

# Visual Grouping of Association Rules for Hypotheses Suggestion

by

Sasha Ivkovic

B Comp(Hons) for the degree of Master of IT

at the

UNIVERSITY of BALLARAT

September 2003

© University of Ballarat, School of ITMS 2003. All rights reserved.

Supervisor: John Yearwood  
Associate Professor John Yearwood, School of ITMS

Supervisor: Andrew Stranieri  
Dr. Andrew Stranieri, School of ITMS

University of Ballard Special  
collection

b 12330565

## Statement of Originality

Except where explicit reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or part from a thesis by which I have qualified for or have been awarded another degree or diploma. No other person's work has been relied upon or used without due acknowledgement in the main text and bibliography of the thesis

S. IVKOVIC

Sasha Ivkovic





## Acknowledgments

I wish to express my gratitude to Dr. John Yearwood, and Dr. Andrew Stranieri for their assistance and guidance in this investigation. I would like to thank my supervisors for many helpful conversations and introducing me to the wonderful world of Data Mining. My gratitude to the School of Information Technology and Mathematical Science for their continuous support during the course.

Many thanks to Diabetes Australia experts, and the diabetes medical research team at the University of NSW. Special thanks to domain experts from Victorian Legal Aid, especially the collaborative research officer Domenico Calabro, who provided his domain knowledge, time and energy so generously during the whole process. This investigation was supported by the Victorian Ministry of Education, Australia.

This thesis is dedicated to my family, my beautiful and supportive wife Debbie, and my children Stephanie and Nicholas.



## **Ethics**

Full approval for this project was granted by the Human Ethics Committee at the University of Ballarat.

Approval number: HREC 02/152

Contact: Sally Boyle

Executive Officer

Human Research Ethics Committee





# Visual Grouping of Association Rules for Hypotheses Suggestion

by

Sasha Ivkovic

## Abstract

Each year more operations are being computerised and virtually all large to mid-size organisations store information as data. For example, hundreds of thousands of electronic bank transactions are now being recorded. However stored (raw) data is rarely of direct benefit for improving business decisions. Moreover, traditional manual data analysis is becoming impractical in many domains as data volumes grow exponentially.

Knowledge Discovery from Databases (KDD) has become an umbrella name for new techniques that intelligently assist humans in analysing large structured data sets. The involvement of domain experts is considered important in all phases of any KDD exercise. However, in practice domain experts have an involvement only in the pre and post mining phases whereas data mining experts drive the entire discovery process. This is largely due to the complexity of current KDD tools.

Most KDD commercial and research tools require extensive training. Furthermore, discovered patterns are often difficult to interpret. Domain experts such as lawyers, health care professionals, engineers and managers require simple-to-use tools that efficiently solve their business problems or guide them to more detailed expert analysis. These experts are usually less interested in using advanced technology than they are in getting clear, rapid answers to their everyday business questions.

In this study, we present a KDD method that is used by non-technical experts with minimal training to discover and interpret patterns that they find useful for their role within an organisation. The approach generates association rules (AR) and then displays them by grouping rules together and visually depicting deviations between groups. While association rules have proven to be useful in practical applications, AR algorithms tend to generate large numbers of rules, most of which are of little interest. In addition, large numbers of rules are difficult to interpret. A visual display of deviations defined as interesting provides new opportunities for hypotheses suggestion and testing.

The approach has been evaluated with the development of a prototype called WebAssociate. This web based tool has been used by domain experts at a government funded aid organisation with favourable responses. The tool has also been used by medical researchers and practitioners to analyse diabetes data.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Information Age . . . . .	1
1.2	Knowledge Discovery from Databases - KDD . . . . .	1
1.3	Organisational use of KDD . . . . .	4
1.3.1	Domain Experts . . . . .	5
1.4	Association Rules . . . . .	6
1.4.1	Pattern Interestingness . . . . .	7
1.4.2	Group Differences . . . . .	8
1.5	Discovery Visualisation . . . . .	10
1.6	Motivation . . . . .	11
1.6.1	Problem 1 - Exploitation of KDD technology within the organisation is not fully implemented . . . . .	12
1.6.2	Problem 2 - Users may be overwhelmed by the number of discovered rules . . . . .	13
1.6.3	Problem 3 - Mapping of a hypothesis to the set of association rules that will confirm or deny the hypothesis is rarely simple . . . . .	14
1.6.4	Problem 4 - Experts need additional methods to explain hypotheses	15
1.7	Research Questions . . . . .	15
1.8	The Organisation of this Thesis . . . . .	16
<b>2</b>	<b>Literature Review - KDD</b>	<b>17</b>
2.1	Data Mining . . . . .	17
2.1.1	Introduction . . . . .	17
2.1.2	Essential Steps to Knowledge Discovery . . . . .	21
2.1.3	Data Mining Techniques . . . . .	24
2.2	Association Rules . . . . .	27
2.2.1	Objective Interestingness . . . . .	29
2.2.2	Subjective Interestingness . . . . .	32

2.2.3	Interestingness of Deviations . . . . .	35
2.2.4	Interestingness of Group Differences . . . . .	37
2.3	Hypotheses and AR . . . . .	41
2.3.1	Hypothesis Suggestion . . . . .	41
2.3.2	Hypothesis Testing . . . . .	42
2.3.3	The link between the Hypothesis and AR . . . . .	42
2.4	Discovery Visualisation . . . . .	43
2.4.1	Visualisation Categories . . . . .	43
2.5	Association Rule Visualisation . . . . .	44
2.6	Chapter summary . . . . .	49
<b>3</b>	<b>Domain Experts</b>	<b>51</b>
3.1	The Role of the Domain Expert . . . . .	52
3.2	KDD Requirements of the Domain Expert . . . . .	54
3.3	KDD Requirements of the VLA Domain Experts . . . . .	57
3.4	KDD Requirements of the Diabetes Domain Experts . . . . .	58
3.5	Chapter Summary . . . . .	59
<b>4</b>	<b>Sample Consultations</b>	<b>61</b>
4.1	Consultation 1 . . . . .	64
4.1.1	Hypotheses suggestion . . . . .	64
4.1.2	Discovery interestingness . . . . .	65
4.1.3	Hypothesis testing . . . . .	68
4.1.4	Suggested Explanations for the Alternate Hypothesis . . . . .	70
4.1.5	Hypothesis Explanation . . . . .	72
4.1.6	Conclusion . . . . .	76
4.2	Consultation 2 . . . . .	77
4.2.1	Country selections for the UK and OZ groups . . . . .	77
4.2.2	Country selection for the Muslim group . . . . .	79
4.2.3	Consequent selection . . . . .	79
4.2.4	Hypothesis exploration . . . . .	80
4.2.5	Conclusion . . . . .	82
4.3	Consultation 3 . . . . .	83
4.3.1	Variable of interest selection . . . . .	83
4.3.2	Consequent selection . . . . .	83
4.3.3	Common matter codes for Vietnam . . . . .	83
4.3.4	Country Vietnam and drug related matters . . . . .	84

4.3.5	Further exploration . . . . .	85
4.3.6	Conclusion . . . . .	86
4.4	Consultation 4 . . . . .	88
4.4.1	Hypothesis suggestion . . . . .	88
4.4.2	Choosing a variable of interest . . . . .	88
4.4.3	Discovery interestingness . . . . .	89
4.4.4	Rule exploration . . . . .	90
4.4.5	Further exploration . . . . .	91
4.4.6	Conclusion . . . . .	91
<b>5</b>	<b>WebAssociate Design and Implementation</b>	<b>93</b>
5.1	Research Methodology . . . . .	93
5.2	VLA Data preparation steps . . . . .	96
5.2.1	Data selection . . . . .	96
5.2.2	Data pre-processing . . . . .	97
5.2.3	Data transformation . . . . .	97
5.3	Diabetes Data preparation steps . . . . .	98
5.3.1	Data selection . . . . .	98
5.3.2	Data pre-processing . . . . .	99
5.3.3	Data transformation . . . . .	99
5.3.4	Section summary . . . . .	99
5.4	Design and Implementation . . . . .	100
5.5	Prototyping WebAssociate . . . . .	101
5.5.1	Rare item problem . . . . .	102
5.5.2	Variable of interest . . . . .	102
5.5.3	Rule set confidence . . . . .	102
5.5.4	Further organisation of discovered AR . . . . .	104
5.5.5	Chi-square test . . . . .	105
5.5.6	WebAssociate -Further Improvements . . . . .	105
5.5.7	Hypothesis - possible explanations . . . . .	108
5.5.8	Three Sections of WebAssociate . . . . .	109
5.6	Discovery Methods of WebAssociate . . . . .	109
<b>6</b>	<b>Evaluation of WebAssociate</b>	<b>116</b>
6.1	Tools for the comparison . . . . .	116
6.1.1	MineSet . . . . .	116
6.1.2	Gnome Data Miner . . . . .	117

6.1.3	WebAssociate . . . . .	117
6.2	Ethics Approval . . . . .	117
6.2.1	Protection of Participants . . . . .	117
6.2.2	Confidentiality . . . . .	118
6.2.3	Informed Consent . . . . .	118
6.3	Software Evaluation Methods . . . . .	119
6.4	WebAssociate Evaluation . . . . .	120
6.4.1	User Credibility . . . . .	121
6.4.2	Software Validation . . . . .	125
6.5	Chapter Summary . . . . .	135
6.5.1	Gnome Data Miner . . . . .	136
6.5.2	SIG MineSet . . . . .	136
6.5.3	WebAssociate . . . . .	136
<b>7</b>	<b>Conclusion</b>	<b>138</b>
7.1	Conclusion . . . . .	138
7.1.1	Organisational use of KDD . . . . .	138
7.1.2	Non-technical domain experts . . . . .	139
7.1.3	Grouping AR for hypothesis suggestion . . . . .	140
7.1.4	Visualisation . . . . .	140
7.1.5	Development of WebAssociate . . . . .	141
7.1.6	Software Evaluation . . . . .	141
7.2	Limitations . . . . .	142
7.3	Further Research . . . . .	142

# List of Figures

2.1	KDD steps . . . . .	21
2.2	Rare Items . . . . .	30
2.3	Association Rules represented by IBM <i>Intelligent Miner</i> . . . . .	46
2.4	Single item Association Rules represented by SGI <i>MineSet</i> . . . . .	46
2.5	Multiple item Association Rules represented by SGI <i>MineSet</i> . . . . .	47
2.6	3-D AR visualisation . . . . .	48
2.7	Mosaic Plot for student data . . . . .	49
2.8	Associative Map . . . . .	49
3.1	Domain Experts and Data Miners . . . . .	52
3.2	Relationships and Dependencies Between Domain Experts and Data Miner	53
3.3	Balanced Relationships and Dependencies Between Domain Experts and Data Miner . . . . .	54
3.4	KDD implementation - Phase 1 . . . . .	56
3.5	KDD implementation - Phase 2 . . . . .	56
3.6	KDD implementation - Phase 3 . . . . .	57
4.1	Three different knowledge discovery approaches . . . . .	63
4.2	Database Selection . . . . .	63
4.3	Table Selection . . . . .	63
4.4	VLA dataset attribute selection . . . . .	64
4.5	Exploring country attribute . . . . .	66
4.6	Refusal rate exploration . . . . .	67
4.7	Variables of interest selection . . . . .	67
4.8	Australian and Italian VLA applicants . . . . .	68
4.9	Mapping Association Rules to the hypothesis . . . . .	68
4.10	Null and alernative hypothesis for the refused Australian and Italian VLA applicants . . . . .	69
4.11	Chi-square test for the refused Australian and Italian VLA applicants . . .	73



4.12	Lift: Attribute-values that contribute to refusal greater than 26.7%	73
4.13	Suggested reasons for the higher refusal rate of the Italian born applicants	74
4.14	Law Type - probability to be refused legal aid	74
4.15	Age Groups - probability to be refused legal aid	75
4.16	40% of all refused - matter code FO (Family Other)	75
4.17	Italy and Australia - Group Characteristics	76
4.18	Attribute selection	77
4.19	UK countries selection	78
4.20	UK group definition	78
4.21	New group definition	78
4.22	OZ group definition	78
4.23	Country selection - Muslim group	79
4.24	Muslim group definition	80
4.25	Consequent selection for the groups of interest	80
4.26	Refusal rate for the UK, OZ and MUSLIM groups	81
4.27	Reason for refusal for the MUSLIM group	81
4.28	Refused on the basis of guidelines	81
4.29	Country Vietnam selection	83
4.30	Consequent selection for country Vietnam	84
4.31	Frequent matter codes for the Vietnamese	84
4.32	The number of Vietnam applicants and drug related matters	84
4.33	Characteristics of the Vietnamese and drug related matters	85
4.34	Matter code RD and age selection	86
4.35	Drug related matters and 19 to 25 years old Vietnam applicants	86
4.36	Vietnam, age 19 to 25, drug related applicants	86
4.37	Characteristics of the Vietnam born, age 19 to 25, drug related applicants	87
4.38	Diabetes attribute selection	88
4.39	Diabetes attribute selection - current smokers	89
4.40	75% of the current smokers are from the DGP sites 573, 847 and 954	90
4.41	Rule: <i>SmokerCurr_1</i> $\Rightarrow$ <i>siteID_573</i> [confidence 29%]	90
4.42	Current smokers and site id 573 - Characteristics	92
6.1	Software Ease of Use	122

# List of Tables

2.1	Calculating statistical significance by using contingency table . . . . .	38
2.2	Example of AR in rule table format . . . . .	44
2.3	Contingency table for student data with two attributes . . . . .	47
5.1	<i>sex</i> file example . . . . .	111
5.2	Unsorted association rules . . . . .	112
5.3	Sorted association rules . . . . .	112
5.4	User specified rule sets . . . . .	113
6.1	User satisfaction - discovered pattern representation . . . . .	123
6.2	User satisfaction - feedback from the tool . . . . .	123
6.3	Task 1 - percentage of completion . . . . .	129
6.4	Task 2 - percentage of completion . . . . .	132
6.5	Task 3 - percentage of completion . . . . .	135

# Chapter 1

## Introduction

### 1.1 The Information Age

Mankind has progressed from the Agricultural age to the Industrial age, and recently to the Information age. Factory workers are being outnumbered by knowledge workers who use computers to gather, exchange and manipulate digital information [32]. The current hardware and software technologies allow efficient and inexpensive data exchange and storage which enables companies to gather and store large volumes of data.

Each year more operations are being computerised and virtually all large to mid-size organisations store information as data. Furthermore, many research based organisations collect enormous amounts of data. For example, each day NASA stores terrabytes of data scanned by various satellites. However keeping large data in warehouses without discovering knowledge from it is rarely of direct benefit. The true value of keeping large amounts of data is predicated on the ability to extract useful information from it in order to improve business decisions and optimize success.

### 1.2 Knowledge Discovery from Databases - KDD

Traditional manual data analysis is becoming impractical in many domains as data volumes grow exponentially. The manual probing of a dataset is slow, expensive and highly subjective. Fayyad and Uthurusamy claim that the human ability to analyse and understand massive storages of data is falling far behind the ability to gather and store the data [26]. The need for automated data analysis techniques and tools is emerging in order to assist humans and improve their analysis capacity. Because current hardware and software have enabled organisations to gather massive volumes of data, it is only natural to use computational techniques to assist analysts in pattern discovery. According to Brachman

et al. [9], Knowledge Discovery from Databases (KDD) had become an umbrella name for all the new computational techniques that intelligently assist humans in analysing large structured data sets. Hence, KDD is an attempt to address a problem that the digital revolution made possible: data overload. At an abstract level, as Fayyad et al. claim, the KDD field is concerned with the development of methods and techniques for making sense of data [25].

Depending on the type of analysis, several KDD methods, such as classification, regression, clustering and association rules, use automated artificial intelligence, mathematical and statistical techniques for this task. Classification is a KDD method that maps (or classifies) a data item into one of several predefined categorical classes. Regression maps a data item to a real-value prediction variable. Clustering finds natural groupings of data items based on similarity metrics and maps a data item into one or several categorical clusters. Association rules (AR) depict how frequently two or more data items appear together and identifies a relationship between those items.

A well known KDD definition by Frawley in [27] states that

*KDD is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data.*

As Fayyad and Uthurusamy [25] claim, the term *non-trivial* means that some search or inference is involved; that is, KDD is not a straightforward computation of predefined quantities like computing the average value of a set of numbers. Therefore knowledge discovery from databases should be seen as a process containing several steps with many decisions made by the user.

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps; data selection, data preprocessing, data transformation, data mining, discovery interpretation/evaluation.

### **Pre data mining steps**

The pre mining steps involve data selection, data preprocessing and data transformation. The aim of pre data mining steps is to prepare the data set for mining. By selecting data of interest, dealing with missing and invalid values and finding useful features to represent the data depending on the goal of the task, the pattern discovery is more likely to be fruitful. In the data selection step, the user is involved in determining which data set will be used in the process. By selecting a data set, the user is focusing on a subset of variables or data samples on which discovery is to be performed. As discussed by Wright, an inappropriate process of data selection will affect the whole project and may introduce uncertainty [81]. Mining irrelevant data with an association rules generator is computationally expensive

and prone to error. Therefore, eliminating irrelevant variables and restricting a data set is important.

Data pre-processing is a necessary step for resolving several problems that occur in large data-sets. Large data-sets typically contain noisy data, missing data or irrelevant data and therefore data reduction is incorporated into the data pre-processing stage. Noisy data is considered as invalid data such as date of birth *2970* instead of *1970*.

Wright [81] suggest that missing data values can present a serious processing dilemma, as they can lead to erroneous conclusions about data. However, substitution of missing values can introduce inaccuracies and inconsistencies. Strategies for handling missing data should be developed during the stage of pre-processing data. The pre-processing step involves correcting those types of errors or discarding records that cannot be corrected.

Data transformation includes finding useful features to represent the data, depending on the goal of the task. Data transformation may be used to represent data in a manner that is acceptable to the data-mining algorithm. For example, data item date of birth *19/03/1963* would be inappropriate for association rule generation. In order to make the values from this field more useful, the user would transform a date of birth to appropriate age group (e.g. transform date of birth *19/03/1963* to age group 40-50).

### **Data mining step**

The actual data mining task involves choosing the data-mining method, selecting the algorithm to be used for pattern discovery and actual mining. This step includes matching a particular data mining method with the overall criteria of the KDD process. For example, the user might be more interested in understanding the model than its predictive capabilities. Subsequently, the user would select appropriate algorithm which is available for the selected method. There are, for example, several association rule algorithms, as advanced in [53, 68, 79, 6, 76, 10, 2] used to depict associations between data items. The data mining step involves searching for patterns by applying the selected method to the data set.

### **Post data mining steps**

The aim of post mining steps is to analyse and interpret discovered patterns in order to make use of the new knowledge. After successful data mining, the generator will produce lots of different patterns that are still in the raw state which is not representing knowledge in a clear and precise form. According to Wright in [81], performing this step properly requires domain knowledge acquired from domain experts or data repository. As Fayyad and Uthurusamy [26] state, interpretation includes understanding the discovered patterns and possibly returning to any of the previous steps, as well as possible visualisation of the

extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users.

KDD has been used in many real life applications. There are a significant number of commercial and academic KDD experts who applied KDD to many different domains such as business [9, 2, 26, 76], scientific [24, 26], education [61], health [41, 68, 1], and law [67, 71, 85, 80, 33, 44]. KDD methods and tools are being gradually transformed from research to a real world organisational use. However, many organisations do not take full advantage of KDD. While some organisations believe that statistical and DBMS query methods are sufficient for data analysis, others use KDD partially. This means that those organisations do not use KDD in all organisational levels.

### 1.3 Organisational use of KDD

According to Brachman et al. [9], an increasing trend in KDD shows that companies rely on the analysis of huge amounts of data to gain competitive advantage. Many organisations use discovered knowledge to gain competitive advantage, increase efficiency, or provide more valuable services to customers. However, in order to use the full potential of KDD, organisations have to pass through several stages. As discussed by Goebel and Le Gruenwald, there are usually three stages at deploying KDD technology in an organisation [32].

The initial stage involves the organisational use of KDD through an external KDD specialist (external consultant). In this stage, an organisation approaches a third party company which is specialised in KDD. The KDD specialist uses domain knowledge, either through a domain expert or a domain knowledge repository, in order to select and preprocess the data set. In the next step the specialist would search for patterns in data by applying an appropriate KDD technique. Finally, the discovered knowledge would be presented (in some meaningful manner) to the organisation, in order to decide which patterns are potentially interesting and useful. The useful patterns would be further used according to the organisational needs.

The second stage involves the organisational use of KDD through an internal KDD specialist or team of analyst experts (e.g. database administrators and data analysts). In this stage, an organisation purchases a KDD technology (hardware and software) which would hopefully meet the analysis requirements of the organisation. In order to use the software, the organisation would traditionally organise appropriate training of its IT department personnel, and subsequently apply the KDD process.

The last stage involves the full exploitation of KDD technology within the organisation.

It includes the use of KDD by enabling domain experts (e.g. lawyers, managers and medical professionals) to perform their own analysis according to their individual needs. This step does not eliminate the use of KDD in any previous stage. Moreover it enhances the use of KDD within an organisation by allowing domain users to search for useful knowledge that would potentially improve their everyday tasks. Although widely still a vision, the necessity for this stage is clearly recognized.

As previously discussed, domain experts play an important role in any KDD process. However, there are some domain experts that are less skilled in complex data analysis and have less knowledge of the nature of the data available than others.

### 1.3.1 Domain Experts

KDD is an iterative process which involves human interaction. There are traditionally two human roles in any KDD process; a domain expert role and a data miner role. A data miner is someone who primarily uses sophisticated KDD technology in conjunction with existing data sources as the basis for discovering useful patterns in data. A domain expert is a person with a comprehensive knowledge of a certain domain. The role of a domain expert was discussed in the literature by [55, 54, 59, 49, 84, 4, 44] and the researchers claim that the availability of actively strong domain knowledge improves efficiency of the knowledge discovery process by reducing the search space and helping to focus on interesting findings. This means that the role of a domain expert is to improve the discovery process by being actively involved in all pre and post mining steps. For small domains, one person can be a domain expert. In larger, more complex domains, a specialist may take over the task of detailing particular partitions of the domain that is he/she is knowledgeable with.

The role of a data miner in the first and second phase of KDD use in organisations is to select an appropriate data mining technique and drive the discovery process. However, some domain experts are data miners (e.g. data analysts and database administrators), some domain experts are not data miners (e.g. lawyers, managers and doctors) and some data miners are not domain experts (e.g. external KDD specialist). In an organisation deploying KDD at stage one, an external KDD consultant has the data miner role and an employee with sufficient domain knowledge has a domain expert role. In the absence of a domain expert it would be also possible for the data miner to use a domain knowledge repository. An individual may have roles. For example, in an organisation deploying KDD at stage two, very often a data analyst would have both KDD roles. Even though data analysts have a primary data miner role, by working with their organisation's data sets, they become domain experts.

As Fayyad et al. [25] claim, the true value of enabling domain experts to discover

patterns in data lies in their ability to extract useful reports, spot interesting events and trends, support decisions and policy based on statistical analysis and inference, and exploit the data to achieve business, operational or scientific goals. However, the ability to use KDD in order to achieve these goals is dependant on the type of domain expert. There are broadly speaking, two types of domain expert; those that practice and understand data mining such as IT personnel (where data mining is their primary role) and those who are non-technical, such as lawyers, managers and health care professionals (domain expert is their primary role). The former are able to perform each of the KDD steps; data selection, data transformation, data mining and interpretation of results. The latter however are non-technical “pure” domain experts and require additional interaction with data miners.

One of the reasons that exploitation of KDD technology within the organisation (stage three) is not fully implemented, is because the majority of KDD tools currently available are expensive, complex adjuncts to database management systems. Their operation typically requires specialist operators. Furthermore, the countless data mining techniques function in such different ways that even KDD experts cannot be expected to be proficient with all approaches. The specialist knowledge required and the cost of KDD tools mitigates against their use by non-technical domain experts.

The majority of KDD tools require a prohibitive amount of training before being useful, and discovered patterns are often difficult to interpret. One can argue that training non-technical domain experts would overcome this problem. However, most non-technical domain experts are usually not interested in using advanced technology, but only in getting clear, rapid answers to their everyday business questions [32]. Non-technical domain experts require simple-to-use tools that efficiently solve their business problems.

In this study, we identify this problem and present a KDD method that is used by non-technical experts with minimal training to discover and interpret patterns that they find useful for their role within an organisation. The approach generates association rules and then displays them by grouping rules together and visually depicting deviations between groups.

## 1.4 Association Rules

Among many different techniques used to extract useful knowledge from databases, Association Rule mining, has in recent years, attracted the attention of data mining communities [2, 49, 68, 26, 76, 10, 6, 79, 53]. Association rule mining is a form of data mining used to discover interesting relationships amongst two or more attributes in data. Association rules were introduced by Agrawal et al. [2] and originated with the problem of supermarket basket analysis. In basket analysis association rules were used to find associations



between the items bought by a customer in order to find which items were frequently bought together. The findings can be used to understand customers' buying habits and preferences in order to increase profits.

An association rule is an expression of the form  $X \Rightarrow Y$  [*support, confidence*], where  $X$  and  $Y$  are sets of items that are often found together in a given collection of data. The attribute group on the left hand side of the arrow is called the antecedent or "left hand side" (LHS) and the group of attributes on the right hand side of the arrow is called the consequent or "right hand side" (RHS). The interestingness and usefulness of an association rule has been based on support and confidence measures. The support is the percentage of transactions in the database containing both  $X$  and  $Y$ . The confidence is the conditional probability of  $Y$  given  $X$ , e.g  $confidence = P(Y | X)$  .

In this study we use Association rules as a descriptive KDD method which allows users to visually explore data sets in order to find possible interesting, previously unknown and useful associations in data.

#### 1.4.1 Pattern Interestingness

While association rules have proven to be useful in practical applications, AR algorithms tend to generate large numbers of rules, most of which are of little interest. Furthermore, large numbers of rules are difficult for analysts to interpret. Confidence and support thresholds were introduced by Agrawal et.al. [2] and represent an objective measure that filters interesting rules. If both support and confidence values are greater than the threshold, the association rule is considered interesting. However, many analysts do not know what an ideal threshold setting should be [59]. If the threshold is set too high, useful rules may be missed. If it is set too low, the user may be overwhelmed by many irrelevant rules. In many real life applications some items appear very frequently in the data, while others rarely appear. If the threshold is set too low those rules that involve rare items will not be found. This dilemma is called the "*rare item problem*". Lin et al. and Liu et al. argue that most of the rules with high support are obvious and well known, and it is the rules with low support that provide interesting new insights [53, 57].

According to Liu et al. and Klementinen [55, 49], objective measures alone are therefore insufficient for determining a discovered rule's interestingness. Additional measures that involve domain expert opinion are needed. Subjective interestingness can be determined by modeling and incorporating domain knowledge into the system [73]. Most existing approaches ask the user to explicitly specify what types of rules are interesting and uninteresting and the rules that match user's specifications are retrieved. The authors in

[6, 38, 75] review several methods proposed for finding the most interesting rules using a variety of metrics.

However, some domain experts are already familiar with common patterns in the data through their years of the experience and their intuition [14]. It is very unlikely for those experts to be satisfied with merely prevalent patterns because presumably the organisation is already exploiting that knowledge.

The greatest benefit to a domain expert would be a tool that can show differences or changes in the relationship between attributes in data.

## 1.4.2 Group Differences

One of the most promising areas in KDD is the automatic analysis of changes and deviations [68]. With deviations we have a simple way to identify things that differ from our expectations. Since they differ from what we expect, they are by definition interesting. Interestingness measures based on deviations were used by [68, 14, 20, 5, 29, 83, 22, 44].

There are two different approaches that use deviations to measure interestingness of a discovery: measuring differences over time and measuring differences between groups. For example, Piatetsky-Shapiro and Matheus in [68] define a deviation as a difference between an *observed value*  $V_o$  and a *reference value*  $V_r$ . The observed value is taken from the most current snapshot of the database. Comparing the observed value to one from the previous time period generates a deviation over time.

In our study we measure interestingness of discovery by measuring differences between groups. A group represents a variable of interest (attribute-value) selected by the user. For example, by selecting attribute-values *sex\_Male* and *sex\_Female* as variables of interest the user is able to look for differences between these two groups. We organise generated association rules by separating the discovered rules into rule sets. Each rule set contains association rules that share a common consequent. For example association rules *sex\_Male*  $\Rightarrow$  *age\_30..40* and *sex\_Female*  $\Rightarrow$  *age\_30..40* represent a rule set because they share the same consequent. If the discrepancy between the confidence values of these two association rules is substantially high, these groups are considered different on the basis of age group 30 to 40, otherwise the groups are considered similar on this age group. Such findings can be useful to suggest hypotheses to the user. For example, if male and female groups are considered similar on the basis of age group 30 to 40, the user may infer the null hypothesis “*There is no difference in the proportion of 30 to 40 age group between the males and females*”. By grouping related association rules into rule sets, (association rules with a common RHS but different LHS), we are able to visually display groups (user selected variables of interest) and their differences.

In our study the value of confidence as a single measure does not reflect interestingness. Our goal is to group related association rules into rule sets, automatically calculate deviations between confidences in each rule set and graphically display findings. Analysts then decide how surprising the finding is. Traditional AR generators discover association rules without grouping the discovered rules. The interestingness of a generated association rule is based on the minimum support and confidence threshold. Association rules that meet the user specified threshold are considered to be interesting. Nevertheless, the rules with high confidence are not necessarily interesting if they were previously known and expected by the user. An expert may be already familiar with the rule “*country of birth **England** ⇒ legal aid refused **Yes** [support 8%, confidence 10%]*”. This rule shows that 10 percent of English born applicants have been refused aid. However, if this rule is grouped and visually displayed together with the rule “*country of birth **Italy** ⇒ legal aid refused **Yes** [support 0.5%, confidence 27%]*”, showing that 27 percent of Italian born applicants have been refused aid, an analyst could find the difference in the refusal rate interesting and seek to invent a hypothesis that will explain the difference.

The link between the hypothesis and rules is therefore just as important when the association rules are used to suggest a hypothesis, as they are when the rules are used to confirm a hypothesis. In order to explain the difference in the rejection rate, the expert may infer several hypotheses: “*1. More Italian born applicants are older. 2. Elderly applicants are wealthier. 3. Elderly and wealthy applicants are refused aid because they have more assets than younger applicants and therefore fail on the means test*”, and seek additional association rules to confirm this.

In this study we identify the requirements of domain experts and approach interestingness from a hypotheses testing point of view. In the process of hypotheses generation, the user is guided by the feedback of the visualizations and quickly learns more about the properties of the data in the database. The confidence value as a single measure does not reflect interestingness. We present a KDD method that displays a set of grouped association rules and their deviations in order to suggest hypotheses to the user. According to deviations between groups for each rule set, an expert is able to visually scan the suggestions, decide how surprising the finding is and subsequently test plausible hypotheses or seek possible explanations.

## 1.5 Discovery Visualisation

Liu et al. in [60] claim that the problem of interestingness is not due to the large number of discovered association rules. The main limitation is with the user's inability to organize and present the rules in such way that they can be easily analysed. What is the most useful way to present data patterns to users? One approach is to organize rules by grouping the related rules together. This approach has been used by Piatetsky-Shapiro and Matheus, Bay and Pazzani, and Aumann [68, 20, 22, 5]. We find the organisation of discovered rules by grouping the related rules together very interesting, promising and useful for our study. However Piatetsky-Shapiro and Matheus, Bay and Pazzani, and Aumann in [68, 20, 22, 5] present the findings as text. We believe that more appropriate way to present these findings is by visualising groups and their deviations. The problem with non-visual representation is that association rules are not always easy to understand particularly if the rules are complex [40]. After useful knowledge is discovered, a visual presentation of the discovery plays an important role in KDD. The assumption is that when users see how data items are related to each other, they will have a better understanding of the discovered patterns [82]. There are several different approaches used to visually display association rules such as directed graphs, 2-D matrix and 3-D visualisation [40, 82, 30, 86, 19, 18, 36, 31, 56, 31].

Directed graphs are used in the IBM (Intelligent Business Machines) KDD software called "Intelligent Miner" to visualise association rules. In a directed graph each node represents a unique attribute-value and an arc connecting two nodes represents the association between the nodes as shown in Chapter 2, Figure 2.3. While directed graphs are useful where only a few nodes and arcs are involved, a great number of discovered association rules would make the graphical display difficult to comprehend. Directed graphs are further discussed in Chapter 2.

A software company, called Silicon Graphics, Inc. (SGI) is the world's leader in high-performance computing, and visualization. KDD software developed by SGI called "Mine-Set" uses powerful graphics for visualising discovered association rules by displaying association rules in a "2-D matrix" format. Two dimensional matrix positions the antecedents and consequent attribute-values on the X and Y axis respectively as shown in Chapter 2, Figure 2.4. The confidence and support of a rule is illustrated by using the height and color of a bar. While the 2-D matrix approach is appropriate for single attribute-value associations (rules with single antecedent and single consequent), 2-D matrix graphs are not designed for rules with multiple attribute-values. The 2-D matrix approach is further discussed in Chapter 2.

Three dimensional (3-D) representation of discovered association rules, introduced by Wong et al. [18], overcomes the problems of 2-D approach. Instead of using "attribute-

value to attribute-value” matrix approach, this technique uses a ”rule-to-item” map. In this approach the rows are attribute-values and the columns are rules as illustrated in Chapter 2, Figure 2.6. The text on the right hand side of the graph display attribute-value labels. The antecedents and consequents are distinguished by using two different colors, while the height of the bar displayed at the end of the matrix represents the confidence of the rule. The 3-D visualisation approach is further discussed in Chapter 2.

While 3-D representation of the association rules is an improvement over the directed graphs and 2-D matrix approaches, none of the visualisation approaches discussed above, is an appropriate visual technique for displaying groups and their differences. In this study we involve domain experts to evaluate three different association rule representations (our approach, MineSet and Gnome Data Miner), in order to test which approach is more appropriate for the representation of grouped association rules and their deviations. Domain experts using our approach report favorable responses.

## 1.6 Motivation

The main focus of our work is to enable a domain expert to investigate hypotheses without assistance. The effectiveness of our methods is demonstrated with a web based data analysis tool that generates association rules from a large data set drawn from over 380,000 applications for legal aid in Australia. Applications for legal aid are made to a semi-government legal aid organization called Victorian Legal Aid (VLA). VLA aims to use KDD techniques illustrated here to further their objectives of providing legal aid in the most effective, economic and efficient manner to those in the community with the greatest need.

Association rules, as a descriptive KDD method, enable VLA domain experts to discover characteristics of the population of VLA applicants, that is to discover who applied for legal aid or who was more refused or approved. Furthermore by grouping association rules the experts are enabled to find differences between groups (e.g. differences in characteristics between distinct sex, country of birth or age groups). For example, by selecting the country of birth attribute (as the antecedent), a VLA expert could find a group of applicants that were refused legal aid more than any other country groups.

The University of Ballarat and Victorian Legal Aid have been working together for some time and established close links between VLA experts and researchers. In our previous work [44] we used the discrepancy between a domain expert’s confidence prediction and

the actual confidence for an association rule as a measure of subjective interestingness. Before association rule generation, a user was asked to select a variable of interest to generate rules containing that variable as consequent. Before displaying confidence values for each generated rule, the user was asked to predict a confidence value. Rules that had confidence values that surprised domain experts were considered interesting.

However, we recognised that users often did not know what the confidence threshold should be, so the need for a new interestingness measure emerged. Furthermore, users had difficulties understanding more complex rules, especially rules with multiple data items. We identified the necessity to organize and present the rules in such a way that they can be easily analysed by the user. This requirements led to a simple visual representation of the discovered rules. We also identified that users have a need and desire to have easy-to-use timely (just in time) data analysis tools.

In order to meet the requirements of a domain expert (especially non-technical), we were motivated to explore different approaches and enable such users to use additional KDD tools for their individual and organisational needs.

### **1.6.1 Problem 1 - Exploitation of KDD technology within the organisation is not fully implemented**

Exploitation of KDD technology within the organisation (stage three) is not fully implemented because the majority of KDD tools currently available are expensive, complex adjuncts to database management systems, requiring an unaffordable amount of training before being useful. Most current KDD tools are being used in organisations by external KDD experts or in house data analysts. This means that most organisations are deploying KDD at stage one or stage two. In order to use the full potential of KDD through all organisational levels, additional KDD tools are needed to cater for the individual needs of non-technical domain experts.

#### **Our contribution**

In this research, we identify the needs of non-technical domain experts and build an easy-to-use web based KDD tool called “WebAssociate” that uses visual aids in order to meet “non-technical” domain expert individual requirements by suggesting or testing hypotheses. We evaluated the tool against two other commercial data mining tools; MineSet (by SGI) and Gnome Data Miner. The evaluation process involved domain experts using the tools to solve a typical business problem. They evaluated the usability, usefulness, user satisfaction and validation of each tool. The experts report favorable responses for “WebAssociate”.

## 1.6.2 Problem 2 - Users may be overwhelmed by the number of discovered rules

The majority of KDD tools use pruning methods in order to reduce the number of discovered rules. Pruning methods are based on interestingness measures such as confidence and support. If both support and confidence are greater than the user's specified threshold the rule is considered interesting. However, many users do not know what the threshold settings should be [59]. Furthermore, if the threshold is set too high, users miss rules that contain rare data items. If the threshold is set too low, users are overwhelmed by the number of discovered rules.

### Our contribution

In this study we organize and visually present the rules in such way that they can be easily analysed by the user.

- Organisation of the discovered rules

In the initial step, we organise discovered association rules by grouping discovered association rules into rule sets. A rule set contains association rules that share the same consequent but different antecedent. We define these rule sets as *iso\_consequent* rule sets. However the antecedents for every discovered rule set are values of the same attribute. For example, male and female antecedents are distinct values of the sex attribute.

We further organise *iso\_consequent* rule sets by classifying them into two classes: similar and different. For each *iso\_consequent* rule set we find an AR with a minimum confidence value and an AR with a maximum confidence value. An *iso\_consequent* rule set is classified as similar if the difference between the maximum and minimum confidence values is smaller than the user specified value (default is 5%), otherwise a *iso\_consequent* rule set is classified as different. This approach allows the users to view the differences and similarities between groups in a data set.

- Discovery Pruning

In contrast to current pruning methods based on the single AR, we allow the user to prune discovered rules by introducing a threshold called a *ruleSet\_confidence* with a value in the range between 0 and 100 (default is 10%). The user can optionally select AND or OR options for this threshold. By selecting the AND option, an *iso\_consequent* rule set will be included in the discovery only if **all** confidence values for each AR in the *iso\_consequent* rule set are higher than the user specified *ruleSet\_confidence*. However, by selecting the OR option, an *iso\_consequent* rule set will be included in the discovery if **at least one** of its ARs has the confidence value

higher than the user specified *ruleSet\_confidence*. Users find this approach useful for two reasons: firstly the users seldomly have to reset the default *ruleSet\_confidence* value, and secondly the threshold is based on the whole rule set and not on the single association rule which gives greater flexibility to the user.

- **Interestingness**

The actual interestingness is based on the visually displayed confidence values for each *iso\_consequent* rule set. By selecting a class of similar *iso\_consequent* rule sets, the user may be surprised by discovered similarities between the groups. However by selecting a class of different *iso\_consequent* rule sets, the user may find differences between the confidence values in each *iso\_consequent* rule set interesting. Moreover, our approach allows him/her to further explore the finding.

### **1.6.3 Problem 3 - Mapping of a hypothesis to the set of association rules that will confirm or deny the hypothesis is rarely simple**

The mapping between a hypothesis and association rules is not one to one and it is not that easy to imagine the correct association rules to investigate a given hypothesis without assistance. For example, in order to test the alternative hypothesis “*Female applicants apply more for family law type matters than for criminal law type matters*”, many users would have difficulties in constructing the appropriate rules. Two rules are needed for testing this hypothesis; *Female*  $\Rightarrow$  *Criminal* and *Female*  $\Rightarrow$  *Family*. However, many experts find this task difficult and very often would map an incorrect set of rules such as *Criminal*  $\Rightarrow$  *Female* and *Family*  $\Rightarrow$  *Female*.

#### **Our contribution**

In this study we introduce a KDD tool that enables the expert to focus more directly on the hypothesis under investigation than on the rules. The user may choose a population for study and the tool will automatically detect all appropriate groups for the selected population. This approach enables the non-technical users to test and explore hypothesis without having the complexity of manually mapping association rules to hypotheses. A tool that adds a connection between AR and hypotheses permits more effective explorations about data.



### 1.6.4 Problem 4 - Experts need additional methods to explain hypotheses

In real world applications, many experts search for explanations. For example, the null hypothesis “*There is no difference in the refusal rate between the Italian born applicants and Australian born applicants*” is rejected because there is a significant difference between the confidence values for association rules  $country\_ITALY \Rightarrow refused\_YES$  (conf. 26.7%) and  $country\_AUSTRALIA \Rightarrow refused\_YES$  (conf. 10%). The experts need additional methods to explain reasons why this null hypothesis is rejected. We believe that current KDD methods do not address this problem and do not have additional methods for this task.

#### Our contribution

We present a KDD tool which models the way that experts seek to explain patterns in data. In this research we introduce additional methods that search for reasons to explain selected hypotheses. The findings are visually represented and enable the user to explain why selected groups are different e.g. what contributed to significant difference in confidence values for the relevant association rules. For example, while evaluating the tool, the experts were able to explain that Italian applicants were rejected not because of their bias but because the majority of Italian applicants are older, male applicants that applied mostly for family matters. Elderly applicants fail to get aid because they tend to be wealthier and fail the means test.

