Chapter 1

Introduction

The main application of databases, which have been in use since the early 1970's, is to support the operational aspects of organisations. Here, everyday transactions are captured and processed via standard database queries and reports. More recently, a new use of database systems has emerged. This trend is mainly due to the demand for information made by management. Database systems now also have to support the decision making process. The desire by organisations to gain a competitive advantage using the knowledge from their corporate data has led to the need for business intelligence. Berthold *et al* (1999), defines business intelligence as "*the gathering and management of data, and the analysis of that data to turn it into useful information to improve decision making*". According to the Palo Alto Management Group the market, by the year 2001, for business intelligence is estimated close to $70 billion, a huge industry by any measure [Berthold *et al* 1999]. However, existing databases have been designed to automate transaction processing, therefore these databases and traditional database query tools are not well suited for this new task [Inmon 1996]. This shortcoming is addressed through the development of intelligent data analysis tools that extracts knowledge from data, as discussed next.

Knowledge Discovery from Data (KDD) focuses on the extraction of knowledge from data repositories i.e. databases and data warehouses, enabling these repositories to support the decision making process of organisations. KDD can be defined as "*the non-trivial extraction of implicit, previously unknown and potentially useful information from data*" [Adriaans *et al* 1996]. In principle, the process of knowledge discovery from data consists of six stages, namely data selection, cleaning, enrichment, coding, knowledge discovery (intelligent data analysis) and reporting. The first four stages concern data pre-processing during which a single, integrated data source for the knowledge discovery phase should be

built. The fifth stage of the KDD process focuses on the actual process of intelligent data analysis, also referred to as data mining. A number of popular data mining techniques are neural networks, as discussed in Section 2.5.6, decision trees, as discussed in Section 2.5.5, rule induction algorithms, as discussed in Section 2.5.4 and generic algorithms. These tools employ statistics and machine learning to extract knowledge from data. Recently, hybrid systems that combine two or more of these data mining techniques into a single system have gained popularity.

The two main objectives of this study are the creation of an intelligent data analysis tool, which combines more than one data mining technique into a unified framework, and the subsequent application of this intelligent data analysis tool to a real-world application. The real-world application concerns, the National Research and Technology (NRT) Audit, as discussed in Chapter 3. This audit was commissioned by the South African Department of Arts, Culture, Science and Technology (DACST) to better understand the capacity, capability and limitations of the country's science and technology system. The rationale behind the audit is discussed next.

South Africa is facing many new challenges in the 21st century – competing in a world economy, growing environmental concerns, social and economic inequalities, an ageing population, low productivity, massive unemployment, and the nation's evolving role in Africa. It is widely recognised in government that the role of technology, defined as *"the set of means embodied in products, processes, machines and related services by which the physical and informational domains are manipulated"* [DACST 1998], is crucial in addressing these challenges. This recognition emphasises the importance of having a robust technology policy. Drucker [in Fairhead 1995] mentioned that the major challenge facing management in developed countries is improving the performance of knowledge and service workers, i.e. the decision makers. However, in a developing country, such as South Africa, the need to improve the performance of decision makers in government is even more important. This improvement will enable decision makers to formulate policies that successfully address the challenges lying ahead. Formulating a technology policy requires a critical assessment of South Africa's strengths and weaknesses in science and technology. This realisation was the reason for the commissioning of the NRT Audit by the South African Government.

This thesis discusses the development of an intelligent data analysis tool and its application to aid decision makers with the interpretation of the findings of the NRT Audit. The tool uses the data repository developed as part of the NRT Audit to verify the findings contained in the synthesis report, Technology and Knowledge [DACST 1998], prepared by the Foundation for Research and Development. In addition, this study investigates the existence of possible additional findings. Here, the use of intelligent data analysis may lead to the discovery of additional conclusions that can be used for decision-making.

The remainder of this chapter provides an introductory overview of the thesis. Section 1.1 contains the problem statement; Section 1.2 briefly outlines the research approach that was followed; and lastly, Section 1.3 presents an outline of the remainder of the thesis.

1.1 Problem statement

According to Hand [in Berthold *et al* 1999], intelligent data analysis does not consist of choosing and applying a technique to match a problem at hand. He argues that data analysis is not a collection of isolated techniques, completely different from one another, waiting to be matched to a problem. On the contrary, these techniques have complex interrelationships that can alter the problem being investigated in subtle ways. Furthermore, very rarely is a problem so precisely stated that a single application of one technique will suffice. Hand describes intelligent data analysis as the *"repeated application of techniques, as one attempts to tease out the structure, to understand what is going on. To refine the questions that the researchers are seeking to answer requires painstaking care and, above all, intelligence. It is a carefully planned and considered process of deciding what will be most useful and revealing"* [Berthold *et al* 1999]. A number of studies have been conducted regarding the selection of the best, most useful technique for solving different types of problems, including: Dhar *et al* 1997, Michie *et al* 1994 and Moustakis *et al* 1996. Some studies reported few differences between the various techniques while others reported a distinct advantage using a particular technique. Moustakis *et al's* (1996) survey indicated the following general trends: A classification task, which is a *task that has the ability to classify*

