Discovering the hidden knowledge in transaction data through formal concept analysis

WATMOUGH, Martin

Discovering the Hidden Knowledge in Transaction Data through Formal Concept Analysis

Martin John Watmough

A thesis submitted in partial fulfilment of the requirements of

Sheffield Hallam University

for the degree of Doctor of Philosophy

August 2013

Abstract

The aim of this research is to discover if hitherto hidden knowledge exists in transaction data and how it can be exposed through the application of Formal Concept Analysis (FCA).

Enterprise systems capture data in a transaction structure so that they can provide information that seeks to align with the knowledge that decision-makers use to achieve business goals. With the emergence of service-oriented architecture and developments in business intelligence, data in its own right is becoming significant, suggesting that data in itself may be capable of capturing human behaviour and offer novel insights from a 'bottom-up' perspective. The constraints of hard-coded top-down analysis can thus be addressed by agile systems that use components based on the discovery of the hidden knowledge in the transaction data.

There is a need to connect the user's human-oriented approach to problem solving with the formal structures that computer applications need to bring their productivity to bear. FCA offers a natural approach that meets these requirements as it provides a mathematical theory based on concepts, logical relationships that can be represented and understood by humans.

By taking an action research and case study approach an experimental environment was designed along two avenues. The first was a study in an educational setting that would combine the generation of the data with the behaviour of the users (students) at the time, thereby capturing their actions as reflected in the transaction data. To create a representative environment, the students used an industry standard SAP enterprise system with the business simulator ERPsim. This applied study provided an evaluation of FCA and contemporary tools while maintaining a relevant pedagogic outcome for the students.

The second avenue was a discovery experiment based on user activity logs from an actual organisations productive system, applying and developing the methods applied previously. Analysis of user logs from this system using FCA revealed the hitherto hidden knowledge in its transaction data by discovering patterns and relationships made visible through the multi dimensional representation of data.

The evidence gathered by this research supports FCA for exposing and discovering hidden knowledge from transactional data, it can contribute towards systems and humans working together more effectively.

Preface

As a Manufacturing Engineer and SAP Consultant with a strong interest in integrating systems with the real world, the application of Formal Concept Analysis has been fascinating particularly as it embraces both technical and human aspects.

Identifying opportunities for driving holistic improvements is the dominant motivator; these can range from very simple changes and basic education through to introducing new technologies and processes. Frequently the first steps in this process are just to listen, understand and apply knowledge.

My background led to studying for personal improvement and in a small way resulted in a contribution to knowledge. Challenging as this was, probably the most rewarding experience through this journey has been gained from the opportunity to teach, share and debate in conjunction with collecting experimental data.

The research drew upon an opportunity to generate experimental data in an educational setting using ERPsim, an enterprise system simulator. This provided a rich set of data in controlled and observable environment. Limitations in the simulator led to a second avenue of experimentation based on user activity using data gathered from an organisation's productive system. A new version of ERPsim for the 2013/14 academic year includes the capability to capture user activity data but this will have to form a topic for further work.

Chapter 1 introduces the motivation for the research establishes the research aim, to discover if hitherto hidden knowledge exists in transaction data and how it can be exposed through the application of Formal Concept Analysis.

Chapter 2 considers the background, intentions and visions within transactional systems and analysis. The difficulties involved in understanding the dynamics of an organisation are discussed. This is due to many factors including the complexity of applications and technology, increasing data volumes and competition forcing constant change.

Chapter 3 focuses on the application of FCA to transactional data. An introduction to Formal Concept Analysis and a synopsis of alternative semantic technologies contrasted with FCA for discovering knowledge in transactional data. .

Chapter 4 provides an overview of an opportunity to observe and evaluate the application of FCA and contemporary techniques; a comprehensive description is included in appendix E. Learning, Teaching and Assessment (LTA) involving two degree modules over several years provided the environment for creating an immersive experience for students and generated experimental data in an ethical manner.

Evaluation of the LTA outputs provided an understanding of the findings and insight into the application of FCA. Moderated assessment criteria provided tangible measures

and text providing a qualitative input for the research and progressive improvements to the LTA cycle. This provided a rich context for the participants and a source of data for analysis.

Chapter 5 draws on user activity data extracted directly from a productive system. This practical application of FCA highlights processes where bottom-up analysis can uncovered useful information and knowledge.

Chapter 6 brings together the results from the LTA and user-activity based experiments. A discussion follows with examples of knowledge discovered from transactional data. Suggested and justified requirements for the successful application of FCA are presented and aligned with domains where FCA is potentially justifiable and cost effective.

Chapter 7 concludes and discusses directions for further work required to develop FCA into a practical application integrated into a BI solution.

Acknowledgements

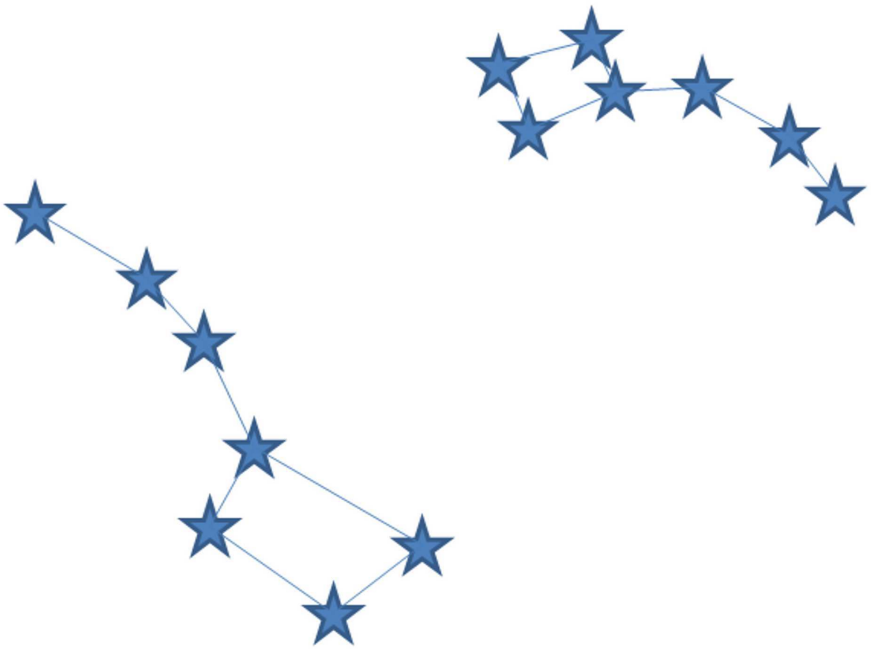
I cannot thank Dr. Simon Polovina enough for the guidance and support he has given me. Truly first class all the way!

I am grateful for the ideas and support Dr. Babak Khazaei has given me, spurring me on through some difficult times.

Thanks also to the students at Sheffield Hallam University who supported my research. I hope my lecturing, enthusiasm and support has helped in a small way towards your own goals, I certainly enjoyed the experience and time spent with you.

Dedication

For my darling Angy, family, friends, Cairo and Teetee.



Contents

List of Figures

Candidate Statement

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university, and to my best knowledge and belief, this thesis contains no material published or written by another person, except where due reference is made in the thesis.

Abbreviations

AR	Action Research
BI	Business Intelligence
BPM	Business Process Modelling
BPP	Business Process Platform
CEA	Cost Effective Analysis
CEP	Complex Event Processing
CKDD	Conceptual Knowledge Discovery In Databases
CRM	Customer Relationship Management
CSV	Comma-Separated Values
DW	Data Warehouse
ERP	Enterprise Resource Planning
FCA	Formal Concept Analysis
FOL	First Order Logic
IOT	Internet of Things
IT	Information Technology
KDD	Knowledge Discovery In Databases
KDP	Knowledge Discovery Process
KM	Knowledge Management
KPI	Key Performance Indicator
LTA	Learning, Teaching and Assessment
OLAP	On-Line Analytical Processing

OLTP	On-Line Transactional Processing
PLM	Product Life-cycle Management
RDBMS	Relational Database Management Systems
RFID	Radio Frequency Identification
S&OP	Sales and Operations Planning
SCM	Supply Chain Management
SOA	Service Orientated Architecture
SQL	Structured Query Language
UI	User Interface
WF	Work Flow

List of Tables

Chapter 1

Introduction and Motivation for Research

1.1 Introduction

The aim of this research is to discover if hitherto hidden knowledge exists in transactional data and how it can be exposed through the application of Formal Concept Analysis (FCA).

With the emergence of service-orientated architecture and developments in business intelligence, data in its own right is becoming significant, suggesting that data in itself may be capable of capturing human behaviour and offer novel insights from a bottom up perspective.

Knowledge discovery is considered to be an active cycle of applying understanding and information. The traditional definition of knowledge as belief, trust and justification and how it is internalised in order to be used is discussed further in section 2.4. This leads onto the application of knowledge; essentially working with knowledge in section 2.5. Discovery is an action or process (Oxford, 2012) and represents the combination of systems and humans. This process uses interaction and visualisation to expose information and knowledge for consumption, reasoning and application in the search for hidden knowledge.

Humans exhibit innovation and intelligence but unlike computers are incapable of processing complex and large volumes of data at multiple levels. There is a need to connect the user's human-orientated approach to problem solving and decision making with the formal structures that computer applications need to bring their productivity to bear.

This chapter introduces the motivation for the research and establishes the research aim. The research approach is outlined, using enterprise systems as an exemplar for transactional data followed by an overview of the thesis. The chapter is completed with a list of prior work contributing towards the research.

1.2 Motivation

The motivation for this research came from observing shortcomings when implementing and supporting complex systems in modern businesses. Enterprise systems are frequently implemented to meet perceived or current requirements but a lack of flexibility hinders change and progress. As a result complexity increases with poorly integrated functions and systems that negatively impact performance. As stated by Plattner (2008) users of enterprise applications should be shielded from complexity. This does not mean that they should be limited in deriving value from systems and data.

The roles and capability of people and technology are constantly changing. Technology has enabled communication, mobile devices and computation power to integrate into everyday life with positive and negative impacts. Softer impacts include how intrusive these technologies can be impacting on both performance and well being. The primary focus we are concerned with is around the harder impacts. Increasing data volumes is challenging enough but it is further complicated by the divergence of formats, structures and distribution. When all factors are combined it is becoming increasingly difficult to source and analysis data effectively.

While change is inevitable, the demise of transaction based systems such as Enterprise System is unlikely in anything but the distant future, therefore research in this domain will not become out of date quickly.

Enterprise systems capture data in a transaction structure so that they can provide information that seeks to align with the knowledge that decision-makers use to achieve business goals. Traditionally this has been achieved by a top down approach whereby the business process is designed then the data is set according to that human-oriented model. However with the emergence of service-oriented architecture and developments in business intelligence, data in its own right is becoming significant, suggesting that data in itself may be capable of capturing human behaviour and offer novel insights from a bottom up perspective. The constraints of hard-coded top-down analysis can thus be addressed by agile systems that use components based on the discovery of the hidden knowledge in the transaction data.

Information is a key business resource that most business people do not recognise (Gordon, 2007). Organisations rely on information for decision making. Capture, storage and most importantly the management and analysis of information is often deemed technical and constrained to Information Technology (IT) departments. This makes connecting the right people with the right data in effective formats and contexts a key motivator.

The consumption and generation of data will enable greater capabilities and performance providing organisations can adapt and benefit from the potential agility offered by these systems. There is a need to connect the user's human-oriented approach to problem solving with the formal structures that computer applications need to bring their productivity to bear.

Data is generated endlessly and a proportion is captured by technology in various forms and locations. This data has a long life and can be repeatedly accessed with no change to the details. Conversely, humans rapidly forget detail and migrate towards a general knowledge of the process or system. Kant (1988) referred to this as the

conscious following of rules that are gradually cognised before they become so familiar that it requires great effort to think them in abstraction. A disadvantage of this is that knowledge can be static and not reflective of changes or developments.

Data must have unambiguous meaning (Gordon, 2007). Sharing and processing data rapidly fails if the means are not aligned. Semantics is about meaning, a simple definition is "semantics = data + behaviour " (McComb, 2004), this suggests that if the semantic content can be identified it may be possible to understand or determine behaviour. Organisations generate a large amount of data across multiple systems and organisations therefore this represents a significant challenge.

Initial thoughts suggested that any pre existing structures in the data are lost during the analysis. It is proposed to investigate if pre existing structures can be maintained while supporting the discovery of knowledge from transactional data. Furthermore, can this also include relationships between loosely connected systems.

Many real-world knowledge discovery tasks are both too complex to be accessible by simply applying a single learning or data mining algorithm and too knowledge-intensive to be performed without repeated participation of the domain expert (Hereth et al., 2003). Better support for knowledge discovery requires data that is both universal and based on practical conceptual models.

An environment that represents modern industry standard systems that supports these requirements and offers control and observation was developed using ERPsim (HEC Montreal, 2011) in an educational setting. This environment provided a framework for assessing student learning through iterative analysis and collaboration. Study of these experiences provided the research with data for evaluating knowledge discovery.

Knowledge discovery is a broad statement, in the context of this research it is constrained to transactional data, people, systems and analysis in a predominantly bottom-up approach, it starts with the data. This alone does not differentiate it from domains such as data mining, a principled method is required that enhances how people interact and view the product of analysis. Formal Concept Analysis has a number of

capabilities that support these functions and is therefore considered worthy of research with a strong emphasis on comparison and evaluation against contemporary techniques.

The primary focus of this research is the analysis of the data this, however, dependencies on technologies and enterprise systems exist. The hardware side of technology and enterprise architecture are peripheral to the main scope of this research but remain an important consideration.

Devlin (1997) states that humans have tried to represent knowledge and understand the laws of thought for thousands of years and that we are still unable to explain exactly how our minds perform such feats . He argues that "our minds are intimately intertwined with the world around us, and that our feelings and perceptions, even our social norms, play crucial roles in the marvellous complex dance of human cognition". The presentation and interactions with data and systems is highly influential and therefore should be a focus of this research.

Transactional systems support organisations in a variety of ways. These include storing and retrieving data, control mechanisms, workflow and reporting to mention a few. If one of the leading vendors is considered they, SAP A.G., consider that 65-70% of the worlds transactions are run on SAP (Clark, 2011). SAP have more than 102,500 customers in 120 countries which is predicted to equate to one billion users by 2015 (Brandt, 2010). To put this in proportion that data acquired and analysed in Chapter 5 represents approximately 35,000 transactions in a one week period for 12 users, considering the anticipated growth in the number of applications and transactions the volume of data will be very large to say the least.

There are numerous definitions of a transaction, commonly these include the act of buying or selling something, the action of conducting business, or an exchange between people (Oxford, 2012). All these definitions are applicable to this research and should be extended to include exchanges with or between systems. This leads to another definition from an enterprise system perspective, a transaction is a fixed sequence of actions with a well-defined beginning and a well defined ending (Plattner and Zeier,

2011). Transactions are a method of starting a function such as a report, data entry, browsing or virtually any other purpose. The use of many transactions in a defined sequence is a process with an aim of meeting the goals of the Enterprise.

Data is created, stored, consumed, transformed, and shared in many ways. A growing proportion is captured electronically however the distribution and range of formats is vast. One projection is that the production of data will grow 44 times by 2020, much of this data being unstructured and an important shift will be towards connecting data to reveal new insights (CSC, 2012). In additional to unstructured data such as audio or visual media, structured data will also expand from traditional relational databases to distributed and integrated systems.

Creating value from data is a fundamental requirements but also a major challenge without new techniques that address the complexity and volumes in a manner that compliments human and systems capabilities. Complimenting humans should be emphasised as replacing the brain is beyond the scope on this research, in fact it is almost beyond the scope of comprehension. Two significant projects are under way at the Allen Institute for Brain Science and Ecole Polytechnique F'd'rale de Lausanne with the aim of reverse engineering and documenting the brain respectively(Evans-Pughe, 2013). These are long term collaborative projects that are many years from reaching their targets, this is why a discover mechanism capable of deriving knowledge without or minimising human input forms the foundations of the research question.

1.3 Research Aim

The aim of this research is to discover if hitherto hidden knowledge exists in transaction data and how it can be exposed through the application of Formal Concept Analysis. This is valuable to the analysis, design and actual usage within enterprise systems. The following objectives are explored within this research:

1. To provide a focus through FCA applied to transactional data allowing an analyst

to discover hidden knowledge within enterprise systems.

2. To provide an approach for teaching FCA and elicit how FCA could be integrated into BI.

3. To provide an improved application of discovery techniques in transactional data, focussing on FCA and evaluated against alternative analysis techniques.

4. To enable knowledge sharing and reuse in order to deepen the understanding of transactional data and processes within enterprise systems.

5. To provide a understanding of knowledge derivable from transactional data and support a paradigm shift for system design.

1.4 Research Approach

As a strategy Action Research (AR) and Case Studies provide a participative and practical means of researching (McKernan, 1996) by capturing examples suitable for generating and testing hypotheses (Yin, 2009). In combination with individual practical experiments and research AR has been used because of its strengths in educational, social and organisation development to research through Learning, Teaching and Assessment (LTA).

An action research and case studies strategy iteratively developed an environment for generating experimental data representative of industry standards. Action research formed the overarching methodology that used case studies for specific sections of the research. Case studies formed a useful method for segmenting and structure the creation and collection of experimental data in an educational environment.

The first experiment utilised LTA and a hybrid combination of Yin's Case Study Method and Bigg's constructive alignment. This applied study and evaluation of FCA and contemporary tools formed cases studies while maintaining enquiry based learning

and a good pedagogic outcome. Table 1.1 summaries the changes implemented during this experiment as a result of iterations around the action research cycle, the case studies structure is also illustrated.

The secondly discovery experiment based on user activity logs from a real organisation applied and developed the analysis method. This data was not available in ERPsim but from the perspective of understanding human behaviour it was important to explore and develop an understanding of the hidden knowledge available. Building on the methods and techniques developed previously desktop research and experimentation applied a structured search for hidden knowledge.

All research has three main purposes (McNiff and Whitehead., 2009), AR has been applied to support these as follows:

1. *Creating new knowledge and making claims to knowledge:* By the developing an environment for applying FCA to transactional data in a formal LTA structure methods can be repeated and outcomes documented. Individual research also contributed and combined into this cycle.

2. *Testing the validity of knowledge claims:* The methods established can be repeated and findings compared iteratively as case studies in an educational environment and by peer review through publications.

3. *Generating new theory:* A combination of the above coupled with a action-reflection cycle, see figure 1.1, that is inherent in the AR method provides a strong foundation to improve understanding of events, situations and problems (McKernan, 1996) while provide intellectual rigour.

Action research links ideas with action, people communicate their ideas as theories of real-world practice, by explaining what they are doing, why they are doing it, what they hope to achieve and ultimately create living theories from personal experiences and theories (McNiff and Whitehead, 2006). Importantly AR requires reflection

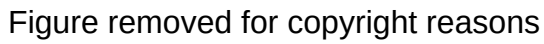


Figure 1.1: An Action-Research Cycle (McNiff and Whitehead, 2006)

(McKernan, 1996), a facet that LTA in particular will embody through the assessment cycle.

The techniques summarised in table 1.1 progressively develop the teaching and assessment methods, the output of iterative evaluation and reflection within the action research cycle. Techniques included learning in conjunction with ERPsim, a mix of individual and group work approaches and comparisons with alternative approaches. Refer to appendix E for a more detailed account.

The graph in figure 1.2 indicates how the assignment marks deviated from the average mark for each module for each case study. The changes implemented aimed to improve the overall teaching and paid particular attention to FCA as this featured a negative deviation. Taking case study 4 as an example, the students achieved higher than the average percentages for the introduction and lower for the FCA sections. The perfect line would run through zero with each student achieving the same percentage for each section of the assignment; as this is based on the average the result does not differentiate between high and low achieving students.

LTA has been used and integrated with individual research in order to discover hidden knowledge in transactional data by incorporating FCA into two degree modules

Case Study	1	2	3	4
Module	SA 2010-11	ES 2010-11	SA 2011-12	ES 2011-12
Average Mark	56.6	58.4	66.8	58.6
Standard Deviation	15.3	21	3.8	11.5
Data Preparation Demonstrated	X	X		X
End to End Data Prepared			X	X
Graphical presentation	X			
Document		X	X	X
Excel and FCA	X	X	X	
BI, Excel and FCA				X
Group discussion	X	X		
Group work		X		
Jigsaw based approach			X	X
Horizontal and Vertical Group work				X
Re-use (multi company)	X	X		
BPM Integration			X	X

Table 1.1: Chronology of Teaching Methods and Results

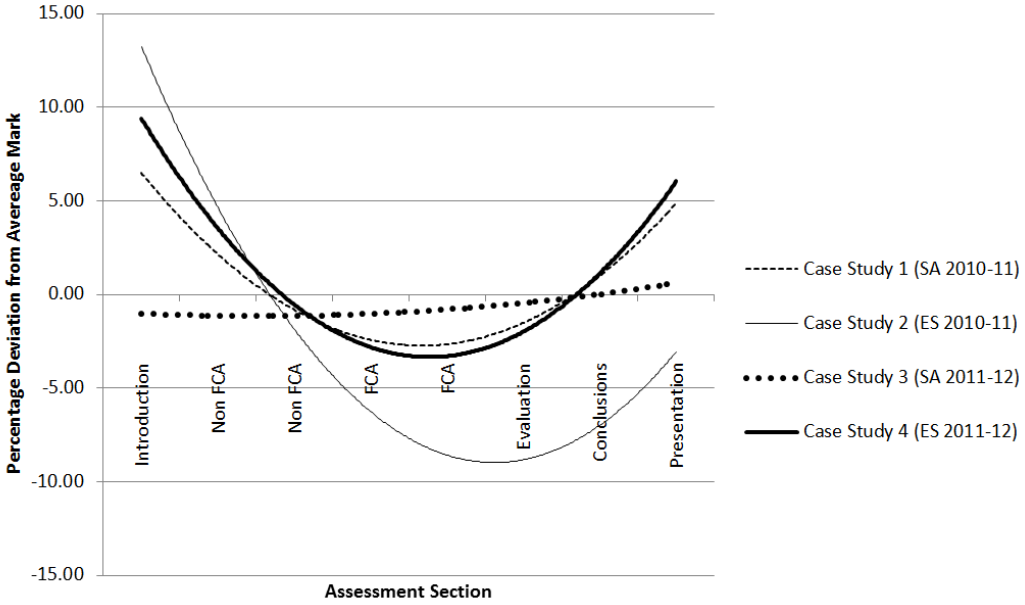


Figure 1.2: Deviation per Section from Student Average Mark

at Sheffield Hallam University. This provides a forum for discussion and a source of quantitative and qualitative data for evaluation. Semantic Technologies, graphical concept modelling and contemporary analysis tools are also included in the course content, these have been applied to the same research questions and contrasted with FCA.

The undergraduate 'Smart Applications' module applies FCA as a potential smart technology and aims to draw comparisons and integrate FCA within this context. The second module 'Enterprise Systems' is postgraduate and aims to forward the research subject area specifically within ERP systems and BI. Enquiry based learning is useful to learning from students who have no previous bias to FCA. An insight into how learning and applying good pedagogic methods was also envisaged for application outside of an educational setting.

A combination of Biggs' Constructive Alignment (Biggs and Tang, 2011) and Yin's Case Study Method (Yin, 2009) was used for the development of the learning environment, see figure 1.3. This is briefly described below refer to appendix E for a more

detailed account. While this required substantial effort and was necessary for the research it is secondary to the main content. It provided a framework for the generation of experimental data while supporting a good pedagogic outcome for the students.

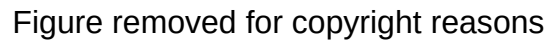


Figure 1.3: Combination of Biggs' Constructive Alignment (Andrews, 2011*a*) (Biggs and Tang, 2011) and Yin's Case Study Method (Yin, 2009)

Yin's method was used to capture and learn from a number of case studies, where each case study represents the respective modules' assignments. Case study methods can feature a number of disadvantages, McKernan (1996) identified that they are time consuming, suffer from researcher bias who can be 'taken in' by respondents as well as a lack of generalisation. Aligning with the degree courses represented a long duration however this suited the time scale of this research and provided rigour and reliable results. Bias and questioning the results had to be considered during the analysis and reflection periods. Generalisation with respect to FCA was considered to be addressed through the inclusion of alternative analysis techniques and challenging students to critically evaluate multiple approaches.

Yin method describes case studies as the preferred method for answering 'how' or 'why' questions where the investigator has little control over events and the focus is

within a real-life context (Yin, 2009). The assessment sets a goal for the students and targets the research aims while only guiding the actual analysis required. To this end the key themes of the assessment's design include graphical analysis, mandatory use of FCA and alternative techniques to contrast and explain how the techniques have been applied. Pedagogy also forms a significant proportion of the LTA cycle as it can significantly influence the outcome. In addition to the marks awarded for the assessments the capability of case studies methods to consider multiple sources of evidence is very suitable and therefore qualitative data including feedback, observations and the assignment contents have also been included.

Biggs' Constructive Alignment has two basic concepts; learners construct meaning from what they do to learn and that the teacher makes an alignment between learning activities and learning outcomes (Biggs and Tang, 2011). The combination of Biggs' constructive alignment and Yin's Case Study Method provides an overall method for aligning the learning activities and learning outcomes for the benefit of future students (Yin, 2009). It was also envisaged that an insight into the introduction of FCA into an organisation's business intelligence capability would be gained.

Siemens and Gasevic (2012) position learning analytics as currently sitting at a crossroads between technical and social learning theory fields . A real life context has been provided by applying ERPsim, this is a large scale business simulation based on Enterprise Resource Planning (ERP) enterprise system by leading provider SAP A.G. (SAP, 2012*a*), a global business software vendor. There is a technical learning aspect with interaction in competitive teams enforcing social interaction.

ERPsim features competitive behaviour and increasing levels of complexity in a highly immersive and demanding atmosphere that reflects industrial practice. It also has strong pedagogic foundations that have been adopted and applied during the development of the degree modules.

1.4.1 Ethics

Ethics in two primary areas was considered, these being the students partaking in the modules and the organisation providing production data. This was created following submission and acceptance of the research approach in accordance with Sheffield Hallam University's Research Ethics Policy, the statement presented is contained in appendix A .

Explicit permission was granted to use the students assessment data; the students were informed that their work was being used for this research both during lectures and in feedback as part of the module review process. The students attention focused on engaging in a business simulation that gave them the opportunity to experience real-life business decision-making and how data that reflects that experience is generated and its value, as if they were industrial practitioners. This is because the research focuses on how data semantics and knowledge may be applied in industry rather than educational research in its own right.

The students learning was intended to be strengthened from this simulated industrial experience enhancing their employability and future careers. All data was anonymised so that it cannot be traced to individual students.

Secondly, transactional data from an organisation using the same underlying SAP ERP system as ERPsim has been used as one source of data. To comply with ethics and privacy all user IDs and non standard identifiers have been replaced with generic labels. In addition permission to use, analysis and present this data for the purpose of this research was given in writing by the Customer Service Manager who was responsible for the data and users discussed in this research.

1.5 Overview of Thesis

Chapter 2 starts by considering some of the intentions and visions within transactional systems and associated analysis. Enterprise systems and business intelligence have been considered as exemplars for transactional data and analysis as they are widely applied in organisations and can be used as a data source and for comparison. Enterprises Systems provide organisations with a capability to capture, process, communication and analyse data in the pursuit of their objectives. System are evolving in both complexity and data volume resulting in an increasing challenge to derive and apply knowledge. There is a need for discovering knowledge though an analysis method capable of discovering relationships that enhances human capabilities whilst being congruent with system-based computation

Chapter3 provides an introduction to the theoretical foundations of Formal Concept Analysis (FCA) used in this research for discovering knowledge in transaction data, the core focus being on the application of FCA.

FCA provides a mathematical theory based on concepts; logical relationships that can be represented and understood by humans, essentially information and knowledge. The capability to analyse large data sets and discover relationships through tabular or graphical analysis provides a useful mechanism that can be applied to transaction data. The steps involved in applying FCA are described, starting from source data through to tabular and graphical lattice representation.

Chapter 4 provides an overview of the learning environment that supports the application of FCA in a situation where observation and evaluation can take place. It reflects both good pedagogy and industrial practice through the use of ERPsim. This large scale, real-world business simulation software is based on the SAP ECC, an enterprise system by global business software vendor SAP A.G.

Drawing upon empirical analysis of assignment material over iterations of the teaching cycle, a range of qualitative analysis methods utilised NVivo to manage data and generate ideas. Querying, modelling and reporting is described including the theories and conclusions developed. Results include the knowledge discovered from transactional data and an assessment of FCA's ability to explorer complex systems.

Chapter 5 explores how the application of FCA as a discovery mechanism to user transaction logs offers an insight into the actual patterns of use. User transaction logs are frequently overlooked as a source of data even though they offer a rich but complex source of data. The data set was captured from a real system carrying out its normal operations.

Chapter 6 combines a mixture of qualitative and quantitative analysis to consider and reflect on the research question, the discovery of hidden knowledge in transactional data. This focusses on the discovery process with actual knowledge discovered as supporting evidence. Quantitative analysis is used to highlight patterns and answer if FCA is capable of helping in the discovery of hidden knowledge. Cost Effectiveness Analysis (CEA) has been applied to understand how and where FCA can add value. Finally reflection is used to gather and consider the requirements for successfully applying FCA as a method for knowledge discovery.

Chapter 7 concludes to what extent the aims of the research criteria have been addressed. Applications for the research are discussed along with the effectiveness of the research approach. Contributions to the research are described along with identifying further areas of research in this field.

Chapter 2

Discovery of Knowledge in Transactional Data

2.1 Introduction

This chapter introduces transactional data and analysis techniques in the context of enterprise systems and knowledge representation. The core components of enterprise systems including business intelligence (BI) solutions are discussed and evaluated. The capabilities of enterprise systems are highlighted as are the competitive forces creating demand for increased integration, intelligence and efficiency of use.

Understanding the dynamics of an Enterprise is difficult due to the complexity of applications and technology, increasing data volumes and the fact the competition forces constant change. The relationship between data, information and knowledge in the context of enterprise systems is discussed and consideration given to shielding users from complexity as far as possible (Plattner, 2008).

Enterprise systems capture data in a transaction structure so that they can provide information that seeks to align with the knowledge that decision-makers use to achieve business goals. With the emergence of service-oriented architecture and developments in BI, data in its own right is becoming significant, suggesting that data in itself may be capable of capturing human behaviour and offer novel insights

Discovering knowledge through combining the processing capability of machines in a form that enhances human capabilities is discussed along with current Business Intelligence solutions. Knowledge is a valuable but expensive commodity. It is delicate, easily lost or misunderstood and difficult to gain; moreover it is hard to computerise. Humans exhibit innovation and intelligence but unlike computers are incapable of processing large volumes of complex data. There is a need for an analysis method capable of discovering relationships that enhances human capabilities whilst being congruent with system-based computation.

2.1.1 Overview

This chapter considers the background, intentions and visions within enterprise systems starting with transactional systems and approaches to analysis. A review of the differences between data, information and knowledge is conducted before discussing the human aspects of understanding and deriving value from discovered knowledge.

Section 2.2 introduces transactional systems, data and analysis in the context of enterprise systems. These are important business tools that represent the real world. The application of transactional based systems will continue for the foreseeable future as they provide integrity, detailed control and a historical repository. The construction and form of enterprise systems is expected to change as data is captured from an increasing array of sources and competitive forces demand higher performance. Technologies and approaches as described in section 2.3.6 will enhance the capabilities and correspondingly the complexity of transactional systems but consequently making understanding them more difficult.

Section 2.3 introduces Business Intelligence as a collective term for obtaining, analysing and distributing information and knowledge. BI is useful for collating large volumes of data from multiple sources and as an environment for applying mathematical calculations and producing visual outputs. There is a need for bottom-up approaches

that can analyse data generated by agile systems and recognise the significance of human behaviour.

Section 2.4 introduces types of data and processes for transforming data into information and knowledge. Through computation, representation, interaction and ultimately human thought data has an inherent value that can be exploited within the context of complex systems and processes. The challenge and context of this research includes the management and effective analysis of large complex data sets for efficiently exposing information and generating knowledge. Knowledge, discussed further in section 2.4, being the ability to internalise [learn] and use information.

Section 2.5 introduces discovery techniques and the importance of knowledge for human and organisational performance. Discovery is an important factor as data is expanding in complexity and volume. Approaches that connect the user's human-oriented approach to problem solving with the formal structures used by computer applications are needed to bring their collective productivity to bear. The final challenge involves reasoning and applying discovered knowledge, making use and deriving value from the effort.

2.2 Transactional Systems

Information systems is a generic term associated with systems that manage data by providing processes and information. They are developed and operated within an environmental context that has a significant effect on them (Avison and Fitzgerald, 2003). Transactions are discrete functions typically used to interact with data and perform calculations, updates, trigger events and many other functions.

Data represents unstructured facts about events, objects and people. When strings of data are associated they can be used to give information. Add a context and they form a basis for decision making Avison and Fitzgerald (2003). Fundamentally this is why information systems in various forms support organisations around the world.

An aspect of system theory is that organisations are open systems, they are not closed and self-contained, therefore the relationship between the organisation and its environment is important, of particular note is the human element (Avison and Fitzgerald, 2003). Information Systems are a representation of the real world, an abstract or model of a process. When implemented at the correct level they can provide a simplified and focussed viewpoint without introducing inaccuracies due to insufficient data. Conversely when implemented incorrectly they lack data or precision leading to incorrect information or misunderstanding.

Enterprise systems are a type of information system typically offering organisations industry specific and best practice functions to support common processes while concurrently reducing the need for technical software and hardware skills. They provide organisations with a capability to capture, process, communication and analyse data in the pursuit of their goals.

Enterprise systems are complex and technology trends suggest that this will continue to grow as the volume of data stored and communicated increases. Systems will need to communicate between themselves, with objects and humans based on loose connections while maintaining the necessary levels of context, trust and reliability.

A significant problem is that market forces make constant change inevitable; systems must be adaptable to meet the information needs of the Enterprise. Mobile devices are now capable of communicating audibly and visually almost on par with any fixed location device. They are also capable of functioning, sensing and communicating without human input, therefore, the potential range of information available to an Enterprise is creating significant challenges around handling and more importantly benefiting from this new environment. ERP is one of the main constants as it traditionally represents the core processes of an Enterprise and forms a central repository to support functions such as the financial reporting. Any new method of analysis requires a framework that is equally flexible to these needs.

It is virtually impossible to define the range of processes that enterprise systems

support as they are both configurable and programmable however there are broad categories. The most common being ERP as discussed above. Similar platforms exist for specific sectors including Customer Relationship Management (CRM), Supply Chain Management (SCM), and Product Life-cycle Management (PLM) that span a range of industries from Aerospace to Wholesale. This is not intended to be exhaustive list but indicative of how products are developed for specific industry sectors and importantly how frequently these are not stand alone systems but integrated to provide an architecture that supports the requirements of an organisation.

Data is one of the fundamental reasons why enterprise systems are deployed. Enterprise systems enables users to share data and information, companies to reduce costs and manage business processes (Aladwani, 2001). The type of data is varied, normalised data in relational database tables is an significant proportion with typical systems running into 100,000 standard tables. These tables contain tens of fields with defined attributes including data type, length, formatting and in some cases links. These links can be to internal long text or external data because, using of SAP ECC as an example, field length is limited to 255 characters. Internal links can represent links to models or documents. External links can integrate with, for example, documentation management systems, URLs or services. Given this flexibility and ability to link date is a complex domain.

Enterprise systems is an umbrella terms for many different types of systems and communication technologies. For consistency Enterprise systems has been used throughout as a general term. Enterprise resource planning (ERP) and business intelligence (BI) are considered sub components, these a discussed in the following sections.

2.2.1 Enterprise Resource Planning

Enterprise Resource Planning systems are prevalent in industry and provide a core transaction based system typically referred to as On-Line Transactional Processing

(OLTP). Based on Relational Database Management Systems (RDBMS) they have been employed since the 1980s to process operational data (Plattner and Zeier, 2011) for organisations around the world.

ERP systems have revolutionised business around the globe; processes are leaner and more efficient, costs are minimised, positive customer service is more prevalent, and government compliance is present (Dunaway and Bristow, 2011).

ERP systems are essentially transactional systems that support a vast array of business functions within the majority of organisations that exist today. They are designed to be explicit and accurate in terms of control and data but often lack the analysis tools and communication methods to meet all functional requirements. This is where value can be added by integrating tools and service.

ERP systems support integration and control across various functional areas of a company, therefore supporting the achievement of the company's plans (Portousal and Dunderam, 2006). ERP is an excellent source of raw data in a relatively well defined format and structure, however the volume and granularity of the data make analysis inefficient or inadequate without the application of BI tools.

Organisations invest significant resources into systems during implementation and through ongoing maintenance and use. These systems are used to control operations, integrate with business partners and should be leveraged to attain any competitive advantage possible.

Transaction based systems such as ERP systems have been relatively static when compared to web based systems for two primary reasons:- standardisation and maintenance. ERP systems are typically internal systems that are not heavily branded, feature standard screens and have limited navigation aids that enhance the user's experience. This is a very different picture to web sites where the user experience can be a fundamental success factor. Secondly, due to the complexity of modern systems, few organisations have the skills or knowledge to support or maintain such systems. Therefore the development of road maps and upgrade cycles by specialist vendors enable even

small organisations to benefit.

Modern Enterprises are complex and rely heavily on people and electronic systems to control and manage their operations. ERP systems are central to many enterprises as they provide an integrated and best-practice set of processes coupled with control and governance. A significant problem is that market forces make constant change inevitable; systems must be adaptable to meet the information needs of the Enterprise.

Mobile devices are now capable of communicating audibly and visually almost on par with any fixed location device. They are also capable of functioning, sensing and communicating without human input; the potential range of information available to an enterprise is creating significant challenges in handling and more importantly benefiting from this new environment.

The predominant trend within ERP solutions has been for process experts to design and architect solutions in a top down manner, in the worst cases with a silo viewpoint. This tenancy to design processes and reporting solutions manually makes it difficult to conquer the challenges of increasing data volumes, process diversity and the range of interactions.

ERP systems provide a transactional capability that forms a fundamental platform upon which the majority of today's organisations operate. ERP provides a detailed and structured mechanism for controlling and capturing operational data and a platform for analysis. ERP is not to be considered as an isolated system, in practice they form part of a complicated architecture communicating and interacting with many other systems. One important system aspect of this is Business Intelligence (BI); by definition this provides decision makers with valuable information and knowledge by leverage a variety of sources of data as well as structured and unstructured information (Sabherwal, 2007).

ERP systems are typically based on a relational database, certainly this is where the majority of transactional activity is captured. Data is normalised to minimise redundancy and remove ambiguity which makes it useful for analysis, however, a great deal

of process logic is embedded within the ERP system and not the database. Relational databases when used for on-line transactional processing (OLTP) are good repositories for detailed information. BI solutions typically transform this data so that it is suitable for on-line analytic processes (OLAP) by converting data into a format that is more applicable for fast analytic applications, frequently at a level where granularity is reduced. This focussed and efficient analysis tool comes at a cost in terms of transforming the data and maintaining its meaning, particularly when data is consolidated across systems or geographical areas.

ERP offers a relatively rigid set of data in a well-structured format, its operation relies on programmed logic that is not necessarily represented in the data. In addition to documents, objects, statuses and relationships ERP systems also capture a variety of log files including user tasks, time stamps and changes. All these in combination will potentially reveal otherwise latent semantics that can be of benefit to the Enterprise and form part of a successful BI application. ERP systems do not typically incorporate semantics and the stored data represents only a proportion of that available within an organisation. Non-integrated but complementary systems and humans form the repositories where the majority of data is stored. Correspondingly the majority of control mechanisms and procedures are not contained directly in the data but encapsulated in programs or dictated by human interaction with the system.

Figure 2.1 illustrates the basic logical layers of a business system (Plattner, 2008). The layers have distinct functions with connectivity and services exposed for further applications. The three core layers are the user interaction, business logic and persistence layer. User interaction is the presentation layer and interface with the user. Business logic forms the core and contains the rules, functions or relationships that define and controls the enterprise. The final layer is the persistency layer that physically stores and manages data. In practical terms this picture may be repeated numerous times within an organisation to support various functions. This environment is starting to represent an integrated platform as depicted on the right hand side of figure

2.2. Application that specialise in and serve specific functions such as Supply Chain or Customer Relationship Management are integrated to provide the platform necessary to support business operations.

Abstraction between layers forms a major factor in the development and application of Enterprise systems. An applications developer does not need to understand how to write to a hard drive in the database, a level of abstraction exists that enables the task to be completed without detailed and expert knowledge.

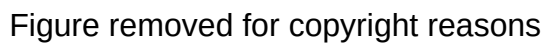


Figure 2.1: Structure of Business Application (Plattner, 2008)

Enterprise systems have developed from two tier platforms in the 1970s to web services with exposed functionality via standard access protocols, see figure 2.2. Various drivers and enablers have driven these changes including competition, technology, users and the growth of the Internet. Access to information is faster and more readily available than every before, none more so than for consumer who can use these technologies to compare and select with relative ease. This in turn drives competition and in turn generates forces that drive improvements in speed, cost and performance.

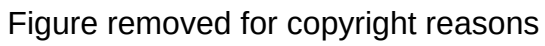


Figure 2.2: Evolution of Enterprise Application Platforms (Plattner, 2008)

2.2.2 Transactions and Relational Databases

Regardless of the form that ERP takes the core function and method of interaction is a transaction. A transaction is a fixed sequence of actions with a well-defined beginning and a well defined ending (Plattner and Zeier, 2011). The use of many transactions in a defined sequence is a process with an aim of achieving the goals of the organisation.