## 1.7 Research Questions

1. What type of KDD tools are needed for non-technical domain experts?
2. How do we model the way that domain experts seek to explain patterns in data?
3. How can association rules be mapped to hypotheses?
4. What is the appropriate method to organise and present the findings to be more understandable to non-technical domain experts?
5. How can easy-to-use KDD tools for non-technical experts be constructed?

We have a particular idea here that has only been applied in the AR context. The full breadth of the question could only be taken up in further work.

## 1.8 The Organisation of this Thesis

The remainder of the thesis is organised as follows. Chapter 2 outlines some existing work related to KDD, human interaction in KDD, association rules, interestingness of findings and pattern visualisation. An overview of the human roles in KDD, such as the domain expert and data miner role, is given in Chapter 3. In Chapter 4 we provide real life sample consultations involving the VLA and Diabetes Australia domain experts using “WebAssociate”. In Chapter 5 we discuss our research questions and describe the methods that we used in order to address them. Chapter 5 also provides an overview of the methods used in the implementation of “WebAssociate”. The methodology used for evaluating “WebAssociate” against two commercial KDD products and evaluation results are discussed and shown in Chapter 6. In the final chapter we make some conclusions based on our experience during this study. In the final chapter we also discuss the limitations of “WebAssociate” and the possibilities for further research.

# Chapter 2

## Literature Review - KDD

This chapter provides an overview of the work done by others in the field of Knowledge Discovery from Databases (KDD). We also closely describe a KDD method called Association Rules (AR) and previous work involved with it. At the end of this chapter a summary and an overview of the remaining chapters is provided.

### 2.1 Data Mining

#### 2.1.1 Introduction

Current hardware and database technologies allow efficient, inexpensive and reliable data storage and access. The amount of data collected and warehoused in all industries is growing at a phenomenal rate [9]. As Fayyad et al. [26] state, raw data is rarely of direct benefit.

Accumulated data contains lots of valuable information for the data gatherers and its true value is predicated on the ability to extract information useful for decision support or exploration and understanding the phenomena governing the data source. As Fayyad et al. [25] claim, the true value of data lies in a user's ability to extract useful reports, spot interesting events and trends, support decisions and policy based on statistical analysis and inference, and exploit the data to achieve business, operational or scientific goals.

The traditional method of turning data into knowledge relies on manual analysis and interpretation. Such manual probing of a dataset is slow, expensive and highly subjective. In fact, manual data analysis is becoming impractical in many domains as data volumes grow exponentially. Our ability to analyse and understand massive datasets is falling far behind our ability to gather and store the data [25]. To overcome the problem of traditionally slow manual analysis, a new generation of techniques and tools is emerging

to intelligently assist humans in analysing mountains of data and finding critical nuggets of useful knowledge, and in some cases to perform analysis automatically [9].

The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of these new techniques based on a brute-force exploration of possible solutions. According to Fayyad et al. [26] a community of researchers and practitioners interested in the problem of automating data analysis has grown steadily under the label of Knowledge Discovery from Databases (KDD). KDD or Data Mining had become an umbrella name for all the new techniques that automated the information retrieval process. We use the term KDD to describe the overall process of discovering useful patterns in data, including not only the data mining step of running specific discovery algorithms but also data selection, transformation and post processing (proper interpretation of the results).

According to Zhi [88] due to its interdisciplinary nature, KDD has received contributions from many disciplines such as databases, machine learning, statistics, information retrieval, data visualization, parallel and distributed computing. The first three in the list, e.g. database, machine learning, and statistics, are undoubtedly the primary contributors. Zhou [88] finds that without the powerful data management techniques contributed by the database community and the practical data analysis techniques donated by the machine learning community, data mining would be seeking a needle in a haystack. It is interesting that even recently, a leading statistician urged that the statistics community should embrace data mining. This suggests that this community had not yet taken data mining seriously at least at that time.

Many KDD techniques are extensions of existing statistical methods. The term KDD is not new to statisticians. Jackson [46] finds that KDD is a term synonymous with *data dredging* or *data snooping* and has been used to describe the process of trawling through data in the hope of identifying patterns. Data snooping occurs when a given dataset is used more than once for inference or model selection as described by White [42]. The connotation is derogatory because a sufficiently exhaustive search will certainly throw up patterns of some kind, since by definition data that are not simply uniform contain differences that can be interpreted as patterns. The trouble is that many of these patterns will simply be a product of random fluctuations, and will not represent any underlying structure in the data. The objective of data analysis is not to model the fleeting random patterns of the moment, but to model the underlying structures that give rise to consistent and replicable patterns and help organisations focus on the most important information available in their existing databases.

Imielinski and Mannila [43] also claim that KDD has been built upon an existing body of

work in statistics and machine learning, but KDD provides completely new functionalities. In summary we can say that data mining does not replace traditional statistical techniques. As [17] states it is rather an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed.

According to Han [37] KDD has become a highly demanding task, attracted lots of researchers and developers, and made good progress in the past several years. In [78] the authors argue that although the term KDD became popular much less than a decade ago, it has become an important field attracting attention from both industrial users and research and development workers. KDD is becoming an indispensable decision-making tool in the ever more competitive business world and challenging applications inspire new techniques and affirm their utility.

Many organizations are using KDD to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers [17]. By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who have bought a particular product it can focus attention on similar customers who have not bought that product (cross-selling). By profiling customers who have left, a company can act to retain customers who are at risk of leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one.

KDD is commonly used for mining business databases using techniques and tools as described in [9, 2, 26, 76]. However, we find the use of KDD in many other domains such as education, health, law, sport and astronomy. Klementtinen et al. in [50] introduce the use of KDD in education by mining a student database in order to find associations between units that students take. In [61] KDD techniques are used to find potentially weak students for remedial classes. Both authors claim that the education domain offers a fertile ground for many interesting and challenging data mining applications.

Hsu [41], Piatetsky-Shapiro [68] and Abidi [1] demonstrate the application of KDD to the medical domain and claim that the medical domain offers a fertile ground for data mining applications. According to [1], today's healthcare enterprise, supported by Telemedicine based high-tech IT systems, also generates volumes of data health related facts. It is our contention that data repositories storing vital health data, need to be charged with harvesting data-driven decision support towards strategic planning and management of the healthcare enterprise. After eight years of gathering diabetic patient in-

formation, [41] apply KDD techniques in order to find useful knowledge that can be used by medical doctors to improve their daily tasks and to understand more about diabetes. Piatetsky-Shapiro and Matheus [68] also apply KDD techniques to the medical domain in trying to automatically acquire medical knowledge from clinical databases.

KDD has been applied in astronomy by Fayyad et al. [24]. In this work the authors demonstrate the application of KDD techniques to scientific data and suggest that digesting millions of data points, each with tens or hundreds of measurements is well beyond a scientist's human capability and can be turned over to data mining. According to Fayyad [26] the manual analysis of data in astronomy is no longer feasible as datasets in this field often exceed  $10^9$  records. In their work, KDD was used to classify huge amounts of data (images from the planet Venus) collected by satellite in order to find volcanoes.

KDD exercises in the legal domain attempted to derive knowledge about decision making processes in the legal domain automatically from data-sets. Zeleznikow and Stranieri [85] demonstrate mining Australian family law legal data in order to predict family law property outcomes. Stranieri and Governatori in [33] demonstrate that the field of KDD called "Association Rules" can be applied to facilitate the discovery of defeasible rules that represent the ratio decidendi underpinning legal decision making. Wilkins and Pillaipakkamnatt [80] apply KDD techniques to predict the time to case deposition. The authors examined data in existing databases from large numbers of cases in order to estimate the number of days that are likely to elapse between arrest and final deposition. There are other researchers who applied KDD to the legal domain such as Pannu [67] who applied KDD technique called "Genetic Algorithms" in order to discover new knowledge from previous court cases. Schweighofer and Merkl [74] use KDD clustering method called "Self Organising Maps" in order to organise queried legal documents returned by the web search engine. Rissland and Friedman [71] applied KDD technique called "Rule Induction" in order to detect changes in legal concepts. By using the ID3 rule induction algorithm the authors built decision trees from bankruptcy cases in the early, mid and late 90's. The decision trees differences were used as indicators for changes in legal concepts. In our previous work reported in [44] Victorian Legal Aid (VLA) domain experts use association rules for data analysis. In this work we apply association rules to a large legal data set in order to assist domain experts to suggest or confirm hypotheses.

KDD methods and algorithms have been used in many other domains including sport. In the recent study by Pingali et al. [69] the authors use visualization techniques to give

new insights into performance, style, and strategy of players. Automated techniques can extract accurate information from video about player performance that not even the most skilled observer is able to discern. For this purpose the authors developed a visualization system called LucentVision. LucentVision uses real-time video analysis to obtain motion trajectories of players and the ball, and offers a rich set of visualization options based on this trajectory data. The system has been used extensively in the broadcast of international tennis tournaments, both on television and the Internet. A similar use of KDD in sport has been reported by Rui et al. [72]. The authors focus on detecting highlights of a baseball game by using audio-track features in order to combine multiple sources of information. The output of KDD algorithms advanced in this study is compared against human-selected highlights for a diverse collection of baseball games with very encouraging results.

Which ever domain KDD is implemented in, the process is interactive and iterative with many decisions made by the user. As Mannila [62] claims, one cannot expect to obtain useful knowledge by pushing a lot of data to a black box. Thus KDD system should not be seen as an automatic analysis system, but rather as an interactive tool involving numerous steps.

### 2.1.2 Essential Steps to Knowledge Discovery

Many factors such as domain knowledge, data preparation (pre-processing), good criteria of interestingness and post processing (understanding of the discovered patterns) make the knowledge discovery process complex. KDD methods such as AR are just discovery tools, just like gold detectors are gold discovery tools. Gold detectors alone do not promise a succesful gold discovery, however other factors such as use of gold maps and domain knowledge could increase chances. In the same way, using AR as a discovery tool does not eliminate the need to know the business, to understand the data, or to understand the analytical methods involved. As an interactive and iterative process knowledge discovery from databases should be seen as a process containing several steps. Figure 2.1 shows an outline of the steps of a KDD process.

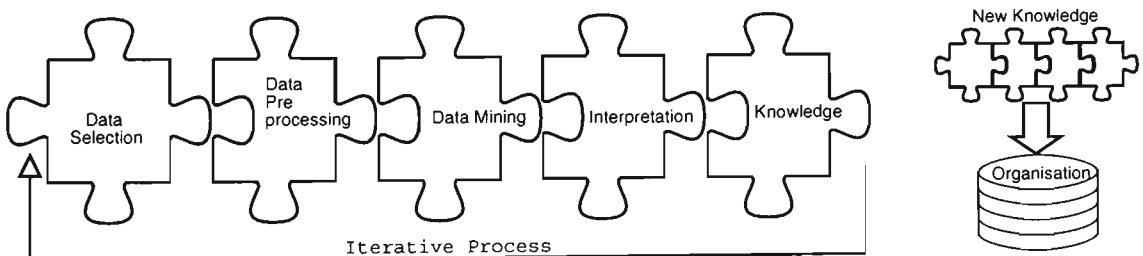


Figure 2.1: KDD steps

As Figure 2.1 shows, the overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

### 1. Data selection

Data selection involves the selection of a sample of data from a database of records. Using the whole dataset is often inappropriate and by deciding not to consider irrelevant data we speed up the discovery process. This step involves selecting a target set of variables for consideration or selecting number of records. For instance, in applying KDD to detect changes in law, Rissland and Friedman [71] restricted their sample to US Bankruptcy cases tried over a 10 year period.

### 2. Data pre-processing and transformation

#### *Pre-processing*

Data pre-processing is a necessary step for resolving several problems that occur in large data-sets. Large data-sets typically contain noisy data, missing data or irrelevant data and therefore data reduction is incorporated into the data pre-processing stage. Many organisations don't have appropriate methods to deal with invalid data entry. The pre-processing step involves correcting those types of errors or discarding records that cannot be corrected.

#### *Transformation*

Transformation includes finding useful features to represent the data, according to the goal of the task. Data transformation may be used to represent data in a manner that is acceptable to the data-mining algorithm. A simple example involves transforming continuous data e.g. date of birth into discrete values such as *age\_18..24*.

### 3. Data Mining

Different data mining methods and algorithms can be chosen to suit the discovery goal as well as the type and structure of data. Fayyad [26] believes that each method and its algorithms typically suit some problems better than others therefore a universally best data mining method is unlikely to exist. Thus, prior to the actual discovery process, an appropriate KDD method and algorithm should be selected.

#### *Selection of data mining approach*

The approach depends on the type and structure of the data being analysed as well as the discovery goal. There are many different KDD methods such as clustering, classification, association rules. The methods are described in section 2.1.3 of this chapter. Wright [81] suggests that often it is more effective to select more than one method or to use a combination of methods. For example, Freitas [28] shows



that if we want to discover prediction rules we would apply classification techniques, dependence modeling or other machine learning functions but we would typically not apply association rules techniques. On the other hand the authors in [77, 76, 2] claim that association rule functions are particularly well suited to the analysis of data for hypothesis suggestion.

#### *Selection of Algorithm*

An important point is that each algorithm typically suits some problems better than others. This process may be iterative and involve trying several algorithms until the most effective one is located [81]. For example, Han and Plank [35] find mining association rules as data intensive, therefore, the authors claim, it is essential to use efficient algorithms. Bayardo and Agrawal [6] addressed this problem and examined different algorithms used for rule generating with association rules. The authors report that almost every recently proposed association rule mining algorithm is a variant of the Apriori algorithm. The Apriori algorithm was introduced by [2] and involves a phase for finding patterns called frequent itemsets. A frequent itemset is a set of items appearing together in a number of database records. Apriori employs a bottom up search that enumerates all frequent itemsets. Hence a large portion of the applications effort should go into properly formulating the problem (asking the right questions) rather than optimising the algorithmic details of a particular data mining method [26].

#### *Data mining*

Once an appropriate KDD technique and algorithm is determined, it must be applied to the data. Data mining is the actual process of searching for interesting patterns. The results of analysis could be fed back into the modelling and hypothesis derivation process to produce improved results on subsequent iterations.

### **4. Interpretation**

After successful data mining, the generator will produce lots of different patterns that are still in a raw state and not in a clear and precise form. Interpretation involves the integration of existing domain knowledge, which may confirm, deny or challenge the newly discovered patterns[9]. This includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualisation of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable to the user [26].

### **5. Knowledge**

#### *Using discovered knowledge*

The use of discovered knowledge includes documenting and reporting the findings to interested parties as well as taking action based on the knowledge. For example, useful knowledge and discovered rules gained through the data mining process may provide crucial information that expert systems need for decision making.

All steps are important in the KDD process as [26] claims. Most previous work on KDD focused primarily on the data-mining step. However, [9] states that the process of data mining typically takes only a small part (estimated 15%-25%) of the effort of the overall process. We can conclude that the other KDD steps are equally if not more important for the successful application of KDD in practice.

As previously discussed, many different data mining methods and techniques can be used to find interesting patterns in data in order to extract useful knowledge from it. The selection of an appropriate technique depends on the the discovery goal as well as on the type and structure of the data.

### 2.1.3 Data Mining Techniques

According to Kumar and Zaki [51] most basic data mining techniques use two types of methods: prediction methods and description methods. Prediction methods use some variables to predict unknown or future values of other variables. Regression, deviation detection and classification algorithms are classified as predictive methods. Description methods find human-interpretable patterns that describe the data. Data mining algorithms such as association rule discovery, sequential pattern discovery and clustering are classified as descriptive methods.

Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. According to Berkhin [7], from a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, Customer Relationship Management (CRM), marketing, medical diagnostics and computational biology.

#### Clustering

Clustering maps a data item into one or several categorical classes (or clusters) where the classes must be determined from the data, unlike classifications in which classes are predefined. Clustering algorithms segment the data into groups of records, or clusters, that have similar characteristics. In marketing for instance, a health insurance company may discover that these characteristics define a segment: 20 to 40 years old, technical worker, fewer than two children, television science fiction fan, and an income of \$50,000 to

\$60,000 per year. The segment can be targeted more effectively with a health insurance package well suited for these people, by using television advertisements in new science fiction episodes.

Berkhin [7] defines clustering as

*A division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. (page 2)*

Pattern proximity is usually measured by a distance function defined on pairs of patterns. A simple distance measure like Euclidean distance can often be used to reflect dissimilarity between two patterns. The authors in [47] give a similar definition of clustering

*Clustering is the organization of a collection of patterns usually represented as a vector of measurements, a point in a multidimensional into clusters based on similarity. (page 265)*

Representing the data with fewer clusters necessarily loses certain fine details, but achieves simplification by modeling data into clusters. In [47] clustering is described as a subjective process; the same set of data items often needs to be partitioned differently for different applications. This subjectivity makes the process of clustering difficult. This is because a single algorithm or approach is not adequate to solve every clustering problem. A possible solution lies in reflecting this subjectivity in the form of knowledge. This knowledge is used either implicitly or explicitly in one or more phases of clustering. Knowledge-based clustering algorithms use domain knowledge explicitly.

Clustering methods are used in a real life applications such as document clustering (finding groups of documents that are similar based on the terms appearing in them) [87], medical diagnosis [3], cancer clustering - finding cancer clusters based on gene expressions [52], spatial database applications (Geographical Information Systems or astronomical data) [64, 11], DNA analysis in computational biology [13] and customer profiling [16].

## **Classificaton**

As [17] outlines, clustering is a way to segment data into groups that are not previously defined, whereas classification is a way to segment data by assigning it to groups that are already defined. Data classificaton involves classifying a set of data based on their values in certain attributes. Classificaton methods are used both to understand the existing data and to predict how new instances will behave.

In the real life example by [15], it is desirable for a car dealer to classify its customers according to their preference for cars so that sales persons will know whom to approach,

and catalogs of new car models can be mailed directly to those customers with identified features so as to maximise the business opportunities. In another example as discussed in [17], a company may want to predict whether individuals can be classified as likely to respond to a direct mail solicitation, vulnerable to switching over to a competing long-distance phone service, or a good candidate for a surgical procedure. Classification is also used in fraud detection, predicting fraudulent cases in credit card transactions (when does a customer buy, what does he buy, how often he pays on time) [8, 65]. Many banks use classification methods to predict the loyalty of a customer by using detailed records such as how often a customer calls, customer's marital status, financial status and age. For example Pan and Lee [66] use classification methods for customer relation management to label the customer as loyal or disloyal.

As the authors [23] claim, the discovery goal is subdivided into prediction, where the system finds patterns for predicting the future behavior of some entities, and description, where the system finds patterns for presentation to a user in a human-understandable form. While classification is described as a KDD method used for prediction, Clustering and Association Rules are KDD methods used for description. Association rules are frequently used to describe data items by showing their associations with other items in the data set.

### **Association Rules**

Association Rules (AR) have been widely used as a powerful data mining method for finding strong associations between attributes in a data set. The rules are displayed as the attribute-value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is Basket Analysis introduced by Agrawal [2]. Such basket databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns. The use of AR is widely accepted in the research domain. Association Rules have been used as a useful descriptive KDD technique. We believe that appropriate pruning, organising and visualisation of the generated association rules can help domain experts to find interesting patterns with a little effort without involvement of an external data analyst. In the next section we describe and review the use of AR in KDD in more depth.

## 2.2 Association Rules

Association rules identify relationships between two or more data variables by providing information in the form of  $X \Rightarrow Y$  or “if-then” statements. The attributes in the “if” part are called *antecedents*, where attributes in the “then” part are called *consequents*. Other common names for the antecedent-consequent pair are “Left hand side-Right hand side” or “LHS-RHS” and “body-head” of the rule.

In our study the definition of association rules is not the same as the formal definition introduced by Agrawal [2]. The definition of association rules introduced in [2] is based on the supermarket “basket analysis”. Each row in the “basket analysis” data set represents a supermarket transaction, with binary values 1 - if item bought and 0 otherwise. In our study the attribute-values are not represented in the binary form. Our definition of an association rule is the following: Let  $A = \{A_1, A_2, \dots, A_n\}$  be a set of attributes. Each attribute can take a finite number of values. Given two attributes  $\{A_1, A_2\} \subset A$ , ( $A_1 \neq A_2$ ) and  $V_1$  (respectively  $V_2$ ) be a particular value for  $A_1$  (respectively  $A_2$ ). An association rule is an implication of the form  $V_1 \Rightarrow V_2$ . For example, Let  $A = \{country, sex, agegroup, lawtype\}$  be a set of attributes. Given two attributes  $\{country, sex\} \subset A$ , ( $country \neq sex$ ) and Australia (respectively Male) be a particular value for country (respectively sex). The implication of the association rule  $Australia \Rightarrow Male$  is that a presence of the country Australia in the dataset, also indicates a possibility of the presence of sex Male. This is the simplest form of a “single item” association rule. Using logical conjunction and logical disjunction operators a more complex “multi item” association rules can be constructed. For example, the implication of the “multi item” association rule  $Australia \text{ and } Male \Rightarrow Criminal$  is that a presence of the country Australia in conjunction with sex Male in the dataset also indicates a possibility of the presence of law type Criminal. In this study we do not use the logical negation operator.

The “basket analysis” definition of an association rule by Agrawal [2] is the following: Let  $I = \{i_1, i_2, \dots, i_m\}$  be a set of items. For example, the set of items may include all items found in a supermarket. Let  $D$  be a set of transactions, where each transaction  $T$  is a set of items such that  $T \subseteq I$ . For example, the transaction represents a supermarket basket of a customer for each visit to the store. We say that a transaction  $T$  contains  $X$ , a set of some items in  $I$ , if  $X \subseteq T$ . For example each basket contains items bought by a customer (e.g. Milk, Bread and Chocolate). An association rule is an implication of the form  $X \Rightarrow Y$  (e.g.  $Milk \Rightarrow Bread$ ), where  $X \subset I$ ,  $Y \subset I$ , and  $X \cap Y = \emptyset$ .

These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. Since their introduction by Agrawal [2] much research has been published that covers almost all aspects of ARs and their discovery. However the

number of discovered rules might be well over several thousand which makes it practically impossible for the user to find interesting rules. For that reason Agrawal [2] has introduced filtering features for the association rule generation. In addition to the antecedent (the “if” part) and the consequent (the “then” part), an association rule has two numbers that express the degree of uncertainty about the rule. The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule as a proportion of all transactions. (The support is sometimes expressed as a percentage of the total number of records in the database.) We calculate the support as

$$s_{X \Rightarrow Y} = \frac{n(X \cap Y)}{n(T)} \quad (2.1)$$

where T is the total number of rows (transactions) in the database.

The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent. We calculate the confidence as

$$\begin{aligned} c_{X \Rightarrow Y} &= \frac{n(X \cap Y)}{n(X)} \\ &= P(Y|X) \end{aligned}$$

For the rule Milk  $\Rightarrow$  Bread (support 12%, confidence 80%), a support value of 12% means that 12% of all customers buy Milk and Bread together, and that 80% of customers that buy Milk also buy Bread. The confidence measure is merely an estimate of the conditional probability of Bread given Milk. However, the reverse association rule Bread  $\Rightarrow$  Milk would have the same support but different confidence value. For example, consider a data set containing 1000 transactions, with 150 transactions containing Milk, 220 transactions containing Bread and 120 transactions containing Milk and Bread together. The support for either rule would be 12%, calculated as  $(120/1000)*100$ . However the confidence for the rule Bread  $\Rightarrow$  Milk would be considerably smaller (54.5%) than the confidence for the rule Milk  $\Rightarrow$  Bread (80%). Many KDD applications generate both association rules (forward and backward) in order to avoid the confusion that the users may have by not knowing where to place an attribute (e.g. 2-D Martix approach). While association rules have proven to be useful in practical applications, AR algorithms tend to generate large numbers of rules. Users have considerable difficulty manually analysing so many rules to identify the truly interesting ones. To overcome this problem various measures of interestingness have been developed to prune all association rules to a subset

of the most interesting to the user. The study of interestingness is related to the study of surprisingness or unexpectedness.

Association rule interestingness is an important, but difficult, problem. Some of the mined rules may be trivial facts such as HUSBAND  $\Rightarrow$  MALE, while some other rules may be redundant e.g TUESDAY  $\Rightarrow$  RAIN. Existing research in rule interestingness focuses on either objective or subjective measures. Objective interestingness measures are data driven and based on some statistical measures. Subjective interestingness measures are user driven and require users to specify whether a rule is interesting.

### 2.2.1 Objective Interestingness

Objective interestingness is not user driven and not domain oriented. This type of interestingness has been used from the very early days of KDD and does not involve domain knowledge. Confidence and support thresholds represent an objective measure that filters interesting rules. If both support and confidence values are greater than the user specified threshold, the association rule is considered interesting. However, many analysts do not know what an ideal threshold setting should be [59, 57, 58]. Many KDD software applications using AR have the default threshold settings for support at 10% and confidence at 50%. However, the suggested threshold settings are more appropriate for the boolean data sets such as supermarket basket analysis, than for other domains where the number of possible values for an attribute is greater than two. A weakness of this approach therefore lies in the difficulty in deciding appropriate thresholds. If the threshold is set too high, useful rules may be missed, but if it is set too low, the user may be overwhelmed by many irrelevant rules.

Another common problem with the user specified support measure is that not all high support rules are interesting. The authors in [53, 57] argue that most rules with high support are obvious and well known, and it is the rules with low support that provide interesting new insight. In many real life applications some items appear very frequently in the data, while others rarely appear. If the support is set too low those rules that involve rare items will not be found. This dilemma is called the *rare item problem*. Hip and Guntzer [39] report that the measure of support is domain dependent. The authors claim that in basket analysis, for example, a minimum support threshold is useful because it probably does not make sense to decide upon special advertisements based on items bought by a very small fraction of supermarket customers. In contrast in the medical domain the death or severe illness of a patient may also be quite rare but is of great importance. For such domains, where rare items are still of a great importance, the minimum support threshold should be avoided [39]. For the same reasons we believe that

rare data items in our domains (Medical and Law) are very important, therefore we are not using the minimum support threshold. For example in the VLA data set, most of applicants were born in Australia, which makes most of *country of birth = Australia* rules meet the high support threshold. However, the number of applicants that were born in other countries is small, therefore these applicants are seen as a rare item group. These rare groups may be very important representatives in the data set but could be overseen due to their low support value.

Figure 2.2 shows the subset of association rules that have low support. The  $X$  axis represents support values from 0 to 100%, while the  $Y$  axis represents confidence values from 0 to 100%. The gray area represented by the rectangle correspond to the group of items that are infrequent. These items have a very low support value and would not be regarded as interesting in the supermarket basket analysis. However in other domains these rare items could be of a great importance.

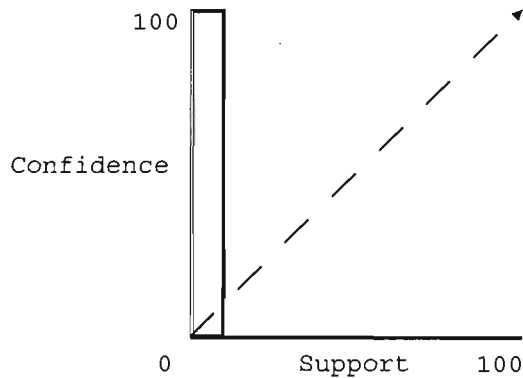


Figure 2.2: Rare Items

To overcome this problem multiple support thresholds are introduced by [57]. Items that are not so frequent in the dataset have *minimum item support* (MIS) lower than items that are more frequent.

Another approach involves mining AR without the support as discussed in [79]. They explore a confidence-based level-wise pruning without the use of support. The confidence-based pruning measures a certain monotonicity of confidence, called the universal existential upward closure, so that only confident rules of larger size need to be examined for generating confident rules of smaller size. The problem of mining confident rules is to find all confident rules for given minimum confidence. Some authors try to automatically specify support without consulting the user. Work in [53] is based on the confidence and lift which provides an automatic support specification. The authors claim that the lift is good at discovering the low support, but high confidence rules. There are many other different methods proposed for finding the most interesting rules as reviewed in [6, 38]. A



common name for this method is *rule pruning*. Shah in [75] use causality based arguments to prune a generated rule set. Prior to the rule generation attributes are categorised as cause or effect and the rules are represented in the format CAUSE  $\Rightarrow$  EFFECT. The pruning technique in [75] is based on minimizing the set of causes for a specific set of effects and maximizing the set of effects for a specific set of causes. In [58] an algorithm that prunes the discovered associations by removing insignificant rules (rules that are already known to have no value) is introduced. This algorithm finds a special subset of the unpruned associations to form a summary of the discovered associations. This subset is called the direction setting (DS). The direction of a rule is the type of correlation computed using the  $\chi^2$  (chi-squared) test. Essentially, DS rules give a summary of the behaviour of the discovered associations and represent the structure (or skeleton) of the domain. The non-DS rules simply give additional details. Using this summary, the user can focus on the essential aspects of the domain and selectively view the relevant details.

Work in [10] is also measuring interestingness of associations via the chi-squared correlation from classical statistics. This measure is upward closed in the lattice of sub-sets of the item space, enabling the authors to reduce the mining problem to the search for a border between correlated and uncorrelated itemsets in the lattice. In other words, if a set  $I$  of items is deemed dependent at significance level  $a$ , then all supersets of  $I$  are also dependent at the same significance level  $a$  and, therefore, they do not need to be examined for dependence or independence.

More recent research shows that discovery interestingness should be based on domain knowledge. Moreover the use of KDD is slowly moving towards the end user who is not necessarily a technical person. The current trend in organisations shows that the number of IT professionals diminishes gradually. For example the number of non-technical domain experts (lawyers) in VLA is a few times greater than the number of IT professionals. This makes it difficult for the lawyers to take advantage of the information stored in the company databases. It is also time consuming for the domain expert to engage an IT professional every time the expert needs to gather some information. The need for new KDD tools built for the purpose of the non-technical domain experts is emerging. In addition, a visual representation of the discovery would make the discovery process more understandable to the user. In the following sections we focus our KDD discussion on these issues.

All approaches discussed above are based on algorithms that prune rule discovery (based on some measures of interestingness) at run time (data mining step) and then show the discovered rules to the user. However, researchers apply further organization and filtering of the generated rules. According to [60] the key problem is not with the

large number of rules because if there are indeed many rules that exist in data, they should be discovered. The main problem is with the inability to organize and present the rules in such way that they can be easily analysed by the user. This organization also allows the user to view the discovered rules at different levels of detail, and to focus his/her attention on those interesting aspects. One of the approaches is to organize rules by grouping the related rules together [68, 20, 22, 60, 45]. By grouping related rules together, the common goal is to identify deviations which can serve as a basis for useful actions. In the next section we explore user-driven subjective interestingness before discussing interestingness of deviations in the following section.

### 2.2.2 Subjective Interestingness

Objective measures alone are insufficient for determining a discovered rule's interestingness [55, 49]. In contrast to objective measures, subjective measures are needed. Such measures involve domain expert opinion. By modeling and incorporating domain knowledge into the system, subjective interestingness can be determined [73]. Most existing approaches ask the user to explicitly specify what types of rules are interesting and uninteresting. The rules that match the user's specifications are retrieved. According to [55] subjective interestingness has two main measures:

- Unexpectedness  
Rules are interesting if they are unknown to the user or contradict the user's existing knowledge (or expectations).
- Actionability  
Rules are interesting if users can do something with them to their advantage.

The authors in [55] claim that the two measures are not mutually exclusive. Subjective interesting rules fall into three categories:

- rules that are both unexpected and actionable
- rules that are unexpected but not actionable, and
- rules that are actionable but expected

The unexpected and actionable rules are regarded as the most interesting to the user because the rules were previously unknown and useful. The second category of rules is not as interesting to the user because the user has no advantage in their use despite their unexpectedness. The third category of rules is not interesting because they confirm to the user's existing knowledge or expectations. The technique reported in [55] is interactive

and iterative. In each iteration the user is asked to specify his/her knowledge about the domain. The system then uses this knowledge to analyse the discovered rules according to some interestingness criteria and identify the potentially interesting rules. Finally the user inspects the analysis result, removes uninteresting rules and saves the interesting rules.

According to Liu et al. [54, 55] there are three types of domain knowledge: general impression (GI), reasonably precise concept (RPC) and precise knowledge (PK). GI and RPC types represent the user's vague feelings that there might be some associations between the attributes. For example the user might have a vague feeling that a student gender and university course taken are associated. However precise knowledge is based on past experience or previous data mining exercises. For example the user knows from census data that there is a strong association between gender and life expectancy (females live longer). In the approach by Klementinen [49] a user's domain knowledge is also used. The user is asked to classify attributes into a class hierarchy of both interesting and uninteresting rules. Templates are used for defining interesting and uninteresting classes of rules. The classes of interesting rules are stored in an inclusive template and the classes of uninteresting rules are stored in a restrictive template. To be interesting, a rule has to match an inclusive template. If a rule, however, matches a restrictive template, it is considered uninteresting.

Instead of trying to find what is interesting, Sahar [73] introduces an algorithm that eliminates a large family of rules that are not interesting from the exhaustive list output by a data mining algorithm. By asking a user to classify only a few rules, specifically chosen, their elimination can bring about the automatic elimination of many other rules. Every time a rule is presented for classification, a user is asked whether: (1) a rule is true, and (2) the user is interested in any rule in the family of the classified rule. The first category, specifying whether a rule is true, is used in the construction of the knowledge base. The rules that represent common knowledge are excluded from the list of interesting rules presented to the user (for example  $\text{HUSBAND} \Rightarrow \text{MALE}$ ). The second category is a broader one characterising the entire family of the rule. Those two categories define four possible classifications for each rule: True-Not Interesting (TNI), Not True-Not Interesting (NTNI), Not True-Interesting (NTI) and True-Interesting (TI). Only TI rules classified by the user are considered valid and interesting.

Another example of using a user's domain knowledge is discussed in our previous work [44]. In this work we use the discrepancy between a domain expert's confidence prediction and the actual confidence for a rule as a measure of subjective interestingness. Before rule generation, a user is asked to select a variable of interest to generate rules containing that variable as consequent. Before displaying confidence values for each generated rule, the

user is asked to predict a confidence value. Rules that had confidence values that surprised domain experts were considered interesting. Although the approach advanced in [44] is appealing, experts were not easily able to predict confidence values for rules that involved more than one or two variables in the antecedent.

As discussed above, subjective interestingness plays an important role in rule discovery. However we believe that any modern data mining tool should include both objective and subjective measures in its rule pruning. The inclusion of objective interestingness measures alone would pass domain expert's knowledge. This leads to a generation of many rules that are expected (previously known) or useless (non-actionable). Furthermore, subjective interestingness measures alone lead to a combinatoric explosion and the user has difficulties winnowing through millions of rules in order to find interesting ones. We conclude that the measure of what is meant to be *interesting* to the user is dependent on the user as well as the domain within which the KDD is being used as discussed in [49, 84, 4, 54, 55]. According to [14] domain experts are already familiar with the common patterns in the data through their years of the experience and their intuitive feelings. They also claim that a domain expert who is already familiar with the application domain is very unlikely to be satisfied with merely prevalent patterns because presumably the company is already exploiting them to the extent possible. A tool that can show differences or changes in the relationship between attributes in data would be of the great benefit to a domain expert. There are several researchers who applied KDD techniques to measure those changes. In the next section we closely look into the deviations as measures of interestingness.

In the modern information based society, more and more information is generated but not used. Most researchers focus on the domain expert as the key player in pre-discovery (data-preprocessing and transformation) and post-discovery (result interpretation) steps. However there is a little work done on domain experts role as the actual discoverer. As the data mining movement is paying closer attention to the end-user, there are few KDD tools available that are easy to use by a non-technical domain expert. The benefits of such tools, in addition to existing KDD tools are manifold. One might argue that there is no need for additional KDD tools. It would be almost similar if one argues that; why are we offering milk, bread and newspapers in small corner shops (or petrol stations) when we have the same items in big supermarkets? The answer is; it is convenient, less time consuming and easy to access. The access to information for domain experts, in modern organisations, should be almost as easy as getting a bottle of milk. Like milk, information can be out of date and become stale if not used in a timely fashion.

In the next section we discuss a different approach to finding interesting rules. By grouping similar rules, differences (deviations) between rules can provide additional interesting cues for the user.

### 2.2.3 Interestingness of Deviations

One of the most promising areas in KDD is the automatic analysis of changes and deviations [68]. With deviations we have a simple way to identify things that differ from our expectations. Since they differ from what we expect, they are by definition interesting. In statistics a deviation is defined as  $d_i = x_i - \bar{x}$  where  $x_i$  is the observed value and  $\bar{x}$  is the expected value (e.g. mean or median). For example, in time analysis the differences are measured over time where the observed value is the current measured value and the expected value is the previously measured value. A commonly used approaches in time analysis is called "trend". Lets consider a few real life examples; The major goal in retail sales analysis is to identify areas in which sales can be increased. In manufacturing, the goal might be to reduce production defects. In healthcare information analysis, goals might include identifying high cost areas or improving quality of care. The common goal in all the above is to identify deviations which can serve as a basis for useful actions. Several systems have been developed for these tasks. Interestingness measures based on deviations were used by [68, 14, 20, 5, 29, 83, 22, 44].

There are two different approaches in the use of deviations to measure interestingness of a discovery: measuring differences over time and measuring differences between groups. The approach by Piatetsky-Shapiro and Matheus [68] is based on the timely analysis and argue that the timely analysis of key patterns that arise in these databases is highly desirable and may often provide competitive advantage. Moreover, the authors in [14] claim that if the analyst is already familiar with the patterns in data, the greatest incremental benefit is likely to be from changes in the relationship between item frequencies over time. In [68] the authors developed a system called Key Findings Reporter (KEFIR) that models the analytic process employed by the expert data analyst. KEFIR is applied to the healthcare database and the central type of interesting pattern is a deviation between an observed value of a measure and reference value e.g. a previous or a normative value. The observed value is taken from the most common snapshot of the database. Comparing the observed value to one from previous time generates a deviation over time. Such a set of deviations is called findings. In addition to uncovering the significant findings, the analyst needs to explain them to the extent possible given the data. The expert analyst performs the further drill down in top-down fashion in order to find an explanation for the findings.

The findings and their possible explanations are then compiled into a text report.

Another use of identifying relationship changes over time is discussed in [14]. The authors explore the problem of boolean customer basket analysis introduced by [2], and claim that analysis of variations of inter-item correlations along time can approximate the role of domain knowledge in the search for interesting patterns. According to [14] a set of  $k$  items is declared as interesting not because its support and confidence exceed a user defined threshold, but because the relationship between the items change overtime. This approach could be used to explore a supermarket customer's behaviour.

Aumann and Lindell [5] use the term *behaviour* to describe changes in a relationship between categorical data attributes. Behaviour of a subset is considered interesting if its distribution stands out from the rest of the population. The subset of the population displaying a distribution significantly different from that of its complement, either in terms of the mean or the variance, is recognised as interesting and noteworthy. The general structure of a rule in [5] is *population-subset*  $\Rightarrow$  *interesting behaviour*. An example of a rule that finds an interesting behaviour is *FEMALE*  $\Rightarrow$  *Wage (mean = \$7.90 p/hr) (overall mean wage = \$9.02 p/hr)*. The authors claim that this rule shows interesting behaviour because it reveals a group of people (females) earning a significantly lower than average wage (\$7.90 p/hr). The authors mentioned above use deviation as an interestingness measure in order to identify relationship changes between attributes in data.

A slightly different approach described in [29] uses deviations to find changes between two different datasets in terms of the models they induce. The authors give a motivating example that: a sales analyst who is monitoring a dataset (e.g. weekly sales from K-mart) may want to analyse the data thoroughly only if the current snapshot differs significantly from previously analysed snapshots. If successive database snapshots overlap considerably, they are quite similar to each other. The calculations of deviations in [29] are discussed as follows: Each frequent itemset  $X$  in the model represents a region in the attribute space (where support is higher than the user specified threshold) whose measure is the support of  $X$ . The set of all itemsets (e.g. a,b,ab) is the *structural component* and the set of their supports (e.g. 05, 04, 025) is the *measure component*. If the structural components of two models are identical (e.g.  $a_1, b_1, ab_1$  and  $a_2, b_2, ab_2$ ), then the deviations are computed for the measure components. However if the structural components are different, the structural components are reduced to their greatest common refinement (union of the sets of frequent itemsets of both models) in order to be identical. The deviation between the datasets is the deviation between them over the set of all regions in the greatest common refinement.

## 2.2.4 Interestingness of Group Differences

Not all algorithms use Association Rules to calculate deviations in order to find interesting patterns in data. The authors in [20, 22] use so called *contrast sets* in order to find differences between groups (subsets). According to Bay and Pazzani [22] a common question in exploratory research is: *How do several contrasting groups differ?* Learning about group differences is a central problem in many domains. These groups can represent different classes of objects, such as male or female students, or the same group over time, e.g. students in 2000 through 2003.

In [20, 22] Bay and Pazzani introduce STUCCO (Search and Testing for Understandable Consistent Contrasts), an algorithm which finds conjunctions of attributes and values that differ meaningfully in their distribution across groups. This means that STUCCO finds rules that have different levels of support for different groups e.g. X is the contrast set for two groups A and B if  $P(A | X) \neq P(B | X)$ . STUCCO finds those contrast sets (cset) where:  $\exists ij P(cset = True | G_i) \neq P(cset = True | G_j)$ .

For example, comparing female and male students across several courses might find that more male than female students are enrolled in Information Technology courses  $P(course = IT | sex = male) = 73\%$  while  $P(course = IT | sex = female) = 27\%$ .

