



DEVELOPMENT OF AN INTELLIGENT ANALYTICS-BASED MODEL FOR PRODUCT
SALES OPTIMISATION IN RETAIL ENTERPRISES

by

COURAGE MATOBOBO

submitted in accordance with the requirements

for the degree of

MASTER OF SCIENCE

in the subject

COMPUTING

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROFESSOR ISAAC O. OSUNMAKINDE

03 July 2016

DECLARATION

Name: Courage Matobobo

Student number: 49116762

Degree: MSc in Computing

Exact wording of the title of the dissertation or thesis as appearing on the copies submitted for examination:

Development of an Intelligent Analytics-based Model for Product Sales Optimisation in
Retail Enterprises.

I declare that the above dissertation is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

SIGNATURE

DATE

DEDICATION

This dissertation is dedicated to Ophellia Kimbini.

ACKNOWLEDGEMENTS

First and foremost, I want to give special thanks to the Almighty God for the wisdom, good health, skills and strength to conduct the research. I am also indebted to my supervisor, Professor Isaac O. Osunmakinde for the ideas, guidance, encouragement and support throughout the writing of this dissertation, as well as during the work on the research articles that we published. I will always be grateful for his supervision.

I owe a debt of gratitude to my mom for her support and prayers. Without her prayers I could not have achieved anything. I also extend my gratitude to my future wife Ophellia Kimbini for her moral support and prayers for the research. I am so grateful for all that she has done for me.

Last but not least I would like to appreciate the following people: Oluwapelumi Giwa for helping me with ideas and the programming part of my research model, aunt Iyabo Adeyemi and Maike Serapins for always reading my articles and correcting language technicalities, Deacon Tinashe Matyatya and Anesu Ruswa, for the ideas and mathematical solutions, Innocent Tawanda Mpofu (Soja) for your support, Kabelo, for his assistance with drawings. Relebogile Monageng for proofreading my work, Progress Sibanda (vatete), for the thorough work on the dissertation, VUT SDAM members, for their prayers, my sister, Sharon Sidume and the rest of the family members for their moral support.

TABLE OF CONTENTS

DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
LIST OF ACRONYMS	x
DEFINITION OF TERMS	xi
DECLARATION OF PUBLICATIONS RESULTING FROM THIS STUDY	xii
ABSTRACT	xiii
CHAPTER 1	1
INTRODUCTION	1
1.1 BACKGROUND	1
1.2 PROBLEM STATEMENT	4
1.3 RESEARCH OBJECTIVES	5
1.4 RESEARCH QUESTIONS	6
1.5 RESEARCH DESIGN	6
1.5.1 Implementation of the ARANN Model in Centralised Retail Enterprises	6
1.5.2 Implementation of the ARANN Model in Distributed Retail Enterprises	9
1.6 CONTRIBUTIONS OF THE RESEARCH	11
1.7 RESEARCH SCOPE AND LIMITATIONS	12
1.8 RESEARCH ETHICS	12
1.9 DISSERTATION OUTLINE	12
CHAPTER 2	14
LITERATURE REVIEW AND THEORETICAL BACKGROUND	14
2.1 INTRODUCTION TO SALES OPTIMISATION	14
2.1.1 Emergence of Market Basket Analysis	14
2.1.2 Benefits of Market Basket Analysis Methods to Retail Enterprises	14
2.2 MARKET BASKET ANALYSIS METHODS	15
2.2.1 Product Combination Analysis.....	15
2.2.2 Product Frequency Analysis.....	15

2.2.3 Next-product Sales Prediction Analysis	16
2.2.4 Product Quantity Analysis.....	16
2.3 CHALLENGES OF SALES OPTIMISATION	17
2.4 ASSOCIATION RULES	17
2.4.1 Introduction to Association Rules	17
2.4.2 Problem Description.....	18
2.4.3 Example of AR.....	20
2.4.4 Classification of AR	22
2.4.5 Areas of Application	23
2.4.6 Apriori Algorithm.....	25
2.5 ARTIFICIAL NEURAL NETWORK	29
2.5.1 Introduction to Artificial Neural Networks	29
2.5.2 Mathematical Basis	29
2.5.3 ANN Architecture	31
2.5.4 Areas of Application	31
2.6 DATA INTEGRATION	34
2.6.1 Concept of Data Integration	34
2.6.2 Data Integration Techniques	35
2.7 CHAPTER SUMMARY	36
CHAPTER 3	38
RESEARCH METHODOLOGY AND PROPOSED SYSTEM MODEL	38
3.1 INTRODUCTION	38
3.2 DATA COLLECTION AND PREPARATION	38
3.2.1 Data Collection in Centralised Retail Enterprises	38
3.2.2 Importance of Data Integration into a Centralised Retail Enterprise	40
3.2.3 Data Collection in Distributed Retail Enterprises	40
3.2.4 Data Preparation Stages.....	41
3.3 PROPOSED ARANN MODEL FOR RETAIL ENTERPRISES	44
3.4 ARANN SYSTEM MODEL FOR CENTRALISED ENTERPRISES	45
3.4.1 Pseudo-code of the ARANN Model for Centralised Analytics	47
3.4.2 Mathematical Description of the ARANN Model for Centralised Analytics	48
3.4.3 Scenario: Arrangement of Products on Shelves for Centralised Retail Analytics	49
3.5 ARANN SYSTEM MODEL FOR DISTRIBUTED ENTERPRISES	52
3.5.1 Pseudo-code of ARANN to Distributed Analytics.....	54

3.5.2 Mathematical Description of ARANN to Distributed Analytics	55
3.5.3 Scenario: Arrangement of Products on Shelves for Distributed Retail Branches	56
3.6 EVALUATION MECHANISM	60
3.7 CHAPTER SUMMARY	61
CHAPTER 4	62
EXPERIMENTAL EVALUATIONS AND RESULTS	62
4.1 OVERALL EXPERIMENTAL SETUP	62
4.2 EXPERIMENTAL EVALUATIONS FOR CENTRALISED ANALYTICS	62
4.2.1 ARANN Experimental Setup in Centralised Analytics	62
4.2.2 Experiment 1: Observations of ARANN with Varying Activations in Centralised Analytics	63
4.3 EXPERIMENTAL EVALUATIONS FOR DISTRIBUTED ANALYTICS.....	66
4.3.1 ARANN Experimental Setup in Distributed Analytics.....	66
4.3.2 Experiment 2: Observations of ARANN with Varying Activations in Distributed Analytics	67
4.4 PERFORMANCE EVALUATION OF DISTRIBUTED AND CENTRALISED ANALYTICS.....	72
4.4.1 Experiment 3: Comparison of ARANN in Terms of Memory and Time Usages.....	72
4.4.2 Experiment 4: Benchmarking ARANN with Classical Models.....	73
4.5 CHAPTER SUMMARY	75
CHAPTER 5	77
CONCLUDING REMARKS	77
5.1 CONCLUSION	77
5.2 MANAGERIAL IMPLICATIONS	78
5.3 FUTURE WORK.....	79
REFERENCES.....	80

