sequence of steps, telling a story about the underlying data. However, one of the points - if not the main point - Tufte makes is to avoid cluttering the representation with unnecessary 'chartjunk' [Tuft83]. Taking this further he introduces the concept of data-ink, defined as:

> *"proportion of the graphics ink devoted to the nonredundant display of data-information"*
>
> OR
>
> *"1.0 - proportion of a graphic that can be erased without loss of data-information."*
>
> *[Tuft83]*

Stating that one should aim to maximise the amount of ink used to represent data in a representation, remove ink not devoted to data, and minimise redundant ink used to represent the data.

These sources provide much of the foundation upon which visualisations are designed. Nevertheless, these are far from the solid laws that are present within other scientific disciplines. They only exist as broad guidelines for the development of visualisations, and they do not form any kind of solid process that could be adhered to in the generation of good visualisations. Furthermore, there is little empirical evidence supporting the use of these guidelines in the development of visualisations. Derived from investigations into diagrams they do not take into account the new possibilities afforded by the modern computer. As statistical

methods of the past struggle to cope with the volumes of data produced today, so the theories derived for diagrams may not prove effective or valid for the development of complex visualisation, especially with the use of animation and three dimensions.

These are truly rocky foundations upon which to build the 'broad church' of visualisation research. Indeed the extensive nature of visualisation could be at least partly due to this fundamental lack of guidance.

3.3 Visualisation Tools and Methods

Diagrams and other predecessors to visualisations are inherently restricted by the nature of the mediums through which they emerge. Largely based on ink and paper, representations took time and effort to produce, limiting the amount of data and detail that could be presented. In this era the use of colour provided an added gloss, but not without extra expense.

Movement from paper to computer technologies has removed many of these restrictions. Once shackled by their medium, visual representations now indulge themselves with the increased computational and graphical capabilities afforded by the computer, providing new possibilities and opportunities in terms of:

- Visual Representations
- Interaction
- Multiple Representations
- Animation
- Display Options

Nonetheless, with these new possibilities also emerge new problems.

3.3.1 Visual Representations

The new generation of computers provide the ability to rapidly reproduce complex, high-quality visual 'masterpieces'. This has opened the way to the design of increasingly elaborate visualisations. By providing greater support for the development of three-dimensional representations, this new technology has also paved way for the investigation of the properties and capabilities of these representations previously not possible. However, developers must be wary of the new and elaborate possibilities now available; their focus must remain on the data and not the representation itself.

3D Representations

The introduction of this extra spatial dimension provides a large number of opportunities that could be taken advantage of by the developer of a visualisation. Three-dimensional representations not only offer another spatial dimension - so effective in encoding all types of data - but also the opportunity to use lighting and other properties to encode data and enhance the representation. The use of three dimensions also provides a number of other benefits to the developer of visualisations:

- Providing the ability to create potentially infinite 3D worlds, giving the developer of a visualisation a lot more 'elbow room' within which to create.

- Taking advantage of human familiarity with 3D spaces and the natural form of abstraction provided through perspective.

Although the use of three dimensions is seen to have great potential for the development of effective visualisations, there is little evidence that demonstrates the extra power afforded. To this date three-dimensional visualisations have seldom been compared methodically, and little is known about the factors that make them usable. Early implementations concentrated more on the new methods of display than the data and its potentially invaluable content, focusing on the effect and impact of the display and not the insight it could provide. Studies of three-dimensional visualisations have also been far from conclusive in providing support. A study carried out by Wiss [Wiss98] evaluating three three-dimensional visualisations found that the visualisation with the greatest use of three dimensions (the Information Cube [Rek93]) to be the least effective. However, in this study it is stated that this could be due user disorientations caused by the type of or lack of global overview provided within each of the tools.

The development of three-dimensional visualisations must also look to overcome problems created by its own display mechanisms. It may provide the ability to place objects of varying size etc. around a potentially infinite space, but this brings with it the following problems:

- Occlusion

- Problems with the direct comparison of objects in the display due to the imposed perspective
- Problems with seeing objects at the limits of the perspective

Navigation and Orientation issues can be a problem with all visualisations. Within two-dimensional visualisations these problems are relatively small with only a single plane of fixed orientation to explore. The introduction of a third dimension and a potentially infinite space accentuates these issues. This problem is well surmised by Pettifer and West:

> *"losing a cursor on the desktop is one thing, losing yourself in cyberspace is quite different" [Pett97]*

Without any kind of outstanding, constant points ('landmarks') of reference in a potentially infinite three-dimensional environment it is extremely difficult to locate the user's viewing position without considering the location of any interesting patterns or trends.

Metaphor and Analogy

Use of the term metaphor is generally limited to literary circles, but it is also used within visualisation research when looking at the process of deriving understanding from visual representations.

Mapping data to visual form provides the means for metaphorical reasoning based on the graphical display; individuals derive their understanding of graphical components based on the context of the component within the display, and the application of their knowledge of similar components. This use of the term "metaphor" within visualisation, as within HCI, ignores most developed cognitive theories [Blac98]. Here, diagrams and other visual representations such as visualisation are seen as of a part of metaphorical processes that can assist learning and form the basis for problem solving.

The design of a suitable metaphor for use in a representation is central to how an individual interacts with the display. It can provide a means of shifting the cognitive load of navigation and visual interpretation to the subconscious mind, and must be

constructed in such a way as to enable the designated tasks to be completed [Knig00a]. In terms of metaphor design Apple suggests:

> *"use metaphors involving concrete, familiar ideas and make metaphors plain, so that users have a set of expectations to apply to the computer environment" [App92]*

Use of naturally occurring metaphor is common, drawing on the human ability to make inferences based on experience with the corresponding 'real world' objects. However, metaphor design must also take into account the data, the structure it imposes on that data and the resultant representation [Dieb94]. Madsen, in looking at user interfaces, presents a guide on how to choose and apply metaphor [Mad94].

The development and use of virtual reality environments represents the use of a natural metaphor, mapping components of the 'real world' into components of the graphical display. Making use of three-dimensional representations, it mimics the spatial qualities of the 'real world' and uses other familiar objects situated in recognised contexts. According to Benford et al. this use of natural metaphor in virtual reality environments can aid usability.

> *"...an attempt to exploit people's natural understanding of the physical world, including spatial factors in perception and navigation, as well as general familiarity with common spatial environments..." [Benf96]*

Virtual Reality environments have thus become popular in the development of visualisations. An example of the use of a Virtual reality environment for data representation is Software World [Knig99, Knig00b].

Figure 3.5 - A district within Software World [Knig00a]

This visualisation makes use of a city metaphor to represent data derived from Java source code.

The benefit of using metaphor in visual representations has many critics and surprisingly little support, possibly due to the difficulty of proving its worth [Blac98]. Experiments undertaken by Dutton [Dutt99] provide support for the use of metaphor, but aims to afford support elsewhere have failed [Blac98]. Criticisms of metaphors include their potential to hide the data, as well as their ability to encourage false inferences about the data and the graphical representation [Mon94]. In his thesis, 'Metaphor in diagrams" [Blac98], Blackwell comments on what in his opinion is the overrated nature of metaphor, believing that the advantages of visual representation are all too often misattributed to the use of metaphor. He instead suggests that the great potential offered by visual representations is linked to the support for direct manipulation.

3.3.2 Interaction

Movement from paper-based to computer technology opened the way for the development of highly interactive representations, a new avenue for investigation and exploration.

Blackwell identifies interaction as the key component to the success of visualisation [Blac98]. Interactions increase the users' intimacy, not only with the display, but also with the underlying data, providing the ability to gain greater insight and understanding. Appearing in various guises, it is commonly assumed that visualisation systems always include rudimentary interaction mechanisms such as zoom. However, more complex interaction mechanisms have developed that play

key roles in visualisations and their support for different tasks. The ability to interact with representations that rapidly update their display based on this user input fundamentally changes the process of understanding the data presented, and make it possible to explore more potential scenarios within the data [Card99a].

In terms of designing interactions for visualisation, Shneidermans' 'Principles of Direct Manipulation' [Shne83] provide a good starting point.

> *"By presenting information visually and allowing dynamic user control through direct manipulation principles, it is possible to traverse large information spaces and facilitate comprehension with reduced anxiety." [Ahl94a]*

Example - Dynamic Queries [Shne97]

Not to be confused with the application of a structured query language at runtime, Dynamic Queries are an interactive technique used to rapidly query and filter data. The key to this mechanism is its support for browsing data sets through the use of rapid, incremental, and reversible queries.

Implementations of Dynamic Queries include the well-documented examples FilmFinder (see Figure 3.6) [Ahl94a] and HomeFinder (see Figure 3.7)[Will92].

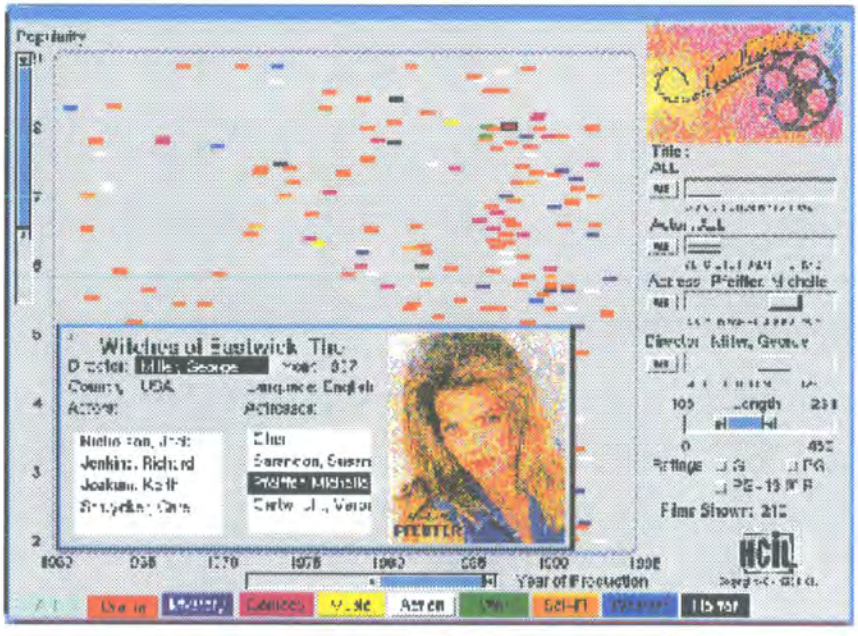


Figure 3.6 - Screen Shot of FilmFinder

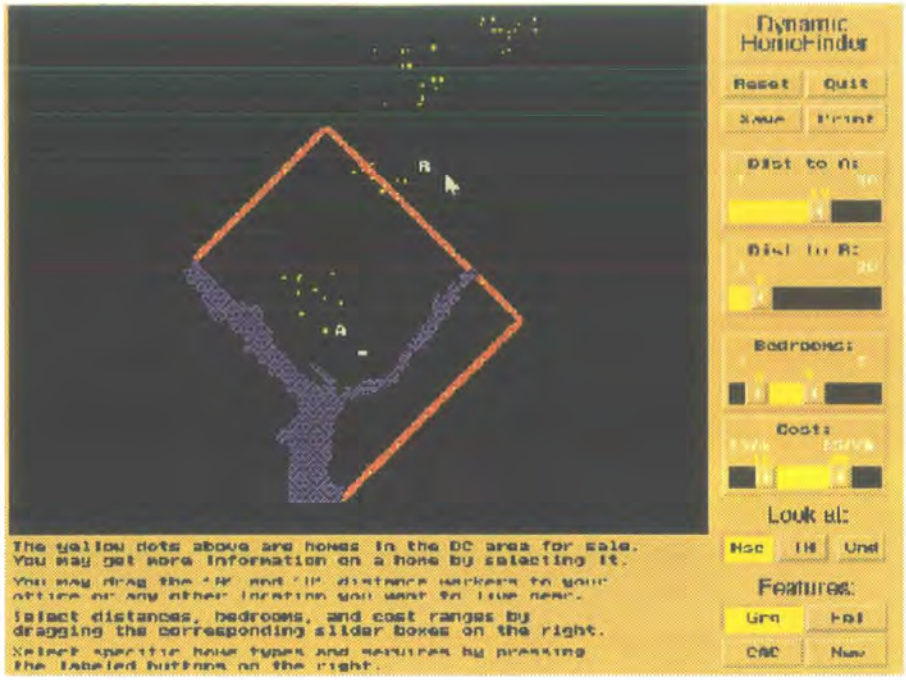


Figure 3.7 - Screen Shot of HomeFinder

In empirical studies implementations of Dynamic Queries have shown benefits in performance, but also user satisfaction [Shne94]. The technique supports the identification of underlying patterns and trends in data by providing the means to quickly and simply filter the data based on constructed queries. A greater understanding of the overall data set can be gained *by "flying through"* [Ahl94a] the data using the responsive controls. As said by Shneiderman this technique *"lets users rapidly and even playfully explore"* [Shne94].

Use of this method offers the potential for improved speed and usability; however, it is not without its restrictions. Current implementations of the concepts such as FilmFinder allow only very simple queries to be constructed. Controls used to construct these queries and manipulate the visualisation, whether present within the representation or not, can over complicate the display and reduce the space available for showing data. Implementations of these controls also do not provide adequate support for all possible data types: nominal types often require an ordering to be imposed on them to enable the use of a slider control [Ahl94a, Ahl94b].

Researchers have already started to look at resolving these issues. The development of Magic Lens filters [Fish95] supports the creation of more complex queries whilst preserving the direct manipulation properties of Dynamic Queries. Eick [Eick94] has looked at making more effective use of the area displaying the

control mechanisms, especially those in the form of sliders, by encoding further data into the sliders themselves.

3.3.3 Multiple Representations

Another method opened up to a greater extent by technology is the capability to provide more than a single representation, displayed either on the same or separate devices; displaying more than a single representation of the data in this way has both its critics and its supporters.

The use of more than a single representation has the potential to support different user tasks and to identify patterns that would otherwise be missed. This may help improve the overall system, but it also introduces the need for increased cognitive processing on behalf of the user:

> *"Visual efficiency is inversely proportional to the number of images necessary for the perception of the data" [Bert83]*

Methods have been developed to help reduce this increased cognitive load by providing tight coupling and common points of reference between the representations, for example:

- Brushing [Henz98, Beck88] – When multiple representations contain the same objects but they are portrayed differently, highlighting an object in one will automatically highlight the corresponding object in others.

Visualisation tools often make use of two displays: one to support the users' need to gain an overview of the data set and the other to provide access to greater detail. Each of these views has associated benefits [Card99a] and tools with separate overview and detail displays are commonplace. However, attempting to reduce the cognitive load involved when dealing with two separate representations, research has developed techniques to represent both in a single visual image. Based on the idea that an individual's interest in detail wanes away systematically as you move further from the focus of attention, Furnas [Furn91] proposed that the space within the representation might be proportioned to the amount of attention from the user, thus

supporting the need for detail and overview in a single 'attention-warped', 'fish-eye' view.

3.3.4 Animation

Desktop computers now have the power to display and move thousands of polygons in a single second. Animation of graphical scenes, automated and at real-time, provide a potentially invaluable tool to the developer of visualisations.

This 'fourth' dimension is a very different property compared to the spatial dimensions of physical space. It offers the means to move away from static representations sufficient only for analysing final results, to more dynamic representations that support the process of finding and understanding patterns and anomalies in the underlying data sets.

A visualisation may use animation to encode properties within a dataset, but animation effects are used to a much greater extent in supporting user interaction. The representation must update based on: users' movement around the display (e.g. change in orientation within a three-dimensional environment), interaction with dynamic parts of the display, and in response to the use of functionality provided such as zoom.

> *"To make 3D work, you need to make it move"*
> *- James Clarke, founder of Silicon Graphics Inc. [Wri95]*

The development and extent of animation has always been linked to the restrictions imposed by technology. These restrictions still remain, but only when considering animating scenes of great complexity. However, ongoing work on new rendering and animation algorithms looks set to reduce these restrictions further.

3.3.5 Display Options

New display technology means that we are no longer restricted to the use of CRT and LCD monitors. Due to the counter-intuitive nature of displaying three-dimensional images in two dimensions on these 'standard' monitors, most work has taken place in the development of displays capable of communicating these images in a more discerning manner.

CAVE environments [Cruz93], Head mounted displays and more recently the development of stereographic LCD displays offer new opportunities to investigate the use of three-dimensional visualisation techniques. Used in combination with naturally intuitive input mechanisms, these displays provide a real opportunity to test the power of the representations and not the difficulty associated with interacting in a three-dimensional environment. However, there is little evidence that demonstrates the extra power afforded by these displays and the technology remains relatively expensive.

3.4 Visualisation Research

The growing problem of 'information overload' has gained much publicity in recent times. More and more people are looking to take advantage of the potentially invaluable information latent within the reams of data produced each day. This in turn has put research topics associated with 'information overload' firmly into the public spotlight, and visualisation is no exception.

With increased attention and an associated increase in funding, visualisation has developed rapidly over the past 15 years. Initially founded around the term 'visualisation in the scientific community' [Def99], with an innate focus on the visualisation of scientific data, the topic has developed into a discipline addressing a much wider range of sources: scientific data, financial data, software, etc. This expansion of focus has led to a sub-division of the research undertaken under the banner of 'visualisation'. The most prominent of these areas defined in literature include:

- Scientific visualisation
- Information visualisation
- Software visualisation

Scientific visualisation is defined as the:

> *"Use of interactive visual representations of scientific data, typically physically based, to amplify cognition" [Card99a]*

When creating visualisations for physically based data the developer of the visualisation can take advantage of the physical properties inherent in the data.

Information visualisation, on the other hand is the:

> *"Use of interactive visual representations of abstract, nonphysically based data to amplify cognition" [Card99a]*

The abstract, non-physical nature of the data under scrutiny within information visualisation does not provide any such clues for the design of the visualisation. The use of the term 'information' within 'information visualisation' should not be compared in any great detail to the definitions provided in chapter 2. 'Information visualisation' may be about the presentation of data, but it is hoped that the visualised form of the data will provide the possibility for information to be derived.

An area of great interest, 'Software visualisation' [Knig99, Ball96] looks to help understand and communicate the features and properties of software systems, to aid comprehension and development. Software visualisation can be viewed as the application of information visualisation techniques to software, as the data collected from all areas of a systems development - code, documentation, user studies, etc. - is abstract and hence has no associated physical structure.

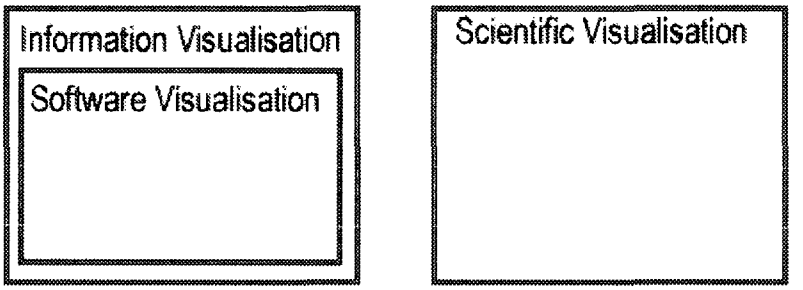


Figure 3.8 - Diagram showing the relationship between Scientific, Information, and Software visualisation

Figure 3.8 and the explanations of the relationships between scientific, information, and software visualisation represents a much-simplified view of the terms and associated research areas. In reality the relationships are much more complex and have, in the past, proved a worthy topic for discussion.

3.4.1 Information Visualisation

Information visualisation research developed around the aim of tackling the vast quantities of abstract data being produced. Created in large quantities, from many sources, including financial markets and software development, this data is a resource with great potential that must not be ignored. This potential has given information visualisation research greater impetus, developing and growing in stature to help quell the need for superior insight into the hoards of abstract data being produced.

Physically founded data offers a direct guide to the development of a visualisation by mapping its physical dimensions directly onto the corresponding properties of the display. Abstract data, however, has no such inherent qualities and thus introduces new challenges to visualisation concerning how to map the variables present onto the different visual properties available for encoding [Bert83].

This lack of guidance can be seen as being at least partly responsible for the considerable development and diversity within visualisation research. Novel approaches in the use of the available graphical properties and the use of metaphor may not have been developed without this complete lack of guidance.

3.4.2 Multi-dimensional Visualisation

The visualisation of abstract multivariate data forms the core of the problems approached by the work in this thesis. The majority of data sets produced and stored are not only massive in size but consist of many different data variables. These large 'multivariate' data sets pose new questions and challenges to the developers of visualisations.

Data sets with three or fewer variables can be directly encoded onto the three spatial dimensions. However, an image has only three dimensions; this is an 'impassable barrier' [Bert81], so when the number of variables exceeds three, the developer of the visualisation is faced with a new problem.

The problem faced is: how do you represent n variables of data graphically where n>3?

Other graphical properties such as Bertins' identified list of retinal properties [Bert83] offer one potential stopgap solution. They can be used to encode variables

in combination with the spatial dimensions; however, these are limited in number and limited in their effective support for different types of data, so they do not provide a definitive solution.

In the form of 'permutation matrices' [Bert81] Bertin proposes a more general solution to this problem. This method enables data with n variables to be displayed in a consistent format, but the method is not without weaknesses: requiring the user to permute through a matrix of representations in order to gain a view of the data set as a whole. A problem exaggerated as the data sets contain more variables.

The capabilities of the spatial dimensions to encode all types of data effectively has encouraged investigation into how they might be used to encode more data and more variables, hence helping to resolve the problem of visualising data with n variables. Several techniques have been developed from this standpoint [Card99a]:

- Composition
- Alignment
- Folding
- Recursion
- Overloading

Novel techniques for the display of large multivariate data sets have developed incorporating these principles, including well documented examples such as Inserbergs' 'Parallel Coordinates' [Ins90] and Feiner's 'World within Worlds' [Fein90] (also see section 3.4.3).

3.4.3 Current State of Research

This section aims to provide a review of information visualisation research by reviewing a number of developed tools and techniques with an innate focus on visualisations capable of dealing with large sets of multivariate data. Making use of a taxonomy, the review aims to emphasise the variety of work present in the research area. Many different taxonomies have already been developed to demonstrate the extent of the work undertaken. These have been developed from all major perspectives within visualisation: task [Shne96], data [Shne96], and visual representation [Keim96, Keim97a].

This review uses a taxonomy devised by Kiem [Keim96, Keim97a], based largely on the visual representation of the techniques. The hope is that this will demonstrate

the vast diversity in terms of display and approach. Included in this review are some of the more prominent and groundbreaking techniques, accompanied by a few of the novel approaches discovered during this research. It must be noted that this in no way represents a complete review of the tools and techniques developed in visualisation research; the aim of this is to provide a flavour of the area as a whole. A more complete survey of visualisation methods can be found here [ATKO97].

Introduction to approach

Kiems taxonomy [Keim96, Keim97a] categories visualisations into the following groups:

- Geometric Techniques
- Icon-based Techniques
- Pixel Orientated Techniques
- Hierarchical Techniques
- Graph-based Techniques
- Hybrid Techniques

This review will provide at least two examples within each section, except for the graph-based techniques and hybrid techniques sections where a brief explanation will be provided.

Geometric Techniques

Scatterplot Matrices

This technique [And72, Clev93] harnesses the power of the classical scatterplot to support the analysis and display of two variables at a time. To support multivariate data with n variables, a matrix of size n*n is produced and within this matrix n*(n-1) scatterplots covering all possible combinations of these variables are presented.

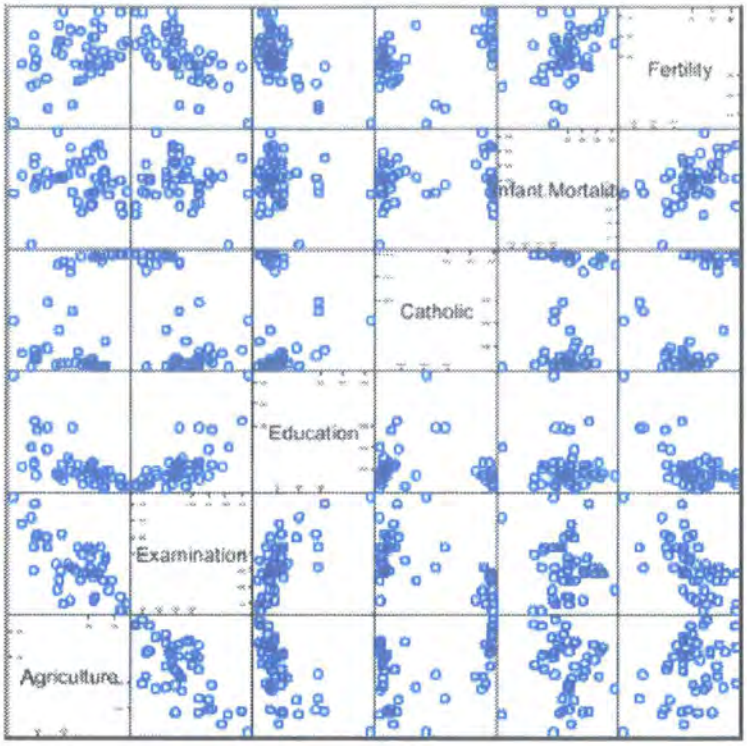


Figure 3.9 - Scatterplot Matrix [Ven00]

As shown in Figure 3.9 each variable defines a single row and a single column, with the corresponding variable names given in the redundant diagonal cells.

This visualisation offers the ability to quickly identify trends between pairs of dimensions in the data. However, in order to gain an overall perspective of the data, the user must permute through all combinations of plots and as the number of variables increases this task becomes more cognitively intensive.

Parallel Coordinates

Developed by Inselberg [Ins90], 'Parallel Coordinates' is a two-dimensional technique that works for n variables. This technique is composed of n equidistant, vertical axes representing each of the n variables in the data set under scrutiny, the scale of which is calculated based on the type and values associated with the corresponding variable in the data set. Each data item in the data set is presented as a polygon line in the display that intersects the n axes at points corresponding to the data item's value for each of the variables.

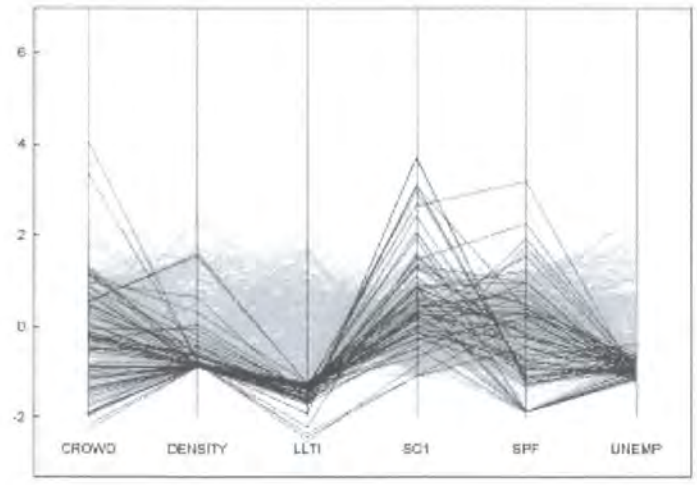


Figure 3.10 - Parallel Coordinates Display [Brun98]

Parallel Coordinates has a low representational complexity and is relatively simple to use and understand. In a single display trends can be identified along with other patterns in the underlying multivariate data set.

However, as the number of data items and variables in a data set increase, the Parallel Coordinates visualisation becomes less effective, becoming cluttered and more difficult to understand. The uniform treatment of all variables forces the technique to introduce false information into the display: an ordering is imposed on qualitative data to enable a scale for the corresponding axis to be created. Some implementations provide support for Dynamic Queries [Siir00]. This helps the display cope with larger sets of data; however the ordering imposed on qualitative data can prove more problematic when constructing the queries, as adjacent points have no real relationship.

Icon-based Techniques

Chernoff Faces

Invented by Chernoff in 1973 [Cher73] for the representation of multivariate data, this novel technique uses the features of the human face to represent the variables of a data set. Each facial feature represents a variable, and the corresponding data values determine its shape and size based on some predetermined scale. A display will be composed of a number of faces, one for each of the data items in the

underlying data set, each presented in a close grouping for direct comparison (see Figure 3.11).