STUCCO searches the space of all possible contrast sets and returns only contrast sets that meet the criteria. The criteria is based on two measures; user specified threshold called the *minimum support difference* and statistical significance criterion. The minimum support difference is calculated as the difference between maximum and minimum supports between the items in the contrast set. The *support* in [20, 22] is equivalent to *confidence* in the AR. The authors use the  $\chi^2$  (chi-square) test as the statistical significance criterion as shown in the equation 2.2. Contrast sets that met the criteria are called *deviations*. Finally, the interpretation of each deviation is displayed as English text to the user e.g. *Rule 1. Students who are male are more likely to enrol in the IT course than students who are female.*

The chi-square test was also used by Liu et al. [59, 54, 58] and supported by [6] as a significance criterion for association rule discovery. The chi-square is a widely used method for testing hypotheses about the independence (or alternatively association) of frequency counts in various categories [12]. The sampling distribution of a proportion is usually a  $\chi^2$  distribution. The frequencies are traditionally displayed in contingency tables. A contingency table over  $r$  is basically a table of counts, in which each count denotes how often a given combination of attribute values occurs in  $r$  [40]. For example, consider female and

male legal aid applicants and the law type of their offence. This relationship can be shown in a contingency table as shown in Figure 2.1.

	Criminal	Family	Civil	row total
female	100	300	50	450
male	400	100	50	550
column total	500	400	100	
grand total				1000

Table 2.1: Calculating statistical significance by using contingency table

In order to determine if the differences in proportions between male and female applicants represent a true relation between the variables, we use the chi square to test for independence  $\chi^2$ :

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2.2)$$

where  $O_{ij}$  is the observed frequency count in cell  $ij$ , and  $E_{ij}$  is the expected frequency count in cell  $ij$ . First we calculate the expected value for each cell as  $\frac{n_{i.}n_{.j}}{n}$  where  $n$  is the total number of observations (grand total) and then calculate  $\chi^2$  as

$$\chi^2 = \frac{(100 - 225)^2}{225} + \frac{(300 - 180)^2}{180} + \frac{(50 - 45)^2}{45} \quad (2.3)$$

$$+ \frac{(400 - 275)^2}{275} + \frac{(100 - 220)^2}{220} + \frac{(50 - 55)^2}{55} = 272.55 \quad (2.4)$$

With 2 degrees of freedom and  $\alpha = 0.05$ , we calculate that  $\chi^2 = 272.55$  and reject the null hypothesis and determine that differences in proportions for males and females are significant.

In this study the chi-square test of significance is an optional measure of interestingness. In contrast to work in [20, 22], where only rules with significant statistical difference are defined as interesting, we use the chi-square test as an optional feature and the user decides if the differences or similarities between groups are interesting. The reason for this approach is twofold; we want to allow the user to drive the process of discovery and decide what is interesting, and statistically justified differences are not always interesting.

There are many similarities between WebAssociate and STUCCO; e.g. both applications use group difference as the main interestingness measure and both applications do not use AR support for pruning. We also adopt the approach by Bay and Pazzani



[20, 22] and use the chi-square to test statistical significance because we are working with proportions. However it is important to say that our work uses different techniques and tools to find differences between groups in order to mimic the exploration methodology that domain experts follow. The main differences between WebAssociate and STUCCO are:

1. In our work we use the brute force Apriori algorithm where [20, 22] do not use any AR techniques. The authors in [20, 22] claim that trying to directly apply AR mining algorithms to find contrast sets is a poor idea. We investigate this claim in Chapter 5.6.
2. WebAssociate is an interactive tool that displays the findings visually where [20, 22] displays the findings in textual format. In the next section we closely explore different visualisation techniques in KDD and present evidence that users prefer a visual presentation.
3. WebAssociate displays findings as similar, different or both where [20, 22] is focusing on the contrast sets which include only rules that are statistically different. We believe that similar findings are as interesting as different ones because they might contradict user's beliefs, despite not reaching statistical significance.
4. WebAssociate suggests why the selected groups may be different.
5. WebAssociate has the ability to further define groups rather than just those in the data set. For example: *country of birth = (Australia, New Zealand, UK)* could be grouped as an *English\_group* and *country of birth = (Italy, Greece, Yugoslavia, etc)* could be grouped as a *Southern\_European* group. VLA experts find this functionality useful. For example to test hypotheses such as *More Southern\_European born applicants get refused legal aid than English\_group applicants*.

In Chapter 5.6 the similarities and differences between WebAssociate and STUCCO are explained in more detail.

According to [21] mining algorithms for finding category or group differences can be classified as characteristic or discriminative. Characteristic miners, such as association rules, attempt to find significant differences in the class descriptions. This can result in rules that are highly predictive as with discriminative mining, but predictiveness is not a requirement of the mined rules. Characteristic miners may also contain information that is not useful for prediction, but nevertheless may be important to an analyst attempting to understand the two groups.

Discriminative miners, such as classification, attempt to find differences that are useful for predictive classification with a high degree of accuracy. However, Bay and Pazzani [21, 22] report some disadvantages with classifiers that use a discriminative approach:

1. Rule learners and decision trees may miss alternative ways of distinguishing one group from another.
2. Rule learners and decision trees focus on discrimination ability and will miss group differences that are not good discriminators but are still important. For example Italian born applicants in the VLA data represent only 0.5% of the dataset but are an important group for VLA experts. This problem is called the *rare item problem* and was discussed in [57].
3. It is difficult to specify useful criteria such as minimum support or an acceptable false positive rate in the classification framework.
4. Rules are usually interpreted in a fixed order where a rule is only applicable if all previous rules were not satisfied. This makes interpretation of individual rules difficult since they are meant to be interpreted in context.

According to Bay and Pazzani [21] there have been many studies which investigate the accuracy of rules that describe differences between groups but very few which investigate how humans interpret results. The authors report that current classification mining algorithms can produce rules which differentiate the groups with high accuracy, but often human domain experts find these results neither insightful nor useful, therefore characteristic miners such as association rules are more useful to domain experts.

As discussed before, with deviations we have a simple way to identify things that differ from our expectations. By identifying things that differ, we can use the deviation measures to test hypotheses. According to Stranieri et al. [77] data analysis with the use of association rules is particularly suited to the generation of hypothesis that may be useful as the subject of further inquiring in order to explain the hypothesis. The authors apply association rules to the legal domain (family law in Australia) in order to demonstrate that AR are invaluable tool for legal analysis.

Many domain experts use hypotheses informally when trying to explain a hunch. This means that domain experts use hypothesis testing in a less standard way that statisticians and mathematicians do. For example, VLA could be interested in finding suggestions for hypotheses involving any single variable or any combination of variables in applications from applicants born in different countries (e.g. Vietnam, Greece, Australia and Italy) or from different cultural groups (e.g. Asians and English). Diabetes experts could be

interested in finding suggestions for hypotheses involving different general practice divisions (e.g. Central division, Wimmera division and Loddon division) in order to find some interesting differences between the diabetic patients from each division. They might find that more diabetes type 3 patients are coming from the Loddon division than any other divisions and decide to further test this suggested hypothesis. The link between the hypothesis and rules is therefore just as important when the association rules are used to suggest a hypothesis, as they are when the rules are used to confirm a hypothesis. The focus of our work is on AR as an invaluable tool for hypothesis suggestion or testing. In the next section we explore the link between AR and hypotheses.

## 2.3 Hypotheses and AR

Association rules show relationships between variables in data and as such have been used to suggest or confirm hypotheses [5, 77]. Stranieri et al. [77] used Association rules in the legal domain to suggest hypotheses generated from previous divorce cases. Aumann and Lindell [5] used association rules to find interesting behaviours in the data set that may be used to suggest hypotheses. A behaviour of the subset is “interesting” if its distribution stands out from the rest of the population. For example, the association rule “*non-smoker*  $\Rightarrow$  *life expectancy = 85 (overall 78)*” identifies interesting behaviour because the life expectancy of non-smokers is higher than the overall life expectancy. This association rule could suggest the hypothesis “*Non smokers are expected to live longer*”. The word *hypothesis* is generally used in a more restricted sense in research to refer to conjectures that can be used to explain observations. A hypothesis is a hunch, an educated guess which is advanced for the purpose of being tested. According to [12] hypotheses are the working instrument of theory. But regardless of its source a hypothesis must:

- be stated so that it is capable of being either confirmed or rejected
- be stated clearly, in correct terminology and operationally
- state relationships between variables

### 2.3.1 Hypothesis Suggestion

In contrast to hypothesis testing, hypothesis suggestion is a preliminary step that uses differences between groups to propose interesting suggestions to the user. As discussed earlier, the essence of KDD is to automatically find patterns in data [27, 2]. Association rules are widely used to discover patterns that are associations between attributes in data [2, 10, 76, 6, 79, 53]. Discovered patterns are further pruned according to some measure

of interestingness. However, many AR algorithms present the discovered rules without any connection between them. The only connection between the discovered rules is that their confidence and support is above the user specified threshold. This approach is not very useful because the connection between discovered association rules is based on their frequency and not based on their content (antecedent or consequent). Furthermore a number of studies reported that analysts typically have difficulty in interpreting rules [40, 86, 56]. In this study, by connecting association rules that have the same consequent, we group rules together into rule sets (*called iso-consequent*) and visually display confidences for each rule set. This approach automatically suggests hypotheses to the user because the rules in the rule set are connected according to their content. For example, a rule set containing association rules  $lawType\_CRIMINAL \Rightarrow refused\_YES$  (conf. 8.4%) and  $lawType\_FAMILY \Rightarrow refused\_YES$  (conf. 15.4%) suggests to the user a hypothesis “*The family law type applications are more refused than the criminal law type applications*”. Struck with an interesting rule set the user is able to further explore and test the suggested hypothesis that s/he finds interesting.

### 2.3.2 Hypothesis Testing

While researchers and statisticians use hypotheses to explore and explain observations (e.g. by rejecting or accepting null hypothesis), domain experts use hypotheses informally when trying to explain a hunch. For example, a legal aid organization interested in gender differences in criminal applications for aid may frame a hypothesis as follows: *There is no gender difference in applicants that apply for legal aid for criminal matters*. Association rules that would test this null hypothesis ( $H_0$ ) are:  $Criminal\ Matter \Rightarrow Female$  (confidence  $X$ ) and  $Criminal\ Matter \Rightarrow Male$  (confidence  $Y$ ). The deviation between confidences would be used to reject or confirm the null hypothesis  $X = Y$ . It takes at least two association rules in order to test this hypothesis.

For example, Aumann and Lindell [5] use AR to identify changes in a relationship between categorical data attributes. The authors test the hypothesis that the mean of the two subsets are not equal (the null hypothesis) against the hypothesis claiming a difference of means e.g.  $sex = FEMALE \Rightarrow Wage$  (mean = \$7.90 p/hr) (overall mean wage = \$9.02 p/hr). Standard statistical methods such as the Z-test are used to establish significance of the inequality of the means.

### 2.3.3 The link between the Hypothesis and AR

The mapping of an hypothesis to the set of association rules that will confirm or deny the hypothesis is rarely simple. The link is not easy to formulate correctly and the mapping

between an hypothesis and an association rule is not one to one. For example in order to test the hypothesis “*More female than male applicants apply for family law type matters*”, we need to formulate two association rules;  $family \Rightarrow female$  and  $family \Rightarrow male$ . However, many experts would have difficulties formatting these rules correctly in order to link the association rules to the hypothesis (e.g. map family as consequent instead of antecedent). Non-technical domain experts need tools that will automatically suggest or test hypotheses. By suggesting hypotheses to the user, we enable him to visually scan through the suggestions and find a hypothesis that he or she wouldn’t think of. This automatic process is the first discovery step that non-technical users find useful. During the evaluation study of our work, experts indicated that several visually suggested hypotheses were interesting because the experts did not know that such hypotheses existed.

WebAssociate applies a visualization mechanism based on the (*iso\_consequent*) rule sets in a way that helps non-technical domain experts to identify rules that are appropriate to test a given hypothesis or to explore suggested hypotheses. In the next section we review different visualisation techniques that aim to present a discovery to the user in more understandable manner.

## 2.4 Discovery Visualisation

Our approach to data mining aims at integrating the human into the data mining process and applying its abilities to the large data sets available in today’s computer systems. For this purpose, techniques which provide a good overview of the data and use the possibilities of visual representation for displaying the findings are especially important. In the process of hypotheses generation, the user is guided by the visual feedback of the process and quickly learns more about the properties of the data in the database.

### 2.4.1 Visualisation Categories

Visualisation is the process of transforming data, information and knowledge into a visual form, making use of a human’s natural visual capabilities [63]. According to Grinstein and Zhang [31, 86] there are three kinds of visualisation categories in KDD. The first category presents the findings obtained from the data mining step, the second category visualises the data before applying data mining algorithms, and the last category uses visualisation to complement the data mining techniques. In this section we focus on the first visualisation category, which aims to visually present findings to the user. Furthermore we focus on visualisation techniques that display findings generated by association rule based algorithms.

Data visualization techniques can be classified into five categories: geometric techniques, icon-based techniques, pixel-oriented techniques, hierarchical techniques, and graph-based techniques as discussed by Keim and Kriegel [19]. However, according to Zhang [86] because of the nature of AR, the graph-based technique is most suitable. Zhang explores many different graph-based visualisation techniques and claims that in the field of AR visualisation, most research focuses on table, directed graphs, two dimensional (2-D) matrices and three dimensional (3-D) visualisations.

## 2.5 Association Rule Visualisation

The most straightforward method for AR visualisation is to use a table representation of rules as shown in Table 2.2. This method demonstrates early approaches in visualising AR when computers were not advanced enough for rich graphic representations. Table 2.2 shows rule representation for hypothetical student data.

LHS	ImPLY symbol	RHS	Confidence	Support
male	$\Rightarrow$	IT	73%	17%
female	$\Rightarrow$	IT	27%	13%
male	$\Rightarrow$	First year, IT	34%	8%

Table 2.2: Example of AR in rule table format

In a real life example, several hundred to several thousand rules could be generated which makes the table AR representation unsuitable. The analyst would have to spend a long time searching for interesting rules even after interestingness based algorithms had pruned the rules. In order to overcome the table rule problem, researchers introduced new graphs that present the findings in a more meaningful way. For example a directed graph is introduced and used in IBM's (Intelligent Business Machines) data mining software "Intelligent Miner". Directed graphs are used to show AR such that nodes represent attribute-values and arcs represent the relationship. Support levels are represented by different colors and confidence levels are represented by length of arcs, longer arcs represent higher confidence. In Figure 2.3 we use directed graphs to represent three different association rules;  $Italy \Rightarrow Male$ ,  $Italy \Rightarrow Female$  and  $Male \Rightarrow Overseas student + It$ . A directed graph is a good visualisation technique when the number of rules is small [86], however we find directed graphs unsuitable for the representation of deviations. For example, the deviation between confidences for rules  $Italy \Rightarrow Male$  [confidence 70%] and  $Italy \Rightarrow Female$  [confidence 70%] could be difficult to distinguish because the length of the arc for each confidence does not clearly identify their differences. Furthermore, additional

rules would make a directed graph cluttered with many additional arcs and nodes.

More advanced visualisation techniques use a 2-D matrix for representing AR. This technique is used by SGI (Silicon Graphics International) in their data mining software *MineSet*. The association rules are represented as a 2-D matrix where user selected attribute-values are displayed on both axes. One axis is labeled as the left-hand side and the other as the right-hand side. The grid intersection between two axis is displayed as the height of a bar and represents the confidence of the rule corresponding to the left-hand side(LHS) and right-hand side (RHS) labels. The support value is represented as a disk attached to the bar and expected probability as the color of a bar. All of these three representations are configurable. Figure 2.4 shows AR presentation for the rules *LawType\_Criminal*  $\Rightarrow$  *Male* and *Male*  $\Rightarrow$  *LawType\_Criminal*

The limitation of this approach is that MineSet is able to visually display only rules that have single left-hand and right-hand side. Although MineSet tried to overcome this problem by allowing rules that have multiple LHS and RHS, as shown in Figure 2.5, the limitations are obvious when the number of items in the LHS or RHS is greater than a couple. Figure 2.5 shows AR presentation for the rules *LawType\_Criminal*  $\Rightarrow$  *Male and Age\_21..25* and *Male*  $\Rightarrow$  *LawType\_Criminal* Despite visualisation limitations by MineSet, the ability of MineSet to adjust the visual display by resizing, rotating and flipping is advanced. However we find this item-to-item 2-D matrix approach unsuitable because it does not clearly identify deviations.

The limitation of 2-D matrix graphs that show AR mapped as item-to-item, resulted in 3-D visualisation techniques advanced by Wong et al. [18]. Those authors represent a rule-to-item approach that shows rows as items and columns as AR rules. The LHS and RHS of a rule are distinguished by two different color represented bars. Confidence and support values are displayed by the height of the bar placed at the end of the matrix. In Figure 2.6 Rule 3 shows the 3-D AR presentation for the rule *Milk*  $\Rightarrow$  *Bread AND Butter* (confidence 84%)

According to [86] the approach advanced by [18] is best for the AR that have multiple RHS and single LHS. That is, the rule body has only one item. When LHS has more than one item, the matrix floor is covered with many blocks. We don't find the approach in [18] suitable for the representation of deviations because it does not clearly identify deviations between groups of rules with the same consequent.

There are many other researchers who addressed the problem of visualising association rules. The authors in [36] visualise the entire process of KDD, while others, as advanced in [40], focus on the post discovery process.

Different visual data mining techniques are available for different stages of the data

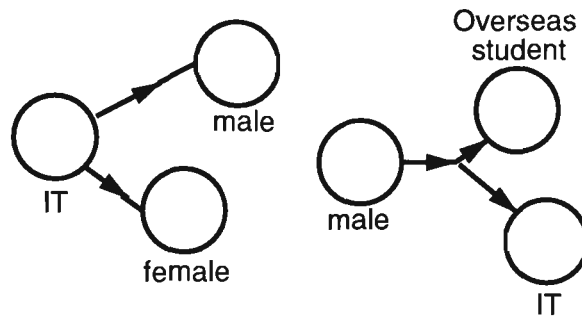


Figure 2.3: Association Rules represented by IBM *Intelligent Miner*

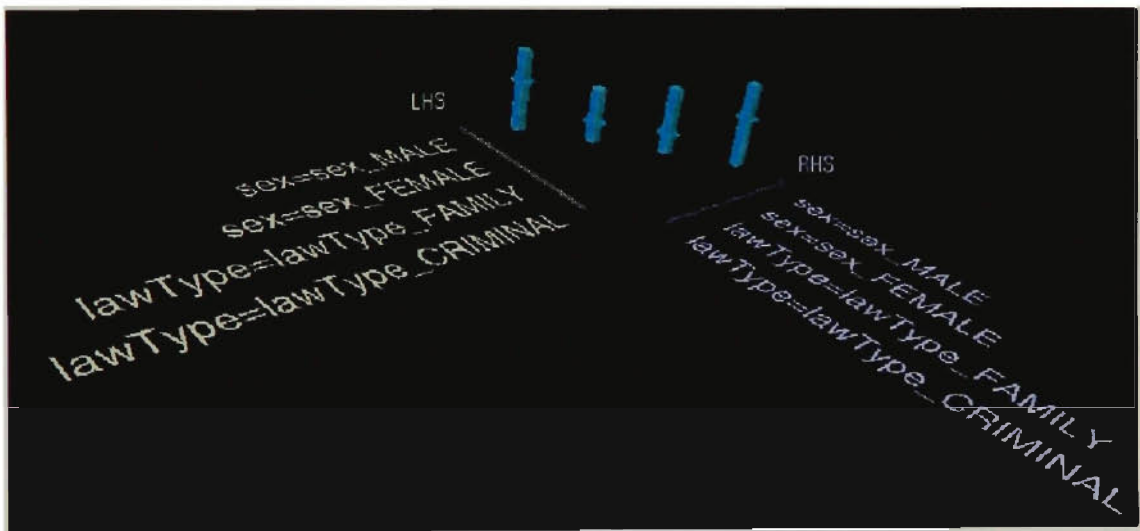


Figure 2.4: Single item Association Rules represented by SGI *MineSet*

mining process. As Hofmann et al. [40] have found AR are not always easy to understand particularly if rules are complex. This led those researchers to develop visual helpers. Visual helpers called "Mosaic plots" aim to support analysts in understanding complex rules. These are visual representations of association rule values represented as contingency tables. As the authors in [40] discuss, a contingency table over  $r$  is a table of counts (frequencies), in which each count denotes how often a given combination of attribute values that occur in  $r$ . The contingency table has a cell for each combination of attribute values of the participating attributes. The simplest case is a two-way table in which two attributes are set against each other is shown in Table 2.3.

The authors claim that "Mosaic Plots" give the analyst a deeper understanding of the nature of the correlation between the left-hand side and right-hand side of the rule. An example of the graphic representation of the contingency table shown in Table 2.3 is illustrated in Figure 2.7.



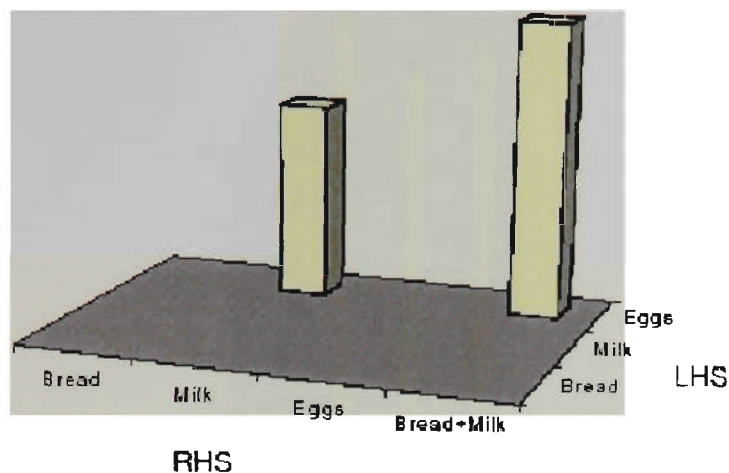


Figure 2.5: Multiple item Association Rules represented by SGI *MineSet*

	sex=male	sex=female
IT=yes	200	100
IT=no	80	120

Table 2.3: Contingency table for student data with two attributes

According to Hofmann et al. [40], providing the context of the rule in such plots, allows users to better assess its quality. Furthermore, these plots give a deeper insight into how results can be used in solving business problems.

A quite different AR visualisation technique is advanced in [82]. The authors apply visualisation techniques to display associations between related web sites returned by a query. The query is sent to a web search engine (e.g. Google and Altavista) by the user in order to search the web for a specific keyword. Visual interfaces are created to let the user explore relationships between related documents returned by the query. An example given by Lin et al. [82] demonstrates a visual interaction between the user and the search engine. Imagine, for example, that a user is searching the web for the keyword “back pain”. The search query displays results visually as shown in the Figure 2.8. A user studies the map and finds that several topics such as “occupational diseases”, “spinal diseases” and “Physical Therapy” are associated with the search keyword. The visual representation of the query results prompt the user to rethink his or her query. The user decides to add “occupational diseases” to the search box. This time the query returns fewer results which are visualised again and presented to the user. The process is iterative until the user is satisfied with the results. Figure 2.8 shows initial visual results returned by the “back pain” query.

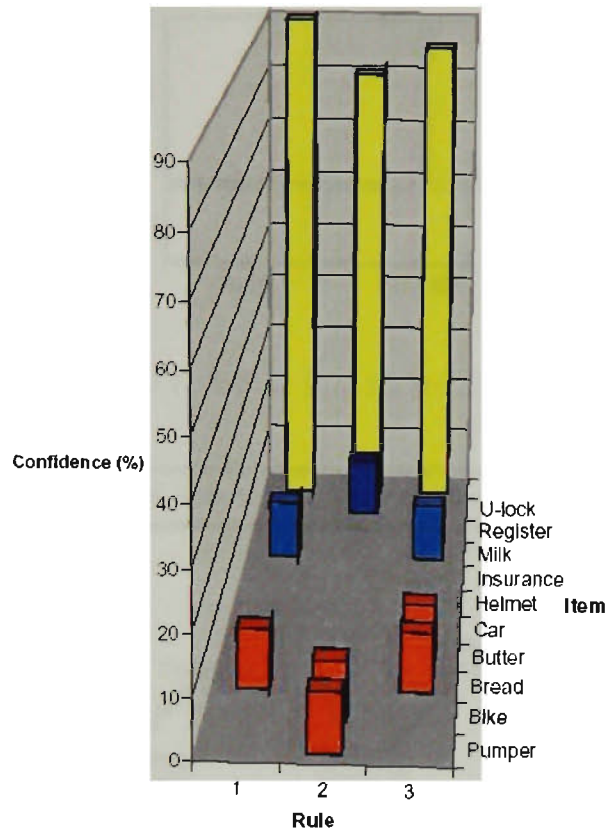


Figure 2.6: 3-D AR visualisation

Visualisation of query results returned to the user demonstrates that users prefer a visual display over a text based display. Advances in [82] model the search reasoning deployed by the user. This approach shows that users use iterative selection process in order to discover information, as Figure 2.8 shows.

Although the approach in [82] is full of promise, its limitation is that it shows relationships between items (documents) but does not represent the strength of a relationship. In association rule mining, the strength of a relationship is traditionally represented by measuring the confidence of a rule. In our approach AR confidences and their deviations are the most important part of the process, therefore we find this approach unsuitable for our use.

The visualization methods advanced in [40, 82] have been shown to support analysts in understanding associations between items. However, it is unlikely that these approaches would identify deviations and facilitate a clearer link between a hypothesis and the rules that would test the hypothesis. Both approaches visually represent rules so that their meaning can be appreciated at a glance, but the task of identifying rules that would suggest or confirm a hypothesis requires knowledge beyond the rules. For example: struck

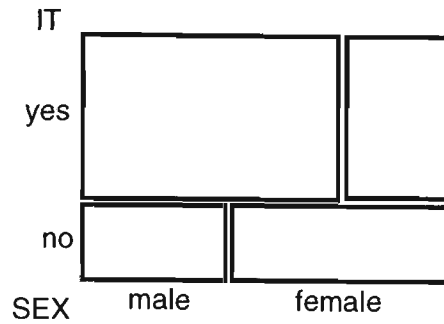


Figure 2.7: Mosaic Plot for student data

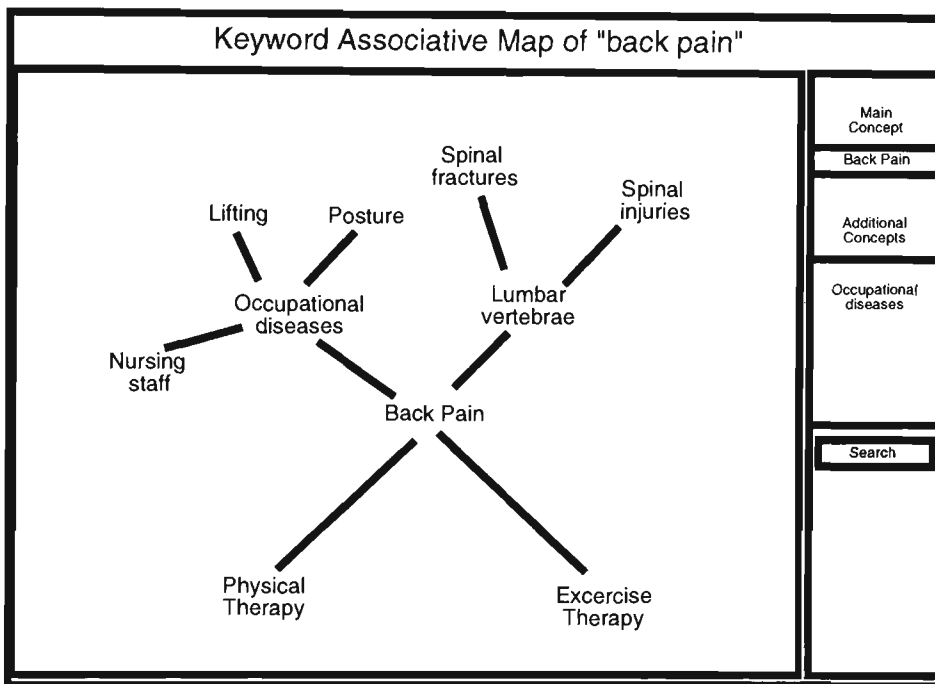


Figure 2.8: Associative Map

with an interesting visually presented rule, an analyst seeks to invent a hypothesis that will explain the rule. For example, if the confidence of the rule *Italian and Elderly*  $\Rightarrow$  *Aid Rejected* is surprisingly high then he or she may infer the hypothesis: *Elderly Italians are wealthy* and seek additional association rules to confirm this. This is explored in greater depth in Chapter 5.

## 2.6 Chapter summary

KDD provides a powerful knowledge discovery tool in the hands of a domain expert. When KDD deploys AR, particular attention is focused on different measures of interestingness. Many interestingness measures in knowledge discovery are based on domain

knowledge in order to classify a discovery interesting. A discovery has to have at least one of the attributes; *unexpected, actionable, surprising, previously unknown, useful and novel*. Interestingness of deviations provides new ways for rules to be grouped and differences between groups to be identified. Deviations offer opportunities for hypotheses testing and can be used to suggest hypotheses. However, many authors claim that AR are sometimes hard to understand, especially for non-technical domain experts. The need for a new set of KDD tools for hypotheses suggestion and testing emerges. Also a visual representation of the discovery plays an important role in the KDD cycle. In the next chapter we closely look at the role of the domain expert and explore the relationship between domain experts and data miners.

## Chapter 3

# Domain Experts

According to the Cambridge and Oxford dictionaries, the term "expert" is defined as:

- A person with a high degree of skill in or knowledge of a certain subject.
- A person with a high level of knowledge or skill; a specialist:
- A person who is well informed or skillful in a subject

For small domains, one person can be a domain expert. In larger, more complex domains, a specialist may take over the task of detailing particular partitions of the domain that he/she is knowledgeable of. In manufacturing, this may result in a departmental partitioning; sheet metal process specialist, a rubber works specialist or electronics specialist. In healthcare, departmental partitioning could include diabetes, cardiology, and radiology experts. In law, experts might be defined as lawyers with technical knowledge in family, criminal or civil law.

There is an increasing number of researchers, [55, 54, 59, 49, 84, 4, 44], who claim that the availability of actively strong domain knowledge improves the efficiency of the knowledge discovery process by reducing the search space and helping to focus on the interesting findings. Depending on their experience, domain experts have a variety of tasks in any KDD exercise. However, in the essence of KDD, domain experts are often end users who will apply discovered knowledge to their use. In order to make the use of KDD more efficient and to meet the requirements of domain experts as end users, we have to understand their characteristics. In the next section we closely look at the characteristics of domain experts.

### 3.1 The Role of the Domain Expert

Currently several successful KDD tools have been reported in many areas of science, business and government. KDD however is an iterative process which is dependent on human interaction. The involvement of the domain expert, as discussed in [50, 59, 54, 73], is important in all KDD steps. This human interaction, for the domain expert, is traditionally twofold; selection of domain data and subsequent interpretation of results. Data selection which involves deciding which attributes should be included for Knowledge Discovery is clearly best achieved by someone who is an expert in that field. The interpretation of results which involves the identification of what is *interesting* is a step which arguably can only be carried out by an expert in that field.

Broadly speaking, there are two types of domain expert; those that practice and understand Data Mining (for example, data analysts and database administrators) and those who are non-technical, such as lawyers and healthcare professionals. The former are able to perform each of the KDD steps; data selection, data transformation, data mining and interpretation of results. The latter however are “pure domain experts” and require additional interaction with data miners. Figure 3.1 shows different types of domain experts and data miners.

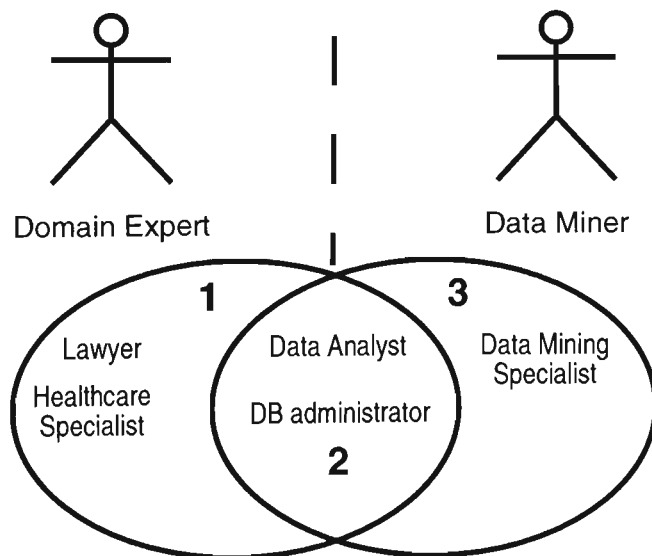


Figure 3.1: Domain Experts and Data Miners

Figure 3.1 illustrates that:

1. Some domain experts are data miners
2. Some domain experts are not data miners

### 3. Some data miners are not domain experts

Due to the complexity of current KDD tools, the vast majority of domain experts are not data miners; they have to collaborate with data miners in order to achieve knowledge discovery. This collaboration can be initiated by either the domain expert or the data miner. In the absence of domain experts, some data miners may seek expert knowledge from other sources such as data dictionaries, business rules, regulations and manuals as discussed in [81, 26]. Figure 3.2 shows the possible collaboration between domain experts and data miners.

The dashed arrow in Figure 3.2 shows that the data miner seeking expert knowledge is

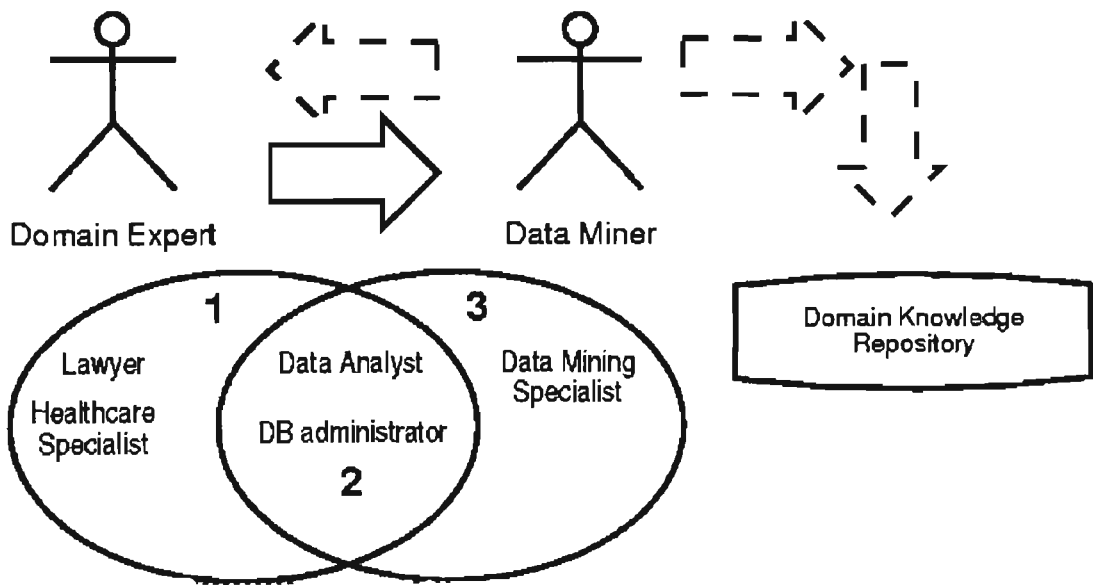


Figure 3.2: Relationships and Dependencies Between Domain Experts and Data Miner

not necessarily dependent on the domain expert. The full line arrow in Figure 3.2 shows that domain experts seeking KDD are dependent on the data miner. As we see, “pure” domain experts seeking KDD have less choice available than do data miners seeking KDD. Moreover, the process of collaboration which a domain expert must enter into with a data miner can be both onerous and error prone, especially when performed iteratively. Meetings have to be scheduled at mutually convenient times, contracts have to be drawn, mistakes can occur due to miscommunications or misunderstandings.

From the point of view of the “pure” domain expert, how can this situation be improved? At first glance, the relationship between “pure” domain experts and data miners is an equal one; the domain expert provides the knowledge whilst the data miner provides the tools and the know-how. However, a closer examination suggests an imbalance. Subsequent sets of data provided to a data miner offer no additional challenges; he already

has the know-how and the tools and just has to process the data. On the other hand even if the tools (software) were given to a “pure” domain expert, that person would still lack the data mining know-how. The complexity of many underlying algorithms result in complex KDD tools requiring specialist knowledge to operate, thus making it unrealistic for the expert data mining know-how to be taught easily. That said, there are a great many day to day “pure” domain expert KDD requirements which require only a fraction of the sophistication of the current KDD tools.

Association rules can provide an intuitive and easy to use set of KDD algorithms which would cater for many everyday tasks at the hands of the “pure” domain experts. The role of new KDD tools is not to replace the use of data miner driven “heavy-duty” tools but to provide an additional set of “pure” domain knowledge driven tools that will meet his/her simpler day to day requirements. Figure 3.3 illustrates that additional data mining resources available to “pure” domain experts lead to a more balanced inter-dependency and relationship.

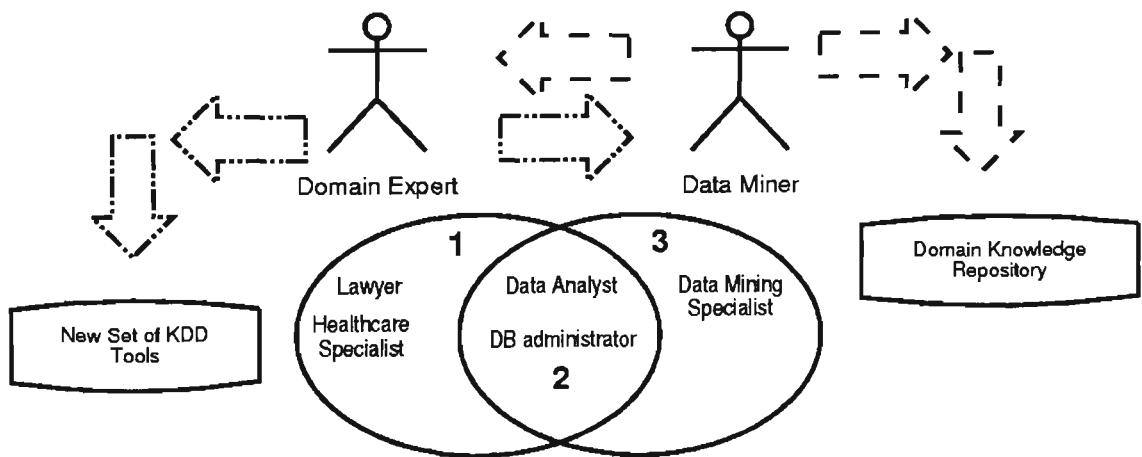


Figure 3.3: Balanced Relationships and Dependencies Between Domain Experts and Data Miner

The relationship between “pure” domain expert and data miners is now optionally dependent on the domain experts requirements. In the next section we explore the requirements of “pure” domain experts.

### 3.2 KDD Requirements of the Domain Expert

In their extensive research, authors in [32] investigate 43 KDD software products which are either research prototypes or commercially available tools. The authors look closely



at the level of support that each software provides for both analysis expert or novice users (“pure” domain experts). According to the study in [32] domain experts such as lawyers, healthcare professionals, engineers and managers require simple-to-use tools that efficiently solve their business problems. Moreover, domain experts expect application to behave similarly to their business practices. This point is stated in the Ergonomics Standard ISO 9241

### **Consistency, Conformity to User Expectations - ISO 9241**

*Conformity to user expectations demands that an application behaves as users expect it to do. This principle goes beyond mere consistency, because it is not restricted to the computer systems but also connects the application with the real world. Note that user expectations can vary largely, depending on the background and learning history of your prospective users. Computer literate users will expect that your application conforms to well-established interface standards, while beginners - who are domain experts - will expect that your application will behave similarly to their business practices.*

Obviously different experts have different demands but most of the available tools are aimed at analysis experts, requiring prohibitive levels of training before being useful to domain experts as novice end users. Most domain experts are usually not interested in using advanced powerful technology per se, but only in getting clear, rapid answers to their everyday business questions.

Many organisations embrace KDD in order to make use of their data set. The evolution is usually carried out through three distinct implementation phases. However, some investigations [32] report that most organisations are at the early stage, implementing phase one or phase two. This means that KDD is still not widely used through all organisational levels. The authors in [32] investigated the use of almost all mainstream commercial KDD products and reported that deploying KDD technology in an organisation is traditionally implemented through the following phases:

1. First, KDD studies are performed by the data mining specialists (external consultants)
2. Once the profitability of KDD is proven, data analysis experts apply the KDD techniques (possibly with the help of a domain expert who has strong domain knowledge)
3. Full exploitation of KDD technology within the organisation. Domain experts are enabled to perform their own KDD analysis according to their individual needs. Although widely still a vision, the necessity for this stage is clearly recognised.

In phase 1, as illustrated in Figure 3.4, an organisation deploys KDD by contracting an external KDD specialist to analyse their data set in order to discover useful patterns. The role of a domain expert in this phase is to provide the specialist with domain knowledge of the organisation. The dashed line in Figure 3.4 illustrates that an external specialist may use a domain knowledge from either knowledge repository or a domain expert.