LIST OF FIGURES

Figure 1.1: Impact of Current Sales Optimisation Models on Retail Enterprises.....	2
Figure 1.2: Problems Caused by Poor Data Quality. Adapted from [12].....	3
Figure 1.3: Data Integration Subsystem	7
Figure 1.4: Data Preparation Subsystem.....	8
Figure 1.5: The ARANN Model for Centralised Retail Enterprises.....	9
Figure 1.6: Data Cleaning and Formatting Layer.....	10
Figure 1.7: Snapshot of ARANN for Distributed Retail Enterprises.....	11
Figure 2.1: Functional Structure of a Two-input Neuron.....	30
Figure 2.2: ANN Architecture.....	31
Figure 2.3: Data Integration Techniques. Adapted from [1].....	36
Figure 3.1: Stages and Outputs of the KDD Process. Adapted from [2].....	43
Figure 3.2: Intelligent Model for Retail Enterprises.....	45
Figure 3.3: Proposed ARANN Framework for Centralised Analytics.....	46
Figure 3.4: The ARANN Model for Centralised Retail Branches.....	49
Figure 3.5. Proposed Intelligent Analytics-based Framework for Distributed Analytics.....	53
Figure 3.6: Intelligent Analytics-based Model for Four Branches.....	56
Figure 4.1: ARANN Rules on Real Life Data in Centralised Analytics.....	62
Figure 4.2: ARANN Rules on Public Data in Centralised Analytics.....	63
Figure 4.3: ARANN Rules on Real-life Data for Branch A.....	66
Figure 4.4: ARANN Rules on Real-life Data for Branch B.....	66
Figure 4.5: ARANN Rules on Public Dataset for Branch C.....	67
Figure 4.6: ARANN Rules on Public Dataset for Branch D.....	67
Figure 4.7: Comparison of the Performance of ARANN in Retail Enterprises.....	72

LIST OF TABLES

Table 2.1: Example of Transactional Data in a Database.....	20
Table 2.2: The Apriori Algorithm.....	25
Table 2.3: Transactional Data.....	26
Table 3.1: Real-life Sample Data.....	39
Table 3.2: Public Sample Data. Adapted from [3].....	39
Table 3.3: Sample of Real-life Data for Branch 1.....	40
Table 3.4: Sample of Real-life Data for Branch 2.....	41
Table 3.5. Sample of Public Data. Adapted from [3].....	41
Table 3.6: Pseudo-code for ARANN Model in Centralised Analytics.....	47
Table 3.7: Market Basket Transactional Data for All Centralised Branches.....	50
Table 3.8: Pseudo-code for ARANN Model in Distributed Analytics.....	54
Table 3.9: Market Basket Transactional Data for Branch 3 of a Retail Enterprise.....	57
Table 3.10: Market Basket Transactional Data for Branch 4 of a Retail Enterprise.....	58
Table 3.11: Confusion Matrix. Adapted from [4].....	60
Table 4.1: Real-life ARANN Results for All Centralised Branches.....	63
Table 4.2: Public ARANN Results for All Centralised Branches.....	65
Table 4.3: Real-life ARANN Results for Branch 1 in Demographic Group A.....	68
Table 4.4: Real-life ARANN Results for Branch 2 in Demographic Group B.....	69
Table 4.5: Public Data ARANN Results for Branch 3 in Demographic Group C.....	70
Table 4.6: Public Data ARANN Results for Branch 4 in Demographic Group A.....	71
Table 4.7: Quantitative Comparison of Three Models on Retail Datasets.....	73
Table 4.8: Quantitative Evaluations of the Cooperative Model in Distributed Branches...	75

LIST OF ACRONYMS

ANN	Artificial Neural Networks
AR	Association Rules
ARANN	Cooperative Model
BI	Business Intelligence
DoB	Degree of Belief
FN	False Negatives
FP	False Positives
LAN	Local Area Network
MBA	Market Basket Analysis
PC	Personal Computer
TN	True Negatives
TOR	Time of Response
TP	True Positives
WAN	Wide Area Network
ECM	Enterprise Content Management
ETL	Extract, Transform & Load

DEFINITION OF TERMS

Business analytics can be defined as groups of methodologies, organisational techniques and tools that are used collectively to gain information, analyse it and predict outcomes of problem solutions.

A **distributed retail enterprise** is a retail enterprise that issues the decision rights to the branches or groups nearest to the data collection.

Centralised retail enterprises are retail enterprises where the decision rights of the branches are concentrated in a single authority.

Market basket analysis (MBA) is a data mining technique that discovers the customers' purchasing patterns by extracting associations or co-occurrences from a retail enterprise transactional data.

Association rules (AR) mining is an unsupervised data mining method used to find interesting associations in large sets of data items.

Artificial neural networks (ANN) is a technique that simulates the behaviour of biological systems and is used to discover complex patterns and relationships.

ARANN is a cooperative intelligent model that combines AR and ANN models.

Data preparation involves all the activities performed on the generated raw data for knowledge discovery so that the data will be ready for modelling in data-mining models.

Data integration is the process of combining data stored at different sources and provides the organisational users with a single view of the data.

DECLARATION OF PUBLICATIONS RESULTING FROM THIS STUDY

The following recent articles have been published from the research work.