Figure 3.11 - Example of Chernoff Faces

Although presented as a group of separate images, the interesting nature of Chernoff faces can overcome some of the problems associated with digesting data from multiple representations [ATKO97]. Nonetheless, the main purpose of the method is to take full advantage of the human ability to recognise small differences between faces.

The power of this method is limited to the possible number of facial features that could be used to encode the data. The method is also open to the subjective interpretation of the faces displayed, and perceived similarity between faces can very much depend on the assignment of variables to particular facial features. However, the introduction of animation to these plots could advance this method further.

DriftWeed

A visualisation developed by Rose and Wong [Rose00] for the presentation of large sets of multivariate data. Each glyph presented in the display denotes a data item from the underlying data set. The glyphs are composed of n segments, representing each of the variables stored, and the horizontal and vertical dimensions of these segments are dependent on corresponding data values. To retain continuity in presentation each of the segments is lined up point-to-point as shown in Figure 3.12.

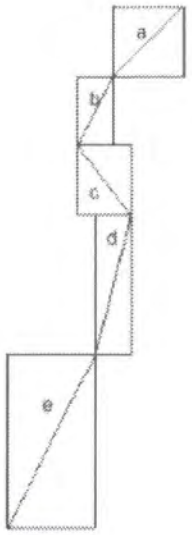


Figure 3.12 - Construction of DriftWeed feature [Rose00]

Figure 3.12 shows an example in which each segment is represented by a line. To avoid undesirable drifting in the glyphs, a later implementation of DriftWeed used an image of a double back line. The set of glyphs are positioned along the x-axis of the representation, but this visualisation proposes stacking several glyphs at a single horizontal location to enable the display to present much larger data sets.

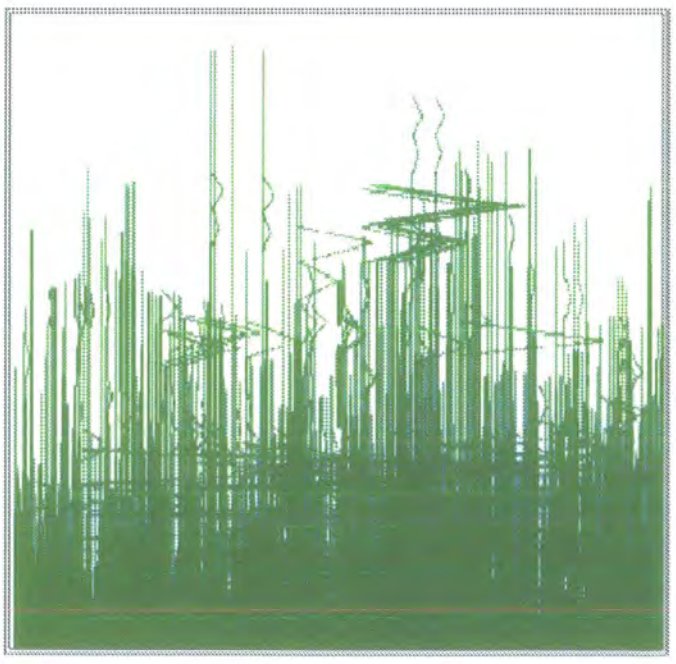


Figure 3.13 - DriftWeed Display [Rose00]

DriftWeed hopes to provide a tool that can help the user quickly and effectively identify trends, patterns and outliers in the data set presented. With a densely populated plot, changes in textures can be identified with patterns in the data and outliers as points in the plot that stand out from the general trend.

However, in displaying large data sets, masses of lines appear that can prevent the user from gaining a more detailed view of the data; individual lines may prove hard to identify amongst a mass of potentially interweaving lines, and certain lines could be at least partly occluded from view. The interactive ability to increase the visible scale of each of the measures hopes to provide at least a partial solution to this problem. Future extensions could also involve the use of colour or transparency to help draw greater attention to these salient features. However, up to this point in time, little work has been done to evaluate tools implementing the DriftWeed, so it is hard to comment upon the usability and effectiveness of this technique.

Pixel Orientated Techniques

Circle Segments

'Circle Segments' is a pixel-per-value technique devised by Ankerst, Keim and Kriegel [Ank96]. To present n variables of data the representation uses a circle partitioned into n segments. Within each segment data values belonging to that variable are arranged from the centre towards the outside, as shown in Figure 3.14. The order in which the data values are entered into the segment can be based on another property of the data, but this ordering must be consistent within each of the other segments displayed. Each of the pixels is assigned a colour based on the value they correspond to, using hue to represent the sense of scale for each variable (see Figure 3.15).

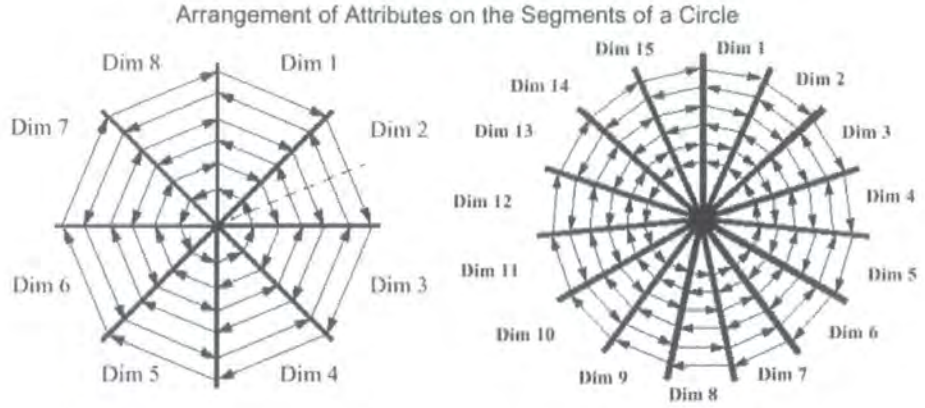


Figure 3.14 - Circle Segments [ATKO97]

This representation has the ability to visualise large sets of multivariate data and is only really restricted by the number of pixels available on the display device. By comparing the location and colour of the pixels within different segments of the circle, potential patterns and trends in the underlying data set can be easily identified. However, identifying these features is greatly dependent on the users' ability to compare and contrast the pixel colour and location within all of the segments, and as the data sets presented grow in size and number of variables, this task becomes increasingly complex and cognitively intensive. In an attempt to resolve some of these issues implementations provide tools with the ability to move and reorder the segments, making it easier to compare selected segments directly.

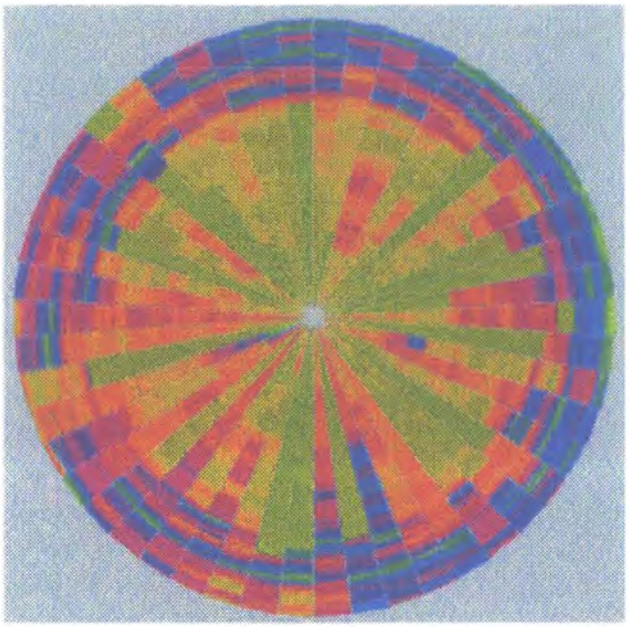


Figure 3.15 - Circle Segment Display [Ank96]

VisDB

This visualisation [Keim94] differs from the other techniques described in that it does not display the underlying data set directly; it instead presents 'relevance factors' for the data calculated from the results of querying the data set. Using a constructed query, relevance factors for each data item are calculated. Arranging data using the query as a focal point, each data item is represented as a pixel organised around this point in a rectangular spiral (see Figure 3.16). The ordering of this arrangement is dependent on the data item's corresponding relevance value, and this value is also used to assign a colour to the pixel.

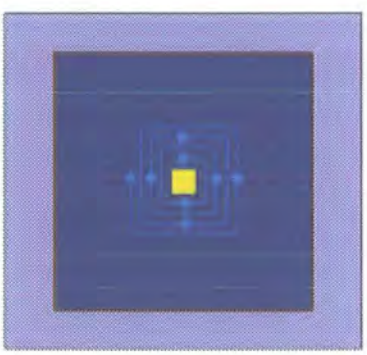


Figure 3.16 - Rectangular Spiral Layout [Keim94]

This display of the results provides access to a very limited amount of information. Its use of colour and location to represent the relevance factor limits the patterns and trends that can be identified, revealing only an insight into how the items in the database relate to the given query. However, this display can be extended by displaying one of these views for each of the variables contributing to the query (see Figure 3.17). Within these views the positions of the data items remains the same as in the original, but the colours used are based on the distance of the data values for the variable from its value at the focal point of the query.

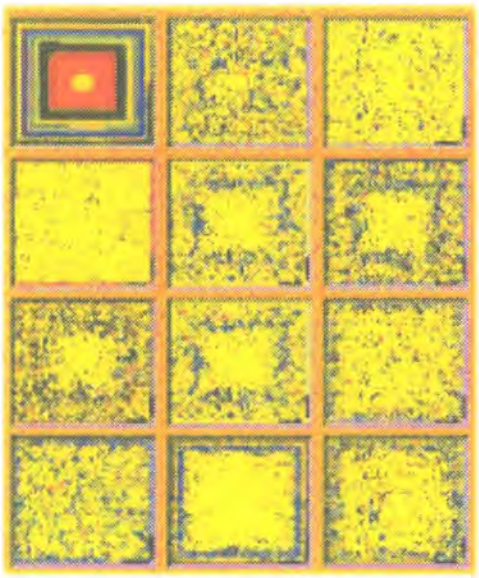


Figure 3.17 - VisDB [ATKO97]

This extended view provides the potential to compare and contrast two or more of the variables involved in the query. From this correlations may be inferred with respect to the query. However, the effectiveness of this visualisation can be questioned for several reasons: Its results are highly dependent on the formulation of the query and not solely on the data, and also the distance calculations, so fundamental to the layout of the display, cannot cope with non-quantitative data.

Hierarchical Techniques

World within Worlds

World within Worlds presents multivariate data sets by continually partitioning the space available until all variables are represented. Developed by Feiner and Beshers

[Fein90], this visualisation technique aims to utilise the three spatial dimensions to much greater effect.

The technique represents an initial three-dimensional coordinate system with x, y, and z-axes; this coordinate system is then overloaded by a second nested within the original, and this process continues until all the variables have been assigned to an axis. Once the final three-dimensional coordinate system has been specified, a surface plot showing the values for the associated variables is displayed (see Figure 3.18); from this display the user can investigate the relationships between these properties in terms of the underlying data set.

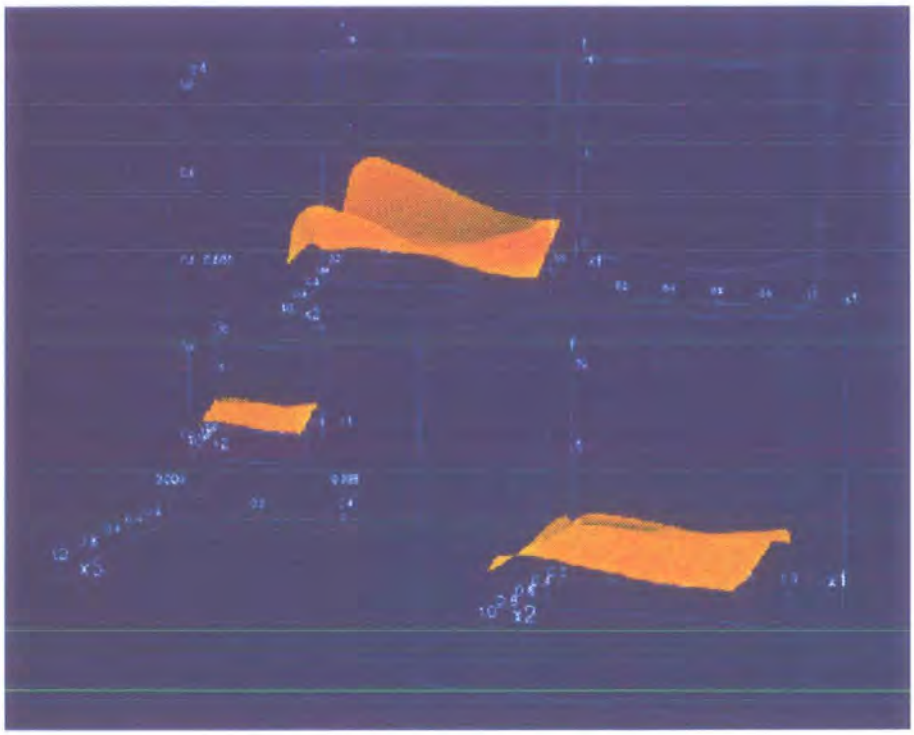


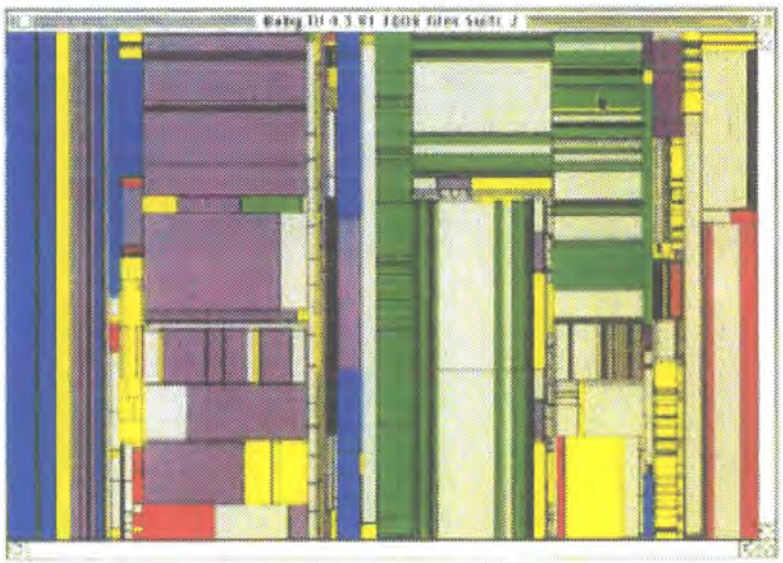
Figure 3.18 - World Within Worlds Display [112]

The technique involves a high proportion of user interaction, not only in controlling the position and orientation of the nested three-dimensional spaces, but also in selecting the variables to be displayed at each level. These different possible mappings can require the user to permute through all possible views in search of any trends or patterns in the data; developments such as AutoVisual [Besh93] are look to resolve these issues. Nonetheless, the lack of any general overview of the data means that users are required to have a definite idea of what exactly it is they are looking for.

Treemap

Shneiderman developed the Treemap visualisation in the early 1990's to display the hierarchical structure of a hard disk in order to help investigate the composition of the disk in terms of the files stored by different users.

The Treemap visualisation presents hierarchical data in two-dimensions by iteratively splitting the screen into rectangles, alternating in horizontal and vertical directions using a two-dimensional space-filling algorithm (see Figure 3.19). Each of the rectangles represents a node in the hierarchical data, the scale and colour of which is determined by a selected property of the data. In the case of the hard disk example, each rectangle represents a file, the scale of which is proportional to the size of the file and the colour is used to indicate the user the file belongs to.



treemap of a file system containing about 1000 files

Figure 3.19 - Treemap [Keim97b]

Many tools have been developed using the Treemap visualisation to represent disk space, including a commercial tool called DiscMapper (Micro Logic Corp Incorporated). However, the potential of this technique is not limited to representations of disk space; the tool has also been used to visualise data sets from Basketball and finance [Shne98].

When using the Treemap visualisation it can be quite difficult to compare the size of the different rectangles because of the different orientations used and the potential spread between the shapes you are trying to compare. Also, from the developer's point of view, this technique when densely populated is very difficult to label.

However, this technique is being extended all the time: implementations have been developed introducing Dynamic Queries and a three-dimensional display [Shne98].

Graph-based Techniques

Developing methods for the optimisation of graph-based displays, this particular category defined by Keim [Keim96, Keim97a] falls outside the scope of work within this thesis. Topics under investigation include: Graph layout optimisation for aesthetic purposes and development of three-dimensional graph drawings.

Hybrid Techniques

As the name suggests, this technique involves the use of two or more methods in combination. The hope is that by incorporating more than one method of viewing the data, the effectiveness of the visualisation will be enhanced.

Tools that incorporate hybrid techniques include:

- Starfield displays [Ahl95a, Ahl95b]
- XmDv [Ward94]

3.5 Conclusions

Visualisation is a developing research area and has yet to reach a state of maturity. Unlike areas such as engineering and physics, now established on solid foundations proven both literally and mathematically, visualisation still remains reliant on sets of guidelines and so called rules-of-thumb. Although this lack of guidance could be seen as the inspiration behind the wide range of visualisations currently produced, it is also the major stumbling block to progress in visualisation research.

Evaluation has a vital role to play in developing and clarifying what a good visualisation should be, and thus in the development of well-founded guidelines for visualisation design. However, evaluation is a second-class citizen in the realm of visualisation research. All too often evaluations are an afterthought and little is put into their design; without due consideration visualisation development will continue to be 'hit and miss', more of an art than a science.

The need to develop increased support for visualisation design is well recognised, but visualisation research should also look to provide improved support for:

- Group working

- Missing Data
- Relationships between data values

The majority of visualisation research tends to focus on supporting individual users in discovering information. Visualisation tools will regularly provide the user with multiple views of data in the hope that more trends and patterns may become visible, but they ignore the number of different viewpoints that could be gained from introducing more people to the discovery process.

Using Tuftes' guidelines, a visualisation should *"above all show the data"* [Tuft83]. However, the majority of visualisation techniques fail to do so, providing little or no support for displaying missing or corrupt data. All too often this type of data is removed or ignored during the visualisation process. Missing or corrupt data can provide potentially invaluable information about the data set and possible problems with its method of collection; information could easily be discovered from patterns and trends appearing in a visualisation.

The use of a multivariate visualisation can help the user identify potential relationships between the different data variables in the data set presented. However, it may be the case that relationships already exist between the data variables, and that details of these relationships are available within the data set. By giving users access to details of these relationships within a visualisation or otherwise, the insight they gain into the data set may alter greatly.

The main issue for visualisation past, present and future is the pivotal role played by humans in the process. Humans are innately different: in the way we perceive things and the way our minds process sensory input. Visualisations will always be a subjective discipline; the many unknowns human participation imposes on the topic make it very difficult to be entirely scientific. Use of metaphor and other conventions used in visualisation may benefit some users but not others.

Visualisation research must look to address the current lack of evaluation effort. Hatch, Smith and Taylor [Hatch01] provide a summary of how evaluation could and possibly should be carried out for software visualisation, and many of these ideas can be taken forward to the more general area of visualisation. The result of such extensive evaluation work would provide a vital feedback mechanism in the

development of new and improved visualisations and design processes. More effort should also be made in developing visualisations with support for group working, representation of missing or corrupt data and incorporating possible relationships between the different data variables in a multivariate data set.

However, the unpredictable nature of humans and the key role we have to play in visualisation means that it may never be an exact discipline.

A question first posed to me in a first year software engineering lecture asked: *"Are software engineers scientists, engineers or artists?"*

This question seems even more relevant here, for visualisation: *"Are developers of visualisations scientists, engineers or artists?"*

In my view the current unguided nature of visualisation development is very much an art. It must be the aim of future work to move visualisation from being an art form into something substantially more scientific and predictable.

Chapter 4 Extending Parallel Coordinates

"Research the literature, find good information designs, and steal them. Remember that talent imitates, genius steals."

– Al Globus

4.1 Introduction

The aim of this chapter is to provide implementation details of a prototype visualisation tool produced as part of this research. Developed to display large sets of multivariate data, the tool hopes to provide the user with the means to interact and explore these data sets, identifying potential patterns and trends.

Previous chapters in this thesis provide background to the problem approached by the implementation presented in this chapter:

- Chapter 1 introduces the more general problem of information overload.
- Chapter 2 introduces visualisation as a possible mechanism for approaching the problem of 'information overload', and focuses on the more specialised problems associated with visualising large sets of multivariate data.

This chapter reiterates the problems that the prototype's implementation looks to address. It then considers the approach taken in developing the visualisation tool, giving details of recognized problems that it attempts to resolve. Based on the results and experience of the implementation, this chapter also includes aspects of possible future work, focusing on some of the problems discovered during the tool's development.

4.2 The Problem

Chapters 1 and 2 provide a gradual introduction to the problems approached by the tool's implementation, including details of related work whose goal it is to solve these and similar problems. Visualising large sets of multivariate data is the problem finally presented and discussed in Chapter 2, and this is the problem considered by this chapter.

The aim of this implementation is to take full advantage of the interactive, graphical medium afforded by the modern computer to produce a visualisation capable of

representing large sets of multivariate data. The prototype tool must also be capable of supporting the interactive exploration of multivariate data sets, helping to identify possible patterns, trends and anomalies in the data, and the means to filter data based on user requirements, providing a means of focusing attention onto data values of specific interest.

4.3 The Approach

The development process for visualisations has little support. No rules or strategies exist that guarantee the effectiveness of a visualisation. Work by the likes of Bertin [Bert83] and Tufte [Tuft83,Tuft90,Tuft97] offer the only semblance of development guidelines, and these were not produced with visualisation in mind. The uncertainty associated with the development of new visualisations prompted Al Globus to say:

> *"Research the literature, find good information designs, and steal them. Remember that talent imitates, genius steals." [Glob94]*

Taking into account the 'hit and miss' nature of visualisation development, the tool's implementation is based on two well-founded and empirically tested techniques:

1. Parallel Coordinates [Ins90]
2. And Dynamic Queries [Shne94]

Note: The following terms are used to describe different features of a data set:

- Data Item – The items whose properties are measured and stored in the data set
- Data Variable – The measured properties of each data item stored in the data set
- Data Value – The measured values for each data item's data variable

	Data Variable (1)	Data Variable (2)	Data Variable (3)
Data Item (1)	Data Value	Data Value	Data Value
Data Item (2)	Data Value	Data Value	Data Value

Table 1 - Table showing the relationship between data items, data variables and data values

If you consider the data set as a table then it would be composed, as shown in Table 1.

4.3.1 Parallel Coordinates

Not a new technique, the concept of 'Parallel Coordinates' has been around for sometime, but its simplicity suggests that it should have been around a lot longer. Many instantiations of this technique have been developed over the years and they are still being produced today.

'Parallel Coordinates' supports the display of multivariate data in a relatively simple and consistent manner that can be easily interpreted by the viewer. Built up of a series of parallel axes, the data values are presented by polygon lines passing over each of these axes. By comparing the different polygon lines and the paths they take, patterns, trends and anomalies in the underlying data set can be identified.

The simple yet effective make-up of Parallel Coordinates means that instead of the tool's implementation focusing on the development of complex display algorithms, effort could be centred on devising new features not present in either of the two base techniques. This basic construction of Parallel Coordinates has remained constant over many years; however, more recently work has been done integrating Dynamic Query mechanisms into the display [Siir00].

4.3.2 Dynamic Queries

Shneiderman and Ahlberg [Ahl94a] reports that both performance and user satisfaction can be greatly improved through the use of Dynamic Queries. By providing a quick and intuitive method for querying and filtering data, they enable users to explore the displayed data sets, helping to identify patterns and trends that otherwise might not be possible. Dynamic Queries map selected variables from a data set onto visual controls, providing the means for users to interactively explore the data displayed. These controls support the introduction of user-defined constraints that filter the data represented in the display, providing the ability to focus on data of specific interest.

Dynamic Query mechanisms have already been incorporated into implementations of Parallel Coordinates. Use the dynamic filtering mechanisms enables this synergy to deal more effectively with much larger numbers of data items.

4.3.3 Points to Resolve

The tool's implementation looks to resolve some of the weaknesses identified in the two base techniques and current faults within visualisation research as a whole.

Problems with the Techniques Employed

'Parallel Coordinates' treats all data no matter what its type in exactly the same way. This degree of consistency is one of its great strengths; however, it is also one of its weaknesses. Although effective in dealing with quantitative data and to a lesser degree ordinal data, the support for qualitative data types is somewhat lacking in the 'Parallel Coordinates' technique. To display data of this type an ordering has to be imposed on the values, allowing a scale for the corresponding axis to be generated. This action introduces false information into the display, not only with respect to the position of the points on the axis (e.g. high=good or larger), but also in terms of the relative position of the values (e.g. closer=similar).

Dynamic Queries provide qualities such as its simplicity and ease of use. However, in its simplicity this technique only supports the construction of relatively simple queries; queries that are constructed through the manipulation of controls provided within the tool. The great strength of Dynamic Queries is the interactivity afforded by these controls, and hence the effectiveness of this technique is largely dependent upon these. Sliders are the main control mechanism employed by Dynamic Query tools and they provide an intuitive, powerful means of querying the data sets represented. Nevertheless, the use of sliders as query controls for qualitative variables is far from intuitive. Data of this type contains no ordering relationships between its values, so the use of a slider to select ranges of data does not map well to the properties of this type. Toggle buttons have also been employed as controls within Dynamic Query tools. These buttons allow the user to select individual data values, an ability that maps well to the properties of qualitative data, but the use of a button to represent all possible data values for a qualitative variable is only feasible when the number of possible values is small.

An implementation of Parallel Coordinates by Siirtola [Siir00] addresses at least one of these points, providing support for the development slightly more complex queries; nevertheless, the constructions allowed are still relatively simple and different types of data are for the most part ignored. Combining Dynamic Query

mechanisms into a Parallel Coordinates display this implementation also does nothing to resolve their joint problems in dealing with qualitative data.

Problems from Visualisation Research

The majority of visualisation techniques fail to take into account the possibility of missing or corrupt data values; in most cases they are simply ignored or transformed into a form accepted by the tools developed.

Multivariate visualisations can help users identify possible relationships between the different variables in a presented data set. However, it may be that these variables already have some kind of defined relationship between each other. This kind of information could be available in the metadata of a data set, and even if in the majority of cases this kind of information is not present, in the cases where it is, visualisation support should be available.

4.3.3.1 Summary

The prototype's development is based on two well-founded techniques: Parallel Coordinates and Dynamic Queries. It aims not only to support users in gaining an insight into the displayed data set, but also to address some of the weaknesses of these two techniques and within visualisation research as a whole.

4.4 Tool Detail

4.4.1 General introduction

The implementation has been developed in an object-oriented fashion and is written in Java, making extensive use of the swing and Graphics2D packages. The final tool will run and compile using JDK1.2 or a later release.

This section details the developed prototype's implementation. An outline of the tool's parallel coordinate display is provided, including details of extensions made to the technique. It then describes how Dynamic Queries have been incorporated into the visualisation, with details of how queries are constructed and organised. The section also includes details of the prototype tool's support for the selection and comparison of data items and outlines the role of the tool's table and Starplot displays in these processes. Finally the section provides a brief description of the tool's remaining functionality along with an overview of the file format used.