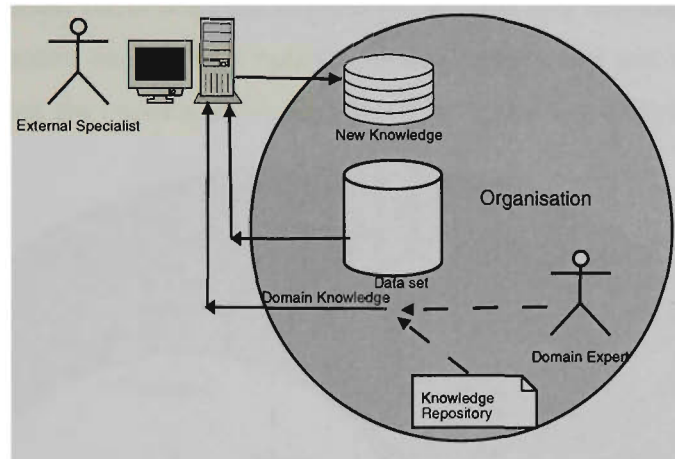


Figure 3.4: KDD implementation - Phase 1

Subsequently in phase 2, as illustrated in Figure 3.5, an organisation deploys KDD by purchasing hardware and software needed for the data analysis, and deploying inhouse data analysts to analyse their data in order to discover useful patterns. In this phase domain knowledge may be sourced from a data analyst or a “pure” domain expert.

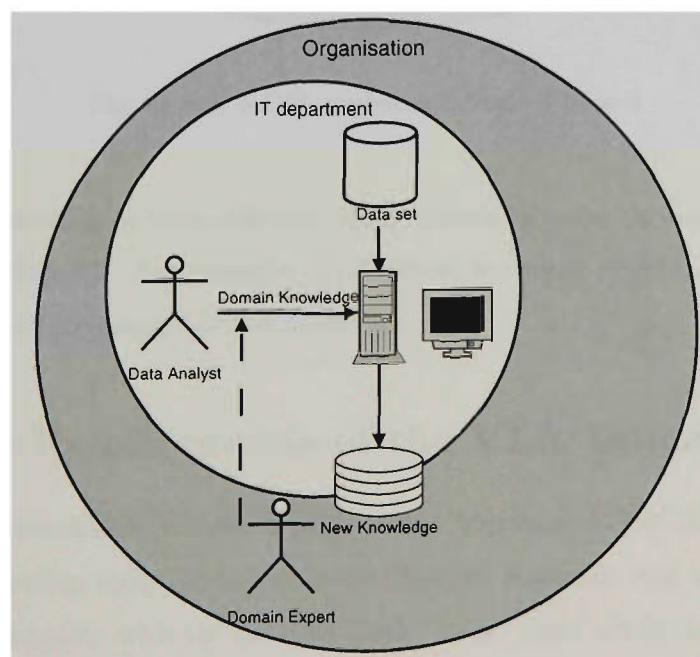


Figure 3.5: KDD implementation - Phase 2

In the final phase, as illustrated in Figure 3.6, an organisation deploys KDD by enabling “pure” domain experts to perform their own KDD analysis. In this phase, each domain expert is using additional KDD tools. The dashed line in Figure 3.6 illustrates that the IT department is still responsible to provide data sets according to the requirements of the “pure” domain experts. The lines pointing to new knowledge in Figure 3.6, illustrate that by using additional KDD tools the experts are making new knowledge contributions. In this study we identify the needs of “pure” domain experts and test implementation of this stage by allowing the experts to perform KDD with the use of WebAssociate.

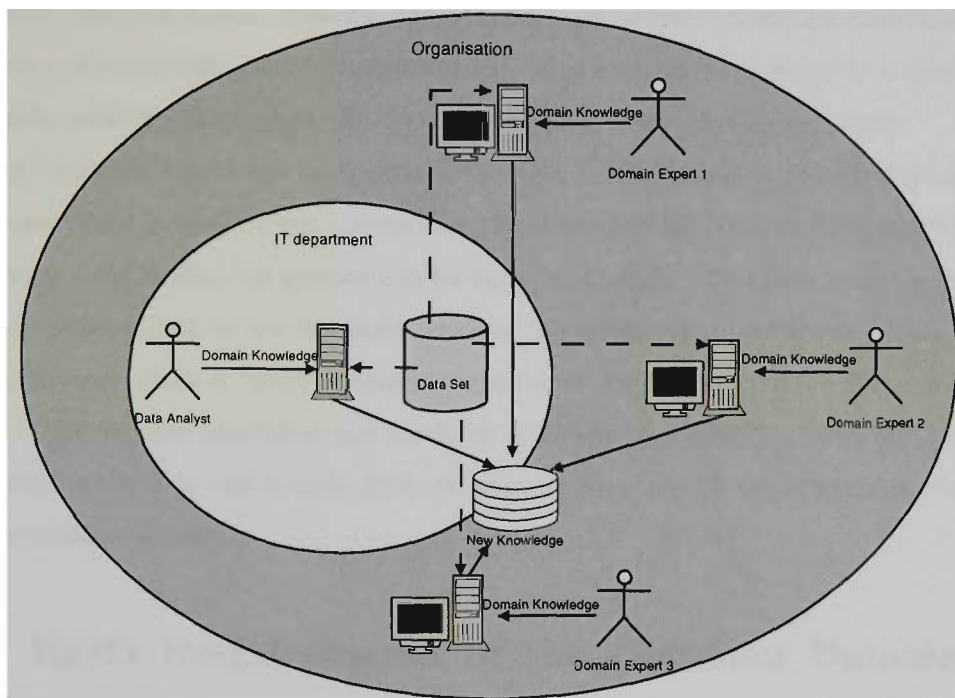


Figure 3.6: KDD implementation - Phase 3

Different organisations have different KDD phases in order to analyse their data as discussed in Section 3.2. For example, VLA offices are using KDD in phase two, where Diabetes Australia are using KDD in phase one.

### 3.3 KDD Requirements of the VLA Domain Experts

The Victorian government created a body called Victorian Legal Aid (VLA) with the objective of providing legal aid in the most effective, economic and efficient manner to those in the community with the greatest need. To get legal aid in the form of funds to cover legal costs or cost free lawyers, a client provides his/her personal, financial and case details. All these details are stored in a VLA data set.

In order to measure to what extent their objectives have been meet, VLA has to anal-

yse their data set. Currently the organisation has over 381,000 records in the database and the traditional method of turning data into knowledge relies on manual analysis and interpretation. Currently VLA creates annual or quarterly reports generated from their data set. The manual probing of a data set is slow, expensive and impractical. The generation of such reports takes time, involves many domain experts and data analysts. Moreover, VLA non-technical domain experts are fully dependent on the data analysts. When a VLA “pure” domain expert (e.g. area manager or lawyer), wants to explore their data set, the expert has to approach the IT department (even for the simplest requirements) and specify his/her needs. The IT department runs queries (using Bi structured query language system developed by *Hummingbird*), then exports the results to a spreadsheet, and finally provides the report. There are few difficulties with this approach.

A “pure” domain expert has to approach the data analyst experts, specify his query and wait (sometimes for more than a week) for the query results. Due to the busy VLA data base server daily traffic, the queries can be only run at night when data traffic is low. Very often the process has to be repeated because the query was insufficient. This problem occurs because often a “pure” domain expert does not know what he/she wants in the report. The experts identified the need for a simple-to-use KDD “desktop” tools that would enable them to run simple data analysis, and use the IT department personnel for more complicated tasks.

### **3.4 KDD Requirements of the Diabetes Domain Experts**

Diabetes Australia is part of a federation of twelve organisations; medical, education and scientific, research and community based. Its key purposes are to facilitate the achievement of the key strategies of Diabetes Australia, coordinate those activities which are best managed at a national level, and advocate for the person with diabetes where a national outcome is sought. The announcement of an integrated national diabetes program through Divisions of General Practice (DGP) was made in May 2001 as part of the Federal Budget. The aim of the program is to assist, through incentives to GPs and Divisions, early detection, diagnosis and effective management of diabetes in the community.

DGP analyse their data sets every two years by contracting a third party data analyst (an external consultant). This approach is equivalent to KDD stage one as discussed in 3.2. The problems with this approach is that possibly interesting findings are “stale” because data is not analysed often enough. The “pure” domain experts (medical professionals)

need to understand more about the diabetic disease, by finding something special about a particular patient population in their domain. By identifying a possible interesting pattern in a particular patient population, the expert can take appropriate action. For example if a certain patient population has an increase in a specific diabetes type, the expert may decide to provide appropriate educational and informational aids in order to reduce the number of diabetic patients in that population.

Data analysis of diabetic patients was advanced in [41] where the authors apply KDD to the National Singapore Diabetes data set. The authors developed a KDD tool that integrates classification with association rule mining in order to predict whether a new patient is likely to have a diabetic related eye disease. The tool also provides a visual (bar-graph) representation of association rules in order to give the doctors a better understanding of their data set. Although the approach advanced in [41] is appealing as “pure” domain experts desktop KDD application, it is different from WebAssociate because it uses confidence and support ( introduced in [2] ) as the measure of interestingness. As we previously discussed, the confidence and support alone are not effective measure of interestingness, especially for non-technical experts. Many experts do not know what an ideal confidence-support threshold setting should be [59, 57, 58]. Another problem with the tool developed in [41] is that, the tool was developed only for the purposes of the Singapore Diabetes Professionals. The most important aspect in [41] is that the authors identify the need for additional set of KDD tools that will assist non-technical domain experts in their search for new knowledge.

### 3.5 Chapter Summary

In this chapter we examined the KDD needs of domain experts, and concluded that there is an increasing need for simple-to-use KDD tools. We identified that there are two main roles in all KDD steps; a domain expert role and a data miner role. Each role has two types:

1. The first type of domain experts are “pure” domain experts (lawyers, managers, engineers, medical professionals) who have the sufficient domain knowledge but do not have technical knowledge needed for the use of current KDD tools. The second type of domain experts are domain experts (DB administrators, data analysts, statisticians) who have the technical knowledge (as primary) as well as the domain knowledge.
2. The first type of data miners are data miners identified as technical domain experts

and the second type of data miners are data miners who are not domain experts.

We also concluded that an organisation deploying KDD technology, evolves through three distinct implementation phases. The initial phase involves the organisational use of KDD through an external KDD specialist (external consultant). The second phase involves the organisational use of KDD through an internal KDD specialist (IT or statistics department personnel) and the last phase involves the organisational use of KDD through the individual needs of domain experts.

We identified that most KDD commercial and research tools are being built for the use of analysis experts, as advanced in [32], and require a prohibitive amount of training before being useful.

Finally we conclude that, in order to use the full potential of KDD through all organisational levels, additional KDD tools are needed to cater for the individual needs of non-technical domain experts.

WebAssociate aims to meet this type of “pure” domain experts everyday requirements. The focus of this study is to create a simple-to-use KDD tool that uses association rules for better proposition of hypotheses. Using association rules as a data mining method for knowledge discovery, discovered rules would be used to:

- Propose better hypotheses to the user
- Assist non-technical domain experts in the analysis of data
- Confirm already known hypotheses

WebAssociate is a generic tool, developed for the use of non-technical domain experts independently of their data set and has been used on two different data sets (VLA and Diabetes). WebAssociate supports the natural discovery and explanation steps taken by a domain expert. By using Association rules, WebAssociate provides a set of tools that enable domain experts to easily explore their datasets. This exploration is mainly based on the visually represented hypotheses suggestions. When a domain expert identifies an interesting hypothesis suggestion, WebAssociate allows him/her to further explore and explain the cause of the hypothesis or test the hypothesis. It is in the nature of the domain expert’s everyday job to explore the data set and try to explain a well educated guess or a hunch.

In the next chapter we provide samples of consultations with VLA and Diabetes domain experts. In Chapter 5 we explore the internals and algorithms of the WebAssociate tool.

## Chapter 4

# Sample Consultations

In this study we used two *real life* data sets; A data set of demographic data from Victorian Legal Aid and a data set of Diabetes attributes from Diabetes Australia.

The Victorian government created a body called Victorian Legal Aid (VLA) with the objective of providing legal aid in the most effective, economic and efficient manner to those in the community with the greatest need. To get legal aid in the format of legal costs or cost free lawyers, a client provides his/her personal, financial and case details. In order to measure to what extent their objectives have been met VLA experts have to analyse their data set. Furthermore, the experts need to monitor their financial resources (e.g. where was the money spent) as well as human resources (lawyer's assignments). All these details are stored in a VLA data. The VLA use cases shown in this chapter, are real examples showing some everyday tasks of the VLA domain experts.

Diabetes in Australia is the fastest growing chronic disease. It is the seventh highest cause of death in Australia. The Australian indigenous population suffers the fourth highest rate of Type 2 diabetes in the world. An average of 55,000 people are diagnosed every year. Diabetes Australia is part of a federation of twelve organisations; medical, education and scientific, research and community based. Its key purposes are to facilitate the achievement of the key strategies of Diabetes Australia, coordinate those activities which are best managed at a national level, and advocate for the person with diabetes where a national outcome is sought. The announcement of an integrated national diabetes program through Divisions of General Practice (DGP) was made in May 2001 as part of the Federal Budget. The aim of the program is to assist, through incentives to GPs and Divisions, early detection, diagnosis and effective management of diabetes in the community.

Different domain experts have various reasons for exploring their data sets. While some experts have a hunch or hypothesis and use a KDD tool in order to test it, others want to explore their data set and search for interesting patterns in data. We identified three distinct approaches that domain experts use for knowledge discovery;

1. A domain expert does not know what associations between the variables in data he/she wants to discover. We call this approach *unkown hypothesis suggestion* and define it as a set of association rules;  $?_{1..n} \Rightarrow ?_{1..n}$  as discussed in the example in Section 4.1.1.
2. A domain expert knows what is his/her variable(s) of interest (antecedent) but does not know what associations he/she wants to discover with it. We call this approach *partial hypothesis suggestion* and define it as a set of association rules;  $X_{1..n} \Rightarrow ?_{1..n}$  as discussed in the example in Section 4.3.
3. A domain expert knows what associations between the variables in data he/she wants to discover. We call this approach "*hypothesis testing*" and relate it a set of association rules;  $X_{1..n} \Rightarrow Y_{1..n}$  as discussed in the examples in Section 4.2 and Section 4.1.3.

Figure 4.1 illustrates a screen of WebAssociate. Each button in Figure 4.1 corresponds to approach 1, 2 or 3. Depending on his/her task and knowledge, a domain expert may use either of the three approaches to start a discovery process as illustrated in Figure 4.1. Furthermore, WebAssociate allows a domain expert to change his/her approach within the discovery process. The aim of the consultation examples used in this chapter is to demonstrate exercercise of all three approaches. In any approach, the user is able to select a database and a table to his/her interest. Figure 4.2 illustrates a database selection and Figure 4.3 illustrates a table selection.



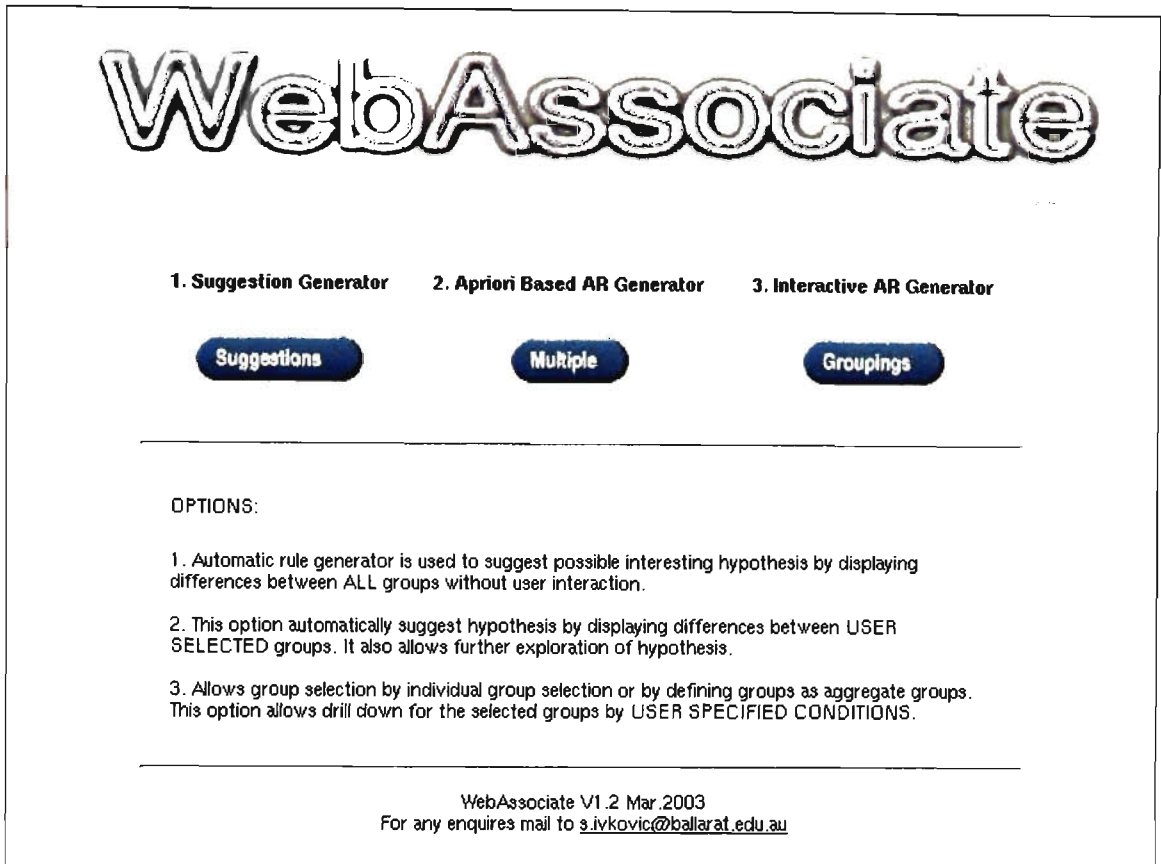


Figure 4.1: Three different knowledge discovery approaches



Figure 4.2: Database Selection

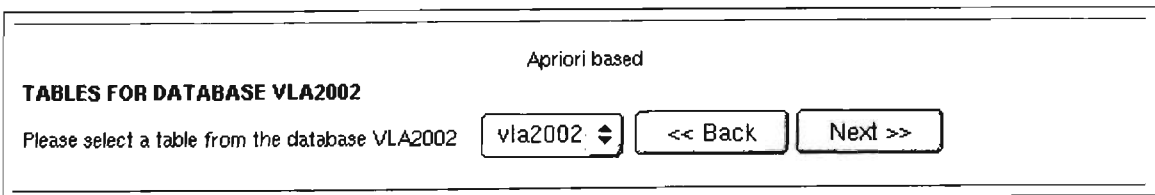
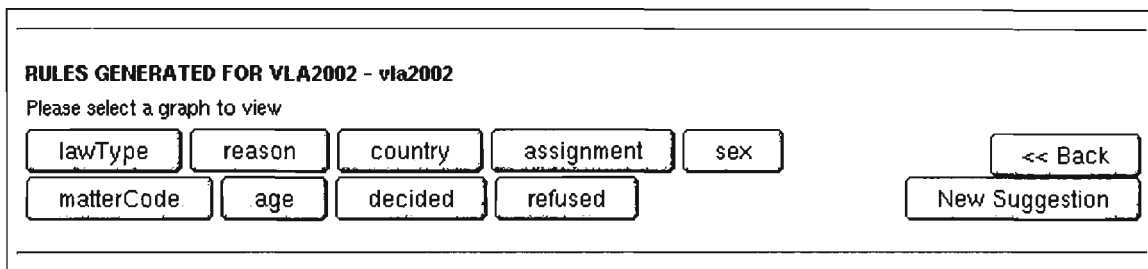


Figure 4.3: Table Selection

## 4.1 Consultation 1

Australia is a multi-ethnic country. VLA receives many legal aid applications from Australian residents born overseas. In order to provide equitable access to services for all applicants, VLA monitors the characteristics of applications from a number of national groups. This task requires deep knowledge of the legal aid process and cultural differences amongst groups and is ideally suited to a senior policy analyst. However, the task cannot be performed by non-technical experts due to the complexity of current VLA data analysis tools. The aim of consultation 1 was to test if WebAssociate enables the senior policy analyst, a non-technical expert and lawyer to monitor the characteristics of applications for equity across all national groups.

### 4.1.1 Hypotheses suggestion



**RULES GENERATED FOR VLA2002 - vla2002**  
Please select a graph to view

lawType   reason   country   assignment   sex

matterCode   age   decided   refused

<< Back

New Suggestion

Figure 4.4: VLA dataset attribute selection

The initial step, taken by the VLA expert in consultation 1, was to use WebAssociate for automatic generation of association rules from the VLA data set in order to explore suggested hypothesis. The expert did not know what associations between the variables in data he wanted to discover. He just wanted to explore suggested hypotheses in order to possibly discover some hypotheses that were useful and previously unknown. After successful automatic generation, WebAssociate provided the expert with a screen that contained nine buttons corresponding to each attribute in data as illustrated in Figure 4.4. In this consultation the VLA domain expert selected “country” in order to focus on hypothesis suggestions related to different national groups. WebAssociate generated the graph illustrated in Figure 4.5. The graph enabled him to explore suggested hypotheses and discover several interesting and possibly useful hypotheses. The expert discovered several interesting hypotheses in Figure 4.5 :

- *More Vietnam born applicants apply for criminal matters than any other national groups (confidence 83%)*

Shown as *lawType\_CRIMINAL iso\_consequent* rule set on the extreme left.

- *Vietnam born applicants apply for drug related matters more than any other national groups (confidence 30%)*

Shown as *matterCode\_RD* iso\_consequent rule set in the middle.

- *Italian born applicants get refused legal aid more than any other national groups (confidence 26.7%)*

Shown as *refused\_YES* iso\_consequent rule set on the extreme right.

The expert was especially interested in hypothesis 3 which identified a high rejection rate for the Italian group. The user identified high rejection rate for the Italians by examining the extreme right iso\_consequent rule set *refused\_YES* illustrated in Figure 4.6. This rule set is a graphical representation of 12 association rules grouped together. The color of each triangle corresponds to the color of each “country” shown in the Legend. In the “refused\_YES” iso\_consequent rule set the antecedent of each rule is a “country” where the common consequent is “refused\_YES”. The highest triangle in this rule set shows that the Italian born applicants have been refused more than any other national groups (26.7%). This is a graph representation of the association rule *country\_ITALY*  $\Rightarrow$  *refused\_YES* (confidence 27%). By visualising grouped association rules we enabled the domain expert to quickly identify possible interesting and useful patterns in data. The height of each triangle represents a confidence value for the corresponding rule. However, we do not use support as a measure of interestingness. Because the Australian born applicants represent the majority in the VLA dataset, other country groups are “rare items” because they represent the minority. VLA considers these “rare” country groups of the same importance as the Australian born applicants. Without the use of support “rare items” are included in the rule discovery.

#### 4.1.2 Discovery interestingness

The VLA expert wanted to compare the Italian population with the benchmark population represented by the Australian born applicants in order to explain why more Italians are refused. After discovering these variables of interest, the VLA user selected a *partial hypothesis suggestion* approach. The next step involved selection of the variables of interest (*country\_Australia* and *country\_Italy* as illustrated in Figure 4.7

After the selection of the variables of interest, the VLA domain expert explored the differences between these two groups, and discovered that there was a deviation in the refusal rate between the Australian and Italian born applicants, as illustrated in Figure 4.8. Each color coded line represents the variables of interest. The height of each triangle represents the conditional probability of each yaxis label given a color coded line. The

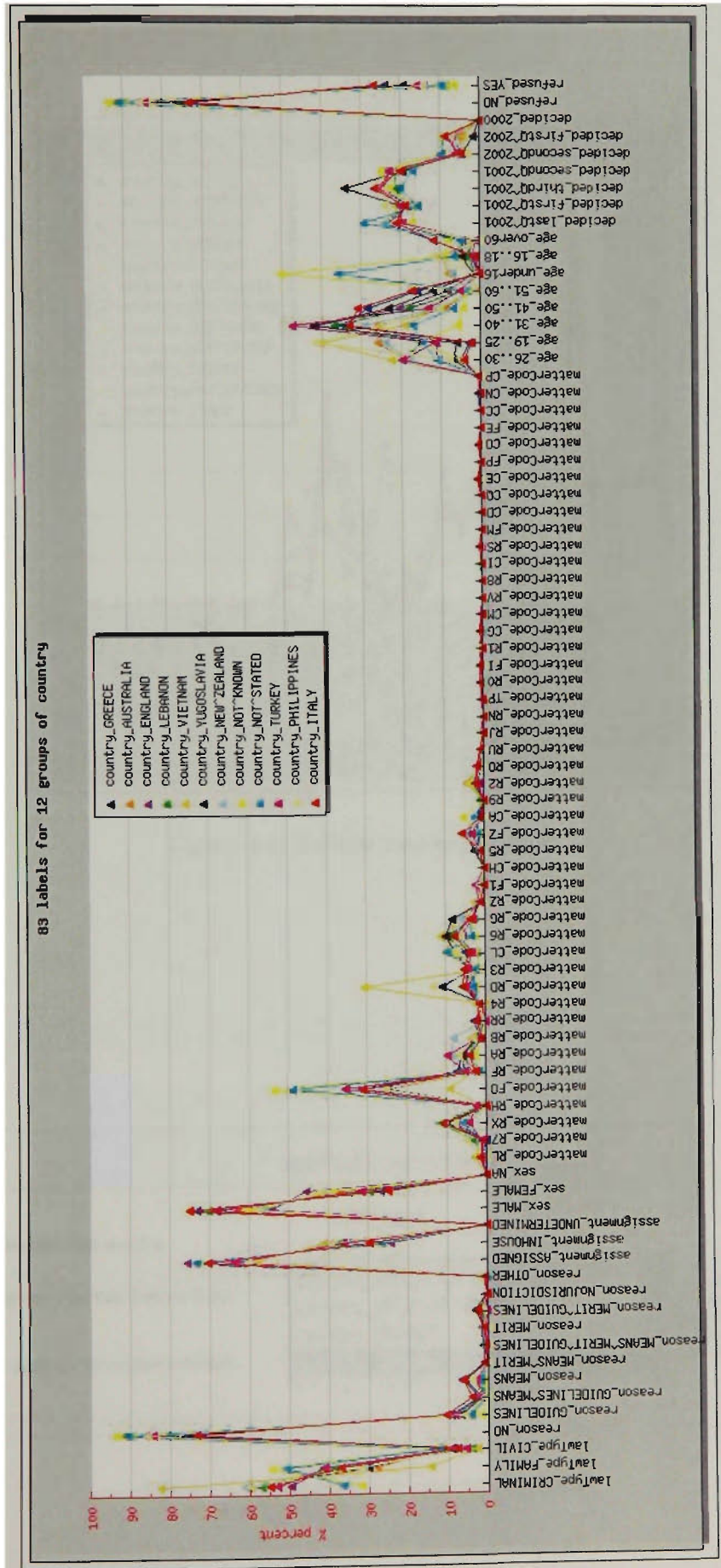


Figure 4.5: Exploring country attribute

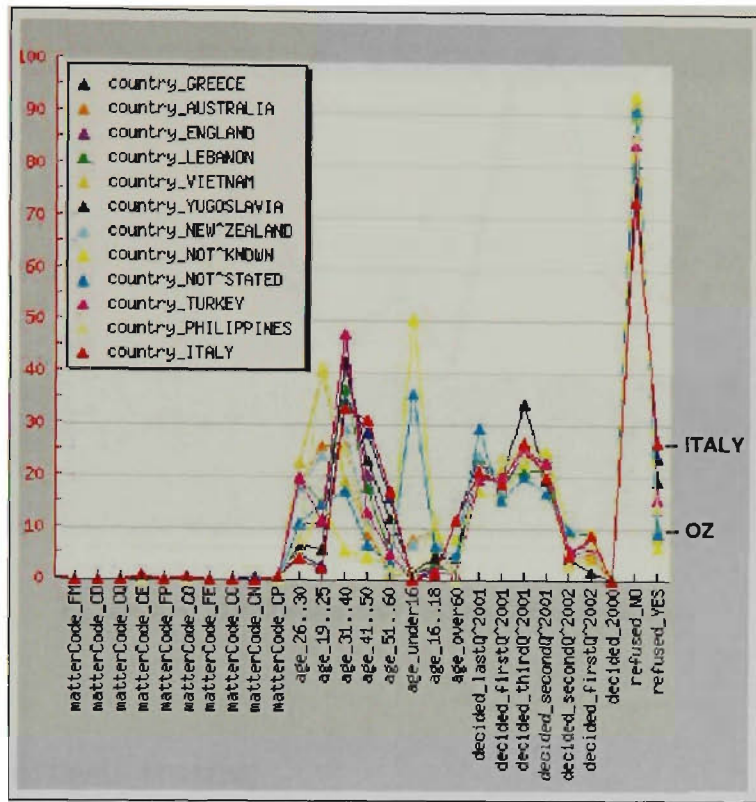


Figure 4.6: Refusal rate exploration

[back](#) | [home](#) | [new search](#)

---

Apriori based

**DISTINCT VALUES FOR country**

Please select group(s) of your interest from **country**

\*\*\* use **ctrl** or **shift** key for multiple selection

- country\_IRELAND^REP
- country\_ISLE^OF^MAN
- country\_ISRAEL
- country\_ITALY
- country\_JAPAN

<< Back

Next >>

Figure 4.7: Variables of interest selection



yaxis label number 14 (*refused\_YES*) shows that 10% of Australian applicants have been refused, compared to 26% of Italian applicants.

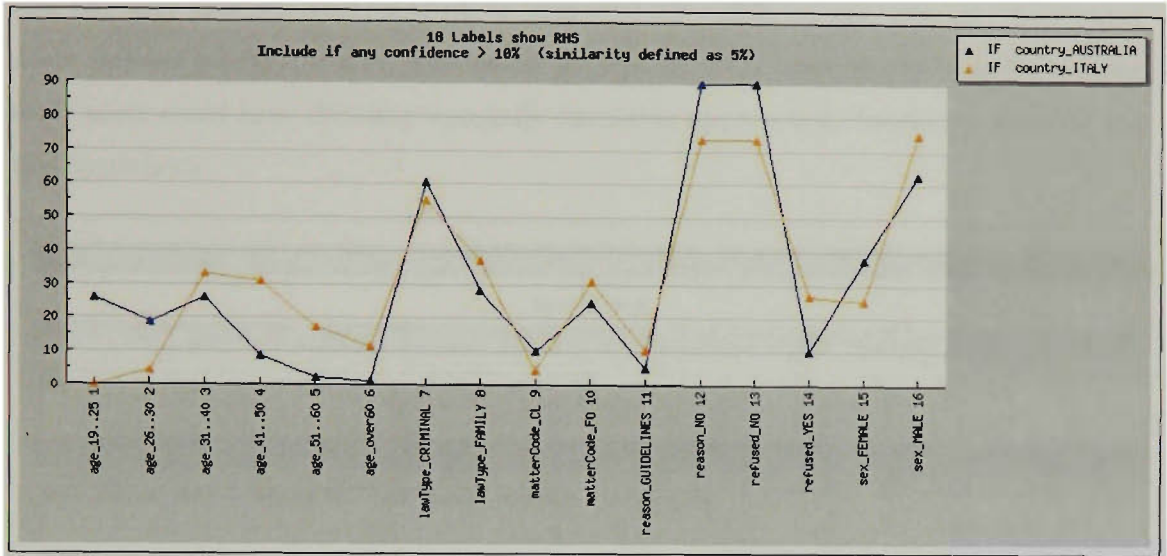


Figure 4.8: Australian and Italian VLA applicants

### 4.1.3 Hypothesis testing

The user decided to test the null hypothesis *"There is no difference in the refusal rate between the Australian and Italian born applicants"* and selected the statistical difference test between the two groups as illustrated in Figure 4.9, for the label number 14 which corresponds to *refused\_YES* in Figure 4.8.

Figure 4.9: Mapping Association Rules to the hypothesis

Prior to the statistical chi-square test of difference, WebAssociate informed the VLA expert of the generated null and alternate hypotheses as illustrated in Figure 4.10. The text entries in Figure 4.10 were automatically generated by WebAssociate. WebAssociate

automatically mapped association rules to hypothesis. As illustrated in Figure 4.10 WebAssociate automatically mapped two association rules  $country\_ITALY \Rightarrow refused\_YES$  (26.7%) and  $country\_AUSTRALIA \Rightarrow refused\_YES$  (10%) to the alternate hypothesis *More Italians get refused than Australians get refused*. This functionality is useful because some users could have difficulty manually formating appropriate association rules to test this hypothesis.

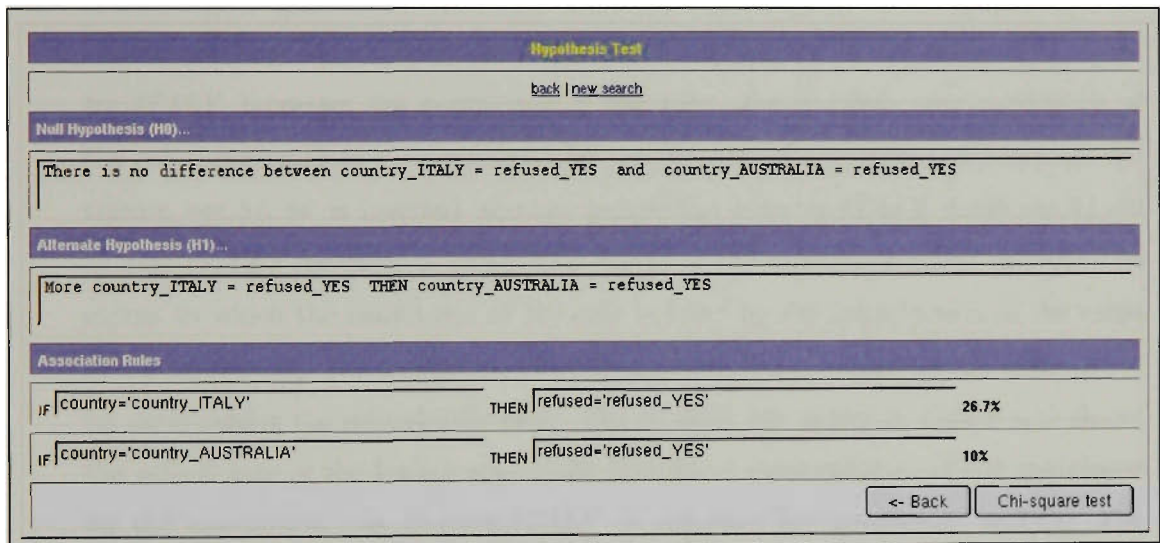


Figure 4.10: Null and alernative hypothesis for the refused Australian and Italian VLA applicants

The expert inspected the suggested hypothesis and selected "Chi-square test" in order to statistically test the null hypothesis. WebAssociate generates a contingency table containing frequencies and calculates the chi-square value as illustrated in Figure 4.11. The null hypothesis "*There is no difference in the refusal rate between the Australian and Italian born aplicants*" was rejected at the 0.05 level of significance because the chi-square value of 58.09 exceeded 3.841 with *degree of freedom 1* as illustrated in Figure 4.11.

After the chi-square test of significance and rejecton of the null hypothesis, the user decided to investigate the alternate hypothesis *Italian born applicants have been refused legal aid more than Australian born applicants*, and seek an explanation for the refusal rate differences between the selected groups. WebAssociate provided the user with possible suggestions for the hypothesis by providing the user with additional information about the selected groups.

#### 4.1.4 Suggested Explanations for the Alternate Hypothesis

By exploring the groups, the VLA expert was looking for group differences that could have contributed to the higher refusal rate of the Italians. This process involved three steps:

- **IDENTIFY LIFT** In this step WebAssociate automatically generated a graph with additional attribute-values for the Italian group which could contribute to the higher refusal rate (higher than 26.7%), as illustrated in Figure 4.12. Initially WebAssociate searches for an attribute that, when conjoined with the antecedent *country\_ITALY*, increases the confidence of the rule. For example, the confidence of the rule *country\_ITALY*  $\Rightarrow$  *refused\_YES* (confidence 26.7%) increases when the attribute *age\_51..60* is inserted into the antecedent *country\_ITALY AND age\_51..60*  $\Rightarrow$  *refused\_YES* (confidence 45%). The extent of the increase is called the *lift*. The extent to which the confidence of the rule is *lifted* by the introduction of the same new attribute into the antecedent conveys a sense of how important the attribute is for determining the refusal difference. The straight line across in Figure 4.12 shows the refusal rate of the Italian applicants (graphical representation of the confidence for the association rule *country\_ITALY*  $\Rightarrow$  *refused\_YES* (confidence 26.7%). The *lifted* attribute-values (X axis labels) are attribute-values with the height of their corresponding triangle above the straight line. These attribute-values are possible contributors for the greater refusal rate. For example, the triangle for label *lawType\_CIVIL* on the extreme left in Figure 4.12 shows that over 47% of Italians who applied for CIVIL matters were refused. This triangle is a graphical representation of the association rule *country\_ITALY AND lawType\_CIVIL*  $\Rightarrow$  *refused\_YES* (confidence 47%). However, the height of the bar for the label *lawType\_CIVIL* shows that only a small number (3.7%) of the Italian applicants who applied for CIVIL matters have been refused. This bar is a graphical representation of the association rule *country\_ITALY*  $\Rightarrow$  *lawType\_CIVIL AND refused\_YES* (confidence 47%. For that reason, even that the height of the triangle for *lawType\_CIVIL* shows lift (the value of 47% is greater than 26.7%), this attribute-value is not a good refusal contributor because the height of it's bar represents only a small number of cases for the Italian group. The circled attribute-values (*lawType\_FAMILY*, *assignment\_INHOUSE*, *matterCode\_FO*, *age\_41..50* and *age\_51..60*) illustrated in Figure 4.12 represent attributed-values that possibly contributed to greater refusal rate for the Italian applicants because their triangles are above the straight line and the corresponding bars show a great number of such cases. In the next step WebAssociate automatically identifies those strong contributors and compares them with the refused Australian applicants. We call this step "QUALIFY LIFT". WebAssociate provided the VLA domain expert with the



graph illustrated in Figure 4.13 automatically, which made the exploration of the graph illustrated in Figure 4.12 optional. However, we provided the graph illustrated in Figure 4.12 in order to enable the VLA expert to understand why WebAssociate generated the suggestions illustrated in Figure 4.13.

- **QUALIFY LIFT** In this step the VLA domain expert was shown a graph illustrated in Figure 4.13 that suggested reasons for the higher refusal rate of the Italian born applicants. As discussed above, WebAssociate automatically identified the attribute-values circled in Figure 4.12 which became labels illustrated in Figure 4.13. The blue squares (connected by the blue line) represent the Italian group, while the red triangles (connected by the red line) represent the Australian group. For example, the height of the blue square with X axis label *lawType\_FAMILY* is a graphical representation of the confidence for the association rule *country\_ITALY AND lawType\_FAMILY ⇒ refused\_YES (confidence 35%)*. Values represented by the blue squares correspond to the same confidence values circled in Figure 4.12. The height of the corresponding red triangle for X axis label *lawType\_FAMILY* is a graphical representation of the confidence for association rule *country\_AUSTRALIA AND lawType\_FAMILY ⇒ refused\_YES (confidence 15%)* as illustrated in Figure 4.13. The height of the bar for X axis label *lawType\_FAMILY* in Figure 4.13 represents the confidence for the association rule *country\_ITALY AND refused\_YES ⇒ lawType\_FAMILY*. For example, the graphical representation for X axis label *lawType\_FAMILY* illustrated in Figure 4.13 suggested to the VLA domain expert that *lawType\_FAMILY* is a good contributor for the greater refusal rate. The reason for this suggestion is supported by the following observations: More Italian applicants that applied for FAMILY matters were refused (35% represented by the height of the blue square) than Australian applicants that applied for FAMILY matters (15% represented by the height of the red triangle) and 49% of all refused Italian applicants were refused legal aid for law type FAMILY (represented by the height of the yellow bar). By using his domain knowledge, the VLA expert knew that some attributes such as older age groups, family law type (especially matter code family other *FO*) would result in a greater refusal rate regardless of the country of birth. This domain knowledge was supported by enabling the user to generate graphs illustrated in Figures 4.14 and 4.15. As illustrated in Figure 4.14, family law type applicants (second last bar from the right) have the greatest probability to be refused legal aid amongst all other law types. Figure 4.14 illustrates that older applicants, especially applicants between 51 to 60 years old, have greater probability to be refused than younger applicants. In order to explain higher refusal rate of the Italian appli-

cants, the VLA domain expert was provided with the characteristics of both country groups, as illustrated in Figure 4.17.

- **EXPLORE** The VLA domain expert explored Italian and Australian groups, illustrated in Figure 4.17 in order to discover their characteristics (displayed as iso\_consequent rule sets). The VLA domain expert was looking for any characteristic of the Italian applicants that had the confidence value (represented by the height of blue triangles) higher than the confidence value (represented by the height of red triangles) of the Australian applicants. For example, the third iso\_consequent from the left labeled *lawType\_FAMILY* in Figure 4.17 shows that a greater proportion of Italian applicants apply for law type FAMILY than Australian applicants. The other characteristics highlighted by the green circle show that a greater proportion of Italian applicants: a) failed guidelines and means tests; b) are male; c) applied for matter code “family other”; d) are older.