Accredited Journal Publications

- C. Matobobo and I.O. Osunmakinde, (2016), **Analytical Business Model for Sustainable Distributed Retail Enterprises in a Competitive Market**, Sustainability Journal, Sustainable Business Models, Vol. 8, no 2, ISSN: 2071-1050, (*ISI journal*)

Accredited Conference Publications

- C. Matobobo and I.O. Osunmakinde, (2015), **The Science of Market Basket Analysis: A Cooperative Business Intelligent Model for Distributed Retail Enterprises**, In Proceedings of the IEEE Pan African International Conference on Information Science, Computing and Telecommunications (PACT) Kampala, Uganda, pp. 32-37, ISBN: 978-1-4673-9230-3. (*DoE accredited*)
- C. Matobobo and I.O. Osunmakinde, (2014), **Comparative Market Basket Analysis for Centralised Retail Enterprises Using the ANN and AR Models**, In Proceedings of the 26th SAIMS Conference, Contemporary Management in Theory and Practice, 14 – 17 September, Vaal River in Vanderbijlpark, Gauteng, pp. 32-43, ISBN: 978-0-86970-784-5 (*DoE accredited*)

ABSTRACT

A retail enterprise is a business organisation that sells goods or services directly to consumers for personal use. Retail enterprises such as supermarkets enable customers to go around the shop picking items from the shelves and placing them into their baskets. The basket of each customer is captured into transactional systems. In this research study, retail enterprises were classified into two main categories: centralised and distributed retail enterprises. A distributed retail enterprise is one that issues the decision rights to the branches or groups nearest to the data collection, while in centralised retail enterprises the decision rights of the branches are concentrated in a single authority. It is difficult for retail enterprises to ascertain customer preferences by merely observing transactions. This has led to quantifiable losses. Although some enterprises implemented classical business models to address these challenging issues, they still lacked analytics-based marketing programs to gain competitive advantage. This research study develops an intelligent analytics-based (ARANN) model for both distributed and centralised retail enterprises in the cross-demographics of a developing country. The ARANN model is built on association rules (AR), complemented by artificial neural networks (ANN) to strengthen the results of these two individual models. The ARANN model was tested using real-life and publicly available transactional datasets for the generation of product arrangement sets. In centralised retail enterprises, the data from different branches was integrated and pre-processed to remove data impurities. The cleaned data was then fed into the ARANN model. On the other hand, in distributed retail enterprises data was collected branch per branch and cleaned. The cleaned data was fed into the ARANN model. According to experimental analytics, the ARANN model can generate improved product arrangement sets, thereby improving the confidence of retail enterprise decision-makers in competitive environments. It was also observed that the ARANN model performed faster in distributed than in centralised retail enterprises. This research is beneficial for sustainable businesses and consideration of the results is therefore recommended to retail enterprises.

Title of thesis:

Development of an Intelligent Analytics-based Model for Product Sales Optimisation in Retail Enterprises.

Key terms:

Analytics; Product sales optimisation; Retail enterprises; Model; Association rules; Artificial neural networks; Data mining; ARANN; Business intelligence; Management; Marketing.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

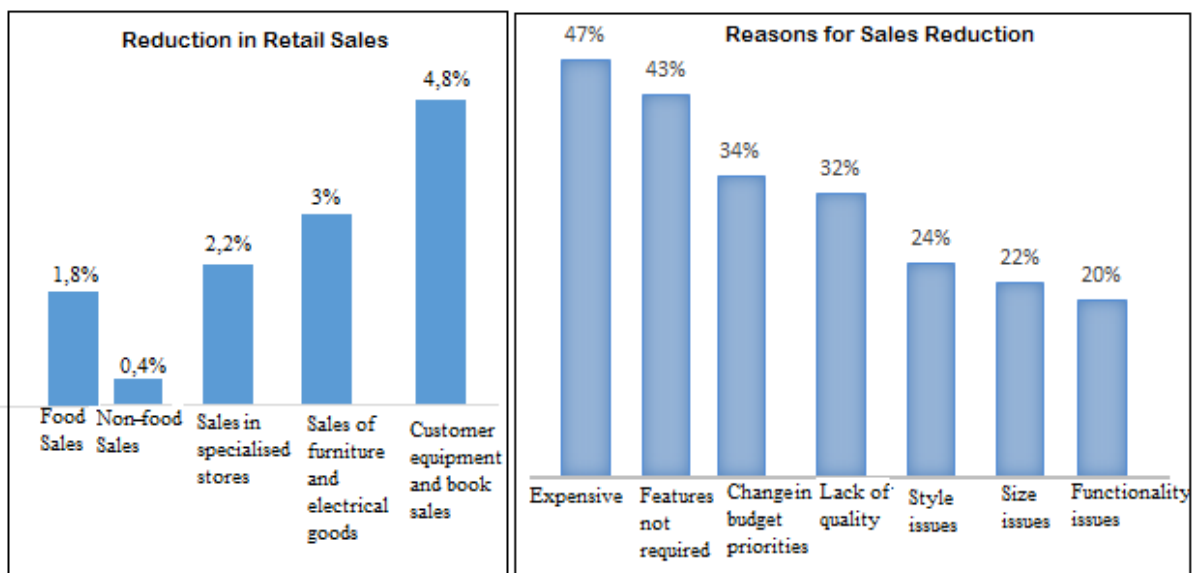
Retailing is growing fast in Africa and the goal is to sell as many products as possible in order to maximise profit levels. A retail enterprise is a business organisation that sells goods or services directly to consumers for personal use through channels such as departmental stores, hawkers and peddlers, stalls, supermarkets, mail order houses, hypermarkets and consumer cooperatives [5]. In Sub-Saharan Africa, retail enterprises face challenges owing to countries' economic struggles, such as credit crises and decreased country exports [6]. Global retail and consumer goods companies have their sights set on South Africa's middle class regardless of the country's own retailing challenges [7]. In South Africa, retailers face intensive competition from informal micro-enterprises such as spaza shops, tuck shops and kiosks shops [8]. This justifies the retail enterprises' attempts to find better ways of improving marketing strategies.

Retail enterprises (such as supermarkets, hypermarkets and department stores), enable consumers to go around the shop, picking the items of their choice from the shop shelves and placing the items into their baskets; the contents of each basket are then captured into transactional systems. The data collected from these transactional systems can be used for analysis purposes. Some retail enterprises apply business analytics to analyse the data generated by these systems. Business analytics can be defined as groups of methodologies, organisational techniques and tools that are used collectively to gain information, analyse it and predict outcomes of problem solutions [9]. The field of business analytics through the use of operational data generated from transactional systems has given decision-makers better insights [10]. These insights can help managers make better and informed decisions. The survival of retail enterprises depends on a good customer response and they should do their best to attract consumers to buy abundantly.

In retailing, retail enterprises can be classified into two broad categories depending on data administration practices. A distributed retail enterprise is one that issues the decision rights to the branches or groups nearest to the data collection [11]. Distributed retail enterprises give each branch management an opportunity to make decisions for each particular branch,

depending on results generated from the data. On the other hand, in centralised retail enterprises the decision rights of the branches are concentrated in a single authority [11]. This might give uniformity in a centralised retail enterprise. Customers may know what to expect in each branch of a retail enterprise, thus improving their shopping experience.