4.4.2 The Parallel Coordinates Display

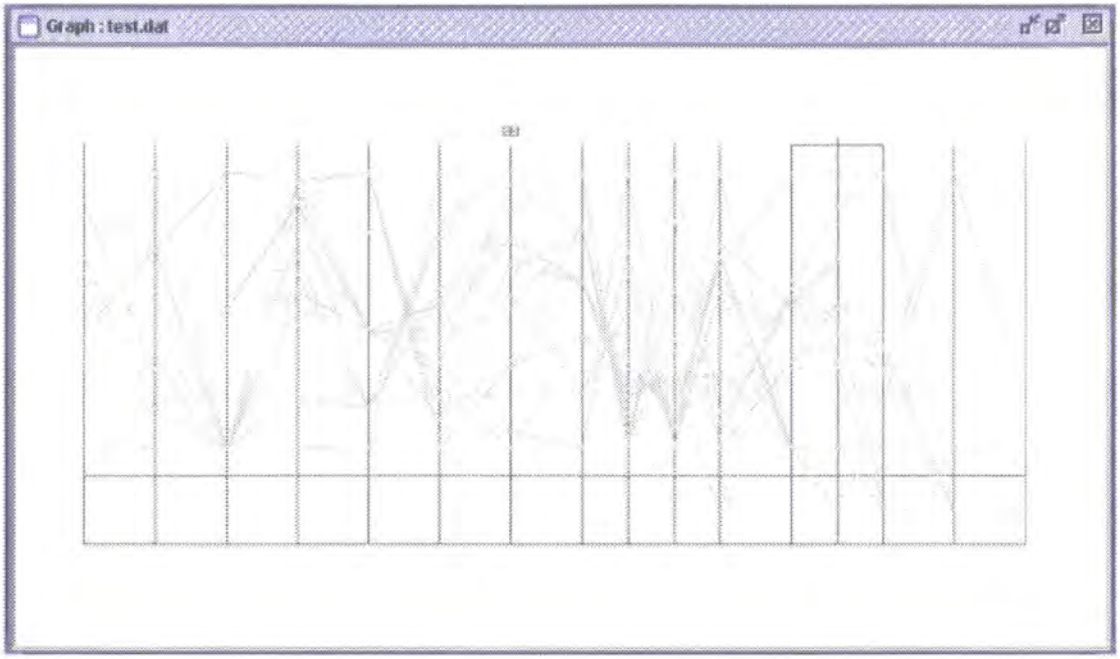


Figure 4.1 – The Parallel Coordinates Display

Description

The implemented Parallel Coordinates display (see Figure 4.1) is much the same as that provided elsewhere. Each data variable is displayed as an axis, the scale of which is determined by the type of the variable and values of the data for that variable. Each data item is displayed as a polygon line moving from left to right along the display; the points at which the line intersects each of the axes is equal to the data item's associated data values.

The tool supports the three basic types of textual data: quantitative, ordinal and qualitative. It breaks these basic types up into 5 defined types, each belonging to one, and only one, of the 3 basic types:

1. Nominal (qualitative)
2. String (qualitative)
3. Ordinal (ordinal)
4. Numeric (quantitative)
5. Date (quantitative)

Each variable in a data set must contain data values of one and only one of these types, otherwise the tool will output an error. Transforming the data variables in the data set into axes in the display, the tool calculates the scale of each axis based on the type and data values associated with the corresponding data variables in much the same way as other instantiations of Parallel Coordinates:

- Quantitative data – The axis scale is derived from the maximum and minimum data values in the set of data for the corresponding data variable.

- Ordinal – Taking the ordered set of all possible values for an ordinal type, the possible values are spaced evenly along the length of the axis.

- Qualitative – Taking the set of all possible values in the order specified in the file, they are placed evenly along the axis.

Different Approaches used to display the data

The current implementation of the tool provides no method for directly displaying underlying data values within the Parallel Coordinates' representation. Without a link or reference to the underlying data it represents, a visualisation can be considered only as an abstract display. Patterns and trends may be visible in the display, but the user would have no way of deriving the meaning behind these features without a frame of reference to the data. Nonetheless, the tool does provide access to the Parallel Coordinates' underlying data values, it just doesn't display them directly in the Parallel Coordinates display.

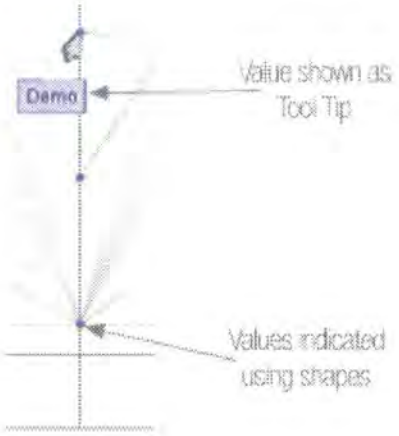


Figure 4.2 - Data Values Indicated By Using Shapes

Taking these issues into account, a separate implementation of the Parallel Coordinates display has been developed. This uses shapes to denote the points where each of the polygon lines crossed the axes (e.g. the data values). When a user wishes to find out the value at any one of these points, they move the mouse over the shape and the point's associated value is presented in a tool tip (as shown in Figure 4.2). This feature provides a useful frame of reference to the underlying data, enabling users to draw inferences about the data based on what they perceive in the visualisation. However, when the data set contains a large number of different data values for a single data variable, the corresponding axis in the Parallel Coordinates display is densely populated with shapes depicting each of the different values. In this situation it would be increasingly difficult to discern the different shapes from one another and hence to gain access to the required data values. The display of so many shapes would also clutter the display and impair the visualisation's effectiveness.

Enhancements to the Parallel Coordinates Display

The Display of Missing or Corrupt data

Compared to other Parallel Coordinates displays, Figure 4.1 has one quite distinct feature. Slightly above the horizontal line depicting the base of the representation is another horizontal line, termed the 'null axis'. The 'null axis' is essential to the representation's display of missing or corrupt data values. Below this line missing or corrupt data values are plotted as shown in Figure 4.3 and Figure 4.1. Above this

line, for all 'valid' data values, the tool's Parallel Coordinates display behaves in the standard manner.

Using a toggle option provided within the tool, missing and corrupt data values can also be emphasised by displaying the line segments passing to and from points below the 'null axis' using dashed lines. This feature enables users to quickly identify missing or corrupt data values without having to follow the polygon line to the point where it crosses the 'null axis', a point that may not be visible within the current viewing area.

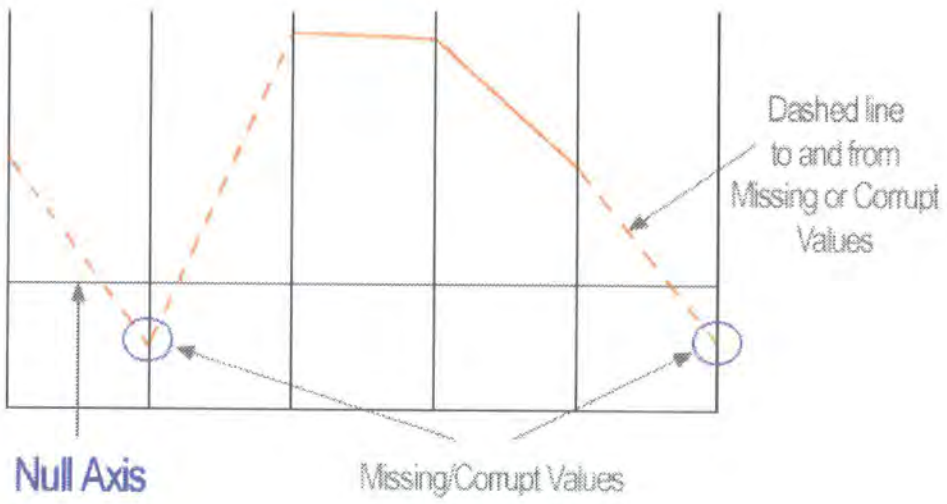


Figure 4.3 - How Null/Missing Values Are Displayed

Other approaches considered for the display of missing or corrupt data values included:

- Leaving a space in the polygon line where the data values for the corresponding axis are missing/corrupt:

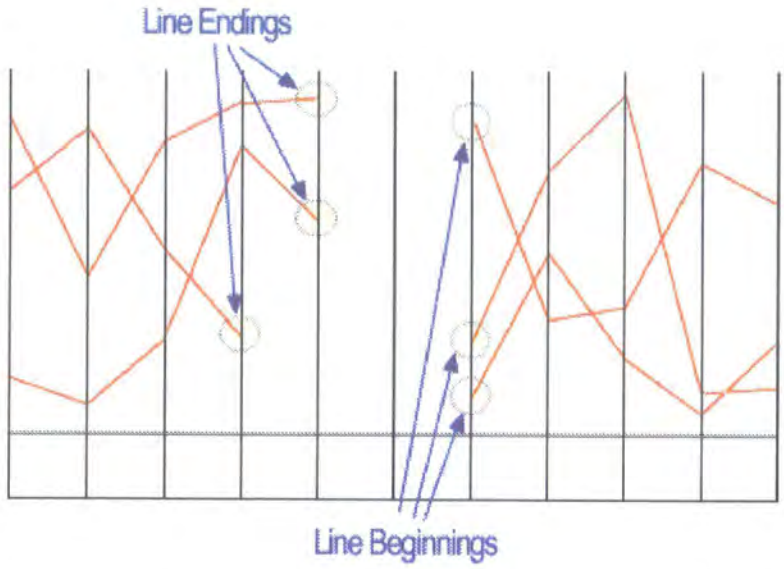


Figure 4.4 - Missing/Corrupt Values Shown as Spaces in the Polygon Line

The problem with this option is that with more than one missing or corrupt data value on an axis, the user would be faced with potentially many line endings and many line beginnings, and with no support for identifying how they relate to each other (see Figure 4.4).

- Continuing the polygon line to the next valid point when a missing/corrupt value is encountered:

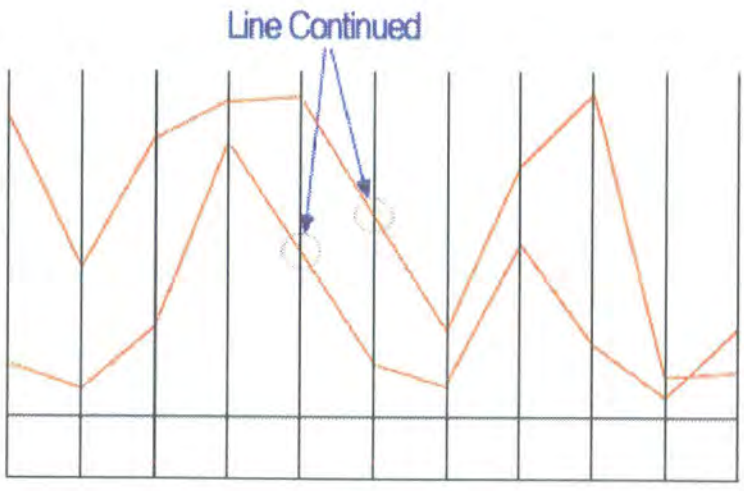


Figure 4.5 - Missing/Corrupt Values Shown as Continuing Polygon Lines

By not plotting the missing or corrupt values, the polygon line would continue in a straight line to the next valid point as shown in Figure 4.5. This introduces the chance that missing/corrupt data values may be misinterpreted as the value where the continued line crosses the axis (see Figure 4.5). It may also be the case that valid data values that happen to form a straight line could also be misinterpreted.

Currently the majority of visualisations ignore the appearance of missing or corrupt entries in the data sets they present, and in doing so remove a potentially valuable source of information. The appearance of missing/corrupt data values in data sets may provide key information relating to possible problems in the methods used to collect, store or process the data. Visualisations that enable the user to view, and more importantly to discern missing/corrupt data values from the remainder of the data provide the potential to access this largely untapped resource.

The Introduction of Relationships between data variables

The developed Parallel Coordinates implementation supports the display of possible relationships between data variables in a data set. Three types of relationship between data variables have been identified and are supported in the tool's implementation.

Note: All the following examples are based on data about software components, so each data item in the data set is a component and each data variable is a component property.

Conceptual Relationship

A conceptual relationship is one in which one or more variables within a data set are seen to embody a single higher-level property of the data items.

For Example:

> 'Mean time before failure', 'number of reported errors', and 'number of changes since the previous release' are all indicators of a component's quality, i.e. they represent the same high-level concept. Thus each of these data variables could be collated within a 'Quality' conceptual grouping relationship.

How it is displayed:

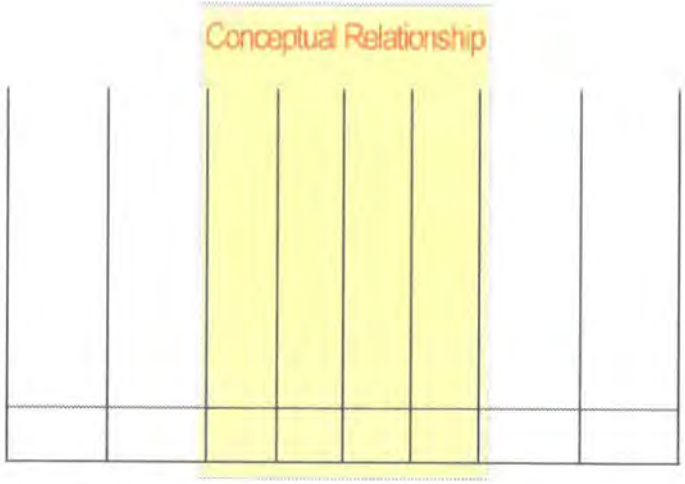


Figure 4.6 - Display of a Conceptual Relationship

Data variables grouped in a conceptual relationship are displayed within the Parallel Coordinates representation as shown in Figure 4.6. Axes corresponding to data variables within a single conceptual relationship are grouped in the Parallel Coordinates display and the spacing between these axes is reduced.

Compound Relationship

A compound relationship is one in which different data variables in the same data set represent alternate views of the same property.

For Example:

> The price property of a component may have several instantiations: 'unit price', 'site license price', 'user license price', etc.; these different possible data variables could be grouped in a compound relationship.

How it is displayed:

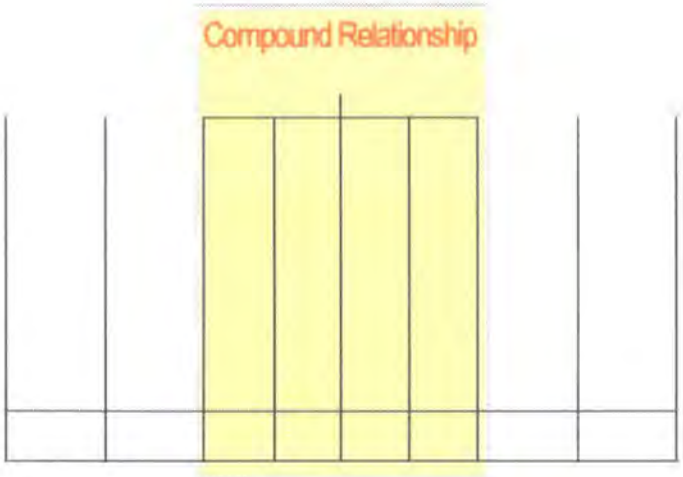


Figure 4.7 - Display of a Compound Relationship

Data variables that are part of a compound grouping relationship are displayed as shown in Figure 4.7. As with data variables within a conceptual relationship, data variables within a compound relationship are grouped within the parallel coordinate display and the space between the axes in the grouping is reduced. However, to distinguish the compound relationships from conceptual relationships, a capping line is placed on top of the grouped axes (see Figure 4.7).

Parent-Child Relationship

A Parent-Child relationship is one in which the values within one data variable are calculated using the values from one or more of the other data variables in the data set.

For Example:

> A component's quality property, appearing as a variable in the data set, may be calculated from values in a number of other data variables such as 'mean time before failure' and 'number of reported errors'.

How it is displayed:

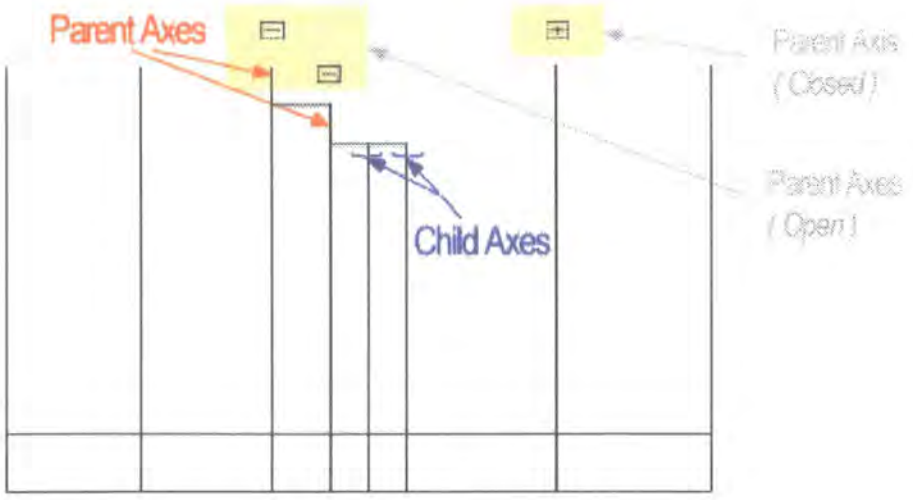


Figure 4.8 - Display of a Parent-Child Relationship

This type of relationship introduces a nest-like structure to the definition of the different data variables within a data set, where data values for a 'high-level' variable may be calculated using the values from one or more 'lower-level' variables. The developed implementation attempts to preserve these nesting structures present in the data variables within the display. Where the values within a data variable are dependent on the values from other data variables, these 'lower-level' data variables are displayed as 'child axes' nested within the axis corresponding to the data variable to which they contribute (see Figure 4.8). Above each axis that contains child axes, an explorer-like tab is displayed, enabling users of the tool to view or hide any child axes.

4.4.3 Dynamic Queries

How they are incorporated into the Parallel Coordinates display

The prototype is based on work by Siirtola [Siir00] that incorporates Dynamic Query mechanisms into Parallel Coordinates visualisations. Dynamic Query controls are included within the parallel coordinate display, removing the need to dedicate separate screen area to their display. The controls appear as pairs of pointers on the axes of the Parallel Coordinates' representation, used to select ranges of values (see Figure 4.9).

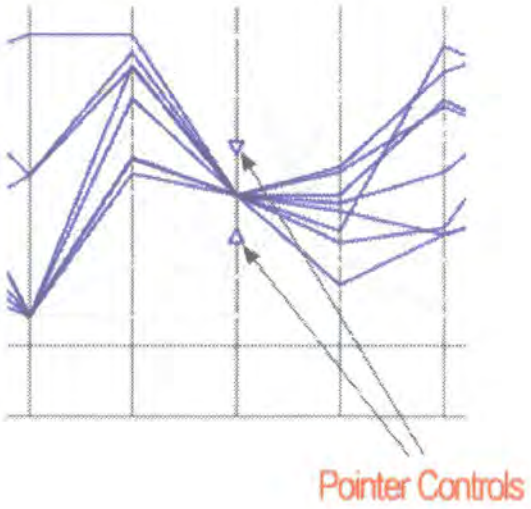


Figure 4.9 - Pointers Used to Control Queries

The parallel coordinate visualisation developed supports the display and construction of multiple, relatively complex, queries. Using these features the tool enables users to focus upon subsets of data of special interest, thereby providing extra support for the identification of trends and patterns in the underlying data set.

How to introduce a query

To introduce a query into the Parallel Coordinates display the user must first highlight a region on an axis containing data values of specific interest (see Figure 4.10).

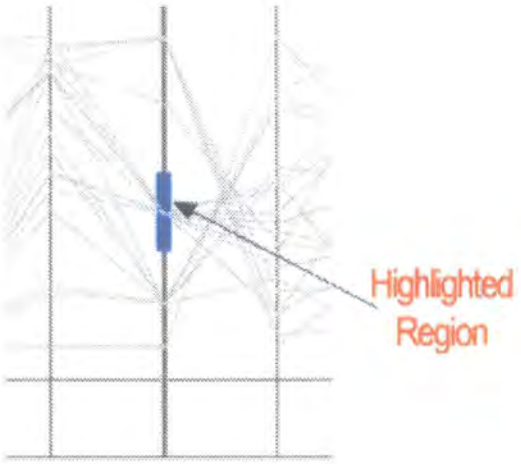


Figure 4.10 - Highlighted Region on an Axis

The user must then choose to introduce a query based on this selected region and, when prompted, select a colour to represent the new query. Once completed the query and the results of applying the query are displayed in the parallel coordinate representation.

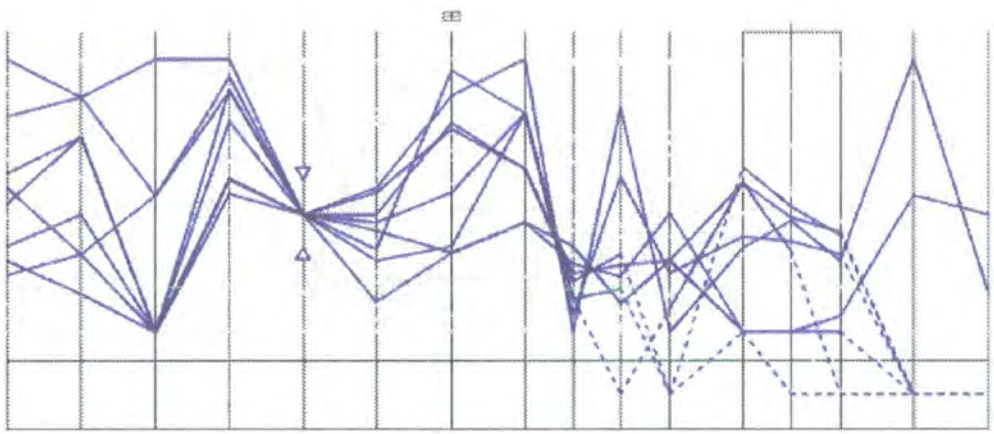


Figure 4.11 – The Results of Introducing a Query to the Display

As shown in Figure 4.11, the region highlighted is now delimited by a pair of pointers in the colour selected for the query. All polygon lines passing within the query bounds are highlighted, again using the colour selected for the query, and all other lines are greyed out and pushed into the background.

How to build up a query

The tool supports the construction of more complex queries by enabling the user to add more constraints to existing queries. First the user must once again highlight a region on an axis. However, instead of creating a new query, this constraint (the highlighted region) can be added to an already existing query. When adding constraints to existing queries the user can select how the new constraints will be added. The tool supports query construction using AND, OR and XOR set operators. Thus if the set of data items (polygon lines) inside the current query is A and the set of data items within the new constraint is B then the set of data items C within the newly updated query will be the results of one of the following:

- A AND B
- A OR B
- A XOR B

Once the user adds a new constraint to an existing query the display is updated to show the new set of values selected by the query. Figure 4.12 shows the result of adding a constraint using the AND operator to an existing query consisting of a single constraint.

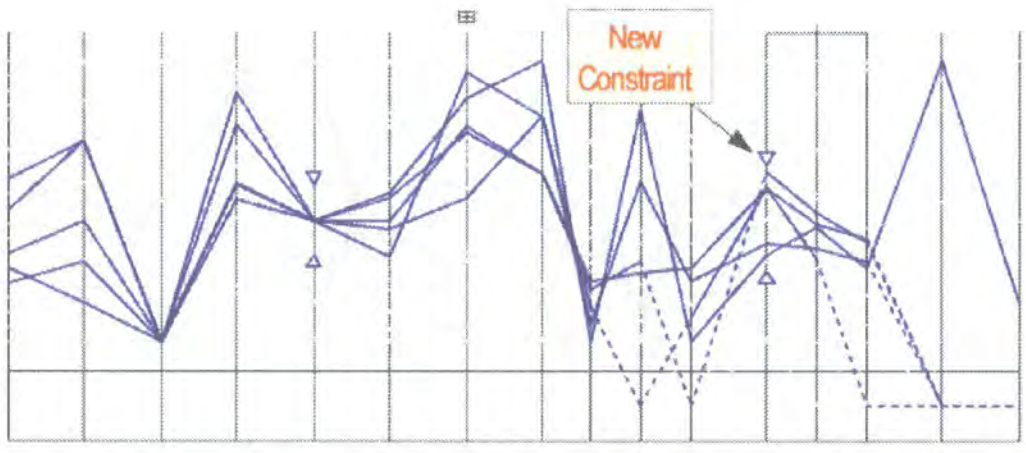


Figure 4.12 – The Result of Adding to an Existing Query Using 'AND'

The Parallel Coordinates implementation supports the ability to create and display multiple queries at any one time. Figure 4.13 shows a display containing two

separate queries: one coloured red, and the other coloured blue. It is possible for a data item (polygon line) to be in the highlighted set of more than a single query, and if this is the case the associated polygon line is coloured black (see Figure 4.13).

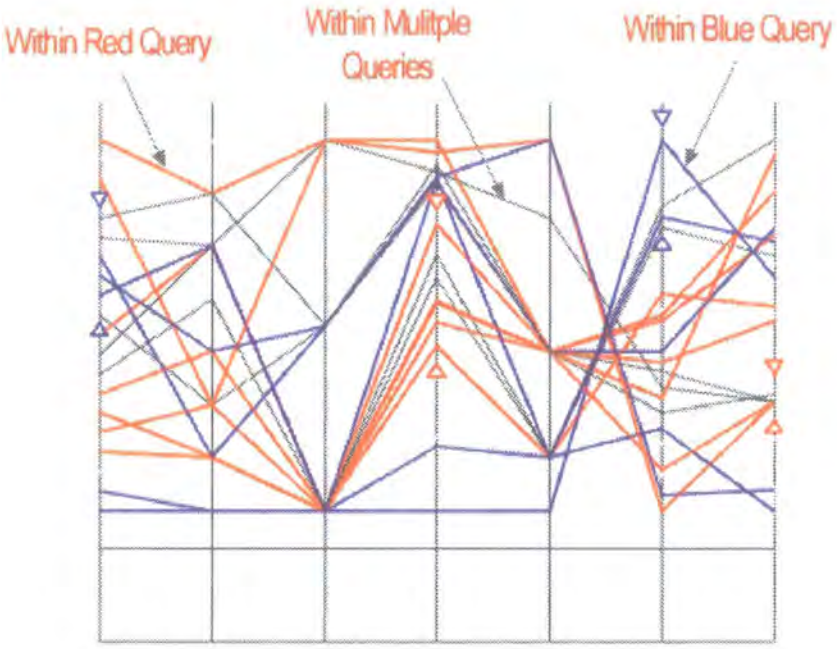


Figure 4.13 - A Display with Two Overlapping Queries

The true power of Dynamic Queries comes from the user's interaction with the query controls and the dynamic update of the visualisation. Within this implementation of Parallel Coordinates, the controls are the pairs of pointers; by clicking and dragging these pointers the user can change the associated query and the display updates to keep pace with these changes.

Approach used for the various data types

Pointers are used to depict the user-defined constraints, forming queries that filter the data set. The pointers are used to define regions on all types of axis. This kind of control mechanism works well with ordinal and quantitative axes, where the scale bares some relevance to the underlying data; however, for qualitative axes the selection of a range of values makes little sense. Instead, it is much more likely that a user would wish to select data values dotted across the entire axis.

To support the selection of data values appearing at different points on an axis, the Parallel Coordinates implementation enables users to define constraints made up of multiple ranges. As shown in Figure 4.14 this enables the user to select any qualitative values of interest by placing them within highlighted regions.

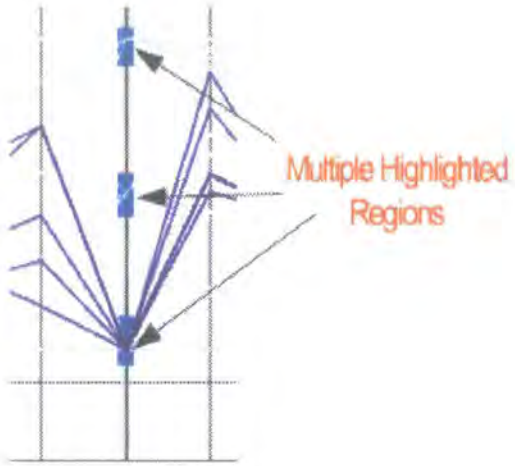


Figure 4.14 – A Display with Multiple Regions Selected

These constraints can then be used in the construction of queries. The resultant display shows pointers marking each of the ranges highlighted, and an updated representation taking into account the newly introduced constraint. Furthermore, the benefits of this are not restricted to qualitative data; it also enables users to specify constraints with multiple regions for any type of data.

4.4.4 Data Item Selection

The tool allows the user to select and investigate data items of particular interest. Each data item is represented as a polygon line in the Parallel Coordinates display, and to select a single data item a user must click on part of the corresponding polygon line. Selected lines appear highlighted in the parallel coordinated display and the user may go on to select multiple data items by holding down the ctrl key. Once a user has selected one or more data items (polygon lines) that they wish to explore further, they can select to expand and view this selection of items in greater detail within the prototype's table and Starplot displays.