The basic method of user or system interaction within an enterprise systems is through processing a transaction. Transactions are a method of starting a program that performs a function such as a running a report, data entry, browsing or virtually any other use. In practical terms many transactions are processed by many users or systems thereby supporting the operations across an entire organisation.

The transaction concept guarantees integrity for all concurrent users (Plattner, 2008). The key properties of a transaction are coined as the acorynm ACID, this stands for Atomicity, Consistency, Isolation, and Durability (Haerder and Reuter, 1983). Transactions defined at this level are detailed but useful from an an analysis perspective as they have definite states, changes are limited and because of this

a second transaction is required to reverse or cancel. In order for a transaction to access data it performs a dialogue step, this essentially requests information from the Database Management System. Importantly this is stored and accessible for analysis purposes.

The relational database model was conceived by E. F. Codd in 1969 using the terms relations, attributes and tuples, they are more commonly referred to with the terms tables, columns and rows (Litwin, 1994). Data is organised in tables as rows and columns with relationships between tables. This creates a powerful and efficient structure for storing large amounts of data and importantly with Structured Query Language (SQL), a method of managing data. An example of these relationships is shown in figure 2.3. Table and field relationships are linked, depicted by the black line between boxes. This is only a small example, database frequently utilise 10,000s of standard tables and even more secondary tables. These are used for various storage and performance reasons including indexing, sorting and hashing.

In addition to table relationships data can also be joined through document flows (in SAP terminology) or within the program itself. These work in a similar manner linking one document to another, eventually linking all documents created along a business process. A business process is defined as a set of logically related tasks performed to achieve a defined business outcome (Davenport and Short, 1990).

For detailed control and processing transactional data is highly appropriate, however, aggregating large volumes of transactional data can be problematic and slow. This is where BI solutions have advantages.

2.3 Business Intelligence

The term Business Intelligence (BI) can be interpreted differently, sometimes as the product of the process and as the process of obtaining, analysing and distributing information and knowledge (Sabherwal and Becerra-Fernandez, 2011). BI is the ability to

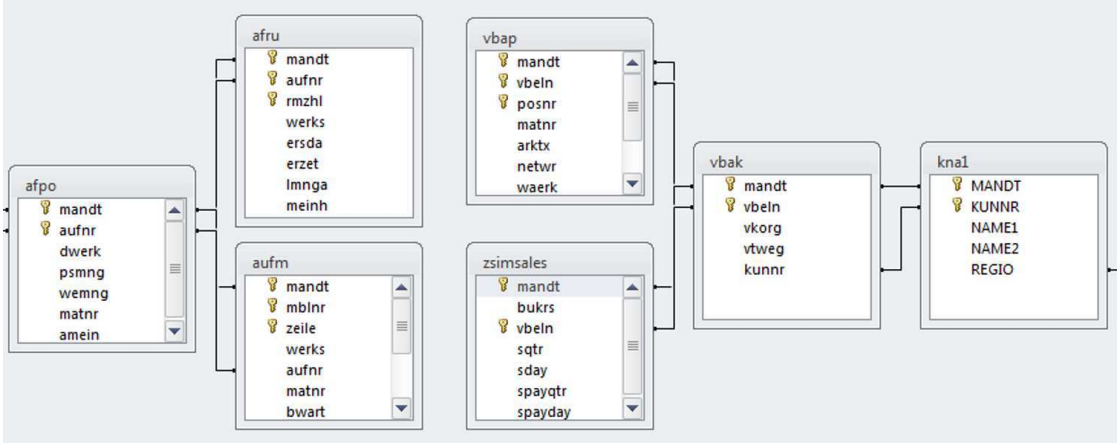


Figure 2.3: Table Joins in a Database

analyse data for decision making purposes using computer-based techniques (Dunaway and Bristow, 2011). Another view is that the goal of BI is to make sense of change, to understand and even anticipate it (Soukup and Davidson, 2002).

BI provides decision makers with valuable information and knowledge by leveraging a variety of data sources including structured and unstructured information. The information and data can reside within or outside the organisation, be obtained from multiple sources, structured in different ways, and be either quantitative or qualitative (Sabherwal and Becerra-Fernandez, 2011).

BI is a set of processes (data gathering, data analysis), technologies (data warehouse), and presentation tools (report generator, dashboard) used by organisations to analyse data (either internal or external) in order to gain new insight on their environment (customers, suppliers) and make better decisions (Mireault, 2011).

Online analytical processing (OLAP) has provided a BI capability to decision making for over twenty years. Initially this required specialist skills, however, advances in software produced simpler graphical interfaces for use by non-technical business users. OLAP is fundamentally based on the same data source as OLTP, however, it supports multidimensional data analysis (Codd et al., 1993). Key drivers behind the success of OLAP have included its flexibility and the ability to 'slice and dice' views of information quickly although it does tend to be an historical analysis tool. Terms such as star

schema or cubes are used to describe aggregates of raw data against key characteristics, these summary levels represent data at a granularity highly suitable for repetitive and fast analysis hiding the details of thousands or millions raw data inputs.

CUBIST, a programme funded by the European Commission aimed at combining the essential features of Semantic Technologies, Business Intelligence and Visual Analytics, argues that the complexity of BI tools is the biggest barrier to successful analysis, particularly because they do not work with the meaning of data (semantics) and are not capable of effectively handling unstructured and structured data (CUBIST, 2010).

Transactional systems provide a core function and support considerable business activity within organisations around the world. They are intrinsically complex and significant effort is required to understand and manage them effectively. Based on the current outlook, system landscapes are evolving and becoming more flexible and agile. Therefore, analysis techniques must follow suit.

Sabherwal and Becerra-Fernandez (2011) describe four key capabilities of BI solutions as shown in the left hand side of figure 2.4 . The distinct capabilities build from data and information through analysis before creating new knowledge with the user at the top of the diagram. The right hand side indicates where Formal Concept Analysis and other technologies could fit in this chain of capabilities. It is clear that any analysis of data will share the inherent problems with respect to data, integration, interpretation and knowledge creation.

BI has many proven applications, holistically it targets the analysis of large data sets with mathematical methods or presentation tools (Mireault, 2011) in order to identify or visualise patterns, classes and characteristics. It is for these reasons BI is considered an important part of enterprise systems as it supports decision making at many levels within an organisation.

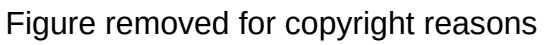


Figure 2.4: Business Intelligence Capabilities after Sabherwal and Becerra-Fernandez (2011)

2.3.1 Business Intelligence and Knowledge Management

There are a number of technologies that share a similar space but have a number of notable differences in terms of the data, information and knowledge they consume and produce. Table 2.1 describes the attributes of technologies designed to support decision making and aid the management of data, information and knowledge.

Starting on the right hand side with DSS and ADS the target users are highly focussed implying the need for specialist skills and knowledge. Consequently these approaches are limited in scope to a small demographic of the total user population. BI on the left hand side is accessible and useful to a much larger demographic using flexible and interactive tools for converting data and information into further data and explicit knowledge. Explicit knowledge is detailed and clearly stated and typically represents actual values, processes and calculations. Section 2.4 expands on the description of knowledge.

What is apparent is that the lines between all of these technologies are blurring as technologies are converging and blending technologies in order to advance. Herschel and Jones have similar views and argue the BI should be a subset of Knowledge Management

(KM) as it can influence the very nature of BI (Herschel and Jones, 2005). This is because KM includes tacit knowledge that is difficult to represent within BI although they clearly state that BI and KM need to be considered together as integrated and mutually critical components.

This research is aiming at a space that includes aspects from BI, Knowledge Management and Data Mining for discovering explicit knowledge and potentially generating tacit knowledge that is accessible to a wide user base while being capable of analysing large data sets. It is expected that tacit knowledge will only be evident from the practical and collaborating aspects of the learning exercises as discussed in chapter 4. It is not expected that tacit knowledge will be defined and documented in a written format.

	Business Intelligence	Knowledge Management	Data Warehousing	Data Mining	DSS or ADS
Inputs	Data, information	Data, information, knowledge	Data (from multiple systems)	Data	Data, information, knowledge
Nature of Inputs	Internal or external, structured or unstructure	Internal or external, structured or unstructured	Internal structured	Internal structured	Data, information, knowledge
Outputs	Information and explicit knowledge	Tacit knowledge and explicit knowledge	Data (in a single logical repository)	Information	Decision recommendation (DSS) or automated decisions (ADS)
Components	Information technologies	Information technologies, social mechanisms, structural arrangements	Information technologies	Information technologies	Information technologies
Users	Across the organisation	Across the organisation	IT personnel	IT personnel, others trained in IT	Specific, targeted users

Table 2.1: Distinctions between BI and Other Related Technologies (Sabherwal and Becerra-Fernandez, 2011)

To be reliable the architecture must be clearly defined and stable for data warehouses to function in production environments (Sabherwal and Becerra-Fernandez, 2011). In the new agile world an approach that has a similar degree of agility is required.

BI solutions are heavily dependent on enterprise architecture (Sabherwal and Becerra-Fernandez, 2011). This includes the databases available and technical capabilities. Viewing this from an enterprise architecture perspective also includes the business,

data, applications and technology. Data will inevitably come from external sources therefore augmenting data from sources such as vendors, service providers and environmental scanning must also be considered under this umbrella. Figure 2.5 contains a number of areas to consider when designing reports. Reports has been used rather than BI or other terms so that there is no predetermination of the technology or application. Areas range of the target user group, the frequency and delivery method, data structures and data sources. The core message is that a range of factors will determine the requirements and therefore solutions will differ with need.

Figure 2.5: Reporting Consideration

2.3.2 The Need for Business Intelligence

Sabherwal and Becerra-Fernandez (2011) cite four reasons for the increasing demand for BI, these are exploding data volumes, increasingly complex decisions, need for quick reflexes and technical progress. They also cite the four benefits of BI as the dissemination of real time information, creation of new knowledge based on the past,

responsive and anticipative decisions and improved planning for the future.

A number of factors contribute towards the need for BI, however, the underlying dynamics have been present for many years. Porter's Value Chains illustrates the relationships between primary and supporting activities for achieving competitive advantage (Porter, 1985). BI can support this by targeting the generic sources of competitive advantage including a focus on cost, reconfiguring the value chains and creating differentiation or uniqueness.

Value chains coupled with technical advances emphasis why data volumes, decision complexity and the need for quick reflexes are critical for organisations today. The integration of systems with systems, sensors, mobile devices and the availability of information coupled with powerful analysis can certainly contribute towards achieving competitive advantage.

2.3.3 Challenges for Business Intelligence

BI can be complex, only a proportion of the data is available, intervening variables called tacit knowledge (Herschel and Jones, 2005) such as culture, bias or conflicting goals can render the results inaccurate. It is viewed as being expensive and business events are not consistently defined throughout the organisation which makes it difficult to utilise organisation wide BI (Sabherwal and Becerra-Fernandez, 2011).

Cook and Cook cite BI's inability to integrate non quantitative data into its data warehouses and relational databases as a major limitation also quoting that up to 80 percent of business information is not quantitative or structured in a way that can be captured in a relational database (Herschel and Jones, 2005) .

2.3.4 Aims of Business Intelligence

BI should help create new tacit knowledge. One way of doing this is by utilising multiple sources of data. Through integration it is possible to create a process that combines

several sources of explicit knowledge into new patterns and relationships (Herschel and Jones, 2005).

Perkin's theory of understanding suggests that a knowledge of the aim, relationships and purpose of the analysis is required to convert a new relationship into tacit knowledge (Herschel and Jones, 2005). This tacit knowledge about an organisation in the context of analysis and decision making starts to encroach onto Knowledge Management topics including organisational memory, strategic alignment and architecture.

2.3.5 Data Mining

Data Mining as a term is generally considered a facet of BI along with are artificial intelligence and machine learning although a definitive demarcation is difficult. Visual Data Mining has benefited enormously from the growth in computation power and graphical capabilities. Recent trends include increased availability, dynamic interaction, complex data visualisation methods and the role industry standards play in the exchange of data (Soukup and Davidson, 2002).

Cios et al. (2007) describe the aim of data mining as making sense of large amounts using mostly unsupervised data in some domain . Firstly the term 'make sense' refers to a non trivial knowledge discovery processes (KDP), the output must be understandable, valid, novel and useful. Organisations are regularly storing and analysing terabytes of data. Quantifying exactly what 'large amounts' represents is highly subjective and topics such as Big Data as discussed in section 2.3.6 have emerged. What is clear is that handling and analysing these volumes is beyond the capabilities of humans without assistance from machines. This highlights the third point about 'mostly unsupervised' that has drivers for lowering resources requirements and costs. The final term is 'domain' this indicates the need for a highly interactive and iterative process for discovering new knowledge from data.

KDP consists of a series of steps executed by practitioners when executing a knowledge discovery process (Cios et al., 2007). These are illustrated in figure 2.6 and represent iterative steps in the pursuit of knowledge generation. This high level view is generic and deliberately lacks details that alternative KDP models include, specifically those actually applied and related to practical or industry. These details relate to topics such as business and data understanding in the context of the domain, data preparation and deploying data mining; these practical topics are address as part of the methods discussed in chapters 3 and appendix E.

Figure 2.6: Data Mining (Cios et al., 2007)

Data mining is an umbrella term but the key feature is that it is data driven as opposed to other methods that are often model driven (Cios et al., 2007). Figure 2.7 represents estimates for the relative effort associated with steps in the knowledge discovery process. Data preparation is decidedly the most time consuming activity with all remaining steps representing only a quarter of the relative effort.

Figure 2.7: Relative Effort Spent of Specific Steps of the KD Process (Pal, 2005) cited in (Cios et al., 2007)

Data mining techniques range include classification, clustering, association, and decision trees. These are briefly described as they have been introduced as a comparison

to FCA. The collection method for this data is discussed in appendix E. FCA can be considered as an association or classification technique and is discussed in more detail in chapter 3.

Clustering is the process of organising objects into groups whose members are similar in some way (Lingras and Akerkar, 2008). Figure 2.8 shows a simple two dimensional set of data containing three clusters produced by k-means, a heuristic algorithm that converges towards an optimum solution.

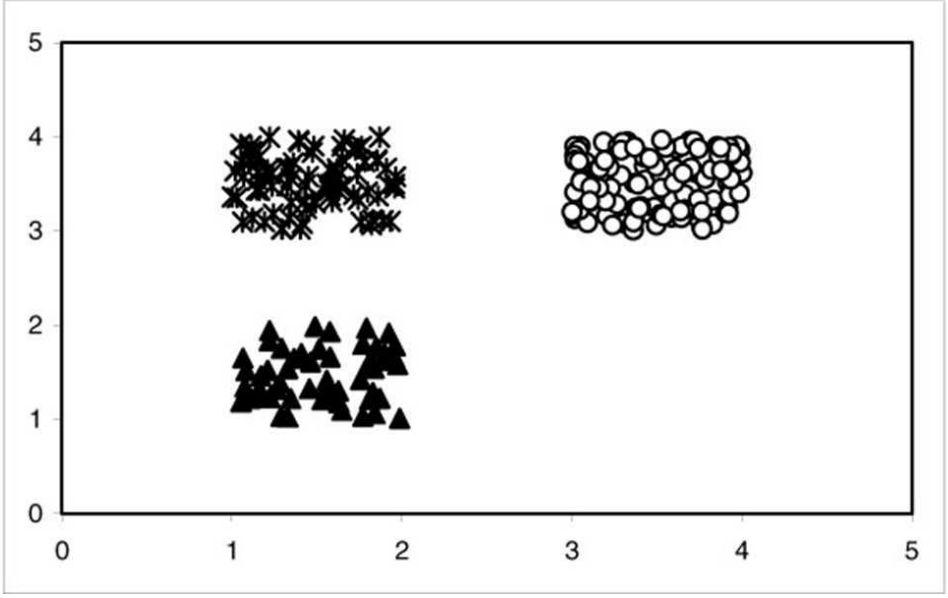


Figure 2.8: Simple Clustering Example

Classification is similar to clustering with the distinction being that classification requires the classes defining before the analysis. The goal of classification is to build a concise model (rules) that can be used to predict the class of the records whose class label is not known (Lingras and Akerkar, 2008).

Decision trees are classifiers in the form of a tree structure, see figure 2.9 for a simple example. Each node is either a leaf or a decision node, decisions move between classes towards an end point. In a similar manner to neural networks training sets can be used to create the structure.

Association is used for discovering relationships within large data sets. A common

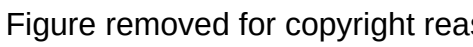


Figure 2.9: Decision Tree (Lingras and Akerkar, 2008)

example is a market-basket analysis; finding relationships between several items within baskets (Lingras and Akerkar, 2008). Forming rules is a useful output that represents knowledge derived from the data.

Discovery as a term implies a number of ideas. Primarily in the context of this research it is about the act or process of finding hitherto unknown knowledge from transactional data. This could range from completely automated functions through to application that promotes learning through interaction and interrogation.

Combining these discovery techniques with search, linked data, semantics and meta data on platforms that include integrated views or mobile devices has the potential for creating powerful applications.

2.3.6 Future Directions

This section refers to recent and future directions in enterprise systems. Real organisations will actually represent a cross section of states as they develop and apply technologies at different rates.

It is fair to say the humans are dependant on systems and tools, in one form or another this has been true for thousands of years but more recently systems and

communication have grown in prevalence. The Internet has experienced massive growth in the past decade and enterprise systems are starting to share many parallels with web interaction (Plattner and Zeier, 2011).

The challenge is to leverage existing systems and data as well as utilising new technologies in order perform better. This has many dimensions but includes topics such as efficiency, integration, decision making, communication and agility.

Plattner states that many customers do not make substantial changes to structures [Enterprise architectures] because they can't predict what effect it will have, he even argues that many are not in a position to master their own landscapes in the first place (Plattner, 2008). Approaches to simplify these complex landscapes risk further longer term complications and imposed limitations when trying to consolidate or integration functions.

Some of the most significant domains from the perspective of this research are highlighted and discussed to the level appropriate for this research. It is beyond the scope of this research to define and document all aspects of these topic. The review ranges from technology, human interaction and decision making, architectural approaches and the representation of data, information and knowledge.

Service-Orientated Architecture (SOA) represents a concept that unites operational business aspects and system architectural aspects thus building a bridge between the business world and IT (Hack and Lindemann, 2008). Business functions are encapsulated and linked together using loose coupling of services for mapping business processes. The significant advantages are flexibility and adaptability where processes can be built by reusing existing encapsulated functions.

SOA makes it possible to integrating ERP and the interoperability disparate systems with abstraction between layers or systems, see illustration in figure 2.10. With SOA it has become possible to distribute functionality across various local and remote systems and enables more flexible and agile applications changing with the demands

of the customer (Plattner, 2008). This distinct capability can address the competitive forces imparted on modern enterprises. Controlling, governing, reporting, even understanding these systems remains a significant challenge.

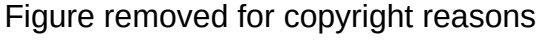


Figure 2.10: Service-Orientated Architecture (Krafzig et al., 2004)

Business Process Platform (BPP) is a model driven development process (Plattner, 2008) and an example of Businesss Process Modelling systems that combines with SOA to form a powerful process design environment. These tools support process design in a flow chart format and the technical capability to create the programs necessary to operate and integrate with other systems.

Business Process Management (BPM) is an approach to modelling processes that can address variations and logic problems. It can accomplish this because it incorporates a mix of metaphors including decisions and events into the model. A metaphor is a way of reducing the dimensions of the descriptions of a process to a more understandable and visible basis (Debevoise and Geneva, 2008). There is no argument against the need for metaphors, however, determining them and communicating them without loss of or mis-understanding is a difficult tasks as modelling is only an abstraction of a real system.

Approaches such as BPM provide enterprises with more responsiveness through flexible and graphical modelling tools. These tools support process design in a flow chart format and the technical capability to create the programs necessary to operate and integrate with other systems. These approaches have the capability to address the needs of the Enterprise however there is a gap - the ability to understand and analysis highly variable processes. This highlights the need to employ techniques that discover information and semantics in order to gain knowledge, insight and pro-actively apply it.

Robustness is a term applied to transactions as discussion in section 2.2.2. This is equally applicable in the domains of BPM and SOA; changes in state through the process must be consistent and defined.

The underlying hardware is constantly evolving. Developments in physical storage devices, in-memory storage and the cloud have an impact on how and where humans and systems interact. These technologies may not change the fundamental principles and processes, however, step changes in volumes, speed, accessibility are certain to occur.

Figures 2.11 and 2.12 illustrate how in-memory technology may transport the landscape that supports ERP and BI. Currently ETL (Extract-Transform-Load) is require to move and convent data between these systems. Typically this requires batch processing and limits two directions movement. The future landscape uses the speed of in-memory technology and processing within the database to enable a step change in analytic capabilities.

A major driving force is for platform independence even as users are pushing for the right to choose and even supply their own hardware(Graham, 2011) (Age, 2013) (Forbes, 2011). This has multiple effect including working practices, greater distribution of information, unknown process interaction that includes social collaboration.

Complex Event Processing (CEP) is about responding to events as they occur or as soon as the data is received. This is quite different to the relational databased model

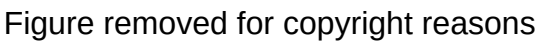


Figure 2.11: Current ERP and BI Landscapes (Muller, 2013)

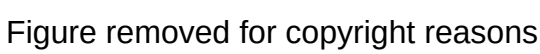


Figure 2.12: Future Landscapes ERP and BI based on in-Memory (Muller, 2013)

where data is first stored before being analysed. The key differences is an 'events processor' that takes incoming messages and runs them through a set of pre-defined continuous queries to produce derived streams or sets of data (Taylor, 2012). CEP appears to offer a platform for integrating FCA in the future but not useful in the context of the research aim.

The Internet of Things (IoT) is a term that has its roots in various network, sensing and information processing approaches (Friess, 2012). Connectivity, smart applications and hardware mobility are three factors that offer the potential to change how organisations and individuals operate. Portable and connected devices enable far greater levels of interaction between humans and systems. In addition to this remote and autonomous devices have become sources of data. Technologies including Radio Frequency Identification (RFID) and sensors are collecting data that ranges from simple linear scales i.e. temperature, geographical through to audio and visual streams.

Query languages are available that define relationships, also known as associations, between objects (Jung, 2013). Relationships between objects such as structures and tables can be defined and reused therefore exposing links and connecting data in an accessible manner. Process logic can be freely defined as data and not programmed in relatively inaccessible formats and locations.

The capabilities of technology to enhance human interaction can directly correlate with success or failure. Human interaction with the systems includes the layout, personalisation and navigation of the user interface. The topic human and system interaction is explorer in section 2.5.

Internet search engines provide a good example and illustrate the point made in the introduction about shielding users from complexity. The algorithms, ontologies, graphs and methods applied are effectively hidden behind a simple user interface. This simple and even crude mechanism forms an invaluable tool for many users and represents a a significant factor behind the success of the Internet.

People quickly become experts within their role by identifying and actively looking for specific information particularly when tasks are regular. This is effective and efficient providing the process is consistent and the information represents all aspects required for decision making and that users are paying attention. Borrowing from a physical example poka-yoke is a term taken on the Toyota Production System meaning "mistake proofing" (Shingo, 1986). The essence of this is to build in helpful features from the very start of the design process preventing users from making incorrect choices. A practical example of this is to incorporate shapes that only permit assembly in the correct orientation.

Big data refers to massive datasets containing billions of information items that are too large to be analysed by conventional database tools (White, 2013). Marz and Ritchie (2011) discuss a paradigm for Big Data with more than 30,000 gigabytes of data generated every second. Data is diverse and virtually unlimited and has pushed relational databases to the limit, therefore a new breed of technology has emerged

grouped under the term NoSQL.

Approaches to Big Data are inherently more complex than relational databases, however, many of the underlying fundamentals about transactions still exist. Summarising and simplifying the approaches segregate and combine data in a more efficient manner by focussing on live or the latest data across layered and distributed file systems with mechanisms to handle latency between sources and processors.

Data itself requires careful examination in order to understand how it can be used. In order to understand data the collection, description and quality are important consideration (Sabherwal and Becerra-Fernandez, 2011) .

Viewed holistically implementing the technologies and approaches described above is a significant challenge but necessary to address competitive forces and gain an advantage. Human dependency on systems will continue to increase along with the need to leverage data from existing and new system will grow. As discussed in section 2.2 transactional systems and hence data will continue supporting fundamental business processes. The evolution of complex architectures and applications with technologies including SOA, BPP, IoT, mobile and BI is a significant challenge especially alongside managing risk, applying effective governance and ensuring security.

Analysis approaches must be flexible and agile in terms of data, navigation, scalability with an intuitive but robust means of navigation. Expectations are continually rising, partially driven by access to consumer applications that feature tailored user experiences, easy adoption and actively target the needs of the user. Implementation cycles are becoming shorter; the ease by which users can change or migrate between systems and applications is a notable feature. Future landscapes and technology solutions are likely to be customised and unique.

Geographical differences will be evident as the physical world differs culturally and technologically there will be many solutions to the same problem, therefore, handling inconsistency is fundamental. Finally and moving on slightly from cultural aspects, users need to learn, understand and interact with systems and other users through these

new technologies and techniques. Controlling and understanding these systems will be an increasingly complex task that requires the application of knowledge. Systems must support users in facing these challenges. A fundamental requirement will be the capability to analysis and understand complex agile solutions.

The future direction will be determined by competitive advantage as Porter (1985) describes these include through differentiation, cost, innovation or operational efficiency.

2.4 Data, Information and Knowledge

2.4.1 Transactional Data

'Master data' and 'transactional data' are two terms used to segregate data in enterprise systems. Master data is relatively static in nature and represents data that relates to structure. Generally it refers to to elements such as customers, vendors and products that are traded. All of these master elements change over long periods of time but from a life cycle perspective they are used time after time but as objects they do not change frequently. Transactional data represents the business operation. It links input data and master data in order to represent real objects that flow through the organisation and beyond. Many transactions are processed in sequence to represent the changing state of an object and associated flow such as information and financial data.

2.4.2 Data, Information and Knowledge

Data, information and knowledge are terms that progressively increase in usefulness and correspondingly difficulty in collecting. Knowledge is considered to be an active human function that can be enhanced by interaction between humans and systems. An iterative process that consumes and creates data, information and knowledge.

Knowledge is complex to define having both explicit and tacit properties. Explicit

knowledge is stated clearly and in detail where tacit knowledge is understood or implied without being stated (Oxford, 2012). Organisation prefer explicit knowledge as this is easier to communicate and measure. The open nature of the real world means closed world systems have to rely on the tacit knowledge of humans.

Epistemology, the theory of knowledge (Oxford, 2012), defines knowledge as propositional, procedural and personal (Fantl, 2012). Propositional is the knowledge of facts. Procedural is the knowledge of how to do something. Personal is knowledge by acquaintance, being familiar with something similar. These terms are similar but more detailed than the explicit and tacit description, however, both represent a move from tangible to intangible.

Data is the simplest and defined as the quantities, characters, or symbols on which operations are performed by a computer, this also include storage and transmission (Oxford, 2012). From a philosophy perspective data are things known or assumed as facts, making the basis of reasoning or calculations (Oxford, 2012). This is mentioned with a thought towards conceptual structures. Davenport and Prusak (2000) argue that there is no inherent meaning in data and that it says nothing about its own importance.

There is no shortage of data per say but there are shortcomings in the quality, availability and usefulness of data. Figures vary but one by Herschel and Jones (2005) estimates that about 80% of business information is available in an unstructured form.

Information is data conveyed or represented by a particular arrangement or sequence of things (Oxford, 2012). Often a context or if some meaning is attributed to data it becomes information (Gordon, 2007).

Information may be derived from data when the data is joined with collective meaning understandable in a community to which the information might be addressed (Wille, 2001).

Knowledge has many dimensions and multiple definitions exist. Knowledge is facts, information, and skills acquired by a person through experience or education; the

theoretical or practical understanding of a subject (Oxford, 2012). The traditional definition of knowledge is based upon belief, truth and justification. The person believes the statement to be true, the statement is in fact true and the person is justified in believing the statement to be true (Ichikawa and Steup, 2013).

Devlin (2001) defines data, information and knowledge as below. Internalise takes the psychology view of making (attitudes or behaviour) part of ones nature by learning or unconscious assimilation (Oxford, 2012):

$$\text{Data} = \text{Signs} + \text{Syntax}$$

$$\text{Information} = \text{Data} + \text{Meaning}$$

$$\text{Knowledge} = \text{Internalised information} + \text{Ability to utilise the information}$$

The creation of knowledge from information can be promoted by proper representation of information which make the inherent logical structure of the information transparent (Wille, 2001).

Knowledge originates and is applied in the people's minds, in organisations it is often embedded in documents, repositories but also in an organisation routines, processes and practices and should be evaluated by the decision or actions to which it leads (Davenport and Prusak, 2000).

Knowledge is a fluid mix of framed experiences, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information (Davenport and Prusak, 2000).

Metadata is a set of data that describes and gives information about other data (Oxford, 2012). Metadata can be useful as it supports additional meaning within applications such as classification, search, discovery and transformation; in essence a level of abstraction. Metadata is text, voice or image and there are at least six types: semantic, storage, process, display, project and program (Burbank and Hoberman, 2011).

Developing understanding through identifying and managing good information is

an enormous challenge. Semantics could provide better insight and improved decision making through improved understanding. McComb (2004) describes semantics as a formula, expressed another way, semantics is about meaning.

$$\text{Semantics} = \text{Data} + \text{Behaviour}$$

Wille (1997) argues that Peirce's pragmatism claims that we can only analyse and argue within restricted contexts where we always rely on pre-knowledge and common sense.

Context considers time, location and historical dimensions, it may change for any given event or user perspective. The definition used here for semantics focusses predominately on the canonical meaning and is less time dependant than context. Making use of semantics and context is certainly a multilevel problem that is not restricted to a unique event or object, it is equally applicable to a sequence of events or a situation.

Relating the research question to these definitions, discovery of knowledge in transactional data is anticipated to traverse all levels. Raw data serves as an input that has meaning, definitions and attributes because of its role in a relational database. Analysis, communication and representation of this data within context can be used to discover and demonstrate knowledge; effectively this is the internalisation and ability to utilise information.

through analysis and communication results in knowledge.

2.4.3 Managing Data, Information and Knowledge

Data management is challenging because data is not static, it constantly flows and changes even if this is just conversions to suit formats as per processing needs. In a simple form and typical of digital storage all data is represented by O and 1 in a binary format. What happens between this state and the data computers and humans interact with is mostly hidden and represented by numeric or symbolic values. From here a range of mechanisms are applied; data is grouped, ordered or linked in files

and databases. It starts to become information when algorithms, programs and indeed users interact with the data.

Managing and analysing data is difficult for a number of reasons. Data is dynamic, it is constantly expanding and containing new features. The quality of data is variable and subject to being incorrect, imprecise, incomplete and containing redundant values. In the context of connected systems this challenge is multiplied as definitions vary and data diverges as it is copied and updated independently.

The diagrams in figures 2.13 & 2.14 indicate how data content and conversion are anticipated to increase, the scale is demonstrated by considering a standard Sales and Operations Planning (S&OP) process. This cycle illustrates the basic steps (1-7) of a S&OP process as shown in table 2.2, steps 1 and 5 are also highlighted on the disaggregation axis in the diagrams. This iterative cycle disaggregates data through various stages and systems before re-aggregating for reporting purposes.

Step	Process	Aggregation	Internal	External	Technology
1	History	High	Yes	No	DW
2	Forecast	High	Yes	No	SCM
3	Plan	Medium	Yes	No	SCM
4	Detailed Plan	Low	Yes	Partial	ERP
5	Execution	Low	Yes	Partial	ERP
6	Feedback	Low/Med	Yes	Partial	WF
7	Reporting	Med/High	Yes	No	DW

Table 2.2: Example of S&OP Process, Data Aggregation and Communication (Watmough et al., 2010)

Movement along arrow A in figure 2.14 could be enhanced by:

- Semantic content in order to facilitate greater system synchronisation, particularly in cross system scenarios using Operational Data Stores or inter organisation processes.

- Capturing more content in process cycle 1 will benefit further iterations of the cycle, particularly qualitative content. Coincidently, greater content may feedback

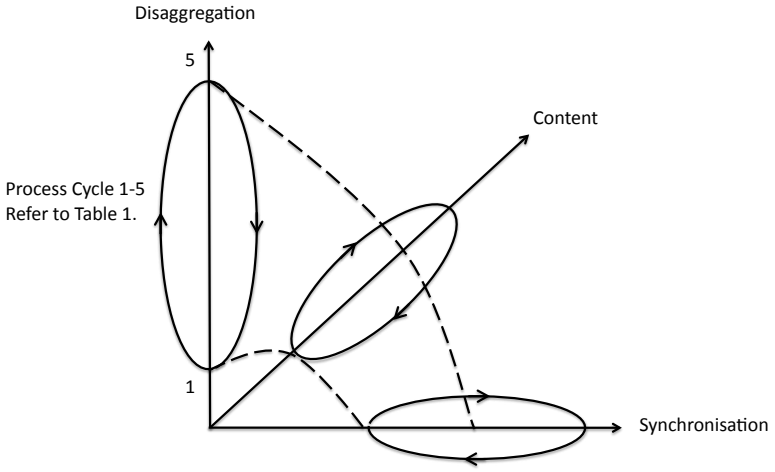


Figure 2.13: Current Supply Chain Data Usage (Watmough et al., 2010)

directly into the process cycle without getting as far as synchronisation between systems.

- Managing change and governance.

Movement along arrow B in figure 2.14 could be enhanced by:

- Higher levels of integration and feedback support by technology developments.
- Managing the life cycle of data.
- Managing change and governance.

The problems that initially triggered the creation of the World Wide Web by Sir Tim Berner-Lee were primarily the need to share, managed, find and change information across complex projects that involved different types of technologies (Dominque et al., 2011*b*). These problems have been power-phrased significantly only to highlight the similarities with enterprise systems, a proportion of which are Internet or Intranet based. Firstly the same technologies may provide an insight or even solution in an

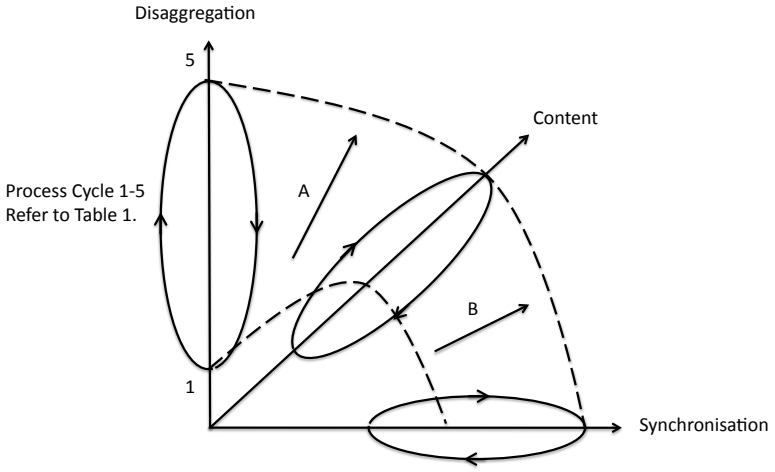


Figure 2.14: Future Supply Chain Data Usage (Watmough et al., 2010)

Enterprise System context, secondly any approach must be compatibly with multiple technologies.

The web is a shared resource and therefore within a machine-readable web meaning should also be shared (Dominque et al., 2011*b*). Machine-readable refers to the semantic web and technologies that enable it to function. This is a web or even application rich environment where computers can perform functions previously only possible by humans through automated analysis, decision making and integration.

The ultimate goal of the web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term Semantic Web refers to W3C's vision of the Web of linked data (W3C, 2013).

Berners-Lee et al. (2001) described human endeavour as being caught in an eternal tension between the effectiveness of small groups acting independently and the need to mesh with the wider community. Semantics has a role to play in ensure concept and context are shared and maintained. Enterprise systems can potentially benefit from a flexible framework that is supportive of knowledge discovery in complex transactional

systems.

The semantic web is not a separate web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation (Berners-Lee et al., 2001). Visions of artificial intelligence are not new however mainstream adoption of semantic technologies in search engines by providers such as Google, Yahoo and Microsoft are appearing.

Semantic technology provides machine-readable (or even better machine processable) descriptions of data, programs, and infrastructure, enabling computers to reflect on these artefacts; there is also the promise of helping people handle increasing amounts of information, application integration and social interaction (Dominque et al., 2011*b*).

Semantic technologies are difficult to define because they are potentially just an extension within the enterprise systems domain as they are in the Internet. There are views that they will be commoditised add-on's for major software vendors such as Microsoft and SAP (Dominque et al., 2011*a*) enabling better processes through efficiency, accuracy, agility and improved analysis rather than a radical change in direction.

Through ontologies and ontology-related technologies, the meaning of and relationships between concepts within published Web pages can be processed and understood by software-based reasoners (Dominque et al., 2011*b*).

2.5 Discovery Techniques and Applying Knowledge

2.5.1 Knowledge and Human Capabilities

This view of enterprise systems presented above is representative of many organisations at the present time but it does not encapsulate the challenges and solutions currently being applied or developed. Debevoise defines the key drivers as visibility, organisation alignment and adaptation (Debevoise and Geneva, 2008), a theme that is continued in this research and echoes the sentiments of exceptions becoming the normal mode of

operation coupled with localised processes and varying data. This is not intended to suggest that core processes are not controlled, it is intended to highlight how processes can be refined locally to reflect the interaction with local partners. The term process within the context of this subject is used to describe a planned sequence of events in order to achieve an outcome and represents the combination of humans and systems. The manner of interaction is expected to differ due to the availability of information and the ability to communicate through a wide range of devices; this includes input from sensors, connected system or even automated decision making.

Everything in nature, in the inanimate as well as the animate world, happens according to rules (Kant, 1988). The exercise of our own powers also takes place according to certain rules while we first follow without being concious of them, until we gradually come to cognise them through experiments and long use of our powers, and finally make them so familiar to us that it costs us great effort to think them in abstraction (Kant, 1988).

Devlin (1997) states that humans have tried to represent knowledge and understand the laws of thought for thousands of years and that we are still unable to explain exactly how our minds perform such feats. He argues that "our minds are intimately intertwined with the world around us, and that our feelings and perceptions, even our social norms, play crucial roles in the marvellous complex dance of human cognition". In the domain of ERP this is equally applicable, systems currently have limited contextual awareness but this is changing. The availability of inputs from sensors and mobile devices is increasing. If it can be harnessed a step change may be possible. Decision making in systems is rule based and constrained therefore, at least at present, systems need to complement human capabilities.

Dreyfus and Dreyfus (1986) introduced a five stage model about human performance based on human skill acquisition. Level one represents novice where rules are followed without context. The middle stages introduce a holistic understanding with some reliance on rules. Level five is where experts function, essentially through skill

and being fully aware of context, with virtually no referral to rules. This represents an interesting choice of paradigm for future enterprise systems; there is a strong argument for rule based systems that are capable of automating or supporting the decision making processes and representing Enterprises in an understandable form. Clearly a business would not want to operate at novice level, somewhere around the middle of this scale where rules are followed with some contextual knowledge would be desirable for systems, however, this is far from the optimum for humans. The converse view of this is that the sheer volume of data and complexity would prevent a human from operating at an expert level across the whole organisation as humans do not posses the communication capability to operate as a collective body, certainly not at any great speed or to the detriment of other activities. This is not intended as a sweeping statement across all areas and there will be many situations where simple rules are all that is needed, but the drive towards expert level will need a union of systems and humans.

In order to discover knowledge it is prudent to understand the foundations. Knowledge Representation (KR) encompasses a range of relevant topics. Sowa states that KR is a multidisciplinary subject that applies theories and techniques from three other fields: logic, ontology and computation (Sowa, 2000). Computation by definition is the action of mathematical calculation and the use of computers, especially as a subject of research or study (Oxford, 2012). Sowa argues that without these KR is vague, ill-defined and cannot be implemented in computer programmes.

Logic by definition is reasoning conducted or assessed according to strict principles of validity (Oxford, 2012). It is a 1,000 year old technology to formally capture meaning with a large number of logics developed, each suitable for a specific purpose (Dominque et al., 2011*b*). Discussing logic in detail is beyond the scope of this thesis, however, the underlying point is that logic can be decipherable. First-order logic, for example, can be used to reason.

First order logic (FOL) deals with predicates, these being functions that maps its arguments to the truth values, the synonym of predicate being relationship (Sowa,

2000). The principle of FOL is beyond the scope of this chapter, suffice to say, it supports the description of relationships in an algebraic form; these in turn can support the formation of rules and inference.

Ontology by definition is the branch of metaphysics dealing with the nature of being (Oxford, 2012). Ontology is about what is real or formal knowledge representation, the metaphysics, this encompasses the first principles of things including abstract concepts such as being, knowing, identity, time, and space (Oxford, 2012).

Ontology is the study of categories of things that exist or may exist in some domain. When combined with logic an ontology provides a language that can express relationships about the entities in the domain of interest (Sowa, 2000). The basis of the analysis contained in this research is a bottom-up discovery of formal ontologies by applying FCA. Lattices constructed by FCA methods are structured in a manner that supports ontology development, this was also indicated by Sowa as hierarchies of categories (Sowa, 2000). FCA is discussed further in chapter3.