Retail enterprises strive for survival in view of current challenging sales optimisation models. These challenging sales optimisation models affect product arrangements in retail enterprises, leading to a decline in sales levels [12], high research and marketing costs, a decline in market share, a wrong product target market and poor management decisions [13]. Figure 1.1 presents the quantitative impact of these challenging sales optimisation models in retail enterprises. Figure 1.1a shows sales decline for a retail enterprise in June 2013. The sales level of computer equipment and books declined drastically by 4.8%, while sales of non-food items had the lowest decline level of 0.4%. Figure 1.1b shows the causes of the reduction in sales level. The highest scoring reason for the reduction in sales was items being expensive (48%), followed by 41% of products in which the desired features were unavailable. The least common reason for a reduction in sales was lack of functionality (20%).



a) Reduction in retail enterprise sales. Adapted from [14] b) Reasons for sales reduction. Adapted from [15]

Figure 1.1: Impact of Current Sales Optimisation Models on Retail Enterprises

Data quality problems also affect the quality of decisions made by managers on different levels of a retail enterprise [13]. Poor data has caused problems in both traditional and e-

business, as shown in Figure 1.2. In both companies, extra cost to prepare reconciliations was seen as the major problem caused by inadequate data, with an impact of 58% and 57% respectively. Inability to deliver an order or a loss of sales was also a poor data quality challenge, with a higher impact in e-business (33%) than in traditional (24%) companies. The lowest scored problem caused by poor data was failure to meet a significant contractual requirement.

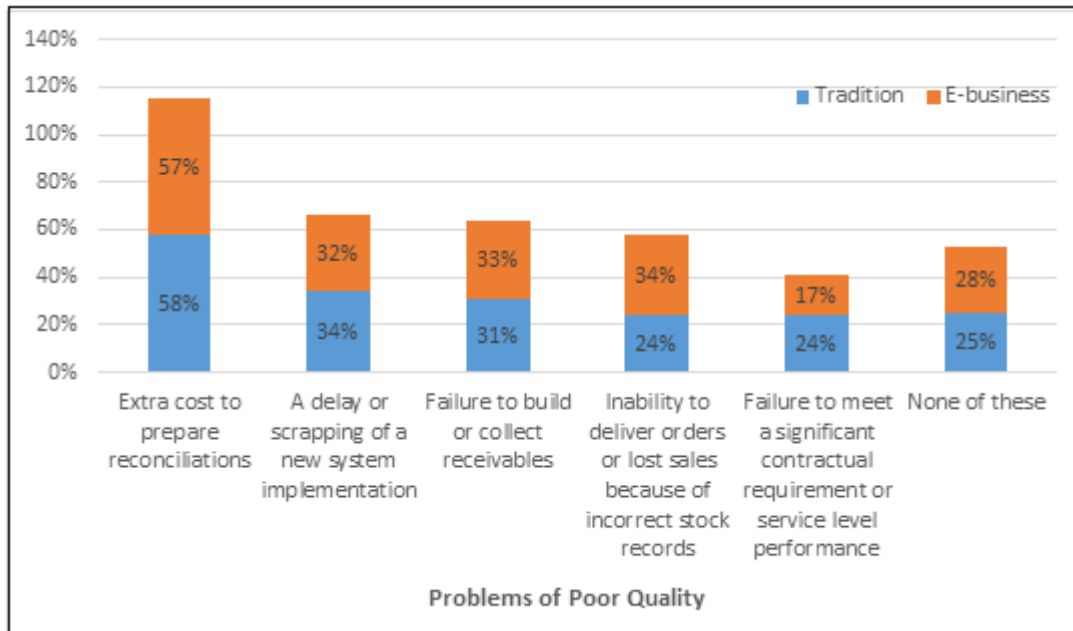


Figure 1.2: Problems Caused by Poor Data Quality. Adapted from [12]

A data consolidation project was undertaken in South Africa by Altron to organise and deliver high-quality data successfully to its executives on their Apple iPads. This quality data set the stage for iPad delivery [16]. In [17], they implemented easy-to-use desktop and server analytics software for the development of several business units. In this project, data from their three departments was integrated and the company had a single version of the truth, which enabled executives to make sound decisions. For example, if back taxes are owed, the system will not grant approval for work to commence until the taxes have been paid. In a survey conducted in [18], it was seen that the cleaning, structuring, reformatting and mashing of data was the lifeblood of any organisation and a source of competitive intelligence. In research done by [19], they suggested that companies should move from descriptive analytics to predictive analytics until they finally get to prescriptive analytics. Although sales optimisation models have been developed in the context of developing countries such as South Africa, several challenges still exist, including:

- Failure to leverage business intelligence (BI) to become ‘matchmakers’
- Lack of analytics-based marketing programmes
- Failure in organisations to assemble data from more enterprise resource planning systems into a single environment so that users can see appropriate transactional data, as well as strategic level data
- Lack of business-driven analytic strategies and failure to test the most effective BI analytics for solving theoretical business problems.

The current study develops a cooperative model that can be used in retail enterprises to implement the best arrangement of shelf products at each retail branch to improve the weaknesses highlighted in Figure 1.1 and Figure 1.2. The cooperative model is built on a machine learning technique which combines the Association Rules (AR) and the Artificial Neural Networks (ANN) models to strengthen the results of these two individual models. To the best of the researcher’s knowledge there is no system or research that has addressed these issues together using the proposed approach. This then justifies the current study.

1.2 PROBLEM STATEMENT

A small retail enterprise business builds relationships with its customers by noticing their needs, remembering their preferences and learning from past interactions how to serve them in the future [20]. This is only realistic in small shops. Large organisations, such as supermarkets, have a large customer base and more products on shelves, which makes it difficult to observe associations among more than a thousand products in a shop. It is difficult for retail enterprises to arrange products on their shelves in a way that would encourage customers to buy more products than planned by merely observing the customers’ buying patterns. Sometimes the way the products are arranged on shop shelves discourages customers from buying more than planned. In this situation a question can be asked: “How can a cooperative model be developed to improve product arrangement on shop shelves for sales optimisation using improved data quality?” According to [21], the authors identified product placement as one variable that affects consumer buying behaviour.