4.4.5 Display data in Table

To present the selected set of data items in greater detail, the first view presented to the user is a data table (see Figure 4.15) containing all the data values of the selected items. This display provides the user with access to the underlying data values presented in the Parallel Coordinates representation, enabling the user to build a greater understanding of how the data and the representation relate. However, it also provides the user with a valuable tool for comparing the selected data items and their associated data values.

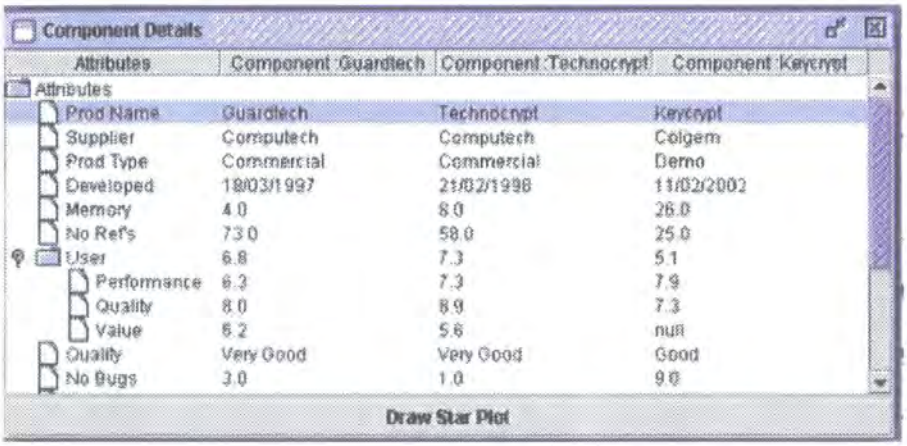


Figure 4.15 - Data Table Containing Data Values for Investigation

Note: Missing or corrupt values are displayed using the value 'null' in Figure 4.15.

As shown in Figure 4.15 the data table can contain nested values, thereby conserving any parent-child relationships represented in the Parallel Coordinates display. Also, to help with the visual comparison of the different data items the user can change the position of each column in the table.

From this view the user can also choose to display any of the data items shown in the table using another visual representation called the 'Starplot'.

4.4.6 The Starplot Display

The Starplot visualisation is based on Parallel Coordinates, but instead of positioning the axes in parallel to one another they are positioned in a circle, as shown in Figure 4.16. Polygon lines that stretched to the horizontal now form shapes that are

coloured and layered on top of each other. In place of the 'null axis, each of the axes start a set distance way from the centre of the display allowing missing/corrupt data values to be plot at this central point (see Figure 4.16).

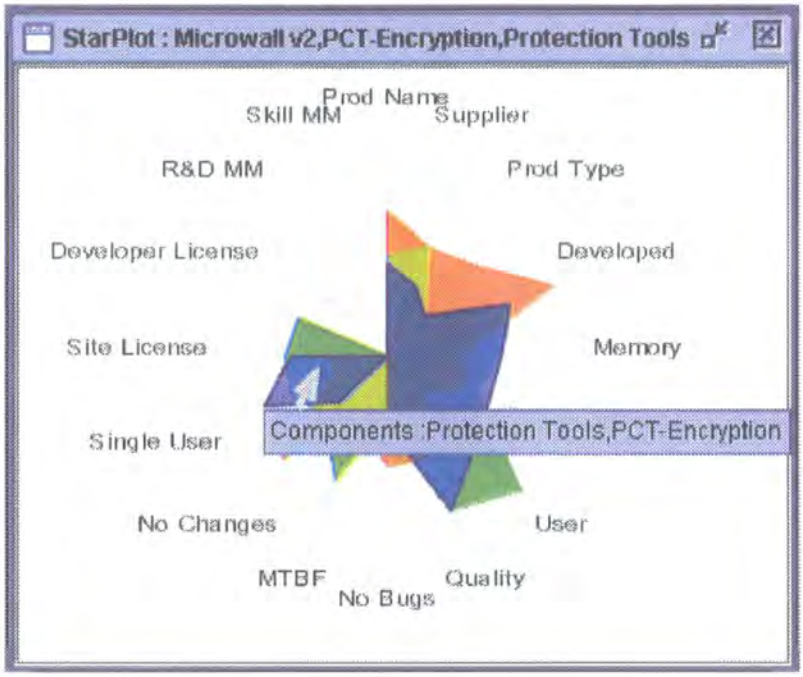


Figure 4.16 - The Starplot Display

The use of transparency enables the Starplot to display several different data items by layering their corresponding shapes on top of one another. A user can change the ordering of these layers by interacting with the display. This provides the means to overcome possible occlusion problems, but also helps support the process of directly comparing the data items presented.

To provide the user with a means of identifying which of the shapes represents each of the data items, the Starplot provides a tooltip display. The tooltip is used to display a list of all data items represented by shapes directly beneath the mouse pointer's location. In cases where more than a single shape lies under the pointer, the list of data items is given in the order the shapes appear, front to back, in the display.

The Starplot display enables data items and their associated data values to be compared visually. By organising its axes into a circle this representation also supports the display of many data variables within a restricted area; thereby removing the need to scroll around a potentially large Parallel Coordinates representation.

4.4.7 Other Tool Details

Zoom Functionality

The tool enables the user to change the scale of the parallel coordinate display in a number of different ways:

- Zoom on an axis – Selecting a densely populated range on an axis the user can rescale the axis based on the area selected. This enables densely packed lines to be viewed more clearly.
- Horizontal Scale – The user can increase or decrease the horizontal proportion of the Parallel Coordinates representation.
- Vertical Scale – The user can change the vertical size of the Parallel Coordinates representation.
- Overall Scale – The user can change the proportions of the parallel coordinate representation.

Note: Changing the scale of the display does not change the thickness of lines representing the axes or the data items, and any labelling visible in the display also remains the same size.

Labelling

The tool provides a standard set of labels for each of the axes in the Parallel Coordinates display:

1. Name of the data variable the axis represents
2. Type of the data variable (e.g. Nominal, String, Numeric, Date, or Ordinal)
3. Max and minimum values of the data items for the variable

The labels can be toggled to be displayed or hidden depending on the user's preference. However, it may be the case that the current scale of the parallel coordinate display doesn't provide sufficient space to display the axis labels. In this case the tool only displays the labels for an axis when the user moves the mouse pointer over it.

Vertical Display

As well as a conventional horizontal representation of Parallel Coordinates, a further implementation of the tool also provides the ability to display the representation vertically. The introduction of Parent-Child relationships into the representation gives Parallel Coordinates a tree-like, nested structure. Although against Tufte's recommendation that graphical representations should tend to the horizontal [Tuft83], the vertical display hopes to make better use of users' familiarity with navigating tree structures like those presented in Windows Explorer. However, the trees presented in the tool's vertical Parallel Coordinates display open from the right, unlike the trees in windows explorer which open from the left. This was done in an attempt to conserve the convention of values increasing from left to right, but also to give all axes a common point of origin.

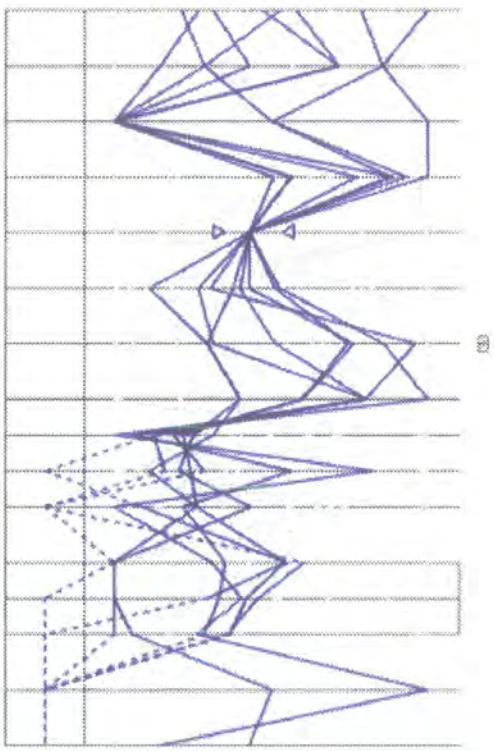


Figure 4.17 - Vertical Display of Parallel Coordinates

This vertical display raises many issues to do with convention, and the result shown in Figure 4.17 represents only one interpretation based on a trade-off of these issues.

However, besides introducing issues of convention, the new display also introduces new problems in terms the orientation and positioning of axis labels.

File Structure

The files used by the tool provide all the data required to produce the Parallel Coordinates display:

- Data item names
- Data variable names and types
- Data values
- Details of any relationships between the different data variables

The file structure is based on tags similar to the form of XML. Some further work has been done by other researchers in developing a format in the form of XML, with a view to using the data sets within other applications.

Note: For more detail of the file structure used by the tool see Appendix A.

4.5 Conclusions

4.5.1 Summary

This chapter has provided an overview of the prototype tool developed as part of this thesis' underlying research. The chapter not only includes an overview of the tool's implementation, but also details some of the rationale and decision making involved in its development: including an outline of the target problem, a description of the Parallel Coordinates and Dynamic Query techniques and a set of currently unresolved issues that the prototype tries to address.

4.5.2 What the implementation has achieved

The developed implementation provides a tool capable of displaying relatively large sets of multivariate data. Using a combination of the Parallel Coordinates and Dynamic Query techniques, the tool supports dynamic filtering of the displayed data, and provides the opportunity to identify potential trends and patterns in this data. Addressing some of the problems identified with the use of these two techniques and

with visualisation research as a whole, the tool's implementation also introduces new features into the Parallel Coordinates display:

- Providing the ability to deal with and make explicit the existence of missing/corrupt data values.
- Identifying the potential for relationships between the different data variables in a data set, the tool extends the Parallel Coordinates still further to incorporate their display.
- Using the logical operators AND, OR and XOR the tool enables multiple, relatively complex queries to be stored and represented.
- Taking into account the mismatch between slider controls and qualitative data types, the tool provides the ability to use sliders to highlight multiple (single or grouped) values appearing on an axis.

The tool also provides essential access to the underlying data values. This enables the user of the tool to relate what they see in the parallel coordinate and Starplot displays to the stored data values. The tool's data tables provide this important link to the data values; nevertheless, as with the Starplot display they also provide mechanisms supporting the direct comparison of data items and their associated values.

The prototype represents an effort to visualise large sets of multivariate data and address some of the issues identified in the underlying research. In attempting to resolve these issues, the tool's development unearthed other potential problems. Where possible these problems were resolved during the tool's development; however, some remain and could be the target of future work.

4.5.3 Identified Problems and Future Work

The very nature of the developed tool is the source of many of the issues still remaining. Developed to showcase new concepts, the tool is not meant to be a 'perfect application' ready for sale and distribution. The innovative ideas presented are untested, and inevitably are the source of some problems. It is only through further investigation and testing that these ideas can mature and the problems can be resolved.

The next section presents some of the more prominent issues identified, including possible avenues for exploration in the future.

Relationships between Data Variables

The tool introduces three types of relationship that may exist between the different data variables in a data set, and it incorporates them into the Parallel Coordinates display. This advancement should not be underestimated, but in hindsight these relationships and their impact on the parallel coordinate plot could have been investigated in greater detail. Future work could look to continue the investigation into the existence of such relationships and how they might be incorporated into visualisation techniques such as Parallel Coordinates.

The prototype visualisation tool enables the display of the three relationships and supports many possible combinations of these relationships. Currently, however, for complex interrelationships the tool often requires data to be repeated in the input files. For example, when one data variable is a member of more than a single relationship the data for this variable, in most cases, currently has to be repeated.

Provision for Complex Queries

Unlike many implementations of Dynamic Queries, the prototype tool supports the construction of multiple, complex queries within a single display. This allows the user more flexibility in the search for data items with specific properties. However, in providing the ability to introduce and store more complex queries the tool can require increased cognitive effort on behalf of the user. The result of the queries is dependent on the order constraints were added to it, and the tool provides no indication of this ordering. Through the experience of interacting with constraints in a query, users can learn to anticipate the impact of modifying these constraints, but no other support is provided to enable the user to better understand the queries constructed. Further work could be done investigating and developing methods to overcome this problem, including the creation of a method capable of storing and communicating details of the displayed queries and their construction, allowing users to gain a good understanding with minimal mental effort.

Displaying plots within multiple queries

Currently, when a data variable - displayed as a polygon line in the Parallel Coordinates plot - lies within the bounds of more than a single query, the line in the representation is shown in black. This enables users to quickly identify data items lying within multiple queries for further investigated. Nonetheless, this approach provides little help in discovering which queries the data items appear within.

Future extensions of this tool should look to develop some method that incorporates this extra information. This could be provided using some kind of drill-down mechanism, enabling the user to select a data item of interest and then select to view which queries the data item lies within.

Access to data values in the displays

The data table is an invaluable component of the developed prototype that provides access to the underlying data values. However, the user of the tool is required to relate what they see in the Parallel Coordinates and Staplot visualisations to the data values presented the data tables. Additional work could investigate the possibilities of displaying the data directly within the Parallel Coordinates plot. An attempt has been made to provide this functionality, but with only limited success. The approach needs to provide a method of accessing relevant data values without further cluttering the display.

Other issues

The tool raises other issues that could be approached through additional implementation (and not research). These include:

- The introduction of coupling between the different displays
    - This would provide a valuable link between the different displays within the tool, providing the user with the ability to relate the displays to one another more easily. For Example: highlighting a data item in one display would cause the same data item to be highlighted within the other displays.

- Speed of the program

- The interactivity afforded by the tool is one of its greatest assets. However, when dealing with large sets of data and many complex queries, the speed of the program will be slowed dramatically. Developed in Java a language not originally designed for such graphical purposes - the speed of the tool is already inhibited. Future work could involve porting the program to a language with additional runtime speed.

- Labelling of the representations

  - During the tool's development the labelling of each of the displays was not considered to any great extent. Labelling is an important component of all representations, helping users to better understand the display. However, labelling can also be a major source of clutter in the display. Future advances in the tool could take into greater consideration the role of labelling and how it is incorporated into the displays.

- XML file format

  - The tag style file format used by the tool is very similar to XML. Future work could look to modify this format into valid XML. This could then form the basis for some kind of standardised data source on which to develop a suite of visualisations.

4.5.4 Conclusions

The developed tool is a synergy of old and new. Making use of tried and tested techniques, the tool builds on the relatively firm foundations they provide, developing new concepts and representations. The motivation for the tool's creation lies primarily in trying to address the problem of information overload: the problems associated with the abundance of data and the need to derive useful information from this resource. Addressing the more refined problem of visualising large sets of multivariate data, the tool aims not only to provide a means of identifying useful patterns within the data, but also to resolve some of the issues identified within existing research.

In approaching these problems the tool provides the ability to display sets of multivariate data, and to identify patterns within the data. However, although the features and functionality afforded by the tool are extensive, the tool was developed only as a proof of concept. Thus both the developed concepts and the implementation itself could be subject to further investigation in the future.

Chapter 5 Case Study

5.1 Introduction

The tool's functionality has already been outlined in chapter 4, and it is foreseen that this range of functionality would be useful for a variety of domains and associated data sets. By developing a scenario-based case study, this chapter hopes to demonstrate the link between the functionality provided and how it might be applied in a 'real world' situation.

The first part of this chapter develops the 'real world' scenario used in this case study, identifying potential users of the tool and their associated goals. With these users and their goals in mind, the chapter then outlines the data set used to support the case study. The case study itself takes each of the identified users in turn, demonstrating how the developed tool could be used to help them reach their goals. Finally the chapter presents a review of the case study, describing what can be concluded from this study.

5.2 The Scenario

The concept of software reuse was once seen as the answer to many of the problems inherent in software development. Promising more reliable software developed in a shorter time span, it has clear benefits. However, for many reasons this technique and the processes involved have not been widely adopted.

The scenario proposed here is based on a component brokerage model of reuse. Central to this model is the role of the broker, creating a market for reusable components by bringing together developers of reusable components (suppliers) and developers looking to reuse components (integrators).

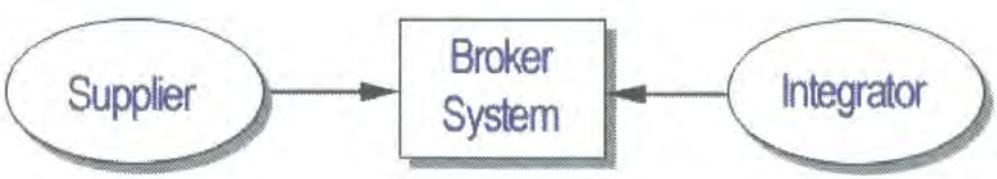


Figure 5.1 - Diagram of Component Brokerage Model

The broker plays an intermediary role between the Integrator and the Supplier (as shown in Figure 5.1), offering the supplier a show room from which to sell their components and the integrator a set of reusable components to help in system development.

The structure of the broker consists of a database storing the details of the available components, and search facilities enabling the integrator to identify components fulfilling their requirements. The key to providing these features is a classification scheme; the foundation for defining the structure and composition of the data stored, as well as the type of searching possible.

This component brokerage model already includes the two roles of the integrator and the supplier, but there is a clear need for a third role to administrate and organise the broker (see Figure 5.2).

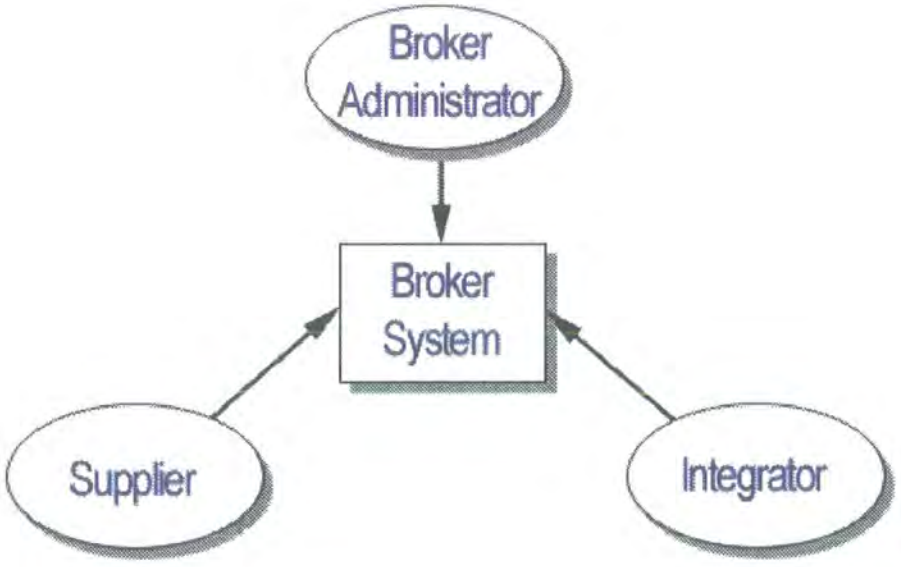


Figure 5.2 – The role of the Broker Adminstrator in the Component Brokerage Model

The broker administrator would be responsible for the standard of service offered to both the integrator and the supplier. This would involve developing and maintaining the classification scheme, providing support for the integrator's search for reusable components, but also the supplier's process of classifying their reusable components and specification of the component data. The role of the broker administrator may

also incorporate some kind of quality of service guarantee, whereby the administrator identifies suppliers of poor quality components and/or incorrect component data, and potentially removes them from the database.

5.2.1 How the tool would fit into the model

The developed tool would provide an interface between the different users (the broker administrator, suppliers and integrators) and the data stored within the broker: It could be used to visualise the data stored in the database, supporting the identification of useful patterns and trends, and would provide a useful aid in supporting the integrator's search for reusable components.

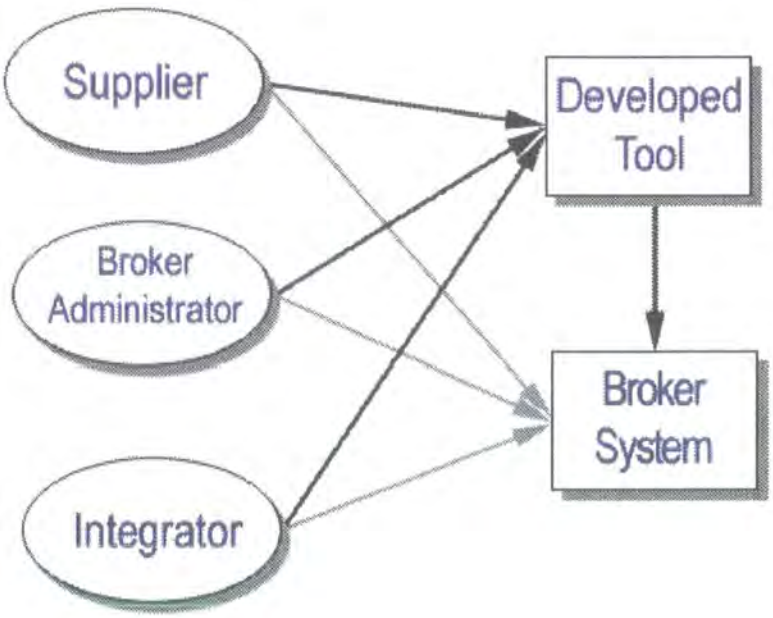


Figure 5.3 - Diagram showing how the developed tool fits into the model

It is envisaged that the broker system would provide some kind of initial query mechanism, such as a keyword search, enabling components of specific interest to be selected. The developed tool would be used to display the data from these results, allowing users to investigate the components' properties further. In the case of the integrator, this investigation would incorporate potential trade-off analysis in selecting candidate components for reuse.

5.3 The Users

As shown in Figure 5.2 there are three main types of user for the tool:

1. Supplier
2. Broker Administrator
3. Integrator

5.3.1 Supplier

It can be questioned whether or not a supplier would be allowed access to the data stored in such a broker, within the representations provided by the tool or directly. This is because suppliers may be able to gain an unfair competitive advantage from information derived from the data. However, if component suppliers were allowed access to the data, the tool would enable them to gain potentially useful and advantageous information from it. This might include identifying competition for their existing components or finding gaps in the component range provided by the broker that could be exploited by future components. If the broker stored customer feedback data, it would also be possible for a supplier to investigate how customers react to their components and the services they provide, enabling them to better understand their clients and their respective needs.

5.3.2 Broker Administrator

The broker administrator is central to the success of the broker. It is their job to develop and maintain the services provided to both the supplier and the integrator clients. Using the developed tool would enable the administrator to identify any alarming trends or patterns in the data stored. These patterns could help the broker focus upon areas of the classification scheme not functioning as previously expected, providing valuable feedback in its development. The displays provided by the tool would also allow the administrator to identify anomalies in the data set, possibly due to misclassification. If customer feedback data formed a portion of the broker database, it would also provide the administrator with the ability to monitor and control the standard of components available: the representations could be used to quickly identify suppliers producing consistently substandard products. Based on this evidence the administrator could investigate their findings further and potentially remove such suppliers from their database and their components.

5.3.3 Integrator

The main aim of the integrator is to identify components for reuse in their software projects. Based on a set of preconceived requirements, the integrator aims to filter and search the data stored in the broker until a component matching these needs is discovered. The developed tool could support these processes by providing the means for the user to filter the data based on their initial requirements. Taking into consideration identified patterns and trends in the data, possibly discovered through the use of the developed tool, the integrator could then go on to manipulate parts of their search criteria until a satisfactory set of components have been found. Presented with a set of components fulfilling most if not all of their requirements, the tool would also provide support comparing the candidate components with the view to identifying the component best matching the needs of the integrator.

5.4 The Data

5.4.1 Introduction to the data

The data, as with all other parts of the scenario, has been created to demonstrate the full range of the developed tool's functionality, and its support for different users and their various tasks. Furthermore, as this case study is meant to demonstrate the tool's application in a 'real world' situation, effort has also been taken to make the data set realistic.

5.4.2 About the data

The data itself constitutes a set of components and their properties (see Appendix C). It is envisaged that this set of components would be the result of some kind of preliminary searching, looking to identify the data for all components with encryption functionality.

5.4.3 Details

The created data consists of:

- 20 Components (Data Items)
- Each with 19 properties (Data Variables)

With the aim of demonstrating the full range of the tool's capabilities, the data contains missing/corrupt data values and data variables of all the types supported and defined in section 4.4.2. It also contains examples of the parent-child, conceptual and compound relationships (see section 4.4.2) between the different data variables.

5.4.4 Data Variables

Data Variable	Description	Type
Product Name	The name of the component	String
Supplier	The Supplier of the component	String
Product Type	The type of the component: Commercial, Evaluation or Demo	Nominal
Developed	The development date of completion	Date
Required Memory	The amount of memory required for the component to function correctly	Numeric
No. Reviews	Number of user reviews of the component	Numeric
User Rating	Average rating of the component by the reviewers Calculated from the 3 rating variables indented below	Numeric
→ Performance	Average performance rating of the component	Numeric
→ Quality	Average quality rating of the component	Numeric
→ Value	Average value rating of the component	Numeric
Quality	Quality of the component: Very Bad, Bad, Ok, Good, Very Good or Excellent	Ordinal
No. Bugs	Number of bugs found in this release of the component	Numeric
MTBF	The Mean Time Before Failure (MTBF) of the component	Numeric
No. Changes	Number of changes made since the previous release of the component	Numeric
Single User	The cost of a single user license for this component	Numeric
Site License	The cost of a site license for this component	Numeric
Developer	The cost of a developer license for this component	Numeric

License		
R&D MM	Research and Development Man Months [Comp]	Numeric
Skill MM	Skill Man Months [Comp]	Numeric

Table 5.1 - Table of Data Variables

Note: In Table 5.1 different shading has been used to indicate the relationships between the data variables:

- Blue – Parent-Child relationship
- Yellow – Conceptual relationship
- Grey – Compound relationship

Presented in a tabular format (see Appendix B) even this relatively small set of data is difficult to analyse: patterns and trends within the data are hidden within the table's uniform simplicity.

This case study hopes to demonstrate that any patterns present in the data can be accessed more readily using the developed tool.

5.5 The Case Study

5.5.1 Introduction

The case study is structured in terms of the 3 identified user types for the system:

- Supplier
- Broker
- Integrator

Following the natural movement of a component through this system, the supplier is looked at first, the broker administrator next and then the integrator. For each of these user types the case study looks at a number of use cases, and details how the tool might be used to support these different tasks.

5.5.2 Supplier

Case: Find a gap in the component market

In this case the objective of the supplier is to use the tool to help identify potential gaps in the component market provided by the broker. Identifying patterns in the data using the Parallel Coordinates display can provide the supplier with an insight into potentially untapped business.

The user is interested in finding a gap in the broker's component catalogue, where no components, or at least no components of a satisfactory standard, exist. The data collected based on past customer ratings of the components is essential to gauging this degree of satisfaction, and thus central to any investigation.

Firstly the user may introduce a constraint into the Parallel Coordinates display, selecting all components rated highly. (See Figure 5.4)

Figure 5.4 - Parallel Coordinates display highlighting components with high User Rating values

The three components highlighted in red, as shown in Figure 5.4, are those with a high rating value. The reliability of these values, as with other rating based data, is highly dependent on the number of times each of the components has been reviewed. On looking at the parallel coordinate plot once again (shown in Figure 5.4), it becomes obvious that the number of reviews for these components is not small enough to bring the rating values into question.

Investigating the child variables of the User Rating variable, the supplier can gain a greater insight into the components' ratings. From this view in the Parallel Coordinates display (see Figure 5.5), the supplier can note that the three components highlighted in red have high quality and value ratings but relatively low performance ratings.

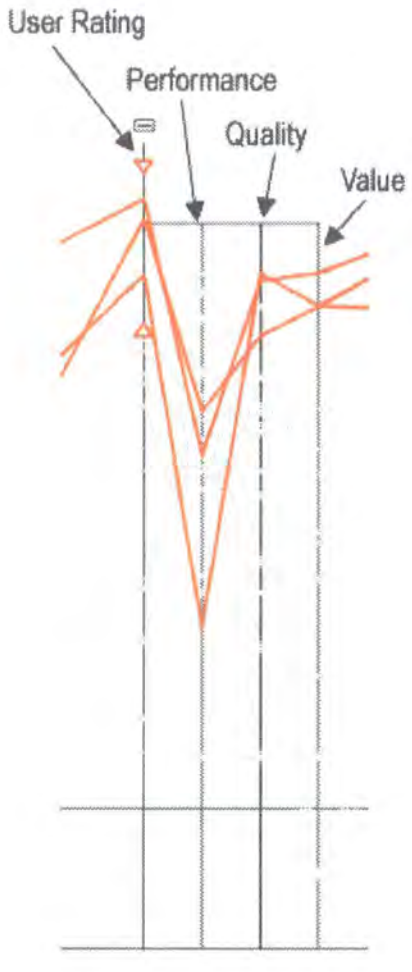


Figure 5.5 – Display showing the expanded User Rating Axis

This demonstrates that there is currently no component in the broker that has a high performance rating and is also rated highly overall. To investigate this situation further the supplier could introduce a second separate constraint to identify components with a high performance rating.