objects as members of known classes, is best supported by neural networks. When analysing the generic soybean data set, containing instances describing soybeans with diseases, a neural network will be the best data mining technique to use if the aim of the task is to classify these diseases. Rule induction algorithms best support a problem solving type of task, which is a *task that describes the procedural and algorithmic aspects of problem solving*. When given a set of constraints describing a system, a rule induction algorithm will be best at optimising the system's parameters.

The question now arises, since there are so many techniques that appear to be applicable to a task, which one should be chosen. Furthermore, if the problem consists of more than one type of task which technique should be used? The study by Moustakis *et al* (1996) draws the conclusion that different types of tasks are best supported by different techniques and that a KDD environment's success as an intelligent data analysis tool depends on the combination and integration of individual data mining techniques.

Real-world problems are complex. It takes an infinite number of attributes to accurately describe the problem domain. This complexity can be simplified by a divide and conquer approach. A problem can be broken into smaller problems, each addressing a particular section of the domain. The data describing each section should then be analysed separately. During the analysis of the data describing each of these smaller domains different data mining techniques will be required depending on the type of task. This implies that, when addressing a real-world problem with the use of KDD, an environment consisting of more than one data mining technique is required. Following Moustakis *et al's* (1996) findings, these techniques should be combined and integrated.

The integration and combination of data mining techniques into a single KDD environment can be viewed from three different perspectives. The environment can be seen as a multi-strategy learning system, a multi-agent learning system or a hybrid learning system, as discussed by Viktor (1999), drawn from Kholsa *et al* (1993), Jacobsen (1998), Jennings (1993), Honavar (1995) and Sun (1995):

- A multi-strategy learning system involves the combination of more than one form of learning strategy into a single system. In this way the strengths of each strategy is enhanced and the weaknesses are alleviated.
- The second perspective is that of a multi-agent learning system. According to Russell *et al* (1995), an agent is *"anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors"*. In a multi-agent learning system the environment consists of a number of agents, residing in a communal environment, interacting with one another as well as the environment.
- Lastly, when viewing the environment as a hybrid learning system, the different data mining techniques are combined into a single system. Operating side-by-side, the techniques co-operate to solve a problem.

Following Viktor (1999), the KDD environment developed during this study will be viewed from the perspective of a multi-agent learning (MAL) system. This approach is chosen since it does not isolate learning from the environment in which the problem exists, but acknowledges the influence of the communal environment on learning. Weiss (1999) states that a multi-agent learning system offers a natural way to view and characterise learning. Multi-agent systems reflect the insight that, learning, intelligence and interaction are deeply coupled.

The agent and multi-agent learning concepts are based on our perception of how humans learn. Humans do not learn in isolation, but participate with other intelligent agents, i.e. other humans or machine learners, in an environment in which they all co-operate. They interact in many ways and most of what they achieve is a result of these interactions. A multi-agent learning system provides a framework for modelling these interactions. Learning within the MAL system will occur by means of the full co-operation of the participating learning agents, applying inductive learning techniques as described in Section 2.1.

In summary this study investigates the use of a KDD environment in a real-world application, focusing on the following:

i. The development of an intelligent data analysis tool, modelled as a MAL system that combines more than one data mining technique into a unified framework for decision support.

ii. The evaluation of the capability of this intelligent data analysis tool, using co-operative inductive learning techniques, in analysing the context as embedded in qualitative data, to be used for decision-making.

The next section discusses the research approach followed during this investigation.

1.2 Research approach

The underlying epistemology supporting this study is a combination of a quantitative and qualitative approach to research, by the triangulation of data from different sources [Kaplan *et al* 1988]. The specific research methodology that has been followed is a combination of case study and action research from an interpretative philosophical perspective.

Case study research is the most common qualitative research method used in information systems research. Yin (1994) defines case study research as follows:

"A case study is an empirical inquiry that:

- *Investigates a contemporary phenomenon within real-life context, especially when*
- *The boundaries between phenomenon and context are not clearly evident"*

The phenomenon investigated in this study is thus: the capability of a multi-agent learning system in analysing a complex real-world domain, defined by a set of qualitative data, for decision-making, as discussed next.

The approach to this study is depicted graphically in Figure 1.