It is a challenge to develop and run a model using unprepared data as there are very high chances of getting model prediction results with low accuracy rate [22]. Unprepared data is data with missing values, wrong values, duplicating columns and/or rows. It can be a

greater risk for an organisation to make decisions based on unreliable model prediction results generated from unprepared data. Data pre-processing is therefore necessary in retail enterprises as it prepares the transactional data so that the miner can produce better and reliable model prediction results that can be used as the base for decision making [22].

Also it is important to test the performance of the model using varying data sets so that the test results can be evaluated based on the matches and deviations [23]. Transactional data generated from the retail enterprises might not be in the format accepted by the model in use, therefore, there is need to format the data. This in turn might leave the data with high chances of becoming dirty and having different formats. To get around this problem, a question can be asked “What would be the performance of the cooperative model in centralised and distributed retail enterprise environments using real-life and publicly available datasets?”

Furthermore, it is crucial to validate new models with popular models in order to ascertain eventual difference. The AR model is the most common method in the market basket analysis (MBA) environment. The technique has the following drawbacks: discovery of non-interesting rules, a massive number of discovered rules and low algorithm performance. There is a challenge of having a single technique that can overcome the drawbacks of the AR model. In [24], the author suggested the creation of intelligent hybrid systems that can be used to solve complex problems that cannot be solved by a single intelligent technique.

1.3 RESEARCH OBJECTIVES

This research seeks to achieve the following objectives:

- To develop a cooperative intelligent analytics-based model to improve product arrangement on shop shelves for sales maximisation
- To develop a data preparatory subsystem for cleaning transactional data for the cooperative model.
 - To test the performance of the cooperative model in centralised and distributed retail enterprise environments using real-life and publicly available datasets.
 - To validate the cooperative model with popular models in order to make an evidential difference.

1.4 RESEARCH QUESTIONS

In order to achieve the specified research objectives, the following research questions should be considered:

- How can a cooperative model be developed to improve product arrangement on shop shelves for sales optimisation using improved data quality?
- How can pre-processed data work effectively in the cooperative model?
- What would be the performance of the cooperative model in centralised and distributed retail enterprise environments using real-life and publicly available datasets?
- How can the cooperative model be validated with popular models for making an evidential difference?

1.5 RESEARCH DESIGN

1.5.1 Implementation of the ARANN Model in Centralised Retail Enterprises

In centralised retail enterprises, the proposed model is composed of three subsystems. These subsystems address the research questions for this research. The proposed model is built on the following subsystems:

- Data integration
- Data preparation
- Development of ARANN model for sales optimisation through product arrangement.

1.5.1.1 Data Integration Subsystem

The data integration subsystem investigates the data integration technique most suitable on a personal computer (PC), local area network (LAN) and wide area network (WAN). In Figure 1.3 transactional data is generated from point of sale systems. This data can be from different PCs, LANs or WANs. The data is integrated using the following techniques:

- Data propagation
- Data federation
- Data consolidation.

The data from different sources is extracted, transformed and loaded into a data warehouse. The purpose of transformation is to improve the quality of the data extracted from these different sources. To improve the quality of data means to remove data impurities. Data impurities can be any unwanted data values, columns or rows. Finally the pre-processed data is loaded into the tables of the data warehouse to make it available for decision support applications. The pre-processed data is then stored in a data warehouse. A warehouse is an integrated data repository put into a form that can easily be understood, interpreted and analysed by decision-makers in order to make informed decisions [25].

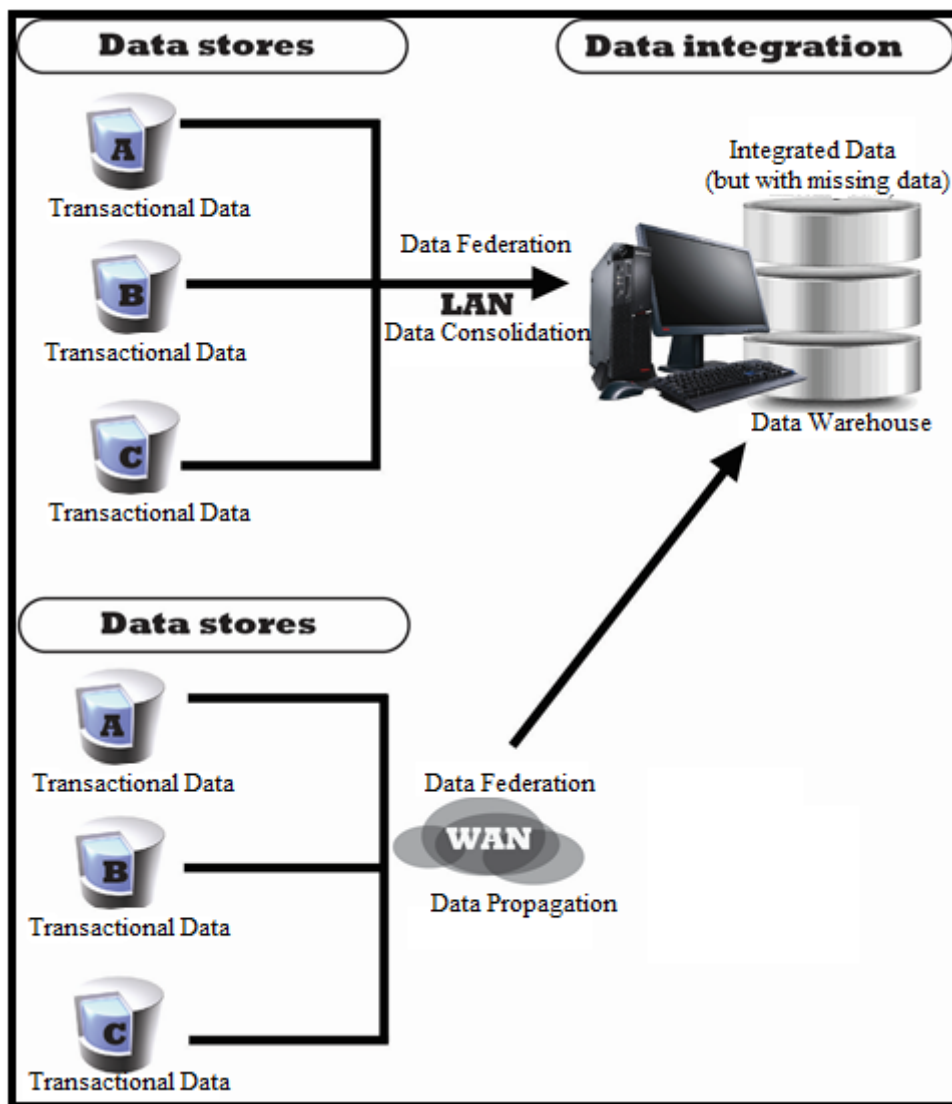


Figure 1.3: Data Integration Subsystem

1.5.1.2 Data Preparation Subsystem

The objective is to develop a data preparatory subsystem that can be used to clean data using the ARANN model. The integrated data from the data integration subsystem is cleaned in Figure 1.4. Data impurities are removed from this subsystem. The processed data from this subsystem is input into the ARANN model for the generation of product arrangement sets. Product arrangement sets are combinations of products with strong associations, used to determine the best ways of arranging products on shop shelves.