The basic principle of ontology can be demonstrated using the example in figure 3.11. The categories of 'Engine', 'Sail', and 'Paddle' have been identified and could therefore be queried using logic. If a secondly lattice or ontology also contained these categories a query could also utilise this knowledge, a useful feature given the complex environments under consideration.

Representing data in a visual form is common and relatively easy with modern systems. Options include tables, many types of graph and charts, hierarchical trees and layer views including graphics such as maps. The importance of visualisation cannot be understated with 80% of BI solution customers finding visualisation desirable (Soukup and Davidson, 2002).

Epistemology contrasts with ontology in that is about perception, in philosophy it is the theory of knowledge and justification between belief and opinion (Oxford, 2012). This contrast highlights a fundamental difficultly within knowledge representation. Knowledge is subjective and is based on truth and facts but also understanding and

beliefs. As Davenport and Prusak (2000) highlight, people cannot share knowledge if the don't speak a common language, more than the language this infers a common understanding of symbols, models, text and any other communication means.

Stamper contributed early to information system design. The subject of Organisational Semiotics is the study of signs and symbols their use and interpretation (Oxford, 2012). Within this field Stamper's semiotic framework or ladder formed a useful tool for understanding information systems at different levels of abstraction.

Stampers early career led him to conclude that an authoritarian organisation such as the army would not allow the proper use of individuals talent though the organisation could be efficient.

Stamper identified an effective theory for information systems inspired by semiotics, after Charles Sanders Peirce, and signs enabling one to perform actions, after Charles Morris (Stampler, 1973).

The Semiotic Ladder, see table 2.3, built on the original Peircian semiotics (syntactics, semantics and pragmatics) by introducing three additional aspects (physical, empirical and social) (Gazendam and Liu, 2005). These additional aspects focus on the technical level and compliment the earlier more human levels. These aspects should be viewed as independent and useful for their intended purpose for planning semiotic analysis, they can be seen a method of connecting ontology and epistemology. Ontology connects the physical world and epistemology with the pragmatic and social levels, typically this is where knowledge is associated.

Semiotic Layer	**Human Level**
Social Layer (Social World)	Cultural norms, beliefs, expectations, functions, commitments, law, culture, contracts, values, shared models of reality, attitudes
Pragmatic Layer	Communications, conversations, negotiations, intentions
Semantic Layer	Meaning, propositions, validity, truth, signification, denotation
Semiotic Layer	**Technical Level**
Syntactic Layer	Formal structure, logic, data, records, files, computer language
Empiric Layer	Noise, entropy, pattern, variety, noise variety, redundancy, codes, efficiency
Physics Layer (Physical World)	Signals, traces, physical distinctions, hardware

Table 2.3: Semiotic Ladder, after Liebenau and Backhouse (1990)

Stamper develop information system analysis as the identification of agents (or actors), affordances and the governing norms (Gazendam and Liu, 2005). Norms are rules at an individual or organisational level, they can be formal or informal but they differ from affordances in that they are determined from social and cultural contexts. This view of information systems as social systems extends the scope of analysis beyond a pure mechanical viewpoint.

Davenport and Prusak (2000) discuss knowledge representation in a more lucid manner as 'working knowledge' where knowledge is a fluid mix of framed experiences, values, contextual information and expert insight that provides a framework for evaluating and incorporating new experiences and information. They have a view that knowledge and the mechanisms to share and use knowledge is embedded across organisations in documents, procedures, processes and norms. This view should be extended into communities, media, social groups and beyond given the open society and information rich environment most people experience as a normal part of life today.

Knowledge by definition is facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject (Oxford, 2012). Turning information into knowledge is best supported when the information

with its collective meaning is represented according to social and cultural patterns of understanding of the community whose individuals are supposed to create the knowledge (Wille, 2001).

Conceptual maps are graphical tools for organising and representing knowledge; concepts in this domain refer to perceived regularity or records of events or objects and relationships or semantic units (Novak and Caas, 2008). Figure 2.15 describes a concept map showing the key features of concept maps, it is read progressing from the top downward. Expressing knowledge as concept maps appears to be a good method for representing knowledge discovered from transactional data. The main features include the ability to focus on questions, context, concepts, relationships and interrelationships that provide an holistic view across the domain under consideration.

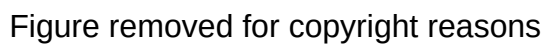


Figure 2.15: A concept map showing the key features of concept maps. (Novak and Caas, 2008)

Wille (2001) discussed turning information into knowledge and how it is best supported when the information with its collective meaning is represented according to social and cultural patterns of understanding of the community whose individuals are supposed to create the knowledge.

Wille (2001) defines knowledge discovery as information discovered combined with

knowledge creation where the combination is given by turning discovered information into created knowledge.

Stockburger (1998) defines a model as a representation containing the essential structure of some object or event in the real world and further differentiates between physical and symbolic. Two key comments are made by Stockburger, firstly that models of the real world are incomplete and secondly that they can be changed or manipulated with ease.

The role of various technologies and underlying theories vary in their application across the Enterprise System landscape. Figure 2.16 neatly highlights the range, starting at the bottom where existing systems perform very well. Clearly definable logic or rules can be computed in a closed world environment; all the inputs and routines required to deliver the outputs are encapsulated in a what could be a black box system. These rules could be abstracted and brought into an SOA architecture and serve multiple customers.

As conceptual activity increase towards the middle and upper section of the diagram so does the traditional level of seniority of the organisation as well as a reduction in frequency. The top of this diagram highlights corporate strategy, an activity involving significant input from humans that occurs infrequently maybe on a 3 to 5 year cycle. The bottom level typically occurs much more frequency with hourly cycles or less. The target of this research is reflected in the middle sections of this chart where data, systems and people interact relatively frequently.

The diagram in figure 2.17 reflects the semantic technology offerings of 100 companies across a range of industries (Dominque et al., 2011*a*). Business intelligence, business process management are two relatively small sections but the expectation is that research in this domain will expand particularly as information access, knowledge management and most significantly as enterprise integration turns into mature technologies.

Figure 2.16: Conceptual and non conceptual activities in a business (Dominque et al., 2011*a*)

2.5.2 Conceptual Knowledge Discovery

Sowa (2000) describes three sources of knowledge about the world; observation, simulation and deduction. He also argues how knowledge acquisition involves a constant cycling of abstraction and reinterpretation by people to approximate the real world. Any approximation of the real world will have flaws, understanding the significance of these flaws is a key challenge. As defined by Sowa (2000), deduction represents the most simple, rule based source with a small range of inputs. Simulation is model based and only limited by the computing power and range of inputs, this is turn limits the complexity of the model and how closely it reflects the subject. The most accurate albeit restricted to the current time is observation of the real world.

Fayyad et al. (1996) describe knowledge discovery as the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. In this definition the data mining stage is used to identify patterns in data that has been transformed into a suitable format, the patterns being interpreted in order to derive knowledge. Interpretation infers human input.

Wille (2001) defines knowledge discovery as information discovery and knowledge

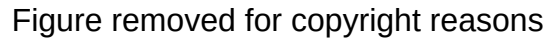


Figure 2.17: Semantic offering according to area (Dominque et al., 2011*a*)

creation, promoted by proper representation of information which make the inherent logical structure of the information transparent.

Rajaraman et al. (2012) define the most commonly accepted definition of "data mining" as the discovery of "models" of data. Models infers more of a structural perspective than patterns with clear interconnections. Patterns suggests movement or time dimensions.

The knowledge discovery cycle within processes, in the context of enterprise systems, varies over the life cycle. Initial design is deductive, a top down approach that reasons with known requirements and mitigates against unknowns. Iterations of the process and analysis identifies issues or cost saving opportunities, stimulus for may be catastrophic failure through to unacceptable variables. Deductive reasoning can progressively focus the analysis through visualisation, statistical or algorithmic techniques but they rely on a base understanding of the intended processes. In practical terms Key Performance Indicators (KPI) provide a means of more immediate monitoring and cognitive understand but suffer from a reliance on a relatively static understanding of the process model (Parmenter, 2010).

Devlin (1997) suggests cognition, reasoning, and communication are simply different ways the the brain processes information, he defines the terms as follows. Cognition can be regarded as a process of acquiring information. Reasoning can be regarded as a means of enlarging our stock of information by deriving new information from existing information. Communication can be regarded as a means of conveying information from one person to another.

This research focuses on reasoning, particularly inductive reasoning, and communication. This research considers human cognition as being outside its scope due to the complexity and unknowns, it is unlikely to contribute to this domain. Inductive reasoning is the focus as this bottom-up approach tries to gain understanding, potentially from raw data, thereby developing knowledge from data or facts. Given the complex nature of data deductively reasoning and determining proven facts is deemed an unlikely outcome.

Retroduction also known as abductive inference should also be mentioned. Peirce argues that retroduction is the first step of any scientific inquiry and that this kind of inference is very human (Burch, 2010). The ability to induce connections between facts that are not stated or connected in a logical manner is typical of human intellect. Where discovery could be really powerful is if inductive reasoning and abductive inference could be harnessed.

Brachman et al. (1993) introduced the notion of data archaeology, a skilled human task in which knowledge emerges only through an iterative process of data segmentation and analysis. Archaeology is a curious term as it implies the analysis of historical data but the key features include identifying patterns, categories and iteration combined with human interaction.

Conceptual Knowledge Discovery in Databases (CKDD) is based on FCA, as a subset of Knowledge Discovery in Databases (KDD) it aims to support a human centred processes of discovering knowledge from data by visualising and analysing the conceptual structure of data (Hereth et al., 2003).

A set of fundamental requirements for human-centered KDD support tools has been compiled by Hereth et al. (2003) from work by Brachman et al. (1993) and Brachman and Anand (1994). They argue that CKDD provides a means to satisfy these requirements. These requirements include:

- Representing and presenting the underlying domain to the user in a natural and appropriate fashion
- The domain representation should be extendible by the addition of new categories formed from queries
- It should be easy to form tentative segmentations of data
- Analysts should be supported in recognising and abstracting common analysis
- There should be facilities for monitoring changes in classes or categories over time
- The system should increase the transparency of the KDD process
- Analysis tools should take advantage of explicitly represented background knowledge of domain experts, but should also activate the implicit knowledge of experts
- The system should allow highly flexible processes of knowledge discovery

It is proposed that these requirements are reviewed with respect to the research question as part of chapter 6.

2.5.3 Knowledge Representation Techniques

Data forms the foundations of knowledge representation, typically this is supported by differing types and structures. Fundamentally data types are quantitative or qualitative, within these there may be standards, known definitions, linear scales or conversely local or unknown definitions, terminology or language.

Cios et al. (2007) describes the main knowledge representation categories as rules, quantified rules, graphs and directed graphs, trees and networks, these are subsequently combined with different types of sets including intervals, rough and fuzzy. An illustration of usefulness in the context of rules with differing levels of granularity is shown in figure 2.18. As rules become more granular the principle is that they become more useful. For example, given any two variables a useful rule would be $A = 2B$, a simple and granular rule capable of being applied to the variables no matter the size or dimension.

Figure removed for copyright reasons

Figure 2.18: Usefulness of Rules and Granularity (Cios et al., 2007)

Rules are akin to mathematical formula, they are used to express a relationship and are frequently readable when stated in computer languages.

Example rule: IF condition THEN conclusion

Graphs and Directed Graphs represent relationships between concepts, with direction the relationship can be quantified. Graphs can be useful for visualising concepts, see figure 2.19, particularly when combined with graphical techniques including colours, hierarchies and weighted relationships. A related idea is the concept lattice, this is discussed in chapter 3 alongside FCA.

Trees are a special category of graph in which there is a single root (A) and a collection of terminal nodes (D E F) and no loops (Cios et al., 2007), see figure 2.20.

Figure 2.19: Example of Graph and Directed Graph

Decision trees are a common application of trees particularly within tools such as expert systems or knowledge databases. Questions or observations can be posed thereby traversing the tree structure, following the path through multiple decision points until a final decision or action is determined.

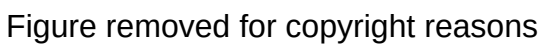


Figure 2.20: Example of Decision Tree

Networks can be regarded as generalised graphs with each node of the graph containing the underlying processing capability (Cios et al., 2007), see figure 2.21. Inputs from the left hand side are processed in nodes (A B) and the output passed to the right hand node (C) for further processing. Networks are capable of logic computing; a common but highly complex example are neural networks that enable people and other living creatures to perform tasks.

All the knowledge representation techniques discussed can be represented as mathematical or logic statements and therefore programmed for use by computer systems. All have been deployed in some form with a range of successes and failures, however,

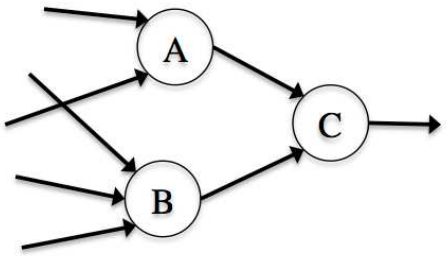


Figure 2.21: Example of Network

further research is still required in many areas.

2.6 Conclusion

This chapter introduced transactional data and analysis techniques in the context of enterprise systems and knowledge representation. The importance of enterprise systems has been highlighted as have the competitive forces creating demand for increased integration, intelligence and efficiency of use. The core components of enterprise systems and BI solutions have been discussed and evaluated.

Technology is contrasted with a review of human capabilities and how important the visual representations and actually working with information is to gain and develop understanding and knowledge. There is a need for an analysis method capable of discovering relationships that enhances human capabilities whilst being congruent with system-based computation. There is no assumption that anything other than complex analysis is required for complex systems.

Enterprise systems capture data in a transaction structure so that they can provide information that seeks to align with the knowledge that decision-makers use to achieve business goals. Traditionally this has been achieved by a 'top-down' approach whereby the business process is designed then the data is set according to that human-oriented model. However, with the emergence of service-oriented architecture and developments in business intelligence, data in its own right is becoming significant, suggesting that

data in itself may be capable of capturing human behaviour and offer novel insights from a 'bottom up' perspective. The constraints of hard-coded top-down analysis can thus be addressed by agile systems that use components based on the discovery of the hidden knowledge in the transaction data.

Knowledge is a valuable but expensive commodity. It is delicate, easily lost or misunderstood and difficult to gain; moreover it is hard to computerise. Humans exhibit innovation and intelligence but unlike computers are incapable of processing large volumes of complex data. There is a need to connect the user's human-oriented approach to problem solving with the formal structures that computer applications need to bring their productivity to bear.

This research therefore proposes to examine and analyse transactional data through knowledge discovery techniques, in particular Formal Concept Analysis (FCA). In order to understand FCA and alternative approaches it is necessary to examine the theoretical foundations. This is discussed in detail in chapter 3.

Chapter 3

Formal Concept Analysis

3.1 Introduction

This chapter provides an introduction to the theoretical foundations of Formal Concept Analysis (FCA) and a brief synopsis of alternative semantic technologies contrasted with FCA in this research for discovering knowledge in transaction data.

FCA provides a mathematical theory based on concepts; logical relationships that can be represented and understood by humans (Wille, 2001). Wille indicates how logical connections in line diagrams of concept lattices can stimulate background knowledge for discovering new knowledge, often produces also critic and self-correction of the present information and knowledge. This capability to analyse large data sets and discover relationships through tabular or graphical analysis provides a useful mechanism that can be applied to transaction data for the discovery of hitherto hidden knowledge.

The focus of this research is on the application of FCA. The analysis steps are described starting with the extraction of source data to tabular and graphical lattice representation. Examples of existing FCA applications are described and considered.

3.2 Formal Concept Analysis

3.2.1 Origins of FCA

"Formal Concept Analysis is a field of applied mathematics based on the mathematisation of concept and conceptual hierarchy. It thereby activates mathematical thinking for conceptual data analysis and knowledge processing. The adjective 'formal' is meant to emphasise mathematical notions." (Ganter and Wille, 1999)

Formal Concept Analysis had its origin in activities of restructuring mathematics, in particular mathematical order and lattice theory (Wille, 2005). For a comprehensive description of FCA , mathematical basis and proof refer to Ganter and Wille (1999). Our focus is on the application of FCA.

FCA is a method for data analysis, knowledge representation and information management (Priss, 2006). The basic steps involve representing data in a formal context, alternatively known as a cross table, this can also be represented as a structure or concept lattice. The following section describes these in more detail.

3.2.2 Formal Concept and Formal Context

A concept is defined as a unit of thought, section 3.3.2 expands on this short definition while this section continues in the context of FCA.

Concepts can only live in relationships with many other concepts and how a mathematical model needs to speak about objects, attributes, and relationships which indicate that an object has an attribute, thus the notion of a "formal context" and Formal Concept Analysis was introduced (Wille, 2005).

FCA is mathematical theory of data analysis using formal contents and concept lattices (Priss, 2006), (Wormuth and Becker, 2004), (Andrews et al., 2011) and has the potential to compliment and advance current forms of analysis.

A concept can be visualised as a pair (A,B), as shown in figure 3.1, and is called

a formal concept of the given context. The set A is called the extent, the set B the intent of the concept (A,B) (Wolff, 1993).

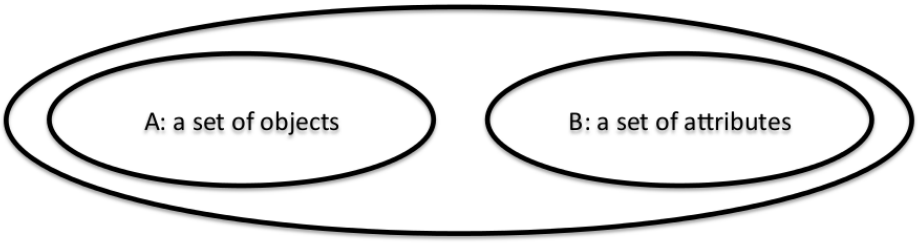


Figure 3.1: A Formal Concept - (A,B)

Many formal concepts in a set is termed a formal context. A formal context $\mathbb{K} := (G, M, I)$ consists of two sets of G and M and a relation I between G and M. The elements of G are called the objects and the elements of M are called the attributes of the context (Ganter and Wille, 1999). The individual elements of G and M are represented by g and m, see figure 3.1.

A simple concept can be represented in a cross table as illustrated in figure 3.2 where the cross in row g and column m indicates that the object g has attribute m. This can also be expressed as gIm or $(g, m) \in I$ and read as 'the object g is in a relation I with an attribute m' (Ganter and Wille, 1999).

When multiple objects have multiple attributes in common the cross table begins to resemble the table in figure 3.3 where A represents a group of objects and B represents a group of attributes, expressed formally this illustrates the extent *(A)* and intent *(B)* of the formal concept (A, B). The pair (A, B) is a formal concept of the context *(G, M, I)* with $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. (Ganter and Wille, 1999). The whole table is referred to as the formal context which may contain many concepts.

If (A_1, B_1) and (A_2, B_2) are concepts of a context, (A_1, B_1) is called a subconcept of (A_2, A_2), provided that $(A_1 \subseteq A_2)$ (which is equivalent to (B_2, B_2). In this case, (A_2, B_2) is a superconcept of (A_1, B_1) and is wrote $(A_1, B_1) \leq (A_2, B_2)$. The relation $\leq$ is called the hierarchical order of the concepts. The set of all concepts of (G, M, I)

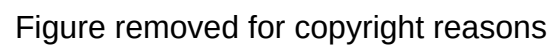


Figure 3.2: Simple Concept in a Formal Context or Cross Table (Ganter and Wille, 1999)

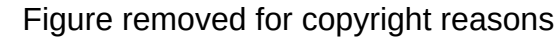


Figure 3.3: Formal Context or Cross Table (Ganter and Wille, 1999)

ordered in this way is denoted by $\mathbb{B}(G, M, I)$ and is called the concept lattice of the context (G, M, I) (Ganter and Wille, 1999). The diagram in figure 3.4 illustrates how concepts are hierarchically connected and represented in a lattice. This is further discussed in the next section.

3.3 Concept Lattice

Graphically represented concept lattices have proven to be extremely useful in discovering and understanding conceptual relationships in given data (Hereth et al., 2003).

Concept lattices with their line diagrams are indeed able to support knowledge discovery in databases; in this way humans can make accessible a rich conceptual

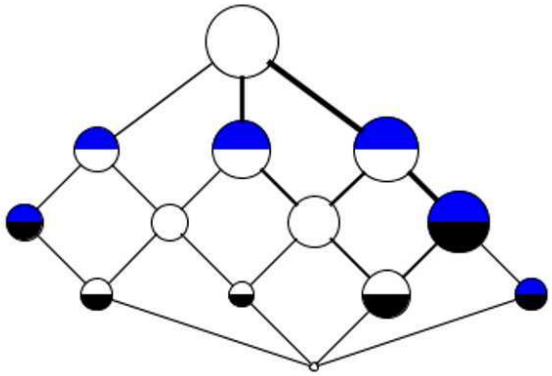


Figure 3.4: Example Lattice

landscape of knowledge (Wille, 2001). An introduction is provided below, refer to Ganter and Wille (1999) for a comprehensive description and proof.

A central notion in FCA is the 'Galois connection', an intuitive way of discovering hitherto undiscovered information in data and portraying the natural hierarchy of concepts that exist in a formal concept (Andrews et al., 2011).

Figure 3.4 represents a lattice visualised within Concept Explorer (ConExp Project, 2006). This simple lattice is represents graphically a formal context where the vertical hierarchy links and thereby visualises relationships between objects, referred to as nodes (circles in the diagram).

A brief description of the lattice within Concept Explorer follows refer to the resource ConExp Project (2006) for the original documentation. Section 3.5 also describes the concept lattice from an application viewpoint.

Each node in the lattice corresponds to a formal concept as depicted in figure 3.1. There are additional options for changing the visualisation and labelling within the Concept Explorer software not described here, as shown the node diameter represents the number of objects matching exactly the intent of the node.

A blue filled upper semi-circle indicates that there is an attribute attached to this concept. If a node is marked by a black filled lower semi-circle then there is an object

attached to this concept. Consequentially a white fill indicates no attachments.

Conceptual scaling is an interpretation process, these include plain, elementary, ordinal, interordinal, biordinal and dichotomic scaling; an example of each is shown in figure 3.5 (Ganter and Wille, 1999). The choice of scale is to an extent determined by the data and analysis. Ordinal scales start to imply hierarchy and can indicate where objects share multiple and ranges of attributes. These may be where attributes share a common characteristic for example shades of a colour or conversely shared or mutually exclusive ranges in the case of interordinal or biordinal respectively. Scaling is mentioned as it has the potential to simplify analysis and benefit the visual representation.

Figure 3.5: Conceptual Scaling Examples (Ganter and Wille, 1999)

Graphically displaying concept lattices in an effect manner can be problematic. Various techniques are available including Minimum Support and Iceberg approaches.

A description of minimum support is included in section 3.3.1. It forms the primary technique applied within this research as it is an accessible tool with no obvious advantages or disadvantages when compared to alternative techniques.

Stumme et al. (2002) describe their iceberg concept lattices approach as consisting of only the top-most concepts of the concept lattice, these being concepts which provide the most global structuring of the domain. Based on frequent items sets they describe this approach as being useful for analysing large datasets where strong correlations exist. Examples include database marketing, transformation of class hierarchies and ontology learning.

3.3.1 Minimum Support

Minimum support is a method of reducing the size and complexity of formal contexts (Andrews, 2011*b*) by removing data that has less than a set number relationships with objects and attributes. In-Close2 is a fast formal concept miner for FCA files in a 'cxt' format that includes minimum support (Andrews, 2011*b*). This format is a text file containing the objects, attributes and formal context produced as an output of FCA. A number is manually entered for the minimum number of intents and extents or attributes and objects respectively. Figure 3.6 and 3.7 demonstrate how D5 is removed by using minimum support to remove intents/attributes or extents/objects in this case with a value of one.

Andrews and Orphanides (2010*a*) applied minimum support to filter out small concepts from a large data set and demonstrate that understandable results are obtainable via this method and software. They described how by applying this straightforward method produced useful insights are gain through the creation of a readable lattice.

	A	B	C	D
1		Engine	Sail	Paddle
2	Yacht	X	X	
3	Dingy		X	
4	Power Boat	X		
5	Kayak			X

Figure 3.6: Before Minimum Support Applied

	A	B	C
1		Engine	Sail
2	Yacht	X	X
3	Dingy		X
4	Power Boat	X	
5	Kayak		

Figure 3.7: After Minimum Support Applied

3.3.2 FCA from a Philosophical Perspectives

Wille provides an insight as to why FCA has been selected as the primary analysis technique for this research in the quote below:

> "The aim and meaning of Formal Concept Analysis as mathematical theory of concepts and concept hierarchies is to support the rational communication of humans by mathematically developing appropriate conceptual structures which can be logically activated (Wille, 2005)".

Wille (2005) describes many insights into relationships between mathematics and philosophy although he also indicates that this is far from an understanding of human thought. Wille presents FCA as an effective means of communication that supports hierarchies that can be generated and manipulated in a graphical form, essentially providing a link between system based calculations and human cognition. Wormuth and Becker (2004) describes FCA as a mathematisation of the philosophical understanding of concept, a human-centred method to structure and analyse data and a method to visualise data and its inherent structures, implications and dependencies.

The adjective "formal" has a delimiting effect; FCA derives its comprehensibility and meaning from its connection with well-established notions of "concept" (Ganter and Wille, 1999). It is the mathematical foundation and expression in a form that provides meaning in a given situation.

Concepts can be philosophically understood as the basic units of thought formed in dynamic processes within social and cultural environments (Wille, 2005). Wille expands on this in-line with philosophical tradition, a concept is constituted by its extension, comprising all objects which belong to the concept, and its intension, including all attributes (properties, meanings) which apply to all objects of the extension.

The words knowledge and representation have provoked philosophical controversies for over two and a half millennia (Sowa, 2000), a warning, challenge and some expectation setting that succeeding in determining new knowledge easily from transactional data is a significant challenge.

Wille highlights the connections between Seiler's concept theory and Formal Concept Analysis which may be taken as arguments for the adequacy of the mathematics of FCA. Seiler's concept theory discusses concept theories in philosophy and psychology where concepts are cognitive structures whose development in the human mind is constructive and adaptive (Wille, 2005). Seiler's concept theory is only available in the German Language, therefore the main points are taken from Wille and shown in italics followed by a brief explanation.

- *Concepts Are Cognitive Acts and Knowledge Units* - Concepts are captured knowledge as a result of a mental act and useful for understanding although they may be independent of language

- *Concepts Are Not Categories, but Subjective Theories* - Concept are abstract models and represent understanding and knowledge

- *Concepts Are Not Generally Interlinked in the Sense of Formal Logic* - Concepts are logical and representative but not linked to situations or domains

- *Concepts Are Domain Specific and Often Prototypical* - Concepts are a reference based on situations and experiences; they represent a typical or standard example
- *Concepts as Knowledge Units Refer to Reality* - A cognitive process relates concepts to reality
- *Concepts Are Analogous Patterns of Thought* - Concepts are an abstract of reality and related by experience
- *Concepts Are Principally Conscious, but Their Content Is Seldom Fully Actualized in Consciousness* - Understanding concepts requires thought, seen as reflexive knowledge an element of learning is required to understand
- *Concepts as Habitual and Virtual Knowledge Can Be Implicitly and Explicitly Actualized* - Concepts represent explicit knowledge but they can also be interpreted based on their logical structure to discover implicit knowledge
- *The Language as Medial System Promotes the Actualization of Concepts* - Language or words support conceptualisation and understanding
- *Concepts Have Motivational and Emotional Qualities* - Manual manipulation can aggregate and focus in a biased direction
- *Concepts Have a History and Go Through a Developmental Process* - Dynamics such as time, discourse and developments results in change
- *Concept Formation Is Not a Formalisable Automatism* - Change is not based on a predetermined set of rule but is unpredictable

Peirce in the 1870s created the first clear formulation of pragmatism as a principle of inquiry and account of meaning (Atkin, 2005). Peirce's pragmatism proposes that for any statement to be meaningful, it must have practical bearings (Peirce, 1878).

Pragmatics is more than semantics, meaning is important however context, language even experience can have a bearing on the meaning in a particular situation.

Pierce combined his pragmatic maxim with notions of clarity from Descartes and Leibniz to identify three grades of clarity or understanding about concepts, these are described below by Atkin (2005).

The first grade of clarity is to have an unreflective grasp of it in everyday experience. Knowing that a switch will turn on a light suggests a basic understanding, young children are capable of this level.

The second grade of clarity is to have, or be capable of providing, a definition of the concept. A knowledge of circuits, electricity and lighting devices would demonstrate an abstract understanding.

The third grade of clarity considers what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object (Peirce, 1878). A far broader understanding and knowledge is demonstrated, continuing the lighting example effect could include payment or light pollution. This is conditional propositions, relationships between states that can imply or deduce further knowledge.

In discussing formal concepts and concept lattices of formal contents Wille highlights a close relationship between logical and mathematical thinking, particularly in the support of human reasoning when represented by concept lattices (Wille, 2001). The need for a circular process is also highlighted by Wille, that it is open for critic and self-correction. The inherent nature of this technique supports experts in identifying errors.

3.4 FCA Applications

FCA has been applied to a wide range of applications; this chapter does not represent a comprehensive review but an illustration of pertinent applications. The approaches

described below benefit from the capability of FCA to analyse large data sets and the discovery of relationships through tabular or graphical analysis.

Andrews and Orphanides (2010*a*) described two ways in which concept lattices can be produced from data by creating sub-contexts by restricting the conversion of data to information of interest, and secondly by removing relatively small concepts from a context to produce readable, yet still meaningful, concept lattices. Based on data they demonstrated that understandable results are obtainable from existing data sources without requiring domain expertise or statistical analysis. The 'Mushrooom and Adult' data set from UCI Machine Learning (Asuncion and Newman, 2007) formed the data source consisting of 8124 records. This represents a classification type problem generating over 220,000 concepts by FCA. Minimum support as introduction in section 3.3.1 reduced the quantity of concepts thereby resulting in useful analysis.

A common application of FCA is the classification of large data sets as in Andrews and McLeod study using FCA (Andrews and McLeod, 2011). Groups of genes with similar expressions profiles were extracted and discovered, essentially identifying classes with similar features or properties.

Poelmans et al. (2010) applied FCA to business processes and data in order to discover variations and best practice in a medical care situation. The ability to identify process anomalies and exceptions was identified but more importantly the advantageous use of FCA as a discovery process. This represented an attempt to analyse physical events and outcomes and is pertinent to the analysis within chapter 5, the significant difference is how the sequence of events are recognised.

Wille discusses a research group "Formale Begriffsanalys" that started a project in 1991 to develop a retrieval system based on a cross table and conceptual views using TOSCANA (Wille, 2001). This normalised vocabulary to produce a concept lattice associating catchwords with documents in a library. The outcomes was an expert system where researchers could iteratively search for documents containing a topic, the catch word, gradually refine the results.

A different application was applied by Poelman et al. for filtering out "interesting persons" for further investigation by creating a visual profile of these persons, their evolution over time and their social environment (Poelmans et al., 2011). Interesting persons in this research involved the search for suspects and victims of human trafficking and forced prostitution. The interesting outcome from this analysis was how it is possible to discover relationships from unstructured data sets, a task that is virtually impossible to achieve manually in a practical period of time.

The closest realistic comparison to the analysis contained in chapter 5 is broadly defined as web usage mining; approaches such as Clickstream or Statviz (Hitzler et al., 2009) are two examples of both existing software and approaches. Clicksteam is typically a web browser based approach to log user clicks when navigating Internet sites. Results are logged and analysed for many reasons, common areas targeted include navigation, marketing and profiling users. Statviz is very similar but this specific software graphically represents the movement between pages and can incorporate measures such as popularity.

The approach taken in chapter 5 is based on a transactional system. This has similarities to Statviz in that it is tracking movement between transactions, this has a likeness to the pages navigated by the user within a web sites. The underlying difference here is that web-based navigation is typically related to fulfilling a single task in its entirety where as from a transactional viewpoint there is not a clear definition. Transactional systems such as ERP form an integral part of the organisations daily function and users differ enormously by breadth of functions, responsibility and understanding. Individuals typically use a limited set of transactions in order to support a sub-section of processes while handling the normal day-to-day exceptions and interruptions.

3.5 Overview of Analysis

The following section contains an overview of the analysis techniques performed. Data has been extracted from an ERP system and processed into a format suitable for FCA. Details about the structure are contained in the following sections. All data used has been extracted directly from the database, there are no steps that could not be mechanised within the data preparation.

The basic steps performed are similar in nature and utilise a set of research tools for creating a formal contexts, reducing complexity and displaying lattices, respectively these are FcaBedrock - figure 3.8 (Andrews and Orphanides, 2010*b*), In-Close2 - figure 3.9 (Andrews, 2011*b*) and Concept Explorer - figure 3.10 (Yevtushenko, 2006). A comprehensive description has been produced by Andrews et al. (2011). Further working examples are contained in chapter 5 as part of the analysis.

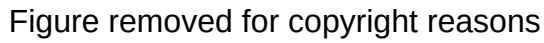


Figure 3.8: FCABedRock

The end result of applying FCA is a concept lattice, figure 3.11 provides an example. The object 'Yacht' has attributes 'Engine' and 'Sail' in contrast to 'Kayak' that does not. Phrased another way, a yacht has an engine and a sail. The lattice is read vertically with lower items sharing the attribute values of higher objects to which they

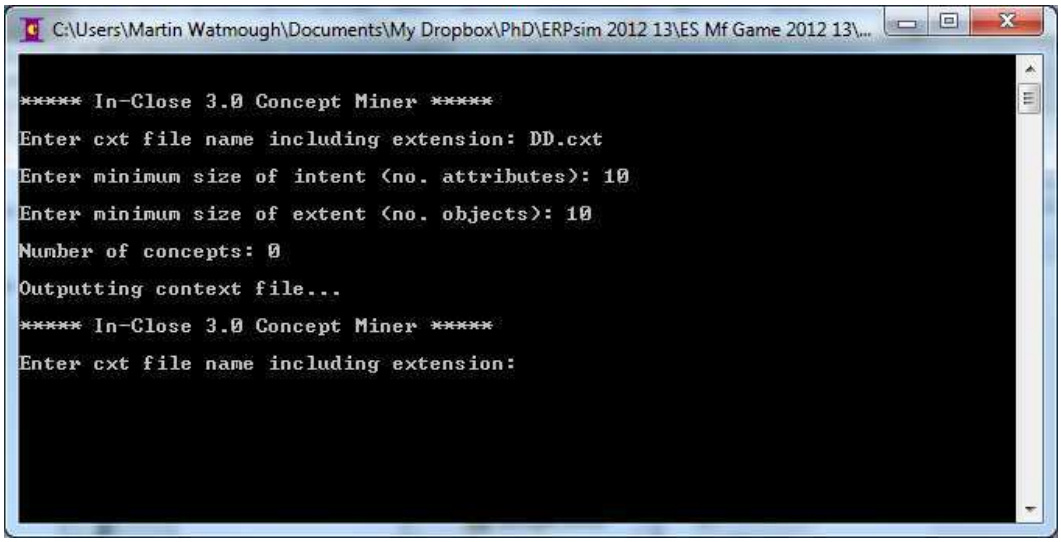


Figure 3.9: InClose2

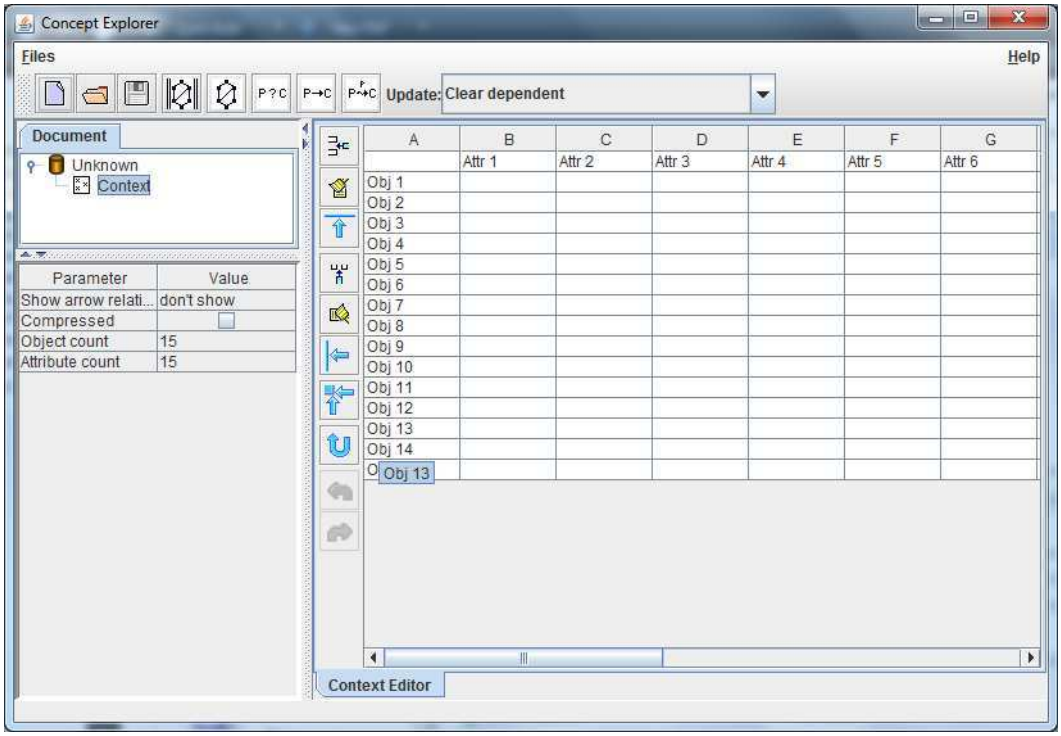


Figure 3.10: Concept Explorer

are linked. FCA is useful as it can be used to analysis raw data, find relationships and produce a converted format suitable for displaying graphically as a lattice. In the example many types of craft could have been analysed in order to produce a simple view of relationships as highlighted in the lattice, this is how it can be a powerful technique when applied to large data sets where relationships are unknown at the outset. Formally this is isomorphism, corresponding or similar form and relations (Oxford, 2012), a feature of the Galois connection.

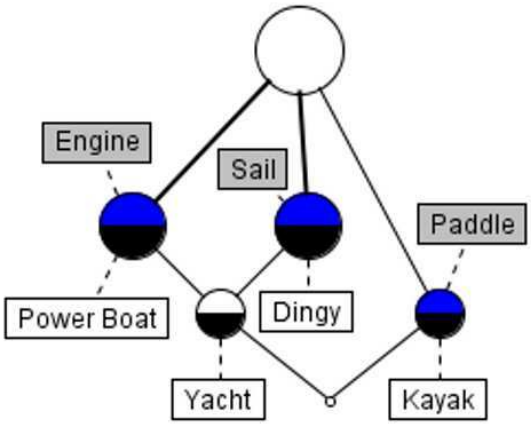


Figure 3.11: Simple Lattice

It is possible to read the concept table used to create the lattice, see figure 3.12. The same relationships between the object 'Yacht' and attributes 'Engine' and 'Sail' are visible in table; the creation of this table is fundamentally what FCA achieves. Two terms are important to understand at this point - Intent and Extent. The extents are the objects such as 'Yacht' and 'Dingy' that share the attribute 'Sail'. Intents are the attributes such as 'Sail' that share the objects 'Yacht' and 'Dingy'. A method of restricting the output to only larger sets of intents or extents is supported by InClose2, this applies a technique known as minimum support and restricts the content of the concept table. If a minimum support of two extents is applied and because 'Kayak' only has one attribute it would be excluded from the concept table, this is useful for

identifying prominent relationships in large data sets as discussed in section 3.3.1.

A	B	C	D
	Engine	Sail	Paddle
Yacht	X	X	
Dingy		X	
Power Boat	X		
Kayak			X

Figure 3.12: Concept Table for Simple Lattice

3.6 Concluding Summary

This chapter introduced FCA, its origins, formal definitions and philosophy. Identifying patterns, variations and classes are problems where FCA has been successfully applied to large and complex data sets.

FCA is an analysis approach that is mathematically sound and represents a useful technique for knowledge representation and information management. Representing formal contexts in tabular and lattice forms illustrates the logical structure, definition and grouping of relationships between data. A demonstration of the analysis steps describes how the analysis tools can be applied in the context of transactional data.

There are many analysis techniques available and no single solution will suit all purposes, however, it is evident that the growth in data is overtaking the speed by which humans and systems can derive understanding and knowledge.

We can therefore conclude that FCA is potentially useful for the discovery of knowledge in transactional data, particularly when applied to large and complex data sets. The complexity of user and system interaction will have a significant impact on successful applications.

Chapter 4

Findings from an LTA Design Experiment

4.1 Introduction

This chapter aims to provided insight into the discovery of hidden knowledge through a bottom-up analysis of transactional data. Using an opportunity to incorporate FCA into two degree courses at Sheffield Hallam University students participated in ERPsim, a business simulator based on an industry standard SAP enterprise system.

Learning, Teaching and Assessment (LTA) acted as an environment for problem solving, learning and study of FCA in action. An action research and a case study strategy iteratively developed an environment for generating experimental data in a controlled and observable situation. This first experiment used situation theory, problem based learning and a hybrid combination of Yin's Case Study Method and Bigg's constructive alignment. This ensured a good pedagogic outcome for the students thereby fulfilling the learning objectives and developing an in depth understanding of data. Appendix E expands on the overview contained in this chapter.