#### 4.1.5 Hypothesis Explanation

By using his domain knowledge and all information from the Figures 4.12, 4.14, 4.13, 4.13 and 4.17 the VLA domain expert was able to infer that the refusal rate of the Italian born applicants was higher than the refusal rate of the Australian born applicants not due to their ethnicity but due to the following reasons:

- Most of the Italian applicants are older - Figure 4.17. Most of refused Italian applicants are older - Figure 4.12. Older applicants regardless of country are refused more often - Figure 4.15. The user explained that older applicants are generally refused aid because of their wealth.
- A great proportion of Italian applicants applied for family law type cases - Figure 4.17. Family law type matters regardless of country are refused more often than civil or criminal matters - Figure 4.14. A great proportion of Italian applicants that applied for family law type cases were refused legal aid - Figure 4.13 and Figure 4.12.
- Italian applicants applied more for family other matters (FO) than any other matters - Figure 4.17. Matters “Family Other” regardless of country have high refusal rate - Figure 4.16. A great proportion of Italian applicants that applied for “Family Other” matters have been refused aid - Figure 4.13 and Figure 4.12.
- The significant number of Italian applicants failed the Means and Guidelines and Means tests.

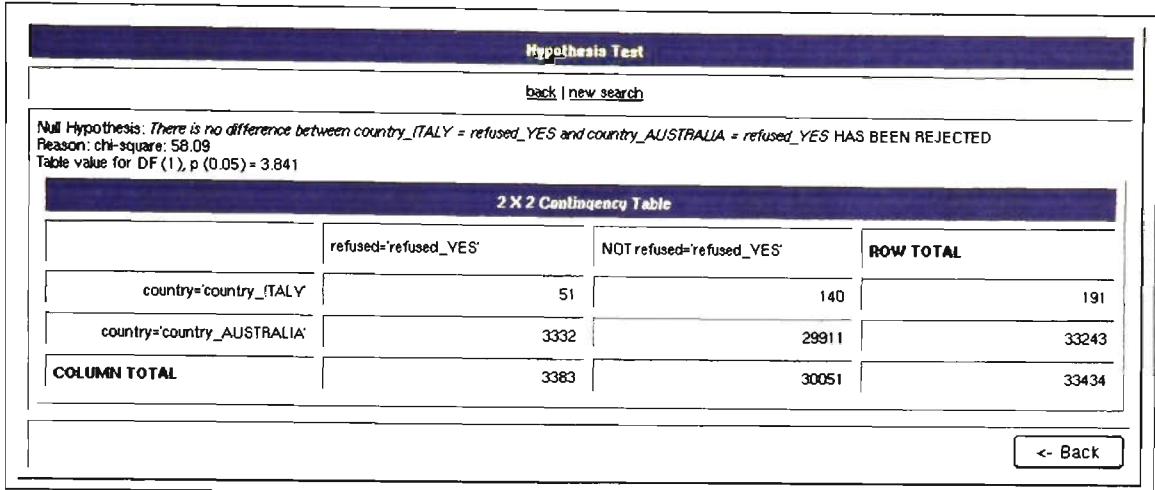


Figure 4.11: Chi-square test for the refused Australian and Italian VLA applicants

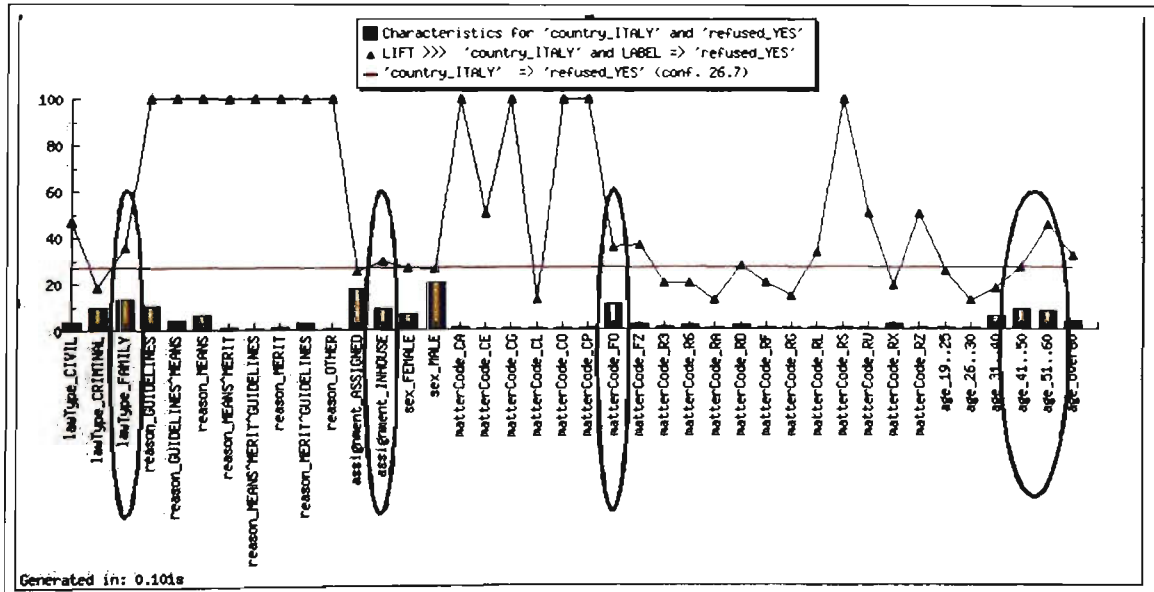


Figure 4.12: Lift: Attribute-values that contribute to refusal greater than 26.7%

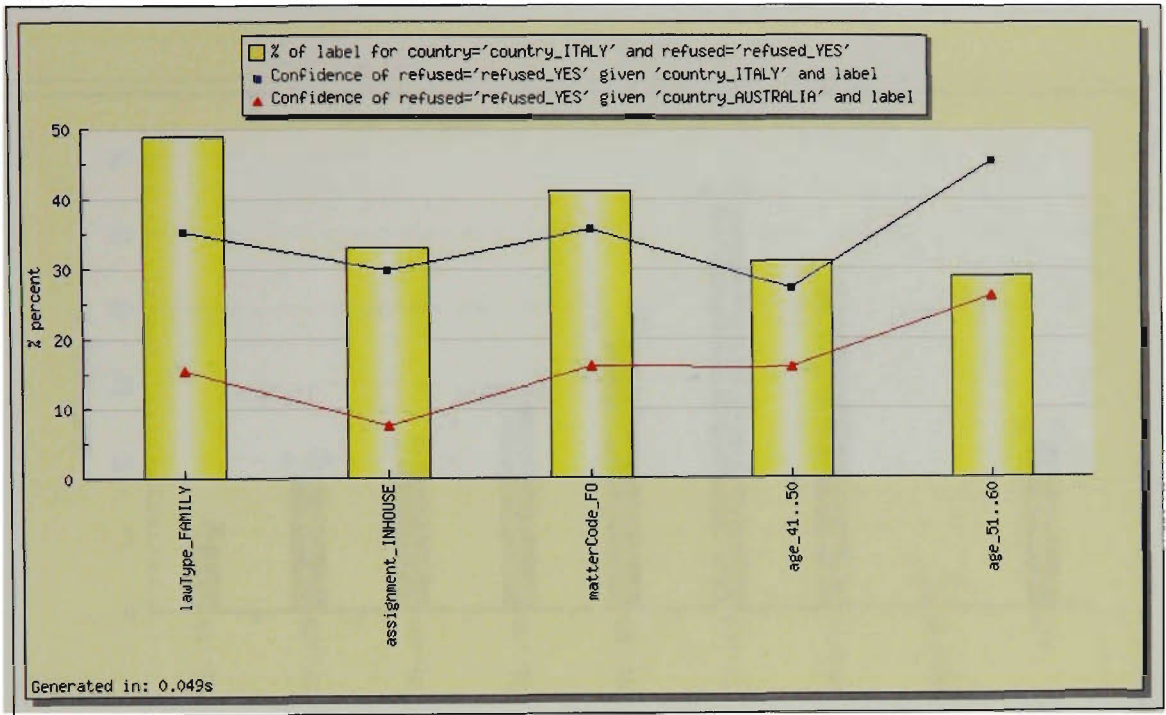


Figure 4.13: Suggested reasons for the higher refusal rate of the Italian born applicants

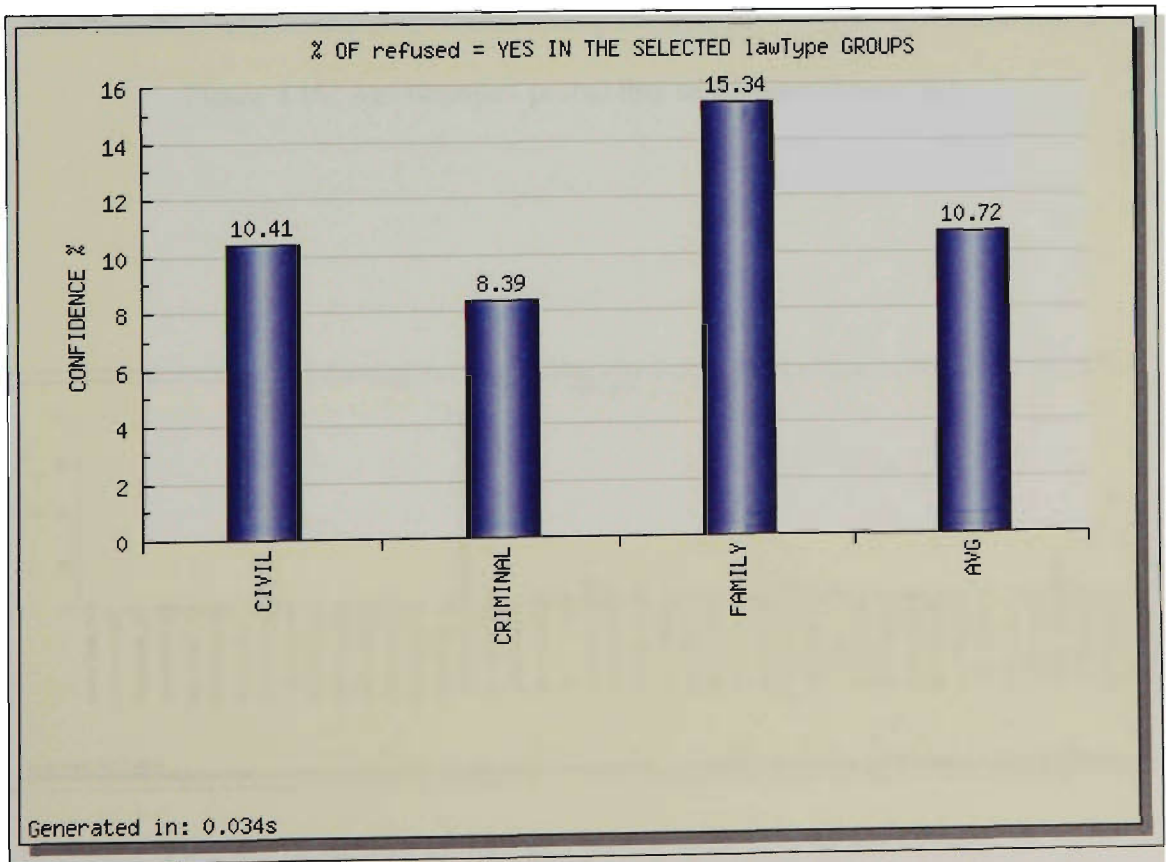


Figure 4.14: Law Type - probability to be refused legal aid

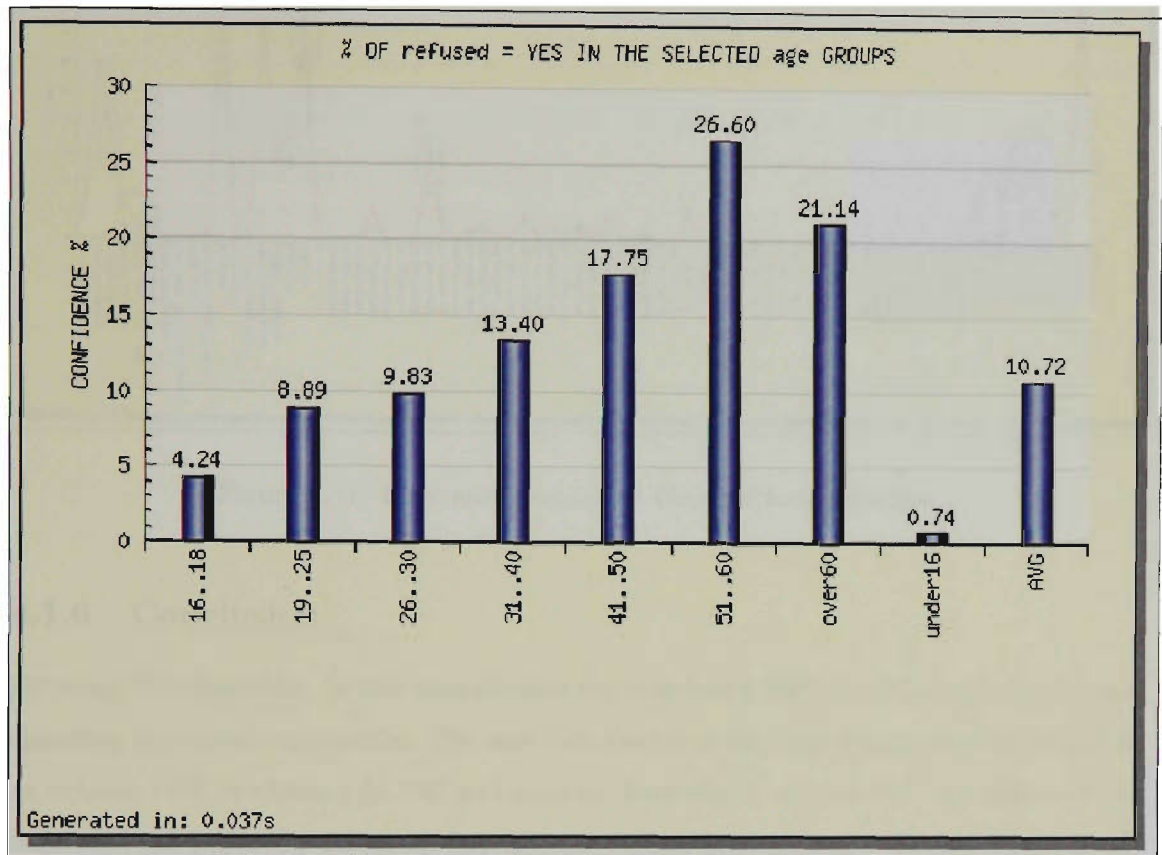


Figure 4.15: Age Groups - probability to be refused legal aid

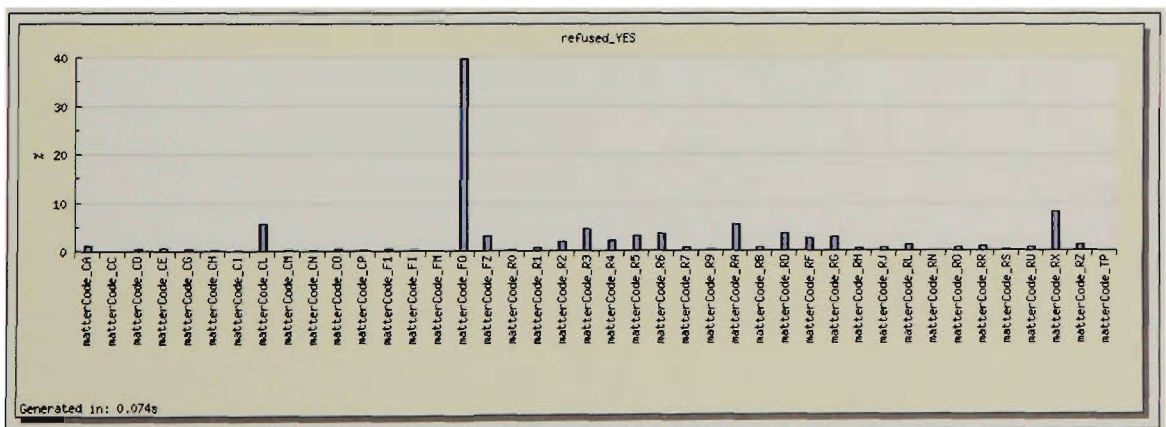


Figure 4.16: 40% of all refused - matter code FO (Family Other)

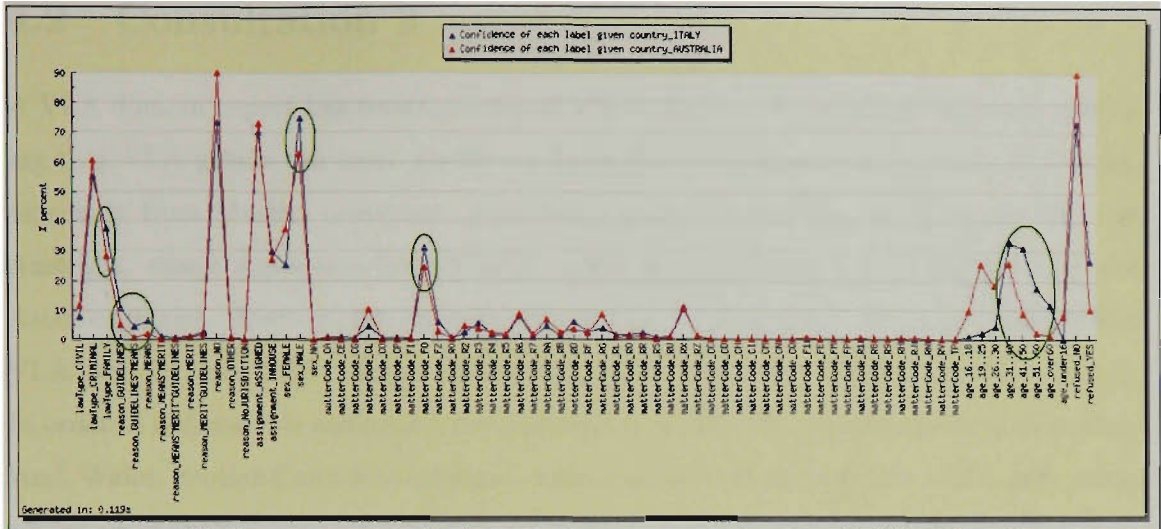


Figure 4.17: Italy and Australia - Group Characteristics

### 4.1.6 Conclusion

By using WebAssociate, in this consultation the user found the discovered rules and corresponding hypothesis very useful. The user indicated that the association rules  $country\_Italy \Rightarrow refused\_YES$  [confidence 26.7%] and  $country\_Australia \Rightarrow refused\_YES$  [confidence 10%] suggested the previously unknown hypothesis “*Italian born applicants are more refused aid than Australian born applicants*”. Furthermore, by using WebAssociate the user was able to identify possible explanations for this hypothesis and conclude that the high refusal rate of the Italian born applicants was not due to their ethnicity.



## 4.2 Consultation 2

A VLA domain expert has received a letter from a disgruntled legal aid applicant, claiming that VLA grants aid more readily to Anglo-Saxon applicants than other applicants, especially from Muslim countries. Most Anglo-Saxon applicants are from the UK and Australia, where applicants from Muslim countries come from several Middle East and Asian countries. However, the data set does not have entries such as UK or Muslim. The VLA data for year 2002 contains 42,434 records with applicants from over 140 countries. In order to address this complaint, the domain expert groups relevant countries (e.g. England, Wales, Scotland and Nth.Ireland are grouped as UK) and explore VLA applications according to their *country of birth* values. In this consultation we guide the domain expert in order to test the hypothesis “*More Muslim applicants get refused aid than Anglo-Saxon applicants*”.

### 4.2.1 Country selections for the UK and OZ groups

Prior to the hypothesis test the VLA expert had to define “UK”, “OZ” and “Muslim” country groups because the VLA data set does not have such data item entries. The expert selected the *country* attribute as illustrated in Figure 4.18 in order to select countries that belong to a group of interest.



Figure 4.18: Attribute selection

WebAssociate allows the domain expert to group data items, in this instance countries, and give a name to the new defined group. The domain expert selected countries that are in the “UK” and selected the option “Define Selected” as illustrated Figure 4.19.

After the selection, the domain expert defined countries England, Northern Ireland, Wales and Scotland as UK as illustrated in Figure 4.20.

In the next step the expert was prompted to define a new group or to continue with the exploration. The VLA expert decided to define a new group as illustrated in Figure 4.21.

The expert decides to define Australian applicants as a separate group in order to identify the refusal rate difference independently from the “UK” applicants. Figure 4.22

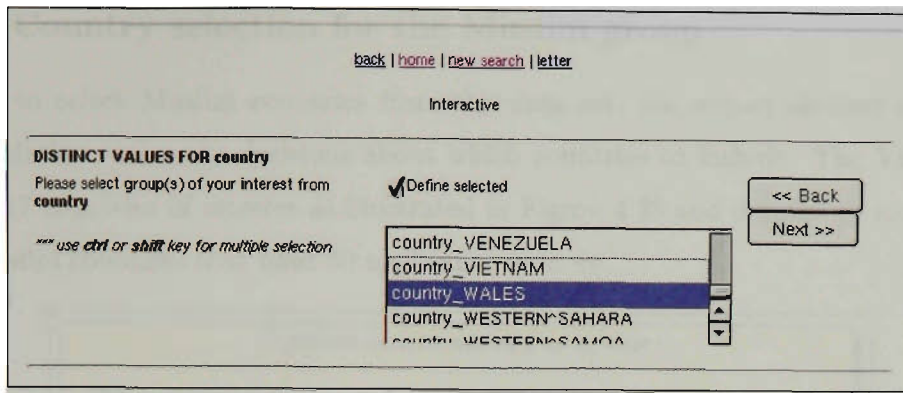


Figure 4.19: UK countries selection

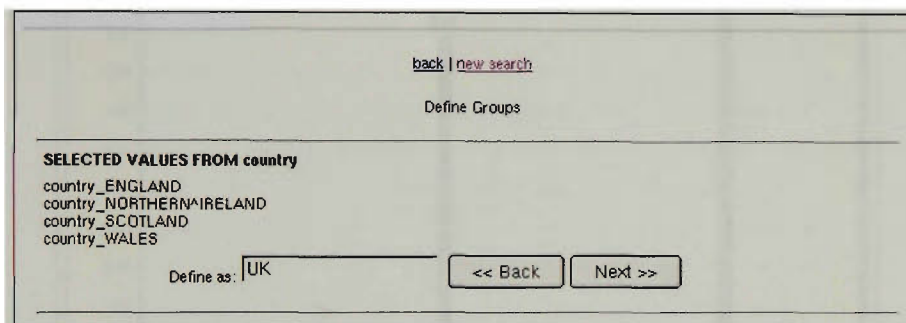


Figure 4.20: UK group definition

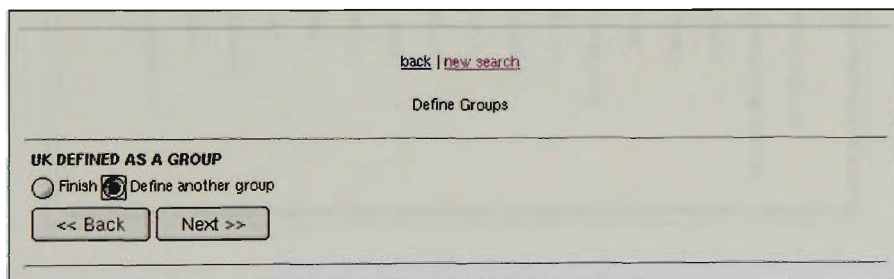


Figure 4.21: New group definition

illustrates the definition of the “OZ” group which contains the Australian applicants.

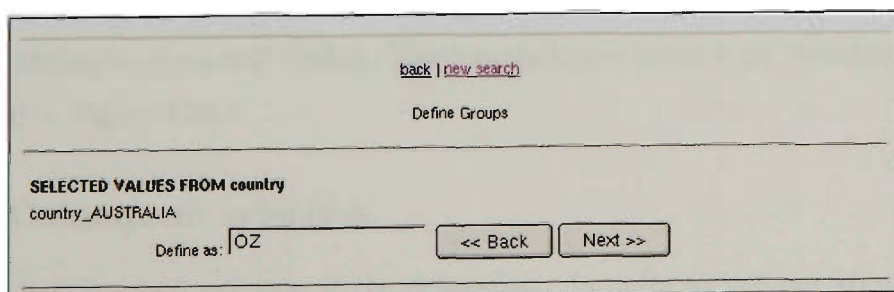


Figure 4.22: OZ group definition



## 4.2.2 Country selection for the Muslim group

In order to select Muslim countries from the data set, the expert decided to explore this population and make decisions about which countries to include. The VLA expert selected 17 countries of interest as illustrated in Figure 4.23 and decided to include only eight Muslim countries that have 30 applicants or more.

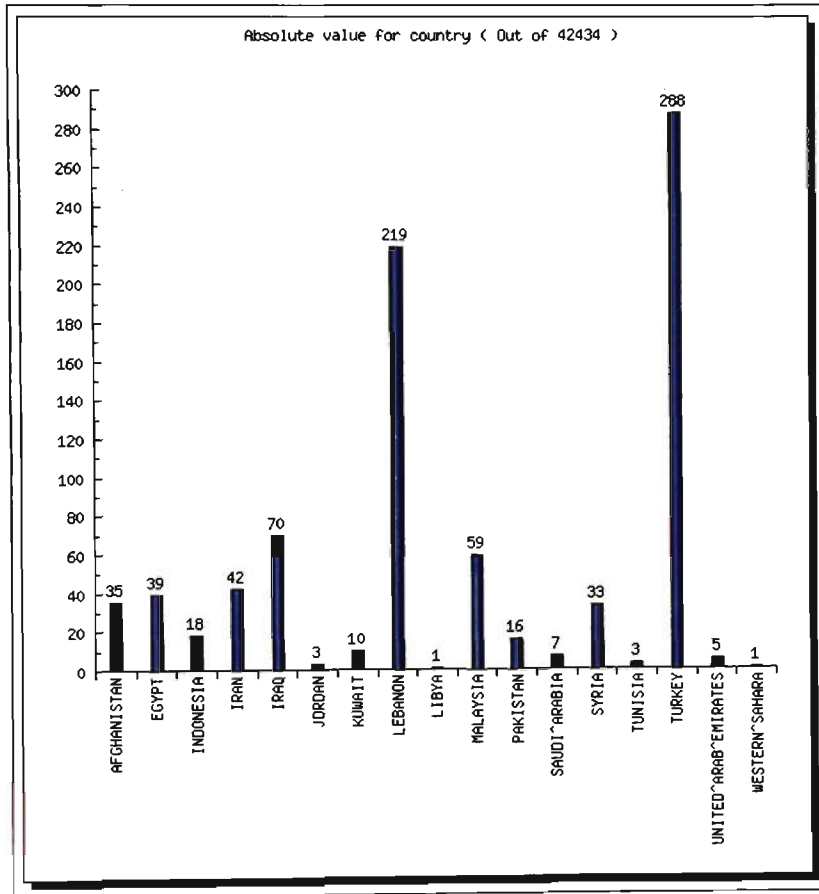


Figure 4.23: Country selection - Muslim group

Eight countries were included in the Muslim group and they represent the majority of the applicants that were born in predominantly Muslim countries. The expert defined “Muslim” group and included the following eight countries: Afghanistan, Egypt, Iran, Iraq, Lebanon, Malaysia, Syria and Turkey. The domain expert defined the “Muslim” group as illustrated in Figure 4.24.

## 4.2.3 Consequent selection

The next step involved selection of the consequent. In order to test the hypothesis, the expert selected *refused\_YES* as illustrated in Figure 4.25.

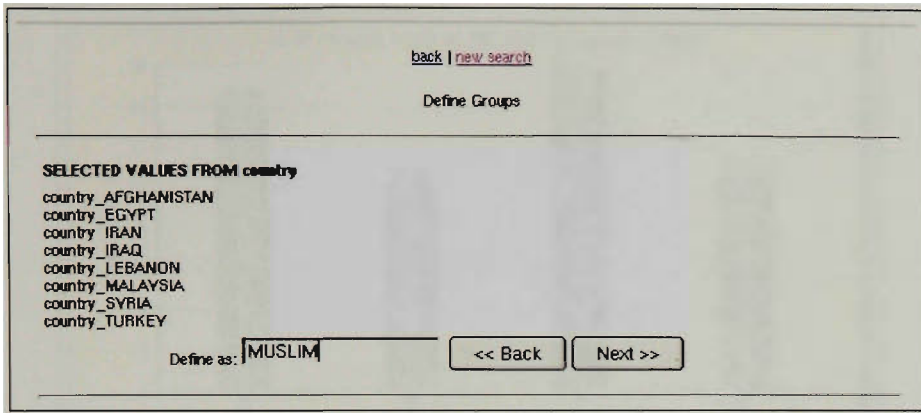


Figure 4.24: Muslim group definition

FIELD	VALUES
lawType	none
reason	none
assignment	none
sex	none
matterCode	none
age	none
decided	none
refused	refused_YES

Figure 4.25: Consequent selection for the groups of interest

#### 4.2.4 Hypothesis exploration

WebAssociate generated the graph showing the refusal rate for the defined groups of interest as illustrated in Figure 4.26. The graph showed that the refusal rate of the “Muslim” group (15.92%) was slightly higher than the refusal rate of the “UK” group (13.96%). The refusal rate of the “OZ” group was slightly lower (10.02%) than the refusal rate for the other two groups.

The VLA expert decided to investigate reasons for refusal for the “Muslim” group and discovered that the high number of the “Muslim” group applicants were refused on the basis of guidelines *reason\_GUIDELINES* (7.01%) as illustrated in Figure 4.27.

The VLA domain expert explained that the reason for the higher refusal rate based on the guidelines amongst the “Muslim” group as illustrated in Figure 4.28, is because most of the applicants from this group apply for legal aid while they still have the “Refugee” status which makes them ineligible for legal aid.

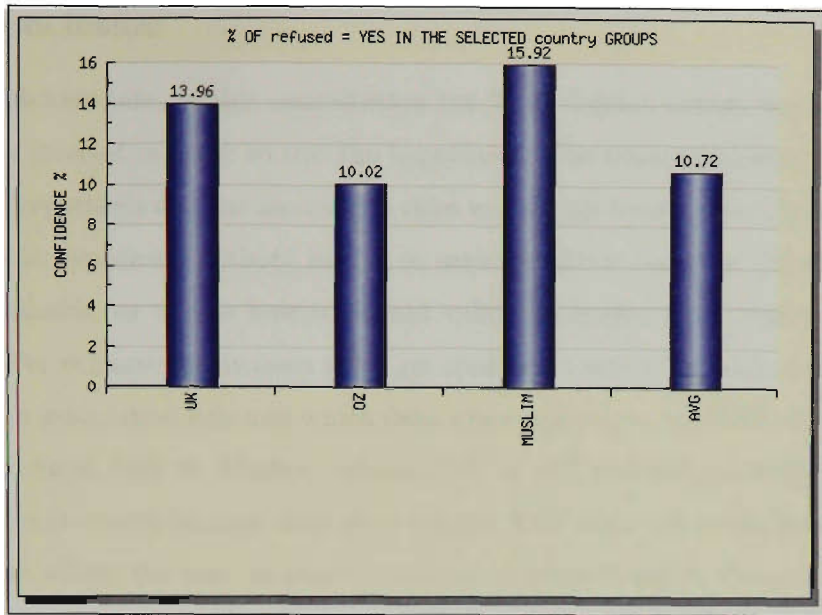


Figure 4.26: Refusal rate for the UK, OZ and MUSLIM groups

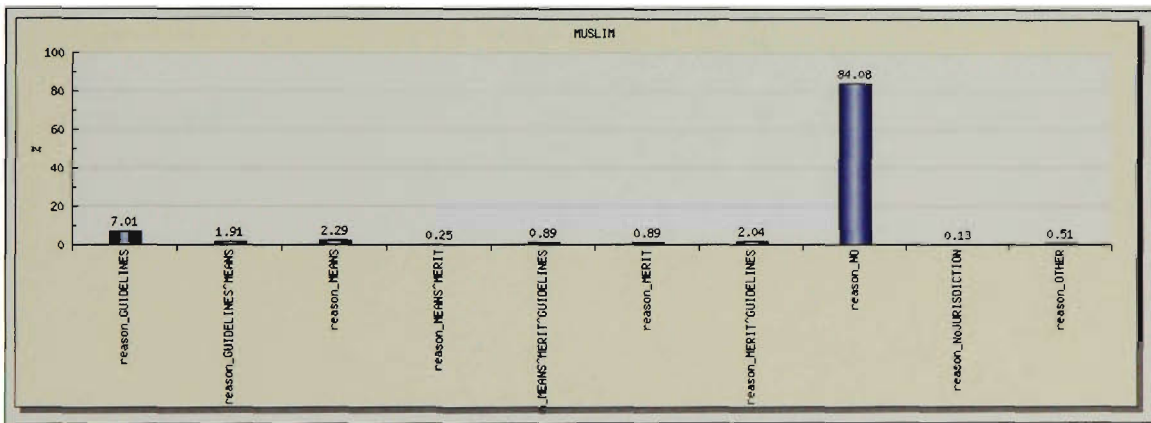


Figure 4.27: Reason for refusal for the MUSLIM group

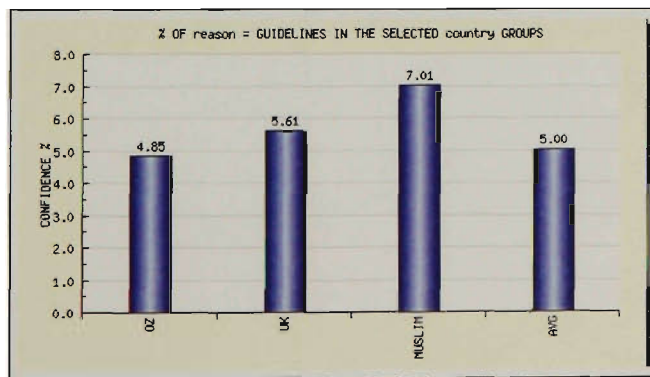


Figure 4.28: Refused on the basis of guidelines

### 4.2.5 Conclusion

By using WebAssociate, in this consultation the VLA domain expert was able to define the groups of interest in order to test the hypothesis. The expert indicated that mapping between the hypothesis and the association rules was simple because WebAssociate allowed him to test the hypothesis without having to make decisions based on the confidence and support thresholds, as well as how to format valid association rules that would test this hypothesis. For example, some users often get confused about which data-items appear on the LHS of an association rule and which data-items appear on the RHS. For instance the set of rules  $refused\_YES \Rightarrow Muslim$ ,  $refused\_YES \Rightarrow OZ$  and  $refused\_YES \Rightarrow UK$  would provide incorrect results because data item  $refused\_YES$  is on the wrong side of each rule. WebAssociate allows the user to select variables of interest which always appear on the LHS of a rule. This approach tends to reduce possible errors in the mapping between a hypothesis and corresponding association rules.

### 4.3 Consultation 3

There is a report from the Victorian police that most drug related offences are from the Vietnamese born community. VLA is interested in finding out more about the Vietnamese applicants and wants to report on the previous VLA cases that involved Vietnamese applicants with drug related cases *matterCode\_RD*. In this consultation we guide a VLA expert to explore this and find characteristics of that population.

#### 4.3.1 Variable of interest selection

In order to explore Vietnam born applicants that applied for the drug related matter types, the user selected *country\_VIETNAM* for variable of interest as illustrated in Figure 4.29.

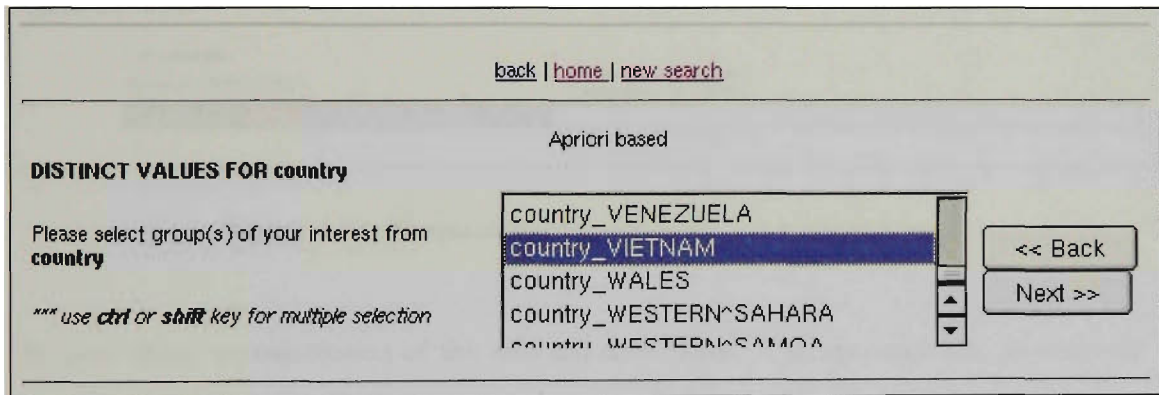


Figure 4.29: Country Vietnam selection

#### 4.3.2 Consequent selection

In the next step the user decided to explore matter codes that were most frequent for the Vietnam born applicants and selected matter code (consequent) as illustrated in Figure 4.30.

#### 4.3.3 Common matter codes for Vietnam

WebAssociate generated association rules containing the most common matter codes for the Vietnam born applicants. As Figure 4.31 illustrates, 30.5% of the Vietnam born applicants applied for drug related matters (*matterCode\_RD*), where 11.3% of these applicants applied for the driving related matters (*driving without licence - matterCode\_R6*).

The user decided to explore the drug related matters *matterCode\_RD* of the Vietnam born applicants. WebAssociate allows the user to find the actual values rather than just

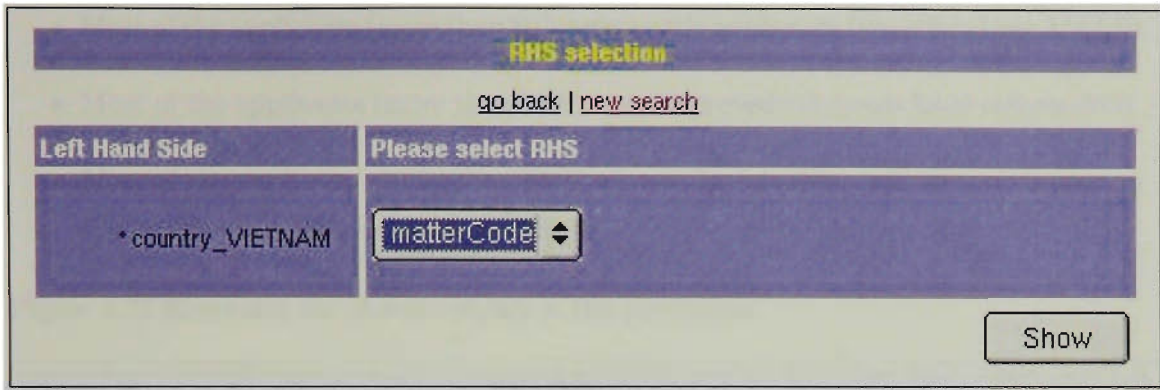


Figure 4.30: Consequent selection for country Vietnam

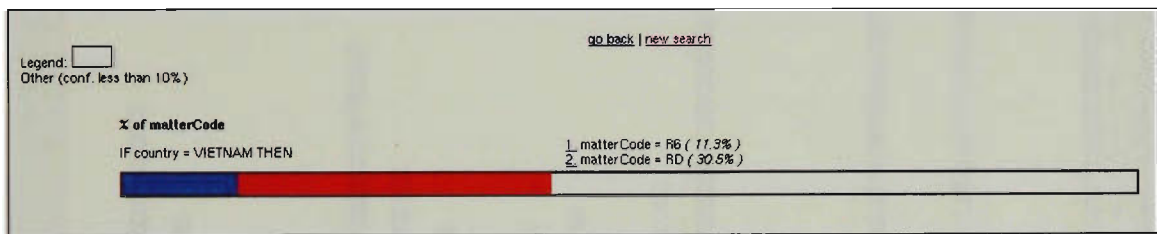


Figure 4.31: Frequent matter codes for the Vietnamese

the percentage representation of the selected data-items. The user was able to discover that 291 of 955 Vietnam born applicants have applied for aid for drug related matters, as illustrated in Figure 4.32.

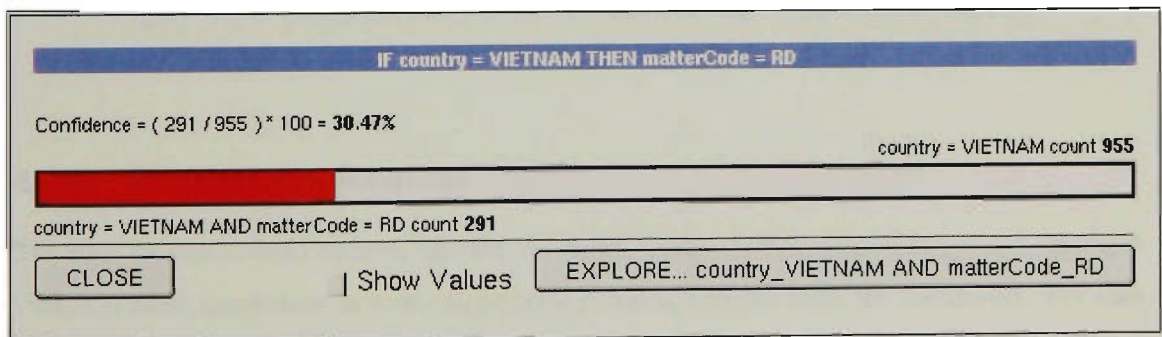


Figure 4.32: The number of Vietnam applicants and drug related matters

#### 4.3.4 Country Vietnam and drug related matters

By deciding to explore the population of Vietnam born applicants that applied for aid for drug related matters, the user discovered the following characteristics:

- Most of the applicants were younger, with 51% of the applicants being age 19 to 25 (yaxis label *age\_19..25*).