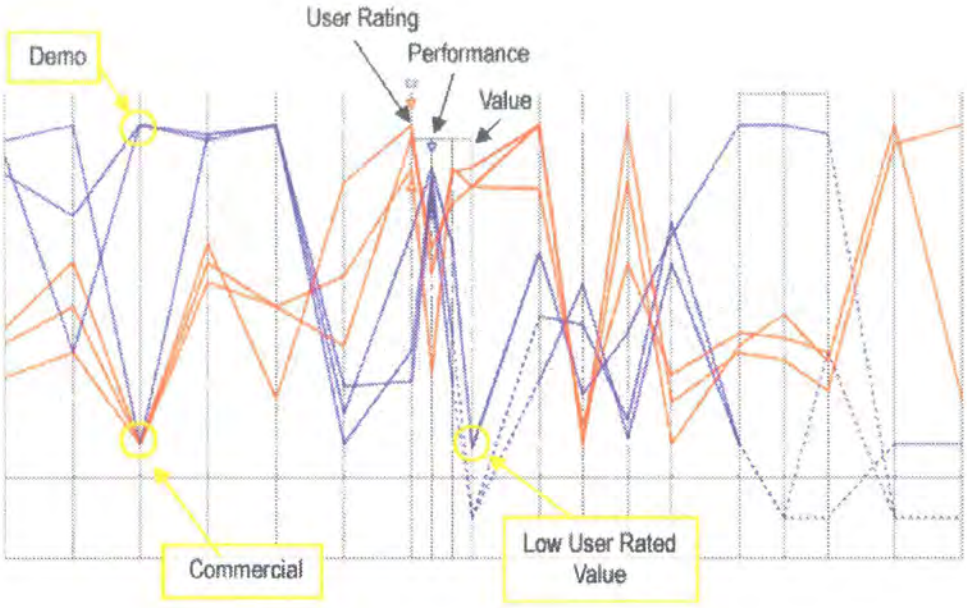


Figure 5.6 - Display highlighting components with high user rating and performance values

Figure 5.6 shows the result of introducing such a constraint. Highlighting three components in blue, the newly added constraint allows the user to focus their attention on the components with high performance rating values. Looking at these three components more closely the supplier would be able to see that two are available only as demos, with relatively low quality ratings and quality values. The remaining component is the only one to be available commercially and although it's quality rating values are higher than the other two components highlighted, they still remain relatively low, especially for the high price value shown that the user would be expected to pay.

The evidence from this investigation suggests there may be an opening in the broker for commercial encryption components with good performance, offering reasonable quality and value for money.

Case: Component Comparison

Here the objective of the supplier is to find out how customers rate their products and how they compare with similar components offered by other suppliers. This could provide valuable feedback to the supplier, enabling them to rectify possible problems and better meet the expectations of their customers (the integrators).

Interested in their components, the first step a supplier could take is to introduce a constraint into the Parallel Coordinates display causing them to be highlighted.

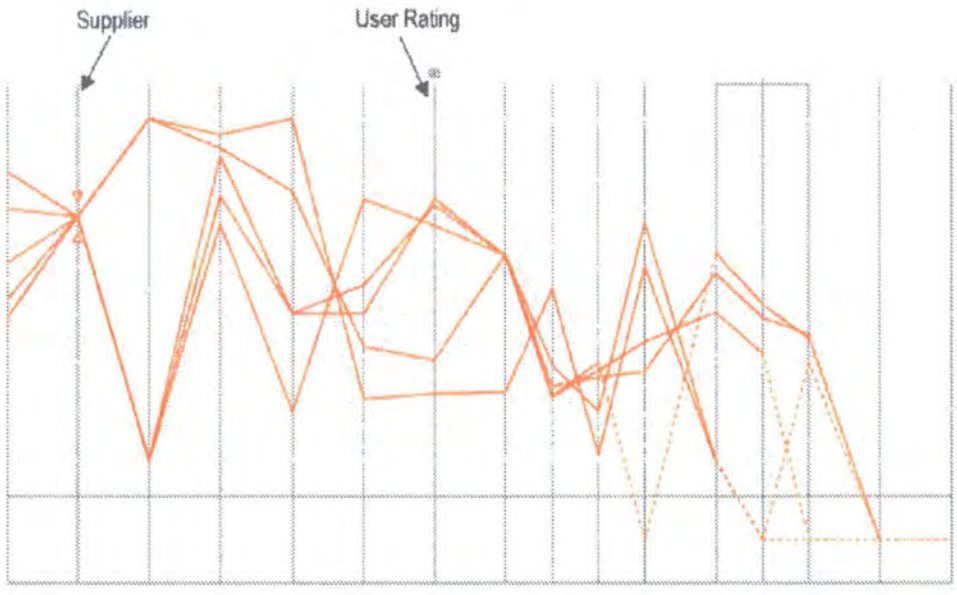


Figure 5.7 - Display highlighting all components supplied by Colgem

Figure 5.7 shows an example of this in which all the components supplied by Colgem have been highlighted in red. Investigation of Figure 5.7 reveals that all of these components have User Rating values in the middle of the range available.

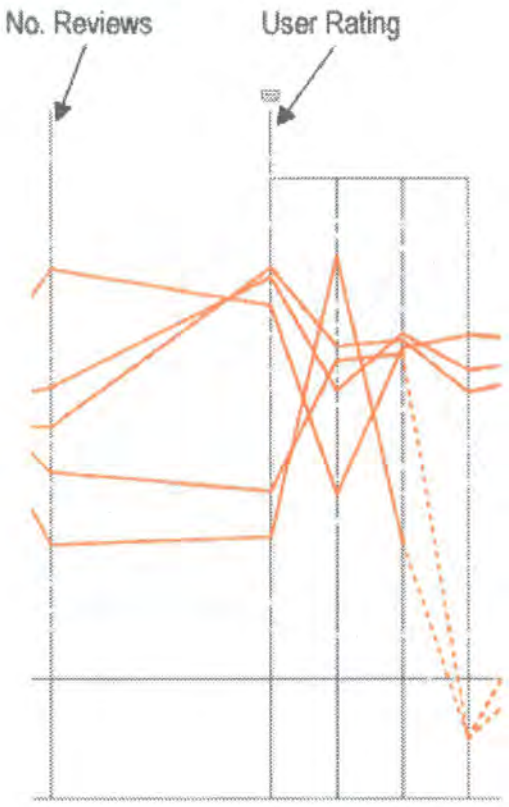


Figure 5.8 - Display of User Rating's child axes

Taking this investigation a step further, and recognising that User Rating is a parent variable, the variable can be expanded in the Parallel Coordinates display causing the axes for the child variables to be presented. Looking at the child axes within the User Rating axis (shown in Figure 5.8), once again the rating values for Colgem's components appear in the middle of the ranges. The No. Reviews axis in the parallel coordinate display (see Figure 5.8) also shows that each of the components has been reviewed enough times to gain a reliable rating.

From this investigation of components produced by Colgem, the supplier would be able to see that their components in general do not seem to excel or fail in any customer rating compared with other components. As a result of this initial investigation, the supplier may wish to look at each of their components individually, developing strategies to allow the components to better meet the expectations of the integrators.

5.5.3 Broker Administrator

Case: Identify possible problems with the classification scheme

Patterns and trends identified through investigation and manipulation of the parallel coordinate display could provide essential feedback to the broker administrator, including details of potential weaknesses in the classification scheme.

The initial Parallel Coordinates display enables the broker to gain a useful overview of the selected data set. This overview provides the broker administrator with the ability to view large amounts of data, identifying trends that may through further investigation be due to a potential fault in the broker's classification scheme.

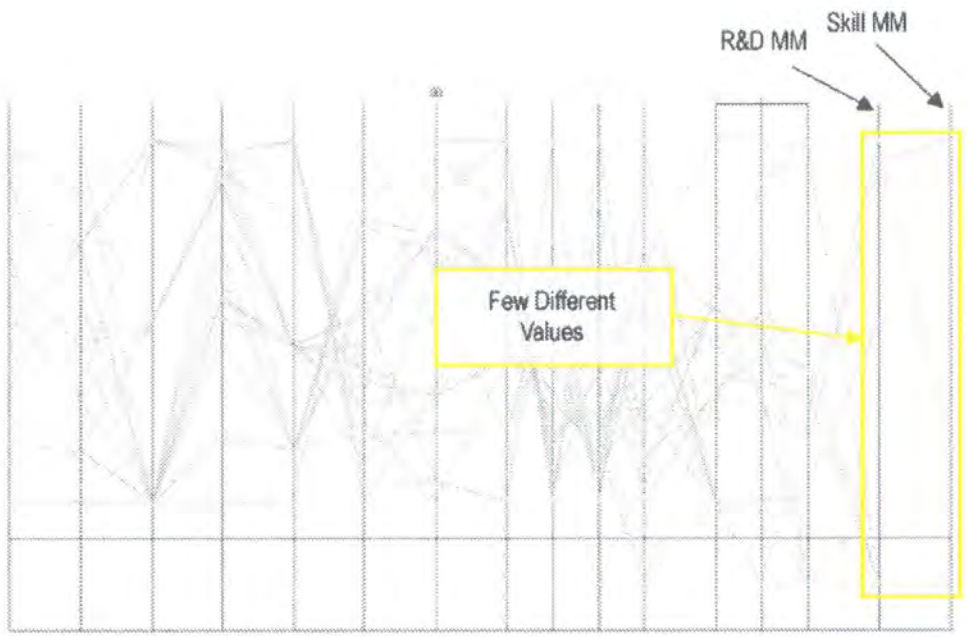


Figure 5.9 - Display showing lack of different values for the Skill MM and R&D MM axes

In Figure 5.9, the lack of different values for the 'R&D MM' and 'Skill MM' is easily recognisable: there are very few points where lines cross these numeric type axes. This small number of points means that a large proportion of the components displayed must have the same data values for these variables, and the appearance of so many missing/corrupt values is a reason for concern for the broker administrator.

The broker administrator could choose to investigate this trend in greater depth. By introducing a constraint into the Parallel Coordinates display, they can highlight

all the components with non-missing/corrupt values on the Skill MM axis (see Figure 5.10).

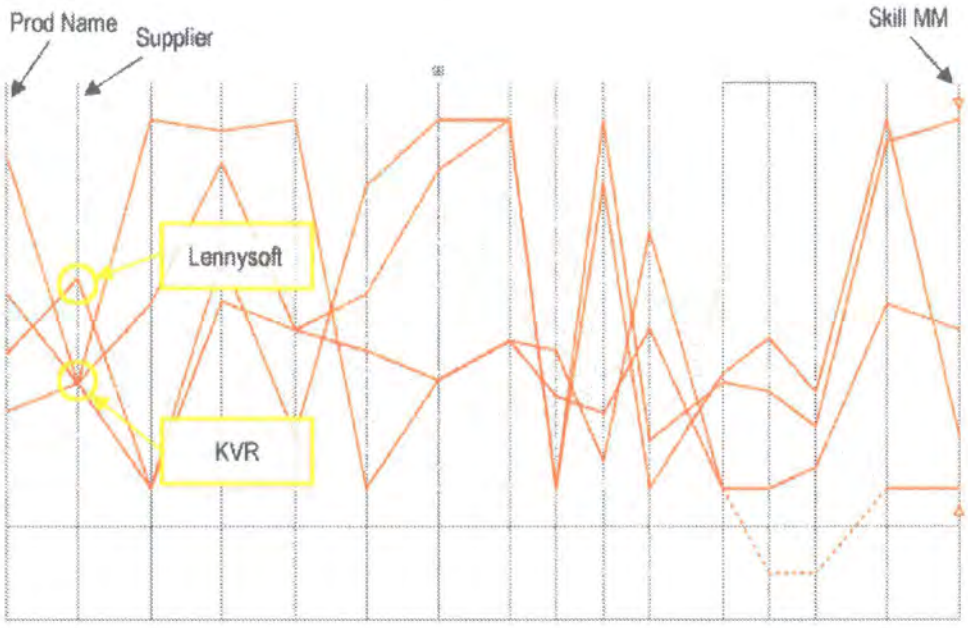


Figure 5.10 - Display highlighting components with non-missing/corrupt Skill MM data values

Looking at the Prod Name axis (where each component has a unique value) it is clear that only four components have valid values for the Skill MM and the R&D MM variables, as there are only four components in the highlighted set (see Figure 5.10). The values on the Supplier axis also show that only two suppliers have provided valid entries for these variables: LennySoft and KVR.

This leaves the broker administrator with many unanswered questions:

Why are there so many missing/corrupt values?

- Is it because the suppliers don't understand how to classify their components for this measure?

Or

- Is it because they do not wish to provide these details?

Either way this could be considered a fault with the classification scheme. To find out more the broker administrator must talk to the suppliers, asking LennySoft and KVR how they are able to provide the values, and all other suppliers why they are unable to provide these values. The result of this investigation may provide further valuable feedback in the development of the classification scheme.

Case: Identifying Misclassified Components

The Broker must be able to identify components that may have been misclassified by the suppliers. As part of the broker's service to integrators, it is essential that components be classified correctly, so identifying incorrectly classified components is important.

Within the classification scheme many variables will not be completely independent of each other; the values of one variable may be linked to the values of another. These relationships can be identified in the Parallel Coordinates display as trends in the polygon lines moving from one variable to another. The tasks of identifying these trends is a relatively simple one when the axes for the variables are close together, but can provide more problematic when further apart.

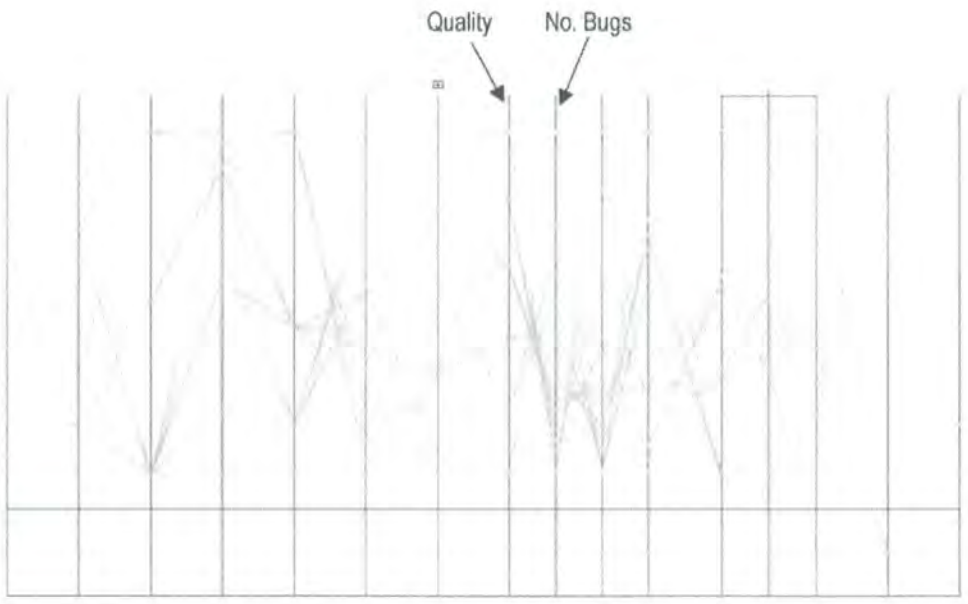


Figure 5.11 - Parallel Coordinates display of the data set

Figure 5.11 shows a number of clear relationships linking values of one variable to another. Looking at the Quality and No. Bugs Axes, it is clear that low values for the Quality variable tend to high values for the No. Bugs variable. This makes sense, as one would expect a program with better quality to have fewer bugs.

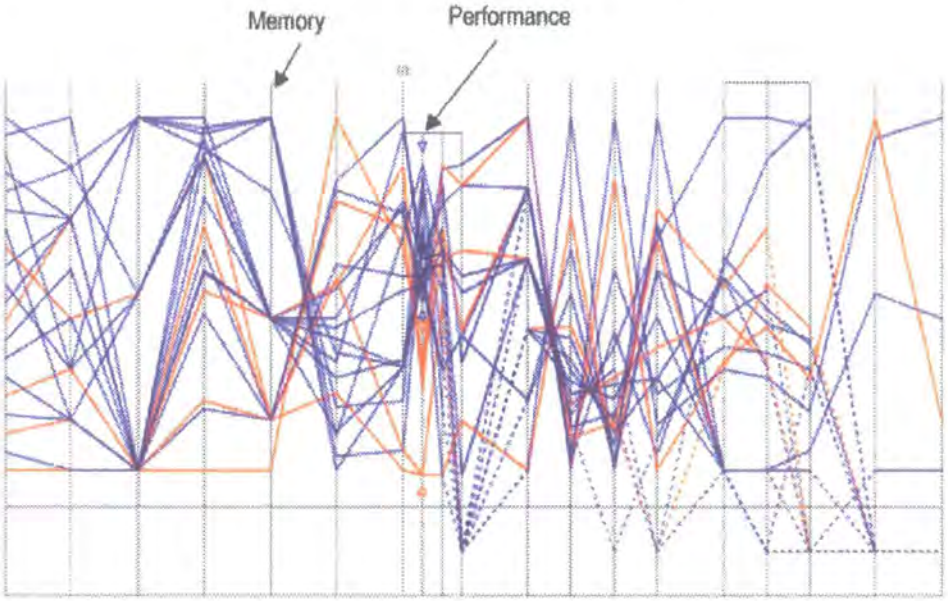


Figure 5.12 - Display highlighting set of components with high and low performance rating

When the axes under investigation are much further apart in the display, constraints can be used to investigate the existence of any linkage. Figure 5.12 shows the Parallel Coordinates display with two constraints: one highlighting the components with low performance values (in red) and the other highlighting those with high values (in blue). Looking at these lines crossing the memory axis, all the red highlighted components - those with low performance values - appear to have low memory values. In turn, the components highlighted in blue - those with high performance - have high memory values. Thus it seems from this data set that components requiring more memory perform much better than those with much lower memory requirements. As the number of plots in the Parallel Coordinates visualisation increases, the display can become overcrowded, making the identification of these patterns much more difficult. Nonetheless, the introduction

and manipulation of constraints in the display enables the user to focus on particular aspects of the data and help make such relationships become more apparent.

The existence and identification of links between the different data variables enables the broker administrator to identify outliers from standard trends that may through further investigation be the result of misclassification.

Comparing the two axes Quality and user rated Quality there is a clear and expected relationship between the two: high values for Quality tending to high values for user rated Quality.

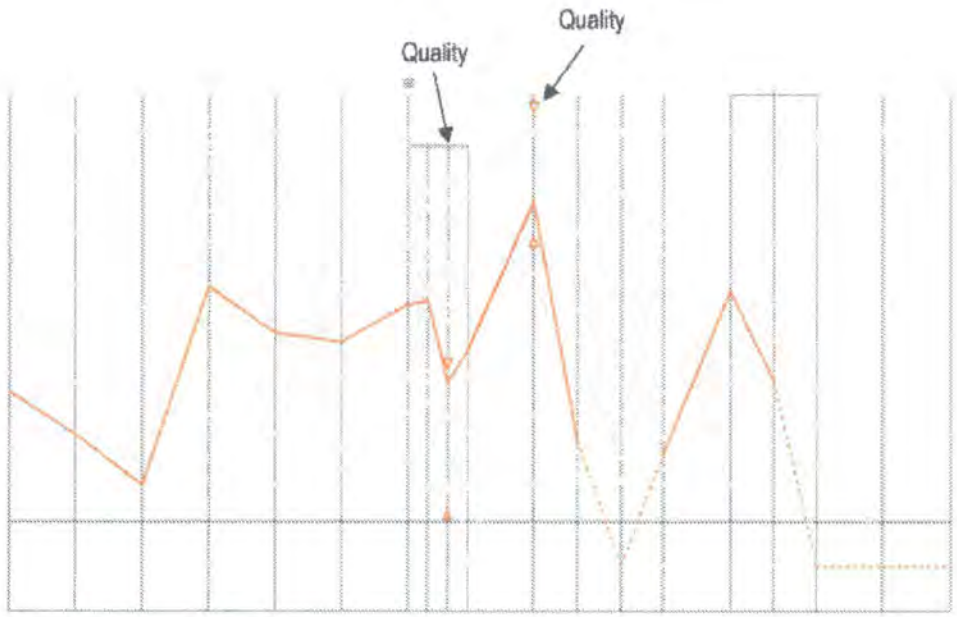


Figure 5.13 - Display highlighting a component that may have been misclassified

Highlighted in Figure 5.13, one component seems to buck this fairly reliable trend. This component, Microwall v2, has a relatively high Quality value and, in complete contrast to the trend with the other components, a low user rating of quality. The No. Reviews value is not so low as to question this rating value, so it could be that this component's quality value has been misclassified. This and similar evidence should provide the broker administrator with a starting point for further investigation: they must check if the component has been misclassified or if the fault lies elsewhere.

Case: Ensuring Quality of Service

It may be a goal of the broker administrator to try and ensure a quality of service to integrators by providing only good standard components. If this is the case then the administrator must be able to identify poor quality components and potentially remove them from the broker.

The quality and user rated quality variables are two of the main indicators of quality in the data set. Thus the broker administrator's definition of a poor standard component may be broadly based on the values for these two variables. However, as stated previously, the reliability of the user rated quality values are dependent on the No. Reviews values. Therefore when investigating the values for user rated quality, the user must also take into account the corresponding 'no. reviews' values.

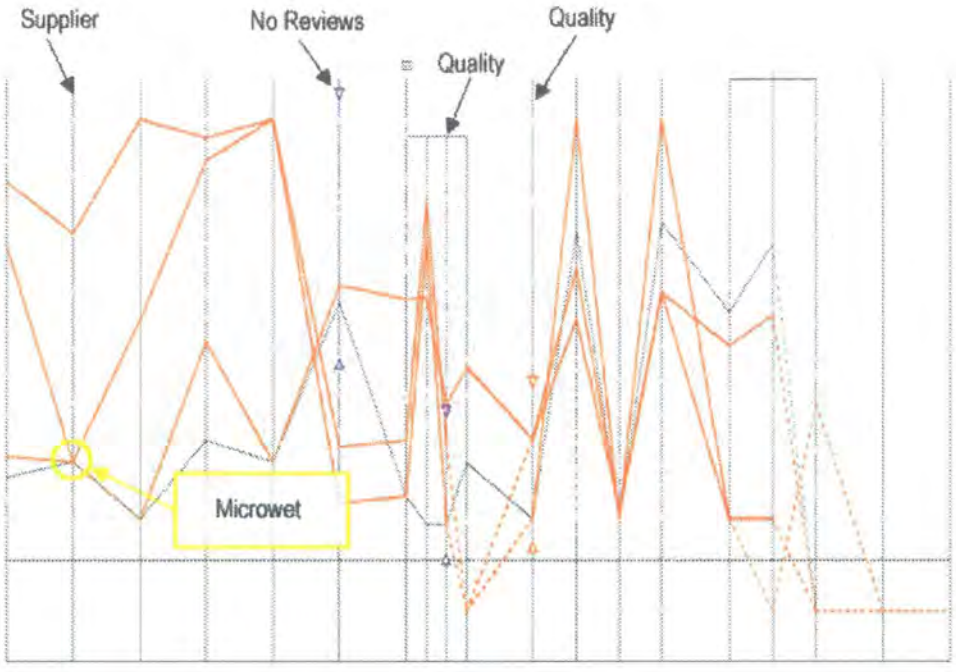


Figure 5.14 - Display highlighting components with low quality values, and low quality rating values with a large number of reviews

In the search for substandard components the broker administrator could introduce constraints as shown in Figure 5.14:

- Constraint (red) – Low Quality values

- Constraint (blue) – Low user rated quality values and relatively high No. Reviews values

Looking at Figure 5.14, there is one component, Microwall v1 (coloured black), identified as being a member of both of these highlighted sets. This component would therefore be a prime suspect for removal from the broker, based on its poor quality. Further investigation of Figure 5.14 shows that Microwet supplies the majority of the poor quality components in the broker. Based on this information the broker administrator could take further action, contacting Microwet and potentially removing their products from the broker altogether.

5.5.4 Integrator

Case: Support for user identification and selection of components

Arriving at the broker, the integrator will have a set of rigid and not so rigid requirements for a component. Their ultimate goal is to identify components fulfilling these requirements with a view to reusing them in their own software development processes.

The developed tool looks to support this selection process by providing the means to:

- Specify and modify their requirements in the Parallel Coordinates display
- Carry out trade-off analysis between the different requirements
- Focus on components of specific interest
- Directly compare components, ensuring the 'best' component is selected

In the form of constraints, the tool provides the integrator with the means to introduce their requirements into the display.

For Example:

A requirement that the component must be able to work with less than 16Mb of memory could be introduced into the display, as shown in Figure 5.15.

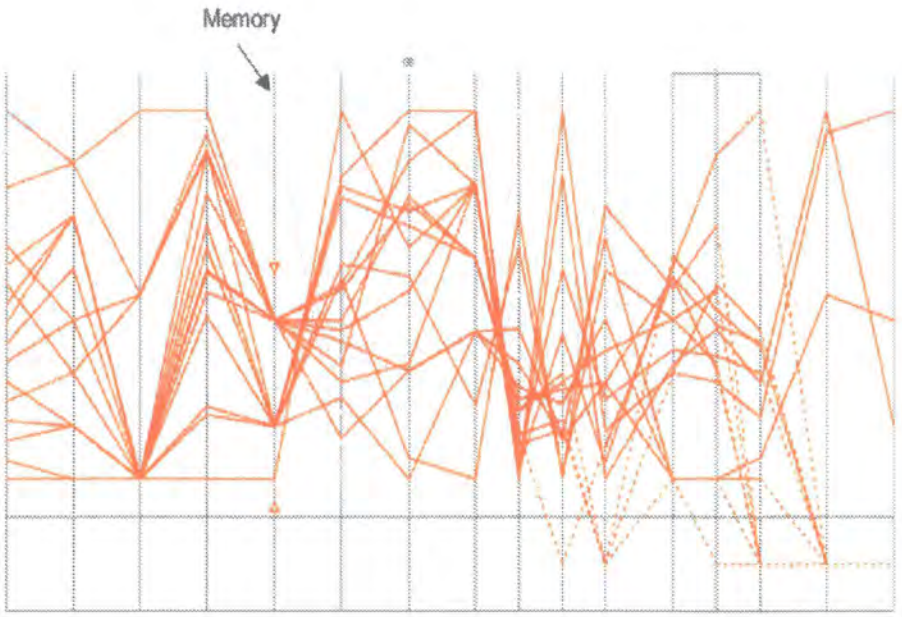


Figure 5.15 - Display showing a memory requirement introduced into the parallel coordinate representation

However, the extent to which the integrator's requirements can be translated into the display depends greatly on how well their requirements map onto the classification scheme of the broker.

Once introduced into the display, constraints also provide a powerful mechanism in identifying patterns and trends in the Parallel Coordinates representation. These could prove useful in the trade-off analysis of the integrator's different requirements.

It may be the goal of the integrator to identify components with good performance and quality values. These requirements could be introduced into the parallel coordinate display using a single query: Good performance rating 'AND' Good quality. The result of applying these constraints could be similar to that in Figure 5.16, where no components lie within these initial bounds.