Moderated assessments provided data, tangible measures and text for qualitative analysis generated in an environment representative of informed and inquiring users. Supported by NVivo, software that supports qualitative and mixed methods of research

(NVivo, 2012), this analysis created an insight into the tools, methods and discoverable knowledge. Results include the knowledge discovered from complex systems and an assessment of FCA's ability to explore transactional data compared with contemporary tools. Permission and ethics approval, see section 1.4.1, was obtained for the use of assignments in this research.

Understanding the iterative steps taken by individuals or groups when performing analysis and discovering knowledge provides insight into the behaviour of users and processes far beyond any discrete answers. The output from the LTA activities discussed in appendix E provide this, they are attempts by the students to discover knowledge and demonstrate this by creating enhanced business process models. A critical evaluation of the tools and techniques also resulted.

4.2 Discovery through an LTA Design

Enterprise Resource Planning (ERP) systems are typically transactional systems that support the core functions within an organisation as introduced in section 2.2.1. SAP A.G. is one of the leading providers (SAP, 2012*a*). The analysis used transactional data generated by students during business simulations in ERPsim. This is an SAP A.G. ERP based simulation game that features competitive behaviour and increasing levels of complexity in a highly immersive and demanding atmosphere that reflects industrial practice (Leger et al., 2007). ERPsim Lab at HEC Montreal created and manage the software (HEC Montreal, 2011).

Gathering data from industrial systems is difficult due to the volumes and complexities involved. Rapidly the task becomes impossible due to hardware constraints and the scalability of tool sets. Moreover data is frequently restricted and organisations are unwilling to provide access. ERPsim is a dynamic and competitive simulation where human behaviour and decision making still determine the outcome but in a controlled and accessible environment. As a source of experimental data it represents a solid

foundation for the research.

Constructed on a standard database design that forms the foundation of SAP ERP systems applied within many organisations, ERPsim shares this common data structure. For this reason ERPsim based on SAP ECC 6.0 (SAP, 2011*b*) is considered as being representative of industry ERP systems. It has the advantage of being repeatable and relatively constrained; it has the disadvantage of representing only a fraction of the volume and complexity of a real enterprise system.

Training people in the concepts and application of SAP ERP systems forms a primary objective of the simulation software, therefore it must be representative in serving this purpose. It is therefore relevant to correlate this with industrial practice.

Creating value using semantic technologies is not significantly different to other technologies, three important aspects are the customers, business model and technology (Dominque et al., 2011*a*). This viewpoint formed part of the education as it is equally applicable to FCA in this context. Principles and ideas to achieve balance and sound business model such as considering Porters Value Chain (Porter, 1996) have also been included.

The simulation generated and captured data on which Business Intelligence was performed. The data generated by the simulation represents typical business activity and is not specifically for FCA, thus it provides a meaningful test of FCA in BI from ERP transactional data.

Permission and ethics approval, see section 1.4.1, was obtained for the use of assignments for this research. As described in appendix E careful consideration and controls were put in place maintain the module learning objectives regardless of this research. The introduction of this leading edge research topic was used to enrich the learning syllabus. Marking was aligned and reviewed against learning objectivities and required students to learn new skills and apply critical thinking, this was not influenced by the research aims which analysed the assignments in a different manner.

ERPsim is based on SAP ECC 6.0 which is an ERP system capable of supporting in

this example logistics and financial activities for a number of competing companies. All sales, procurement, master data, inventory, marketing and financial transactions are captured real time in additional to a limited number of reports to show sales, inventory, balance sheet and the profit and loss statement. These are transaction based reports and offer no analysis without the application of further tools. A detailed description of the design is contained in appendix E, an overview is presented in the following section.

Kang et al. (2011) proposed an approach for real-time monitoring of business processes through the application of FCA and reachability lattices. This is an interesting viewpoint that generates a probability based representation of historical event-patterns with the intention of applying these in real time for decision making. The core focus revolves around the probability of changing states and progressing toward the outcome, as a result, the reachability lattices are directional. This is based on the sequence of process steps and where it differs from this research, it is unknown if the sequence is an influencing factor in all cases. Kang et al. (2011) refers to the selection of the events, performance and handling unobserved events as the issues requiring further research.

4.2.1 Pedagogy

This research interests arises from how to discover the hidden knowledge within transactional systems i.e. how useful information or knowledge can be identified from mainstream database systems by applying and combining analysis techniques. To assist, at Sheffield Hallam University this research has been incorporated into two Computing degree course modules. The aim for the research is to be informed by the student's experiences, whilst enriching the student's knowledge in this topical area.

By applying and developing the approach to teaching transactional systems and analysis, two benefits are envisaged. Firstly, an insight into how learning these methods benefits the modules and students. Secondly, to engender a creative arena that encourages open answers from the students. Formal Concept Analysis is a technique

for analysing data in order to discover information and knowledge. FCA is particularly attractive in that offers an automated means of eliciting these concepts from the data (Wille, 1997) (Wolff, 1993). Therefore, FCA was selected as the underlining technique for designing learning in order to research the hitherto hidden knowledge in transactional data.

ERPsim has a strong pedagogic foundation that has been adopted and applied during the development of the degree modules. ERPsim is designed for active learning in that it achieves long-term retention. ERPsim takes advantage of Situation Cognitive Theory and Problem-based Learning (Feldstein, 2012).

Situation theory states that activities, tasks, and understanding do not exist in isolation, but rather are part of broader relation systems and that situated cognition is associated with a higher level of engagement and motivation in learners, thus generally leads to a better understanding and transfer of knowledge (Leger et al., 2011). Problem-based learning is a widely applied technique that has its origins dating back to 1966 in medical education (Hillen et al., 2010). It is a teaching strategy to promote self-directed learning and critical thinking through problem solving in which active participation and challenging problems in a relevant context are key (Ginty, 2007).

Furthermore, the learning environment created by ERPsim has been carried into the analysis of its output by comparative techniques. These techniques, described later, are used in order to evaluate the comparative value of FCA for transactional data.

To assess the effectiveness of this learning, teaching and assessment (LTA) we examined the marks achieved and learning objectives; the findings and feedback from the students have also been considered. It should be clarified that the students on these modules' did not have a significant mathematical background; rather the modules focussed on the business application of FCA. For this reason and to preserve consistency we ensured that the FCA tools were explained and applied according to their understanding. The fact that the raw data structure was constant aided the process.

4.2.2 Discussion

This section draws on the work and findings documented in appendix E. The intended learning outcomes have been achieved reasonably successfully. Students grasped the fundamental theories and applied them in a simulated context that is a representative example of real-work operations, particularly when the actual time scales are considered. A distinct understanding was developed between the simplicity of models and the challenge of identifying useful data and outcomes from a large data set.

The context and energy developed during ERPsim is inherently valuable in achieving the learning outcomes; it promotes the group dynamics, rapid learning and knowledge retention. The complete cycle, including a range of contemporary through to research level analysis methods was key to achieving the learning outcomes. Data preparation was a highly cited problem; however, it is anticipated that this will simplified into a data selection task with future generations of the software. The difference between industry produced solutions such as 'BI On demand' and the FCA tools was quickly highlighted by the students.

The case studies in the following text refers to four iterations of the module and assessment cycle between academic years starting in 2010 and finishing in 2012. The construction of the coursework is explained in section 4.3.4.

A number of unintended but valued learning outcomes were also highlighted in line with Biggs Constructive Alignment. There emerged an inherent value in the analysts (students) being involved in the data preparation, despite their raising this as an issue. Rather than just implying that it was unduly time consuming they appreciated the value in understanding the context, source and calculations that help discern towards extracting the transactional data. In passing there was little to differentiate the results from the analysis for the case study 1 where students complete the whole preparation task with case study 4 where students modified a generic preparation routine, thus enabling the students to focus on FCA. A further unintended but valued learning

outcome was how effective the tools would be when used in conjunction with each other instead of the separation of the tools as originally directed by the assignments.

Certain students found it difficult to grasp an 'incomplete' picture. The idea of determining rules helped somewhat but the data only provided a fraction of the complete set of rules. Partial cognitive models will probably be more common than a comprehensive understanding as the rate of change and volume of information increases making this potentially a topic for further research.

The capability of FCA for discovering the concepts and relationships in transactional data was repeatedly identified as a key reason for applying it. A lack of confidence was also cited by the students, particularly in the context of understanding what the analysis actually indicated. Frequently a number of repetitions were needed to clarify and subsequently accept the result. There were also some interesting remarks that expressed unexpected negatives about more familiar tools (Excel) when considering large data sets or potential 'big data' problems. This demonstrated that the key messages of the modules had been learnt and applied usefully in comparison with FCA.

The propositions of Presthus in describing why teaching Business Intelligence is challenging from the perspective of students and lecturers also emerged (Presthus, 2012). It is interesting (and comforting) to note that the approaches taken in this study happened to address to an extent these propositions. The propositions included providing a mechanism for reducing the level of abstraction when teaching and demonstrating the business value of BI. This lead to generating interest, effective learning based on suitable data sets, and the value of case studies.

The method and tools applied in Case Study 3 represents the most successful teaching methods to date for FCA in Sheffield Hallam's modules in terms of marks, table E.5 contains a breakdown. There was an evident improvement in the marks of the FCA sections but there are still opportunities to develop and improve the application of FCA based tools. The learning environment largely succeeded in providing students with those meaningful experiences that business analysts need. In particular it equipped

them with a well rounded experience, which is a significant factor. This reinforces Gartner's findings that analysis will be controlled by business units and not technical experts (Gartner, 2009).

4.3 Aims of Empirical Analysis

The empirical analysis in this section is intended to develop an understanding of the data generated and address a number of research aims described below.

As discussed in section 4.2.1, progressively refining the education and delivery method has enabled the students to focus to be on the creation of ideas and knowledge discovery. This has included refining the starting point provided, improving guidance, refining KPIs and providing or creating models to demonstrate discovered knowledge.

The assignments form the source data for this analysis. Generated within the LTA environment described in section 4.2, they have been designed to support the research question of discovering knowledge from transactional data using FCA. A principle aim of this chapter is to analysis, document and assess this knowledge.

Contrasting FCA and conventional techniques aims to understand and differentiate between the techniques, rationalise the advantages and reflecting on the disadvantages identified. Deciphering the operation of ERPsim can be considered knowledge discovery and an aim as gaining an understanding of forces leading to successful outcomes is desirable.

Developing an understanding of how conceptual models are used and created is the final aim. The system features tangled decision points, quantitative inputs requiring interpretation of complex systems with multiple dimensions and non-linear relationships. Deriving clarity and understanding in an effective and efficient manner is a challenging task.

4.3.1 Qualitative Data Analysis with NVivo

Text and visual information formats are not generally suitable for quantitative analysis; qualitative methods are more suitable particularly in situations where a detailed understanding of a process or experience is wanted (Bazeley, 2007). Qualitative research can be complex and is generally time-consuming Mason (2002). More efficient analysis can be achieved by combing the rigour of a methodological approach and functions available in the software tool NVivo.

NVivo formed the core analysis tool. Converting document formats can be required as some formats are not supported. By converting MS Powerpoint documents into Portable Document Format (PDF), text and graphics can be selected and organised within NVivo. Secondly, Excel has been used for calculations based on data extracted from NVivo.

QSR International, the developers of NVivo, have provided a set of tools that will assist in the undertaking of qualitative analysis in five principled ways (Bazeley, 2007). These are listed below and described in the context of this research:

- *Manage data:* Assignment documents follow some guidelines but they not rigid enough to support automated analysis methods, therefore, some manual input and selection is required for the inclusion and analysis of text and graphical elements.

- *Manage ideas:* Thoughts can be logged during the analysis such as observations during the coding exercise.

- *Query data:* Tools for structuring and analysing data are included.

- *Graphically model:* Results are combined with graphical representations and models.

- *Report from the data:* Sources, categorised and coded data is summarise and used to present information in a structured and refined format.

4.3.2 Assumptions and Reflection

Good practice is to to record assumptions as part of qualitative methodologies (Bazeley, 2007) and reflect upon. With this in mind factors potentially affecting the analysis have been considered.

Terminology is a significant factor, ERPsim is based on SAP ECC and heavily influenced by SAP terminology. The students are generally not from an industry background or experienced in such application domains. This combination can potentially lead to inaccuracies when interpreting and applying terminology. Expanding on this point slightly, English is not necessarily the students first language and a small number have recognised communication problems such as dyslexia; these factors did not appear to influence the outcome but it was considered.

Bias is considered to be an important factor. Bias due the use of ERPsim was expected as it is designed to be an absorbing and self contained environment. The core research focus of FCA can impart a bias, this was mitigated through the comparison against contemporary tools and open questions. The aim for students is to discover knowledge, null answers providing the steps are demonstrated, explained and argued are equally valid.

The act of preparing working data models will have an influence that should be gauged, outcomes that are unexpected are noted. Personal conceptual models and knowledge are also sources of bias and possibly conflict with perceptions and knowledge about the operation of ERPsim.

During the actual ERPsim experience the students focussed on achieving the aim of the game in a competitive environment. Analysis came after with contextual knowledge and experience of partaking in the game.

The approach itself is a bottom up approach, starting with the data using emergent theory linking into action research. This risk losing focus on the objectives and goals are magnified as interesting aspects of analysis are followed.

4.3.3 Design of FCA Method

To understand the challenges associated with applying FCA to transactional data a set of representative data, itself generated by student groups using ERPsim, provides the data for experimentation in order to develop a process and draw conclusions. Figure 4.1 illustrates the basic decision points, process generating data and finally the application of KPIs. A combination of output data, essentially data extracted directly from the database, in combination with KPIs is termed the source data for analysis.

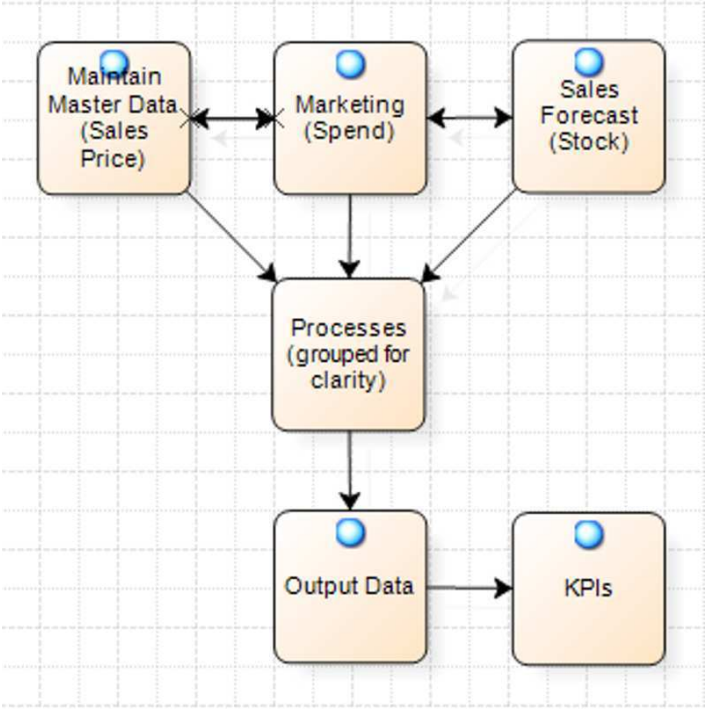


Figure 4.1: Simplified Process and Data Model

The usefulness of this approach is discussed with respect to its application from the students perspective, future iterations of the simulation and in the context of real organisations. This section refers specifically to the data preparation and analysis, the basic FCA tool set applied is as described in section 3.5; FcaBedrock, In-Close2 and Concept Explorer.

Data that contributed to this strategic goal was extracted and subsequently analysed in order to identify formal concepts. The intention of the approach was to provide

a semi-structure mechanism that also facilities the retrieval of unstructured data, any data with a loose association could be retrieved using this method regardless if it contributed to the outcome or not.

A simple chaining format linked the extracted the data in a format suitable for analysis by FCA. The aim of the simulation was to make a profit by controlling three decision points. These being selling price, marketing spend and forecast as shown in figure 4.1. As the goal of the organisation was to make a profit the final profit figure at the end of the simulation linked the outcome with key transactional data.

As the ERP system is effectively a relational database with data held in joined tables it is possible to extract data that contributed towards a goal via a SQL query. The table relationships supported the extraction of data that contributed to towards the outcome. For example, all sales transactions within the time-period could be found via the connection from billing through the outbound shipments to the sales orders. Correspondingly individual sales order profit based on the materials cost price could also be extracted.

The approach discussed in appendix E, based on ERPsim and FCA, achieved an analysis of transactional data from a mainstream ERP system. All data being generated without knowledge or architecture designed for FCA. The results have been validated by conventional analysis and an intuitive understanding of the simulation.

This approach has concluded that applying FCA to ERP systems has merit, however, an underlying understanding of the data from a relatively simple model is a key factor in the successful. Analysing a complex system would require an understanding of the data and careful consideration of values and ranges for representation in the lattice. It is difficult to navigate analysis and focus on key topics in order to derive this understanding and software development would need to support this.

It is clear that calculated fields and the inclusion of performance measures aid the graphical analysis and help to differentiate data with different meanings. This could be viewed as a primitive means of adding semantic information. The calculations and

performance measures had to be added within MS Access as this provided the tools for adding logic and queries. No reason has been identified why these could be determined or added at any stage of the process particularly within the graphical stage, this would effectively enable real time interrogation of the lattice.

When starting the analysis it is difficult to understand what to analyse, a situation that will be dramatically more complicated in real-world applications than within the constrained simulation with limited input and output variables. With further refinement relationships within the data could be used to determine the process at multiple levels, holistically as a business process model and focused by including individual users and transactional events. This would enable FCA to build an understanding of the processes, instead of requiring knowledge of the relational database. This idea is developed further and applied in chapter 5.

A significant advantage of the extraction method and using SAP ECC as a basis is the standardised construction making it relatively easy to utilise different systems.

4.3.4 Coursework Design and Alignment

As a reminder of the situation, students have played the game, they have a context and expectation of the questions that forms the analysis. They understand that it is a discovery exercise aimed at understanding the rules, decisions, inter-company dynamics, competitive forces and even strategies for deploying in the future.

Two courses at Sheffield Hallam University have been used as a vehicle for this research; under graduate 'Smart Applications' and post graduate 'Enterprise Systems' modules'. The aim is to be informed by the student's experiences while enriching the student's knowledge, as discussed in appendix E the development of an environment capable of supporting this was integral to the research.

Smart Applications aims to introduce frameworks and techniques for representing and reasoning with knowledge for smart applications (Sheffield Hallam University,

2010). Enterprise Systems aims to build an appreciation of current and future thinking on enterprise wide systems and ERP software (Sheffield Hallam University, 2012). Both modules share an interest in analysis for supporting applications that enhance individual and organisation performance. Smart Applications focuses on knowledge and reasoning with a broad technology base. Enterprise Systems as the name suggests focuses on ERP and BI technologies in the context of modern, competitive organisations.

Evolution of the environment and approach resulted in changes to the structure and content, however, the alignment and coverage of the research aims against the learning environment was maintained. The alignment is shown in table 4.1 and table 4.2 for Smart Applications and Enterprise Systems respectively. This was used to ensure consistency and relevancy for the research while being flexible enough to support the modules' learning objectives.

It should be noted that only the elements of the coursework relating to this research have been covered, a multiple choice phase test and group presentation also formed part of the assessment criteria for Enterprise Systems with content targeting the general module content.

Maintaining a document structure suitable for automatic coding in NVivo was considered but deemed uncontrollable and not pursued given the nature of the assignments, particularly in respect to the visualisation required. Instead, the intent behind the assignments was to encourage students to discover patterns and classify data or information in search of knowledge and support the use of coding in NVivo.

Visualisation is an important theme and links with the presentation of discovered knowledge, it also supported an understanding of underlying concepts from an early stage. SAP ECC basic process flows, figure 4.2, represents a basic overview of the processes that generated data and an indication where decisions take place. This example has been created to represent the simpler version of the game ERPsim Distribution. It was the intention that this would formed the background knowledge and a basis to

Assignment Section	**Node Reference to Research Aims**
Introduction	1
Excel	2 3
FCA	2 3
Evaluation & Conclusion	4 5
Presentation	4

Assignment Section	**Node Reference to Research Aims**
Introduction	1
Analysis	2 3
Defining Rules	2 3 5
Evaluation	3 5
Conclusion	4 5
Presentation	4

Table 4.1: Smart Applications 2010-11 (top) and 2011-12 (bottom)

start constructing a conceptual map, in effect the context and relationship aspects.

In practice, the processes flow continually around in various directions, loops and repetitions with different rates and frequencies. Representing a complex organisation as stated is an over simplification for the purposes of understanding and education, it is useful for highlight relationships and context.

4.3.5 Working with Data in NVivo

The assignments form a set of data sources that NVivo can hold as sources in both Word and Adobe documents formats, see figure 4.3. The module and year also formed part of the folder structure, this was replicated in the document classification. It is necessary to convert Powerpoint documents into Adobe for NVivo, however this maintains text and graphics therefore supporting analysis in NVivo. Figure 4.4 contains a view of an imported Powerpoint document and diagram.

Saldana (2009) describes a code in qualitative inquiry as a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data. Salana continues to describe

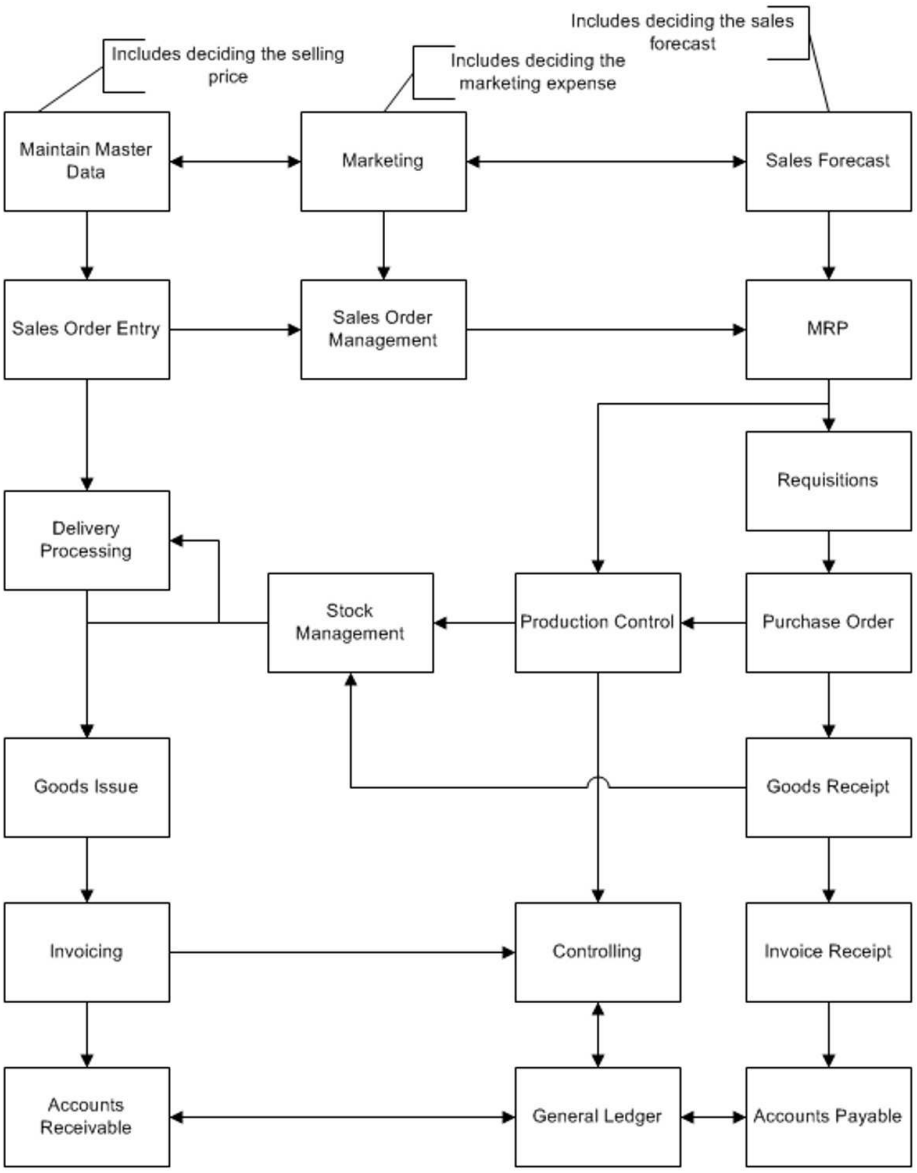


Figure 4.2: SAP ECC Basic Process Flows

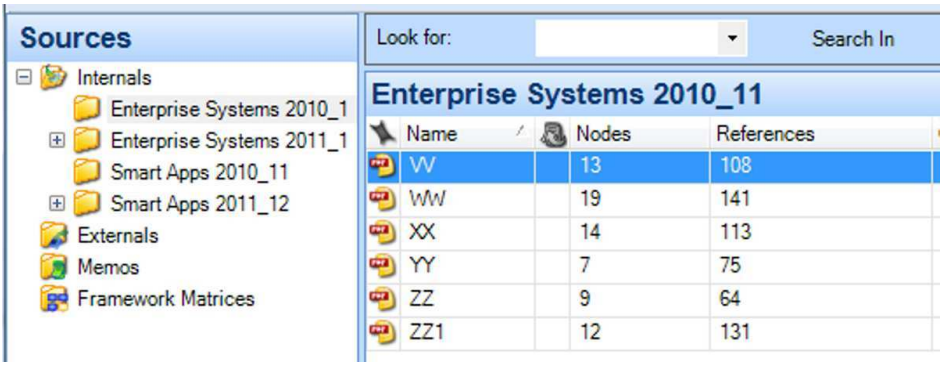


Figure 4.3: Sources in NVivo

Assignment Section 2010-11	**Node Reference to Research Aims**
Introduction	1
Individual Company Analysis (Excel)	2 3
Multi Company Analysis (Excel)	2 3
Individual Company Analysis (FCA)	2 3
Multi Company Analysis (FCA)	2 3
Evaluation and Conclusion	4 5
Presentation	4

Assignment Section 2011-12	**Node Reference to Research Aims**
Introduction	1
Analysis (Excel, BI and FCA)	2 3
Solution	2 5
Justification and Detailed Benefits Case	4
Conclusion	4 5
Presentation	4

Table 4.2: Enterprise Systems 2010/11 (top) and 2011/12 (bottom)

an iterative process of coding in a variety of methods, the primary content and essence of the datum being represented by the code. Bazeley expresses this slightly differently in that coding is a way of linking data to ideas and from ideas back to supporting data (Bazeley, 2007) although both agree that coding is not a precise science, it is primarily an interpretive act.

Figure 4.5 displays a screen shot of the top level codes applied to the assignments and cross references to the research aims as defined in tables 4.1 and 4.2. For convenience these are noted in brackets against the header title. A complete list is contained in appendix D. The source and references columns represent statistics about the coding. 'Source' represents the number of documents sources coded, in this case assignments. 'References' is the count of all coding from the documents. This number is generally significantly larger than the number of sources and does not represent a direct link between the selected topic and node.

Coding with reference to the research aims provides a good basic structure and starting point. The quantity of nodes expanded during reflection as patterns emerged

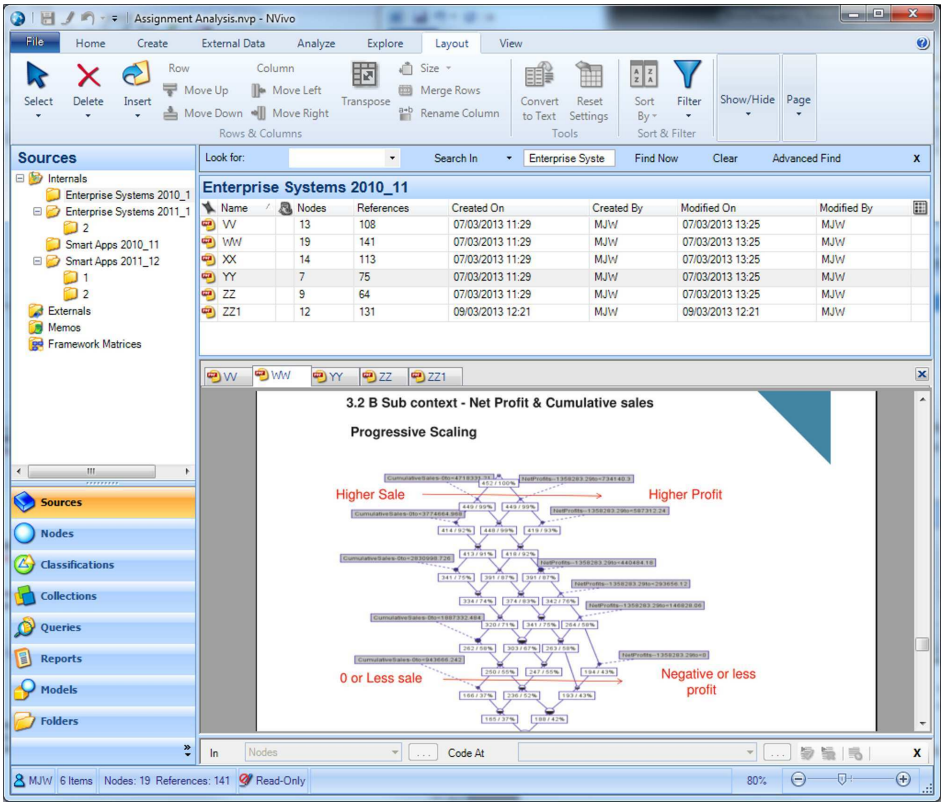


Figure 4.4: Sources in NVivo

leading to ideas and further classification. Multiple review cycles, categorisation and the qualitative analysis methods applied aided the identification of missing or misinterpreted items.

Coded text or diagrams can be viewed in a collated format, see figure 4.6. Typically this is generated on demand and analysed iteratively during the coding exercise. Reading and coding every document forms a time consuming activity. As the node structure represented a hierarchy aggregating statistics to the top level provided an overview while maintaining detail and sub-division at lower levels.

The coding represents the data and basis for further analysis, reviewing and checking for omissions and consistency requires a methodological approach. Calculating the percentage of sources coded and secondly coding similarity provided measures and a level of confidence in the coding exercise. These two dimensions indicate differences in the application and spread of coding across the source documents.

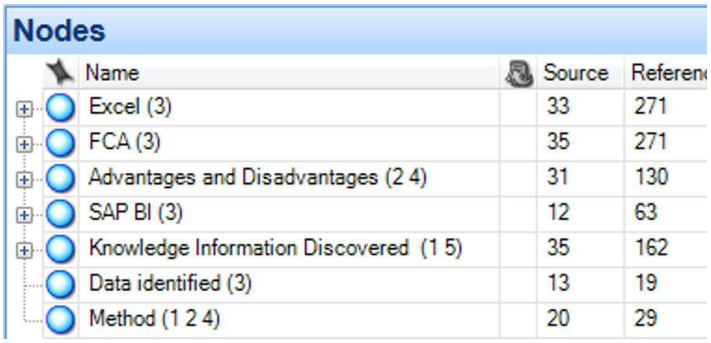


Figure 4.5: Top Level Codes Applied in NVivo

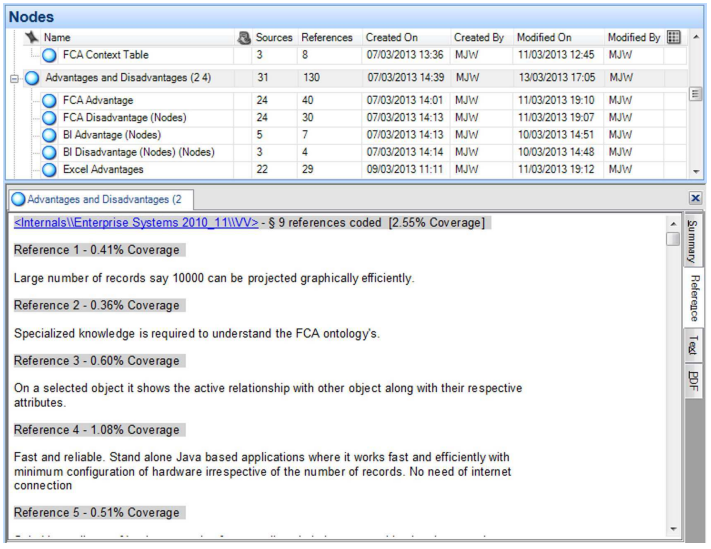


Figure 4.6: Nodes and Example Coding in NVivo

Table 4.3 contains the coded nodes, number of sources referenced and a calculated percentage for the number of nodes as a proportion of the total number of sources. Three nodes of source coding have relatively low percentages, the others are acceptable as a basis for continuing the analysis particularly in the knowledge that these are based on student assignments where there is a spread of marks achieve.

The low percentage for 'SAP BI' is explainable as it was only introduced as an analysis tool for one year of one module. The low scores percentage for 'Data identified' is an indication of where the approach taken was not identified in the assignment, possibly indicating pragmatically or intuitively gained knowledge The final section, 'Method', leans towards actual or suggested improvements to the analysis method and as such an advanced topic for the students.

Figure 4.7 displays a section of a hierarchical diagram relating and clustering source by coding similarity. This query indicates the range and depth of coding against each individual source, for convenience each source is labelled with course and year.

A total of 46 sources are included in the analysis, of these 34.8% are clustered with 5-7 levels of coding similarity and 60.9% with 8-12 levels. The remaining 4.3% represented sources with very little coding, these assignment achieved a low grade and therefore correlates with a low level of coding.

Considering these points the coding exercise is deemed suitable as a basis for further analysis.

Coded Nodes	**Count of Sources**	**Percentage**
Excel (3)	33	72%
FCA (3)	35	76%
Advantages and Disadvantages (2 4)	31	67%
SAP BI (3)	12	26%
Knowledge Information Discovered (1 5)	35	76%
Data Identified (3)	13	28%
Method (1 2 4)	20	44%

Table 4.3: Summary of Sources Coded

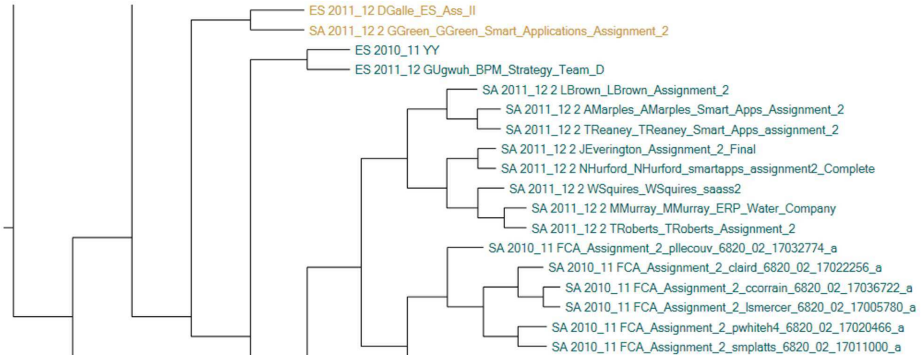


Figure 4.7: Sources Clustered by Coding Similarity

4.4 Empricial Anaylsis

4.4.1 Word Frequency

Word frequency is a method for identifying possible themes or words used by a particular demographic (NVivo, 2012). As a significant proportion of the assignments are text based word, frequency based on coding reflecting the research aims instead of demographics aims to reveal differences between FCA and contemporary tools.

The "Knowledge Information Discovered" node, see figure 4.6, has been coded to with the intention of enabling a comparison between FCA and contemporary techniques. NVivo's word frequency query has been performed for each node, see figure 4.9, the top one hundred words with a minimum length of four are found. Words that add little value are excluded by applying stop words, see figure 4.8. Stop words are common words that are useful for grammatical purposes but to a much lesser extent when searching, the list applied is a set created by Lingras and Akerkar (2008).

Two sets of words have been created applying the the query criteria on the respective nodes of data, FCA and Excel/BI. A short sample is shown in table 4.4. The core or exact word is shown on the left and stemmed words and synonyms along the row. This level of matching is deemed to be appropriate as meanings are at a similar level to the core word.

Further options include specialisms and generalisation but these could match based

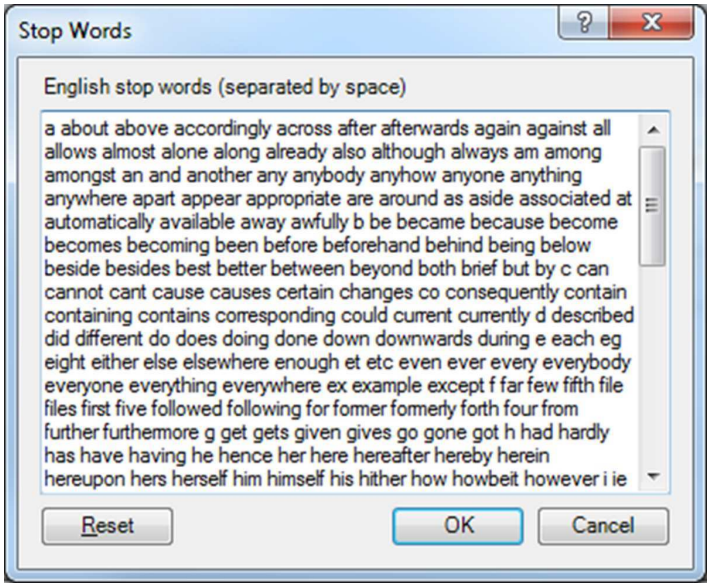


Figure 4.8: Stop Words Applied in NVivo

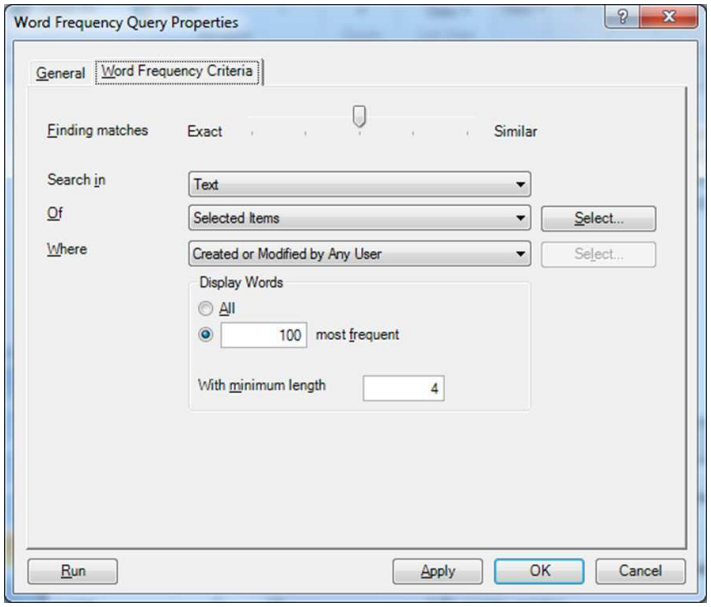


Figure 4.9: Word Frequency Query in NVivo

Core Word	**Stemmed Words and Synonyms**					
lattice	lattices					
give	giving	leave	makes	making	passed	reached
part	percentage	region	regions	section	sections	
found	establish					

Table 4.4: Section from FCA Word Frequency

on criteria that are to generic and overlap. These points have been tested by expanding the matching criteria. This caused the number words returned to expand to forty one from nine. This in turn made subsequent analysis more complex without any obvious benefit. Overlap was observed where words relating to the research question, for example 'lattice', is grouped with unrelated words including company, future, balance and material. Occurrences of words belongs to multiple sets became a common feature making a comparison as described virtually impossible.

By comparing the contents of each row against each list the words have been categorised as appearing in both or one node. Starting with word sets in common, table 4.5 displays the top 25 frequent words found in the 'FCA' node that are also found in the 'Excel/BI node'. The complete set of synonyms has been limited slightly for presentation purposes, the complete word set for all combinations are contained in appendix B. The column labelled 'Matched' indicates the words that appear in the top 25 frequent words query for both sets of words, a high proportion at 68%. There is clearly a correlation, an inspection of the words suggests a number of reasons for this.

Words including profit (goal), marketing/price/stock (user input) and decision (game) forms the foundations of ERPsim, the source of the data. It is no surprise that the these feature highly, however, it does provide confidence that the word frequency method applied is generating a sensible output.

The opposing analysis view to words found in 'FCA' and 'Excel/BI' is where words only appear in one or the other. The same method has been applied to matching any stem or synonym within the set of words and select the top 25 by frequency; the complete list is contained in appendix B. Examining the word sets relied on a manual

Matched	Core Word	Stemmed Words and Synonyms			
X	profit	positive	product	products	profit
X	marketing	market	marketing	sell	selling
X	shows	appears	establish	proved	proves
X	quarter	quarter	quarters		
X	amount	amount	number	quantity	total
X	companies	companies	company		
X	price	price	priced	prices	
X	results	answered	answers	effect	leads
	daily	daily			
	made	made			
X	products	output	product	products	
X	sales	sale	sales		
	units	combination	combined	units	
	high	extreme	extremes	high	highs
	time	time			
	cover	continue	cover		
X	spend	expenditure	passed	spend	
X	data	data	information		
X	higher	higher			
X	stock	inventory	stock		
X	analysis	analysis			
	figures	figures	forecasting	forecasts	number
	average	average	mean	meaning	medium
X	need	demand	necessarily	need	needs
X	decisions	decision	decisions	finally	

Table 4.5: Matched Word Sets between FCA and Excel/BI

process based on knowledge of both analysis techniques and ERPsim in an attempt to categorise the words.