- Most of the applicants (more than 80%) were male applicants (yaxis label *sex\_MALE*).
- Most of the applicants (more than 95%) were approved aid (yaxis label *refused\_NO*).
- Most of the applicants (more than 65%) were assigned a lawyer from outside sources (yaxis label *assignment\_ASSIGNED*).

Figure 4.33 illustrates the characteristics of the population.

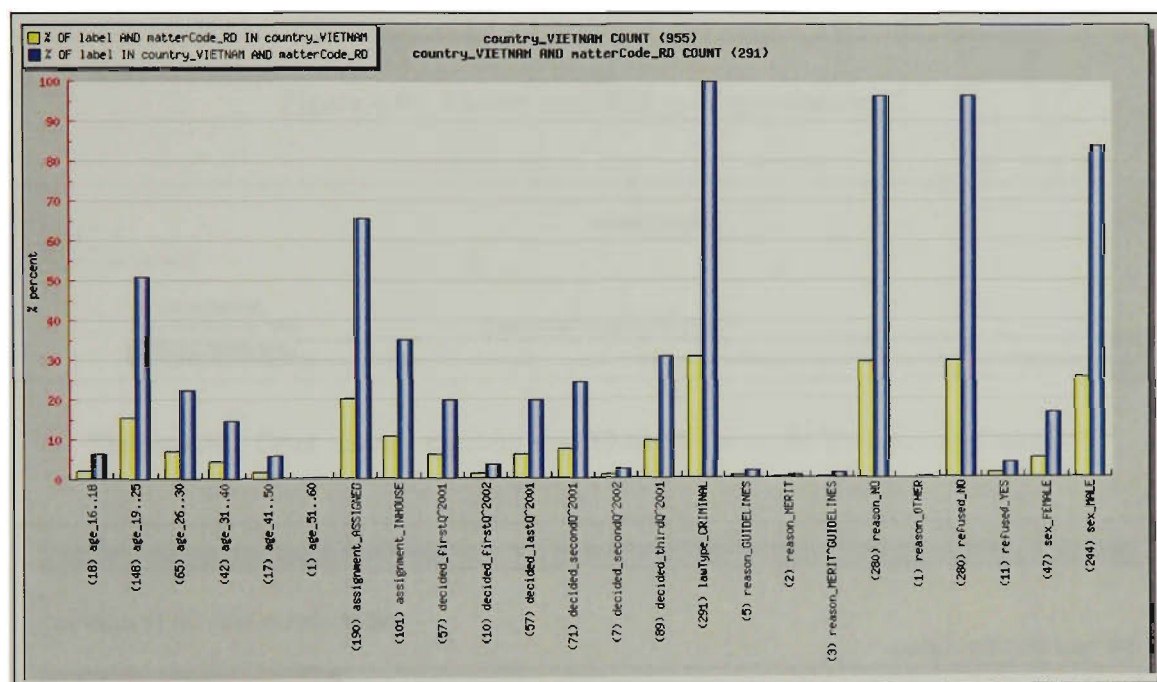


Figure 4.33: Characteristics of the Vietnamese and drug related matters

### 4.3.5 Further exploration

The user was interested in investigating the most common drug related age group amongst Vietnam born applicants in order to provide some additional facts for the report. For this purpose the user selected matter code and age data items as illustrated in Figure 4.34.

WebAssociate generated the graph, illustrated in Figure 4.35, showing that 15.5% of the Vietnam born applicants are age group 19 to 25 (young applicants) who applied for aid for drug related matters.

As illustrated in Figure 4.36, 148 applicants were young Vietnamese (19 to 25 years old) who applied for drug related cases (matterCode RD). The user found that the actual number of these cases (148), rather than just the percentage value (15.5%) provided useful information to be included in his report.

Finally, the user was able to discover the following characteristics of this group of young applicants:

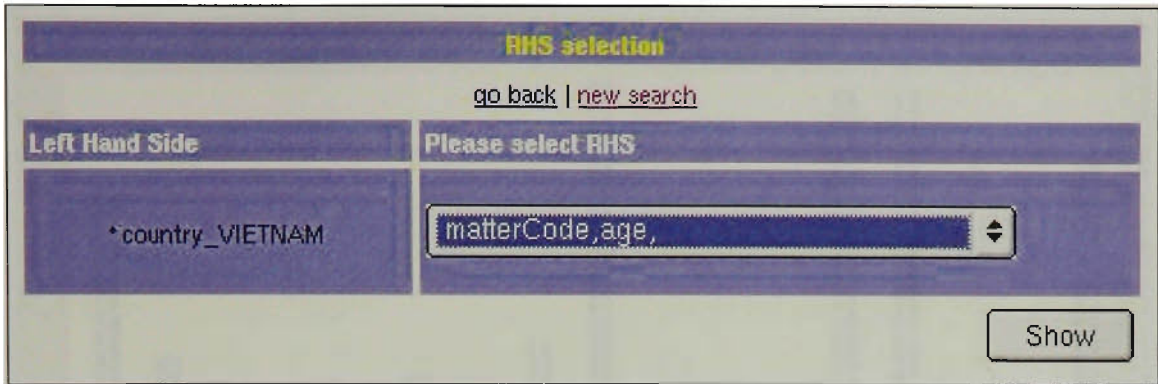


Figure 4.34: Matter code RD and age selection

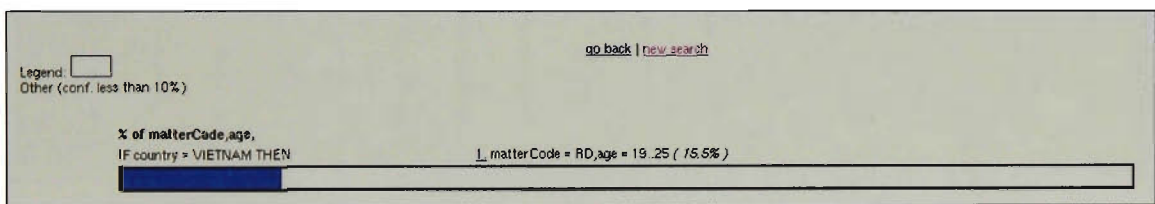


Figure 4.35: Drug related matters and 19 to 25 years old Vietnam applicants

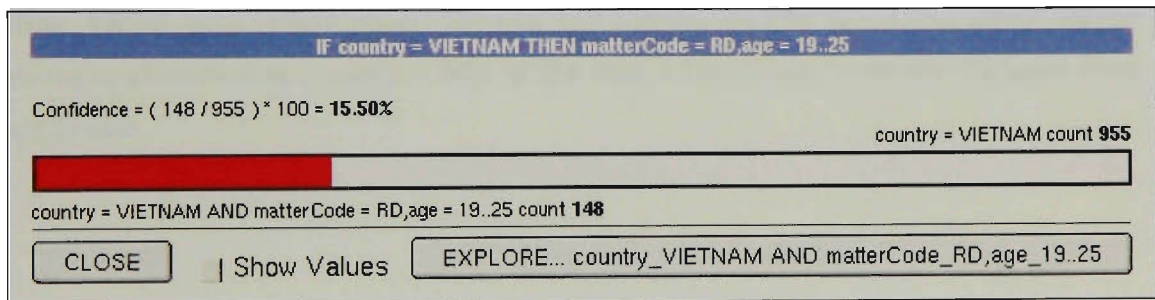


Figure 4.36: Vietnam, age 19 to 25, drug related applicants

- Most of the applicants (79%) were male applicants (yaxis label *sex\_MALE*) where only 21% of the applicants were female (yaxis label *sex\_FEMALE*).
- Most of the applicants (more than 95%) were approved aid (yaxis labels *refused\_NO* and *reason\_NO*).
- Most of the applicants (more than 70%) were assigned a lawyer from the outside sources (yaxis label *assignment\_ASSIGNED*).

### 4.3.6 Conclusion

In this consultation the user was able to explore the population under study in order to provide a report on the previous VLA cases that involved Vietnamese applicants with



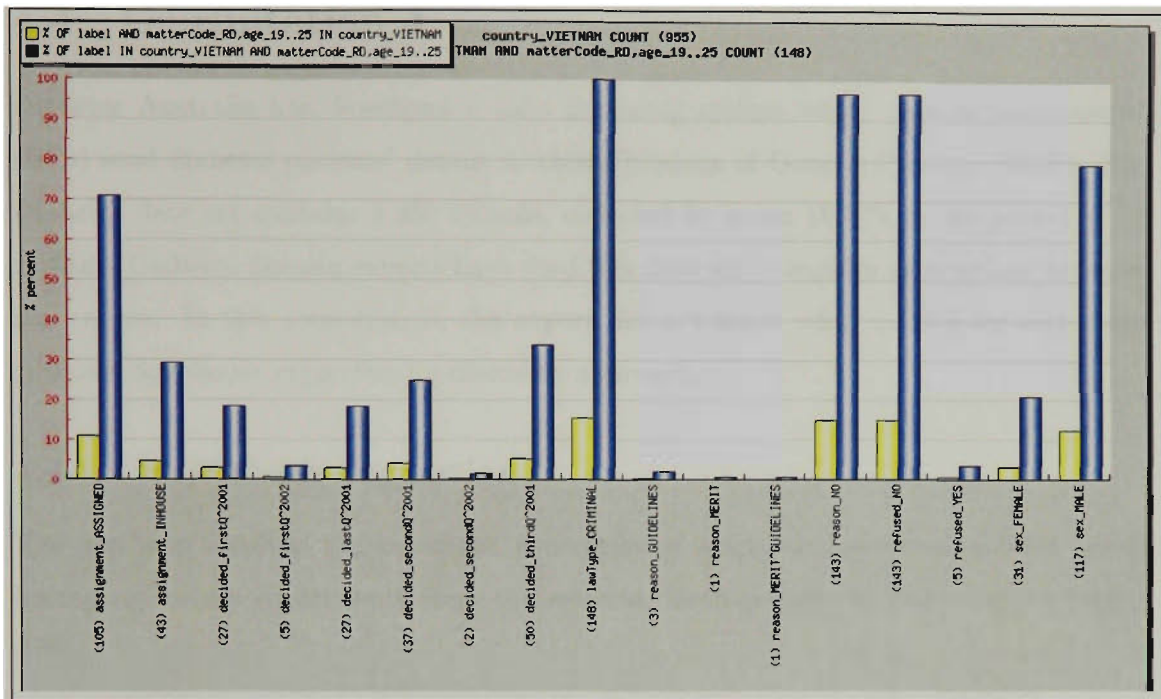


Figure 4.37: Characteristics of the Vietnam born, age 19 to 25, drug related applicants

drug related cases. Furthermore, the user was able to describe the characteristics of the young applicants that constitute 50% of the drug related matters for the Vietnam born applicants.

## 4.4 Consultation 4

Diabetes Australia has developed a data gathering system where general practitioners (GPs) send diabetes patients' details to their Divisions of General Practice (DGP). The Diabetes data set contains 4,359 records, collected by seven DGP's in the period of 12 months. Diabetes domain experts have used this data set to explore associations between data items. In this consultation, the expert did not know what to look for and chose (*unkown hypothesis suggestion*) a discovery approach.

### 4.4.1 Hypothesis suggestion

The first step involved the automatic generation of single item association rules. Each button represents an attribute from the selected database-table as illustrated in Figure 4.38.

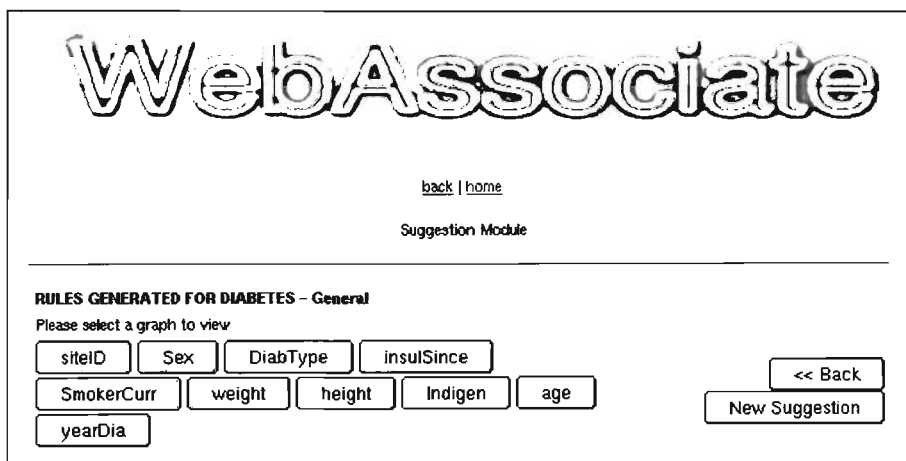


Figure 4.38: Diabetes attribute selection

WebAssociate allows the user to scan through the suggested graphs (displayed by click on a button) and find possible interesting patterns. Each button represents a corresponding attribute from the Diabetes dataset.

### 4.4.2 Choosing a variable of interest

The click of a button generates a graph containing the values for the selected attribute as the antecedents and non-selected attributes-values as the consequents. The expert visually examined all available graphs and selected *SmokerCurr* as an attribute of interest. The graph illustrated in Figure 4.39, shows three distinct values for this attribute; *SmokerCurr\_0* (never smoked), *SmokerCurr\_1* (current smoker) and *SmokerCurr\_2* (ex-smoker),

represented by a color coded line. The height of each point represents a confidence value (from 0 to 100 percent) for the corresponding label (e.g. *siteID\_430*, *Sex\_1* and *DiabType\_2*) for a given current smoker value.

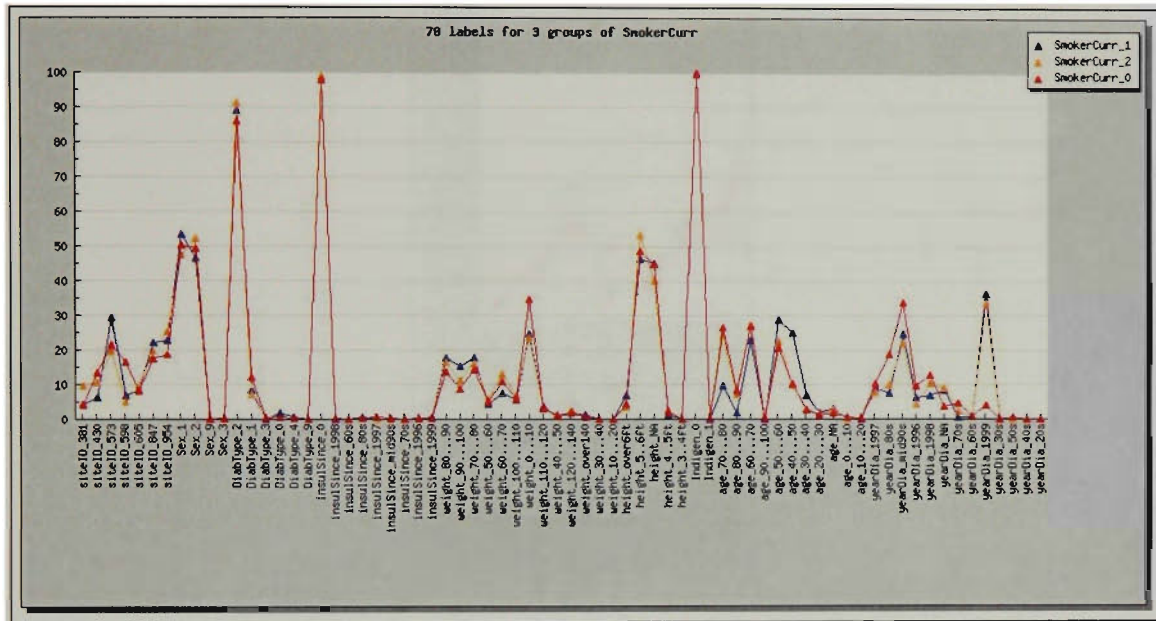


Figure 4.39: Diabetes attribute selection - current smokers

The user was interested in exploring a few characteristics of the current smokers (*SmokerCurr\_1*) represented by the blue line. By visually exploring the characteristics illustrated in Figure 4.39, the user discovered that most of the current smokers were older patients (labels *age\_40..60*, *age\_50..60* and *age\_60..70*), with diabetes type 2 (label *DiabType\_2* with 80% confidence) and almost one third of the patients who are current smokers were from one DGP (label *siteID\_573* with 29% confidence).

#### 4.4.3 Discovery interestingness

The user identified the rules  $SmokerCurr_1 \Rightarrow siteID_573$  [confidence 29%],  $SmokerCurr_1 \Rightarrow siteID_847$  [confidence 22%] and  $SmokerCurr_1 \Rightarrow siteID_954$  [confidence 23%] very interesting, and inferred the hypothesis “Three quarters (75%) of the current smokers are from DGP with sites id 572, 847 and 954”. Figure 4.40 shows the conditional probability for each DGP.

The graphical presentation of the interesting rules is illustrated in Figure 4.40. The third, sixth and seventh labels from the left are representing *siteID* as the consequents (RHS) and the lines represent *SmokerCurr* as the antecedents (LHS). The user was especially interested in the rule  $SmokerCurr_1 \Rightarrow siteID_573$  [confidence 29%], represented by the intersection of the third label from the left and blue triangle, because this rule shows

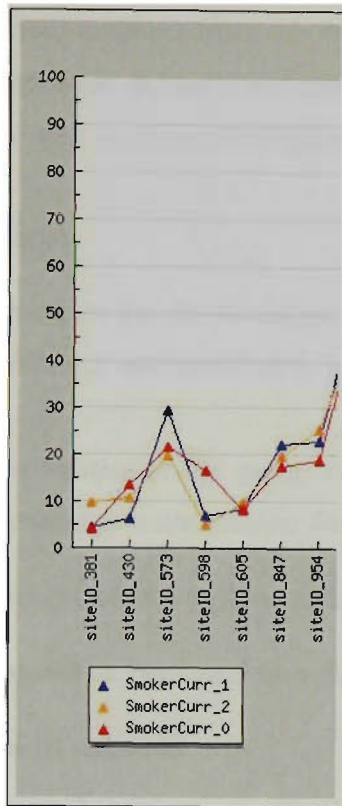


Figure 4.40: 75% of the current smokers are from the DGP sites 573, 847 and 954

that almost one third of the current smokers are from DGP 573. The user indicated that additional educational aids are needed for this DGP in order to reduce the number of current smokers amongst those diabetic patients. The next step involved further exploration of the *SmokerCurr\_1* and *siteID\_573* population in order to target the appropriate patients as shown in Figures 4.41 and 4.42.

#### 4.4.4 Rule exploration

Figure 4.41 shows the rule  $SmokerCurr_1 \Rightarrow siteID_573$  and actual counts which the user found useful in order to plan for the additional anti-smoking educational aids.

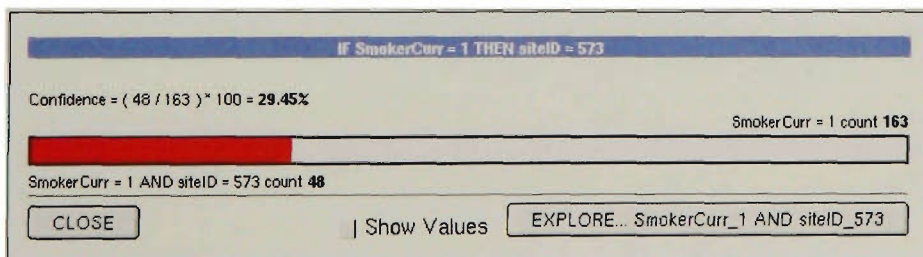


Figure 4.41: Rule:  $SmokerCurr_1 \Rightarrow siteID_573$  [confidence 29%]

As illustrated in Figure 4.41, only 48 out of 163 patients that are current smokers

represent the population of current smokers from the DGP 573. However the user has indicated that 48 patients still represent an important sub-population that have increasing risk of cardiovascular diseases as well as diabetic complications due to smoking.

#### 4.4.5 Further exploration

In order to target appropriate patients for the educational program, the user selected to explore this population (*SmokerCurr\_1 and siteID\_573*) as illustrated in Figure 4.42, and find their characteristics.

Figure 4.42 illustrates characteristics of the group under study and shows that the current smokers in division of general practice number 573

- are diabetes 2 type patients (46 of 48 - 96%) and they make 28% of all smokers
- have 52% of patients diagnosed in 1999 and they make 15% of all smokers
- are mostly between 40 and 70 years old
- are not correlated with gender (50% are male and 50% are female)

#### 4.4.6 Conclusion

By using WebAssociate, in this consultation the user found the discovered rules and corresponding hypothesis very useful. The user indicated that previously unknown patterns such as *SmokerCurr\_1  $\Rightarrow$  siteID\_573 [confidence 29%]* helped the user focus on this population in order to provide additional educational aid which can reduce the number of smokers amongst the diabetic patients. The user also indicated that the tool helped to focus on the particular patients such as diabetes 2, male and female patients that are older than 40 and younger than 70, rather than involving other patients that would be irrelevant for this task (e.g younger patients).

In this consultation we conclude that hypothesis suggestion provides a useful discovery approach when the users are not sure what to look for in the data set.

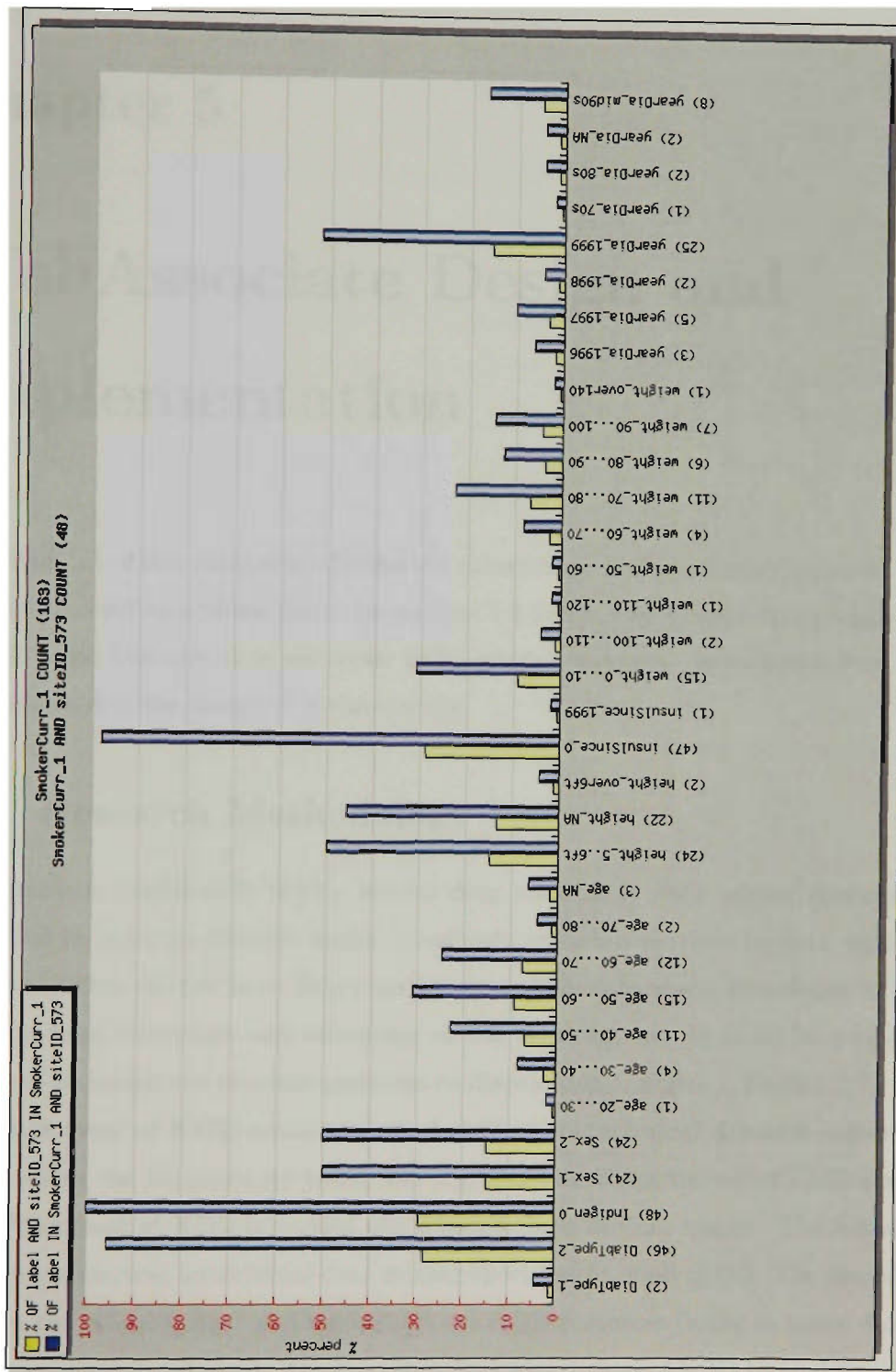


Figure 4.42: Current smokers and site id 573 - Characteristics

## Chapter 5

# WebAssociate Design and Implementation

In Section 5.1 of this chapter we discuss our research questions and describe methods that we used in order to address them. In sections 5.2 and 5.3 we discuss data preparation of the VLA and Diabetes data sets prior to knowledge discovery. In section 5.4 we discuss methods used in the design of WebAssociate.

### 5.1 Research Methodology

Organisations traditionally deploy several data analysts or data mining specialist that use KDD in order to discover useful, previously unknown patterns in data while other employees often do not have direct access nor sufficient technical knowledge to exploit KDD to their advantage and advantage of the organisation. In order to address this problem, we revisit the research questions as discussed in Chapter 1, Section 1.7.

#### **What type of KDD tools are needed for non-technical domain experts?**

By reviewing the literature we found that organisations adopt the use of KDD gradually. The deployment of KDD is carried out through three distinct stages. The initial stage involves contracting an external data mining specialist to apply KDD. The second stage involves the establishment and use of in-house KDD resources (using in-house data analysts, hardware and software). The final stage includes the use of KDD by enabling domain experts (e.g. lawyers, managers and medical professionals) to perform their own analysis according to their individual needs.

We further investigated the use of KDD by contacting several Australian organisations (VLA, Diabetes Australia and Australian Timken PTY LTD) who were willing to partici-

pate in our study. We investigated the current use of KDD in the organisations mentioned above and identified that none of these organisations deploy KDD to its full potential.

In order to make KDD more useful to an organisation we believe that the full potential of KDD could be reached only if organisations deploy all three stages and enable the use of KDD through all organisation levels. The domain experts involved in our study pointed out that they were unable to explore their data sets regularly (through the use of data analysts) because they were fully dependent on the IT departments and professional data analysts but those departments were usually under resourced. Furthermore, the organisational non-technical domain experts pointed out that the current KDD tools in their organisation (if any) were complicated tools that used advanced methods and algorithms, required high technical knowledge and abtruse training. The experts were interested in extracting useful reports, spotting interesting events and trends and supporting their decisions without necessarily becoming technical experts. The experts expressed the need for an easy to use tool that will automatically suggest hypotheses, rather than another sophisticated analytical tool. Furthermore, the non-technical domain experts had a demand for a tool that uses descriptive rather than predictive methods. The purpose of this study is not to build a tool which should replace the current KDD tools that are used for more sophisticated data analysis. It is to build an additional tool which would use descriptive KDD methods to suggest hypotheses and enable non-technical users to explore their data sets.

### **How do we model the way that domain experts seek to explain patterns in data?**

By reviewing the literature we found that Association Rules are a suitable descriptive data mining method that provides a simple mechanisms for data exploration. Furthermore, by finding associations between attributes in data, association rules can be grouped together in order to suggest hypotheses. Grouped association rules help domain experts to explain patterns in data. Furthermore, we decided to use Association Rules because we have had previous experience with this technique. In order to test this research question we developed a KDD application called “WebAssociate” which groups discovered association rules in order to enable non-technical domain experts to seek and explain patterns in data. However, we believe that the valuable experience gained during this study would provide us with the basis for further research and further development of easy to use KDD tools that utilize KDD techniques such as rule induction and neural networks.

### **How can association rules be mapped to hypotheses?**



Association rules discover associations between data attributes. An association rule consists of two parts: antecedent “left hand side” and consequent “right hand side”. For example, in rule R1 “IF *age18..25* THEN *watch football* (confidence 70%)”, *age18..25* is the left hand side of the rule while *watch football* is the right hand side of the rule. An association rule algorithm might discover hundreds, even thousands of rules. However some rules might have common left hand side, while other rules might have a common right hand side. For example, rule “IF *age70..80* THEN *watch football* (confidence 10%)” and R1 have common right hand side *watch football*. By grouping those two rules together we suggest the hypothesis “People between 18 and 25 years old watch football more often than people between 70 and 80 old do”. We call this rule set *iso\_consequent* because all association rules in this rule set share the same consequent. The method of grouping corresponding rules together automatically maps association rules to hypotheses.

**What is the appropriate method to organise and present the findings to be more understandable to non-technical domain experts?** By reviewing the literature we discovered that the most appropriate method to organise and present the findings is by displaying findings visually as graphs. Many researchers have shown that users understand visual display better than text. Imperical evidence from our study (Chapter 6) also shows that users understand visual display better than text. Visual presentation of grouped association rules allows users to have clearer understanding of discovered patterns.

**How can easy-to-use KDD tools for non-technical experts be constructed?** The significant number of organisations that use current KDD tools make them available only for small number of people. IT departments usually employ professional data analysts who use those sophisticated tools in order to discover useful patterns in data. Those tools are generally password protected and available on the Local Area Network (LAN) or just available on a single machine. The purpose of our tool is to be available for wider use in an organisation. However, non-technical domain experts need an easy to use tool that does not require installation or configuration. For this reason we developed a client-server web based KDD tool “WebAssociate” that does not require installation nor configuration. The tool is available via users web browser and does not require any additional plugins or applications on the users machine. The organisations involved in our study found the web approach useful because many domain experts are not stationed in a single organisational branch. Furthermore, the use of server based tool does not require a powerful PC. “WebAssociate” can be accessed via the web from any organisational PC regardless to its CPU power.

## 5.2 VLA Data preparation steps

### 5.2.1 Data selection

In this first phase we asked the VLA experts to select variables of interest that would be potentially useful for knowledge discovery. The availability of strong domain knowledge improves the knowledge discovery process by reducing the search space and helps to focus on the interesting findings. The experts selected nine variables that were important and interesting for rule generation. The variables were: lawType, reason for refusal, country, assignment, sex, dob, matter type, decision date and refused.

#### 1. lawType

Three distinct values of this variable are indicating law type of the offence. Distinct values are: Criminal ,Civil and Family.

#### 2. reason for refusal

There are eight distinct reasons for the refusal of legal aid. Reason for refusal “NO” indicates that user has not been refused aid. Other seven reasons indicate that the applicant was refused legal aid on some criteria such as Means, Guidelines and Merit or combination of these.

#### 3. country

This attribute indicates a country of birth of the applicant. Most of the applicants in this data set were born in Australia, while the minority applicants were born overseas. The data set contains records of applicants from over 140 countries. Even though overseas born applicants make only 26.6% of the data set, they still represent an important population for the data analysis.

#### 4. assignment

This attribute indicates the type of assignment of the applicant. There are three assignment types in the data set: INHOUSE, ASSIGNED and UNDETERMINED. The inhouse assignment indicates that the applicant was assigned an inhouse lawyer to represent him/her at court. The value of assigned indicates that the applicant was given money to deploy an outside lawyer.

#### 5. sex

Gender of the applicant.

#### 6. dob

Date of birth of the applicant.

#### 7. matter type

There are over 500 different matter types describing the charges. For example five different matter types are used to describe homicide: attempted murder, murder, conspiracy to murder, manslaughter and culpable driving.

#### 8. decision date

A date indicating the decision date. For example, if an officer made the decision “refused legal aid”, data entry for the field “refused” will automatically set the date of decision.

#### 9. refused

This attribute has two distinct values: “YES” indicating that the user was refused legal aid and “NO” indicating that the user was granted aid.

### 5.2.2 Data pre-processing

Data pre-processing is a necessary step in dealing with large data sets that contain noisy, missing or irrelevant data. However, the VLA data set is presumed to contain minimal noise. Noise is generally regarded as values that are recorded incorrectly because of the data entry processes. The database management system in use by VLA performs integrity checks on all input data so absurd values cannot be entered. Nevertheless, some anomalies were observed by VLA analysts. Analysts suspected that the support for *matter code FO*, (*family matters-other*) was far too high to be accurate. Subsequent investigation revealed that data entry operators deployed the FO type whenever a matter type was unknown.

### 5.2.3 Data transformation

Data transformation includes finding useful features to represent the data, depending on the goal of the task. This means that data has to be transformed in the form that is accepted for the data mining algorithm. Association rules are only useful if numeric (quantitative) or continuous attributes are transformed to intervals. For example, mining each date of birth value would result in too many columns for useful association rules. In this study several VLA data attributes had to be transformed in order to be useful for association rule generation. The attributes “matterCode”, “age” and “decided” were derived from “matter type”, “dob” and “decision date” attributes. VLA domain experts were approached to define the intervals of interest. Moreover, the “matter type” attribute had to be transformed from over 500 different matter types into 47 matter group codes in order to generate useful association rules. The transformed VLA attributes are: decided, matterCode and age.

- decided

This attribute identifies seven distinct time periods. The values of this attribute were derived from the actual decision date. For example decision date “12/02/2002” was transformed to value “decided\_firstQ2002” indicating that the decision was made in the first quarter of 2002. The period values were suggested by the VLA domain experts.

- matterCode

Due to the great number of offence matter types (over 500), we transformed “matter type” values to appropriate matter codes. The definition list of matter types and matter codes was supplied by the VLA experts. There are 47 distinct matter codes, with some of most frequently used codes such as FO (family matters - other), RD (drug and related offences), RX (theft and related offences) and R6 (breach offences).

- age

In order to apply KDD algorithms and generate association rules we transformed the attribute dob (date of birth) into eight age groups. The attribute *date of birth* is too fine grained for meaningful processing in that a rule that indicates if *date of birth = 12/05/1959* then *decision = refused NO* (with a support x%, confidence y%) is unlikely to be useful. To meaningfully mine such quantitative (numerical) variables the values had to be partitioned into intervals. Domain experts were again approached to define the intervals. Partitions identified as useful were: under 16, 16...18, 19...25, 26...30, 31...40, 41...50, 51...60 and over 60.

In the next section we discuss data preparation of the Diabetes data set. However due to the large number of attributes in this data set and their medical nature we are not discussing these attributes into fine details.

## 5.3 Diabetes Data preparation steps

### 5.3.1 Data selection

The integrated national diabetes program through Divisions of General Practice (DGP) commenced in May 2001 and by the May 2002, over 4300 patient records were collected. The records contain 57 variables representing many different patient characteristics. The patient records are classified into ten different categories such as: general, central circulation, peripheral circulation, treatment, results, referrals, renal, eyes, tests and sexual. In this study we included all 57 variables. Each of the ten categories contains on average six attributes. For example category “general” contains the attributes containing

personal details of a patient such as date of birth, gender, weight, height, diabetes type, year diagnosed and site ID. The “site ID” attribute identifies a GP division number from which the patient is coming from. The “results” and “tests” categories include attributes that contain laboratory results of the patient such as sugar level, cholesterol and blood cell counts. The other categories such as “renal”, “referrals” and “eyes” include attributes containing binary values (e.g. 1 indicating “YES” and 0 indicating “NO”). For example, “amput\_0” indicates that the patient is not an amputee, “refPodiat\_1” indicates that the patient was referred by a podiatrist and “new\_blindness\_1” indicates that the patient lost his/her sight. However in order to generate useful association rules, we divided the data set into ten tables (files) representing ten categories. Each table contains the attributes corresponding to their category (e.g. “impotence” attribute is stored into the “sexual” category) as well as additional attributes from the “general” category (e.g gender, height, weight and diabetes type).

### 5.3.2 Data pre-processing

The DGP data set was already used and analysed by their data analysts (the analysts used statistical methods), therefore it did not contain invalid or noisy records.

### 5.3.3 Data transformation

Many attributes in this data set were numeric (e.g HighuAlb = 14.6). In order to apply the KDD method of association rules we had to transform such quantitative attributes to intervals. Sydney based diabetes domain experts were involved in this phase and suggested interval values for all numeric attributes. For example the patients’ weight values were transformed into groups of ten (e.g 68kg transformed into weight group 60-70). We also transformed date based attributes such as dob into age groups. The transformation technique was identical as discussed in section 5.2.

### 5.3.4 Section summary

Data preparation for the analysis was a time consuming and complex process requiring high technical and domain knowledge. The domain experts made important decisions including attribute selection and interval definition. The technical part of this process was time consuming and required significant IT knowledge. For example, preprocessing and transformation of the records stored in the database required a knowledge of SQL (structured query language). Moreover, preprocessing and transformation of the records stored in a CSV (comma separated value) file required knowledge of many different Unix-Linux utilities such as awk, sed, tr, cut, paste, vi and tar. We spent many hours with these

utilites in order to successfully prepare data sets for the analysis.

While preprocessing data sets used in this study we identified that other researchers at the University of Ballarat experienced similar problems. Many university researchers experienced problems opening or editing large data files on the Win32 platform therefore we decided to use the Linux platform in order to handle such large data sets. Furthermore many researchers were not technically advanced or lacked tools for this task. In order to improve the current situation and speed up the process we initialised an open source project at the University that will allow researchers to prepare (select, pre-process and transform) their datasets with a minimal effort. The project is Linux based (Open Source) and includes a development of the wizard alike (point and click) application that requires minimal technical knowledge. The application is web based and allows researchers to upload a data file to the server, preprocess the data (e.g. select certain columns or rows, transform numerical attributes to bins and select rolumns or rows based on certian conditions) and save the preprocessed file. The author is reporting on a promising early stage open source project for pre-processing large datasets using a web based GUI.

In the following section we discuss the design and implementation of WebAssociate.

## 5.4 Design and Implementation

Many KDD developers design and implement software without being involved with the end users. Traditional life-cycle model of KDD software development (especially in the research domain) often assume very little interaction between design team and eventual users. Even if it does, the end users are assumed to have a high technical knowledge. This type of development tend to be machine orientated, paying scant attention to the algorithms used for the tasks to be performed.

Design and implementation of WebAssociate was an iterative and interactive process involving many demonstration and evaluation phases. Usability evaluation was not a last minute process performed before release of the system to give it a cosmetic gloss. Several testing sessions involved VLA domain experts to test the software. During the testing phase we introduced several different approaches that would assist the user in data analysis and observed users reactions. Observations and discussions with the end users made valuable feedback that provided the basis for further development. As discussed in Hall and Zeleznikow [34], many KBS are developed iteratively using new development methodologies or prototyping, without software requirement specifications. It is in the nature of the research based development to use prototyping. WebAssociate was developed using a similar methodology. Our methodology was based on the “bazaar” development

rather than generic “cathedral” development. This two development types were discussed by Raymond [70], and their main differences are:

- In the “bazaar” development users feedback provides the basis for further development, where the “cathedral” development assumes that the architecture of the software is known and the design and functionality are well understood before development begins.
- In the “bazaar” development users feedback is wanted early and as frequent as possible, while “cathedral” development provides external feedback only during alpha and beta testing.

In [70], Eric Raymond points out the importance of frequent software evaluation and claims that

*If you treat your software testers as if they're your most valuable resource, they will respond by becoming your most valuable resource.*

For example, we started the development of WebAssociate without predefined user specifications. During our presentations and demonstrations of WebAssociate we observed responses of the users and identified some areas which users found useful (e.g. graphical representation of the discovered association rules rather than text representation) and other areas which users had difficulties with (e.g. specifying the confidence and support thresholds). The user feedback from the each presentation provided bases for the further development. In the next section we identify steps and methods used for the design and development of WebAssociate.

## 5.5 Prototyping WebAssociate

In the initial stage of the WebAssociate development we used several methods suggested by other researchers that were possibly useful for non-technical domain experts in order to apply data analysis to their data set. The methods addressed the following issues:

- Visualisation  
Visualisation of the discovered patterns in data provide more information to the user than textual presentation.
- Group differences  
By grouping relevant association rules (e.g. sets of rules with common antecedent or common consequent), the user is enabled to identify differences between attributes (e.g. *Male*  $\Rightarrow$  *Drug related charges* and *Female*  $\Rightarrow$  *Drug related charges*).

- Hypothesis suggestion

Association Rules allow the user to identify associations between attributes in data and as such are useful for hypothesis suggestion or hypothesis testing.

### 5.5.1 Rare item problem

The initial testing of WebAssociate identified some issues. Domain experts had difficulties setting up confidence and support values. If the threshold was set too low the users were overwhelmed by too many generated rules. If the threshold was set too high, less frequent data-items were overseen (rare item problem). In order to overcome the “rare item” problem we decided to dismiss the measure of “support”. The literature shows that “support” is useful in the supermarket domain (basket analysis) where rare items are not much of importance to the user. However in the medical and legal domains the rare items are as important as more frequent items which makes little use of “support”. Furthermore we decided to introduce the “variable of interest” notation.

### 5.5.2 Variable of interest

A variable of interest selected by the user corresponds to the population under study and as such is represented as antecedent by default. For example selection of the *country\_ITALY* and *country\_GREECE* values, defaults to setting these values as antecedents (LHS). The corresponding AR are generated and grouped together (e.g. *country\_ITALY*  $\Rightarrow$  *Male* and *country\_GREECE*  $\Rightarrow$  *Male*). Each group of AR that have common consequent is called a “iso-consequent rule set”. By using this approach the user is guaranteed to discover the main characteristics of the selected attribute-values even if the population under study is very small (Italians make only 0.45% of the whole dataset). Note that backward rule *Male*  $\Rightarrow$  *country\_ITALY* would be considered uninteresting because Italians make a very small percentage amongst all males. In order to overcome the problem of setting the minimum “confidence” threshold we introduced the notation of “rule set confidence”.