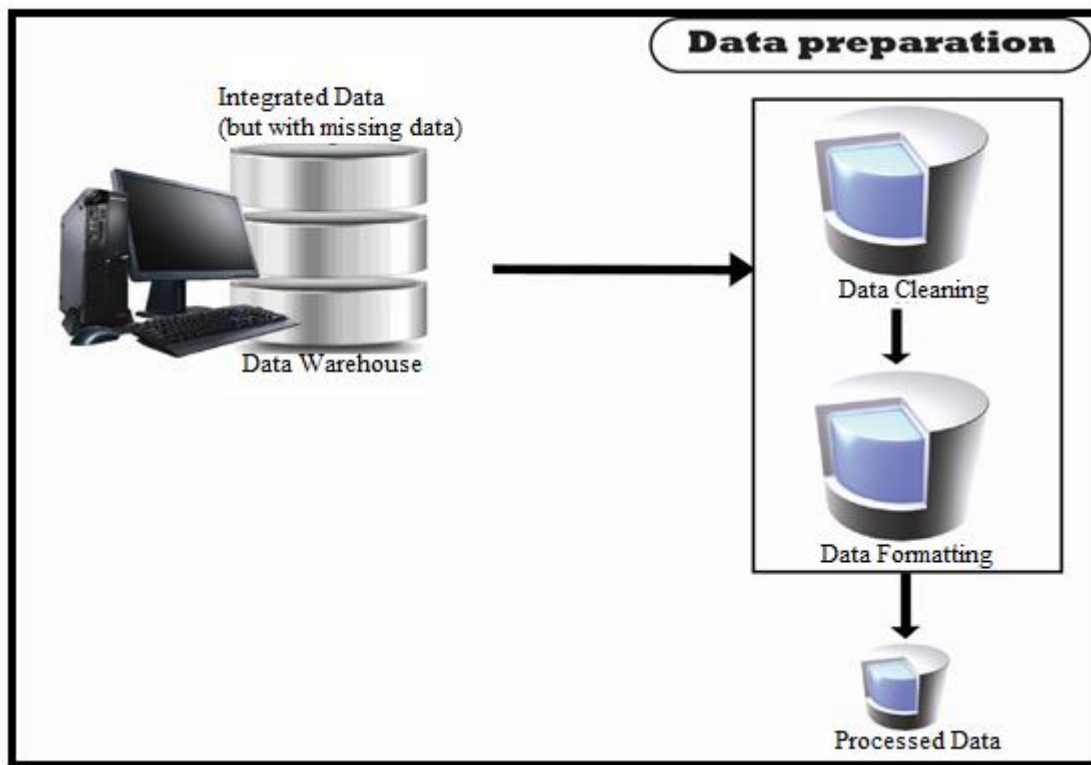


Figure 1.4: Data Preparation Subsystem

1.5.1.3 The ARANN Model

The prepared data from the data preparation subsystem is taken as the input into the ARANN model. The ARANN model is composed of the AR and ANN models, as shown in Figure 1.5. The data from the data preparation subsystem is entered into the AR model first for the generation of support ($sup()$) and confidence ($con()$) values. These generated values are passed into the ANN model to output the resultant degree of belief (DoB) values. The DoB values determine the product arrangement sets to be adopted in all branches of a retail enterprise. The products are uniformly arranged on shelves of all branches of a centralised retail enterprise according to the accepted product arrangement sets.

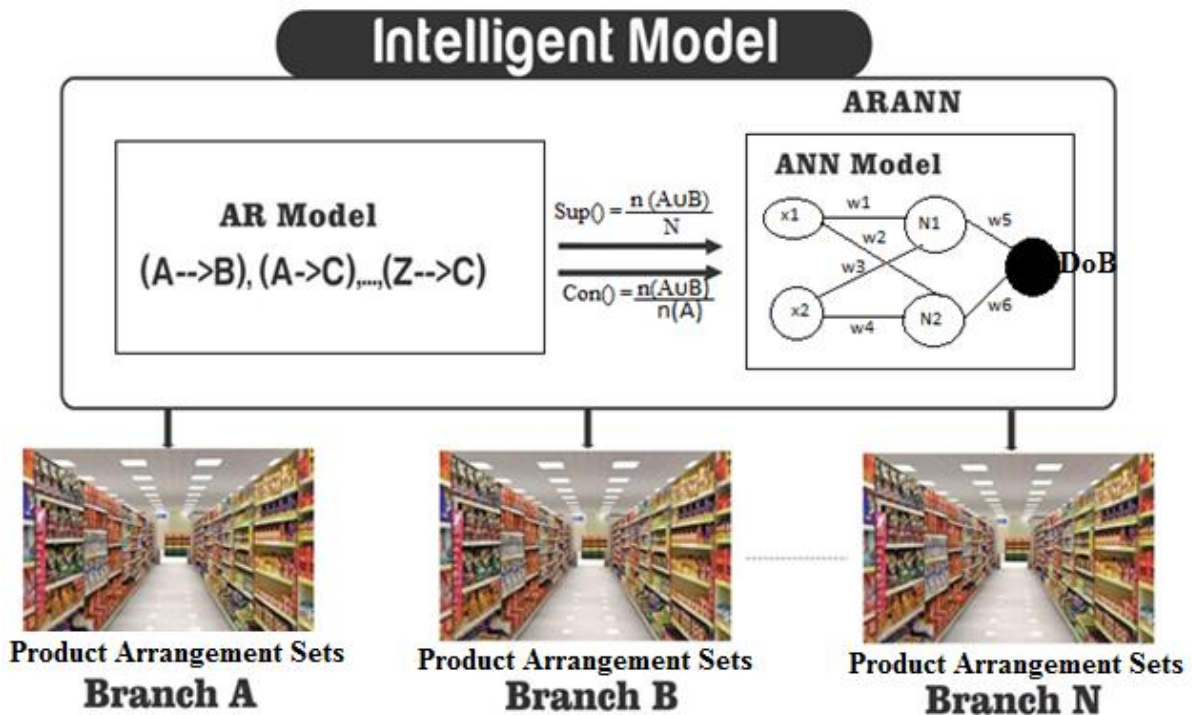


Figure 1.5: A Snapshot of the ARANN Model for Centralised Retail Enterprises

1.5.2 Implementation of the ARANN Model in Distributed Retail Enterprises

The ARANN model for distributed retail enterprises has three layers, namely data cleaning and formatting, intelligent subsystem and distributed product shops.