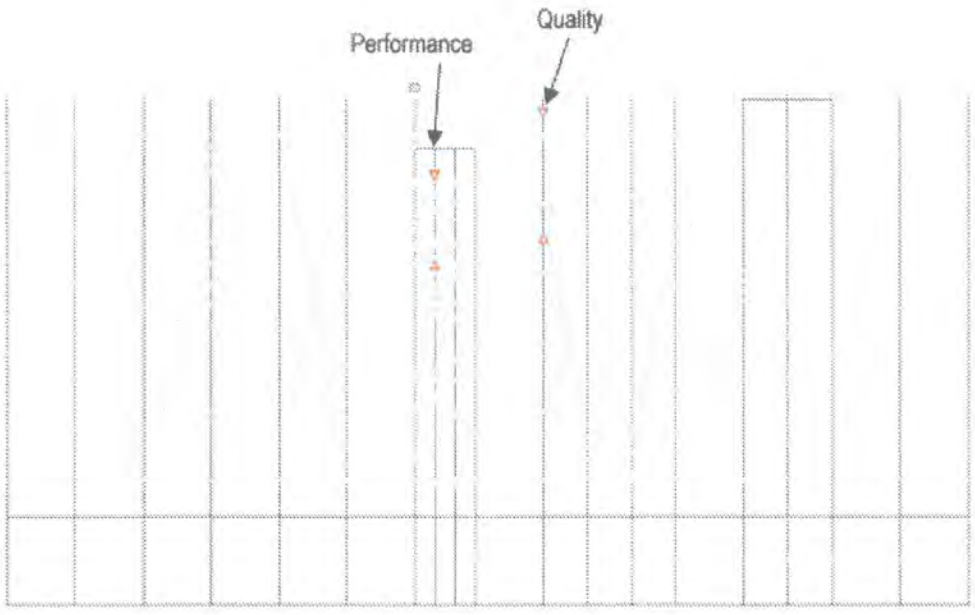


Figure 5.16 - Display showing the result of a search for components with high performance rating and quality values

Nevertheless, if these requirements are relatively open, the user can directly manipulate the constraints in the display until a number of components are highlighted, and the constraints still match their 'fuzzy' requirements. This 'fuzzy' introduction and manipulation of the requirements (as constraints) is well supported by the tool that not only displays the components in the highlighted set, but also all other components greyed out in the background. The integrator may have to perform several iterations of this kind to their requirements, before reaching a satisfactory conclusion.

Hopefully this process of introducing and manipulating constraints in the display will leave the integrator with one or more components highlighted in the tool that match their requirements. It is these components that are of specific interest to the integrator. At this point the tool provides two additional representations, enabling the integrator to view, compare and contrast the details of these components, providing a platform for the final selection of a component.

1. Data Table

This provides the user with access to the underlying data values of the parallel coordinate and star plot displays. By placing the values of selected components next

to each other in a table, this display provides a useful means of directly comparing the components (see Figure 5.17).

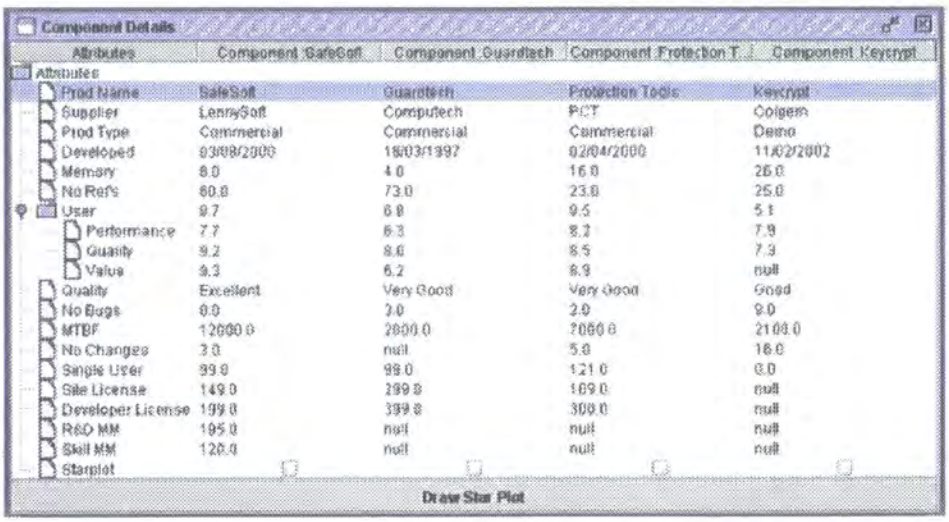


Figure 5.17 - The data table display

2. Star Plot

A visual display that would enable the integrator to quickly compare and contrast the different selected components. In this display the tool represents all the data values for each of the selected components, allowing the integrator to get a better overall view of the components and their properties (see Figure 5.18). By viewing and comparing the shapes defined by the components' data values, this display enables the integrator to select the component that best meets their requirements.

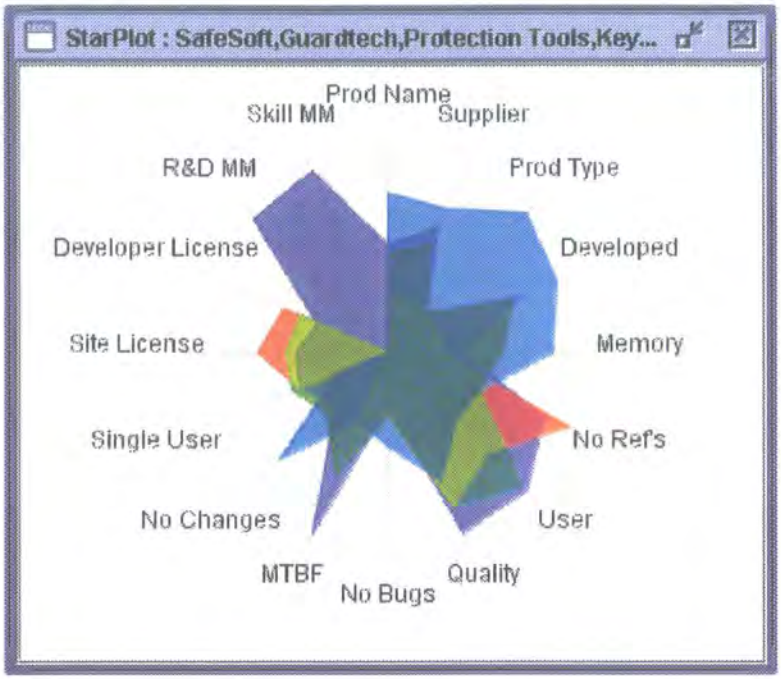


Figure 5.18 - The Starplot display

Based on this final comparison the integrator may decide to take their interest further by purchasing a component from the broker.

5.6 Conclusions

This case study has demonstrated how the implemented tool could function within the specified scenario of a component brokerage system, identifying how the tool would fit into the overall system and the processes involved.

The case study outlines how the tool's functionality provides support for the different users identified and their associated objectives. However, the case study does not describe a set of specific use cases defining a single situation with exact details of how the tool could be used; this does not match the exploratory nature of the tool. Instead it presents much more generalised cases that provide an insight into how the tool could be used in each case. The developed set of data provides a base for the discussion of these cases and an important source of examples to back up the explanations given.

It is clear from the case study that the tool does not provide definitive answers to all questions. In most cases the patterns and trends revealed by the tool provide only a starting point for further investigation, describing 'what could be', not 'what is'. Nevertheless, couldn't these same patterns and trends be identified using much more

basic displays such as a spreadsheet with searching and sorting facilities? The answer is yes, if enough effort is employed. The point of the tool is that it provides a means to access these details with much less effort and in a reduced period of time. The power of this tool, as with other visualisations, is the impact of the visual display. Individuals can look at this and quickly identify areas with interesting patterns and features. The tool also provides support for fuzzy searching and filtering strategies. Spreadsheets and similar tools provide support for only definite searches that cannot be manipulated. The interactivity afforded by the tool provides users with the means to manipulate these searches and filters until a satisfactory result is achieved.

This case study has gone some way to demonstrating how the features of the tool could be used within a 'real world' scenario. However, this tool has been developed with such a scenario in mind, therefore it is no surprise that it provides support for the different users, and especially the integrator, whose objectives were central to the tool's development. Nevertheless, it is believed that the tool's functionality is generic enough to cope with many different situations. To provide evidence of this, further case studies would need to be developed with scenarios not previously considered.

Chapter 6 Evaluation

6.1 Introduction

The aim of this chapter is to provide an evaluation of the tool outlined in chapter 4 and the functionality offered by this tool. Using a set of developed criteria, this evaluation hopes to present a review of the tool and the work undertaken in its development, providing a guide as to the progress made and an insight into opportunities for future work. Derived from the work undertaken and the associated goals, these criteria form the main basis for this evaluation. By considering each of these criteria, the chapter discusses the extent to which they have been addressed by the tool: outlining how and if they are addressed, and if they are not addressed, how the issues might be resolved.

Many strategies are used to evaluate visualisations, but to evaluate a visualisation effectively is an extremely difficult task. Hatch et al. [Hatch01] provide a summary of evaluation strategies used in the field of software visualisation, strategies that can also be applied to the more general field of visualisation. These strategies include the use of design guidelines, feature based frameworks, scenarios and walkthroughs, and user and empirical studies.

The approach taken in this evaluation is based on that used in feature based evaluations. This evaluation is composed of a nested framework of criteria against which the extent that these are fulfilled by the prototype tool and the visualisations it contains are discussed. Largely derived from the work of Bertin [Bert81, Bert93], Tufte [Tuft83, Tuft90, Tuft97], and Larkin and Simon [Lark87], these criteria aim to focus the evaluation on the visual representations used and not the processes involved in the use of the visualisations, or the visualisations suitability for any particular task. This method was chosen with the aim of provoking discussions to identify points that are fulfilled by the tool and its visualisations, along with those points that are not addressed and could therefore be the subject of further work on this project.

Empirical and user evaluation approaches were considered, however, these strategies require a great deal of planning and resources. The need to gain measurable values also means that many empirical and user evaluations focus on the processes involved in the use of the visualisations under scrutiny. For this reason,

these strategies could be seen of greatest benefit when comparing visualisations with the same or very similar functionality.

The criteria presented in this chapter incorporate issues at varying levels of abstraction, imposing a nested, tree-like structure onto the discussions. These issues focus on the developed tool and the external entities involved in its use. Data plays an important role in these considerations: the criteria themselves and the discussions are based on no specific data sets, only on the assumption that it is a multivariate data set of some kind. However, where necessary support for the discussions is provided, giving examples using theoretical data sets based on the Component Brokerage theme introduced by the case study presented in chapter 5.

In this evaluation the criteria have been split up into two groups:

1. Problem Specific Criteria
2. General Issues

Problems Specific Criteria focus on the task that has been addressed: the development of a visualisation for large sets of multivariate data. These criteria represent issues that affect the tool's ability to support this task and hence are key to evaluating the tool.

General Issues represent those issues external to the tool that nevertheless have an impact on the tool, its effectiveness and how it is used. The main sources for these issues are the external entities most involved in the tool and its use: the data, the user and the tasks performed. These issues have a much wider context than the specific objectives of the developed tool, and are points that must be considered when developing and evaluating all visualisations and other tools.

These two sets of criteria enable this evaluation to focus on different aspects of the tool whilst taking into consideration wider issues that may have some impact on the tool and its use. However, although the criteria have been split into two distinct groups, they are not completely independent of one another: many of the general issues described directly impact many of the problem specific criteria.

6.2 Problem Specific Criteria

The tool and the visualisations within it must be capable of dealing with large sets of multivariate data

This is an extremely broad point to make, nevertheless it could be considered as the ultimate objective of the developed tool, and thus has an important role to play in this evaluation.

In terms of this criterion, the tool does provide the means to visualise large sets of multivariate data within the Parallel Coordinates display. However, with such an open declaration it is difficult to make more complete and specific remarks on the tool's support for visualising large multivariate data sets.

To enable a more complete analysis of the tool's support for visualising large sets of multivariate data, this criterion has been broken down into two sub-criteria:

1. The tool must display the data
2. The tool must support the derivation of information from the underlying data

These points are key components that contribute to the tool's ability to visualise large sets of multivariate data, and thus provide the basis for further investigation.

1. The tool must display the data

This criterion relates closely to Tufte's comment that visual representations must above all "show the data" [Tuft83]. By "show the data" Tufte refers to the need to emphasise the data in a visualisation and not the method of representation.

The tool provides two visualisations that can be used to display multivariate data: the Parallel Coordinates and Starplot displays. However, once again to comment on the extent to which these visualisations support the display of large sets of multivariate data, this criterion must first be broken down into more specific and approachable points; breaking this criteria down further provides the basis for considering the different components and properties of a multivariate data set and how the tool and the visualisations it incorporates deal with these features.

This point has been broken down into the following:

1.1 The visualisation should enable the user to distinguish between all data values
1.2 The visualisation should be able to deal with and display data of various types
1.3 The visualisation should be able to support the possible existence of relationships within data sets

The Parallel Coordinates visualisation is the user's main point of interaction with the tool and the bulk of the tool's functionality is provided through this representation. For this reason, the discussions in this section focus their attentions on the Parallel Coordinates visualisation and its support for presenting the different facets of multivariate data sets.

1.1 The visualisation should enable the user to distinguish between all data values

To gain a more complete insight into a data set, it is important that users of a visualisation are provided with a clear view of all aspects of the data. Encoded as visual features in a visualisation, being able to distinguish these features is crucial to the user's ability to perceive and interpret the data represented.

The Parallel Coordinates visualisation contained within the tool displays all data values associated with each data item as a segmented line moving along the horizontal. The orientation of this line is dependent on the data values for the corresponding data item, and the position and scale of the axes used to plot these values. These axes appear evenly spaced across the display area, with the same vertical position and proportions, distinguishable from one another by the labelling provided. This powerful display format has the potential to support the visualisation of many data items with many variables and therefore many values, affording the user a view of all values through the use of the different axes and intersecting segmented lines.

The ability of the Parallel Coordinates visualisation to provide access to each data value is limited: it is possible that the values stored in a data set could produce a display where many of the data items presented are partially occluded by one

another. This could affect the user's ability to follow the path of the segmented lines representing each data item and thus identify its associated data values. Furthermore this could also result in the user misinterpreting the path of a data item's segmented line, potentially leading to false assumptions being made about the data set. The possibility of such problems arising is dependant upon the relative closeness of values associated with each variable. Data sets with large numbers of data items cause the resulting Parallel Coordinates display to appear crowded with segmented lines, and as the number of data items considered grows it becomes increasingly difficult to track the path of each data item. The number of data variables that can be represented effectively is also limited. The fixed size of the display permits only a certain number of axes to be represented within a single view without scrolling. This need to scroll around the display when the number of data variables is large has an impact on the user's ability to gain an overview of the entire data set, and thus make assumptions based on this view. By moving the axes closer together the need for scrolling can be greatly reduced, however the result of this is often a more crowded and difficult to interpret display.

The Parallel Coordinates visualisation incorporated in the tool provides a number of mechanisms aimed at overcoming some of these issues:

- The ability to highlight the path of any data item in the display
    - This provides the user with the ability to quickly identify any path and therefore the associated values for each data item in the display.

- The introduction of constraints
    - This allows the user to focus their attention on a subset of the data items, filtering out much of the complexity of the original display, and highlighting the items of specific interest, making them much easier to identify and trace.

- The ability to zoom in on a range of values on an axis
    - This allows the user to zoom into a specific range of values on an axis, providing the ability to investigate values appearing close together within this range more effectively.

- The ability to zoom in on the entire representation
    - This allows the user to spread out all the features of the display, providing a means of resolving many of the problems with overcrowding. However, zooming in on the representation also increases the size of its display area, potentially leading to the need for increased scrolling in order to traverse the display in its entirety.

It can also prove helpful in the Parallel Coordinates visualisation to provide an axis where each data item has a unique value. This ensures that each data item is distinguishable from all others by at least one segment of its associated line. It is clear that the Parallel Coordinates visualisation and functionality provided by the prototype tool support the ability to deal effectively with relatively large sets of multivariate data. Nevertheless, this visualisation scheme as with many others can become less effective when the data set under consideration is much larger. Other methods based on the results of some kind of pre-processing on the data could help resolve these issues with scale, introducing abstraction mechanisms and reducing the amount of complexity within any single display.

1.2 The visualisation should be able to deal with and display data of various types

Data can exist in many different forms and it is important that a visualisation makes best use of the available data to help benefit the user and fulfil their information needs. No data should be ignored, even if it cannot be transformed and taken into account within the visualisation, it should be made available to the user directly via other means.

The Parallel Coordinates visualisation provides support for the 3 basic types of textual data: qualitative, quantitative and ordinal. It is assumed that each variable in a multivariate data set can contain data of only one of these types. Nevertheless, each variable is displayed as an axis in the Parallel Coordinates representation, distinguishable from one another by their associated label. The scale of each axis is calculated based on the type and values associated with the corresponding variable in the data set; any scale present within the data is preserved by the axis within the

Parallel Coordinates display. The Parallel Coordinates visualisation also incorporates mechanisms that support the display of missing/corrupt data values. In many visualisations these values are often ignored and removed from consideration, even though they may provide valuable insight into the data set.

The display scheme utilised by the Parallel Coordinates visualisation imposes a linear ordering onto all data it displays. For this reason when dealing with qualitative and ordinal data, the visualisation introduces a certain amount of false information into the display that may lead to false inferences being drawn about the data.

Qualitative data has no inherent ordering, but the Parallel Coordinates display imposes an order onto the data to enable it to be displayed on an axis. This provides the possibility for false inferences to be made based, not only on the ordering introduced, but also the proximity of the values to one another.

Examples:

- Values appearing higher on an axis may be considered greater in value than those lower down
- Values close to one another on an axis may be considered as being similar or close to each other

Ordinal data already has an ordering, but does not provide any scaling details. The Parallel Coordinates display imposes a scale, as with qualitative data types, by spacing the values evenly along the variables corresponding axis. This imposed scale could once again form the basis for false inferences about the relationship between the values and the differences between them.

This method introduces the possibility of false inferences being made about the data displayed. However, the tool relies on the investigative qualities of the user to explore the other details of the display, identifying the context for the data displayed (i.e. its type).

The work done in the development of the tool and its visualisations makes the assumption that the data provided is in a textual form. It could be the case that a data set contains data other than text, such as pictures. However, this is considered

outside the scope of the work undertaken in this thesis and represents an extensive research area in its own right.

1.3 The visualisation should be able to support the possible existence of relationships within data sets

It may be the case that a data set includes data about itself, so called 'meta-data'. This meta-data could provide details of relationships existing between the values in the data set. In terms of multivariate data sets, meta-data is already provided associating the data values with the different variables and data items. However, this same meta-data could be used to describe more complex relationships in the data set, and these relationships, if presented in a visualisation could greatly alter the insight a user gains into the data.

The Parallel Coordinates display supports the existence of three distinct types of relationship between the data variables of a multivariate data set identified and described in Chapter 4 (see section 4.4.2): Conceptual, Compound and Parent-Child. As demonstrated in the case study outlined in chapter 5, these relationships can have a great and beneficial impact on investigations into data ands its properties. However, providing support for the existence of such relationships within the Parallel Coordinates display introduces increased complexity into the representation that may ultimately hinder and confuse the user: potentially requiring axes, and hence its associated data, to appear more than once in the display.

Currently the Parallel Coordinates visualisation developed for the tool provides support for three identified types of relationship between the data variables of a data set. This represents an important step forward in acknowledging the possible existence of meta-data other than that defining the data items and variables within a multivariate data set. However, it may be the case that data sets include meta-data outlining relationships that cannot be incorporated into the display, not only between the different data variables, but also between the data items of a data set.

For example:

- A data set containing details of software components may also contain details of relationships between these components. One such relationship could be

used to indicate that a software component (data item) in the data set is composed of one or more different software components (data items) also appearing in the data set.

The tool does not address this kind of 'data item-data item' relationship, but these relationships could also have a dramatic impact on the user of the system and their understanding of the data. Ultimately, however, the ability to take advantage of a visualisations support for the existence of such 'meta-data' is dependent on this data being present in the first place.

2. The tool must support the derivation of information from the underlying data

This criterion represents the objective of any visualisation: to help extract information from a data set. Visualisations hope that by presenting data visually and providing mechanisms for interaction with the display that useful information can be derived more quickly and easily.

Once again this is a very broad issue and there are many factors that contribute to this discussion: the different types of information that can be derived, the different processes involved, and the extent of the support for these processes. This section breaks up the point into sub-issues, providing the basis for a much more thorough investigation of the tool's support for deriving information from the underlying data set.

This point has been broken down into the following:

2.1 The visualisation should support comparisons between associated data values
2.2 The visualisation should support the identification of patterns and trends in the data
2.3 The visualisation should provide a link between the visual representation and the underlying data set
2.4 The visualisation should provide the means of identifying data with specific properties of interest

2.1 The visualisation should support comparisons between associated data values

An important part of any visual representation of data is its support for the visual comparison of displayed values. It is this comparison between the displayed values that forms the basis for all information derived from the display. Visual representations have the power to encourage such comparisons by providing fast and easy access to data values most likely to be compared to one another. This relates closely to the comments made by Larkin and Simon [Lark87] (also see chapter section 3.2.4) on how grouping data that is related or used together improves access and encourages comparison.

The Parallel Coordinates visualisation is constructed based on grouping associated data values:

1. Grouping all values associated with a data item in the form of the item's corresponding segmented line.
2. Grouping all values associated with a data variable on the variable's corresponding axis.

These features make it easy to access data associated with each data item and each data variable. By grouping values in this way the Parallel Coordinates representation also encourages comparisons to be made between:

- The different data items, comparing their values for each data variable
- The different data variables, comparing associated values for each data item

These comparisons provide the basis for identifying trends and patterns within the data set represented.

The Parallel Coordinates display encourages comparisons to be made between those data values appearing close together in the display, but it may be the case that the user wishes to compare values appearing much further apart.

For Example:

Compare the values on an axis at one end of the Parallel Coordinates display to those on an axis at the other end of the display

In this case it is much more difficult to directly compare the values appearing on each of the axes. The introduction and movement of constraints in the Parallel Coordinates visualisation can help overcome some of these difficulties, highlighting sets of specific interest, enabling values distributed over greater distances in the display to be compared more easily (see Chapter 5 for examples). It would also be possible to develop the Parallel Coordinates visualisation with increased interaction: allowing users to move axes around to suit their comparison needs.

2.2 The visualisation should support the identification of patterns and trends in the data

The identification of patterns and trends within a visualisation provide a valuable insight into the data set under scrutiny, allowing information to be inferred. Therefore the identification of patterns and trends within a visualisation must be considered in order to comment on the extent to which the tool supports the derivation of information.

The Parallel Coordinates visualisation within the prototype tool provides extensive support for the identification of patterns and trends in the underlying data set. By presenting data in the form of segmented lines moving across the horizontal, the display reveals patterns and trends through the position and orientation of these lines.

For Example:

> Correlations between the values of two data variables may be identified by looking at the general trend followed by data item lines moving from one variable to the other.

The visualisation also supports the identification of more complex patterns and trends through the introduction and movement of constraints. The constraints allow the user to focus attention on subsets of data, which are of some specific interest, providing the means to reveal interesting patterns not visible when considering the data set as a whole. Movement of these constraints and the resultant changes in the

display can also help to emphasise the existence of patterns that could otherwise be missed.

The patterns and trends identified in the display provide only an insight into the information inherent in the data. They do not provide definitive results, only assumptions that might be investigated further. The tool's Parallel Coordinates display provides methods for the identification of patterns and trends, but does not include any other functionality that enables the data and the potential patterns contained within to be analysed further. In later versions of the tool, it would be possible to incorporate statistical calculations supporting further analysis of data that could potentially provide details of patterns not immediately visible in the display.

2.3 The visualisation should provide a link between the visual representation and the underlying data set

A visualisation without any link to the underlying data set provides no way for the user to relate what they see in the visualisation to the data. The visualisation may show the existence of many patterns, but without a link to the data the user will be unable to understand the meaning these patterns.

The Parallel Coordinates visualisation does not provide direct access to the underlying data values from which it is constructed. The labels for the axes provide some indication that the display is more than just a collection of intersecting and adjoining lines, but no other details of the underlying data are provided. Methods for providing access to the underlying data values directly within the Parallel Coordinates implementation have been considered, but for reasons of clutter they were not included within the tool. Nevertheless, the tool does provide access to the underlying data values on the request of the user within a separate table display: by highlighting data items in the display the user can choose to view the data values for the item within a separate table display. This method provides the crucial link between the visual representations presented by the tool and the underlying data set, without overcomplicating and cluttering the displays.

2.4 The visualisation should provide the means of identifying data with specific properties of interest

The ability to identify data with specific properties in a data set can be considered as another form of information.

For example:

Discovering the set of software components that will function with less that 16Mb of memory

The Parallel Coordinates visualisation supports this process by enabling the user to introduce constraints into the display, filtering the data. Constraints can be used to introduce predetermined filters into the display, enabling the user to identify and focus their attention upon data items fulfilling their specified requirements. The Parallel Coordinates visualisation also provides support for the following type of scenario:

- A user wishes to identify a set of components with a price less than £100, but they do not wish to remove those components with missing/corrupt price values from consideration, in case on further investigation one of these components proves to be the 'best' and within budget.

It does so by allowing the user to introduce a constraint composed of two ranges, one encapsulating the missing/corrupt data values and the other containing all values less that £100. The Parallel Coordinates visualisation also offers support for filtering the data based on less than definite, 'fuzzy' specifications. Constraints can be introduced into the display and then altered dynamically, taking into account the data items currently outside the filter (greyed out in the display) and the set of data items within the filter until the results in the display satisfy the user. In terms of the tool's support for filtering the data based on 'fuzzy' specifications, the tool indicates when a data item is and is not within the defined constraints, but it provides no indication as to how far the data items lie outside the constraints.

For Example:

A User may wish to identify components with good performance and quality values that functions with less than 16Mb of memory.

The introduction of such constraints into the Parallel Coordinates visualisation may result in a number of data items lying within the defined constraints to become highlighted. However, the tool provides no indication as to how far the remainder of the data items are from satisfying the specified constraints.

6.3 General Issues

6.3.1 Data

Visualisation as well as other research areas such as data mining developed to try and take advantage of the increasing volumes of data being produced and stored. The ultimate goal of any visualisation is to provide some 'useful' insight into the data set under investigation. However, the properties and features of this data have a great impact on the visualisation and how it is used.

First and foremost, for the tool to be able to display the data, it must be in the correct format as outlined in appendix B. This requirement quickly reduces the amount of data available for investigation using the tool. Multivariate data sets may be commonplace but not specifically in the format used by the tool, and even if the data could be processed into a compatible form it would take some effort to do so. It is also expected that even fewer data sets would contain meta-data indicating relationships between the different data variables of the type used by the tool. One approach to making more data available for use with the tool would be to change the file format to a valid form of XML. This would enable data specified in other XML formats to be transformed automatically into the format required by the tool.

The 'quality' of the available data will also have a great impact on the effectiveness of the tool, and users' opportunities to investigate the data set. For the purposes of this discussion, the 'quality' of a data set is a measure of how well the data matches the requirements of the tool and users' tasks. The Parallel Coordinates visualisation is best suited to the display of quantitative data, therefore it would be most effective with data sets containing more values of this type and fewer qualitative and ordinal values. Other problems with the quality of a data set can

originate from the way in which it is collected and formatted. However, use of a visualisation or other analysis method could help highlight such problems with a data set. An example of this would be the appearance of many missing data values within a large multivariate data set, a likely scenario when many of the variables are only applicable to a small number of the data items.

Data may exist in many formats, both textual and pictorial. The developed tool is able to support and display 3 basic types of textual data: qualitative, quantitative and ordinal. It is also capable of dealing meta-data outlining relationships between the different data variables. However, many other relationships may exist that are not taken into account, and the appearance of such meta-data in a data set may be rare.

There is no question that data is central to the visualisation process, and has a great impact on visualisations and their effectiveness. The developed tool and its Parallel Coordinates visualisation is no exception. In an era where increased value is being assigned to the massive amounts of data being produced and stored, emphasis is being given to techniques that can provide insight into this data. The processes and methods used in the collection and organisation of data are the primary drivers that govern what processes (such as visualisation) may be applied to the resultant data set, but all too often these methods are considered only as an afterthought.