The first category to eliminate are those specific to the analysis method such as lattice as the reason for not matching seems obvious, 'Excel/BI' in this example is unlikely to include lattices.

The second category to eliminate are those due to errors or technical terms that should actually relate to another term. Examples of these respectively are 'dailyprofit' that should be two words and 'matnr' which is the database definition for 'material'. Also included in this group are words that should really be included on the stop words list.

Determining possible themes is subjective, however, ideas for further research and investigation are identified. Words unique to 'Excel/BI' involved many that could be associated with measurement, movement or trends. Examples of this include highest, observation, lowest, running and long. This is in contrast to 'FCA' with words that express a more discrete basis for the analysis such as days, affected, attributes and part. This would appear to indicate that FCA has been used to focus in a manner more akin to cause and effect than trends. Given that chart based analysis over a time period was a common approach within 'Excel/BI' this is not overly surprising but interesting when considering that they shared the same source data. It may be possible to focus FCA on trends but this may require further intermediary transformations and a change to the method.

Attempts to model tools against decision points proved unsuccessful, see figure 4.10. Starting with the basic model, figure 4.1, links between tools and decision points only indicated that FCA, Excel and BI were applied across all areas. There is no clear demarcation of areas or topics evident.

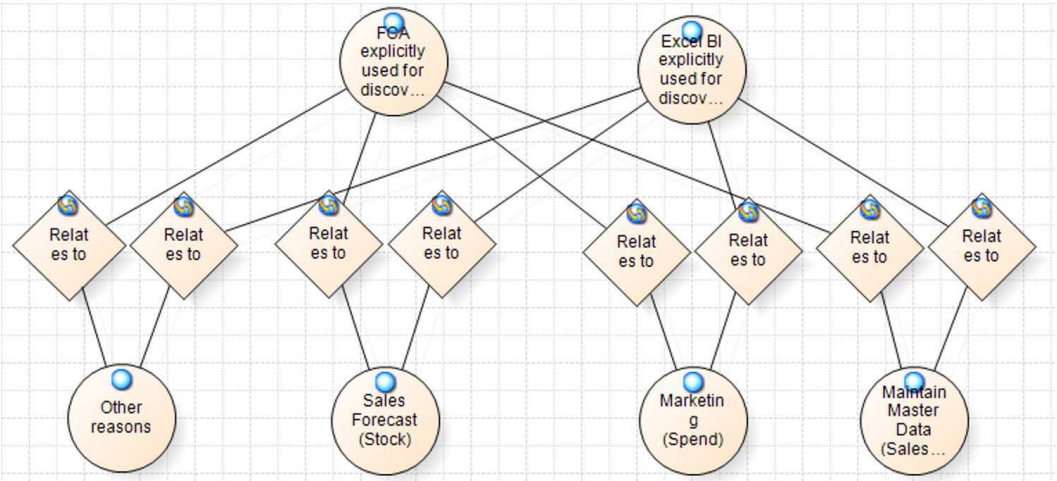


Figure 4.10: Decision Points Mapping Model

4.4.2 Choice of Data

Thirteen sources are coded as having identified data, see figure 4.5. This can be categorised into three types, each gradually refining the level of detail. Firstly raw values (inputs) that are viewed to contribute to the output (goal) are identified and applied as the basis for the analysis. Secondly, the process is considered and known forces, following principles similar to Porter's five forces analysis (Porter, 1985), are used to identify data. Finally insight is included thereby intuitively including factors that might determine the result, in the examples differentiation by product falls into this category. This represent a pragmatic approach based on experience, conceptual models and intelligence to create ideas for inclusion in the analysis.

It is difficult to draw any conclusions about the use of raw data versus KPI's, both have been demonstrated to be useful but this appears to be very subjective and dependant on the analysis and visualisation technique applied.

4.4.3 Choice of Tools and Visualisation

The aim of this section is to analyse the discovery process applied by comparing the tools and techniques. It is acknowledged that failed attempts are unlikely to be feature in the assignments. Also the BI tools were only available for one module (14 assignments). Observations collated during the coding are discussed followed by a comparison based on the coding.

During the coding exercise, the same pattern of analysis can be repeatedly observed. Repeating the use of a certain chart type between three and five times with only minor variations. This appeared to be an effort to reuse successful techniques as much as possible rather investing in the time attempting a different approach. It could also be perceived as a lack of imagination or confidence. The difficulty involved in learning or performing an analysis technique could also limits the range versus the requirements and marking scheme of the assignment.

The use of tables and charts with annotation to present or compare facts is a common approach. Representing knowledge through an assignment is a challenging task but many examples stop after highlighting a fact rather than continuing, expressing it as knowledge by using techniques such as rules, models or as a measure of usefulness.

A repeated failing is trying to present large volumes of information that leads to confusing and uninterpretable or irrelevant charts. Irrelevant in this context is taken to mean not linked or mentioned within the assignment, it may have played a role in the discovery process but it's inclusion does not support the point being made.

Table 4.6 contains statistics relating to the source coding of analysis tools. The percentage represents occurrences of each software package contained in the students' assignments. This data is exported from NVivo and percentages calculated in Excel. Figures under 'Total' indicate the sum of percentages within the box purely for convenience, some rounding differences are evident as decimals are not shown for clarity. Pivot tables and decision trees are not available in the version of SAP BI available therefore totals have been reduced by this 21%, for example 13% reduced by 20% equals 10%, in order to balance the comparison.

The use of data in a tabular form is low for Excel (9%) and FCA (8%) suggesting a preference for graphical analysis even though process of preparing data has utilised a table in some form as an intermediate step for both. It should be noted that SAP BI typically includes table and charts in parallel therefore zero occurrences of only table use is not significant.

A notable difference is the application of bar charts and line charts, almost opposite in there frequency 17% to 56% and 44% to 14% between Excel and SAP BI respectively. It is unclear if the tools themselves influences this favouritism or if the inherent combination of table and chart within SAP BI had a bearing on the knowledge representation. The frequency of annotation was virtually identical across Excel (21%) and SAP BI (22%).

Mandating the use of FCA as an analysis method obviously influenced the high

figures, the low frequency associated with the context table indicated perhaps a missed opportunity for discovery.

The next section discusses if the tools used have an influence on knowledge discovered and if it can be proved that a mix of approaches is beneficial.

Tools	**Excel**	**Total**	**SAP BI**	**Total**	**FCA**
Table	9%	-	-	-	-
Bar Chart	9%	-	22%	-	-
Bar Chart Annotated	4%	-	0%	-	-
Bar Chart and Table	4%	-	35%	-	-
Bar Chart and Table Annotated	0%	17%	13%	56%	-
Line Chart	28%	-	9%	-	-
Line Chart Annotated	17%	44%	9%	14%	-
Bubble Chart	2%	-	0%	-	-
Pie Chart	7%	-	0%	-	-
Pie Chart and Table	0%	7%	13%	10%	-
Pivot Table	4%	-	-	-	-
Decison Tree	17%	-	-	-	-
FCA Lattice	-	-	-	-	92%
FCA Context Table	-	-	-	-	8%

Table 4.6: Frequency of Coded Applications by Analysis Tools

4.4.4 Discovered Knowledge

To ascertain what knowledge had been discovered and documented by the students relationship nodes based on the coded nodes were applied in NVivo. These compiled and ordered documented knowledge with the aim of creating a conceptual model. Relationship nodes record a connection of a particular kind between two project items (Bazeley, 2007); a simple 'relates to' statement creates a relationship between coded nodes and the tool set applied. One practical mechanism for demonstrating and applying knowledge was communicated through process flow charts. An example created by one of the students is shown in figure 4.11 and demonstrates how logical statement from discovered knowledge have been incorporated into process for decision making.

Appendix C contains the complete results from qualitative analysis of knowledge

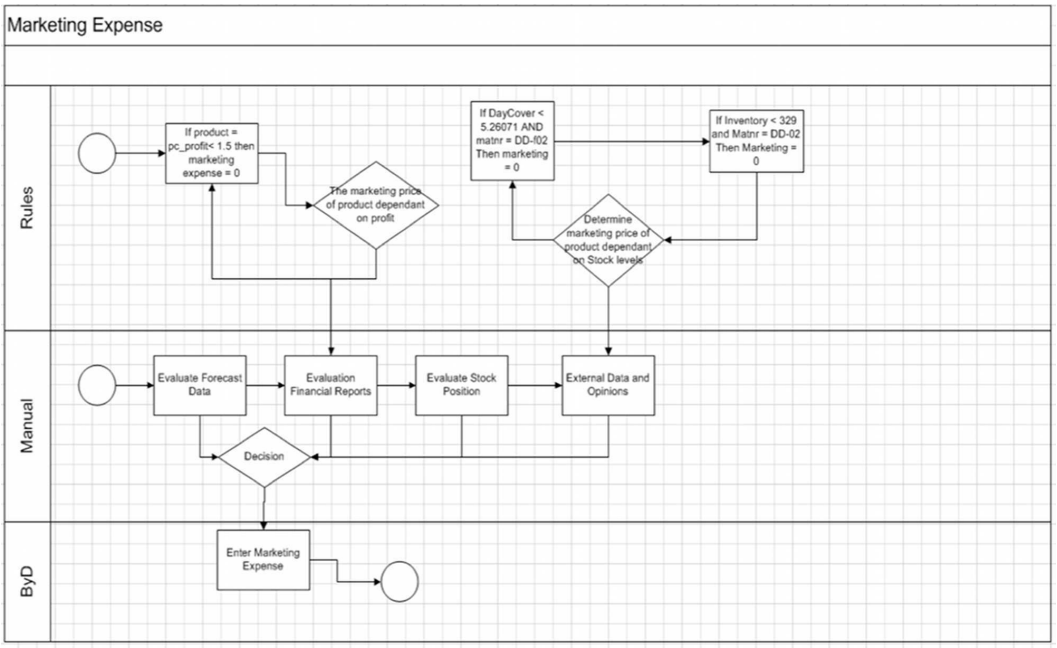


Figure 4.11: Example of Student applying Knowledge

discovered and documented in the students' assignments. The results are based only on knowledge that can be explicitly associated with the technique and demonstrated within the assignment. The figures are not deemed large enough for any quantitative analysis, however, they can contribute partially towards a conceptual model and discussion points.

Each point identified is categorised against the descriptions shown in the key, see Appendix C. These are listed in a descending order of complexity from facts, through general rules (rules of thumb), formulaic representation, consideration of other factors and finally with points that build on the method applied or highlight factors not presented as variables in the assignment.

The findings discussed below are not comprehensive but they highlight a number of points for further investigation. All methods discovered knowledge about ERPsim to some extent. 'Excel/BI' appears to generate more general rules, the hypothetical reason for this is due to the general level of understand and pre-existing skills.

Knowledge discovery within FCA appears to more useful with respect to discrete

data, indicated by a larger proportion of facts versus lower counts of rules and relationships.

Knowledge discovery within FCA also identified areas where the method could be changed or improved, a point not identified in relation the other techniques. It is acknowledge that the FCA tool-set is more difficult to apply than Excel or BI. There may also be an unstated assumption that these tools are complete and that there is no scope for modification or improvement.

The discovery of 'unexpected' knowledge highlights a number of interesting points, see table C.8. Unexpected is defined as an aspect of the game that is not documented as a feature of ERPsim. Alternatively it is novel or unique. The actual operation and competitive elements of ERPsim is not published beyond the decision points highlighted to the students, see annotations at the top of figure 4.2.

Result can differ depending on the medium (FCA, Excel or BI) used for the analysis. Assuming all methods are applied to the same data set without errors during the analysis, it seems illogical that the actual result can be different. What this may infer is that the interpretation of results is different, a point that is logical and understandable. The second unexpected discovery is the identification that certain products sold better than others, this is not a known feature of the simulation.

4.5 Method Evaluation

This section collates and summaries content coded across all assignments that relate to the knowledge discovery methods applied; for clarity text copied directly from the assignment is indicated by *italics*. The intention is to minimise misinterpretation and bias by adopting grammar and language copied directly from the assignments, spelling and grammar within the *italics* has not corrected.

Comparisons, difficulties and positive comments are highlighted and evaluated

against theories in the context of the data source, discovery techniques and requirements for expert knowledge. The range and alignment of coding against relevant theories, FCA and knowledge management provides a level of confidence in the evaluation.

4.5.1 Assignment: Data

The data source featured erratic and noisy data, typically this was true at the start of each game and deriving meaning was difficult. In the context of ERPsim this is true and essentially part of the learning cycle. Reinforced by the requirement to derive rules and understanding about the processes involved. The ability to identify and move beyond this data or time period is very positive.

FCA helps to understand the meaning of data and the key knowledge behind it, even across perceived unrelated categories. This acknowledgement of discovery and maintaining context across large data sets demonstrates the usefulness of relationships and aligns with the interest in FCA as an approach.

4.5.2 Assignment: Discovery Techniques

The analysis techniques applied with the aim of discovering knowledge resulted in different opinions about their suitability. MS Excel particularly useful features include sorting, filtering and the range of visualisations available. Comments about suitability echo the points made about knowledge discovery in section 4.4.4, primarily differentiating between discrete and continuous data. The ability to manipulate and dissect data by discrete ranges either in tabular or graphical format is highlighted as a key feature supporting interactive knowledge discovery.

Microsoft based tools also integrated well and enabled the production of simple and readable graphics easily although difficulties analysing and managing large data sets were experienced. The ease and reliability with which data can be transferred between applications is testament to the maturity of the applications but given rigorous design

this should be achievable between any applications assuming fundamental foundations are shared. Simplicity is a valued feature, it can be inferred that when the management of data is manual it can become a challenge.

The strength of visualisation was cited across all techniques and analysis by visual inspection of graphics is common. Decision trees formed that only exception to this because the text based rules are generated. Frequently knowledge was presented using this formula based representation as it can be easy to document and understand. Common conclusions were derived from both decision trees and FCA but excessive detail in the output from decisions trees caused complications and a lack of understanding. There is not an ideal answer, formulaic rules are valued at the appropriate level of detail, determining this level is difficult and frequently reliant on knowledge working and interactive tools.

Difficulties with FCA included managing the number of concept and creating meaningful lattices notwithstanding the effort require preparing data and actually using the tool set provided. A better user experience may be possible by combining Excel and FCA was suggested. This highlights the well designed navigation functions and presentation options. Potentially by combining views of data in various forms including tabular and graphical further improvements can be made to FCA's usability.

Finally FCA was viewed to support better insight into trends, focussed analysis and provides different perspectives than possible through Excel and BI. Generally FCA was considered a better technique for the discovery of hidden concepts from large data. An interesting viewpoint attributed this to this is the *manner that is matches the conceptual approach to problem solving taken by the analyst, partially by the visual link and hierarchy maintained between multiple objects.*

Minimum support within FCA, the ability to refine strong relationships, aided the identification of influencing factors and facilitate reasoning into their effect on outcomes. A similar comment was made under the data section, refinement and identification of pertinent information, or at least prominent relationships is an aid to users.

4.5.3 Assignment: Expert Knowledge

Identifying specific data to analyse was difficult and required expert knowledge. In some respects this is to be expected with a bottom up approach.

The inclusion of KPIs, particular cited with reference to FCA, made lattices easier to interpret. Intuitive KPIs made identifying, for example low profit, relatively easy. Determining the KPI categories initially is the challenging aspect.

SAP BI provided fast and efficient overviews without requiring expert knowledge. The user experience and automated data preparation significantly contributed towards this comment. Importing SAP data into an SAP product carries a level of meta data and quickly translates into graphical outputs.

4.6 Concluding Summary

In this chapter, empirical analysis of assignments has been used to compare and contrast FCA against contemporary analysis tools and elicit discovered knowledge.

This LTA approach has revealed the need for integrated tools that support knowledge discovery in a collaborative model with complex data; the success of these tools will be based on far more than their discrete technical capabilities.

This approach has concluded that applying FCA to ERP data has merit, however, an underlying understanding of the data from a relatively simple model is a key factor in the successful. Analysing more complex systems would require an understanding of the data and careful consideration of values and ranges for representation in the lattice. Navigation and analysis while maintaining a focus on key topics is challenging and software development to support this is needed.

A case is developed for deploying discovery techniques in transactional systems to compliment conventional and current industry based solutions. Actual deployment of the techniques is still in need of refinement, however, the underlying principles have

merit and practical application.

Applying FCA is shown to be challenging but results included more reflection and observations around the method than with contemporary tools. Findings outside of the anticipated or model answers to the assignments were also more prevalent. This is partially due to the less prescribed nature of the analysis and students not having pre-conceived ideas about how FCA should be performed, this is in contrast to spreadsheets where they typically have had experience and education.

The challenges faced through teaching and also those experienced by the students has clear parallels with the implementation and adoption of such tools in the work-place. Comparing and contrasting the techniques that have proven to be successful in the classroom to the business world would be an interesting research topic, as would addressing the problem of managing incomplete information and models. From this education experiences we can envisage that FCA has an important role to play.

Chapter 5

Knowledge and Relationship Discovery from User Activity

5.1 Introduction

ERP systems support the core and many other business functions. Enterprises invest significant resources into systems during implementation, ongoing maintenance and actual use. These systems control operations through to integrating with business partners and should be leveraged for any competitive advantage available. A frequently overlooked source of data are the user transaction logs, this is the use case for this chapter.

This chapter explores how the application of FCA as a discovery mechanism to user transaction logs offers an insight into the actual and patterns of use. The logs are not available within the ERPsim system used in chapter 4 and appendix E, therefore data extracted directly from an anonymous organisation's productive SAP ECC 6.0 system provides an actual industrial scenario and sample of real-world data.

5.2 Enterprise System Use Case

Enterprise Resource Planning (ERP) systems contain massive amounts of data that is frequently under utilised. This chapter presents a method for discovering useful and semantic data and knowledge in a practical manner that demonstrates the potential applications of this approach. Data extracted directly from a productive ERP system forms the use case and represents the challenges of handling large volumes along with the variations and noise associated with real data.

Formal Concept Analysis (FCA) has been applied in order to hidden discover knowledge from transactional data. This chapter describes the method applied alongside examples of the analysis in order to support the theory and demonstrate a practical application. The findings from the analysis are discussed and it is explicated how knowledge discovery from ERP is a useful exemplar for the Internet of Things and ERP development.

A method for discovering information and knowledge is described in this chapter based on data collected directly from an ERP system employed by a real Enterprise; all data has been anonymised. User interaction with the system has been captured and processed using Formal Concept Analysis (FCA), as introduced by Rudolph Wille and Bernhard Ganter (Ganter and Wille, 1999) as the mechanism for discovery.

The aim of the following section is to analyse data sourced directly from an ERP system using techniques that are supported by computational systems and complementary to human understanding. For this reason FCA has been applied as it embodies many of the desired properties. Wille describes FCA's roots as being in philosophical logic with a pragmatic orientation and formalisation of concepts (Wille, 1997). As discussed above the recognition of a situation, the capability of maintaining an understanding of the position and relationships through formal concepts in order to provide a context for the analysis is vital for understanding. Winograd and Flores, as discussed by Devlin (1997), developed systems to complement human communicative

skills. This is undoubtedly the direction of this chapter, it is fundamentally an alliance between humans and interactive system tools, it is not targeting a solution to Artificial Intelligence.

5.2.1 Rationale

ERP is a compelling data source for this analysis due to the wide number of applications, industries, the sheer volume of data available and finally because it is structured and consistent. Data is one of the key components within ERP systems; it ranges from highly structured to virtually unstructured. In ERP the bias is towards structured data, if a view of data across the whole Enterprise is considered the bias would be towards unstructured data.

The principle of this research is to discover if ERP user data can reveal any useful knowledge or information. Potential applications included user management, authorisations, process design, interface design and understanding general patterns of use. Cross referencing this with respect to a time period or geographical data may also reveal useful information. The data used in the examples include both content and structure that supports both of these considerations. Traditional BI solutions do not focus on user transactional data; therefore, this approach and data set presents an interesting dimension.

5.2.2 Data Preparation

The data for this analysis has been extracted directly from an organisation's productive SAP ECC 6.0 system. Therefore, it is a representative sample of real-world data; all data has been anonymised. SAP A.G. is one of the leading vendors for ERP systems and accounts for a significant proportion of the worlds transactions, this is estimated to be between 60 - 70% (Poonen, 2012) (Forbes, 2011). This is clearly a very large quantity and given the growth in data and communications a volume that is likely to

increase. ERP systems are typically rigid systems that represent the core functions of an Enterprise, although with the development of advanced technologies and architectures ERP systems are slowly increasing in flexibility at significantly lower costs than in previous times. This does not imply that ERP only represent simple systems, they are employed across global organisations with complex and unpredictable behaviours caused by internal and external factors.

A useful set of data captured in SAP ECC 6.0 are the transaction logs of its users, this forms the primary data source and use case. It has been combined in part with other data contained in the system, this will be described in detail but an example is a lookup for the description associated with a transaction. In general this is all data available within SAP ECC 6.0 that can be extracted by query. The steps taken in this analysis are consistent, the primary differentiator being the manner in which data is prepared and the graphical manipulation of the lattice in Concept Explorer. The first example includes all steps in detail, subsequent examples only show the differentiating factors.

In order for a transaction to access or change data it performs a dialogue step, this essentially requests information from the Database Management System. These requests are logged and available for analysis by the Business Transaction Analysis tool that displays kernel statistical data for user transactions or background processing (SAP, 2012*c*). The basic data structure is a time stamped record of the transaction or program executed by a user. The basic constraint is that data is only written at the end of the dialogue step, essential after successful completion of the activity, however performance related data is also include which enables the calculation of the start time for the dialogue step. The raw data is based on dialogue steps to the database therefore an individual transaction may result in a number of dialogue steps as the user retrieves data and makes updates, however, it is simple to identify the first dialogue step and filter out any secondary steps.

Details about the structures are contained in the following sections. All data used

Subject	Predicate	Object
171	USER	User 171
942	USER	User 171
(Repeat for all users)		
171	TRANSACTION	Report_1
942	TRANSACTION	Report_1
(Repeat for all user / transaction)		

Table 5.1: Example Input File

has been extracted directly from the database, there are no steps that could not be mechanised within the data preparation.

All data in this section has been prepared to a 3-column CSV format, this represents triples in the form subject-predicate-object. The data was also restricted to one week of transactional activity within a single department. This structure enables the inclusion of additional data potentially from other sources without requiring it to be contained in the ERP system itself.

5.2.3 Analysing Transactional Activity

The aim of this analysis is to discover knowledge from 'transactions by user in a time period'. This is intended to be an elementary question to test the method, it could be answered with a simple query but it is necessary to understand FCA in this context. The raw data was queried in order to produce a text file (.csv) as shown in table 5.1. Two queries have been combined to retrieve file contents from the raw data, the file contained approximately 35,000 rows with a run time of only a few seconds. The first section of the data describes the user reference with the user name, for anonymity the names have been replaced by 'User-number'. The second section of the data describes the user and transaction.

The tool set applied does not offer the capability to scale this approach indefinitely, an overview is provide in section 3.4. Tools such as RDF Databases, also known as Triple Stores, would be required to support scaling up to the volume and performance

levels required for an entire organisation. These are relational databases design with the intention of holding data in this subject-predicate-object format (W3C, 2004). Although this is potentially required to support further work and larger volumes of data it is not deemed necessary or detrimental to this research.

In this example the context is 'transactions by user in a time period', it does not include any chronological data. This text file (.csv) is read into FcaBedrock, figure 5.1, and processed to produce a context file (.cxt). Without any further processing the output files contains the data as represented in figure 5.3, the image is taken from Concept Explorer rather than the data file purely for presentation purposes. In its simplest form the task of creating a formal context file is complete.

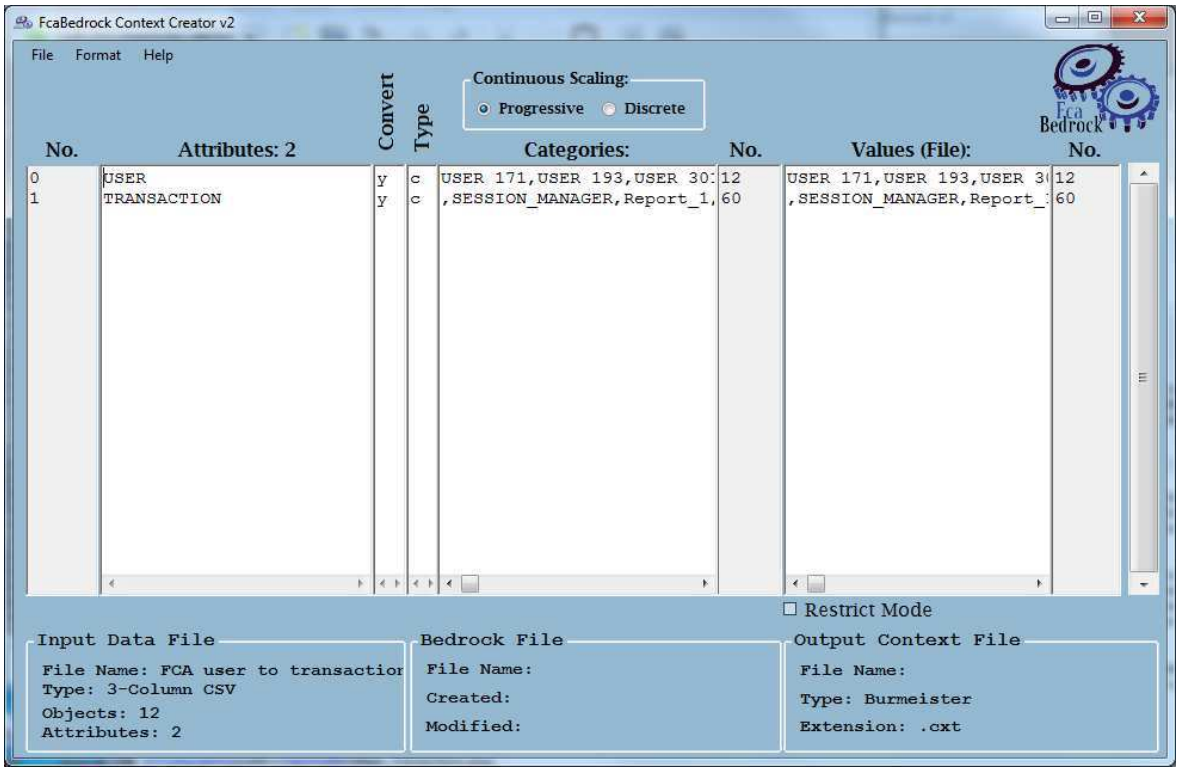


Figure 5.1: FcaBedrock

Figure 5.2 displays a lattice created with an attribute count reduced by 50% within the context editor, in this example the attributes are the transactions. The initial

lattice produced was complex and unreadable therefore simplification through reducing the number of attributes was attempted. Although this could be deemed as a visual improvement it has a significant drawback in that a proportion of attributes are simply removed, there is no ranking or measure of importance. With expert knowledge it is known that 'VA01' is a transaction to create a sales orders and therefore fundamental to the operation of the department in question, however, it has been lost from the analysis. This is where minimum support can be applied through the use of InClose2 to control the intents and extents therefore simplifying the lattice in a more controlled manner.

The second point to note is highlighted in the lattice, see figure 5.2. Reading from the bottom left 'User 954' is connected to the transactions 'Report_3', 'MD13' etc. as are other users including 'User 171'. The strand highlighted indicates that only half the users use the same transactions. Although all users all perform the same function within the same department it is clear that over the time period there was a difference in the actual transactions used. There may be other circumstantial reasons why there is a difference as the team is essentially performing the same task many times per day. To have such a number of differences is concerning and further investigation may indicates deviations from the defined process or ineffective interaction with the system.

By collating use over an observation period or by comparison with the designed process, individual deviations could be highlighted and addressed. Objects or attributes can be hidden within Concept Explorer making it easier to reveal common patterns or differences.

From this simple example a number of applications are evident including BPM, interface design, process monitoring and training to mention a few. The most obvious problem is the complicated and bordering on unusable lattice produced. Simplification of the lattice is not a trivial task, significant data or a focus can be lost from the analysis without realising it.

Figure 5.3 represents the complete context table from which figure 5.2 was produced.

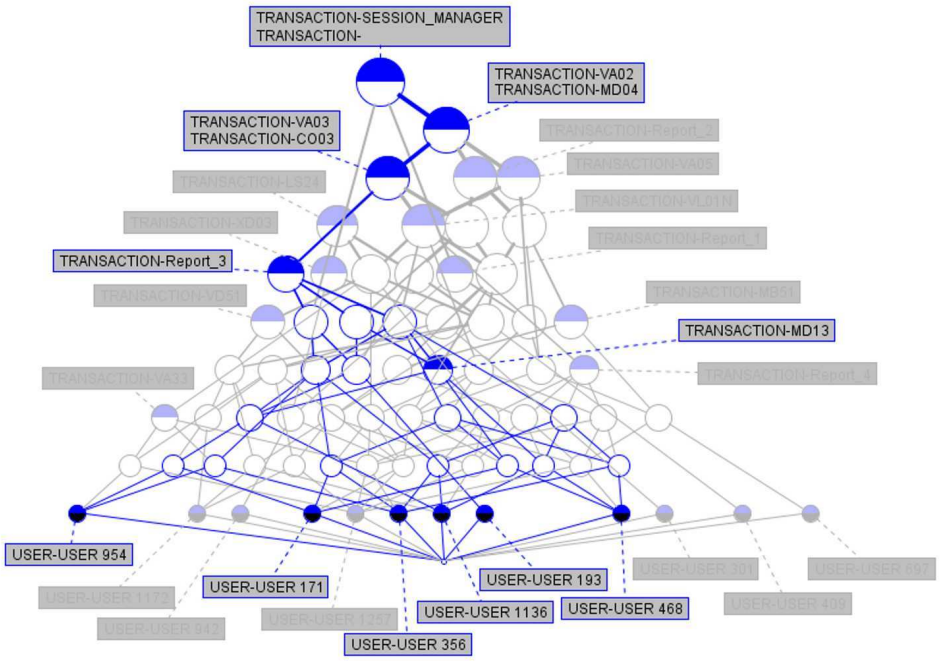


Figure 5.2: Transaction Based Lattice

The columns represent the transactions and rows the user, in simple terms, the 'X' marks an occurrence of user and transaction (the column headers have been reduced in width for clarity). The interesting aspect of this is the heavily populated groups particularly towards the middle of the table where many users are using the same transactions. One potential opportunity would be to understand why users are not using certain transactions, this may indicate a process feature or training requirement.

Figure 5.4 contains the same data as figure 5.2 manipulated into a form suitable for Excel. In this example, 'X' is replaced by '1' to support calculations, it is also shaded for clarity. The column on the far right shows 'transactions usage by user'. By example, 'User 171' is performing 93% of the transactional range within the department. The extent row along the bottom is reflected in the lattice and represents the range of

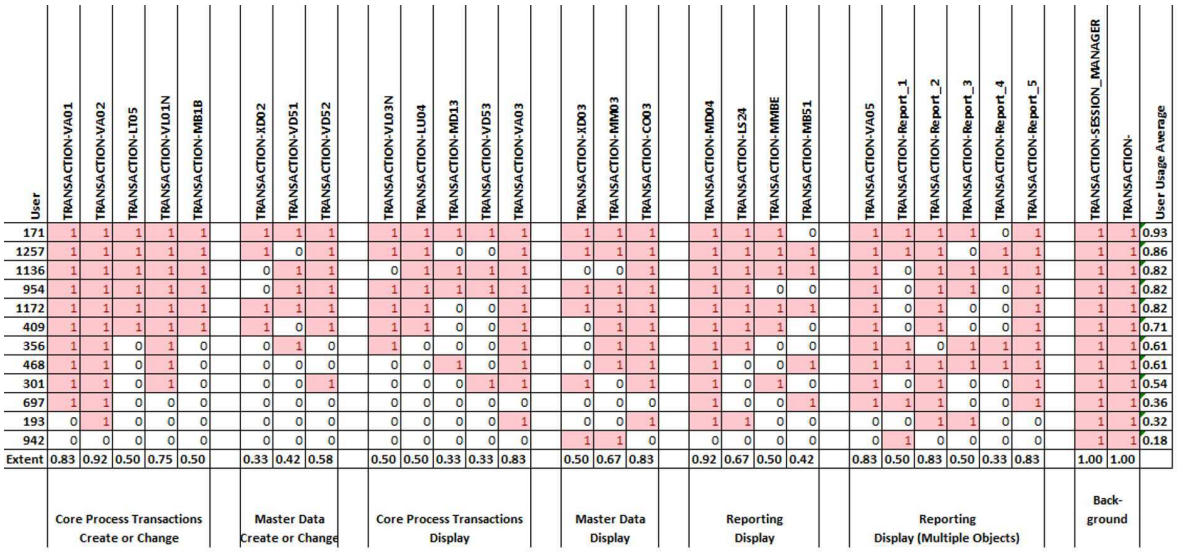
Figure 5.3: Concept Table for Transaction Based Lattice

users performing a transaction; this figure is the same as the extent value available in the lattice diagram. The core function of the department is known to be transactions 'VA01' and 'VA02' (create and change sales orders). These activities have high extents indicating virtually all users perform these transaction. A large data set could be analysed using minimum support to select only concepts with high extents. This is potentially useful and is demonstrated in subsequent examples where further attributes are introduced. Potentially this could include geographical, performance measures or job title/role attributes to mention a few.

An additional application of this tabular view is in the definition of the standard processes. Over a period all activities associated with a department or process step could be identified, capturing the actual process, supporting activities and providing feedback on the designed process. This point is explored later when transactional flow is considered.

5.2.4 Analysing Transactional Activity with Descriptions

The aim of this analysis is to discover knowledge from 'transaction descriptions by user in a period'. Figure 5.5 is based on the same data set as figure 5.2. In this example the transactions have been replaced by their high level description, for example the transaction 'VA01' has been replaced by 'Create', see table 5.2. In order to produce



User	TRANSACTION-VA01	TRANSACTION-VA02	TRANSACTION-LT05	TRANSACTION-VL01N	TRANSACTION-MB1B	TRANSACTION-XD02	TRANSACTION-VD51	TRANSACTION-VD52	TRANSACTION-VL03N	TRANSACTION-LU04	TRANSACTION-MD13	TRANSACTION-VD53	TRANSACTION-VA03	TRANSACTION-XD03	TRANSACTION-MM03	TRANSACTION-CO03	TRANSACTION-MD04	TRANSACTION-LS24	TRANSACTION-MMBE	TRANSACTION-MB51	TRANSACTION-VA05	TRANSACTION-Report_1	TRANSACTION-Report_2	TRANSACTION-Report_3	TRANSACTION-Report_4	TRANSACTION-Report_5	TRANSACTION-SESSION_MANAGER	TRANSACTION-	User Usage Average
171	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0.93
1257	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0.86
1136	1	1	1	1	1	0	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0.82
954	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	0	1	1	1	0.82
1172	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	0.82
409	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	1	1	1	1	0	1	0	1	0	0	1	1	1	0.71
356	1	1	0	1	0	0	1	0	1	0	0	0	1	0	1	1	1	1	0	0	1	1	0	1	1	1	1	1	0.61
468	1	1	0	1	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0.61
301	1	1	0	1	0	0	0	1	0	0	0	1	1	1	0	1	1	0	1	0	1	0	1	0	0	1	1	1	0.54
697	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	1	1	1	0.36
193	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	1	1	0	0	1	1	0.32
942	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0.18
Extent	0.83	0.92	0.50	0.75	0.50	0.33	0.42	0.58	0.50	0.50	0.33	0.33	0.83	0.50	0.67	0.83	0.92	0.67	0.50	0.42	0.83	0.50	0.83	0.50	0.33	0.83	1.00	1.00	
	Core Process Transactions Create or Change					Master Data Create or Change			Core Process Transactions Display					Master Data Display			Reporting Display				Reporting Display (Multiple Objects)						Back-ground		

Figure 5.4: Excel Concept Table with Calculated Values in Excel

this file the query was modified to replace the transaction with its description. This data is available within SAP ECC as each transaction is held as a record including its description, associated program and other information. For the purposes of this example the first key word has been selected but with the addition of methods such as stemming and key word search this approach could be extended.

In this configuration, the lattice has taken a considerable step towards being more usable and a distinct hierarchy is starting to become visible, this is illustrated by a highlighted strand in figure 5.5. Although this analysis is based on the same data the use of general descriptions has caused a grouping effect and simplified the lattice. This is effectively applying metaphors and cognitive models in order to understand and generalise complex processes or events. The highlighted strand indicates transactions that support the core process with sub processes or supporting activities appearing beneath them and not highlighted. An understanding of lattice construction and the ability to graphical interact and highlight strands is starting to make lattices an intuitive method of analysis.

Subject	Predicate	Object
171	USER	User 171
942	USER	User 171
(Repeat for all users)		
171	TRANSACTION	Create
942	TRANSACTION	Report
(Repeat for all user / transaction)		

Table 5.2: Example Input File with Descriptions

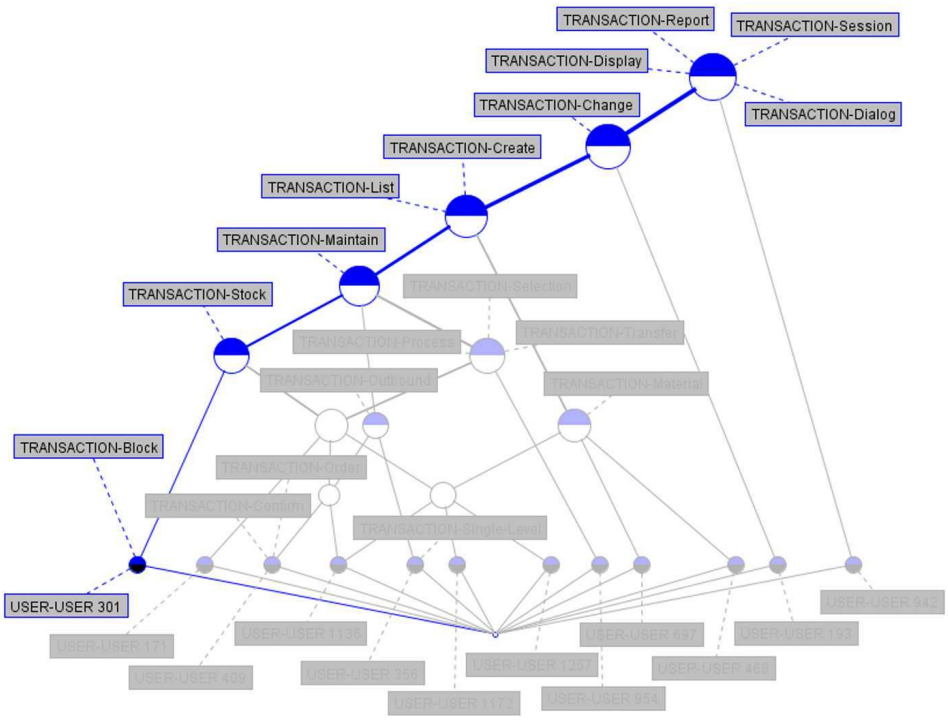


Figure 5.5: Transaction Lattice Represent by Basic Transaction Descriptions

5.2.5 Analysing Transactional Activity with Multiple Attributes

The aim of this analysis is to discover knowledge from transactional activity using 'multiple attributes descriptions by user in a period'. Figure 5.6 is based on the same data set as figure 5.2. In this example, attributes have been added rather than replaced, see the date structure in table 5.3. The USER attribute remains unchanged from the original example. The ACTION attribute has now been added to represent the description. The AREA attribute has been introduced to represent the business function. The final attribute is transaction date; TRANSACTION has been used to represent the date a transaction was used. As before, all data used has been extracted directly from the system.

This analysis has been restricted to a single user by using the 'restrict mode' in FcaBedrock, this essentially applies a filter. Because the AREA attribute has been derived from the transaction it is expected that they will align with each other. Using the same example as previously 'Create' and 'VA01' will share the same node. By visual inspection the lattice indicates that there is not a standard daily process, transaction 'VA01' is performed daily but the sub tasks vary. If it is assumed that the core function of the department is 'VA01', then other tasks could indicate inefficiencies or distractions from the goal.

This capability to input multiple objects supports two features, firstly the ability to choose the level of detail displayed and focus the analysis. Figure 5.7 has all transactions hidden apart from 'VA01'. The AREA and ACTION attributes remain for interrogation but the complexity caused by displaying all transactions is removed. The second feature is the continued maintenance of a context, a detailed focus is surrounded by a context in order to support a cognitive model.