### 5.5.3 Rule set confidence

The “rule set confidence” is the confidence threshold for a group or set of rules. It has the range of 0 to 100 (representing 0% to 100%). The default value is set to 10% as suggested by the experts. The “rule set confidence” filter has two possible boolean options: “AND” and “OR” option. Selection of the “AND” boolean results in inclusion of the rule set only if the confidence values of each association rule in the rule set are equal or above the “rule set confidence”. Selection of the “OR” boolean results to inclusion of the rule set is at least one confidence value of each association rule in the rule set is equal or above the “rule



set confidence”. For example, iso\_consequent rule set called “male” has two association rules R100 *country\_GREECE*  $\Rightarrow$  *Male* (confidence 77%) and R200 *country\_AUSTRALIA*  $\Rightarrow$  *Male* (confidence 47%) . If the user has set rule set confidence to 50% and selected the “OR” option, iso\_consequent rule set “male” would be included in the discovery. The “OR” option looks for at least one association rule in the rule set that has the confidence value higher than 50%. However, selection of “AND” option for the rule set confidence of 50% would not include iso\_consequent rule set “male” in the rule discovery because the confidence value for rule R200 is not above 50%.

In the further demonstration and evaluation of WebAssociate we observed the users and concluded that the users found grouping of discovered association rules very useful. Furthermore users hardly needed to readjust the default “rule set confidence” threshold. If users wanted to set the “rule set confidence” to a higher value (e.g. 25% instead of 10%) in order to refine the findings (include fewer rule sets), the users had no difficulties deciding what the threshold should be. However the users were still swamped by the great number of discovered association rules. In order to address this problem we decided to include further filtering of the discovered rules and introduced the notation of “similarity” with the default value of 5%. This option allows the user to define similarity in order to categorise discovered rule sets as similar or different. Rule sets with deviations between the maximum and minimum confidence values greater than user defined “similarity” are categorised as different where the sets with deviation equal or smaller than user defined “similarity” are categorised as similar. The user explores the discovered rules by selecting either categories or both. An example of a rule set in the similar group with a “similarity” value of 5% includes the following association rules:

- *country\_GREECE*  $\Rightarrow$  *refused\_YES* (confidence 25%) (R1)
- *country\_ITALY*  $\Rightarrow$  *refused\_YES* (confidence 27%) (R2)
- *country\_YUGOSLAVIA*  $\Rightarrow$  *refused\_YES* (confidence 24%) (R3)

In this example the rule set (rules with common consequent) is categorised as similar because the deviation between the maximum confidence (rule R2) and minimum confidence (rule R3) is less than or equal to 5%. The user may infer the null hypothesis: “*There is no difference in the refusal rate between Italian, Greek and Yugoslav born applicants*” because the variables of interest are similar on the bases of refusal rate.

#### 5.5.4 Further organisation of discovered AR

By observing domain experts using WebAssociate we also concluded that domain experts were initially more interested in exploring the association rule sets that contain single consequent (e.g. *country\_GREECE*  $\Rightarrow$  *refused\_YES*) and then further exploring the rule sets with multiple consequents (e.g. *country\_GREECE*  $\Rightarrow$  *refused\_YES AND age\_51..60*). In order to address this problem we decided to split the generated rule sets into two groups, the rule sets that have single consequent (level one rules) and rule sets that have multiple consequents. This approach allows the users to explore either groups in order to discover interesting and useful associations. Furthermore, we decided to introduce combo boxes that allow the users to select consequent attributes for the generated rule sets. For example level one rules combo box for the country variables of interest (e.g. *country\_GREECE*, *country\_ITALY* and *country\_YUGOSLAVIA*) allows the user to select any of the single consequent attributes (e.g. sex, law type, matter code, refused and age). By selecting one consequent attribute (e.g. sex) a rule set will be included if it contains attribute values (e.g. Male) that meet the user specified confidence threshold. Selection of a consequents with multiple attributes (e.g. sex and refused) from the multiple attribute combo box results to the rule sets that contain attribute values as consequents from both attributes (e.g. sex Male and refused YES) and meet the user specified confidence threshold. An example of such a set with “rule set confidence” of 15% and “OR” boolean (at least one confidence from the rule set above the threshold) is the set of following association rules:

- *country\_GREECE*  $\Rightarrow$  *refused\_YES AND sex\_MALE* (confidence 14%) (R4)
- *country\_ITALY*  $\Rightarrow$  *refused\_YES AND sex\_MALE*(confidence 17%) (R5)
- *country\_YUGOSLAVIA*  $\Rightarrow$  *refused\_YES AND sex\_MALE* (confidence 16%) (R6)

Note that if the user selected the “AND” boolean filter, this rule set would not be included because “AND” filter includes only the rule sets that have all confidence values greater than the “rule set confidence” of 15% (confidence for the rule R4 does not meet the requirements of this filter).

Due to limitations of the Christian Borgel’s apriori algorithm used in this study WebAssociate AR graphs are limited to maximum of five items per rule. We did not further explore possibilities of AR with more than five items because research shows that users find difficulties in understanding ARs that contain more than five items in their antecedent or consequent. However, we believe that our visualisation methods would allow user to understand AR with more than five items in the consequent.

### 5.5.5 Chi-square test

In the subsequent evaluations of WebAssociate we decided to enable users to use the statistical test of significance for any hypothesis that suggest differences between two groups. Some suggested differences in confidence values between two ARs may not be statistically different even that visual representation suggest that the difference is significant. For example, discovered association rules *country\_ITALY*  $\Rightarrow$  *refused\_YES* (confidence 26.7%) and *country\_AUSTRALIA*  $\Rightarrow$  *refused\_YES* (confidence 10.7%) suggested to the user the alternate hypothesis H1 “*More Italian born applicants are refused legal aid than Australian born applicants*”. In order to test this hypothesis and confirm statistical difference between those two groups we use chi-square test. WebAssociate generates a contingency table containing frequencies and calculates the chi-square value as illustrated in Figure 4.11. The null hypothesis “*There is no difference in the refusal rate between the Australian and Italian born applicants*” was rejected at the 0.05 level of significance because the chi-square value of 58.09 exceeded 3.841 with *degree of freedom 1* as illustrated in Figure 4.11. By applying the chi square test of significance in WebAssociate the user was able to test the hypothesis H1 and find out whether two groups are statistically different or not.

### 5.5.6 WebAssociate -Further Improvements

By this development stage the users were quite confident using WebAssociate. The non-domain experts experienced the satisfaction of being enabled to perform data analysis without being dependent on the IT department. Furthermore, the experts found that generated rule sets were useful in suggesting a hypothesis with additional features such as filtering and chi-square test of significance. However the experts agreed that more sophisticated data analysis were still needed to be done by the professional data analysts from the VLA IT department. The professional analysts were also thrilled by not having to assist the non-technical experts in their simpler analysis requirements. The users found our approach of “variable of interest” defaulting to antecedents easier to use than the generic two dimensional “matrix” approach (also used by SGI MineSet), where selected variables of interest appear as antecedents as well as consequents. The users also found our approach of using “rule set confidence” and boolean (“AND” and “OR”) filters more understandable and useful than generic approach of setting the confidence and support thresholds.

Nevertheless, we still found several areas of WebAssociate that needed further improvements. The users still had difficulties knowing which “variables of interest” were useful

for the exploration. For example, some variables of interests appeared in the data set so rarely that were not valuable to be explored (e.g. there were just two VLA applicants that were born in country *ANGOLA*). Selecting country *ANGOLA* as a variable of interest would be of little use to the expert. In order to overcome this problem we introduced additional options that allow the users to view bar graphs representing percentage (e.g. country *ANGOLA* makes only 0.0047% of the dataset) or absolute count values for the selected variables of interest. Prior to the association rule generation the users were able to use this additional options and decide which selected variables of interest are of little or no use. Furthermore, by observing experts using WebAssociate we found that experts sometime preferred to have actual values rather than percentage values for the rule confidence representation. We decided to enable the users to view confidence values as percentage (e.g. 50%) or as absolute value (e.g. 130 of 260).

Further observations of the non-technical experts using WebAssociate uncovered additional needs of the experts. Even though experts found hypothesis suggestion and testing methods very useful for the analysis requirements of their data set, additional methods that follow their reasoning were needed. For example, the discovered association rules that suggested the alternate hypothesis H1 *“More Italian born applicants are refused legal aid than Australian born applicants”* were useful but the users needed additional methods in order to explain the suggested hypothesis (e.g. explain refusal rate differences between the two country groups). The experts were able to give some explanations for the causes for the higher refusal rate of the Italian born applicants by being experienced in the field and using their domain knowledge. However, we identified the need for further data analysis that would support their explanations and distinguished several steps that followed their reasoning.

## 1. SUGGEST HYPOTHESIS

Identify the association rules that suggested the hypothesis

- a)  $A_1 \Rightarrow B_1$  (conf.%)
- b)  $A_2 \Rightarrow B_1$  (conf.%) Example:

- a) country\_ITALY  $\Rightarrow$  refused\_YES (conf.26.7%)
- b) country\_AUSTRALIA  $\Rightarrow$  refused\_YES (conf.10.7%)

where difference between confidences exceeds a threshold.

## 2. IDENTIFY LIFT

A search is made for an attribute that, when inserted into the antecedent of the rule

- a), increases the confidence of the rule.

c)  $A_1 \text{ AND } X \Rightarrow B_1$  (conf.%) Example:

$c_1$ ) country\_ITALY AND age\_51..60  $\Rightarrow$  refused\_YES (conf 45.45%)

$c_2$ ) country\_ITALY AND lawType\_FAMILY  $\Rightarrow$  refused\_YES (conf 35.5%)

We automatically identify additional attribute-values for country\_ITALY which contribute to a refusal rate higher than 26.7

### 3. QUALIFY LIFT

For the identified attribute-values we find the strongest contributors.

d)  $X \Rightarrow B_1$  (conf.%) Example:

Find the strongest contributors for the identified attribute *age*.

age\_under16  $\Rightarrow$  refused\_YES (conf 00.74%)

age\_16..18  $\Rightarrow$  refused\_YES (conf 04.24%)

age\_19..26  $\Rightarrow$  refused\_YES (conf 08.89%)

age\_26..30  $\Rightarrow$  refused\_YES (conf 09.83%)

age\_31..40  $\Rightarrow$  refused\_YES (conf 13.40%)

age\_41..50  $\Rightarrow$  refused\_YES (conf 17.75%)

age\_51..60  $\Rightarrow$  refused\_YES (conf 26.60%)

age\_over60  $\Rightarrow$  refused\_YES (conf 21.40%)

We identified that age group *age\_51..60* is the strongest contributor amongst all age groups for *refused\_YES*. This step is important for the user because we want to check if lifted attribute-values are also identified as strong contributors.

For example, WebAssociate found that lawType\_FAMILY, age\_51..60, age\_41..50 and matterCode\_FO were qualified attribute-values and strong contributors for greater refusal rate amongst the Italians. Consequently we generate association rules that contain “country\_AUSTRALIA AND X” as antecedent, where “X” is qualified attribute-value, and “refused\_YES” as consequent. Finally, WebAssociate calculates deviations between corresponding confidence values for each rule.

Example:

$e_1$ ) country\_ITALY AND age\_51..60  $\Rightarrow$  refused\_YES (conf 45.45%)

$e_2$ ) country\_AUSTRALIA AND age\_51..60  $\Rightarrow$  refused\_YES (conf 26.1%)

$f_1$ ) country\_ITALY AND lawType\_FAMILY  $\Rightarrow$  refused\_YES (conf 35.5%)

$f_2$ ) country\_AUSTRALIA AND lawType\_FAMILY  $\Rightarrow$  refused\_YES (conf 15.0%)

### 4. EXPLORE GROUP CHARACTERISTICS

Explore the characteristics of both groups under study - find their similarities and differences for each contributor

This step involves identifying characteristics of the group with higher refusal rate (Italians) which had higher confidence value than the characteristics of the group with lower refusal value (Australians). For example the confidence of 35% for the association rule *country\_ITALY*  $\Rightarrow$  *lawType\_FAMILY* compared with the confidence of 26% for the rule *country\_AUSTRALIA*  $\Rightarrow$  *lawType\_FAMILY* identifies to the user that the greater proportion of Italian applicants apply for aid for Family matters. Incorporated with the previous suggestions 1) great number of Italian applicants have been refused for Family matters; 2) Family matters get refused more than any other matters, we have a logical approach in trying to explain a suggested hypothesis.

By using our method the experts were able to make conclusion that the Italian born applicants were more refused aid than Australian born applicants for the following reasons:

- a) The greatest contributor for higher refusal rate among Italian applicants is age group 51 to 60 (45.5%). 17% of Italian applicants are in this age group compared to 2.1% of Australian applicants.
- b) Another contributor for high refusal rate among Italian applicants is matter type FO (35.5%). 31% of Italian applicants applied for this matter code compared to 24% of Australians.
- c) The last contributor for higher refusal rate among Italian applicants is Family law type. 37.1% of Italian applicants applied for aid for Family matters compared to 28% of Australian applicants.

### 5.5.7 Hypothesis - possible explanations

The characteristics of Italian and Australian applicants showed that most of the Italian applicants were older, male, and applied for family law matters, while most of the Australian applicants were younger applicants and applied for the criminal matters. The characteristics also showed that the Italian applicants were mostly refused aid on the basis of guidelines (10%), means (4%) and guidelines and means (6%) while only 5% of the Australian applicants were refused aid on the basis of the guidelines, and very small percentage on the other bases.

The characteristics of refused Italian and refused Australian applicants showed that most of the refused Italian applicants were older, male, and applied for family law matters, while the most of the Australian applicants were younger applicants and applied for criminal matters.

## 5.5.8 Three Sections of WebAssociate

In the last evaluation of WebAssociate we identified the need for three different approaches in order to enable the experts to analyse their data set (as discussed in Chapter 4). In order to cater for these needs we split WebAssociate into three separate sections:

Section 1:

Domain experts that do not know what are they trying to discover use the *unkown hypothesis suggestion* approach.

Section 2:

Domain experts that know what are their variables of interest (antecedent) but don't know what associations they want to discover use the *partial hypothesis suggestion* approach.

Section 3:

Domain experts that know what associations they want to discover use the *hypothesis testing* approach.

Domain experts found all three approaches useful for their discovery tasks. Moreover, the experts favoured ability to independently mix and match approaches within the discovery process.

In the next section we discuss WebAssociate algorithms and methods used for each discovery approach.

## 5.6 Discovery Methods of WebAssociate

### Development Tools

WebAssociate is a web based tool running on the Linux platform. All the development tools used in this application are mostly Open Source tools released under GPL (General Public Licence). We used the following development tools:

- Linux Distribution - RedHat 8.0  
(<http://redhat.com>)
- Web Server - Apache  
(<http://apache.org>)
- Server Side Development - PHP4  
(<http://php.net>)
- Client Side Development - Java Script
- Data Base - MySQL  
(<http://mysql.com>)

- Data sorting, searching and editing - Linux Utilities  
(e.g. awk, sed, grep, tr, cut, tar and vi)
- Graphic presentations - PHP JpGraph Library  
(<http://www.aditus.nu/jpgraph/>)
- AR generator - Christian Borgelt's Apriori implementation  
(<http://www.fuzzy.cs.uni-magdeburg.de/borgelt/>)

### Data Sets

In this study we used four data sets containing three real life data sets and one fictional set. The real life data sets include two VLA data sets (380,000 records collected in 1999-2001 and 42,434 records collected in 2002) and the Diabetes Australia data set (4,359 records collected in 2002). The fictional data set is a small set representing tourist records in Ballarat area. Purpose of using several different data sets is to test if WebAssociate is domain independent (e.g. not just suitable for the legal domain). The data sets are stored in two formats: relational data base (RDB) format and comma separated value format (CVS). The CVS format is used by Christian Borgelt's Apriori implementation, while RDB format is used for data querying purposes (using structured query language - SQL).

### WebAssociate Sections

As discussed in Section 5.4 WebAssociate is divided into three separate sections:

Section 1:

Domain experts that don't know what are they trying to discover use the *unkown hypothesis suggestion* approach.

Section 2:

Domain experts that know what are their variables of interest (antecedent) but don't know what associations they want to discover use the *partial hypothesis suggestion* approach.

Section 3:

Domain experts that know what associations they want to discover use the *hypothesis testing* approach.

#### Section 1

After users selection of a database and table of interest , WebAssociate generates association rules by querying the selected RDB data set using SQL. For each attribute  $A$  in the selected data set a file is generated containing vectors for each attribute-value  $A.V$ . Each vector contains confidence values for single association rule represented as  $A.V \Rightarrow \neg A.V$ . For example for the set of attributes *sex*, *lawType* and *refused*, we have the following attribute values *sex* (*male and female*), *lawType* (*criminal, civil and family*) and *refused* (*yes*



and no). File *sex* is generated with the two vectors, vector *male* and vector *female*. The *male* vector contains confidence values for the following association rule sets: *sex\_male*  $\Rightarrow$  *lawType\_family*, *sex\_male*  $\Rightarrow$  *lawType\_criminal*, *sex\_male*  $\Rightarrow$  *lawType\_civil*, *sex\_male*  $\Rightarrow$  *refused\_yes* and *sex\_male*  $\Rightarrow$  *refused\_no*. The *female* vector contains confidence values for the same consequents as rules for the *male* vector, but the antecedent is *female*. In Table 5.1 the *male* and *female* vectors are illustrated as rows two and three and each rule set is illustrated as a column.

	lawType criminal	lawType family	lawType civil	refused yes	refused no
male	75.08	17.41	7.51	9.79	90.21
female	31.39	52.25	16.36	12.29	87.71

Table 5.1: *sex* file example

After successful file generation for each attribute, WebAssociate graphically represents the contents of each file.

## Section 2

After the selection of a database, table and attribute of interest, WebAssociate prompts the user to select a variable of interest from a list. The list of variables of interest corresponds to distinct values of the selected attribute. For example, attribute *sex* contains the list of three distinct values: male, female and NA (gender not specified). The user has option to select one or more variables of interest.

**AR generation** Subsequently, WebAssociate activates the Christian Borgelt's Apriori implementation with variables of interest appearing as antecedents (left hand side). We are not discussing Apriori algorithm in this work because we could have used any other algorithm that generates multiple association rules. At this stage we use the brute force Apriori where minimum confidence and support threshold is not specified (zero) and all possible rules are generated. The output of all association rules generated by Apriori is redirected to a temporary file. The file contains a set of all possible association rules containing the variables of interest as antecedents. Table 5.2 shows an example of unsorted association rules generated by Apriori:

## Rule Matching

In the next step we use our sorting algorithm which matches the generated association rules. The purpose of this algorithm is to find all rules that have common consequents. Each set of matched rules is called a "rule set". Table 5.3 shows an example of sorted

antecedent	imply	consequent	confidence
sex_female	$\Rightarrow$	X and Y	(conf %)
sex_male	$\Rightarrow$	Y	(conf %)
sex_male	$\Rightarrow$	Y and Z	(conf %)
sex_male	$\Rightarrow$	X and Y	(conf %)
sex_female	$\Rightarrow$	Y and Z	(conf %)
sex_female	$\Rightarrow$	Y	(conf %)

Table 5.2: Unsorted association rules

association rules:

antecedent	imply	consequent	confidence
sex_male	$\Rightarrow$	Y	(conf %)
sex_female	$\Rightarrow$	Y	(conf %)
sex_male	$\Rightarrow$	X and Y	(conf %)
sex_female	$\Rightarrow$	X and Y	(conf %)
sex_male	$\Rightarrow$	Y and Z	(conf %)
sex_female	$\Rightarrow$	Y and Z	(conf %)

Table 5.3: Sorted association rules

#### Pruning and categorising discovered AR

In the next step we prompt the user to optionally set the filters that prune generated association rules. The pruning algorithm includes the rule sets that meet the user specified settings. There are three inputs for this algorithm: a "rule set" confidence (a value from 0 to 100), a boolean flag (AND or OR) and similarity value (0-100). Depending on the boolean flag, the rule sets that that meet the minimum confidence threshold are included in the output. For example, if OR boolean flag was selected a rule set that contains at least one association rule with confidence higher or equal than the user specified minimum threshold will be included. For the AND boolean filter, a rule set will be included only if all associations rules in the set have confidence higher or equal than the user specified minimum threshold.

The similarity function is used to categorise discovered rule sets into "similar" and "different". This function finds a minimum and maximum confidence value for a given rule set. If the deviation between minimum and maximum values is equal or smaller than the user specified "similarity value", the rule set is categorised as similar otherwise the rule

set is categorised as "different". Furthermore, the rule sets are categorised as "level one" and "multiple level" AR. Level one category contains the rule sets with a single item AR (e.g.  $X \Rightarrow Y$ ), where multiple level category contains the rule sets with multiple item AR (e.g.  $X \Rightarrow Y \text{ and } Z$ ). The reason for the further categorisation of discovered rule sets is because the literature shows that users often start the exploration of discovered patterns from the simplest (single item AR) to more complex formats.

#### Presentation of the discovered rule sets

The user has several options to view the discovered rule sets that met the user specified filters. The rule sets are displayed graphically and the user is able to view graphs that contain similar, different, level one or multiple level rule sets. Further more the user is able to select rule sets of his interest. For example, if the user selected *sex\_male* and *sex\_female* attribute-values as variables of interest, the user may be interested to select only rule sets containing refused and lawType attribute-values as consequent. Table 5.4 illustrates an example of the rule sets containing association rules with refused and lawType attribute-values as consequents:

antecedent	imply	consequent	confidence
sex_male	$\Rightarrow$	lawType_Family AND refused_YES	(conf %)
sex_female	$\Rightarrow$	lawType_Family AND refused_YES	(conf %)
sex_male	$\Rightarrow$	lawType_Criminal AND refused_YES	(conf %)
sex_female	$\Rightarrow$	lawType_Criminal AND refused_YES	(conf %)
sex_male	$\Rightarrow$	lawType_Civil AND refused_YES	(conf %)
sex_female	$\Rightarrow$	lawType_Civil AND refused_YES	(conf %)
sex_male	$\Rightarrow$	lawType_Family AND refused_NO	(conf %)
sex_female	$\Rightarrow$	lawType_Family AND refused_NO	(conf %)
sex_male	$\Rightarrow$	lawType_Criminal AND refused_NO	(conf %)
sex_female	$\Rightarrow$	lawType_Criminal AND refused_NO	(conf %)
sex_male	$\Rightarrow$	lawType_Civil AND refused_NO	(conf %)
sex_female	$\Rightarrow$	lawType_Civil AND refused_NO	(conf %)

Table 5.4: User specified rule sets

Each rule set is defined as the set of association rules that have the same consequent. WebAssociate generates a graphical representation of the rule sets illustrated in Table 5.4. Subsequently, WebAssociate enables the user to select a single association rule and further drill down in order to find interesting patterns. For example, selection of the first

association rule in Table 5.4, would allow the user to explore all possible association rules that have additional attribute-values in the consequent (e.g. *sex\_male*  $\Rightarrow$  *lawType\_Family AND refused\_YES AND age\_21..30 (conf %)*)

### Hypothesis Testing

If the user elected to view the graph containing different rule sets (rule sets that have a deviation between its AR confidences greater than the user defined “similarity”), the user is able to test the suggested hypothesis. WebAssociate uses several methods for this task. The initial method uses chi-square test of significance by calculating the frequencies for the selected association rules as discussed in details in Chapter 4. The frequencies are obtained by using SQL queries. Additional methods that provide a set of association rules for further exploration of the selected hypothesis also use SQL query approach. At this point the filters and constraints that were specified by the user are not implemented.

### Section 2 summary

In this study we used the efficiency of the Apriori algorithm in Section 2 only for initial association rule generation. Any further explorations of the discovered rule sets generate association rules by querying the database with SQL queries.

### Section 3

Section three of WebAssociate is used by the users who know what is their discovery target. In this section we use SQL approach in order to get frequency counts. The association rules are generated by calculating the conditional probability of the user selected attribute-values for the given conditions. Additional functionality of this section allows the user to define new groups (variables of interest). This option is useful when users are interested in finding how new defined groups differ. For example all age groups specified as attribute-values in the data set (e.g. age under 16, 16..20, 21..30, 31..40, 41..50, 50..60 and over 60) could be redefined by the user into new groups such as young (e.g. by selecting age groups under 31 years old), middle age (e.g. by selecting age groups between 30 and 50 years old) and older (e.g. by selecting age groups over 50). The conditional probability values for the newly defined groups are in fact calculated by using SQL queries including each selected group in the SQL “WHERE” statement.

In this Chapter we discussed development and implementation methods of Web Associate. However due to the complexity of the software, we were unable to describe all algorithms of WebAssociate. WebAssociate contains over 30 code files and over 7000 lines of code. The software is available for use over the web at: <http://141.132.69.25/WebAssociate/>.

The password required for the access of the WebAssociate web site can be obtained by email (s.ivkovic@ballarat.edu.au). Please note that due to the University of Ballarat regulations this URL is subject to change. If URL is inactive contact the author by email provided above. In the next Chapter we describe evaluation methods used for testing usability and functionality of the WebAssociate, Gnome Data Miner and MineSet.

# Chapter 6

## Evaluation of WebAssociate

### 6.1 Tools for the comparison

In this study we were looking for AR software that was freely available and simple to use. It was important that at least one tool displayed AR visually so that the grouping aspect of the WebAssociate visual display could be assessed. It was also important that at least one AR tool displays AR textually so visualisation aspect of displayed AR could be assessed. In the early stage we considered open source KDD tools such as Weka (<http://www.cs.waikato.ac.nz/ml/weka/>) and MLC++ Machine Learning Library (<http://www.cald.cs.cmu.edu/software.html>). However, further testing indicated that those tools were too complicated for the non-technical domain experts that participated in our study. Finally we decided to use MineSet to assess the grouping aspect of visually displayed AR and use Gnome Data Miner (GDM) to assess the visualisation aspect.

#### 6.1.1 MineSet

MineSet is a data mining platform and development environment, which provides the user with the tools to perform data access, data transformation and analysis. MineSet suite of tools lets you analyze, mine, and graphically display data so that you can visualize, explore, and understand your data. We used SGI MineSet in this study because it was freely available at the University of Ballarat. Furthermore, MineSet 3D graphical display of generated AR is easily understandable by a novice user.

### 6.1.2 Gnome Data Miner

Gnome Data Miner Apriori (<http://www.act.cmis.csiro.au/edm/resources/gdmapriori.html>) is part of the CSIRO Enterprise Data Mining Tools (EDMtools) toolkit developed for GNU/Linux. GDM Apriori is used for building association rules from transaction data. The package includes both the Gnome GUI and the apriori command line from Christian Borgel (<http://fuzzy.cs.uni-magdeburg.de/borgelt/software.html>). Gnome Data Miner generates association rules according to user specified threshold (confidence and support) and displays generated AR textually. We decided to use GDM in order to evaluate the visualisation aspect. We believe that visualisation of discovered AR is essential because non-technical users could have great difficulties understanding textual display of generated AR.

### 6.1.3 WebAssociate

WebAssociate is a web based tool designed to allow non-technical users to search for interesting patterns in data. It is designed for exploratory tasks, without using a heavy-duty statistical approaches. The interestingness of discovered patterns in most AR tools is based on the frequency values such as confidence and support. In WebAssociate the interestingness is based on the content. By grouping AR that share a common consequent (single or multi-item) but different antecedent (distinct values of an attribute) we are able to visually observe differences in confidence value between two or more AR belonging to the same rule set regardless to their confidence value. WebAssociate should not be used as AR tool for decision making or prediction. It rather suggests interesting patterns that should be further confirmed by using more significant analysis (e.g. statistical tests of significance).

## 6.2 Ethics Approval

Prior to the software evaluation, ethics approval was sought from the University of Ballarat ethics committee. Full approval for this project was granted by the Human Ethics Committee at the University of Ballarat on the April 10, 2003. In this section we discuss human issues addressed in the ethics application.

### 6.2.1 Protection of Participants

- Prior to the evaluation process, 30 minutes training will be provided to familiarise participants with each of software tools used in this study. This should alleviate

anxiety about the novelty of each task and provide more uniform starting knowledge for each of the participants.

- These is a very low risk of emotional or physical harm greater than or additional to risks encountered in the participant's normal lifestyle.
- Participants in this study (VLA domain experts) do use in-house software on a regular basis to analyse their data set and this task is part of the participant's daily job.
- In case of participants being distressed during the testing, VLA emergency procedures will be used and testing will be ceased or postponed.
- An additional step would involve calling emergency services (phone number: 000) or lifeline (phone number: 131114) if required.
- Participant names will not be used in this study, and there is no need to collect any private information of the participants.

## 6.2.2 Confidentiality

- There is no collection or use of personal data from the participants involved.
- The question and answers in this study will not involve any other information than information of the usability and effectiveness of the tested software.
- In case that a participant answers a question in the form that divulges any of the participant's information (e.g. "I have been working for VLA for ten years and I have been using similar tools..."), such information ("...working for VLA for ten years") would be discarded.
- To take precautions, it will be stated in the informed consent form taht no personal information should be included in answers.

## 6.2.3 Informed Consent

### Explanation of project given to participants

- The purpose of this study is to evaluate a visualisation tool *WebAssociate* developed by Sasha Ivkovic. The evaluation process will involve using WebAssociate and other data mining tools (MineSet and Gnome Data Miner) in order to test given hypothesis (e.g. "Is there a certian age group of applicants that gets more rejected than other age groups?").



- Appropriate 30 minutes training will be conducted prior to the evaluation task.
- In any case of discomfort and possible hazards involved, the study will be ceased or postponed and appropriate services contacted (e.g. emergency and GP).
- By evaluating the software there are potential benefits for the domain experts and VLA in using visual software for data analysis.
- A participant is free to withdraw consent and to discontinue participation in the study at any time.

### **Usability**

- You will use each tool and fill a simple questionnaire at the end of the evaluation process.
- Questions will ask for your opinion on the usability.
- Each of five tasks will take 15 minutes for each hypothesis test with 10 minutes break between the tasks.
- Answers to the questionnaire should not include any personal information that may identify you (e.g. your age, name and position).

### **Effectiveness**

- Each participant will write a short report on the outcomes of the investigation of each hypothesis with each software tool.

## **6.3 Software Evaluation Methods**

According to ISO/IEC 9126 : Information technology - Software Product Evaluation Standard (<http://www.cse.dcu.ie/essiscope/sm2/9126ref.html>), there are six quality characteristics of a software artifact:

1. Functionality; Checks for the validation of software e.g. are the required functions available in the software?
2. Portability; Is the software platform dependent and how easy is it to transfer the software to another environment?
3. Reliability; Is the software stable?
4. Maintainability; How readily can the software be modified ?
5. Efficiency; How efficient is the software?

## 6. Usability; Is the software easy to use?

Many KDD tools would have been evaluated for all six characteristics, however the usability and functionality characteristics are to some degree subjective. If the end user is a sufficiently technical domain expert, the KDD software could be rated highly, where a non-technical domain expert is likely to rate the software that requires technical knowledge poorly. WebAssociate, SGI MineSet and Gnome Data Miner (CSIRO GUI for Apriori) have been evaluated for the subjective characteristics such as functionality and usability. The evaluation process used in this study was partly based on the *Context, Criteria and Contingency* (C.C.C.) framework reported in [34], especially on the user credibility and validation. Hall and Zeleznikow [34] explore how conventional and general Knowledge-Based Systems (KBS) are evaluated and the C.C.C. evaluation framework is introduced.

As discussed in [34], many KBS are developed iteratively using new development methodologies or prototyping, without software requirement specifications. It is in the nature of the research based development to use prototyping. WebAssociate has been developed iteratively. Several testing sessions involved VLA domain experts in testing the software. This valuable feedback provided the basis for further development. WebAssociate has been built for the use of non-technical domain experts (not just VLA), and the software evaluation process aimed to explore the hypothesis that *"Domain experts find WebAssociate easier to use and more useful than other KDD tools for their individual everyday requirements"*. The evaluation process used in this study was limited due to the inability to include more KDD tools. Most commercial KDD tools available for evaluation (available for download) were not suitable due to their complexity and additional requirements (e.g. additional client-server installations). In the next section we describe the software evaluation process conducted in this study.

## 6.4 WebAssociate Evaluation

As discussed in [48] one of principles in software evaluation is to outsource the evaluation task to those who might be considered the best to carry them out, typically those who are most familiar with the domain i.e. the end users. Bay and Pazzani [21] also claim that it is important to use domain experts for evaluation, because they represent the intended users of the software and will have a distinct purpose in mind when evaluating the software. Due to the inability to additionally involve Sydney based Diabetes domain experts in the evaluation process, only VLA domain experts were used. The limitation is that domain experts are inherently rare [21]. Thus we were only able to get responses from five VLA experts.

The software evaluation was conducted in the VLA Melbourne office with three separate sessions. Five VLA domain experts were involved, with three domain experts having technical background and two “pure” domain experts who are non-technical. Organisational role of the technical domain experts involved in this evaluation, is to analyse the VLA data set and provide reports to their peers. This task is carried out by using an in-house query system (Bi Query by Hummingbird Inc.). Organisational role of the non-technical domain experts involved in this evaluation, is to manage a branch of VLA and maintain legal issues (lawyer in-charge duties). It is also a duty of these non-technical experts to create business reports which include some simple analysis of the legal aid applications. However due to the complexity of the query system currently used, the non-technical experts are not able to complete this task without the assistance of in-house analysts, which makes them to be fully dependent on the technical experts.

Non-technical domain experts did not use Gnome Data Miner (association rules represented as text) because it was impossible for them to check the hypothesis without visual aids. One hour training was provided prior to each evaluation session. The software usability evaluation process included three hypothesis questions that the participants tried to answer by using each of the software tools and a simple questionnaire on the usability of each tool. Software effectiveness evaluation process involved each participant to write a short report on the outcomes of the investigation of each hypothesis. The hypotheses used in the evaluation process were representing potential everyday requirements of the domain experts.

The names of domain experts were not used in this study, therefore we used uppercase letters A,B,C,D and E instead.

### 6.4.1 User Credibility

According to [34] the user credibility is Micro/People (end-user) oriented and is subdivided into three main areas:

1. User Satisfaction
2. Usability (ease of use)
3. Utility (usefulness or fitness for purpose)

#### Usability

In order to test the usability of the software, Question 1 “Overall, was this tool easy to use?” was included. The possible answers to question 1 were *very easy, moderately easy, easy, moderately hard, hard and very hard*. For the purpose of graphically representing

the participants responses, we valued the answers from 1 (corresponding to very hard) to 6 (corresponding to very easy) respectively. Figure 6.1 shows a bar graph representation of the results for the question 1.

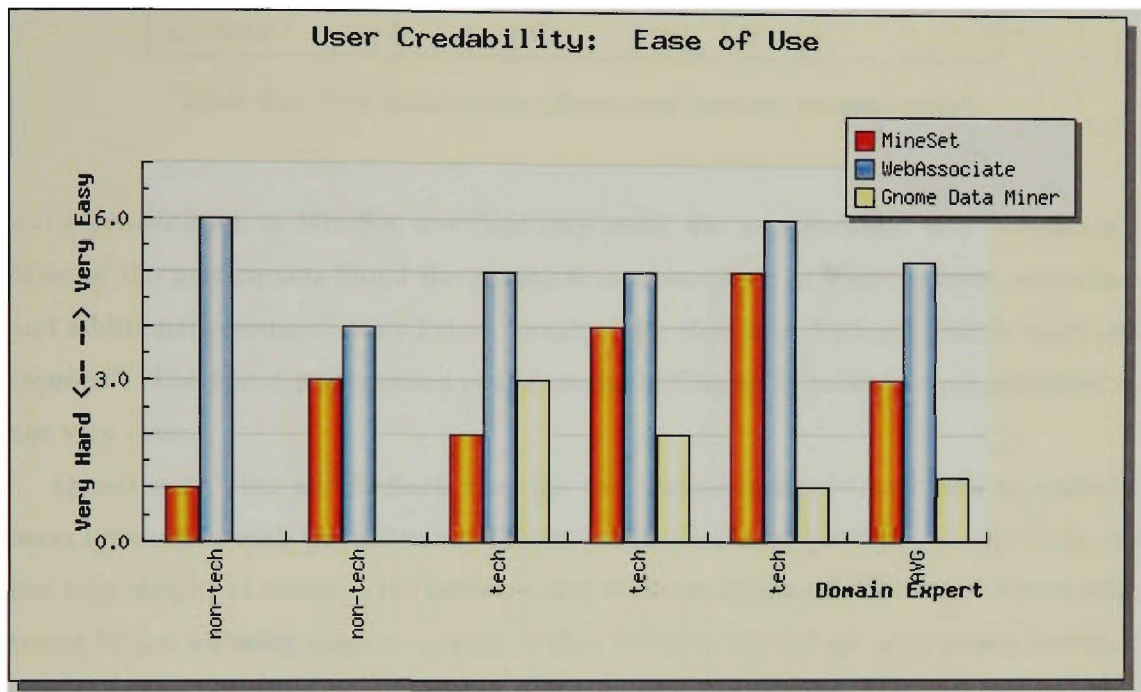


Figure 6.1: Software Ease of Use

As Figure 6.1 shows, both non-technical and technical domain experts found WebAssociate moderately to very easy to use. MineSet was evaluated by the non-technical experts as very hard to moderately hard to use, where the technical experts described it as hard to moderately easy. Gnome Data Miner was evaluated as very hard to moderately hard to use, due to the participants struggle with non-visual pattern discovery representation.

### User satisfaction

The purpose of Question 2 "Were intermediate results from the tool clear and understandable?" was to evaluate the user satisfaction. Four possible responses included *very*, *not very*, *to some degree* and *not sure* options. Intermediate results correspond to the representation of the discovered patterns. MineSet and WebAssociate represent the patterns graphically, where Gnome Data Miner represents discovered patterns as text. Many KDD tools have state of art graphical representation of discovered patterns but users have difficulties understanding them. The responses for question 2 are shown in Table 6.1.

Additional comments by some technical experts identified that matrix based graph-

	MineSet	WebAssociate	Gnome Data Miner
very clear	0	4	0
to some degree	3	1	0
not very clear	2	0	3
not sure	0	0	0

Table 6.1: User satisfaction - discovered pattern representation

ical representation in MineSet was *"not very clear, not user friendly"* and *"confusing"*. Most of the participants found the graphical representation in WebAssociate, very clear and additional comments stated that *"graphs were clear and displayed exactly what was required"*. The text representation of the results in Gnome Data Miner was described as not very clear.

Question 3 *"Was the feedback from the tool understandable?"* was used to evaluate users interaction with the software. The interaction included prompts, on line hints, on line help, graphical layout of the software, and feedback in general. The experts were frustrated by not knowing what to do next, if they felt they did not get appropriate feedback from the software. Four possible responses for question 3 included *very, not very, to some degree and not sure* options.

	MineSet	WebAssociate	Gnome Data Miner
very understandable	1	5	0
to some degree	1	0	1
not very understandable	2	0	2
not sure	1	0	0

Table 6.2: User satisfaction - feedback from the tool

All participants described the feedback from WebAssociate as very understandable. The feedback from MineSet was described by one of the non-technical domain experts as not very understandable, where the other non-technical expert was not sure. Technical domain experts described the feedback from MineSet not very understandable to very understandable. Two experts did not find the feedback from Gnome Data Miner very understandable, where one expert described the feedback as understandable to some degree.

## Utility

The last question from the evaluation questionnaire was used to evaluate the usefulness of the software. The question *"What features were particularly useful in this tool?"* aimed to allow the participants to give a short description of usefulness corresponding to their individual experiences.

### Mineset

1. expert A - non technical

*"None"*

2. expert B - non technical

*"The 3D graphics are nice, but problem with viewing. Nothing else much usable."*

3. expert C - technical

*"You would really need a technical background to use this package. Not very user friendly at all."*

4. expert D - technical

*"Lack of familiarity of this package was an issue, as the graphical layout looked simple and very effective. However, manipulation of the graphical results was difficult. Obtaining text output quickly problematic, since it involved highlighting every output item. Neat graphs but clumsy extraction."*

5. expert E - technical

*"I don't think this was a particularly useful tool. Assumptions are made that you know all symbols e.g. when writing expressions. Unless taught I don't believe the 'expert' would find this easy."*

### Gnome Data Miner

1. expert C - technical

*"Too complicated. Cannot read results properly."*

2. expert D - technical

*"The layout of the module was fairly basic, therefore it can be picked up quickly. However the tool is not very useful in extracting only relevant results. Too much redundant data provided in result."*

3. expert E - technical

*"Very time consuming application."*

1. expert A - non technical

*"I was able to group things together and also able to identify several different scenarios for one query. I was also able to break info down to microlevel. Very user friendly."*

2. expert B - non technical

*"All of the WebAssociate was extremely easy to use, understand and than explicate the information. The charts were extremely useful."*

3. expert C - technical

*"I really found it great that I could easily drill down in reports easily enough, especially at the click of a button. I really liked that I was able to group things - i.e. define groups."*

4. expert D - technical

*"The graphical representation was clear, useful and precise values were able to be extracted easily using only a few additional steps. Because the tool was able to generate more detailed output, it offered more options which took a while to get accustomed to."*

5. expert E - technical

*"Much more user friendly. Able to visualise data and understand easier. With further training would be a useful tool."*

### 6.4.2 Software Validation

According to [34] Validation and Verification is Micro/Technical oriented and concerned with validity, software design, knowledge representation, inferencing, learning and provision of explanations. Software verification has not been carried out due to the nature of the research, that the development was based on prototyping and used "bazaar" approach [70].