6.3.2 Usability

Visualisation is very much a human driven activity; the human role is central to the investigation of data, but also to the tool and how it is utilised. However, the heavy involvement of a human user in these processes introduces many unknowns based on differences between all individuals:

- Abilities
- Cultural
- Conventions

The use of colour can have a great impact on an individual, based not only on their ability to perceive colour, but on both their personal preference and cultural beliefs. To avoid many of the problems associated with colour use, the bulk of the tool's displays are presented in black and white, leaving the selection of other display colours to the user's preference.

In terms of conventions, the Parallel Coordinates representation contained within the tool presents quantitative data with values moving from low to high as you move up the variable's corresponding axis. Within the tool's Starplot display a similar convention is used displaying lower values closer the representations origin. However, these are quite common conventions, which should not be confusing to the users in anyway.

In an number of ways the tool attempts to take advantage of user familiarity with other window based programs: presenting each display within windows, providing menu options at the head of the main window and an accompanying Toolbar. By providing such a familiar environment the tool hopes to enable the user to focus on the more complex features included in the tool and its interactive displays. The zoom feature provided for the Parallel Coordinates representation is one that a user may expect to find in such a tool and provides some important functionality. It is however implemented in a way that may not be familiar with the user. The zoom functionality provided enlarges the display and resets the viewing position of the user to the top left corner of the representations display area. It does not allow the user to select an area in the representation on which to apply the zoom directly as within other packages. Further work could be done extending the functionality to match this and other user expectations, but the main point of the developed tool is as a proof of concept, and inclusion of such 'polished' features would not in keeping with this purpose.

One of the tool's greatest and most important features is the interactivity afforded by the Parallel Coordinates display it contains. This interactivity allows the user to actively investigate the data set presented, identifying potential patterns and trends to be considered further. The user is able to interact directly with the display, adding, removing and changing constraints, causing the display to update showing the impact of these changes. Nonetheless, the effectiveness of such interaction mechanisms is reliant on the display being able to update within a short time of the users' input. This requirement allows the user to associate the updates they see in the display with their previous actions. However, the ability to fulfil this requirement is largely dependent on the speed of the program; when dealing with very large data sets, the number of calculations required may not allow the display to update itself quickly enough. This in turn could cause the user to become confused, unable to associate their action with the results (eventually) displayed.

Usability is a hard property to comment upon and measure. Empirical user tests provide one avenue with which to explore the usability of any tool. However through such tests it is difficult to gain objective results that can be measured and compared with other results. The nature of the developed tool: 'a proof of concept' also leads to the question of how important the issue of usability should be. Is it something that can be taken into account later, when fine-tuning a tool for a commercial environment? Or, as the user is central to the use of any visualisation, should their needs be considered at a much earlier date?

6.4 Conclusions

This evaluation represents a review of the progress made in the development of the prototype tool outlined in chapter 4. Using a set of criteria representing problem specific and more generic issues, this chapter provides a discussion as to the extent to which these points are taken into account and addressed by the tool.

Focusing for the most part on the Parallel Coordinates display provided by the tool, the evaluation outlines its support for visualising large sets of multivariate data: it recognises the visualisations support for the display of different forms of textual data and its support for the identification of patterns and trend within the data. However, issues that the current tool does not address are also introduced and details of possible avenues for future developments are provided.

By taking into account wider issues, the chapter outlines the great impact that both data and users have on a visualisation. It identifies how the effectiveness of a visualisation is dependent on the data and its quality. The evaluation also considers the many unknowns a user introduces into the investigation of data using a visualisation, describing how these points are considered, if at all, by the tool.

In terms of an evaluation, this chapter does provide an indication of the progress made by the work undertaken in the development of the tool and its Parallel Coordinates visualisation. It provides details of what the work has achieved, along with points that the work has not addressed and potential areas for the future development of the work. Nevertheless, this entire evaluation is a collection of subjective discussions based on the developed set of criteria. It does not provide any measurable values that could be used to compare the tool and its Parallel Coordinates visualisation to other tools and visualisations. However, evaluating visualisations

effectively is a difficult task - a research topic in its own right - and the development of such meaningful values is not easy.

Chapter 7 Conclusions

7.1 Introduction

The aim of this chapter is to provide a review of the work detailed within this thesis. It hopes to provide an indication as to the progress made by work undertaken, and how the work contributes and fits into the wider scopes of visualisation and other information related research.

The chapter begins by providing an overview of the work completed: research, development and evaluation. The progress of this work is then discussed using the initial criteria for success set out in chapter 1. Based on these discussions and the work contained in each of the previous chapters, opportunities to extend this work are identified and a conclusion based on these findings is presented.

7.2 Overview of Study

The role of information in modern society is growing in importance, and the terms 'information age' and 'information society' are now commonly used to describe this period of time and the nature of our society respectively. The growing significance assigned to 'information' has brought about a growing amount of interest in, and discussion of associated subject matters.

This thesis is part of this growing trend towards information related research. Nevertheless, the meaning associated to the term information in this research is often confused. For this reason the thesis begins by outlining the interpretation of 'information' and related terms assumed by the work presented, providing a useful frame of reference supporting discussions and removing ambiguity in the later use of these terms. Based on these foundations, the chapter presents a survey of information, outlining how it is becoming an integral part of our lives and our 'information society'. Discussing the impact of becoming an 'information society', it outlines the potential benefits and problems society faces, and how the requirements placed on citizens of this society are changing. Emphasis is placed on the problem of 'information overload', and the chapter introduces various approaches that are looking to resolve this problem and associated issues.

Narrowing the focus of this research, chapter 3 presents visualisation as an approach to resolving some of the issues associated with 'information overload'. It

introduces the field of visualisation research, providing background to its development and the theories that underpin the work. Narrowing the focus of this research still further, the chapter outlines the more specialised problems associated with the development of visualisations for large multivariate and abstract data sets. The chapter presents an overview of existing visualisation techniques that look to resolve these issues with the aim of demonstrating the diversity present within this research area, as well as recognising the relative strengths and weaknesses of each technique.

Chapter 4 outlines the details of a prototype tool developed based on the findings and principles identified in the previous chapters. First the chapter re-emphasises the problems that the tool aims to address and outlines the approach taken in the development of the prototype, including the rationale behind these decisions. The chapter then provides details of the tool and the functionality it offers, and identifies how the tool could be developed further in the future.

With the aim of demonstrating how the developed tool could function within a real world situation, chapter 5 presents a case study that explores how the tool could be employed within a component brokerage system (see chapter 5). Based on the developed scenario and identified roles within this scenario, details are provided of how the tool could be used to benefit these different roles, making use of examples to demonstrate the points made.

Chapter 6 presents an evaluation of the tool, aiming to measure the progress made by the work undertaken in its development. Using a set of criteria, the chapter discusses the extent of the tool's support for visualising large sets of multivariate data, but also outlines how wider issues such as data and usability have been considered. In doing so the evaluation also highlights points that the tool does not take into full consideration, and goes on to propose how these points may be approached differently in future work.

7.3 Criteria for Success

7.3.1 Introduction

The main point of this work was to examine how visualisation research can help overcome some of the problems associated with 'information overload', focusing on the specific problems faced when developing visualisations for large sets of

multivariate data. Based on the findings from this research, the work aimed to develop and demonstrate a tool capable of visualising and providing useful insights into large multivariate data sets.

Chapter 1 outlines four criteria for success for this project:

1. To Investigate visualisations capable of presenting large amounts of multivariate data
2. To develop a prototype tool to demonstrate
3. To explore the use of these visualisations in a component brokerage system
4. To develop visualisation mechanisms to handle relationships between data sets

Revisiting each of these points, the following sections 7.3.2 to 7.3.5 discuss the extent to which these criteria have been met, and thus provide some indication as to the progress and success of this project relative to these points.

7.3.2 Criterion 1 - To Investigate visualisations capable of presenting large amounts of multivariate data

By providing an overview of existing techniques and strategies used to visualise large sets of multivariate data, including details as to how each is considered to succeed and fail, the hope is to demonstrate the broad nature of the approaches used, but also to form support for the later development of a prototype visualisation based on its findings.

Chapter 3 presents an overview of existing visualisation techniques providing details of their construction and the relative advantages and disadvantages associated with their use. This overview presents a number of visualisation techniques most of which provide means of presenting large multivariate data sets. Based on a taxonomy of visualisations developed by Kiem, the overview also aims to demonstrate the broad range of approaches that have developed to deal with the problems associated with visualising data sets. Also included in section 3.4.2 of this chapter is an outline of some of the generalised approaches employed to deal with the specialised problems faced when visualising multivariate data.

The overview provided in chapter 3 in no way represents a complete review of the techniques that have developed to visualise large sets of multivariate data; so many

techniques to approaching this problem have developed and continue to emerge in this fledgling research area of visualisation. The overview given instead provides a flavour of the variety present in the techniques that have developed, and an insight into the effectiveness of these techniques.

7.3.3 Criterion 2 - To develop a prototype tool to demonstrate

The development of a prototype tool hopes to provide the ability to showcase and test new concepts and ideas. Based on the findings of the research, the tool aims to incorporate proven ideas, whilst at the same time developing new and novel methods of addressing those points identified in the research as requiring further attention.

Chapter 4 contains the details of a prototype tool developed as part of the work contained in this thesis. It provides details of: the problems the tool looks to address, the approaches taken, and the novel concepts developed. The tool itself provides the ability to visualise large sets of multivariate data specified in the format required by the tool. It also provides the opportunity to test the new concepts incorporated in the tool and the tool's ability to deal with these large sets of data.

However, the tool itself does not provide any indication as to the extent of its support for visualising large sets of multivariate data and the potential success of the features incorporated. This can only be achieved through the implementation of tests and evaluations, and the implemented prototype provides the basis for such activities to take place.

7.3.4 Criterion 3 - To explore the use of the tool in a component brokerage system

By placing the tool within a theoretical 'real world' scenario, the hope is to demonstrate how the tool might be used and provide an insight into the potential benefits of using the tool.

Chapter 5 describes a component brokerage scenario, identifying the different roles in such a scenario and their associated objectives. It then demonstrates how the developed tool could be used to support the different roles in attaining their objectives, using examples to support these comments based on a 'realistic' data set; a data set constructed with the aim of demonstrating the full potential of the tool.

The case study contained in chapter 5 clearly demonstrates how the developed tool could function within a component brokerage system: demonstrating how it could be

used and giving details of the potential benefits it would provide to all human roles in this scenario.

7.3.5 Criterion 4 - To develop visualisation mechanisms to handle relationships within data sets

Based on the research undertaken, it is clear that the possibility of relationships existing between the data values of a data set is for the most part ignored. Nevertheless, details of these relationships included within a data set could, if available to an individual, greatly alter the insight they gain into the data under scrutiny. Within multivariate data sets basic relationships between data values give the data structure, defining details of the different data items and data variables. When taken into account, these details allow the set to be investigated to a much greater and useful extent.

Chapter four identifies the possibility of three different relationships that could exist between the data variables of a multivariate data set: Compound, Conceptual and Parent-Child (see Chapter 4 - section 4.4.2). The chapter outlines the meaning associated with each of these relationships and how each has been incorporated into the prototype's Parallel Coordinates display.

However, this work does not consider the possible existence of relationships between the different data items within a multivariate data set. Further investigation of data sets may also identify relationships between data variable that cannot be categorised as any one of the three defined types. The impact of including these relationships into a visualisation must also be considered more greatly because as you make more information available to the user, the display inevitably becomes more complex and more difficult to understand.

7.4 Future Work

The work in this thesis represents an effort to identify and resolve some of the problems involved in developing visualisations for large sets of multivariate data, with a view to tackling the issue of information overload.

However, even considering the reduced set of problems this thesis focuses upon, there is definite scope to further the work in all areas:

- Research
- Development

- Evaluation

The work in this thesis could be viewed as a starting point for the development of further work. Details of possible further developments have already been mentioned briefly within the relevant chapters, extensions that could in the future form projects in their own right.

7.4.1 Research

There is definite scope to extend the research undertaken as part of this thesis. The topic of visualisation and the problem of information overload are vast research areas that this thesis only scratches the surface of. In terms of visualisation research – the main research component in this thesis – this could be extended in a number of ways:

1. Investigating points of specific interest in greater depth
2. Broadening the research

The thesis concentrates on the problems associated with the development of visualisations for large sets of multivariate data. By presenting a review of existing techniques using a taxnonomy, it emphasises and the broad range of ideas and techniques that have developed. However, further work could be done reviewing a greater range of existing visualisation techniques capable of displaying large multivariate data sets. This would not only provide an invaluable source of ideas on how to extend or develop visualisations, but also a set of visualisations that could be used in comparisons with the prototype's Parallel Coordinates representation. The work also concentrates on the lack of consideration given to missing/corrupt data values and the potential existence of relationships between data variables; identifying the benefits of incorporating these details into a representation and how they can be incorporated into visualisations. Further research could take place looking at the relationships existing between data variables, the possibility of relationships between data items, and the potential impact of introducing these into a visualisation. Other methods of displaying missing/corrupt data and the relationships between the data variables could also be identified and investigated, providing alternatives and points of comparison with those already implemented.

The broad extent of visualisation research is emphasised by the work detailed in this thesis: outlining how resources from many different research areas are being incorporated into visualisation. These different areas contribute to the field of research and hence provide alternative paths through which this work could be extended. Research could be undertaken looking to investigate how this work may be extended to take advantage of three-dimensional representations and the corresponding display and interaction technology: it could investigate how this work may be extended to take advantage of three-dimensional representations and the corresponding display and interaction technology. Perhaps more significantly this research – and visualisation research in general – could focus on human abilities and how best the harness them. This movement into the arena of the social sciences, investigating humans and their behaviour, may provide details enabling alternative evaluation strategies to be explored and tested; a key component for the advancement of visualisation research.

The scope of research that could be undertaken under the banner of visualisation is seemingly endless and it is easy to get overwhelmed by this boundlessness. It is only possible to focus effort on specific subsets of the problems approached by visualisation research, and there are always opportunities for further work and expansion.

7.4.2 Development

The tool developed and the concepts contained within could also be the subject of further work. In terms of the tool that has been developed, work could be done adding additional functionality, modifying the existing functionality and polishing the tool – making it ready for commercial use. Points that further implementation on the tool may consider include: -

- The display and access of the underlying data values within the Parallel Coordinates display.
- The development of support for the task of constructing queries and understanding queries represented in the tool's Parallel Coordinates display.
- The update of the tool's zoom functionality. The current implementation contains only the most primitive zooming capabilities that can disorientate.

- A change in file format to valid XML. This would potentially allow other multivariate data sets specified in XML to transformed automatically into the required format.
- A change in the methods used to display labels in both the parallel coordinate and Starplot displays. This would aim to make labels easier to access and interpret, as well as making the display more aesthetically pleasing.
- Increase in the tool's support for the display of the different relationships between data variables. This would enable the tool to deal with all possible combinations of the three defined relationships, not only within the Parallel Coordinates display but also possibly within the Starplot display.
- Adding more interaction into the Parallel Coordinates display, enabling axes to be moved and removed based on user preference.

The prototype tool is an extensive implementation that includes a large amount of functionality, but there is definite scope to develop it further. However, some of these possible additions would go against the principle of the tool. The tool was developed to help test the concepts contained within, it was not meant to be a fully marketable final implementation. Also, if any further development work is to be undertaken, it should only take place based on the results of some initial investigation or evaluation of the prototype tool.

7.4.3 Evaluation

If further work were to be undertaken then a more in-depth evaluation of the tool and the concepts it presents would be a good starting point. The case study presented provides extensive evidence for the tool's application to 'real world' problems. The evaluation that follows in chapter 5 provides some indication as to the progress made by the work. It presents discussions reviewing the work with respect to a set of criteria, identifying the extent of the tool's support for visualising large sets of multivariate data, but also the extent to which it considers wider issues such as data and usability. This provides an extensive review of the prototype tool and the features it incorporates. However, it does not provide an objective foundation on which to extend this work or base future work upon. Only by putting in place a more complete and thorough evaluation can judgements be made on the tool and its constituent components, and progress be made.

A more complete evaluation would endeavour to gain measurable values from the tool, enabling a comparison to be made with other existing visualisation tools and techniques. These values could be gained by taking direct physical measures from the tool and the representations contained within, but also through empirical testing. These values and the results of the comparisons would support the identification of potential strengths and weaknesses in the tool, enabling strategies to be formed to take this work further. However, other than with the use of empirical testing it is very difficult to derive useful, measurable values from visualisations and the tools that contain them. For this reason evaluating visualisations effectively is an extremely difficult and challenging task, and is a research topic in its own right.

There is no question that this work could be furthered in all areas of the project. The work presented here is only the beginning, and it could be argued that until visualisation becomes more of a science, all work will have a similar fait. The development of evaluation techniques has a critical role to play in supporting the progression of visualisation research, as it is only through analysing results of evaluations that more solid guidelines for the development of visualisations can emerge. However, the way forward for this evaluation effort is not clear. It may involve the incorporation of a still broader range of ideas and theories, including those from social sciences and other human focused studies. It is the unpredictability introduced into visualisation by the large amount of human involvement that makes it so difficult to be truly scientific.

7.5 Conclusions

The work in this thesis represents an effort to identify and resolve some of the many problems inherent in the development of visualisations for large sets of multivariate data. Research detailed in the early chapters of this thesis provide background to the problem of information overload, visualisation as an approach to resolving some the problems associated with information overload, and the more specialised problems faced when developing visualisations for large sets of multivariate data. The prototype tool developed is an attempt to address some of the findings from the research undertaken: making use of proven techniques, and identifying and addressing points overlooked by the majority of the research. Focusing on this tool and the functionality it affords, the thesis provides a case study demonstrating how

the tool could be used, and an evaluation outlining the extent of the tool's support for visualising large sets of multivariate data and wider issues such as data and usability.

In terms of visualisation research as a whole, this work represents an attempt to resolve some of the problems inherent in the visualisation of large sets of multivariate data. The work also identifies and looks to address the lack of attention given to missing/corrupt data values and the existence of possible relationships between data variables in a multivariate data set. However, as with many other visualisation projects, the success of this work is difficult to gauge.

Visualisation research is in a state of paralysis, unable to move or develop; work continues to engage the same problems but with little or no progress. Few guidelines exist to aid in the development of visualisations, and those that do, exist only as rules of thumb, the majority of which were not developed with visualisation in mind, some conceived well before the birth of visualisation research. For this reason the development of a visualisation still remains more of an art than a science, a hit or miss affair where progress is stumbled upon and not planned. In order for the discipline to become more scientific, there is a need to develop more concrete strategies to help govern the development of visualisations. Critical to the development of these strategies is the ability to recognise what a good visualisation is composed of, and effective evaluation is key to this. However, the evaluation of a visualisation is far from an easy task, and is only just beginning to gain the attention it deserves from the research community. One of the greatest problems the development of evaluations face is the sheer diversity present in the visualisations produced. For this reason it is unlikely that a single set of measurable and comparable points could be developed to describe the properties all good visualisations should have: two dimensional, three dimensional, animated, etc. Nevertheless, it is clear that the development of evaluation methods for visualisation has a key role to play in taking the research area forward; moving it from the art form as it exists today, into the scientific arena.

Appendix A - Prototype Tool's File Structure

Text-Based File

Text-based files provide the program with the data used to construct an initial parallel co-ordinate plot. The file uses tag-based identifiers to define the data used to build a representative plot.

File Format

The text-based file format can currently be seen to break up into two discrete sections: -

1. [Axis/NonAxis Definitions]

This part of the file defines all component attributes; these attributes may or may not be presentable as an axis in the final plot. To provide the program with the ability to display the inter-relationships between the attributes, this part of the file is also responsible for defining these relationships.

i) <Axis,'Name','Type'> .... </Axis>

Used to define a component attribute to be plotted as an axis within the parallel co-ordinate plot. The definition must include the name of the attribute and the type of the data expected for the attribute (e.g. Date). If the type specified is 'Ordinal' then the user must also include information specifying the ordinal types possible values, and their ordering.

Different Types: Nominal, String, Ordinal, Numeric and Date

Examples:

<Axis,Unit Price,Numeric> .... </Axis>

<Axis,Quality,Ordinal,Very Bad,Bad,OK,Good,Very Good,Excellent> .... </Axis>

ii) <SubAxis,'Name','Type'> .... </SubAxis>

This tag provides the ability to develop a nested structure of attributes, allowing parent-child relationships between attributes to be specified.

Example:

<SubAxis,Developed,Date> .... </SubAxis>

iii) <NonAxis,'Name'> .... </NonAxis>

Used to define a component attribute that is not represented as an axis within the parallel co-ordinate plot, but who's values will nevertheless be made available to the user of the program. As the program does not plot these nonaxis values the type of the values should not be specified.

Example:

<NonAxis,Description> .... </NonAxis>

iv) <SubNonAxis,'Name> .... </SubNonAxis>

As with the 'SubNonAxis' tag this provides the ability to define a nested attribute structure involving 'nonaxis' attributes.

Example:

<SubNonAxis,Bug Report> .... </SubNonAxis>

v) <Group,'Name'> .... </Group>

Used to define a conceptual group. All attributes related to each other by a conceptual relation are placed within the Group tags.

Example:

```
<Group,Quality>
        <Axis,Mean Time Before Failure,Numeric>
        </Axis>
        <Axis,Number of Reported Errors,Numeric>
        </Axis>
        <Axis,Changes in Release,Numeric>
        </Axis>
        <NonAxis,Bug Report>
        </NonAxis>
</Group>
```

vi) </Compound> .... </Compound>

These tags are used to define a Compound relationship. Once again, as with the 'Group' tag, attributes within the 'Compound' tags have been defined as being related to each other by a Compound relation.

Example:

```
<Compound,Price>
        <Axis,Unit Price,Numeric>
        </Axis>
        <Axis,Site License,Numeric>
        </Axis>
</Compound>
```

Notes: Although the basic constructs for the axes are defined above there are a number of rules that must also be followed: -

1) Conceptual and compound grouping can only be defined at the top-level of definition.
2) All top-level attribute definitions have to be Axis or NonAxis.
3) A NonAxis definition can only contain definitions of SubNonAxis.
4) A SubAxis definition can contain definitions of SubNonAxis and SubAxis.

5) Axis and NonAxis defined at the same level require unique names.

Top-Level: - by a top level I refer to the main list of attributes within the definition and not any lower sub-definitions.

2. [Component Definitions]

This provides the means to define the components that will populate the parallel co-ordinate plot.

1) <Component,'Name','AttName'='AttValue', ... ></Component>

The definition of a component consists of a unique ID and a number of attributes, attribute value pairs.

Example:

<Component,C23,
Type=Filer,Producer=Charlie Ltd,Developed=01/3/2000,Quality=Good,Cost=120>

Notes - Specifying nested attribute values – Not every attribute has to have a unique identifier (name). However, the identifier must be unique with respect to the other definitions at the level within the attribute definitions. Thus to identify each attribute you specify its location in terms of its 'attribute path' (parents).

e.g.

'ParentOf'.'ParentOf'.'Attribute'

Empty/Null values – If an attribute value does not exist for component then a null value can be specified using one of two methods: -

1) Do not mention the attribute in the component definition
2) State the value of the attribute as 'null'

Example: -

For this Axis definition –

```
<Axis,Performance,Numeric>
        <SubAxis,Average Response Time,Numeric>
                <SubNonAxis,Functions Tested>
                </SubNonAxis>
        </SubAxis>
        <SubAxis,File IO Time,Numeric>
        </SubAxis>
</Axis>
```

Component definition –

```
<Component,C1,Perfomance=10,Performance.Average Response Time=30.4>
</Component>
<Component,C2,Performance=5,Performance.Average Response Time.Functions Tested=A
list of functions tested,File IO=34>
</Component>
```

Appendix B - Case Study Data set

Product Name	Supplier	Product Type	Developed	Required Memory	No. Reviews	User Rating	User - Performance	User - Quality	User - Value
Guardtech	Computech	Commercial	18/3/1997	4	73	6.8	6.3	8.0	6.2
Technocrypt	Computech	Commercial	21/2/1998	8	58	7.3	7.3	8.9	5.6
Microwall v1	Microwet	Commercial	12/4/1998	8	40	3.9	4.2	3.1	4.5
Encryption Suite	Microwet	Commercial	07/08/1999	8	43	6.4	7.8	5.2	6.1
Data Protector	KVR	Commercial	10/12/1999	16	39	8.0	5.7	9.3	8.9
Microwall v2	Microwet	Commercial	24/3/2000	16	30	6.2	7.6	5.2	5.7
Protection Tools	PCT	Commercial	2/4/2000	16	23	8.5	8.2	8.5	8.9
SafeSoft	LennySoft	Commercial	3/8/2000	8	60	8.7	7.7	9.2	9.3
VirtualBarrier	Colgem	Commercial	2/12/2000	8	56	7.1	6.1	7.4	7.7
FileSafe	Colgem	Commercial	12/5/2001	16	38	7.4	7.5	7.6	7.2
KVR-Crypt	KVR	Evaluation	14/12/2001	16	28	5.1	7.9	7.3	
XYZEncrypt	Colgem	Commercial	26/12/2001	16	32	7.5	8.1	7.5	6.9
PCT-Encryption	PCT	Evaluation	4/1/2002	8	17	3.6	4.2	6.5	
Encryption Lab	Microwet	Evaluation	14/1/2002	32	4	3.9	8.7	3.1	
Keycrypt	Colgem	Demo	11/2/2002	16	25	5.1	7.9	7.3	
EncodeAll	V Logic	Evaluation	29/3/2002	16	20	5.2	7.7	8.0	
FileGuardian	Colgem	Demo	1/5/2002	32	14	4.6	9.3	4.5	
Pro-Encryption	KVR	Demo	2/6/2002	32	1	5.1	9.6	5.6	
EncryptIT	AllThingsIT	Commercial	6/6/2002	32	8	7.0	9.8	7.6	3.5
CryptoLogic	V Logic	Demo	4/8/2002	16	9	5.1	8.4	6.8	

Case Study Data set (Continued)

Quality	No. Bugs	MTBF	No. Changes	Single User	Site License	Developer License	R&D MM	Skill MM
Very Good	3	2000		99	299	399		
Very Good	1	5000	1	199	499	899		
Very Bad	23	500	17	178	389			
Bad	20	580	13	150	289			
Excellent	0	10000	0	107	230	279	200	90
Very Good	4		2	189	167			
Very Good	2	7000	5	121	189	300		
Excellent	0	12000	3	99	149	199	195	120
Good	6	3500	8	149	178			
Good	7	3250	6	189	237	369		
Ok	8	2800	10	0	0	109	160	100
Good	6	3750		209	259	359		
Ok	13	1800	15	0	0			
Very Bad	32	450	23	0	0			
Good	9	2100	16	0				
Very Good	4	2300		0	0	59		
Bad	16	660	13	0		299		
Ok	12	1300	16	0			120	85
Good	5	4500	15	345	567	875		
Ok	10	1700	6	0				

Appendix C – Case Study Data File

```
Axis,Prod Name,String>
</Axis>
<Axis,Supplier,String>
</Axis>
<Axis,Prod Type,Nominal>
</Axis>
<Axis,Developed,Date>
</Axis>
<Axis,Memory,Numeric>
</Axis>
<Axis,No Reviews,Numeric>
</Axis>
<Axis,User Rating,Numeric>
        <SubAxis,Performance,Numeric>
        </SubAxis>
        <SubAxis,Quality,Numeric>
        </SubAxis>
        <SubAxis,Value,Numeric>
        </SubAxis>
</Axis>
<Group,Quality>
        <Axis,Quality,Ordinal,Very Bad,Bad,Ok,Good,Very Good,Excellent>
        </Axis>
        <Axis,No Bugs,Numeric>
        </Axis>
        <Axis,MTBF,Numeric>
        </Axis>
        <Axis,No Changes,Numeric>
        </Axis>
</Group>
<Compound,Price>
        <Axis,Single User,Numeric>
        </Axis>
        <Axis,Site License,Numeric>
        </Axis>
        <Axis,Developer License,Numeric>
        </Axis>
```