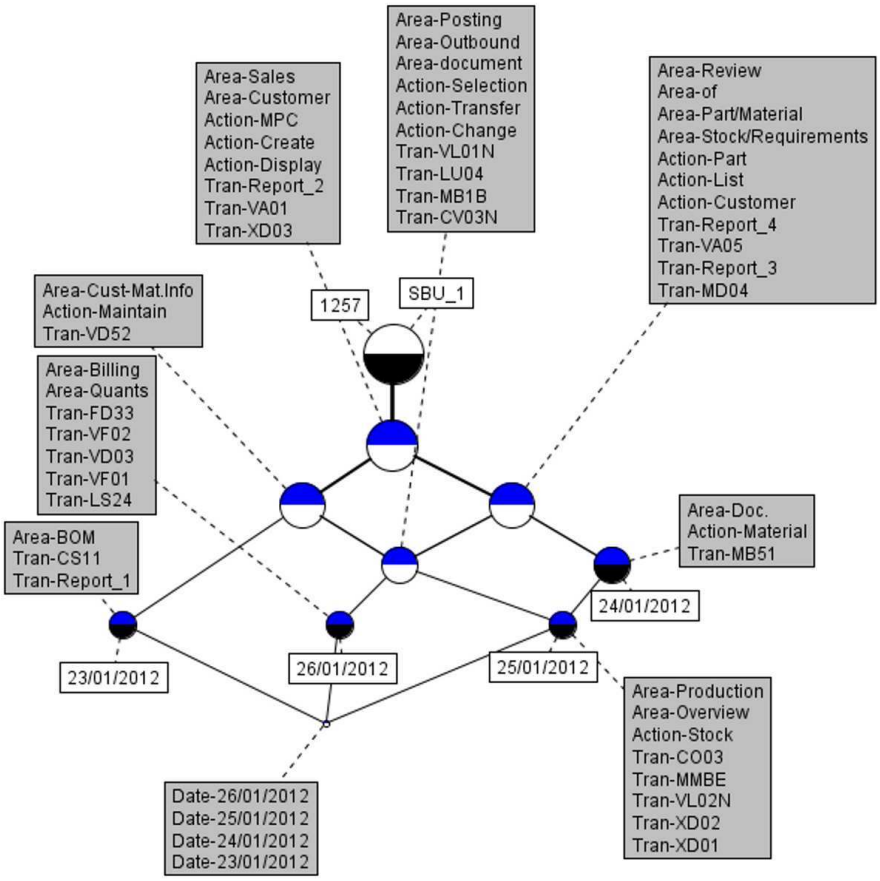


Figure 5.6: User Transactions

Subject	Predicate	Object
1257	USER	User 171
(Repeat for all users)		
VA01	ACTION	Create
(Repeat for all transactions / actions)		
VA01	AREA	Sales
(Repeat for all transactions / areas)		
Date	TRANSACTION	VA01
(Repeat for all dates / transactions)		

Table 5.3: Example Input File containing Multiple Attributes

5.2.6 Analysing Transactional Activity with Direct Comparison

The aim of this analysis is to discover knowledge from transaction activity between 'users in a period' and support a detailed level of investigation. Two users have been compared side by side in order to demonstrate this point, see figure 5.8. The preparation steps are identical to 5.6 apart from focussing on a different user.

A distinct difference can be observed when the 'create' node is highlighted, the user on the left shows that 'create' is a frequent used transaction and occurs every day. Conversely the user on the right used only display and reporting transactions for the first two days of the observation period.

Side by side comparisons of lattices has the potential to produce useful knowledge. The level of detail and granularity is very significant in the analysis. This example is highly detailed but the same method of comparison could equally be compared to groups of users or Enterprises. The ability to support large data sets, multiple attribute descriptors and graphical analysis is fundamental.

The ability to group and generalise features is fundamentally important to support a cognitive understanding for visualisation and comparison. Grouping could be used to combine nodes that share similar features, effectively promoting dissimilar features that

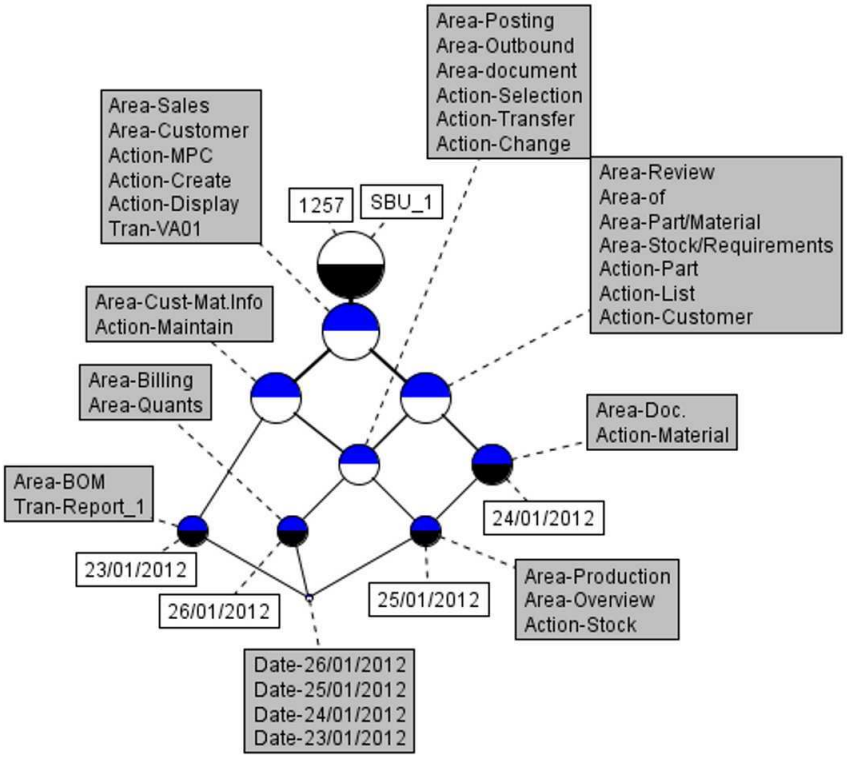


Figure 5.7: User Transactions limited to Target Transaction VA01

may be at a different level of detail. Replacing attributes with alternative descriptions or values from the system or via ontologies is theoretically possible and would represent a useful feature if added to the software.

5.2.7 Analysing Transactional Sequence

The aim of this analysis is to discover knowledge and patterns in transactional sequences. The analysis methods applied so far have focussed on the use of transactions over a time period; this section will focus on the sequence of transactions. The sequence of transactions closely represents how users actually interact with the system.

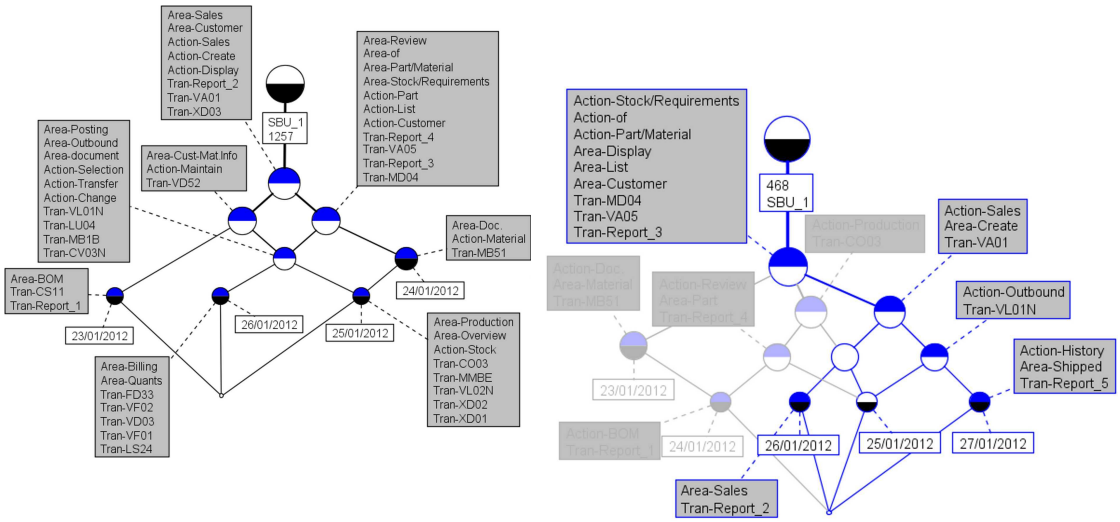


Figure 5.8: User Transactions for User 1257 (left) 468 (right)

Subject	Predicate	Object
VA01	FLOW	Report_108
Report_108	FLOW	MMBE
(Repeat for all transactions)		

Table 5.4: Example Input File containing the Sequence of Transaction

It has the potential to be used for comparing planned and actual usage and highlight patterns of usage. This analysis could be useful in both BPM and Interface design.

The format in table 5.4 contains the sequence in which transactions have been used by a single user. The object is the transaction and the attribute is the transaction that followed. The obvious point is that a transaction, particularly a report, may remain on screen and in use for a long period, therefore it may overlap other transactions. Any refresh would be captured but, just like if the report was printed, a static display is not represented.

It is intriguing to note from the highlighted section of the lattice in Figure 5.9 that one of the core functions 'VA01 - create sales order' has such a strong relationship with five reports and one maintenance transaction. This clearly indicates a need to investigate the consolidation of these report. For the purposes of this explanation 'XD03 - display customer' has been included in the report count, 'VD52' is the only

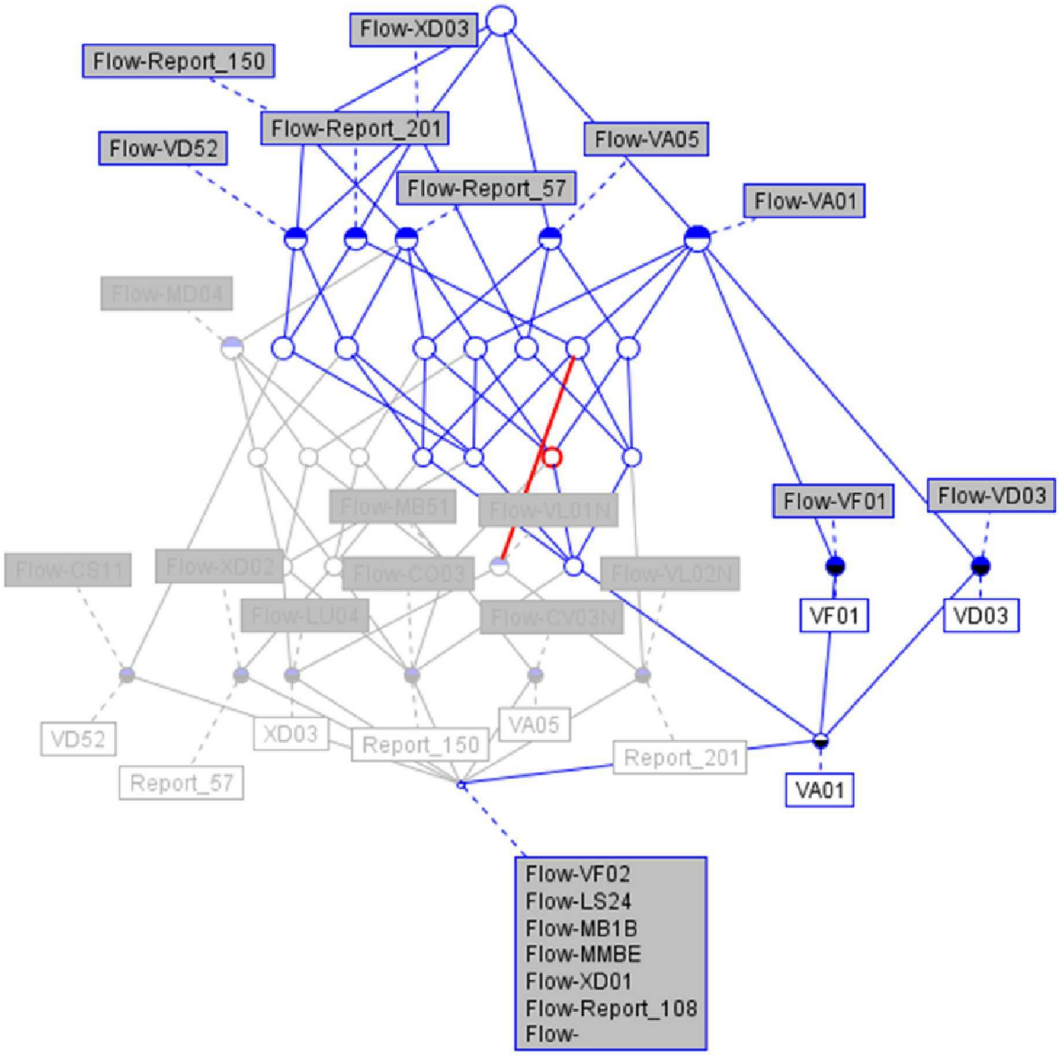


Figure 5.9: Lattice for Transaction Flow for an Individual User

transaction with a maintenance function. It should be also be noted that this only addresses the data available within the ERP system, it is likely that there are even more sources that should be considered.

From a BPM perspective it should be questioned why so much non value adding time is being incurred, an obvious question to ask is what areas are the reports reflecting? Using the ability to add multiple levels figure 5.10 has been produced. This focusses on a specific transaction 'VA01 - create sales order' and applies the AREA attributes for all other transactions. From an analysis viewpoint this provides an easier conceptual model for understanding and reduces the need for expert knowledge. This example indicates that reports are being run against a range of other departments when creating a sales order. The salient point here is that these are separate transactions being used to collect information beyond the minimum that is required to process the transaction.

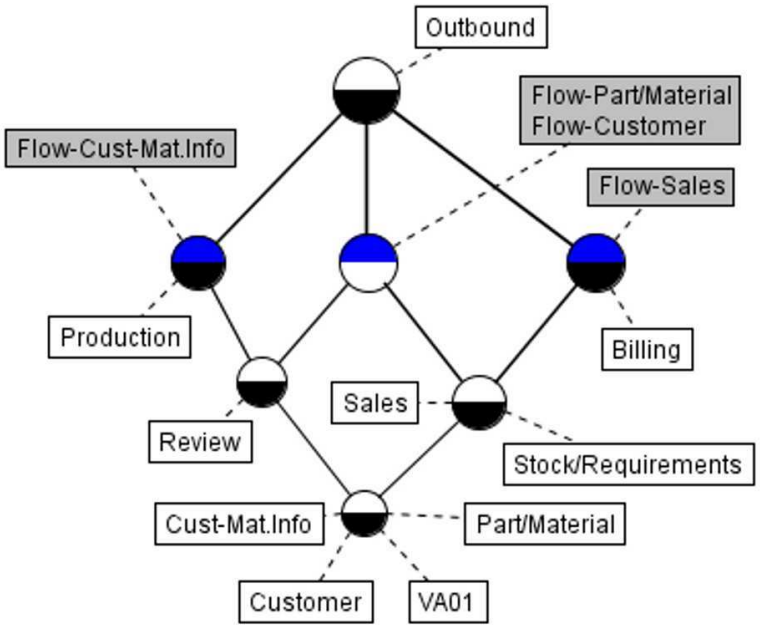


Figure 5.10: Lattice with VA01 and Related Transactional Areas

5.3 Evaluation

Many useful features have been demonstrated, more importantly the potential for including this type of analysis as part of an organisation's Business Intelligence capability has been highlighted. Two points of view have considered; the results of the analysis in the context of the data set and secondly the potential for this analysis method as a BI application. The range of techniques applied have shown how areas within an Enterprise can be identified as targets for improvement efforts. Specifically this relates to interface design, process adherence, training and common transactional patterns.

The analysis techniques applied are separated from the actual process; the unique objects processed have not been considered. Key Performance Indicators (KPIs) such as the product quality or delivery performance have not been included in the analysis at present. The analysis of transactional data has been considered in chapter 4 and could potentially be combined in future work as the analysis of actual work patterns and performance would be of interest.

It is clear that this approach complements human capabilities and is in keeping with the philosophical ideas of context and semantics. The principles and structure appear to complement human capabilities and in combination could push towards the attainment of expert levels as defined by Dreyfus and Dreyfus (1986) and discussed in section 2.5. A contextual perspective can be maintained and a detailed or targeted analysis combined within a view containing more descriptive and general information. The method demonstrated supports multiple labels for a transaction or event and has the potential to include data from multiple sources while maintaining relationships. There is still significant work required to enhance the users experience, particularly around data preparation and lattice navigation or interrogation.

From an analysis viewpoint, it would indicate that the sales department is querying information from many areas including production, inventory and billing during the order capture process. This in itself is not unusual when accepting customers orders,

using separate transactions to access this information suggest a non optimal system or a process that is out of control, certainly highlighting an area for further investigation.

Applications of this method could indicate how resources such as reports and transactions should be unified or combined to support specific job functions, there is a potential for automated mash-up reports in the context of the task being performed. Another move forward would be to associate the success of objects against the patterns associated with the specific job function. The individual manner by which a user approaches a tasks could be attributable to the outcome.

In memory techniques such as SAP HANA (SAP, 2011*a*) could enable this analysis directly from the data source without intermediary data preparation steps. The data preparation steps will not differ significantly from those describe. The concept of holding all ERP data in memory coupled with significant performance gains in database access speeds as described by Plattner and Zeier (2011) have the potential to make this a real time analysis technique.

Process and interface design could benefit from the sequence and overlapping use of transactions. For example, transactions that are used concurrently to perform a task could be candidates for process and or interface redesign. Integrated interfaces such as SAP Business Suite 7 are based on an open SOA platform and aimed at enhancing integration (Muir and Kimbell, 2010), incorporating this analysis and evolving alongside the users actual job function is a possibility.

User management could include grouping users based on actual use and comparing commonalities within hierarchies of groups. The result may be useful in defining authorisations that are tailored to the users job function and far less generic than is common in organisations today.

5.3.1 Framework

Finally the rudimentary capabilities required in a framework for discovering knowledge from ERP system data are listed below. These have been discussed in the main text but are presented in a concise list below.

Data Requirements

- Hierarchical levels of granularity in the source data
- Single or standardised data repository
- Simple data format for representing facts

Semantics and Context Requirements

- Traverse through data in multiple dimensions and maintaining focus
- Prominently retain and promote context
- Support for combined human and system discovery

Graphical Interaction Requirements

- Promote or eliminate common features between lattices
- Graphical filtering to remove or replace attributes
- Control of layer transparency
- Ability to record analysis at key points

It is suggested that ERP is positioned as the temporal reference for an organisation, a reference for the orchestration and choreography involved. The captured knowledge using this approach does not mandate a direct reference to the transactional system providing the data and models have been constructed correctly. The data sources

may be external to this system and a standard reference is necessary particular when considering sequence and usage. ERP typically represents the core of the business functions and is central to an organisations operation; this makes it a logical choice as the central reference of interactions and sequencing.

Post analysis the reference maybe redundant, particularly if the analysis if focussed on relationships, classification, sequence or usage; however, the ability to construct these consistently requires a frame of reference. With the growing trend towards ubiquitous computing the ability to align processes accurately will become even more fundamental, it may not be possible to maintain a direct link and navigate between the results and source data. It is highly likely that as data volumes continue to increase much of the source will be lost or loosely coupled data. After the completion of the analysis and storage of results, this repository may effectively become the storage and application for contextual knowledge.

The data used for the analysis could be extracted and stored in a triple store (W3C, 2004) and combined with data from other sources. Extraction and storage of user data into a triple store would also negate the need to permanently store the source log files within the ERP system. The data set analysed contained approximately 35,000 records, a fraction of the four million records for the whole enterprise over the same period. In the context of big data, the method described is highly suitable.

The knowledge represented in the lattices could be used to build a useful ontology for logical deduction and analysis. The examples constructed have demonstrated that it is possible to extract and build knowledge that represents the actual use of systems from large data sets relatively easily and with differing levels of granularity. The hierarchical capability provides this from a time frame and sequential perspective. For specific examples, as demonstrated, the use of lattices to display results has enabled meaningful graphical analysis. Lattice diagrams can add value in this environment, the limiting factor is the graphical complexity.

Data preparation at this point in time is manual, however, everything demonstrated

has only used database queries and data available from the source SAP ECC 6.0 system.

5.4 Concluding Summary

In this chapter, the use case has revealed that FCA has a practical application for the development of ERP systems and potentially incorporating with BPM. Capturing transaction activity related to the core functions of an Enterprise coupled with the capability of representing multiple levels and data sources makes the method and use case a complementary pair.

A emergent theme is the agility that this method supports, a view supported generally of Semantic Technologies (Dau, 2011). This analysis has demonstrated that it is possible to discover useful knowledge and semantics through applying FCA, lattice diagrams and graphical exploration. Important capabilities demonstrated include the interactive interrogation and discovery enabled by traversing levels and maintaining context with a focus.

A number of distinct applications have been highlighted in BPM and system interaction. These include process design and monitoring based on actual use and the individuality of users. Applying the same method to data captured from the Internet of Things could equally result in useful semantics and knowledge within topics such as Network Horizons (Liere, 2007) which is the interaction between Enterprises and also towards within Supply Chain viewpoint focussed on the actual objects handled or traded.

An important topic only lightly discussed in this chapter is a comparison between the usefulness of tables and lattices. Section 4.5 indicated a combination of techniques is potentially useful, the representations in table form, figure 5.4, and lattice form, figure 5.8, illustrates a similar point. When conversions are understood tables are useful and capable of displaying detail and summarised data. The underlying mechanics of tables are familiar to most business users through orderly formatting in predominately

straight rows and columns. Lattices can lack this consistent presentation and layout therefore requiring manipulation for visual comparisons.

Chapter 6

Discovery of Hidden Knowledge

6.1 Introduction

This chapter combines a mixture of qualitative and quantitative analysis to consider and reflect on the research question, the discovery of hidden knowledge in transactional data. This focusses on the discovery process with actual knowledge discovered as supporting evidence.

Quantitative analysis is used to highlight patterns within the students assignments from chapter 4 and answer if FCA is capable of helping in the discovery of hidden knowledge. Cost Effectiveness Analysis (CEA) has been applied to understand how and where FCA can add value. Finally reflection is used to gather and consider the requirements for successfully applying FCA as a method for knowledge discovery.

6.2 FCA as a Knowledge Discovery Approach

Understanding how FCA can aid people performing analysis and discovering knowledge examines the module outputs as a record of the students journey through learning and applying FCA. This is investigated by calculating and analysing correlations between assignment grades as discussed in chapter 4 and appendix E. Further dimensions to this question include how successful application of FCA correlates with contemporary

tools for developing an understanding of complex relationships.

Positive correlations indicate a relationship between two sets of variables (x,y) identified as triangles and a trend line, see figure 6.1. Correspondingly negative figures reveals a inverse relationship indicated by squares and a trend line. The example demonstrates positive and negative correlations of 95% between for two small data sets. High negative correlations in this context have not been observed in the results. Potentially this could have identified problems within the assignment structure; students fulfilling their potential would be expected to perform relatively consistently across all sections of the assignment resulting in a positive correlation.

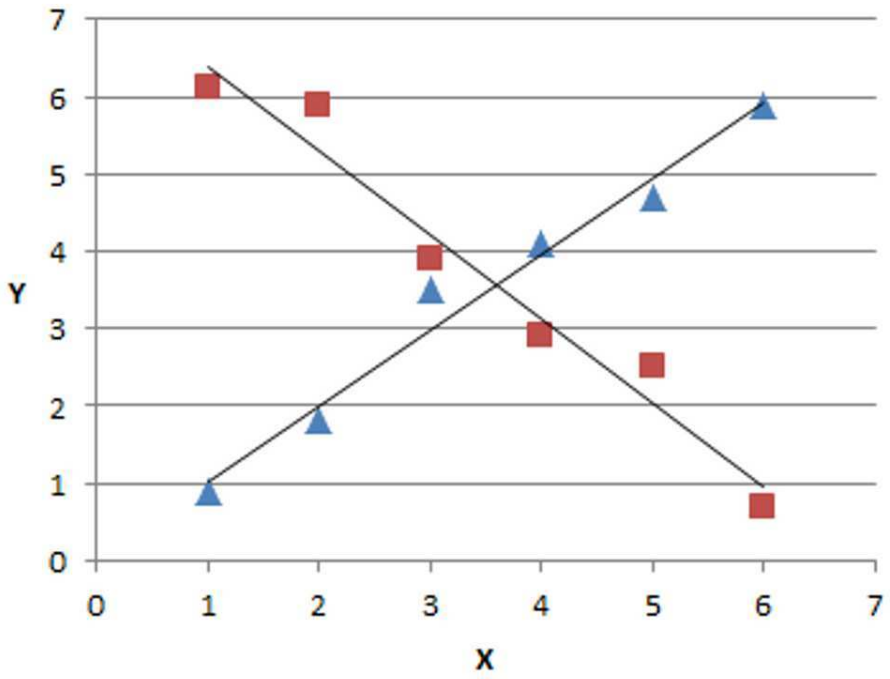


Figure 6.1: Examples of Correlations

Based on the data collected from 'Smart Applications' and 'Enterprise Systems' modules for academic years 2010-11 and 2011-12, this research carried out statistical analysis of the marks awarded. Refer to section 4.3.4 for an overview of the assignment structure.

The coefficient between each assignment section indicates the correlation between the marks achieved across all sections in a matrix format. For example, in table 6.1,

the correlation coefficient between marks achieve for the 'Excel' and 'FCA' sections of the assignment is 80%; this indicates a strong linear relationship between the marks achieved for applying 'Excel' and 'FCA'.

Correlations from Smart Applications 2010-11, table 6.1, and Enterprise Systems 2010-011, table 6.2, shared a similar structure and directly compared 'Excel' and 'FCA'. In both case studies the correlation coefficient indicated a strong positive linear relationship (80% and 98%). A high correlation between these two variables can imply one or more of the following reasons:

1. Iterative analysis cycles with alternating methods resulted in findings being shared and targeted. Form these results it cannot be assumed that the outcome is directly associated with an individual technique due to the iteration cycles.

2. Familiarity with the data increased awareness and the likelihood of successful analysis. Similarly to the first point, knowledge can be applied that influences the outcome.

3. Successful analysis correlates with student ability. Better analysis skills in general lead to better performance.

4. Successful analysis correlates with students ability to document and communicate. In contrast to the previous point, the ability or not to communicate effectively may effect the marks awarded for analysis.

These last two points intuitively have some merit, however, a much lower correlation exists between 'Excel' and 'Evaluation & Conclusion'. In 'Smart Applications 2010-11' a figure of 53%, see table 6.1, is significantly different to 94% for 'Enterprise Systems 2010-11', see table 6.2. It is noted that this Enterprise Systems assignment was group based and therefore a blend of student abilities can be expected, this is a possible reason for the higher correlations calculated between all sections.

Other than concluding that a strong 'Evaluation & Conclusion' have a high linear correlation with the total mark achieved there is little to indicate any significant advantages or disadvantages between the techniques. For this reason the structure of the assignments along with the reason discussed in appendix E changed to combine the analysis into a single section with respect to the assignment marking.

	Excel	FCA	Eval/Conc.	Pres.	Total
Introduction	-24%	-12%	-15%	-12%	-16%
Excel		80%	53%	85%	85%
FCA			83%	87%	99%
Evaluation & Conclusion				62%	62%
Presentation					89%

Table 6.1: Correlation between sections: SA 2010-11

	FCA	Eval/Conc.	Pres.	Total
Introduction	-	-	-	-
Excel	98%	94%	85%	98%
FCA		98%	83%	99%
Evaluation & Conclusion			83%	98%
Presentation				86%

Table 6.2: Correlation between sections: ES 2010-11 (Group)

Correlations from section in 'Smart Applications 2011-12', table 6.3, displayed three linear correlations that are relatively high in comparison to the remaining values. The first set with a high correlation includes the 'analysis' and 'conclusion' (61%) / 'total mark' (71%). The second set includes 'defining rules' and the 'evaluation' (61%). The apparent gaps are between the 'analysis' and 'defining rules' and also 'evaluation and conclusion'. Implicitly these should contribute in sequence towards the overall total indicating that it is possible to skip or lightly answer some middle sections of the assignment; while maintaining a level of knowledge. It is surmised that conceptual or mental models are enabling people to traverse sections while maintaining understanding and effectively achieving one of the goals.

Correlations from section in 'Enterprise Systems 2011-12', table 6.4, displayed four linear correlations that are relatively high compared to the remaining values. Firstly

	Analysis	Define Rules	Eval.	Conc.	Pres.	Total
Introduction	10%	-32%	-43%	51%	13%	-3%
Analysis		-17%	-11%	61%	20%	71%
Define Rules			61%	-24%	-11%	48%
Evaluation				-25%	0%	43%
Conclusion					47%	55%
Presentation						47%

Table 6.3: Correlation between sections: SA 2011-12

between 'analysis' and 'solution' (90%). Secondly between 'Solution' and 'Conclusion' (82%) / 'Presentation' (81%). Finally between 'Justification' and 'Conclusion' (95%). Generally the correlation is much higher across all sections which is in line with the development of the LTA cycle. It also implied that successful analysis is fundamental to overall success. This is in contrast to 'Smart Applications 2011-12' were successful analysis did not imply overall success.

	Analysis	Solution	Justification	Conc.	Pres.	Total
Introduction	66%	72%	62%	71%	79%	76%
Analysis		90%	76%	76%	68%	95%
Solution			75%	82%	81%	94%
Justification				95%	63%	89%
Conclusion					72%	91%
Presentation						77%

Table 6.4: Correlation between sections: ES 2011-12

In summary the implied points include:

1. Analysis is only one part of the knowledge discovery and communication process. A range of factors influence the outcome regardless of the technique applied including experience, bias, iterations and analytical skills.

2. Mental processes and cognitive models enable continuity across missing or weak sections of the assignments. The understanding gained during the analysis is utilised throughout the assignment even where supporting intermediary sections seem weak.

6.3 Discovered Knowledge

Two areas in particular stand out in the assignments, understanding information and representing knowledge. Devlin describes information in a curious manner with an analogy to the Cheshire Cat where the "Cheshire Cat's grin remains after the rest of the cat has vanished" (Devlin, 1997). This analogy reflects the representation of physical objects and how quickly information disappears or looses meaning when the physical representation is lost.

Wille (2001) summaries Devlin's outline about how to approach a basic science of information as "what is information and how does it flow?". With this Wille emphasised that information may be derived from data when the data is joined with collective meaning understandable in a community to which the information might be addressed. One can say that information exists in the collective mind of a social group.

Davenport and Prusak (2000) state that knowledge management must be an integral part of knowledge processes and relating knowledge to the physical world or workplace is pivotal.

In this context data from an enterprise systems formed the basis for creating figure 6.2, a simple model of information. When read by an informed user who understands the context this is useful knowledge, in a sentence format it could be read as"high profit is achieve when marketing spend is high and days cover is low, except in distribution channel twelve". This model is derived from raw data in a defined period with a low level of granularity; potentially good for high level pattern matching which would have some computational efficiencies.

Increasing the level of information presented without changing the level of granularity produces the lattices shown in figure 6.3. This contains a different level of insight and knowledge. Starting at a high level the object count versus sample size provides guidance, 131/298 (44%) in this example, the sample being a count of raw data lines extracted from the data base.

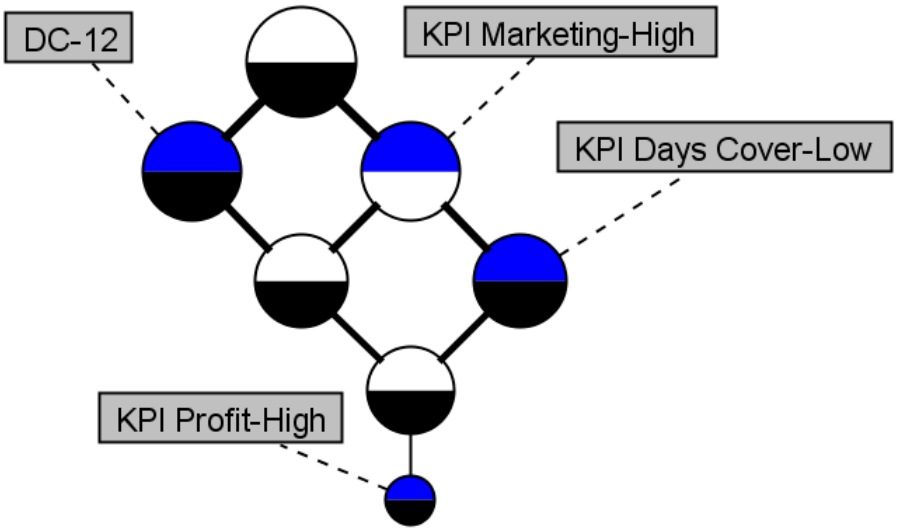


Figure 6.2: KPI Relationships

Examining the critical success measure in this example, profit, indicates that high profit is represented by 21 objects / 16% of the concepts.

Applying FCA in this manner comes close to answering discovering knowledge in transactional data. By recognising situations and presenting a sub view of a lattice and overlaying current raw data onto conceptual structures.

Refining these views further produces a simpler format, see figure 6.4. Two applications are evident, firstly in recognising a situation and bringing it to the users attention. Secondly, by applying process knowledge the situation could be manipulated to attain the goal. By example, the three KPIs for marketing, days cover and profit are directly controlled by human decisions made in ERPsim. The actually effect involves a lead time, in the instance of days cover a stock replenishment cycle takes a number of days to complete. The application of knowledge suggests that probabilistically the best option is to increase marketing spend and selling price when days cover is low.

The important point is that the known variables and situations pertaining to a probably outcome can be presented and linked with decision points. Presenting this

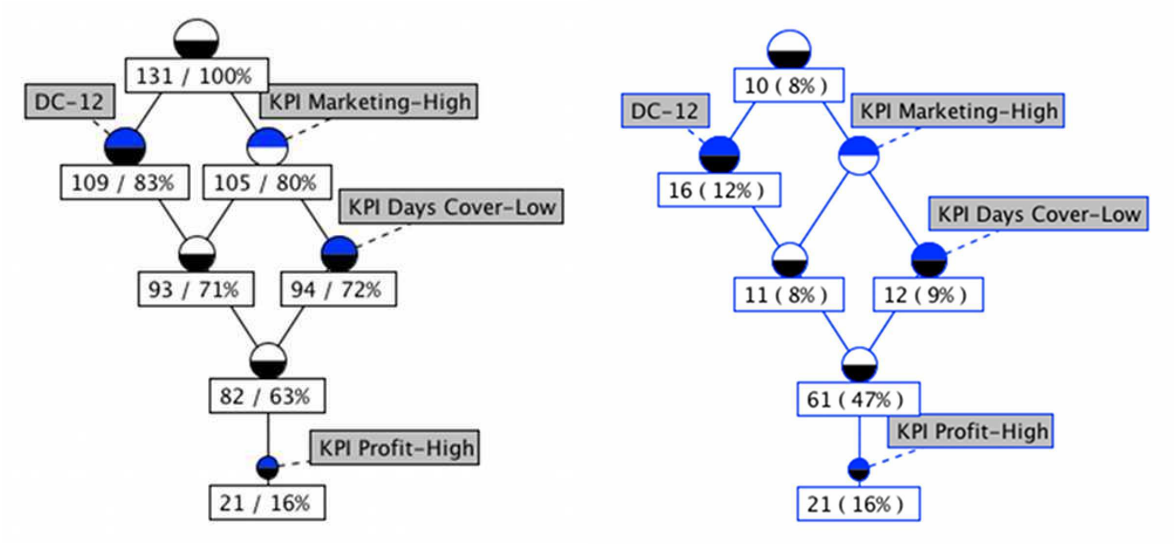


Figure 6.3: Relationships: Object Count - Extent (left), Own Objects (right)

information and consuming it as a knowledge process is expected to be dynamic therefore lattices must be pertinent and concise.

As graphical proof discrete values for 'Average Profit' and the KPI for 'High profit' are shown in figure 6.5. Both discrete values are contained in the lattice below the KPI. This was achieved by including both sets of information in the formal context. Visually this will become complicated as the number of discrete figures increases particularly due to the lattice software applied. Alternative options such as a drill down into tabular data may provide a more usable solution.

6.4 Cost Effectiveness Analysis

Considerable effort has been consumed performing analysis in the search for hidden knowledge using FCA, including comparisons with conventional techniques. To position FCA and justify the investment and effort required Cost Effectiveness Analysis is used as a vehicle to evaluate the benefits and costs involved based on the experiments and experience gained over the course of this research.

Cost Effectiveness Analysis (CEA) and Cost Benefit Analysis (CBA) are considered

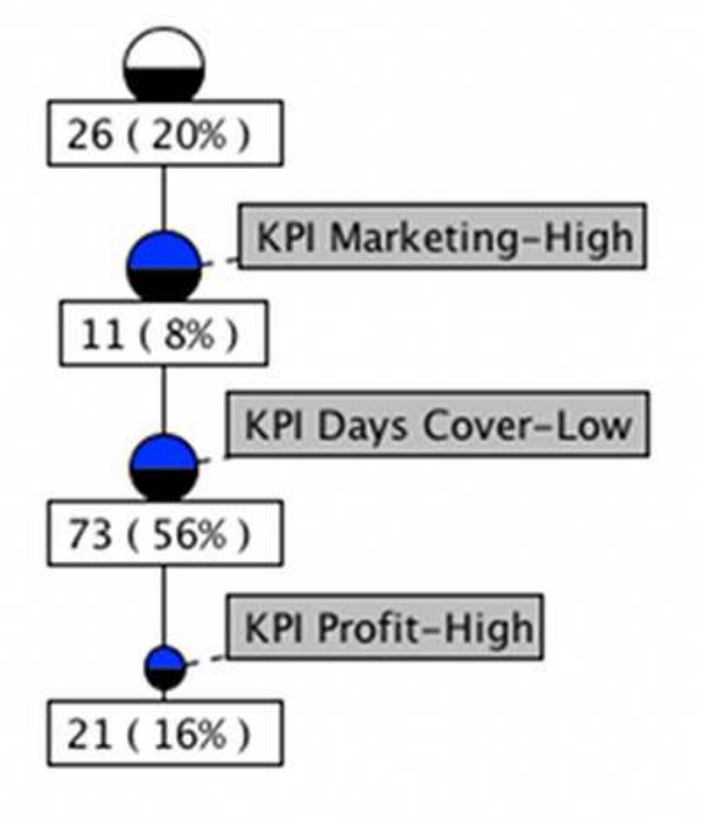


Figure 6.4: KPI Relationships: Own Objects

a learning framework for understanding benefits and costs through the process of attempting to identify, measure and value them (Cellini and Kee, 2010). CBA generally requires actual monitory figures, obtaining accurate figures is not possible given the predominantly educational data source.

CEA has advantages where the desired outcome is known at the start but intangible or difficult to monitories, however, the output requires subjective judgement (Cellini and Kee, 2010).

Following a modified series of steps (indicated by *italics*) based on Cellini and Kee (2010) CEA has been applied using relative measures. These steps acted as a guide from initial scoping through collecting benefits and costs before finally making a recommendation.

Step 1 - Set the framework for the analysis: An assumption made at this point is that analysis, regardless of the technology, will continue to be applied and subject to

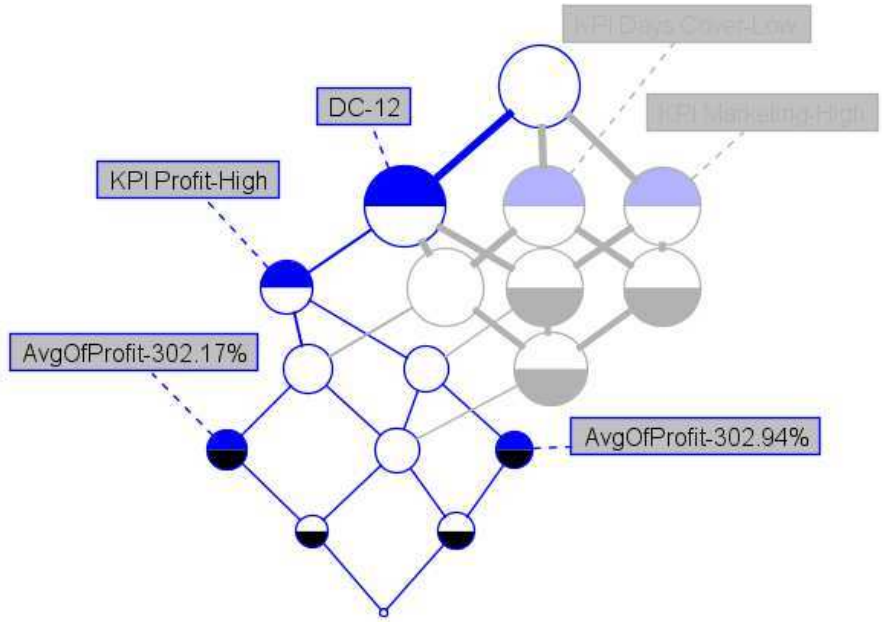


Figure 6.5: Grouping KPIs and Discrete Values

competitive forces therefore driving continual improvement and evolution.

Considering the source of data within this research, the majority from assignments, determining actual costs is deemed unreliable. Document statistics, such as editing time, contain too many unknowns to support any analysis. A key time not captured is the actual time spent performing analysis within each package. For effective cost benefit analysis (CBA) this level of data would be required in combination with a notional monitory figure.

Step 2 - Decide whose costs and benefits should be recognized: The scope is guided by the research question 'the discovery of hidden knowledge in transactional data through FCA'.

The process can be viewed as a simple progression and iteration, see figure 6.6. Inputs include the raw data resulting from ERPsim, time, resources, learning and discussions between students and staff. The final output are the submitted assignments, the end product of many iterations as part of the analysis. This transforming

of information into knowledge is the heart of this research. Davenport and Prusak (2000) describe this process as C-words including comparison, consequences, connections and conversations, all words that have been suggested or incorporated throughout the teaching. Students have been encouraged to share and discuss in addition to the actual method taught and questions posed as part of the assignments.