In order to test the software validation "*Validation: Are we building the right product?*", the evaluation process included three "real life" tasks of the domain experts. The validation process aimed to test if the software is correct and meets job requirements of the participants.

During the development phase of WebAssociate, many VLA employees suggested a real life task examples which could be tested with this software. All tasks used in this evaluation study derive from these suggestions.

Three “real life” tasks were presented to the participants with requirements for a short report after completing each task with each software. Non-technical domain experts were not required to use Gnome Data Miner due to its complexity. Gnome Data Miner does not visualise discovered patterns, and non-technical experts find manual text search too difficult.

### Task 1

**Description:** WebAssociate has suggested that there is a difference in the rejection rate between Italian born applicants and Australian applicants.

**Requirements:** Test the hypothesis, show the results (50% of the task) and explore the causes (50% of the task).

**Purpose:** To explore the cause of the hypothesis (i.e. Why are more Italian born applicants refused aid than Australian born applicants). The answer to this question is that most Italian born applicants are older, applying for family matters, where Australian born applicants are younger, applying for criminal matters. Older applicants are generally refused aid because of their wealth. Family matters are typically more often refused aid compared to criminal matters.

### MineSet

1. expert A - non technical

**Completed 50% :** Hypothesis tested, correct results shown.

**Not completed 50% :** Unable to explore the causes.

**Additional comments:** *Software not intuitive. Confusing to adjust confidence and support. Too many rules shown. Not very easy to use for non-technical person.*

2. expert B - non technical

**Completed 50% :** Hypothesis tested, correct results shown.

**Not completed 50% :** Unable to explore the causes.

**Additional comments:** *Unable to group.*

3. expert C - technical

**Completed 50% :** Hypothesis tested, correct results shown.

**Not completed 50% :** Unable to explore the causes.

**Additional comments:** N/A

4. expert D - technical

**Completed 50% :** Hypothesis tested, correct results shown.



**Not completed 50%** : Unable to explore the causes.

**Additional comments:** *Would like to be able to run separate reports for each country.*

5. expert E - technical

**Completed 50%:** Hypothesis tested, correct results shown.

**Not completed 50%:** Unable to explore the causes.

**Additional comments:** *Tried to find a major refusal matter code.*

## WebAssociate

1. expert A - non technical

**Completed 100%** : Hypothesis tested, correct results shown, the causes explored and explained.

**Not completed:** N/A

**Additional comments:** *Most Italian born applicants are older. They mainly apply for aid for "family other" matters, of which are mostly male applicants refused on means and guidelines. Most of the cases were assigned.*

2. expert B - non technical

**Completed 100%** : Hypothesis tested, correct results shown, the causes explored and explained.

**Not completed:** N/A

**Additional comments:** *Very easy steps to drill down to test and explore the hypothesis. More Italian born applicants are refused for the reasons that more Italians are older, refused for "guidelines and means", males who apply for "family other" matters. Family matters have very high refusal rate, especially "family other".*

3. expert C - technical

**Completed 90%:** Hypothesis tested, correct results shown, 40% of the causes explored and explained.

**Not completed 10%** : Causes not fully explained

**Additional comments:** *I would need a little bit more training for the hypothesis explanation graphs.*

4. expert D - technical

**Completed 100%** : Hypothesis tested, correct results shown, the causes explored and explained.

**Not completed:** N/A

**Additional comments:** *The results were cross-referenced to complimentary fields such as matter code, law type and age group.*

5. expert E - technical

**Completed 100%** : Hypothesis tested, correct results shown, the causes explored and explained.

**Not completed:** N/A

**Additional comments:** *The refused Italian born applicants are largely older 41-50 and 51-60 and who applied in Family Law Matters.*

## Gnome Data Miner

1. expert C - technical

**Completed 50%:** Hypothesis tested, correct results shown.

**Not completed 50%** : Unable to explore the causes.

**Additional comments:** *Results only provided in text format. Difficult to verify and cross reference. Problems with setting confidence and support threshold; if too high - Italians are not shown, if too low, too many rules shown.*

2. expert D - technical

**Completed:** N/A

**Not completed 100%:** Unable to tested the hypothesis, show results or explore the causes.

**Additional comments:** *This application is very time consuming, unless you have all day. I have no idea of results you are expecting. I wouldn't recommend using this application.*

3. expert E - technical

**Completed 50%** : Hypothesis tested, correct results shown.

**Not completed 50%** : Unable to explore the causes.

**Additional comments:** *Unable to complete. Too hard.*

As table 6.3 shows, by using MineSet or Gnome Data Miner, the participants have not been able to explore the causes for the deviation between Italian and Australian refusal rate. By using WebAssociate, almost all participants were able to explain that Italian born applicants were refused due to fact that Italian born applicants were older than Australian born applicants. Italian born applicants applied for mostly family law matters, where Australian born applicants applied for criminal matters. Family matters are in general refused aid compared to criminal matters. The experts explained that older applicants are wealthier and therefore more often refused aid, and that Australian born applicants are mostly younger applicants with fewer assets.

As Table 6.3 shows, most of technical experts had problem testing and explaining the

	non-technical	non-technical	technical	technical	technical
MineSet	50%	50%	50%	50%	50%
WebAssociate	100%	100%	90%	100%	100%
G. Data Miner	0%	0%	50%	0%	50%

Table 6.3: Task 1 - percentage of completion

hypothesis with Gnome Data Miner which outputs association rules as text. The experts also struggled setting the confidence and support thresholds, required by MineSet and Gnome Data Miner. Only WebAssociate methods allowed the experts to search for the cause of the hypothesis. Only WebAssociate enabled the participants to explore cause of the hypothesis.

## Task 2

**Description:** The company has received a racist letter from a disgruntled legal aid applicant. The applicant claims that legal aid decisions are biased in terms of ethnic background for Southern European applicants versus Anglo-Saxon applicants (Australian and UK born).

*Requirements* Group Wales, Scotland and England born applicants as UK group, Greek and Italian born applicants as Southern European group and Australian born applicants as OZ group (50% of the task). Test the hypothesis and show the results (50% of the task).

**Purpose:** To be able to define new groups. For example applicants that are in "under16", "16 to 18", "18 to 25" age groups could be defined as group "YOUNG" and applicants that are in "25 to 30", "30 to 40" and "40 to 50" age groups could be defined to group "MIDDLE AGE" group. The new defined groups could be used as population under study, rather than just data defined groups.

## MineSet

1. expert A - non technical

**Completed 25%** : Hypothesis tested, correct results shown.

**Not completed 75%** : Unable to group. Not correct results shown.

**Additional comments:** *Very confusing because I couldn't group countries. Graphics difficult to manipulate. Had to manually calculate the average refusal rate for the countries. Difficult to set the confidence and support thresholds.*

2. expert B - non technical
 

**Completed 25%** : Hypothesis tested, correct results shown.

**Not completed 75%** : Unable to group. Not correct results shown.

**Additional comments:** *Unable to group countries. Needs manual calculations.*
3. expert C - technical
 

**Completed 0%** : N/A

**Not completed 100%** : Unable to test hypothesis, group countries and show results. shown.

**Additional comments:** N/A
4. expert D - technical
 

**Completed 25%** : Hypothesis tested, correct results shown.

**Not completed 75%** : Unable to group. Not correct results shown.

**Additional comments:** *Unable to define new groups.*
5. expert E - technical technical expert
 

**Completed 25%** : Hypothesis tested, correct results shown.

**Not completed 75%** : Unable to group. Not correct results shown.

**Additional comments:** N/A

## WebAssociate

1. expert A - non technical
 

**Completed 100%** : Hypothesis tested, correct results shown, groups defined.

**Not completed 0%** : N/A

**Additional comments:** *Very easy to select groups and to view results (as manager). Easy to use and to understand information.*
2. expert B - non technical
 

**Completed 100%** : Hypothesis tested, correct results shown, groups defined.

**Not completed 0%** : N/A

**Additional comments:** N/A
3. expert C - technical
 

**Completed 95%** : Hypothesis tested, correct results shown, groups defined.

**Not completed 5%** : Incorrect result for "Southern Europe" group. The participant included country Spain in this group which resulted 24.6% refusal instead of 25.2% refusal.

**Additional comments:** N/A

4. expert D - technical

**Completed 100%** : Hypothesis tested, correct results shown, groups defined.

**Not completed 0%** : N/A

**Additional comments:** N/A

5. expert E - technical

**Completed 100%** : Hypothesis tested, correct results shown, groups defined.

**Not completed 0%** : N/A

**Additional comments:** N/A

### Gnome Data Miner

1. expert C - technical

**Completed 0%** : N/A

**Not completed 100%** : Unable to test hypothesis, group countries and show results. shown.

**Additional comments:** N/A

2. expert D - technical

**Completed 0%** : N/A

**Not completed 100%** : Unable to test hypothesis, group countries and show results.

**Additional comments:** *Unable to complete, too hard.*

3. expert E - technical

**Completed 0%** : N/A

**Not completed 100%** : Unable to test hypothesis, group countries and show results. shown.

**Additional comments:** *Time consuming exercises. I would not recommend using this application.*

None of the participants were able to complete Task 2 by using Gnome Data Miner. It was obvious that text representation of the discovered rules did not allow the participants to understand the rules, nor to group them. The largest problem with MineSet was inability to define a new group (aggregate values) for non-numeric data. However, MineSet has option for group definition based on the mathematical functions such as AVG, MIN and MAX, but this option was not useful for this task. Additionally, the participants again had difficulties to set the confidence and support threshold required by MineSet. Most of participants have completed the task by using WebAssociate, as shown in Table 6.4.

	non-technical	non-technical	technical	technical	technical
MineSet	25%	25%	0%	25%	25%
WebAssociate	100%	100%	95%	100%	100%
G. Data Miner	0%	0%	0%	0%	0%

Table 6.4: Task 2 - percentage of completion

As table 6.4 shows, Gnome Data Miner was not able to meet grouping requirements of the domain experts. MineSet has allowed users to explore the hypothesis individually, by finding the refusal rate for each country. However this approach involved manual calculation of the refusal rate for each group. The participants claimed that only WebAssociate met the experts requirements by enabling them to define new groups and test the hypothesis.

### Task 3

**Description:** Provide a short report on the Vietnamese applicants. Especially focus on the most common matter code amongst those applicants.

**Requirements:** Find what is the most common matter code for the Vietnam born applicants (50% of the task) and write a short report on this population (50% of the task).

**Purpose:** To be able to drill down and explore the population. Correct approach to this task is to find that Vietnamese applicants applied mostly for drug related offences *matterCode\_RD*. Majority of such applicants are younger males that were approved aid.

### MineSet

#### 1. expert A - non technical

**Completed 50% :** Found the most common matter code for the Vietnam born applicants.

**Not completed 50% :** Unable to drill down. Report not provided.

**Additional comments:** *Gets a bit easier with settings, but hard to remember all steps. Difficult to remember filters and how to use it. Difficult to read screen, can't manipulate image. Rules generation difficult to understand. Difficult for manager to know how to drill down quickly and effectively.*

#### 2. expert B - non technical

**Completed 50% :** Found the most common matter code for the Vietnam born applicants.

**Not completed 50% :** Unable to drill down. Report not provided.

**Additional comments:** *Required manual search through textual output. Very hard to read.*

3. expert C - technical

**Completed 50% :** Found the most common matter code for the Vietnam born applicants.

**Not completed 50% :** Unable to drill down. Report not provided.

**Additional comments:** *The methodology used was to lower confidence and support in order to capture all the data. Then based upon a graphical representation of the results, particular data was pinpointed.*

4. expert D - technical

**Completed 0% :** N/A

**Not completed 100% :** Most common matter code for the Vietnam born applicants not found. Report not provided.

**Additional comments:** *Unable to complete report. Message "End of input. Data file is empty".*

5. expert E - technical

**Completed 50% :** Found the most common matter code for the Vietnam born applicants.

**Not completed 50% :** Unable to drill down. Report not provided.

**Additional comments:** *I think that information is difficult to read the way the results are presented. I don't like that you have to point the mouse on the data in order to read the values.*

## WebAssociate

1. expert A - non technical

**Completed 70% :** Found the most common matter code for the Vietnam born applicants. Was able to drill down but did not include the results in the report.

**Not completed 30% :** Report not completed.

**Additional comments:** *Easy to extract all the information. Only confusion related to whether percentage is of all Vietnam born applicants or of all Vietnam born applicants and matter code "RD". Good data visualisation.*

2. expert B - non technical

**Completed 100% :** Found the most common matter code for the Vietnam born applicants. Full report provided.

**Not completed 0% :** N/A

**Additional comments:** *291 of 955 Vietnam born applicants applied for aid for matter code "RD" which makes 30.5%. Most of them are males, aged between 19-25, and have been assigned cases.*

3. expert C - technical

**Completed 70%** : Found the most common matter code for the Vietnam born applicants. Was able to drill down but did not include the results in the report.

**Not completed 30%** : Report not completed.

**Additional comments:** N/A

4. expert D - technical

**Completed 100%** : Found the most common matter code for the Vietnam born applicants. Full report provided.

**Not completed 0%** : N/A

**Additional comments:** *291 Vietnam born applicants out of 955 had matter code "RD", which makes 30.5%. Most of those are age 19-25, most are approved and most are male.*

5. expert E - technical

**Completed 100%** : Found the most common matter code for the Vietnam born applicants. Full report provided.

**Not completed 0%** : N/A

**Additional comments:** *291 out of 955 applications submitted by Vietnamese are RD matters - 30.5%. Most of those are age 19-25, most are assigned, most are male and very little were refused.*

## Gnome Data Miner

1. expert C - technical

**Completed 0%** : N/A

**Not completed 100%** : Most common matter code for the Vietnam born applicants not found. Report not provided.

**Additional comments:** N/A

2. expert D - technical

**Completed 0%** : N/A

**Not completed 100%** : Most common matter code for the Vietnam born applicants not found. Report not provided.

**Additional comments:** *Unable to complete. Too hard.*



3. expert E - technical

**Completed 0% :** N/A

**Not completed 100% :** Most common matter code for the Vietnam born applicants not found. Report not provided.

**Additional comments:** *Very time consuming application. I would not recommend using this application.*

By using Gnome Data Miner for the Task 3 the participants were not able to complete the task. The text based rule representation was too complicated for the experts. By using MineSet majority of experts were able to partially complete the task, by finding "RD" as the most common matter code for the Vietnam born applicants. However, by having to decide on the confidence and support threshold settings, the experts struggled to find the small population of Vietnam born applicants (only 955 cases). By using MineSet the experts were not able to further drill in order to provide report for the Vietnam and "RD" group. By using WebAssociate most of the experts were able to complete the task and provide a short report. Table 6.5 shows completion percentage for Task 3.

	non-technical	non-technical	technical	technical	technical
MineSet	50%	50%	50%	0%	50%
WebAssociate	70%	100%	70%	100%	100%
G. Data Miner	0%	0%	0%	0%	0%

Table 6.5: Task 3 - percentage of completion

As Table 6.5 shows, the experts have identified that WebAssociate provides adequate methods for exploration of hypotheses. MineSet provided nice graphics but experts had difficulties with support and confidence threshold settings. If the threshold was set to high, the experts were not able to identify smaller population of the Vietnam born applicants. Lower threshold settings resulted to too many rules. Gnome Data Miner wasn't adequate for this task because it had rule representation as text, which experts found confusing.

## 6.5 Chapter Summary

In this chapter we evaluated three KDD applications. The evaluation process included Usability, User Satisfaction, Utility and Validation tests. The participants were given three "real life" tasks to complete by using each application. Consequently, the participants were asked to provide a short report and questionnaire.

### 6.5.1 Gnome Data Miner

The evaluation results show that Gnome Data Miner was very hard to use by all experts and failed the usability test. The experts were not satisfied by the result representation of this tool and claimed that intermediate results were not very clear. The feedback of this tool was evaluated as not very understandable by the majority of experts. The participants claimed that none of Gnome Data Miner features were found particularly useful, and the application was too complicated and time consuming. By using Gnome Data Miner the participants were able to meet only a small fraction of the task requirements.

### 6.5.2 SGI MineSet

Non-technical domain experts have found MineSet moderately hard to very hard to use, while technical experts described it as hard to moderately easy. The results show that MineSet has high technical requirements and non-technical experts would require substantial training in order to make MineSet useful for their individual requirements. Intermediate results provided by this tool were not very clear to non-technical experts and clear to some degree to the technical experts. Some experts claimed that matrix rule representation was not very clear, not user friendly and confusing. Due to different level of technical knowledge, each participant gave a different evaluation for the feedback provided by MineSet. According to the participant comments, 3D graphics were one of the most advanced features, however the participants claimed that the graphs were not particularly useful and manipulation of the graphical results was difficult. The participants were able to meet some task requirements, however the matrix representation of the discovered patterns and support and confidence threshold settings were described as too complicated and difficult by the experts.

### 6.5.3 WebAssociate

All domain experts described WebAssociate moderately easy to very easy to use, and claim that WebAssociate could be used as an additional KDD tool for their everyday requirements. Intermediate results provided by this tool were described as very clear by the majority of experts. The feedback from WebAssociate was described as very understandable by all experts. The participants comments identified WebAssociate as much more user friendly, with easier to understand and clear data visualisation. Many experts found WebAssociate rule representation as extremely useful. Especially the method of grouping rules together. Even that some experts had small difficulties in interpreting the results, most of the experts were able to complete all three tasks.

We believe that this evaluation test has suggested that domain experts find WebAssociate easier to use and more useful than other KDD tools for their individual everyday requirements.

# Chapter 7

## Conclusion

In this thesis, we performed a comprehensive study on association rule mining, developed a web based KDD tool “WebAssociate”, proposed a new association rule discovery approach which groups association rules based on their content. We also evaluated our tool and provided evaluation results. In this chapter we will conclude with a summary of this study, discuss the limitations and propose some future research directions in this field.

### 7.1 Conclusion

#### 7.1.1 Organisational use of KDD

Many organisations are embracing KDD in order to analyse their ever growing data sets. However, the process of deploying KDD into an organisation is not straight forward. There are usually three phases of deploying KDD technology in an organisation. In phase one and two, sophisticated KDD tools are used for analysing data sets of an organisation, requiring analyst experts with high technical knowledge. In phase three end users (e.g. managers, lawyers and medical professionals) are able to perform their own analysis according to their individual requirements.

However we believe that most organisations implement only phase one and two which results in failure to use the full potential of KDD. In this study we explore the use of KDD in several Australian organisations and conclude that organisations have not used KDD to its full potential, because the KDD tools in these organisations were not suitable for the third phase of the implementation. We found that the current KDD tools in these organisations were built for technical experts, requiring technical training before being useful to the end users. We discovered that the non-technical domain experts do not have adequate KDD tools and are not able to drive a knowledge discovery process without the

help of the technical people.

### 7.1.2 Non-technical domain experts

In this research, we identified the inability of non-technical domain experts to use the current KDD tools that are highly technical and sophisticated and we built an easy-to-use web based KDD tool called “WebAssociate”. We demonstrated that KDD tools used to support non-technical domain experts can be constructed for:

1. hypotheses suggestion
2. hypotheses exploration
3. assisting in explanation
4. pattern visualisation

Our KDD tool has been found to be useful to non-technical domain experts. These users are less skilled in complex data analysis and have less technical knowledge, but have a thorough understanding of their domain. The experts identified that they are usually not interested in using advanced powerful technology per se, but only in getting clear, rapid answers to their everyday business questions. For example, VLA domain experts need to regularly analyse their data in order to provide business reports. The reports should contain answers to the following questions:

- Who applied for legal aid?
- Who was refused legal aid and why?
- What were certain age groups applying for?
- What were certain national groups applying for?

In order to assist non-technical users to find the answers, our tool utilises a widely used KDD technique called “Association Rules” to discover patterns in data. However “WebAssociate” tackles the problem of discovering interestingness by grouping discovered association rules into “rule sets”. The interestingness of the discovered rules is based on the visually displayed deviations between the confidence values of items in a rule set which can be used to suggest or test hypotheses. Visual display of grouped ARs allows the user to inspect the findings and identify interesting ones according to their similarities and differences.

### 7.1.3 Grouping AR for hypothesis suggestion

By grouping association rules we provide a direct connection between ARs and hypotheses which permits more effective data exploration. Using this approach we enable the expert to focus more directly on the hypothesis under investigation than on the rules. A “rule set” in our study contains a set of discovered association rules that have common consequent and different antecedent. The antecedents in a rule set are attribute-values from the same attribute. For example, lets consider a rule set RS1 containing sex MALE as the common consequent, and law type attribute with three attribute-values: CRIMINAL, FAMILY and CIVIL as antecedents. The rule set RS1 than contains three association rules:  $lawType\_CRIMINAL \Rightarrow sex\_MALE$ ,  $lawType\_FAMILY \Rightarrow sex\_MALE$  and  $lawType\_CIVIL \Rightarrow sex\_MALE$ . Similarities or differences between the confidence values for association rules in RS1 could suggest previously unknown hypothesis to the user. For example, if confidence values for ARs in the rule set RS1 are similar (e.g. 70%, 71% and 69% respectively), their similarity suggests the hypothesis “*There is no difference in the proportion of male applicants between all three law types*”. This hypothesis is considered interesting if it contradicts users’ expectations (e.g. the user expected a greater proportion of females in family law applications).

Most approaches connect discovered ARs according to user specified a minimum threshold. The threshold is based on the frequency value e.g. confidence, support and lift. Only ARs that meet the minimum threshold are discovered and considered interesting. The problem with these approaches is that users usually do not know what the minimum threshold should be. If the threshold is too low, the user is overwhelmed by too many rules, if the threshold is too high, less frequent items are not included in the discovery (rare item problem). Furthermore, the frequency connection between discovered ARs is not likely to enable users to map AR to hypotheses. Therefore this does not support the hypothesis suggestion phase.

In our study the connection between the discovered association rules is based on their content and not frequency. Our approach overcomes the “rare item” problem. By connecting rules according to their content we enable users to automatically map AR to hypothesis. For example, the association rules in the rule set RS1 are connected because they share the same consequent “sex\_MALE”, and have a common antecedent attribute “law type”.

### 7.1.4 Visualisation

Many researchers as well as our own evaluation of WebAssociate suggest that users find a visual representation of a discovery easier to understand than a textual representation. In

this study, by grouping discovered ARs into rule sets we can easily visualise the discovered rules. For example, the RS1 rule set (discussed in Section 7.1.3) can be visualised by plotting their confidence values. The label “sex\_MALE” on X axis shows the common consequent for AR in this rule set. The range of 0 to 100 on the Y axis corresponds to possible confidence values. Confidence values for each AR in RS1 would be plotted as points with values 70%, 71% and 69% respectively. Each point is color coded (e.g. blue, red and green) and used to distinguish the antecedents. The legend shows each antecedent and its corresponding color. For example, a user can visually inspect the rule set RS1 and easily identify the following: according to three points on the Y axis (70, 71 and 69) for X axis label “sex\_MALE”, and antecedent values in the Legend (CRIMINAL, FAMILY and CIVIL), there is a very small proportional difference of male applicants between all three law types. The evaluation in this study shows that the non-technical experts prefer our AR visualisation approach to the “2D-Matrix” used by SGI-MineSet and the textual approach used by Gnome Data Miner (CSIRO GUI for Apriori).

### 7.1.5 Development of WebAssociate

Our web based KDD tool was developed by using prototyping. The focus of the development was to create an easy-to-use tool that will be primarily used by non-technical domain experts. The evolution of WebAssociate was to mimic how domain experts think. In order to answer their everyday questions, domain experts needed a tool that would show differences between groups in a data set. The tool was developed to investigate simple groups (e.g. Males vs Females, Australians vs Italians and refused vs approved) or more complex groups (e.g. Males and refused vs Males and approved). Differences or similarities between such groups can be used to suggest unknown hypotheses or allow users to test previously known hypotheses.

### 7.1.6 Software Evaluation

In this study we evaluated “WebAssociate” with two other commercial data mining tools; Silicon Graphics - MineSet and Gnome Data Miner. In the evaluation process we involved domain experts using the tools to solve a typical business problems. The experts evaluated the usability, usefulness, user satisfaction and validation of each tool. The experts found “WebAssociate” a very useful and easy to use KDD tool which can be used to answer their everyday business questions. The tool allows three levels of validation. In the first level the tool might aid the user in suggesting hypotheses. In the second level the tool allows hypotheses exploration and confirmation by explanation with evidence from the data. Finally, the third level allows hypotheses testing.

## 7.2 Limitations

Our approach is not suitable for Basket data analysis because we do not use support. The aim of basket data analysis is to discover frequent item sets. Frequent item sets are sets of items that frequently appear together (e.g. Milk and Bread). Some researchers claim that it would not be useful for a supermarket manager to discover items that are rarely bought (e.g. Caviar and Champagne). However, some data sets containing low frequency items are still important (e.g. adverse drug reactions) and are suitable for our approach.

The second limitation of our approach is that we are constrained by the number of variables (columns) in a data set. By using the apriori algorithm we are limited to less than twenty variables in a data set. However, research shows that users find difficulties in understanding ARs that contain more than five items in their antecedent or consequent.

WebAssociate does not limit the number of rows. In this study we used several data sets ranging from over 4000 records to over 380,000 records. We did not find a great difference in the processing time between the various number of records.

## 7.3 Further Research

In this study we built a web based KDD tool that uses ARs for pattern discovery. However we identified the need for additional KDD tools. In further research we would attempt to develop additional KDD tools for non-technical experts that use other data mining methods such as clustering and classification. We believe that by building easy-to-use additional tools that do not need a high level of technical knowledge, we would enable organisational use of KDD tools at all levels.



# Bibliography

- [1] Syed Sibte Raza Abidi and Zaharin Yusoff. Data-driven healthcare management: From a philosophy to an info-structure. In *In International Conference on Multimedia and Information Technology, Kuala Lumpur, Malasia, August 1998*, pages 11–19, 1998.
- [2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD - Int. Conf. Management of Data*, pages 207–216, 1993.
- [3] A.M. Rubinov A.M. Bagirov and J. Yearwood. Using global optimization to improve classification for medical diagnosis and prognosis. *Topics in Health Information Management*, 22(1):65–74, 2001.
- [4] Sarabjot S. Anand, David A. Bell, and John G. Hughes. The role of domain knowledge in data mining. In *CIKM*, pages 37–43, 1995.
- [5] Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. In *Proceedings of the KDD-99 conference, San Diego CA - USA*, pages 261–270, 1999.
- [6] Roberto J. Bayardo and Rakesh Agrawal. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international booktitle on Knowledge discovery and data mining*, pages 145–154. ACM Press, 1999.
- [7] Pavel Berkhin. Technical report: Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.
- [8] F. Bonchi, F. Giannotti, G. Mainetto, and D. Pedreschi. A classification-based methodology for planning audit strategies in fraud detection. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California - USA*, pages 175–184. ACM Press, 1999.

- [9] Ronald J. Brachman, Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, and Evangelos Simoudis. Mining business databases. *Communications of the ACM*, 39(11):42–48, 1996.
- [10] Sergey Brin, Rajeev Motwani, and Craig Silversteint. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the SIGMOD 97 AZ - USA*, pages 265–276, 1997.
- [11] Thomas Brinkhoff. Using a cluster manager in a spatial database system. In *Proceedings of the ninth ACM international symposium on Advances in geographic information systems, Atlanta, Georgia - USA*, pages 136–141. ACM Press, 2001.
- [12] Robert B. Burns. *Introduction to Research Methods*. Addison Wasley Longman Australia Pty Limited, third edition edition, 1998.
- [13] Tom Burr, James R. Gattiker, and Gregory S. LaBerge. Genetic subtyping using cluster analysis. *ACM SIGKDD Explorations Newsletter*, 3(1):33–42, 2001.
- [14] Soumen Chakrabarti, Sunita Sarawagi, and Byron Dom. Mining surprising patterns using temporal description length. In *Proceedings of the 24th International Conference on Very Large databases VLDB98, New York - USA*, pages 606–617. Morgan Kaufmann, 1998.
- [15] Ming-Syan Chen, Jiawei Han, and Philip S. Yu. Data mining: an overview from a database perspective. *Ieee Trans. On Knowledge And Data Engineering*, 8:866–883, 1996.
- [16] Paul B. Chou, Edna Grossman, Dimitrios Gunopulos, and Pasumarti Kamesam. Identifying prospective customers. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, Massachusetts - USA*, pages 447–456. ACM Press, 2000.
- [17] Two Crows Corporation. Booklet: Introduction to data mining and knowledge discovery, third edition, 1999.
- [18] C.Wong, P.Whitney, and J.P.Thomas. Visualizing association rules for text mining. In *Proceedings of the 1999 IEEE Symposium on Information Visualization, San Francisco, California -USA*, pages 120–123, 1999.
- [19] D.A.Keim and H.P.Kriegel. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):923–938, 1996.

- [20] Stephen D.Bay and Michael J.Pazzani. Detecting change in categorical data: Mining contrast sets. In *The conference proceedings of KDD-1999 San Diego CA USA*, pages 302–306, 1999.
- [21] Stephen D.Bay and Michael J.Pazzani. Discovering and describing category differences: What makes a discovered difference insightful? In *In Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society.*, 2000. [<http://newatlantis.isle.org/sbay/papers/eval.pdf>] last visited 06/09/2003.
- [22] Stephen D.Bay and Michael J.Pazzani. Detecting group differences: Mining contrast sets. In *The conference proceedings of KDD-2001*, pages 213–246, 2001.
- [23] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *Ai Magazine*, 17:37–54, 1996.
- [24] Usama Fayyad, David Haussler, and Paul Stolorz. Mining scientific data. *Communications of ACM*, 39(11):51–57, November 1996.
- [25] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.
- [26] Usama Fayyad and Ramasamy Uthurusamy. Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11):24–26, 1996.
- [27] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases. In *Knowledge Discovery in Databases*, pages 1–27. AAAI Press, Menlo Park, CA, 1991.
- [28] Alex A. Freitas. Understanding the crucial differences between classification and discovery of association rules: a position paper. *ACM SIGKDD Explorations Newsletter*, 2(1):65–69, 2000.
- [29] Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. A framework for measuring changes in data characteristics. In *Proceedings of the PODS-99 conference, Philadelphia PA - USA*, pages 126–137, 1999.
- [30] N. Gershon, S.G.Eick, and S.Card. Information visualization. *ACM Interactions*, 5(2):9–15, March/April 1998.
- [31] G.Grinstein and D.Thuraisingham. Data mining and data visualization, database issues for data visualization. In *Proceedings of IEEE Visualization 95 Workshop Phoenix Arizona-USA*, pages 54–56, 1995.

- [32] Michael Goebel and Le Gruenwald. A survey of data mining and knowledge discovery tools. *SIGKDD Explorations*, 1(1):20–33, June 1999.
- [33] Guido Governatori and Andrew Stranieri. Towards the application of association rules for defeasible rules discovery. In *Legal Knowledge and Information Systems*, pages 63–75, Amsterdam, 2001. JURIX, IOS Press.
- [34] Maria Jean Hall, Richard Hall, and John Zeleznikow. A process for evaluating legal knowledge-based systems based upon the context criteria contingency-guidelines framework. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, June 24-28, Edinburgh, Scotland, UK*, pages 274–283, 2003.
- [35] Jia Liang Han and Ashley W. Plank. Background for association rules and cost estimate of selected mining algorithms. In *Proceedings of the fifth international conference on Information and knowledge management*, pages 73–80. ACM Press, 1996.
- [36] Jianchao Han and Nick Cercone. Ruleviz: a model for visualizing knowledge discovery process. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 244–253. ACM Press, 2000.
- [37] Jiawei Han. Conference tutorial notes: Data mining techniques. In *Communications of ACM-SIGMOD International Conference on Management of Data, Montreal, Canada*, 1996.
- [38] R. Hilderman and H. Hamilton. Knowledge discovery and interestingness measures: A survey. technical report cs 99-04. Technical report, Department of Computer Science, University of Regina, October 1999.
- [39] Jochen Hipp and Ulrich Gntzer. Is pushing constraints deeply into the mining algorithms really what we want?: an alternative approach for association rule mining. *ACM SIGKDD Explorations Newsletter*, 4(1):50–55, 2002.
- [40] H. Hofmann, P.J.Siebes, and F.X.A Wilhelm. Visualising association rules with interactive mosaic plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, August 20-23, 2000, Boston, MA, USA*, pages 227–235, 2000.
- [41] Wynne Hsu, Mong Li Lee, Bing Liu, and Tok Wang Ling. Exploration mining in diabetic patients databases: Findings and conclusions. In *The conference proceedings of Knowledge Discovery and Data Mining, KDD-2000*, pages 430–436, 2000.

- [42] H.White. A reality check for data snooping. In *The conference proceedings of Econometrica-2000*, pages 1097–1127, 2000.
- [43] Tomasz Imielinski and Heikki Mannila. A database perspective on knowledge discovery. *Communications of ACM*, 39(11):58–64, 1996.
- [44] Sasa Ivkovic, John Yearwood, and Andrew Stranieri. Discovering interesting association rules from legal databases. *Information and Communication Technology Law*, 1(8):35–47, 2002.
- [45] Sasa Ivkovic, John Yearwood, and Andrew Stranieri. Visualizing association rules for feedback within the legal system. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, June 24-28, Edinburgh, Scotland, UK*, pages 214–223, 2003.
- [46] Joyce Jackson. Data mining: A conceptual overview. *Communications of the Association for Information Systems*, 8:267–296, 2002.
- [47] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.
- [48] J.J.Dijkstra. User interaction with legal knowledge based systems. In *Legal Knowledge and Information Systems, Jurix 2000 - Amsterdam*, pages 11–21. IOS Press Amsterdam, 2000.
- [49] Mika Klementtinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and Inkeri Verkamo. Finding interesting rules from large data sets of discovered association rules. In *The conference proceedings of CIKM'94 Gaithersburg MD, USA*, pages 401–408, November 1994.
- [50] M. Klemettinen, H. Mannila, and H. Toivonen. Interactive exploration of discovered knowledge: A methodology for interaction and usability studies - technical report c-1996-3. Technical report, University Of Helsinki, Department of Computer Science, 1996.
- [51] Vipin Kumar and Mohammed Zaki. High performance data mining (tutorial pm-3). In *Tutorial notes of the sixth ACM SIGKDD international booktitle on Knowledge discovery and data mining*, pages 309–425. ACM Press, 2000.
- [52] Wentian Li and Ivo Grosse. Gene selection criterion for discriminant microarray data analysis based on extreme value distributions. In *Proceedings of the seventh annual international conference on Computational molecular biology, Berlin - Germany*, pages 217–223. ACM Press, 2003.

- [53] Wen-Yang Lin, Ming-Cheng Tseng, and Ja-Hwung Su. A confidence-lift support specification for interesting association mining. In *Proceedings of the PAKDD international booktitle on Knowledge discovery and data mining*, pages 148–158. Springer-Verlag Berlin Heidelberg, 2002.
- [54] Bing Liu, Wynne Hsu, and Shu Chen. Using general impressions to analyze discovered classification rules. In *The proceedings of the Third International Conference on KDD*, pages 31–36, 1997.
- [55] Bing Liu, Wynne Hsu, Shu Chen, and Yiming Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.
- [56] Bing Liu, Wynne Hsu, and Shu Chen Ke Wang. Visually aided exploration of interesting association rules. In *In Proceedings of the 3rd Pacific-Asia Conferences on Knowledge Discovery and Data Mining, PAKDD-99*, pages 380–389, Heidelberg, Germany, 1999. Springer.
- [57] Bing Liu, Wynne Hsu, and Yiming Ma. Mining association rules with multiple minimum supports. In *The conference proceedings of KDD-99 San Diego CA USA*, pages 337–341, 1999.
- [58] Bing Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In *The conference proceedings of KDD-99 San Diego CA, USA*, pages 125–134, 1999.
- [59] Bing Liu, Wynne Hsu, and Yiming Ma. Identifying non-actionable association rules. In *The conference proceedings of ACM SIGKDD-2001 international booktitle on Knowledge discovery and data mining*, pages 329–334. ACM Press, 2001.
- [60] Bing Liu, Minqing Hu, and Wynne Hsu. Multi-level organization and summarization of the discovered rules. In *The conference proceedings of SIGKDD-2000 - Boston, USA*, pages 208–217. ACM Press, 2000.
- [61] Yiming Ma, Bing Liu, Ching Kian Wong, Philip S. Yu, and Shuik Ming Lee. Targeting the right students using data mining. In *The conference proceedings of SIGKDD-2000 - Boston, USA*, pages 457–462. ACM Press, 2000.
- [62] Heikki Mannila. Methods and problems in data mining. In *The booktitle proceedings of ICDT-97*, pages 41–55, 1997.
- [63] N.Gershon, S.G.Eick, and S.Card. Information visualization. *ACM Interactions*, 5(2):9–15, March/April 1998.

- [64] Beng Chin Ooi, Kian-Lee Tan, Tat Seng Chua, and Wynne Hsu. Fast image retrieval using color-spatial information. *The VLDB Journal The International Journal on Very Large Data Bases*, 7(2):115–128, 1998.
- [65] Thomas Ormerod, Nicola Morley, Linden Ball, Charles Langley, and Clive Spenser. Using ethnography to design a mass detection tool (mdt) for the early discovery of insurance fraud. In *Proceedings of the extended abstract conference on human factors and computing systems, Ft. Lauderdale, Florida - USA*, pages 650–651. ACM Press, 2003.
- [66] Shan L. Pan and Jae-Nam Lee. Using e-crm for a unified view of the customer. *Communications of the ACM*, 46(4):95–99, 2003.
- [67] A. Pannu. Using genetic algorithms to inductively reason with cases in the legal domain. In *ICAIL'95 Proceedings of the Fifth International Conference on Artificial Intelligence and Law*, pages 175–184, 1995.
- [68] Gregory Piatetsky-Shapiro and Christopher J. Matheus. The interestingness of deviations. In *The conference proceedings of KDD-94 Knowledge Discovery in Databases*, pages 25–36, 1994.
- [69] Gopal Pingali, Agata Opalach, Yves Jean, and Ingrid Carlbom. Visualization of sports using motion trajectories: providing insights into performance, style, and strategy. In *Proceedings of the conference on Visualization 2001*, pages 75–82. IEEE Press, 2001.
- [70] Eric S. Raymond. *Book: The Cathedral and the Bazaar*. O'Reilly and Associates, 1999.
- [71] Edwina Rissland and Timur Friedman. Detecting change in legal concepts. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law. ICAIL'95*, pages 127–136. IAAIL, ACM Press, 1995.
- [72] Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 105–115. ACM Press, 2000.
- [73] Sigal Sahar. Interestingness via what is not interesting. In *The conference proceedings of Knowledge Discovery and Data Mining, KDD-1999, San Diego, California - USA*, pages 332–336, 1999.
- [74] Erich Schweighofer and Dieter Merkl. A learning technique for legal document analysis. In *ICAIL'99 Proceedings of the Seventh International Conference on Artificial Intelligence and Law*, pages 156–164. ACM Press, 1999.

- [75] Devavrat Shah, Laks V. S. Lakshmanan, Krithi Ramamritham, and S. Sudarshan. Interestingness and pruning of mined patterns. In *1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1999.
- [76] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal - CANADA*, pages 1–12, Jun 1996.
- [77] Andrew Stranieri, John Zeleznikow, and Huch Turner. Data mining in law with association rules. In *Proceedings of IASTED International conference on Law and Technology.*, pages 129–135, 2000.
- [78] David Waltz and Se June Hong. Guest editors' introduction: Data mining: A long-term dream. In *The conference proceedings of Intelligent Systems IEEE-99*, pages 30–31, 1999.
- [79] Ke Wang, Yu He, and David W. Cheung. Mining confident rules without support requirement. In *The conference proceedings of CIKM'01 - Atlanta Georgia, USA*, pages 89–96. ACM Press, 2001.
- [80] Dawn Wilkins and Krishan Pillaipakkamnatt. The effectiveness of machine learning techniques for predicting time to case disposition. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law*, pages 106–113. AIL, ACM Press, 1997.
- [81] Peggy Wright. Knowledge discovery in databases: tools and techniques. *Crossroads*, 5(2):23–26, 1998.
- [82] X.Lin, H.White, and J.Buzydlowski. Associative searching and visualization. In *International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet - SSGRR'2001, L'Aquila - Italy*, pages 182–189, 2001.
- [83] Jiong Yang, Wei Wang, and Philip S. Yu. Infominer: Mining surprising periodic patterns. In *Proceedings of the KDD-01 conference, San Francisco CA- USA*, pages 395–400, 2001.
- [84] Suk-Chung Yoon, Lawrence J. Henschen, E. K. Park, and Sam Makki. Using domain knowledge in knowledge discovery. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 243–250. ACM Press, 1999.



- [85] John Zeleznikow and Andrew Stranieri. Knowledge discovery in the split up project. In *Proceedings of Sixth International Conference on Artificial Intelligence and Law*, pages 89–97, 1997.
- [86] Hongen Zhang and Sharma Chakravarthy. Mining and vizualisation of association rules over relational dbmss. In *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC), March 9-12, 2003, Melbourne, Florida, USA*, pages 922–926, 2003.
- [87] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management, McLean, Virginia - USA*, pages 515–524. ACM Press, 2002.
- [88] Zhi-Hua Zhou. Technical report: Three perspectives of data mining. Technical report, National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China, 2001.