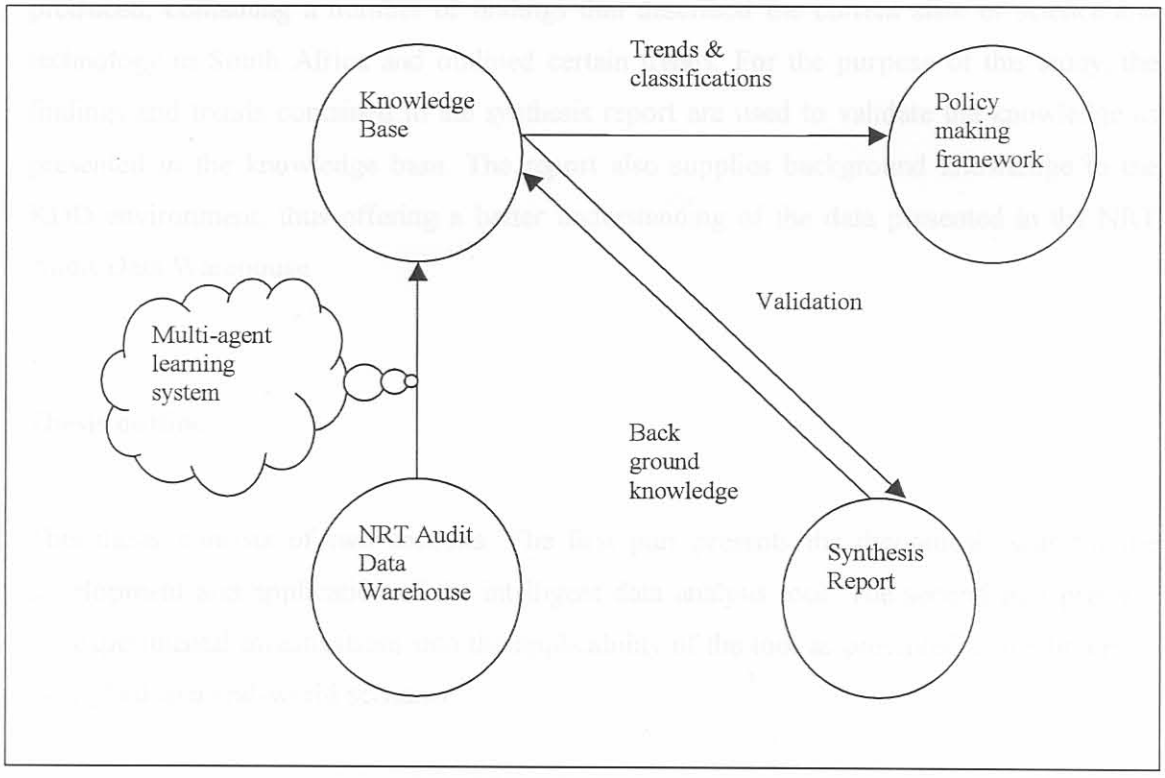

Figure 1 Research Approach

The real-world domain analysed for this study is the South African Science and Technology System. This science and technology system is defined by a set of qualitative data as contained in, the NRT Audit Data Warehouse. The NRT Audit Data Warehouse is one of the outputs of the NRT Audit and is a data repository of all the data gathered during the audit. Knowledge should be extracted from this data repository by an intelligent data analysis tool and stored into a knowledge base. This knowledge base, containing the potential useful information from the data can then be used as a basis for policy making directed at the development of science and technology in South Africa. The knowledge will be represented by means of rule sets consisting of trends and classifiers, as defined in Section 2.1. Recall that, learning within the MAL system will occur by means of the full co-operation of the participating learning agents, applying inductive learning techniques. Therefore, the intelligent data analysis tool used for extracting the knowledge is modelled as a co-operative inductive learner team - multi-agent learning system (CILT-MAL system), as described in Chapter 2.

During the NRT Audit a synthesis report, Technology and Knowledge [DACST 1998], was produced, containing a number of findings that described the current state of science and technology in South Africa and outlined certain trends. For the purpose of this study, the findings and trends contained in the synthesis report are used to validate the knowledge as presented in the knowledge base. The report also supplies background knowledge to the KDD environment, thus offering a better understanding of the data presented in the NRT Audit Data Warehouse.

1.3 Thesis outline

This thesis consists of two sections. The first part presents the theoretical basis for the development and application of the intelligent data analysis tool. The second part presents the experimental investigations into the applicability of the tool as presented in the first part, as applied to a real-world scenario.

Chapter 2 lays the theoretical foundation for the experimental work conducted in Chapters 4 and 5. In Chapter 2 the CILT-MAL system is defined and the learners that participated in the CILT-MAL system are described. Chapter 3 describes the NRT Audit case study. This case study is used for the experimental work as documented in Chapters 4 and 5. Chapter 4 describes the first experiment that used a CILT-MAL system to conduct a classification task required by the audit. Chapter 5 documents the second experiment, the aim of which was to process a problem solving type task, as needed by the audit. Chapter 6 summarises the main results of this thesis and discusses areas of future research.