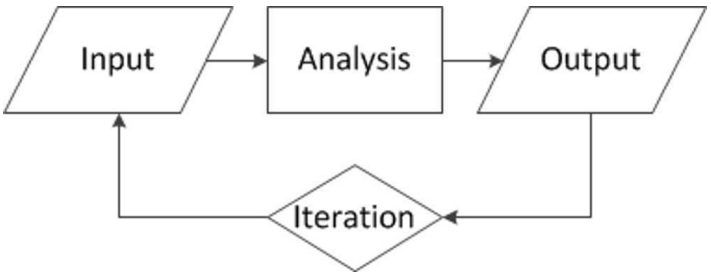


Figure 6.6: Process Iteration Cycle

Step 3 Identify and categorize costs and benefits: By considering costs and benefits relative to each other a CEA view of knowledge discovery emerges. Tangible and intangible factors are weighted from the viewpoints of the analysts and the organisation they represent over a life cycle from implementation to normal use. Tables 6.5 and 6.6 detail this breakdown.

Step 4 Considering Costs Over the Life of the Project: The aim is to discover hidden knowledge from transactional data with emphasis on the mechanisms and actual knowledge. For this reason the life is considered to be relatively short and align with the assignments rather than hypothesising about it in the context of a real organisation.

Step 5: Quantify Benefits (for CEA): A relative scale differentiates the methods applied instead of attempting to determine a monitory figure. Costs are contained in table 6.5 and benefits in table 6.6. As this is a framework for learning and understanding, differentiating between the methods with a relative indicator is considered appropriate. As an example the 'time to learn' Excel is considered direct and low cost to the analyst in contrast to FCA which is high cost, these effectively correlate to prior knowledge of the tool and the difficulty level in this case.

Cost	Measure	Excel	SAP BI	FCA
Cost to Analysts				
Direct / Real	**Time to Learn**	Low	Medium	High
	Data preparation	Medium	Low	Very High
Indirect	**Context: ERPsim**	Medium	Medium	Medium
	Education	Medium	Low	Medium
	Work experience	Medium	Low	Low
Cost to Company				
Tangible	**Time**	Medium	Low	High
	Resources	Medium	Low	Medium
	Software	Low	High	High
Intangible	**Incorrect analysis**	High	Medium	High
	Inefficient analysis	Medium	Low	High
	Opportunity costs	Medium	High	Low
	Failure	Low	Medium	High
Recurring cost				
Financial	**Iterations of analysis**	Low	Low	Medium
	Confidence in results	Medium	Low	High
	Software	Low	Medium	Medium

Table 6.5: CEA Costs

Benefits	**Measure**	**Excel**	**SAP BI**	**FCA**
Benefits to Analysts				
Tangible	**Large data sets**	Low	Medium	High
	Discovery	Medium	Medium	Very High
	Governance	Low	High	Low
	Learning curve	Medium	Low	High
	Data Inclusion	Low	Low	High
Intangible	**Graphical**	High	High	Medium
	Standardisation	Low	High	Low
	Intuitive	Medium	High	Low
Transfer	**Standardisation**	Medium	High	Low
Benefits in General				
Financial - Tangible	**Time savings**	Low	High	High
	Automation	Low	High	Low
	Error limitiation	Low	High	Medium
	Reuse	Medium	Medium	Medium
Financial - Intangible	**Governance**	Low	High	Low
	Flexibility	Medium	Low	Medium
	Agility	High	Medium	Medium
	Improved performance	Medium	High	Medium
Social	**Learning**	Medium	High	Medium
	Sharing (models)	Low	Low	High

Table 6.6: CEA Benefits

The charts in figures 6.7 and 6.8 display the CEA costs (table 6.5) and benefits (table 6.6) respectively. The bars are stacked with the lowest cost or benefit at the base.

Costs attributable to Excel and SAP BI are predominantly low or medium and are in contrast to the high and very high cost of FCA. The significant areas that create this picture include the availability of the tools, hence pre-existing skills and level of software maturity. For FCA to succeed it must address the topics of data preparation, the user experience and education; developing into a more mature and integrated product.

Benefits attributable to Excel and SAP BI are more varied, both have a range, however, the user experience and guided analysis of SAP BI provides an easier but perhaps more constrained approach. For FCA to succeed it must add value to be justifiable as an analysis tool. These are predominantly in the discovery and data management areas. Specifically this relates to processing large and varied data sets in a flexible manner.

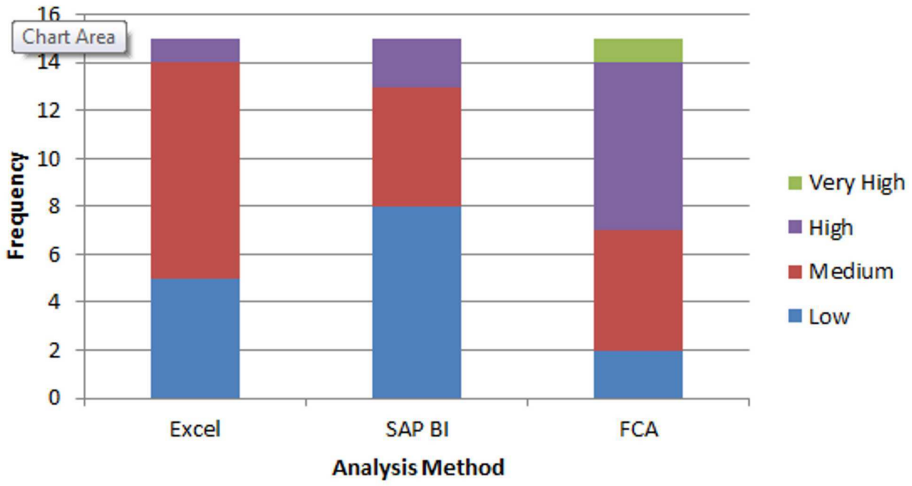


Figure 6.7: CEA Costs

Step 6: Perform Sensitivity Analysis: The perceived or experienced benefits are from a small viewpoint, that of an analyst in a company. There are other viewpoints ranging from much simpler needs to highly complex tasks.

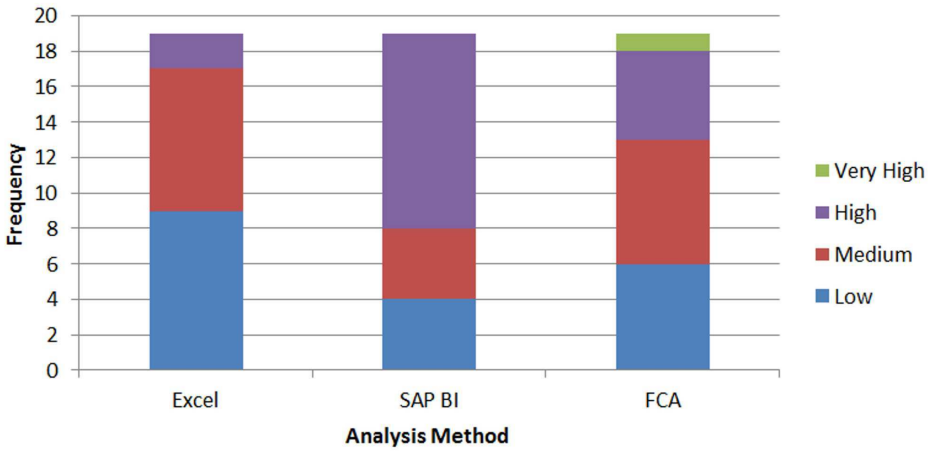


Figure 6.8: CEA Benefits

The more flexible a tool set becomes the harder it is to govern effectively, an example of this is Excel. The factors that make Excel the first choice for many applications are also its limitations. The ability to directly change data, local storage, lack of governance and single user control make it both powerful, flexible and conversely inefficient, inaccurate and subject to errors.

Development effort and costs associated with maturing FCA based tools are perhaps unfairly represented as all the tools considered have travelled through a development cycle that could be offset by mass market economics. Examples of this already happening between Excel and SAP BI in commercially available products such as SAP Lumira. Previously Visual Intelligence, this supports self service data combination and manipulation in a visual form (SAP, 2013). It is anticipated that FCA will follow a similar route providing the benefits discussed can be achieved.

Step 7: Make a Recommendation: FCA can be applied to a range of scenarios and provide useful outputs, a proportion overlaps with the capabilities of more conventional and mature applications. Within this overlaps integrating into existing applications as another means of analysis would be more advisable than a standalone product.

The benefits are clearest when integrated with process control and decision making. This includes identifying patterns and relationships within complex data and

subsequently presenting useful and highly specific information. Interaction alongside physical or representations of physical events is advantageous.

6.5 Requirements for Successful FCA

Any form of analysis involves at least some level of subjectivity and as such requires understanding, procedures and standards. A common financial transactional process such as account payable can be used to demonstrate this problem. From one perspective this is is a very mechanical processes, relatively simple and predominantly a three way checking process to ensure what you have physically requested, physically received and requested to pay in return all balance before making the actual payment. The classification and allocation within this process is the subjective issue, reasonable definitions exist but for complex products, individual understanding and a host of other factors can introduce variations. This type of scenario is replicated many times within an organisation and only key details of the process are actually captured and stored.

Utilising data within a database such as the high level descriptions in section 5.2.4 is simple to achieve and provides a useful context. Linking semantic or meta data in a similar manner is a realistic option and suitable for inclusion in the same data structure and therefore database. The challenge is to include the capability for traverse through such dimensions within a single view rather than returning to data preparation as was required for the research.

Addition or removal of data easily from the analysis view is required to support enquiry and discovery. This is intend to reflect data from many different sources or systems utilising technologies such as SOA. It is expected that this will require a number of stages where existing links are explored such as shared document numbers, references, standards, time dimensions and manual or automated linking of the remaining dimensions.

Two ideas should be recalled about databases; they are only a representations of

reality and incomplete. The ability to add data for analysis purposes may also come from manual input or from ontologies for the automated application of predefined classification. Practically this was demonstrated in the assignment as KPI definitions in appendix E. Multiple definitions could equally exist therefore tailoring the viewpoint for a specific purpose, for example a pessimistic or optimistic outlook.

Organisational processes are complex, integrated, operate at different speeds and functional at different levels of aggregation. To further complicate this they involve many different people and systems with differing objectives, capabilities and success criteria. Discovering patterns and relationships interactively is one key aspect, suggestions as a result of pattern matching, applying previous or learned routines would be a major aid to knowledge discovery. Effectively sharing prior knowledge driven from the data or applied analysis.

Cross referencing the KDD requirements first mentioned in section 2.5.2, approximately half are directly relevant to the requirements above. Requirements such as monitoring changes over time and transparency are limited by the tool set capabilities and time involved preparing data. It is reasonable to say that the approach applied is starting to meet the requirements for human-centered KDD using FCA.

6.6 Concluding Summary

This chapter combines a qualitative and quantitative analysis to understand the application of FCA for the discovery of hidden knowledge in transaction data.

Analysis is only one part of the knowledge discovery and communication process. Together with mental processes and cognitive models formed during the analysis, continuity across missing or weak sections of the analysis is possible. This relies on the presentation and consumption or information as part of the knowledge process therefore lattices must be dynamic, pertinent and concise.

FCA can be applied to a range of scenarios and provide useful outputs, a proportion

overlaps with the capabilities of more conventional and mature applications. Within this overlaps integrating into existing applications as another means of analysis would be more advisable than a standalone product.

The benefits are most apparent when integrated with process control through learning and suggestions in real-time for active decision making. This needs to be supported through presentation of useful and highly specific information alongside physical or the representations of physical events for context. This relies on the easy addition or removal of data to or from the analysis data set and graphical representation.

Linking descriptions, semantic or meta data is a realistic option and suitable for inclusion in the same data structure and therefore database. The ability to traverse through data definition levels within a single view is required.

Chapter 7

Conclusions and Further Work

The aim of this research is to discover if hitherto hidden knowledge exists in transaction data and how it can be exposed through the application of Formal Concept Analysis.

This chapter concludes to what extent the aim and research objectives have been addressed. Applications for the research are discussed along with the effectiveness of the research approach. Contributions to the research are described along with identifying further areas of research in this field.

Conclusions support the case for deploying FCA as a discovery technique. The application of FCA has been demonstrated through practical application; the inherent interaction drove questioning and knowledge creation providing insight into large datasets.

7.1 Retracing the Events

Chapter 1 introduces the motivation for the research and establishes the research aim. The research approach is outlined, using enterprise systems as an exemplar for transactional data followed by an overview of the thesis. The action research methodology applied links people with real-world events and is supported by a case study approach. Qualitative analysis and desktop research further developed the discussion and conclusions. Ethics are considered in the context of leveraging a learning environment for

data collection and the application of FCA.

Chapter 2 considered the intentions and visions within enterprise systems. enterprise resource planning and business intelligence have been considered as exemplars for transactional data and analysis as they are subcomponents of enterprise systems, widely applied in organisations and provide an accessible source of data. Enterprises System are evolving in complexity, data volumes are growing and correspondingly the challenge of deriving and applying knowledge is increasing. There is a need for discovering knowledge though an analysis method capable of discovering relationships that enhances human capabilities whilst being congruent with system-based computation.

Chapter 3 provides an introduction to the theoretical foundations of FCA used in this research for discovering knowledge in transaction data, the core focus being on the application of FCA.

FCA provides a mathematical theory based on concepts; logical relationships that can be represented and understood by humans, essentially this can be considered as information and knowledge. The capability to analyse large data sets and discover relationships in an interactive manner provides a useful mechanism that can be applied to transaction data. The steps involved in applying FCA are described, starting from source data through to tabular and graphical lattice representation.

Chapter 4 provides an overview of the learning environment that supports the application of FCA in a situation where observation and evaluation can take place. It reflects both good pedagogy and industrial practice through the use of ERPsim. This large scale, real-world business simulation software is based on the SAP ECC, an enterprise system by global business software vendor SAP A.G.

Drawing upon empirical analysis of assignment material over iterations of the teaching cycle, a range of qualitative analysis methods utilised NVivo to manage data and

generate ideas. Querying, modelling and reporting is described including the theories and conclusions developed. Results include the knowledge discovered from transactional data and an assessment of FCA's ability to explorer complex systems.

Chapter 5 explores how the application of FCA as a discovery mechanism to user transaction logs, applying and developing the methods applied previously. User transaction logs in enterprise systems are frequently overlooked as a source of data even though they offer a rich but complex source of data. The data set was captured from a real system carrying out its normal operations and resulted in an insight into the actual patterns of use and divergence from stated top-down processes.

Chapter 6 combines a mixture of qualitative and quantitative analysis to consider and reflect on the research question, the discovery of hidden knowledge in transactional data. This focusses on the discovery process with actual knowledge discovered as supporting evidence. Quantitative analysis is used to highlight patterns and answer if FCA is capable of helping in the discovery of hidden knowledge. Cost Effectiveness Analysis (CEA) has been applied to understand how and where FCA can add value. Finally reflection is used to gather and consider the requirements for successfully applying FCA as a method for knowledge discovery.

7.2 Lessons Leant from Action Research

Action research (AR) supported the three main purposes of research as stated in section 1.4, in summary these are creating new knowledge, testing validity and generating new theory.

AR worked well within the practical setting of the simulation environment supporting four iterations of the cycle, incorporating documented outputs and also the

contribution made by individual research. AR addressed the challenge of researching in this complex domain.

The application of AR is a time consuming activity both for collating data and performing analysis. The evaluation and reflection time required to derive an understanding of the students personal experiences and actual results suited the small number of case studies generated.

The average results analysis, figure 1.2, demonstrate how the AR cycle has improved the environment for generating experimental data in a controlled manner, targeting both pedagogic objectives and those of this research. AR successfully focussed on people and linking ideas with action. Disadvantages include the lack of support for any quantitative predictions such as the outcome of the next cycle. As concluded the third iteration was the most successful with a weaker result for the forth and final cycle, this was not the desired outcome.

It was not possible to deploy all possible alternatives within the AR cycle. An interesting experiment would be for groups to compete with only one analysis technique and subsequently compare results. Unfortunately this would have negatively impacted the pedagogic outcome judging from the lower than average FCA marking results, refer to the negative values in figure 1.2 for FCA. Further extensions of this AR cycle with experiments without the pedagogic constraints are a possibility.

It is expected that the findings will have parallels with applications in the real world. It is also acknowledged that the simulation is constrained and the knowledge produced is not a proven generalisation. The methodology suits situations where the AR cycle has time to operate and benefits from having a framework for deploying additional techniques. This included cross case comparisons as demonstrated in this research. Deploying this approach in a real organisation with real time needs and constraints would be challenging unless a level of automation could be achieved for collating and comparing results.

The approach itself is bottom-up, starting with data and using emergent theory

linking into action research. This risks losing focus on the objectives as directions are affected by factors including personal interests, peer guidance or differing levels of understanding. These have been considered in section 4.3.2 during the design phase but should be revisited within every cycle of AR. Maintaining alignment with the aim and strategy is a fundamental requirement.

7.3 Contribution to Knowledge

The primary contributions of this research are as follows. These are aligned with the research objectives as stated in section 1.3.

1: To provide a focus through FCA applied to transactional data allowing an analyst to discover hidden knowledge within enterprise systems.

Applying FCA in a multi dimensional simulation based in a learning environment demonstrated that knowledge discovery from transaction data is possible although further development will be required to make it practical proposal. Possible application include the development of ERP systems, aligning user training and BPM.

The importance of transactional systems, transaction data and analysis is highlighted and the vital role enterprise systems perform for modern organisations. There is a need to connect the user's human-orientated approach to problem solving with the formal structures that computer applications need to bring their productivity to bear. The emergence of service-orientated architecture and developments in business intelligence is making data itself significant. Bottom-up approaches that complement agile systems are capable of capturing human behaviour and offer novel insights into complex data and processes.

2: To provide an approach for teaching FCA and elicit how FCA could be integrated into BI.

The LTA approach described has developed a case for integrated tools that support knowledge discovery from transactional data to complement conventional and current

industry based solutions. This mechanisms is useful for identifying areas for analysis where conventional techniques lack a discovery capability but are better suited for a particular purpose.

The challenges faced through teaching and also those experienced by the students has clear parallels with the implementation and adoption of such tools in the workplace. Comparing and contrasting the techniques that have proven to be successful in the classroom to the business world would be an interesting research topic, as would addressing the problem of managing incomplete information and models. From this education experiences we can envisage that FCA has an important role to play.

3: To provide an improved application of discovery techniques in transactional data, focussing on FCA and evaluated against alternative analysis techniques.

Applying FCA has been challenging with outcomes including more reflection and observations about the method of analysis than with contemporary tools. Findings outside of the anticipated or model answers to the assignments were also more prevalent. This is partially due to the less prescribed nature of the FCA analysis methods and students not having preconceived ideas about how FCA should be performed. This is in contrast to Excel based analysis where the typically user had prior experience and education.

FCA has the potential to enhance learning and knowledge through the discovery of relationships, rules and unknowns from complex systems data in an effective and efficient manner. There is potential for FCA to be an integral part of future BI solutions; providing a link between complex events or relationships and conventional analysis techniques or human functions. The inherent interaction drives questioning, discovery and knowledge creation.

4: To enable knowledge sharing and reuse in order to deepen the understanding of transactional data and processes within enterprise systems.

Analysis is only one part of the knowledge discovery and communication process.

Together with mental processes and cognitive models formed during the analysis, continuity across missing or weak sections of the analysis is possible. The success of these tools will be based on far more than their discrete technical capabilities, a holistic view is necessary.

FCA can be applied to a range of scenarios and provide useful results, a proportion overlaps with the capabilities of more conventional and mature applications. Within this overlaps integrating into existing applications as another means of analysis would be more advisable than a standalone product.

5: To provide a understanding of knowledge derivable from transactional data and support a paradigm shift for system design.

The user logs data set has revealed that FCA has a practical application for the development of enterprise systems. Capturing transaction activity related to the core functions of an Enterprise coupled with the capabilities demonstrated in chapter 5 including interactive interrogation and discovery enabled by traversing levels and maintain context. This use case and FCA represent a realistic opportunity for real productivity improvements in an organisation. The method itself has the potential for applications where data sources are agile and varied including Business Process Modelling (BPM), process control, user training and decision making. Knowledge discovered includes factors effecting process performance and insight into the users actual interaction with the system and behaviour.

The application of FCA based on transactional data is feasible and suitable for multiple purposes. The area with the most potential for cost effect application is in domains where data and processes are complex and variable. Presenting graphical representations for consumption and interaction in a pertinent and concise manner alongside representations of physical events highlights an opportunity for adoption. This relies on the easy addition or removal of data to or from the analysis data set. It is reasonable to say that the FCA approach applied is starting to meet the requirements for human-centered knowledge discovery in database as complied by Hereth et al.

(2003), see section 2.5.2.

The paradigm shift for system design is that bottom-up analysis based on transactional data has the potential to function with agile systems and not constrain the architecture, form and function.

7.4 Conclusion

The aim of this research was to establish if it is possible to discover hidden knowledge in transaction data through formal concept analysis.

The research in this thesis leads to the conclusion that the above statement is supported. FCA can aid the discovery of hidden knowledge from transactional data. With guidance and graphical interaction, information and understanding can be explored and refined to be both useful and context sensitive. Knowledge discovered includes factors effecting process performance and insight into the users actual interaction with the system and behaviour.

This application of FCA provides a means of identifying patterns and relationships from complex data by informed and inquiring users with minimal FCA experience. FCA visualised through tables and concept lattices can contribute towards systems and humans working together more effectively. A practical real-time solution will require significant further development, however, the application of FCA could deliver significant time savings and benefits by supporting a link from complex events and relationships in transactional data to humans and contemporary analysis techniques.

7.5 Limitations and Further Work

This research has led to a better understanding of the domain and resulted in the identification of topics for further work, these are described below.

A solution to reduce the amount of data preparation is required to support the

real-time applications of FCA. Analysing more complex systems will require an understanding of the data content utilising additional techniques such as ontologies and semantic technologies. Navigation, analysis, and maintaining a focus on specific objects or attributes is challenging and software development is required to support this level of graphical interaction.

More advanced forms of analysis should be directly supported, this includes utilising qualitative data, better visualisations such as lattice on lattice comparisons, concept clustering, and alternative scaling methods. Some of these approaches have been modelled within this research through data preparation and multiple applications running concurrently. An integrated solution would be much more efficient.

Integrating the techniques described with process based transactional data may produce even more insight into the knowledge discoverable from transactional data. Differences between users (Chapter 5) could be directly related to operational performance (Chapter 4), highlighting useful knowledge and semantics at various stages of the process. This suggestion to collect the relevant data has been incorporated into the latest version of ERPsim 2013-14 as stated in the latest release notes (HEC Montreal, 2013). Albeit too late for this research it should be explored in further work.

Utilising a solution that integrates directly with the data instead of the relatively constrained data set supported by the existing extract is definitely a requirement. Plug-ins that directly connect to SAP systems already exist, this is how the data extraction tool for copying ERPsim data to MS Access functions. Developing services for integrating an FCA applications with data sources for real-time, or certainly close to real-time analysis would provide an interesting research topic. It would be interesting to understand the impact and capabilities of FCA when applied real time within the simulation. This also has the potential to start scaling towards actual production systems and applications.

A range of reduction techniques for simplifying lattice diagrams exist beyond the minimum support technique applied in this research. CUBIST (2010) is a project

that aims to combine the essential features of Semantic Technologies and Business Intelligence. Features also include novel ways of applying visual analytics. Applying tools developed as part of this project would represent an interesting topics for further work.

Areas where the application of FCA may offer benefits include SOA environments, BPM, IoT, and topics such as training through developing a better understanding of actual system use. Ideas such as re-designing or re-orchestrating processes as scenarios unfold with inputs from sensors and loosely coupled systems represent longer term visions. Clearly these are large topics but ultimately they could form subjects for further work given developments in the tool set as discussed above.

7.5.1 Word Count

I confirm the word count (excluding appendices) is 40,000.

References

Age, I. (2013), 'Dealing with the Empowered User'.

URL: *http://www.information-age.com/channels/management-and-skills/features/2143933/dealing-with-the-empowered-user.thtml [Accessed 2/2/13]*

Aladwani, A. M. (2001), 'Change management strategies for successful erp implementation', *Business Process Management Journal* **7**, 266–275.

Andrews, S. (2011*a*), Aligning the Teaching of FCA with Existing Module Learning Outcomes, *in* 'ICCS 2011. LNAI 6828', Spring-Verlag Berlin Heidelberg., pp. 394–401.

Andrews, S. (2011*b*), *In-Close2, a High Performance Formal Concept Miner*, Vol. Conceptual Structures for Discovering Knowledge, Springer Berlin / Heidelberg, pp. 50–62.

URL: *http://sourceforge.net/projects/inclose/*

Andrews, S. and McLeod, K. (2011), 'Gene co-expression in mouse embryo tissues', *CEUR Workshop Proceedings* **753**, 1–10.

Andrews, S. and Orphanides, C. (2010*a*), 'Analysis of large data sets using formal concept lattices', *CEUR Workshop Proceedings: Proceedings of the 7th International Conference on Concept Lattices and Their Applications* .

Andrews, S. and Orphanides, C. (2010*b*), FcaBedrock, a Formal Context Creator, *in*

F. S. Croitoru, M. and D. Lukose, eds, '18th International Conference on Conceptual Structures (ICCS).', Springer, pp. 181–184.

Andrews, S., Orphanides, C. and Polovina, S. (2011), Visualising computational intelligence through converting data into formal concepts, *in* 'To appear in: Next Generation Data Technologies for Collective Computational Intelligence, Bessis, N., Xhafa, S. (eds.), Studies in Computational Intelligence book series, Springer.', Springer.

Aronson, E. (2012), 'Jigsaw classroom: Overview of the technique'.
URL: *http://www.jigsaw.org/overview.htm [Accessed 18/8/13]*

Asuncion, A. and Newman, D. (2007), 'Uci machine learning repository'.
URL: *ttp://www.ics.uci.edu/∼mlearn/MLRepository.html*

Atkin, A. (2005), 'C.S. Peirces Pragmatism'.
URL: *http://www.iep.utm.edu/peircepr/*

Avison, D. E. and Fitzgerald, G. (2003), *Information Systems Development: Methodologies, Techniques and Tools*, McGraw-Hill Education.

Bazeley, P. (2007), *Qualitative Data Analysis with NVivo*, Sage Publications Ltd.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), 'The semantic web', *Scientific American* **May**, 28–37.

Biggs, J. and Tang, C. (2011), *Teaching for Quality Learning at University*, fourth edn, McGraw-Hill.

Brachman, R. J. and Anand, T. (1994), 'The process of knowledge discovery in databases: A first sketch', *AAAI-94 Workshop on Knowledge Discovery in Databases* pp. 1–11.

Brachman, R. J., Selfridge, P. G., Terveen, L. G., Altman, B., Borgida, A., Halper, F., Kirk, T., Lazar, A., McGuinness, D. L. and Resnick, L. A. (1993), 'Integrated

support for data archaeology', *International Journal of Intelligent and Cooperative Information Systems* **2**, 159–185.

Brandt, W. (2010), 'Deutsche bank european tmt conference 2010'. Accessed: 11-01-13.
URL: *http://www.sap.com/corporate-de/investors/pdf/2010-09-08_WB_DB_London.pdf*

Burbank, D. and Hoberman, S. (2011), *Data Modeling Made Simple*, Technics Publications, LLC.

Burch, R. (2010), *Charles Sanders Peirce: The Stanford Encyclopedia of Philosophy (Fall 2010 Edition).*
URL: *http://plato.stanford.edu/entries/peirce/#dia [Accessed 28/02/12]*

Cellini, S. R. and Kee, J. E. (2010), Cost-effectiveness and cost-benefit analysis, *in* J. S. Wholey, H. P. Hatry and K. E. Newcomer, eds, 'Handbook of Practical Program Evaluation', 3 edn, Jossey-Bass, chapter 25.

Cios, K. J., Pedrycz, W., Swiniarski, R. W. and Kurgan, L. A. (2007), *Data Mining A Knowledge Discovery Approach*, Springer Science+Business Media, LLC.

Clark, T. (2011), 'Sap's bill mcdermott: It doesn't take two years to create a good strategy'.
URL: *http://www.forbes.com/sites/sap/2011/10/07/saps-bill-mcdermott-it-doesnt-take-two-years-to-create-a-good-strategy/ [Accessed: 11-01-13]*

Codd, E. F., Codd, S. B. and Salley, C. T. (1993), 'Providing olap to user-analysts: An it mandate', *ComputerWorld* **32**, 3–5.

ConExp Project (2006), 'The concept explorer'.
URL: *http://conexp.sourceforge.net/users/documentation/index.html*

CSC (2012), 'The rapid growth of global data'.
URL: *http://www.csc.com/insights/flxwd/78931-big_data_growthjust_beginning_to_explode [Accessed 01/12/12]*

CUBIST (2010), 'CUBIST'.
URL: *http://www.cubist-project.eu/ [Accessed 11/12/12]*

Dau, F. (2011), Semantic technologies for enterprises, *in* S. Andrews, S. Polovina, R. Hill and B. Akhgar, eds, 'Conceptual Structures for Discovering Knowledge', Vol. 6828 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 1–18.
URL: *http://dx.doi.org/10.1007/978-3-642-22688-5_1*

Davenport, T. H. and Prusak, L. (2000), *Working Knowledge: How Organizations Manage what they Know*, Havard Business School Press.

Davenport, T. H. and Short, J. E. (1990), 'The new industrial engineering infomation technology and business process redesign', *Sloan Management Review* pp. 11–27.

Debevoise, T. and Geneva, R. (2008), *The Microguide to Process Modeling in BPMN 2.0*, 2 edn, Advanced Component Research.

Devlin, K. (1997), *Goodbye Descartes - The End of Logic and the Search for a New Cosmology of the Mind*, John Wiley and Sons, Inc.

Devlin, K. (2001), *InfoSense: Turning Information into Knowledge*, W. H. Freeman and Company.

Dominque, J., Fensel, D. and Hendler, J. A., eds (2011*a*), *Handbook of Semantic Web Technologies*, Vol. 2, Springer-Verlag Berlin Heidelberg.

Dominque, J., Fensel, D. and Hendler, J. A., eds (2011*b*), *Handbook of Semantic Web Technologies*, Vol. 1, Springer-Verlag Berlin Heidelberg.

Dreyfus, H. and Dreyfus, S. E. (1986), *Mind over Machine: the power of human intuition and expertise in the age of the computer*, Oxford, Basil, Blackwell.

Dunaway, M. M. and Bristow, S. E. (2011), Importance and impact of erp systems on industry and organization, *in* 'Readings on Enterprise Resource Planning', HEC Montral, chapter 1, pp. 7–18.

Evans-Pughe, C. (2013), 'Mapping the mind'.
URL: *http://eandt.theiet.org/magazine/2013/02/mapping-the-mind.cfm*

Fantl, J. (2012), '"knowledge how", the stanford encyclopedia of philosophy (winter 2012 edition)'.
URL: *http://plato.stanford.edu/archives/win2012/entries/knowledge-how/*

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *AI Magazine* **17**, 37–54.

Feldstein, H. (2012), 'Interview with harvey feldstein erpsim'.
URL: *http://erpsim.hec.ca/learning/#/curriculum/153 [Accessed 20/8/12]*

Forbes (2011), 'It doesn't take two years to create a good strategy'.
URL: *http://www.forbes.com/sites/sap/2011/10/07/saps-bill-mcdermott-it-doesnt-take-two-years-to-create-a-good-strategy. [Accessed 1/4/12]*

Friess, P. (2012), Towards dynamism and self-sustainability, *in* I. G. Smith, ed., 'The Internet of Things 2012; New Horizons', Halifax, UK.

Ganter, B. and Wille, R. (1999), *Formal Concept Analysis Mathematical Foundations*, Springer-Verlag, Berlin.

Gartner (2009), 'Gartner reveals five business intelligence predictions for 2009 and beyond'.

URL: *http://www.gartner.com/it/page.jsp?id=856714 [Accessed 15/12/10]* (accessed 2010-12-15)

Gazendam, H. and Liu, K. (2005), 'The evolution of organisational semiotics: A brief review of the contribution of ronald stamper', *Studies in organisational semiotics* .

Ginty, A. (2007), *Problem Based Learning*, Higher Education Academy, Escalate.

Gordon, K. (2007), *Princples of Data Management: Facilitating Information Sharing*, The British Computer Society.

Graham, F. (2011), 'BYOC: Should employees buy their own computers?'.
URL: *http://www.bbc.co.uk/news/business-12181570 [Accessed 2/2/13]*

Hack, S. and Lindemann, M. A. (2008), *Enterprise SOA Roadmap*, SAP Press and Galileo Press.

Haerder, T. and Reuter, A. (1983), 'Principles of transaction-oriented database recovery', *ACM Computer Survey* **15**, 287–317.

HEC Montreal (2011), 'ERPsim Lab'.
URL: *http://erpsim.hec.ca/*

HEC Montreal (2013), 'Release Notes - ERPsim 2013-2014'.

Hereth, J., Stummey, G., Wille, R. and Willez, U. (2003), 'Conceptual knowledge discovery - a human-centered approach', *Applied artificial intelligence.* **17**, 281–302.

Herschel, R. T. and Jones, N. E. (2005), 'Knowledge management and business intelligence: the importance of integration', *Journal of Knowledge Management* **9**(4), 45–55.

Hillen, H., Scherpbier, A. and Wijnen, W. (2010), *History of problem-based learning in medical education*, Oxford University Press.

Hitzler, P., Krtzsch, M. and Rudolph, S. (2009), *Foundations of Semantic Web Technologies*, Chapman & Hall/CRC.

Ichikawa, J. J. and Steup, M. (2013), *The Analysis of Knowledge*, Stanford.

Jung, T. (2013), 'Introduction to Software Development on SAP HANA'.
URL: *https://open.sap.com/course/hana1*

Kang, B., Jung, J.-Y., Cho, N. W. and Kang, S.-H. (2011), 'Real-time business process monitoring using formal concept analysis', *Industrial Management & Data Systems* **111**(5), 652–674.

Kant, I. (1988), *Logic; translated with an introduction by Robert S. Hartman and Wolfgang Schwarz*, Dover Publications.

Krafzig, D., Bank, K. and Slama, D. (2004), *Enterprise SOA: Service-Oriented Architecture Best Practices*, Prentice Hall.

Leger, P.-M., Robert, J., Babin, G. and andD. Lyle, R. P. (2011), 'ERP Simulation Game with SAP ERP: Logistics Game (Platinum Version)', **ERPsim Lab, HEC Montral**, 44.

Leger, P.-M., Robert, J., Babin, G., Pellerin, R. and Wagner, B. (2007), 'Erpsim', *HEC Montral, Montral, Qc.* .

Liebenau, J. and Backhouse, J. (1990), *Understanding information: an introduction*, Palgrave Macmillan.

Liere, D. W. V. (2007), 'Network horizon and the dynamics of network positions', *ERIM Electronic Series Porta, Erasmus University Rotterdam* .

Lingras, P. and Akerkar, R. (2008), *Building an Intelligent Web: Theory and Practice*, Jones & Bartlett Learning.

Litwin, P. (1994), 'Fundamentals of relational database design', Microsoft TechEd.
URL: *http://www.deeptraining.com/litwin/dbdesign/fundamentalsofrelationaldatabasedesign.as*
[Accessed 9/2/13]

Luhn, H. (1958), 'A business intelligence system', *IBM Journal of Research and Development* .

Marz, N. and Ritchie, S. E. (2011), *Big Data: Principles and best practices of scable realtime data systems*, Manning Publications.

Mason, J. (2002), *Analysing Qualitative Data*, Routledge, chapter Linking qualitative and quantitiative analysis.

McComb, D. (2004), *Semantics in Business Systems: The Savvy Manager's Guide*, San Francisco, US, Elsevier.

McKernan, J. (1996), *Curriculum Action Research*, Kogan Page Limited.

McNiff, J. and Whitehead, J. (2006), *All you need to know about Action Research*, SAGE Publications Ltd.

McNiff, J. and Whitehead., J. (2009), *You and your action research project 3rd ed.*, Jean McNiff And Jack Whitehead.

Mireault, P. (2011), Business intelligence, *in* 'Readings on Enterprise Resource Planning', HEC Montral, chapter 16, pp. 209–232.

Muir, N. and Kimbell, I. (2010), *Discover SAP*, 2 edn, Galileo Press.

Muller, J. (2013), 'In-memory data management'.
URL: *https://open.sap.com/courses/1/wiki/downloads?module_item_id=57*

Novak, J. D. and Caas, A. J. (2008), The theory underlying concept maps and how to construct and use them, Technical report, Florida Institute for Human and Machine Cognition.

NVivo (2012), 'NVivo 9 Help'.
URL: *http://help-nv9-en.qsrinternational.com/nv9_help.htm*

Oxford (2012), 'Oxford English Dictionary'.
URL: *http://oxforddictionaries.com/definition/english/logic?q=logic*

Pal, N., ed. (2005), *Advanced Techniques in Knowledge Discovery and Data Mining*, Springer Verlag.

Parmenter, D. (2010), *Key Performance Indicators*, John Wiley & Sons Inc.

Peirce, C. S. (1878), 'How to make our ideas clear', *Popular Science Monthly* **12**, 286–302.

Plattner, H. (2008), 'Trends and concepts', *Hasso Plattner Institure Lecture Notes* .

Plattner, H. and Zeier, A. (2011), *In-Memory Data Management*, Springer-Verlag Berlin Heidelberg.

Poelmans, J., Dedene, G., Verheyden, G., Mussele, H., Viaene, S. and Peters, E. (2010), Combining business process and data discovery techniques for analyzing and improving integrated care pathways, *in* P. Perner, ed., 'Advances in Data Mining. Applications and Theoretical Aspects', Vol. 6171 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 505–517.
URL: *http://dx.doi.org/10.1007/978-3-642-14400-4_39*

Poelmans, J., Elzinga, P., Dedene, G., Viaene, S. and Kuznetsov, S. (2011), A concept discovery approach for fighting human trafficking and forced prostitution, *in* S. Andrews, S. Polovina, R. Hill and B. Akhgar, eds, 'Conceptual Structures for Discovering Knowledge', Vol. 6828 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 201–214.
URL: *http://dx.doi.org/10.1007/978-3-642-22688-5_15*

Poonen, S. (2012), 'SAP Solutions Portfolio - "The Big Picture"'.
URL: *http://www.sap-tv.com/video/7344/sap-solutions-portfolio-the-big-picture-presented-by-sanjay-poonen. [Accessed 01/04/2012]*

Porter, M. E. (1985), *Competitive advantage :Creating and sustaining superior performance*, Free Press.

Porter, M. E. (1996), 'What is strategy?', *Harvard* **November-December**, 61–78.

Portousal, V. and Dunderam, D. (2006), 'Business processes: Operation solutions for sap implementations', *Idea Group Inc* .

Presthus, W. (2012), 'Never giving up: Challenges and solutions when teaching business intelligence', *The Norwegian School of IT.* .

Priss, U. (2006), 'Formal concept analysis in information science', *Annual Review of Information Science and Technology* **40**(1), 521–543.
URL: *http://dx.doi.org/10.1002/aris.1440400120*

Rajaraman, A., Leskovec, J. and Ullman, J. D. (2012), *Mining of Massive Datasets*, Stanford University.

Sabherwal, R. (2007), 'Succeeding with business intelligence: Some insights and recommendations', *Cutter Benchmark Review* **7(9)**, 5–15.

Sabherwal, R. and Becerra-Fernandez, I. (2011), *Business Intelligence: Pracitces, Technologies and Management*, John Wiley & Sons.

Saldana, J. (2009), *The Coding Manual for Qualitative Researchers*, Sage Publications Ltd; 1st edition.

SAP (2011*a*), 'In-memory computing; Better insight faster with the SAP in-memory applicance (SAP HANA)'.

URL: *http://www12.sap.com/platform/in-memory-computing/in-memory-appliance/index.epx*

SAP (2011*b*), 'SAP A.G.'.

URL: *http://www.sap.com/*

SAP (2012*a*), 'About SAP A.G.'.

URL: *http://www.sap.com/about-sap/about-sap.epx [Accessed 8/10/12]*

SAP (2012*b*), 'BI on Demand powered by HANA'.

URL: *http://www.biondemand.com/businessintelligence [Accessed 15/8/12]*

SAP (2012*c*), 'SAP Help - STAD'.

URL: *http://help.sap.com/saphelp_nwpi711/helpdata/en/ec/af4ddc0a1a4639a037f35c4228362d*

content.htm. [Accessed 10/10/2012]

SAP (2013), 'SAP Lumira Solution Brief', *SAP A.G.* .

Sheffield-Hallam-University (2009), 'Smart applications'.

Sheffield Hallam University (2010), 'Smart applications module descriptor'.

Sheffield Hallam University (2012), 'Enterprise systems module descriptor'.

Shingo, S. (1986), *Zero Quality Control: Source Inspection and the Poka-Yoke System*, Productivity Press.

Siemens, G. and Gasevic, D. (2012), 'Learning and knowledge analytics', *Journal of Educational Technology & Society* **15(3)**, 1–2.

Soukup, T. and Davidson, I. (2002), *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, John Wiley & Sons Inc.

Sowa, J. F. (2000), *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA.

Stampler, R. (1973), *Information in Business System Adminstration*, John Wiley & Sons.

Stockburger, D. W. (1998), *Introductory Statistics: Concepts, Models, and Applications*, WWW Version 1.0, Missouri State University.
URL: *http://www.psychstat.missouristate.edu/introbook/sbk00.htm [Accessed 1/2/12]*

Stumme, G., Taouil, R., Bastide, Y. and Lakhal, L. (2002), 'Conceptual clustering with iceberg concept lattices', *Data & Knowledge Engineering* **42**.

Taylor, H. (2012), *CEP (Complex Event Processing) for Dummies, SAP Special Edition*, John Wiley & Sons, Inc.