</Compound>
<Axis,R&D MM,Numeric>
</Axis>
<Axis,Skill MM,Numeric>
</Axis>
<Component,Demo,Prod Name=Guardtech,Supplier=Computech,Prod Type=Commercial,Developed=18/3/1997,Memory=4,No Reviews=73,User Rating=6.8,User Rating.Performance=6.3,User Rating.Quality=8.0,User Rating.Value=6.2,Quality=Very Good,No Bugs=3,MTBF=2000,Single User=99,Site License=299,Developer License=399>
</Component>
<Component,Technocrypt,Prod Name=Technocrypt,Supplier=Computech,Prod Type=Commercial,Developed=21/2/1998,Memory=8,No Reviews=58,User Rating=7.3,User Rating.Quality=8.9,User Rating.Performance=7.3,User Rating.Value=5.6,Quality=Very Good,No Bugs=1,MTBF=5000,No Changes=1,Single User=199,Site License=499,Developer License=899>
</Component>
<Component,Microwall v1,Prod Name=Microwall v1,Supplier=Microwet,Prod Type=Commercial,Developed=12/4/1998,Memory=8,No Reviews=40,User Rating=3.9,User Rating.Performance=4.2,User Rating.Quality=3.1,User Rating.Value=4.5,Quality=Very Bad,No Bugs=23,MTBF=500,No Changes=17,Single User=178,Site License=389>
</Component>
<Component,Encryption Suite,Prod Name=Encryption Suite,Supplier=Microwet,Prod Type=Commercial,Developed=7/8/1999,Memory=8,No Reviews=43,User Rating=6.4,User Rating.Performance=7.8,User Rating.Quality=5.2,User Rating.Value=6.1,Quality=Bad,No Bugs=20,MTBF=580,No Changes=13,Single User=150,Site License=289>
</Component>
<Component,Data Protector,Prod Name=Data Protector,Supplier=KVR,Prod Type=Commercial,Developed=10/12/1999,Memory=16,No Reviews=39,User Rating=8.0,User Rating.Performance=5.7,User Rating.Quality=9.3,User Rating.Value=8.9,Quality=Excellent,No Bugs=0,MTBF=10000,No Changes=0,Single User=107,Site License=230,Developer License=279,R&D MM=200,Skill MM=90>
</Component>
<Component,Microwall v2,Prod Name=Microwall v2,Supplier=Microwet,Prod Type=Commercial,Developed=24/3/2000,Memory=16,No Reviews=30,User Rating=6.2,User Rating.Performance=7.6,User Rating.Quality=5.2,User Rating.Value=6.1,Quality=Very Good,No Bugs=4,No Changes=2,Single User=189,Site License=167>
</Component>
<Component,Protection Tools,Prod Name=Protection Tools,Supplier=PCT,Prod Type=Commercial,Developed=2/4/2000,Memory=16,No Reviews=23,User Rating=8.5,User Rating.Performance=8.2,User Rating.Quality=8.5,User Rating.Value=8.9,Quality=Very Good,No Bugs=2,MTBF=7000,No Changes=5,Single User=121,Site License=189,Developer License=300>

</Component>
<Component,SafeSoft,Prod Name=SafeSoft,Supplier=LennySoft,Prod Type=Commercial,Developed=3/8/2000,Memory=8,No Reviews=60,User Rating=8.7,User Rating.Performance=7.7,User Rating.Quality=9.2,User Rating.Value=9.3,Quality=Excellent,No Bugs=0,MTBF=12000,No Changes=3,Single User=99,Site License=149,Developer License=199,R&D MM=195,Skill MM=120>
</Component>
<Component,VirtualBarrier,Prod Name=VirtualBarrier,Supplier=Colgem,Prod Type=Commercial,Developed=2/12/2000,Memory=8,No Reviews=56,User Rating=7.1,User Rating.Performance=6.1,User Rating.Quality=7.4,User Rating.Value=7.7,Quality=Good,No Bugs=6,MTBF=3500,No Changes=8,Single User=149,Site License=178>
</Component>
<Component,FileSafe,Prod Name=FileSafe,Supplier=Colgem,Prod Type=Commercial,Developed=12/5/2001,Memory=16,No Reviews=38,User Rating=7.4,User Rating.Performance=7.5,User Rating.Quality=7.6,User Rating.Value=7.2,Quality=Good,No Bugs=7,MTBF=3250,No Changes=6,Single User=189,Site License=237,Developer License=369>
</Component>
<Component,KVR-Crypt,Prod Name=KVR-Crypt,Supplier=KVR,Prod Type=Evaluation,Developed=14/12/2001,Memory=16,No Reviews=28,User Rating=5.1,User Rating.Performance=7.9,User Rating.Quality=7.3,Quality=Ok,No Bugs=8,MTBF=2800,No Changes=10,Single User=0,Site License=0,Developer License=109,R&D MM=160,Skill MM=100>
</Component>
<Component,XYZEncrypt,Prod Name=XYZEncrypt,Supplier=Colgem,Prod Type=Commercial,Developed=26/12/2001,Memory=16,No Reviews=32,User Rating=7.5,User Rating.Performance=8.1,User Rating.Quality=7.5,User Rating.Value=6.9,Quality=Good,No Bugs=6,MTBF=3750,Single User=209,Site License=259,Developer License=359>
</Component>
<Component,PCT-Encryption,Prod Name=PCT-Encryption,Supplier=PCT,Prod Type=Evaluation,Developed=4/1/2002,Memory=8,No Reviews=17,User Rating=3.6,User Rating.Performance=4.2,User Rating.Quality=6.5,Quality=Ok,No Bugs=13,MTBF=1800,No Changes=15,Single User=0,Site License=0>
</Component>
<Component,Encryption Lab,Prod Name=Encryption Lab,Supplier=Microwet,Prod Type=Evaluation,Developed=14/1/2002,Memory=32,No Reviews=4,User Rating=3.9,User Rating.Performance=8.7,User Rating.Quality=3.1,Quality=Very Bad,No Bugs=32,MTBF=450,No Changes=23,Single User=0,Site License=0>
</Component>
<Component,Keycrypt,Prod Name=Keycrypt,Supplier=Colgem,Prod Type=Demo,Developed=11/2/2002,Memory=26,No Reviews=25,User Rating=5.1,User Rating.Performance=7.9,User Rating.Quality=7.3,Quality=Good,No Bugs=9,MTBF=2100,No

Changes=16,Single User=0>

</Component>

<Component,EncodeAll,Prod Name=EncodeAll,Supplier=V Logic,Prod Type=Evaluation,Developed=29/3/2002,Memory=16,No Reviews=20,User Rating=5.2,User Rating.Performance=7.7,User Rating.Quality=8.0,Quality=Very Good,No Bugs=4,MTBF=2300,Single User=0,Site License=0,Developer License=59>

</Component>

<Component,FileGuardian,Prod Name=FileGuardian,Supplier=Colgem,Prod Type=Demo,Developed=1/5/2002,Memory=32,No Reviews=14,User Rating=4.6,User Rating.Performance=9.3,User Rating.Quality=4.5,Quality=Bad,No Bugs=16,MTBF=660,No Changes=13,Single User=0,Developer License=299>

</Component>

<Component,Pro-Encryption,Prod Name=Pro-Encryption,Supplier=KVR,Prod Type=Demo,Developed=2/6/2002,Memory=32,No Reviews=1,User Rating=5.1,User Rating.Performance=9.6,User Rating.Quality=5.6,Quality=Ok,No Bugs=12,MTBF=1300,No Changes=16,Single User=0,R&D MM=120,Skill MM=85>

</Component>

<Component,EncryptIT,Prod Name=EncryptIT,Supplier=Commercial,Prod Type=Commercial,Developed=6/6/2002,Memory=32,No Reviews=8,User Rating=7.0,User Rating.Performance=9.8,User Rating.Quality=7.6,User Rating.Value=3.5,Quality=Good,No Bugs=5,MTBF=4500,No Changes=15,Single User=345,Site License=567,Developer License=875>

</Component>

<Component,CryptoLogic,Prod Name=CryptoLogic,Supplier=V Logic,Prod Type=Demo,Developed=4/8/2002,Memory=16,No Reviews=9,User Rating=5.1,User Rating.Performance=8.4,User Rating.Quality=6.8,Quality=Ok,No Bugs=10,MTBF=1700,No Changes=6,Single User=0>

</Component>

Glossary

Data Item	The items whose properties are measured and stored in a multivariate data set.
Data Value	The measured values for each data item's data variables in a multivariate data set.
Data Variable	The measured properties of each data item stored in a multivariate data set.

References

Ack97 Ackoff, R.L., Transformational Consulting. Management Consulting Times, 1997. Vol. 28 No. 6. http://www.imcusa.org/Times/1997/imcTimes97JulAug.html

Ahl94a Ahlberg, C. and B. Shneiderman, Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In CHI '94. 1994. New York: ACM. p. 313-317

Ahl94 Ahlberg, C. and B. Shneiderman, The Alphaslider: A Compact and Rapid Selector. In CHI '94. 1994. Boston: ACM. p. 365-371

Ahl95a Ahlberg, C. and E. Wistrand, IVEE: An Environment for Automatic Creation of Dynamic Queries Applications. In CHI'95. 1995. ACM.

Ahl95b Ahlberg, C. and E. Wistrand, IVEE: An Information Visualization and Exploration Environment. In Information Visualization 95. 1995. Atlanta, Georgia: IEEE. p. 66-74

And72 Andrews, D.F., Plots of High-Dimensional Data. Biometrics. 1972. Vol. 29: p. 125-136.

Ank96 Ankerst, M., D.A. Keim, and H. Kriegel, Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets. In Visualisation '96. 1996. San Francisco.

App92 Apple Computer Inc., Macintosh Human Interface Guidelines. 1992. Reading, MA: Addison-Wesley.

Atk68 Atkinson, R. and R. Shiffrin, Human memory: A proposed system and its control processes. In The psychology of learning and motivation: Advances in research and theory, K. Spence and J. Spence, Editors. 1968. Academic Press: New York.

ATKO97 ATKOSoft, Survey on Visualisation methods and software tools. 1997. europa.eu.int/en/comm/eurostat/research/supcom.96/30/result/a/visualisation_methods.pdf

Ball96 Ball, T. and S.G. Eick, Software Visualization in the Large. IEEE Computer, 1996. Vol. 29 No. 4: p. 33-43.

Bate79 Bateson, G., Mind and Nature: a necessary unity. 1979, New York: Dutton.

Beck88 Becker, R.A. and W.S. Cleveland, Brushing Scatterplots. In Dynamic Graphics for Statistics, W.S. Cleveland and M.E. McGill, Editors. 1988. Wadsworth. p. 201-224.

Belk76 Belkin, N.J. and S.E. Robertson, Information Science and the Phenomenon of Information. Journal of the American Society for Information Science, 1976(July-August). p. 197-204.

Bell74 Bell, D., The coming of post-industrial society: a venture in social forecasting. 1974, London: Heinemann Educational Books.

Bell97 Bellinger, G., D. Gastro, and A. Mills, Data, Information, Knowledge, and Wisdom. 1997. www.outsights.com/dikw/dikw.htm

Benf96 Benford, S., Brown, C., Reynard, G. and Greenhalgh, C., Spaces: Transportation, Artificiality and Spatiality. In Computer Supported Cooperative Work '96. 1996. ACM. p. 77-86.

Bert81 Bertin, J., Graphics and Graphic Information Processing. 1981. Berlin: Walter de Gruyter.

Bert83 Bertin, J., Semiology of graphics: Diagrams, Networks and Maps. 1983. University of Wisconsin Press.

Besh90 Beshers, C. and S. Feiner, World within worlds n-Vision and AutoVisual. Computer Graphics, 1990. ACM. Vol. 24 No. 2: p. 37-38

Besh93 Beshers, C. and S. Feiner, Autovisual: Rule-based design of interactive multivariate visualizations. IEEE Computer Graphics and Applications, 1993. Vol. 13 No. 4: p. 41-49.

Blac97 Blackwell, A.F. Correction: A Picture is Worth 84.1 Words. In Proceedings of the First ESP Student Workshop. 1997. www.cl.cam.ac.uk/users/afb21/publications/Student-ESP.html

Blac98 Blackwell, A.F., Metaphor in Diagrams, Phd. Thesis. 1998, Cambridge University.

Borg00 Borgman, C.L., The Premise and Promise of a Global Information Infrastructure. First Monday, 2000. Vol. 5 No. 8. www.firstmonday.dk/issues/issue5_8/

Brac96 Brachmann, R. and T. Anand, Advances in Knowledge Discovery and Data Mining: A Human-Centered Approach. In Advances in Knowledge Discovery and Data Mining. 1996. AAAI Press: California. P. 37-58.

Broo74 Brookes, B.C., Robert Fairthorne and the Scope of Information Science. Journal of Documentation, 1974. Vol. 30: p. 139-152.

Broo80 Brookes, B.C., The foundations of information science. Part I. Philosophical Aspects. Journal of Information Science, 1980. Vol. 2: p. 125-133.

Broo86 Brookes, F.P. No Silver Bullet - Essence and Accidents of Software Engineering. In Information Processing 86. 1986. Dublin: Elsever Science Publishers. p. 1069-1076

Brow00 Brown, J.S. and P. Duguid, The Social Life of Information. First Monday, 2000. Vol. 5 No. 4. www.firstmonday.dk/issues/issue5_4/

Brun98 Brunsdon, C., A.S. Fotheringham, and M.E. Charlton, An Investigation of Methods for Visualising Highly Multivariate Datasets. In Case Studies of Visualization in the Social Sciences, D. Unwin and P. Fisher, Editors. 1998.

Buck91 Buckland, M., Information as a Thing. Journal of the American Society of Information Science, 1991. Vol. 42 No. 5: p. 351-360.

Cab97 Cabena, P., P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, Discovering Data Mining: From Concept to Implementation. 1997. Upper Saddle River, NJ: Prentice Hall.

Card99a Card, S.K., J.D. Mackinlay, and B. Shneiderman, Readings in Information Visualization: Using Vision to Think. 1999. Morgan Kaufmann Publishers.

Card99b Card, S.K., P. Pirolli, and J.D. Mackinlay, The Cost-of-Knowledge Characteristic Function: Display Evaluation for Direct-Walk Dynaimic Information Visualisations. In Readings in Information Visualization: Using Vision to Think. 1999. Morgan Kaufmann Publishers. p. 582-588.

Chec90 Checkland, P.B. and J. Scholes, Soft Systems Methodology in Action. 1990. New York: John Wiley & Sons.

Cher73 Chernoff, H., The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, 1973. Vol. 68: p. 361-367.

Clev93 Cleveland, W.S., Visualizing data. 1993, Hobart Press: New Jersey.

Cleve85 Cleveland, W.S. and R. McGill, Graphical perception and graphical methods for analysing scientific data. Science, 1985. Vol. 229: p828-833.

Comp ComponentSource, http://www.componentsource.com/Services/Glossary.asp (11/08/2003)

Cruz93 Cruz-Neira, C., D.J. Sandin, and D.T. A. Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. In SIGGRAPH '93. 1993. ACM

Def99 DeFanti, T.A., M.D. Brown, and B.H. McCormick, Visualisation - Expanding Scientific and Engineering Research Opportunities. In Reading in Information Visualisation: Using Vision to Think. 1999. Morgan Kaufmann Publishers. p. 39-52.

Derv86 Dervin, B. and M. Nilan, Information Needs and Uses. In Annual Review of Information Science and Technology, M. William, Editor. 1986. White Plains: New York. p. 3-33.

Dieb94 Dieberger, A., Navigation in Textual Virtual Environments using a City Metaphor. Phd. Thesis. 1994. Faculty of Technology and Sciences, Vienna University of Technology.

Dill95 Dilly, R., Data Mining An Introduction: Student Notes. Parallel Computer Centre, The Queen's University of Belfast. 1995.

Dutt99 Dutton, R.T., J.C. Foster, and M.A. Jack, Please Mind the Doors - Do interface metaphors improve the usability of voice response services? BT Technology Journal, 1999. Vol. 17 No. 1: p. 172-177.

Eick00 Eick, S.G. and A.F. Karr, Visual Scalability. IEEE Transactions on Visualization and Computer Graphics, 2000. Vol. 6 No. 1: p. 44-58.

Eick94 Eick, S.G. Data Visualization Sliders. In User Interface Software and Technology (UIST) '94. 1994. Monterey, California: ACM. p. 119-120

Fab01 Fabrikant, S.I. and B.P. Buttenfield, Formalising Semantic Spaces For Information Access. Annals of the Association of American Geographers, 2001. Vol. 91: p. 263-280.

Fayy96 Fayyad, U.M., et al., Advances in Knowledge Discovery and Data Mining. 1996. California: AAAI Press.

Fein90 Feiner, S. and C. Beshers. Worlds within Worlds: Metaphors for Exploring n-Dimensional Virtual Worlds. In UIST'90. 1990. ACM.

Feld98 Feldens, M.A., Moraes, R.L., Pavan, A., Castiho, J.M.V., Towards a Methodology for the Discovery of Useful Knowledge Combining Data Mining, Data Warehousing and Visualization. 1998. citeseer.nj.nec.com/243358.html

Fish95 Fishkin, K.P. and M.C. Stone. Enhanced Dynamic Queries via Movable Filters. In Conference on Human Factors in Computing Systems. 1995. Denver, Colorado: ACM.

Flor02 Floridi, L., Is Information Meaningful Data? 2002. www.wolfson.ox.ac.uk/~floridi/papers.htm.

Furn91 Furnas, G.W., The FISHEYE View: A New Look at Structured Files. 1981. Bell Laboratories Technical Report. Murray Hill, New Jersey.

Glob94 Globus, A., Principles of Information Display for Visualization Practitioners. 1994. NASA Ames Research Center. www.nas.nasa.gov/Research/Reports/Techreports/1994/HTML/NAS-94-002.paper.html

Goeb99 Goebel, M. and L. Gruenwald, A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations, 1999. Vol. 1 No. 1: p. 20-33.

Gog99 Goguen, J., Tossing Algebraic Flowers down to the Great Divide. In People & Ideas in Theoretical Computer Science, C.S. Calude, Editor. 1999, Springer-Verlag Singapore Pte. Ltd. p. 93-129.

Guv98 Guvenen, O., The Impact of information and communication technologies on society. Journal of International Affairs, 1998. Vol. 2 No. 4.

Hamm73 Hamming, R.W., Numerical Analysis for Scientists and Engineers. 1973. New York: McGraw-Hill.

Hatch01 Hatch, A.S., M. P. Smith, C. M. B. Taylor, M. Munro, No Silver Bullet for Software Visualisation Evaluation. In International Conference on Imaging Science, Systems, and Technology (CISST). 2001. Las Vegas, USA.

Heal93 Healey, C.G., Visualization of Multivariate Data Using Preattentive Processing. In Computer Science. 1993. University of British Columbia.

Henz98 Henze, C., Feature Detection in Linked Derived Spaces. In Visualization '98. 1998. IEEE.

Hill97 Hillis, D., A Time of Transition. Communications of ACM. 1997. Vol. 40 No. 2: p37-39

Ins90 Inselberg, A. and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. In Visualization '90. 1990. IEEE. p. 361-378.

John83 Johnson-Laird, P., Mental models: Towards a Cognitive Science of Language, Inference and Consciousness. 1983, Cambridge: Cambridge University Press.

John97 John, G.H., Enhancements to the Data Mining Process, Phd. Thesis, Computer Science. 1997. Stanford University.

Keim94 Keim, D.A. and H. Kriegel, VisDB: Database Exploration Using Multidimensional Visualization. IEEE Computer Graphics and Applications, 1994. Vol. 14 No. 5: p. p40-49.

Keim96 Keim, D.A. and H.P. Kriegel, Visualization Techniques for Mining Large Databases: A Comparison. IEEE Trans. Knowledge and Data Engineering, 1996. Vol. 8 No. 6: p. 923-936.

Keim97a Keim, D.A., Visual Database Exploration Techniques. In Tutorial KDD'97 Int. Conference on Knowledge Discovery and Data Mining. 1997. Newport Beach, CA. http://romblon.dbs.informatik.uni-muenchen.de/~daniel/index_personal.html

Keim97b Keim, D.A., Visual Techniques for Exploring Databases. In Int. Conference on Knowledge Discovery in Databases (KDD'97). 1997. Newport Beach, CA. http://www.dbs.informatik.uni-muenchen.de/~daniel/publication.html

Kidd94 Kidd, A., The Marks are on the Knowledge Worker. In CHI '94. 1994. Boston, MA: ACM Press.

Knig00a Knight, C. and M. Munro, Mindless Visualisations. In ERCIM Workshop "User Interfaces for All". 2000.

Knig00b Knight, C. and M. Munro, Virtual but Visible Software. In International Conference on Information Visualisation (IV00). 2000. London, England: IEEE.

Knig99 Knight, C. and M. Munro, Visualising Software – A Key Research Area. In International Conference on Software Maintenance. 1999. Oxford, England: IEEE.

Lark87 Larkin, J. and H.A. Simon, Why a Diagram is (Sometimes) Worth Ten Thousand Words. Cognitive Science, 1987. Vol. 11 No. 1: p. 65-99.

Lehn97 Lehner, F. and R. Maier, Can Information Modelling be Successful without a Common Perception of the Term 'Information'? In 6th European - Japanese Seminar on Information Modelling and Knowledge Bases. 1997. Copenhagen. p. 181-198

Mach83 Machlup, F., Semantic Quirks in Studies of Information. In The Study of Information: Interdisciplinary Messages, F. Machlup and U. Mansfield, Editors. 1983. New York: Wiley. p. 641-471.

Mack86 Mackinlay, J., Automating the Design of Graphical Presentations of Relational Information. ACM Transactions on Graphics, 1986. Vol. 5 No. 2: p. 110-141.

Mad94 Madsen, K.H., A guide to metaphorical design. Communications of the ACM, 1994. Vol, 37 No. 12: p. 57-62.

Marc95 Marchionini, G., Information seeking in electronic environments. 1995. New York: Cambridge University Press.

March99 Marchionini, G., Educating Responsible Citizens in the Information Society. In Educational Technology. 1999. Vol. 39 No. 2: p. 17-26.

Mizz96 Mizzaro, S., On the Foundations of Information Retrieval. In AICA'96 The Annual Italian Computer Science Conference. 1996. Rome. citeseer.nj.nec.com/mizzaro96foundations.html

Mizz98 Mizzaro, S., How many relevances in information retrieval? Interacting with Computers, 1998. Vol. 10 No. 3: p. 305-322.

Mon94 Monin, N. and D.J. Monin, Personification of the Computer: A Pathological Metaphor in IS. In SIGCPR '94. 1994. Alexandria, VA, USA: ACM Press.

Norm93 Norman, D.A., Things that Make Us Smart. Reading MA. 1993. Addison Wesley.

Petr98 Petre, M., A.F. Blackwell, and T.R.G. Green, Cognitive Questions in Software Visualisation. In Software Visualization: Programming as a Multi-Media Experience, J. Stasko, et al., Editors. 1998. MIT Press. p. p453-480.

Pett97 Pettifer, S. and A. West, Deva: A coherent operating environment for large scale VR applications. In Virtual Reality Universe. 1997. Santa Clara, California.

Popp72 Popper, K.R., Objective Knowledge: An Evolutionary Approach. 1972. Oxford: Claredon Press.

Pran98 Prang, T., Unsupervised Data Mining in Nominally-Supported Databases. 1998. www-lehre.informatik.uni-osnabrueck.de/~ftprang/papers/tproject/tproject.html

Pree94 Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., Carey, T., Human Computer Interaction. 1994. Addison Wesley.

Prop99 Proper, H.A. and P.D. Bruza, What is Information Discovery About? Journal of the American Society of Information Science, 1999. Vol. 50 No. 9: p. 737-750.

Quig99 Quigley, E.J. and A. Debons. The Interrogative Theory of Information and Knowledge. In SIG-CPR. 1999. ACM.

Rek93 Rekimoto, J. and M. Green. The Information Cube: Using Transparency in 3D Information Visualization. In Third Annual Workshop on Information Technologies & Systems (WITS'93). 1993. p. 125- 132

Rose00 Rose, S. and P.C. Wong, DriftWeed: A visual metaphor for interactive analysis of multivariate data. In Visual Data Exploration and Analysis (IS&T/SPIE). 2000. San Jose, California.

Rosz86 Roszak, T., The Cult of Information. 1986. New York: Pantheon.

Shap96 Shapiro, J. and S. Hughes, Information Literacy as a Liberal Art: Enlightenment proposals for a new curriculum. Educom Review, 1996. Vol. 31 No. 2. www.educause.edu/pub/er/review/reviewarticles/31231.html

Shne83 Shneiderman, B., Direct Manipulation: A Step Beyond Programming Languages. IEEE Computer, 1983. Vol. 16 No. 8: p. 57-69.

Shne94 Shneiderman, B., Dynamic Queries for Visual Information Seeking. IEEE Software, 1994. Vol. 11 No. 6: p. 70-77.

Shne96 Shneiderman, B. The eyes have it: Task by Data Type Taxonomy for Information. In IEEE Workshop on Visual Languages. 1996. p. 336-343.

Shne97 Shneiderman, B. Direct Manipulation for Comprehensible, Predictable, and Controllable User Interfaces. In Intelligent User Interfaces (IUI97). 1997. Orlando, FL. ACM. p. 33-39.

Shne98 Shneiderman, B., Treemaps for space-constrained visualization of hierarchies. 1998. http://www.cs.umd.edu/hcil/treemaps/

Sholl99 Sholle, D., What is Information? The Flow of Bits and the Control of Chaos. In Media in Transition. 1999. MIT.

Siir00 Siirtola, H., Direct Manipulation of Parallel Coordinates. In Information Visualisation (IV2000). 2000. London, England: IEEE.

Sten01 Stenmark, D., The Relationship between Information and Knowledge. In Proceedings of IRIS 24. 2001. Ulvik, Norway

Tayl86 Taylor, R., Value-added processes in information systems. 1986, Norwood, NJ: Ablex Publishing Corporation.

Toum99 Toumi, I., Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organisational memory. Journal of Management Information Systems, 1999. Vol. 16 No. 3: p. 107-121.

Trei80 Treisman, A., Preattentive processing in vision. Computer Vision, Graphics and Image Processing, 1980. Vol. 31: p. 156-177.

Trei86 Treisman, A., Properties, Parts and Objects. In Handbook of Perception and Human Performance, K.R. Boff, L. Kaufman and J.P. Thomas, Editors. 1986. Wiley: New York. Ch. 35

Tuft83 Tufte, E.R., The Visual Display of Quantitative Information. 1983. Cheshire, CT: Graphics Press.

Tuft90 Tufte, E.R., Envisioning Information. 1990. Cheshire, CT: Graphics Press.

Tuft97 Tufte, E.R., Visual Explanations: Images and Quantities, Evidence and Narrative. 1997. Cheshire, CT: Graphics Press.

Unw00 Unwin, A., Visualisation for Data Mining. International Conference on Data Mining, Visualization and Statistical System. 2000. www1.math.uni-augsburg.de/~unwin

Ven00 Venables, B., Trellis Graphics. 2000. www.cmis.csiro.au/statline/2000/feb2000.htm

Ward94 Ward, M., Integrating multiple methods for visualizing multivariate data. In Visualization '94. 1994. Los Alamitos, California: IEEE.

Wea49 Weaver, W. and C.E. Shannon, The Mathematical Theory of Communication. 1949. Urbana, Illinois: University of Illinois Press.

Wers79 Wersig, G., The Problematic Situation as a Basic Concept of Information Science in the Framework of the Social Sciences: A Reply to Belkin. In New Trends in Informatics and its Terminology. 1979. VINITI: Moscow. p. 48-57.

Wick87 Wickens, C.D., Information processing, decision making, and cognition. In Handbook of Human Factors, G. Salvendy, Editor. 1987. John Wiley & Sons: New York. p. 126-127.

Will92 Williamson, C. and B. Shneiderman, The Dynamic HomeFinder: Evaluating Dynamic Queries in a Real-Estate Information Exploration System. In SIGIR Conference on Research and Development in Information Retrieval. 1992. Copenhagen, Denmark: ACM.

Wiss98 Wiss, U., D.A. Carr, and H. Jonsson, Evaluating Three-Dimensional Information Visualization Designs: A Case Study of Three Designs. In IV'98. 1998. London, England: IEEE.

Wri95 Wright, W. Information Animation Applications in the Capital Markets. In InfoVis '95. 1995. IEEE.