1.5.2.1 Data Cleaning and Formatting Layer

This layer is found at the bottom of the ARANN model. In this layer, data is collected from separate transactional systems per branch, as shown in Figure 1.6. The data is cleaned and formatted to the appropriate file type accepted by the ARANN model. The processed data is then passed to the middle layer of the Intelligent model.

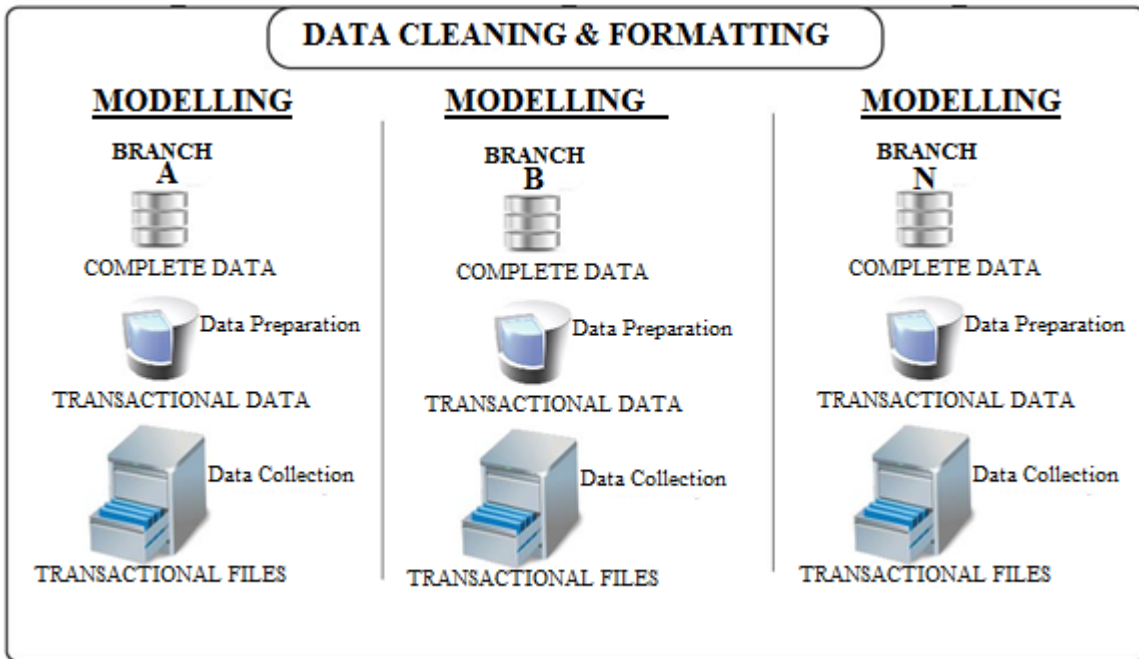


Figure 1.6: Data Cleaning and Formatting layer

1.5.2.2 Intelligent Subsystem

In this layer, processed data from different branches is input into the ARANN model, as shown in Figure 1.7. The processed data from the bottom layer is passed onto the AR model and it outputs confidence and support values. These values are then passed on to the ANN model as inputs, in order to get the DoB. The intelligent subsystem generates product arrangement sets for each particular branch and the accepted product arrangement sets are sent to the distributed product shops layer.

1.5.2.3 Distributed Product Shops

This layer involves the application of the generated product arrangement sets from the intelligent subsystem. The accepted product arrangement sets are applied branch per branch, as shown at the top layer of Figure 1.7.