W3C (2004), 'Resource description framework'.
URL: *http://www.w3.org/RDF/ [Accessed 9/2/13]*

W3C (2013), 'W3C'.
URL: *http://www.w3.org/ [Accessed 9/2/13]*

Watmough, M., Polovina, S., Khazaei, B. and Hill, R. (2010), Future supply chain processes and data for collective intelligence, *in* F. Xhafa, L. Barolli, H. Nishino and M. Aleksy, eds, 'Proceedings of the 2010 international conference on P2P, parallel, grid, cloud and internet computing (3GPCIC)', IEEE Computer Society, pp. 179–185. Paper presented at the 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3GPCIC), Fukuoka Institute of Technology, Fukuoka, Japan on 4 November 2010.
URL: *http://shura.shu.ac.uk/2719/*

White, S. (2013), Conceptual structures for stem data, *in* H. Pfeiffer, D. Ignatov, J. Poelmans and N. Gadiraju, eds, 'Conceptual Structures for STEM Research and Education', Vol. 7735 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 1–21.

URL: *http://dx.doi.org/10.1007/978-3-642-35786-2_1*

Wille, R. (1997), *Conceptual Graphs and Formal Concept Analysis*, Springer.

Wille, R. (2001), 'Why can concept lattices support knowledge discovery in databases?', *Proceedings of the CLKDD*01 Workshop on Concept Lattices-based Theory, Methods and Tools for Knowledge Discovery in Databases* **42**, 7–20.

Wille, R. (2005), *Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies*, Vol. 3626 of *Formal Concept Analysis, Lecture Notes in Computer Science*, Springer Berlin Heidelberg, chapter Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies, pp. 1–33.

Wolff, K. E. (1993), 'A first course in formal concept analysis: How to understand line diagrams', *Faulbaum, F. (ed.) SoftStat'93, Advances in Statistical Software 4, 429-438.* .

Wormuth, B. and Becker, P. (2004), Introduction to formal concept analysis, *in* '2nd International Conference of Formal Concept Analysis', Springer.

Yevtushenko, S. (2006), 'Concept explorer'.

URL: *http://sourceforge.net/projects/conexp/ [Accessed 15/9/10]*

Yevtushenko, S. (2010), 'Concept explorer'.

URL: *http://sourceforge.net/projects/conexp/ [Accessed 15/9/10]*

Yin, R. K. (2009), *Case Study Research: Design and Methods,*, Vol. 5 of *Applied Social Research Methods*, Thousand Oaks.

Appendix A

Ethics Statement Presented to Students

The work you are undertaking for your assignment represents some of the leading work in this field, so you could also be making a valuable contribution to advancing knowledge through your own learning. In accordance with Sheffield Hallam University's Research Ethics and Standards (http://www.shu.ac.uk/research/ethics.html), Martin Watmough would thus like to notify you of his wish to use your findings to support his PhD research 'Discovering the Hidden Knowledge in Transaction Data through Formal Concept Analysis'. Martin is an SAP Consultant by profession and will assist you during this assignment as a guest tutor particularly with respect to SAP knowledge and understanding. It is hoped that his use of your work is acceptable. If you have any concerns however, please contact Martin (mjwatmou@my.shu.ac.uk). Additionally this will be communicated verbally during any sessions where Martin will be the guest tutor. Any concerns or questions you have will be dealt with to your satisfaction. Your assessment will not be affected in any way, and you will not be referred to by name directly or indirectly in any outcomes of this work.

Appendix B

Word Frequency NVivo

Core word	Stemmed Words and Synonyms						
lattice	lattices						
spent							
days							
ordered	ordered	orders	range				
give	give	giving	leave	makes	making	passed	reached
part	part	percentage	region	regions	section	sections	
matnr							
raise	proved	proves	raise	rises			
happened	material	occur	occurred	passed			
affected	affected	affecting					
Found	establish	found					
regions	areas	region	regions				
attributes	attributes						
begin	beginning	begins					
behaviour							
dailyprofit							
euros							
highlight	highlighting	highlights					
money							
relationship							
test	proves	test					
answered	answered	answers					
small	small						
exceed	extreme	extremes	passed				
selecting	take						

Table B.1: Words in FCA and not in Excel/BI

Core word	Stemmed Words and Synonyms								
multiple	multiple	time							
chart	chart	charts	graph	graphs					
point	direct	item	items	levels	peaked	periods	place	point	
highest	highest								
pcprofit	pcprofit								
going	breaking	functionality	going	lead	leads	loss	passed	running	runs
observation	find	note	noticed	observation	observations				
setting	defined	location	lots	place	sets	setting			
deciding	deciding	definitively							
generally	general	generally	popular						
help	help	helped	helps						
lowest	lowest								
spritz	spritz								
running	campaigns	carry	course	functionality	lead	leads	passed	plays	runs
factors	factor	factors							
family	family								
goes	goes								
altering	altering	change	changing						
ctree	ctree								
entering	entering	figure	figures	introduced					
individual	individual	person	single						
interestingly	interestingly								
long	long								
longest	longest								
negative	negative								

Table B.2: Words in Excel/BI not in FCA

Core word	Stemmed Words and Synonyms								
profit	positive	product	products	profit	profits				
marketing	market	marketing	sell	selling					
shows	appears	establish	proved	proves	read	showed	showing	shows	
quarter	quarter	quarters							
amount	amount	number	quantity	total					
companies	companies	company							
price	price	priced	prices						
results	answered	answers	effect	leads	leave	outcome	result	resulted	results
daily	daily								
made	made								
products	output	product	products						
sales	sale	sales							
units	combination	combined	units						
high	extreme	extremes	high	highs					
time	time								
cover	continue	cover							
spend	expenditure	passed	spend						
data	data	information							
higher	higher								
stock	inventory	stock							
analysis	analysis								
figures	figures	forecasting	forecasts	number					
average	average	mean	meaning	medium					
need	demand	necessarily	need	needs	required	take			
decisions	decision	decisions	finally						

Table B.3: Matched in FCA

Core word	Stemmed Words and Synonyms						
price	cost	costs	price	priced	prices	pricing	
marketing	market	marketing	sell	selling	sells		
sales	sale	sales					
profit	benefits	gained	product	products	profit	profits	
company	companies	company					
stock	carry	standard	stock				
shows	point	proves	proving	show	showed	shows	
result	effect	effected	effects	lead	leads	result	resulted
rule	find	rule	rules				
high	high						
amount	amount	amounts	number	total	totally		
product	product	products	yield	yields			
increase	gained	increase	increased	increasing			
clear	clear	gained	make	makes	making	passed	
spend	dropped	passed	spend				
quarter	quarter	quarters	tailed				
analysis	analysis						
area	area	areas	regional				
data	data						
higher	higher						
decision	conclusion	decision	decisions	final			
order	consistent	logically	order	ordering	orders	place	
lower	lower	lowered	lowering				
created	created	make	makes	making	produced		
need	demand	necessarily	need	want			

Table B.4: Match in Excel/BI

Appendix C

Knowledge Discovery

Key:

Key	Description
F	Fact
G	General Rule
P	Pivot Point
2R	Linear Relationship - 2 Variables
3R	Linear Relationship - 3 Variables
U	Unknown identified
O	Outside Influence
M	Method Suggestion
E	Unexpected Discovery

C.0.2 Excel/BI Explicitly used for Knowledge Discovery

Key	Coded Text
F	No stock = no sales
U	Price is not the only factor effecting sales

Table C.1: Sales Forecast

Key	Coded Text
G	The higher the price, the lesser the sales and vice versa
U	That suggests that the price does not explain everything
G	Low selling price may not always have the highest sales
G	Medium sales price sells the most
2R	When companies lowered the prices it improve sales
G	Lowest prices don't necessarily attract the highest sales
3R	Marketing had a negative effect and it was the selling price of the products that had the biggest effect on profit

Table C.2: Maintain Master Data (Sales Price)

Key	Coded Text
2R	There is a steady constant between marketing costs and profit gained
2R	With marketing in place the company makes more profit.
G	Marketing had a negative effect and it was the selling price of the products that had the biggest effect on profit
2R	If sales = low => increasing marketing
3R	If profit = negative and selling price > standard price => decrease marketing
2R	Regional sales = strong => decrease marketing
2R	Regional sales = weak => increase marketing

Table C.3: Marketing

Key	Coded Text
2R	Marketing had a negative effect and it was the selling price of the products that had the biggest effect on profit
O	The rise in profits is due to Company HH selling lots of products, this is probably the result of other companies running out of stock at that particular time.
O	These observations combine to show that other factors have a large impact on profit - e.g. marketing spend, other company behaviour

Table C.4: Other

C.0.3 FCA Explicitly used for Discovery

Key	Coded Text
G	We can see that on the lattice there is an association between Days_Cover and PC_Profit
F	When the days cover is low then profit is also low

Table C.5: Sales Forecast

Key	Coded Text
U	The price changes explain some of the behaviour during the test, but is not enough to explain the number of sales
U	We can see that decisions to raise or lower prices had little direct effect on the profit

Table C.6: Maintain Master Data (Sales Price)

Key	Coded Text
F	Higher marketing the higher profit
F	Marketing expenses at 0 managed to receive the least profit
2R	It also appears that the duration of marketing also affected the outcome of the profit.
F	most popular amount spent on marketing per day was over 2000 at 48%.
P	This leads to the assumption that there is an upper and lower limit on marketing, which when passed has an adverse effect on the daily profit.
M	Do not market any products in the first quarter, test the market regions to observe how high the selling price could be set without affecting profits.
M	Once the optimum price has been found, begin small increments of marketing, if the amount of units sold begins to decrease.
2R	If Day Cover = 5 THEN Pc_Marketing = Very Low
2R	If Day Cover = 15 THEN Pc_Marketing = Low
2R	If Day Cover = 25 THEN Pc_Marketing = Average
2R	If Day Cover = 30 THEN Pc_Marketing = Very High

Table C.7: Marketing

Key	Coded Text
O	A direct relationship between marketing and daily profit cannot be identified. To fully understand the reason for the change in daily profit, other areas of the 'game' need to be analysed
E	Showing that results can differ depending on the medium used for analysis.
U	Marketing cannot alone justify this wide range of daily profit values because there is little change in it - item price and stock levels would need to be considered in order to explain these fluctuations.
E	The results returned showed that certain products sold better than others
E	Clearly, this second lattice is far more manageable, but unfortunately no useful rules can be derived or proved from it either.

Table C.8: FCA: Other

Key	Description	Count: Excel/BI	Count: FCA
F	Fact	1	4
U	Unknown identified	2	3
G	General Rule	5	1
P	Pivot Point	0	1
2R	Linear Relationship - 2 Variables	7	5
3R	Linear Relationship - 3 Variables	2	0
O	Outside Influence	2	1
M	Method Suggestion	0	2
E	Unexpected Discovery	0	3

Table C.9: Summary

Appendix D

Complete set of Coding Nodes Applied in NVivo

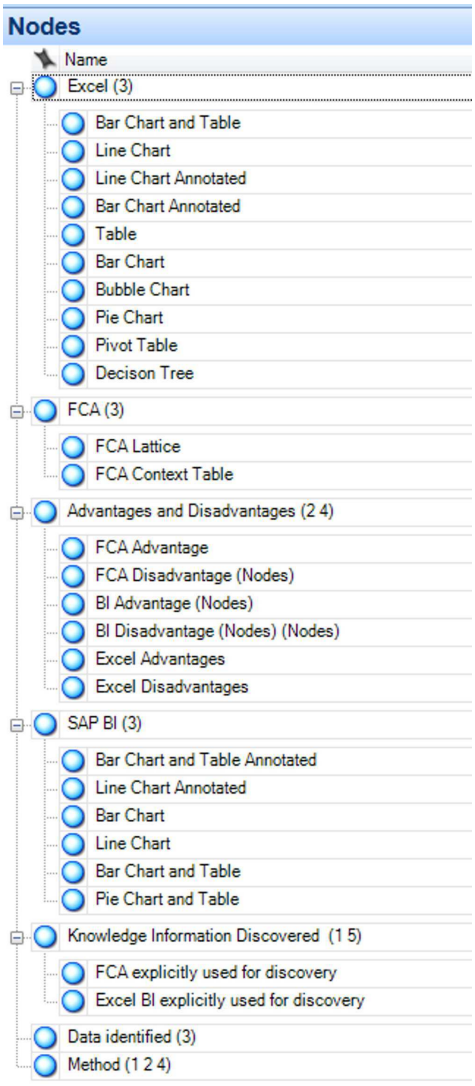


Figure D.1: Complete set of Coding Nodes Applied in NVivo

Appendix E

Discovery through an LTA Design

E.1 Introduction

This appendix provides a more detailed description of the discovery through an LTA design, section 4.2. Utilising an opportunity to incorporate Formal Concept Analysis (FCA) into two degree courses at Sheffield Hallam University, Learning, Teaching and Assessment (LTA) acted as an environment for problem solving, learning and study of FCA in action.

Structured cases formed the data collection mechanism for this applied study of FCA and contemporary tools, additionally assisting in maintaining a good pedagogic outcome. Fundamentally the student's learning objectives focus on education about systems and applications, in parallel developing an in depth understanding of data. The outcome with respect to the research question did not influence their learning or development of opinions.

Action research and a case studies strategy iteratively developed an environment for generating experimental data representative of industry standards. This first experiment utilised situation theory, problem based learning and a hybrid combination of Yin's Case Study Method and Bigg's constructive alignment resulting in a critique of tools and techniques in the process.

Applying FCA in a multi dimensional simulation based in a learning environment

demonstrated that knowledge discovery through of transaction data is practical and that there is scope for FCA to be an integral part of a Business Intelligence (BI) solution. The exercise and been pertinent to students learning experience and relevant to the research question.

E.2 Creating an Environment

E.2.1 Generating Useful Data

Enterprise Resource Planning (ERP) systems are typically transactional systems that support the core functions within an organisation. SAP A.G. are one of the leading providers (SAP, 2012*a*). The analysis on ERPsim (Leger et al., 2007), an SAP A.G. ERP based simulation game that features competitive behaviour and increasing levels of complexity in a highly immersive and demanding atmosphere that reflects industrial practice.

Creating value using semantic technologies is not significantly different to other technologies, three important aspects are the customers, business model and technology (Dominque et al., 2011*a*). This viewpoint is equally applicable to FCA and can be overlaid onto Porters Value Chain (Porter, 1996), essentially this tries to achieve balance and sound business model.

The source data was the result of a game between competing student teams undertaken on a mainstream ERP system provided by the business software vendor SAP A.G. (SAP, 2011*b*) and using the ERPsim software provided by ERPsim Lab at HEC Montreal (HEC Montreal, 2011). The simulation generated and captured data on which Business Intelligence was performed. The data generated by the simulation represents typical business activity and is not specifically for FCA, thus it provides a meaningful test of FCA in BI from ERP data.

ERPsim is based on SAP ECC 6.0 which is an ERP system capable of supporting in

this example logistics and financial activities for a number of competing companies. All sales, procurement, master data, inventory, marketing and financial transactions are captured real time in additional to a limited number of reports to show sales, inventory, balance sheet and profit and loss. These are transaction based reports and offer no analysis without the application of further tools.

As an ERP system is effectively a relational database with data held in joined tables it is possible to extract data that contributed towards a goal via a query. Therefore a query using the table relationships was able to extract all the transactional data available that contributed towards the outcome. For example all sales transactions within the time period could be found via the connection from billing through the outbound shipments to the sales orders. Correspondingly individual sales order profit based on the materials cost price could also be extracted.

The chart in figure E.1 provides an example of the input and output variables plotted to highlight the relationships that can exist in the simulation game. On the right hand side cumulative profit and percentage profit above cost per sale are shown. Cost is indexed at 100%, therefore 105 equates to 5% profit over cost. On the left hand side days inventory cover and the percentage of sales price attributable to marketing spend is shown. In summary the data could represent a number of relationships including:

- Increasing cumulative profit has an inverse relationship with decreasing days of inventory cover (how many days the stock will last given the sales forecast). [Holding less stock will result in more profit]
- Increasing profit has a direct relationship with increasing marketing spend. [Spending on marketing leads to more sales, therefore more profit]
- Increasing profit has a direct relationship with increasing profit per sale. [More profitable individual sales leads to higher overall profit]

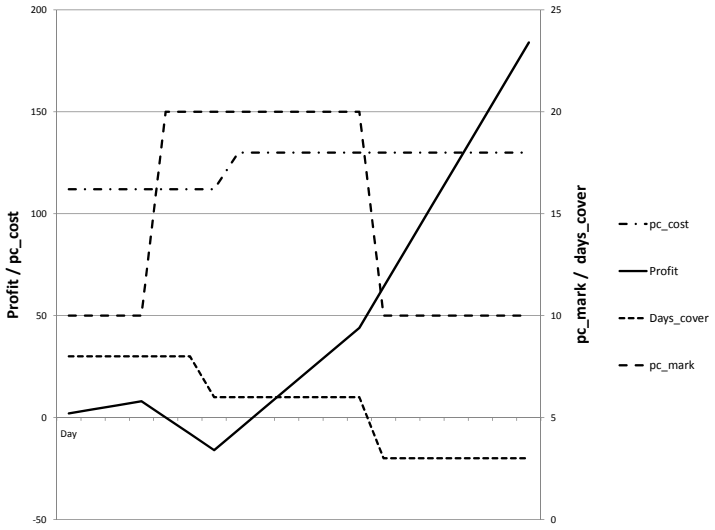


Figure E.1: Example of Input and Output Variables

E.2.2 Pedagogy

This research interests arise from how to discover the hidden semantics within transactional systems i.e. how useful information or knowledge can be identified from mainstream database systems, by applying and combining analysis techniques. To assist, at Sheffield Hallam University we have incorporated this research into two Computing degree course modules. As such we aim for the research to be informed by the student's experiences, whilst enriching the student's knowledge in this topical area. Accordingly this paper focuses on the incremental design of the learning environment in the pursuit of discovering hidden semantics and the results achieved.

By applying and developing the approach to teaching transactional systems and analysis, two benefits are envisaged. Firstly, an insight into how learning these methods benefits the modules and students. Secondly, to engender a creative arena that encourages open answers from the students. Formal Concept Analysis (FCA) is a technique for analysing data in order to discover information and knowledge. FCA is particularly attractive in that offers an automated means of eliciting these concepts

from the data (Wille, 1997) (Wolff, 1993). Therefore, FCA was selected as the underlining technique for designing learning in order to research the hitherto hidden semantics in transactional data.

ERPsim has a strong pedagogic foundation that has been adopted and applied during the development of the degree modules. ERPsim is designed for active learning in that it achieves long-term retention. ERPsim takes advantage of Situation Cognitive Theory and Problem-based Learning (Feldstein, 2012).

Situation theory states that activities, tasks, and understanding do not exist in isolation, but rather are part of broader relation systems and that situated cognition is associated with a higher level of engagement and motivation in learners, thus generally leads to a better understanding and transfer of knowledge (Leger et al., 2011). Problem-based learning is a widely applied technique that has its origins dating back to 1966 in medical education (Hillen et al., 2010). It is a teaching strategy to promote self-directed learning and critical thinking through problem solving in which active participation and challenging problems in a relevant context are key (Ginty, 2007).

Furthermore, the learning environment created by ERPsim has been carried into the analysis of its output by comparative techniques. These techniques, described later, are used in order to evaluate the comparative value of FCA for transactional data.

To assess the effectiveness of this learning, teaching and assessment (LTA) we examined the marks achieved and learning objectives; the findings and feedback from the students have also been considered. It should be clarified that the students on these modules' did not have a significant mathematical background; rather the modules focussed on the business application of FCA. For this reason and to preserve consistency we ensured that the FCA tools were explained and applied according to their understanding. The fact that the raw data structure was constant aided the process.

E.3 Comparing FCA to Contempory BI tools

This section describes the evaluation of coursework set for final year degree students designed to teach Formal Concept Analysis (FCA). The assessment applied a set of FCA tools and conventional Business Intelligence (BI) using graphical or statistical methods in Microsoft Office Software. There were to distinct objectives to this activity, firstly to fulfil the students learning objectives and secondly to support an action research project about the application of FCA within Enterprise Resource Planning (ERP) systems. The fulfilment of the learning activity was assessed in two ways, firstly as a comparison between the use of conventional BI analysis and FCA tools and secondly as a comparison of FCA against established theory. Topics for the action research project are highlighted in the conclusions and further work sections, however, this is not the primary focus of this paper.

The usefulness of this approach is discussed with respect to its application in future iterations of the coursework. The source data was the result of a simulation between competing student teams undertaken on a mainstream ERP system provided by the business software vendor SAP A.G. and using the ERPsim software provided by ERPsim Lab at HEC Montreal. The simulation generated data on which Business Intelligence (BI) is typically based and is representative of business activity. The data generated by the simulation exercise was not specifically for FCA, thus it provides a meaningful test of FCA in BI.

The rationale for selecting this research is due to the demands being placed on BI systems to improve and the difficulties in identifying semantic data. A simple definition is"semantics = data + behaviour" (McComb, 2004). This suggests that if the semantic content can be identified it may be possible to understand or determine behaviour.

The coursework is described in more detail later, however the principle is to introduce frameworks and techniques for representing and reasoning with knowledge for

smart applications (Sheffield-Hallam-University, 2009). The principle of the coursework is to compare how analysis using tools such as Microsoft Excel compares to a FCA tool set using data generated through the realistic use of an ERP system. The students entered into the analysis with a practical knowledge of the processes that generated the data set but having performed no analysis or reflection on the impact of decisions made during the simulation.

The need for analysis and decision making within enterprises is not new but competition and complexity do combine to make the task vast and difficult to execute efficiently or accurately. Business Intelligence (BI) is frequently used to support analysis and decision making and can be traced back at least as far as 1958 (Luhn, 1958), however, it remains a field that is subject to much ongoing research and development. Gartner (Gartner, 2009) predicts that business units will control at least 40 per cent of the total budget for BI, a reason cited for this is that a significant percentage of companies regularly fail to makie insightful decisions about their business and markets. This implies that tools must be suitable for non technical users while encompassing the reliability and flexibility for application in modern environments.

E.3.1 Method

The primary problem is how to analyse data and identify semantic data or relationships from a generic transactional data set. The coursework addressed three of the learning outcomes from the course (Sheffield-Hallam-University, 2009):

1. Describe the notion of representing and reasoning with knowledge for smart applications.

2. Draw on one or more frameworks and techniques for representing and reasoning with knowledge for smart applications.

3. Identify the practical use of software tools for developing smart applications.

The scenario presented to the students was:

You are performing the role of a business analyst who has been tasked with analysing the performance of your ERP Water Company by understanding how a) your decision-making and that of others has impacted the organisation and b) identifying rules that could be used to help this decision making in the future. You are also evaluating the method of analysis in order to refine the approach employed for future iterations of this process. It is therefore less the intention to learn ERP; rather through this experience you will explore business intelligence and the role that Formal Concept Analysis (FCA) might play in this context.

The coursework had three main sections consisting of conventional BI, FCA and Evaluation / Conclusions. The BI analysis would use MS Access and Excel in order to familiarise the students with the data using tools that would already be familiar and offer graphical analysis techniques that are common and taught at a school level of mathematics. Secondly FCA tools are applied based on essentially the same data with any calculated values added to support the analysis. This is expected to be an iterative process in order to produce the best results possible but the core section of the data extract should be stable and reusable. The final section is an evaluation of the two approaches and conclusions.

It is acknowledged that the goal of both the BI and FCA approach is to identify potentially the same relationships, this is deliberate in order to encourage an understanding of the data using tools and techniques in applications such as Excel that will be familiar and well supported with documentation and guides. An understanding of the data and relationships was deemed necessary given the students had no prior knowledge of ERP systems or an understanding of the processes in operation.

The tools set consists of five key software packages: MS Access as a mechanism for extracting data from the ERPSim SAP system and creating the initial data file (CSV) for analysis, MS Excel and FCA tools including: FcaBedrock (Andrews and

Technique	% Occurrences
Line chart (2 variables)	100
Graph on graph comparisons	69
Cumulative and actual data charts	62
Detailed Focus with annotation.	46
Line chart (3 variables)	23
Pie chart	23
Data table	15
Pivot table	15
Summary table (annotated)	8
Use of trend lines	0

Table E.1: Methods Applied under BI

Orphanides, 2010*b*), In-close (Andrews, 2011*b*) and Concept Explorer (Yevtushenko, 2010).

The method selected was generally an experimental and iterative approach in order to extract and analyse key data, gradually refining the method to explore the anticipated relationships and evaluate the capabilities of the tool set. The aim was to supply a consistent set of data to the FCA tool set making it a repeatable process.

E.3.2 Student Results

The basic analysis methods applied across all the course work are shown in Table E.1 and E.2 with a percentage occurrences. It is noted that the marking of the coursework did take into account more than the range of techniques applied.

A minority of students also attempted to identify rules that explicitly stated relationships and could be reused in future iterations of the simulation.

The average mark achieved was 57 % with a standard deviation of 15.3.

Tables E.3 and E.4 contains a summary list of points made within the Evaluations and Conclusions section of the coursework for BI and FCA respectively.

Technique	% Occurrences
Analysis over 2 data ranges	69
Percentages of occurrences	54
Identification of Relationships	54
Analysis by product profitability	38
Use of Ranges	38
Analysis over 3 data ranges	38
Analysis by Profit by quarter	23
Performance measures / KPIs	23
Graph on Graph Comparison	8

Table E.2: Methods Applied under FCA

Pros	Cons
Good compatibility with data sources / MS Access etc	Data can be manipulated / changed manually at the interpretation or error of the user
Easy to learn	Have to drive the analysis and discover trends, no automation
Can manipulate data and combine with charts/diagrams	Required manual input to compare multiple charts etc
Hands on, easy to manipulate data.	Difficult to represent hierarchies in the data
Graphical options give quick visual descriptions of any rules/trends	Tools do not replace expert knowledge
Handles different data types, formulas	Data can be misunderstood
Reliable software	
Widely available	
Reuse / Refresh of charts etc	

Table E.3: Pros and Cons for BI

E.3.3 Discussion

The initial reaction of the students was one of confusion in how to tackle the coursework, this is reflected to a degree in Tables E.1 and E.2, these show that less complex forms of analysis were prevalent in all work, for example line charts with two variables, but relatively few progressed onto considering more complex selections such as line charts with three variables. A little trial and error coupled with confidence could have eliminated most problems, this could also be supported better with guided examples attempting the coursework.

Pros	Cons
Good for analysing small data sets	Difficult to refine data, particularly large data sets
Data can be refined in FCA	Involved manual manipulation of data source
Good for displaying large amounts of data	Difficult to identify anomalies in the data and to correct.
Lattice covers all possible aspects (with Concept Explorer)	Many different formats, applications time consuming
Relationships are highlighted visually	Difficult to pin point trends/rules in concept form (for this example)
A level of interaction with the data	Any data must be calculated for going into FCA and was therefore reliant on other tools to structure the data, i.e. Excel
Analysis of relationships between unconnected data categories.	Comparing multiple lattices etc. is not supported directly.
Good for viewing hierarchies	Lack of statistics or alternative graphical analysis or drill down to raw data
	Data has to be consolidated to a large extent (to much) before the lattice is readable.
	Difficult to reuse not integrated with source data.

Table E.4: Pros and Cons for FCA

The marking of the coursework produced a normal distribution of marks with an unexpected enthusiasm for FCA although this was tempered by the difficulties in using the tool set. It is not surprising that they experienced difficulties given the difference between the development effort behind the FCA tool set and BI tools from providers such as Microsoft. It could be surmised that the students understood the advantages of analysing large and relatively unstructured data without expert knowledge or time consuming analysis. It would have been nice to see the students experimenting more with the data and discovering or at least looking for less obvious relationships.

A consistent criticism of the FCA tool set, see table E.4, was the difficulty level involved in data preparation and use of the tools. It would have been nice to eliminate

some of the repetitive tasks required by the exercise as the students struggled to grasp and achieve a reusable data extraction mechanism, therefore consuming time that could have been spent more productively on the analysis. A problem that is not uncommon in real life applications.

The presentations produced for assessment made it relatively easy to mark however it was sometimes difficult to understand what was trying to be communicated especially where annotations or additional notes were not present or of low quality. The graphical nature of the presentation medium did form a good basis for presenting the analysis and forced a summary rather than lengthy descriptions of the process and mechanisms involved.

It was clear from the conclusions in Tables E.3 and E.4 that an appreciation about the difficulties involved in delivering BI was achieved even from this relatively small data set.

The students generally struggled to identify data or relationships outside of the key parameters provided, this is partially due to the data available as it was only a partial extract of ERP systems. Even so there are many factors that could have been offered for consideration even if they could not directly be included in the analysis. Examples of this could include the team structure or the decision making of certain individuals being categorically better or worse in outcome to others.

Graph on graph comparisons featured highly in the BI analysis, essentially this included graphical comparisons that were either overlaid or annotated to illustrate an event or relationship. Considering the frequency of this type of analysis when it came down to the FCA tools set it was hardly applied, even though the concept lattices are primarily a visual tool. The reasons for this were not clear and possibly related to the difficulty experienced in using the tool set. This feature was not supported in the tool set but it was easily possible to capture and present images side by side within the presentation.

Discrete values proved much easier to understand than ranges, in order for ranges

to be understood manual input is required in order to create meaningful sub ranges. Progressive scaling was applied but the definition of the discrete values was not appropriate to take advantage of this. With this in mind a bi-ordinal scale would be more useful when representing such values but this will require a different approach when extracting the data or within FCA.

As soon as the analysis required calculations to be performed it started to face many of the challenges also faced by BI. Firstly there may be differences in the calculations between analysts, regions or indeed of interpretation. Secondly, calculated figures and performance measures can lack scale. The analysis was more successful when focus was given to a specific attribute, this was achieved by restricting the data being analysed. The down side of this was that it was a manual process with relatively long iterations even though the source data set did not alter. This limits the scope of data available and potentially the results obtained which could be a significant disadvantage.

It was clearly difficult to analyse the lattices unless a specific feature was chosen as the focus for the analysis, primarily due to their size and complexity. A possible side effect of focussing would be the accidental exclusion of data that could highlight unknown or unexpected relationships which should have been a major benefit for this type of analysis. The whole problem of visualising and exploring or "concept exploration" as termed by Stumme (Priss, 2006) is proving to limit the usefulness of this approach graphically at this time but alternative methods of applying the results may be possible that either solve this issue or do not require graphical representation.

The analysis was limited as it only included attributes that could be attributed to a strategic goal within the ERP system. Making the link within relational database is relatively straight forwards however a far greater challenge would be including data from sources with less well defined relationships. This maybe possible using tentative links such as times and dates but further work is required. This could be achieved within the data extract query as applied in MS Access for this approach.

E.3.4 Review of Learning Outcomes

Learning objective 1 - *Describe the notion of representing and reasoning with knowledge for smart applications.* This was visible in the coursework by the use of techniques such as performance measures / key performance indicators (KPI) within the data extraction on graphical interrogation of the outcome.

Learning objective 2 - *Draw on one or more frameworks and techniques for representing and reasoning with knowledge for smart applications.* This was visible in the coursework by the application of the tools and presentation of the analysis in the form of the coursework. The range of techniques applied further demonstrated the depth of analysis. There are a wide range techniques available and a reasonable range have been applied but only the minority of students have applied them.

Learning objective 3 - *Identify the practical use of software tools for developing smart applications.* This was visible in the coursework clearly by the conclusions where the ability to interact with the analysis and discover relationships was a clear advantage for FCA tools.

An emergent learning outcome was with regard to a developed appreciation of how the application of relatively simple analysis can highlight major flaws in the decision making processes employed during the game therefore resulting in poor performance. A number of teams indicated this and identified where mistakes had been made due to a lack of analysis or assumptions based on incomplete knowledge.

The learning outcomes have been achieved with all students appreciating the value and difficulties associated with analysing ERP data. The results did reflect a reasonable range of marks being awarded with all students able to perform both BI and FCA over the data set provided.

The difficulty involved in data preparation had a significant impact on the analysis performed, particularly with respect to the application of more complex analysis techniques and semantic discovery. This was the main factor that detracted from the

learning outcomes.

The coursework would benefit from more focus on the analysis and less effort required for the preparation of data. It is expected that significant manual input will still be required in terms of defining any calculations and manipulation of graphical outputs.

A structured criteria for the analysis techniques expected could lead to an improvement in the marks awarded. This could include a pre-configured solution containing the basic forms of analysis, therefore forcing the use of more advanced analysis methods as a minimum criteria for the coursework. This could be achievable by reducing in the amount of data preparation activity required, however this must not place a constraint on the experimental aspect of this coursework and the ability to perform an open analysis.

There is a continued value in applying two methods of analysis, the BI approach is already familiar to the audience and clearly help understanding of the data set. Applying purely an FCA approach would be very challenging at this point in time.

As part of the action research aspect a number of factors should be changed for the next deployment of this coursework in order to permit the students to progress towards more advanced use of FCA, this is detailed in Further Work.

It was clear from the conclusions that the notion of applying BI and FCA was understood and the value it has in real life applications. The value of good analysis and the ability to evaluation unknown relationships was imparted. Equally the potential for error, misunderstanding and potential lack of uptake because of the complexity was clear and echoed the comments from Gartner in the introduction with respect to what how analysis will be controlled by business units and not technical experts (Gartner, 2009).

E.4 Incremental Development of LTA

E.4.1 Method

The method was based on previous work that applied Biggs' Constructive Alignment and Yin's Case Study Method (Andrews, 2011*a*), (Biggs and Tang, 2011), (Yin, 2009). This method was modified to better support the learning outcomes and theory relevant to the analysis of business transactional data using FCA tools.

Yin's method was applied to capture and learn from a number of case studies, where each case study represents the relevant modules' assignments. There were four case studies, one from each module for the academic years 2011-11 and 2011-12. An overview is contained in Table E.5. Two aspects have been used for evaluation, firstly, the assignment marks per section have been compared with the teaching and learning techniques applied. Secondly, the student's evaluations and conclusions have been used for qualitative analysis. Biggs' Constructive Alignment has two basic concepts; learners construct meaning from what they do to learn and that the teacher makes an alignment between learning activities and learning outcomes (Biggs and Tang, 2011). The combination of Biggs' constructive alignment and Yin's Case Study Method provides an overall method for aligning the learning activities and learning outcomes for the benefit of future students (Andrews, 2011*a*). It was also envisaged that an insight into the introduction of FCA into an organisation's Business Intelligence capability would be gained.

The FCA tools used for this study were FcaBedrock (Andrews and Orphanides, 2010*b*), In-Close (Andrews, 2011*b*) and Concept Explorer (Yevtushenko, 2010). An overview of the steps is described in section 3.5.

To begin with, ERPsim stores the raw data from the game in a relational database. This raw data was extracted directly into a Microsoft (MS) Access database that mirrors the tables and relationships of the ERPsim system. MS Access queries were

then used to extract data into a comma seperated values (CSV) text format that contained the key attributes and meta data. From this file, FcaBedrock was used to created a Formal Context. In-Close was then used to provide minimum support by reduce the number of formal concepts, before graphically presenting its results using Concept Explorer as a concept lattice. Without the application of In-Close the output from FcaBedrock can be too complex for meaningful visualisations; in effect the less dominant relationships are removed.

The techniques summarised in Table E.5 were employed in order to develop the teaching and assessment methods. These have included learning in conjunction with ERPsim, a mix of individual and group work approaches and comparisons with alternative approaches.

The graph in Figure E.2 indicates how the assignment marks deviated from the average mark for each module. Taking case study 4 as an example, the students achieved higher than the average percentage for the introduction and lower for the FCA sections. The perfect line would run through zero with each student achieving the same percentage for each section of the assignment; as this is based on the average the performance does not differentiate between high and low achieving students.

E.4.2 Case Study Review

Beginning with Case Study 1, generic processing steps were intended to be reused over different subsets of the data incorporated. This however appeared to have only generated repetition and not an improvement in marks or learning. For comparison, the students were required to target the same data with FCA and Excel. The results achieved did not differ to a noticeable extent. Comments by the students suggested that significantly more time was required in order to apply and understand FCA, although its graphical nature did lend itself effectively to creating content for inclusion in the assignment.

Case Study	1	2	3	4
Module	SA 2010-11	ES 2010-11	SA 2011-12	ES 2011-12
Average Mark	56.6	58.4	66.8	58.6
Standard Deviation	15.3	21	3.8	11.5
Data Preparation Demonstrated	X	X		X
End to End Data Prepared			X	X
Graphical presentation	X			
Document		X	X	X
Excel and FCA	X	X	X	
BI, Excel and FCA				X
Group discussion	X	X		
Group work		X		
Jigsaw based approach			X	X
Horizontal and Vertical Group work				X
Re-use (multi company)	X	X		
BPM Integration			X	X

Table E.5: Chronology of Teaching Methods and Results

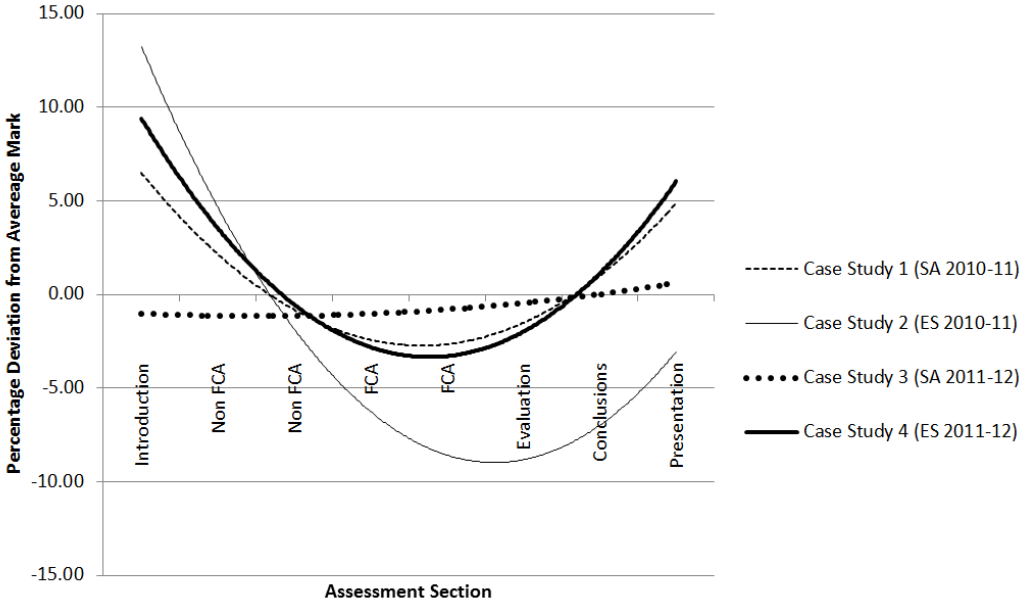


Figure E.2: Deviation per Section from Student Average Mark

Reducing or simplifying the quantity of data preparation was suggested by the students, but it was not clear if the challenge of performing this task was beneficial to the learning process. Instead of eliminating the data preparation task it was decided to reduce the individual workload through group work. Therefore, the group would still retain any learning from the experience while sub-dividing the manual effort required.

Case Study 2 generated clear comments and queries during the module indicating that a method for reducing the amount of time for preparing data is required, even with a group-based approach as collaboratively preparing data proved difficult to achieve. The majority of students managed the task but probably at the expense of actually performing and evaluating the analysis. An approach to improving group work and networking was also identified as the use of communication tools, predominantly the discussion boards and blogs were limited.

Learning the principles of generic design and reuse was successful but repeating the analysis for multiple scenarios did not add significant value. There were enough opportunities to repeat and tailor the analysis in a single section of the assignment to

support the learning outcomes. Reintroducing group based assignments would be an interesting choice to return to as the FCA tools develop and focus can be shifted to collaborative or even social topics; however, in the current context it did not have a positive effect upon the results.

Case Study 3 applied a technique for cooperative learning called Jigsaw (Aronson, 2012) that encourages participation and emphasises the value of every student's contribution towards the outcome. Jigsaw was intended to develop group problem-solving while maintain the individual's contribution to the task. An emergent outcome was identified in that this group cooperation has parallels with working in current or future workplaces that feature more diverse skills requirements, physical distributed teams and all manners of collaboration and communication mechanisms.

A complete set of data was prepared for each group through all stages, including instructions about how to modify, refine and enhance the analysis. Preparing the initial data for each individual team resulted in less creativity and fewer variations across the assignments. However, the average mark for FCA did improve.

Rule definition was introduced as a mechanism for expressing the findings as a formula or in a logic form. Deriving rules from the analysis was challenging for the students, however it appeared to complete the cycle back to source transactional data. The students demonstrated an understanding of the relationships discovered and how they could be applied to ERPsim processes.

Case Study 4 delivered the most comprehensive learning judging by feedback from the students. The Jigsaw (Aronson, 2012) structure employed resulted in the most frequent use of the discussion boards and collaboration. The groups were organised in two directions, vertically to promote interaction and team work within the group and horizontally to create cross group knowledge sharing, almost like expert communities between groups. A number of students found advantages and disadvantages from the analysis tools and envisaged a blend of approaches in order provide their ERPsim organisation with the best tools for future improvements.

An alternative to Excel was made available in the form of SAP A.G.'s product 'BI On Demand' (SAP, 2012*b*). This tool was chosen as it is representative of the leading vendors' retail products for BI. Although the interface featured a wide range of display options in an integrated package it did not dominate or significantly improve the marks achieved.