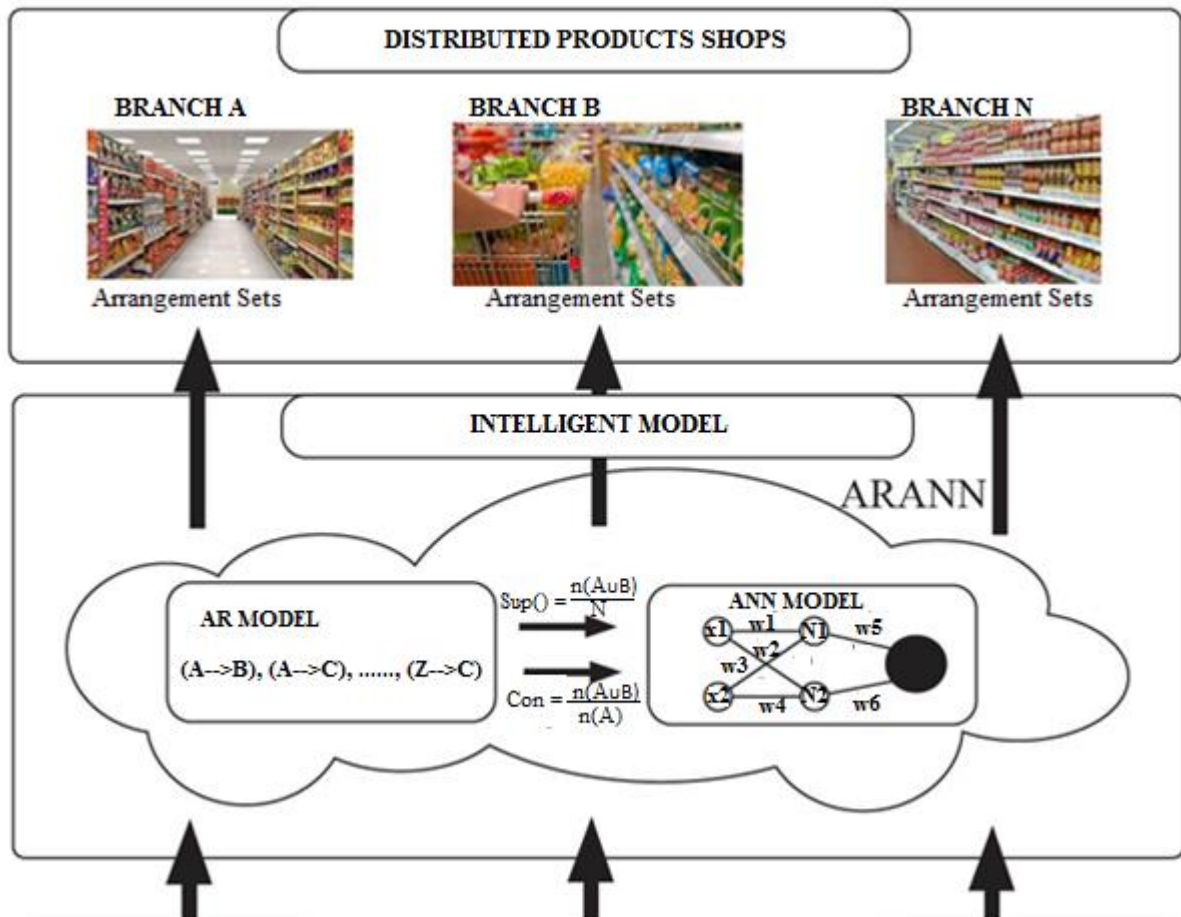


Figure 1.7: Snapshot of ARANN for Distributed Retail Enterprises

1.6 CONTRIBUTIONS OF THE RESEARCH

The major contributions of this study are as follows:

- Development of a newly proposed ARANN model that could intelligently assist retail enterprises' management to arrange products on the shop shelves to improve customer product purchases and maximise profits.
- Development of a data integration subsystem that provides a single view of enterprise data.
- Development of a data preparatory subsystem for cleaning transactional data.
- Provision of a reference knowledge guide (knowledge generation) to help professional managers understand product sales optimisation models and their application in retail enterprises.
- Logical demonstration of working scenarios of the ARANN model for simplified implementation and managerial practices.

1.7 RESEARCH SCOPE AND LIMITATIONS

The section sets the scope of the research. The research is carried out within the following research scope:

- The study only focuses on centralised and distributed retail enterprises in the context of a developing country.
- A prototype of this research project has been implemented as proof of the concept.

The study also has the following limitation:

- Implementation of a full-scale project requires more manpower and it is time-consuming.

1.8 RESEARCH ETHICS

Ethics are norms for conduct that distinguish between acceptable and unacceptable behaviour [26]. Research ethics is defined to be the ethics of the planning, conduct and reporting of research [27]. These policies are formulated by professional bodies, government agencies, and academic institutions to guide the conduct of research. The data was collected for academic purposes only and is treated as private and confidential. The data will not be given to any third parties. The research complies with the UNISA research ethics policy.

1.9 DISSERTATION OUTLINE

The dissertation is organised as follows:

Chapter 1 gives the introduction to the thesis. It introduces the retail enterprises' motivation and challenges. It covers the problem statement, research questions, research objectives, research design, research scope and limitations, research ethics and the contributions of the research.

Chapter 2 gives the background and the significance of the study. This includes a brief introduction of sales optimisation, MBA methods that can be used in retail enterprises, challenges of sales optimisation in retail enterprises, discussion on AR mining and ANN models.

Chapter 3 describes the research methodology and the proposed system models implemented in centralised and distributed retail enterprises. It also describes how

data was collected, prepared and integrated for the experiments and model evaluation mechanisms.

Chapter 4 presents experimental evaluations and results acquired from the experiments conducted. It presents sets generated in both types of retail enterprises and evaluates the ARANN against the classical models. A performance comparison of distributed and centralised retail analytics was also presented.

Chapter 5 gives the conclusion of the study. It also presents some managerial implications and some areas for future work.

CHAPTER 2

LITERATURE REVIEW AND THEORETICAL BACKGROUND

2.1 INTRODUCTION TO SALES OPTIMISATION

2.1.1 Emergence of Market Basket Analysis

MBA is a data mining technique that discovers customers' purchasing patterns by extracting associations or co-occurrences from a retail enterprise's transactional data [28]. The idea is to discover purchasing patterns from the transactional data generated from point-of-sale systems. The MBA technique was initially applied to supermarket transactional data sets and has expanded to industries selling multiple products, such as banks, catalogues, direct marketers and many more [29]. The technique studies customers' shopping baskets in order to discover their buying habits through recognising associations among various customer baskets [30]. The output of MBA consists of a series of product association sets, for example, if a customer buys product A, the chances are high that the customer will buy product B as well [29]. The aim of MBA is to give retail enterprise decision-makers associated product pairs with minimal human interaction [29].

2.1.2 Benefits of Market Basket Analysis Methods to Retail Enterprises

- MBA can be used to discover meaningful associations between the different products that are purchased by customers. These associations can help retail managers to develop appropriate marketing strategies by gaining insight into which products are frequently purchased together by customers [31].
- MBA can be used to determine the relationships within the data itself even without having some specified direction for search, i.e. it is best for undirected data mining.
- Another benefit of MBA is its computational simplicity. Data mining techniques commonly used to perform MBA computations are simpler than other techniques being implemented in retail enterprises, such as ANN.
- MBA can help decision-makers to make informed decisions about pricing, product placement, promotions and profitability [32].

- MBA helps retail enterprise management realise customer expectations and how to serve them best in future.
- MBA makes market segmentation easier. It makes marketers target the right segments of customers. This can even lead to a reduction in marketing costs, involving for example promotional materials and a wrong target market.
- MBA results in marketing managers understanding customers better, i.e. their purchasing habits and behaviour, in order to optimise marketing and sales operations.
- It enables discovery of significant associations and patterns that were not possible even with experienced marketing managers.

2.2 MARKET BASKET ANALYSIS METHODS

2.2.1 Product Combination Analysis

This is an analysis of the frequency of certain product combinations in different customer baskets, for example, a combination of bread and a drink in different customers' baskets [12]. It analyses how frequently items are purchased together by different customers, i.e. the association among products. Product combination analysis looks at the number of transactions performed against the frequency of a certain combination. A high product combination analysis rate might indicate strong product associations. This analysis can be used by retail enterprise management implementing marketing strategies, such as product combination sales and the arrangement of products on shelves. Product combination sales can help promote products that are not frequently purchased to be sold together with strongly associated products. An example might be if product A is not selling well and products B and C are selling well, then managers can create a combination sale of products A, B and C in order to improve the sales levels of product A. Implementation of product combination analysis results might improve the marketing strategies of a retail enterprise, reduce marketing costs and improve sales levels.

2.2.2 Product Frequency Analysis

Product frequency analysis is an analysis of how frequently an item is found in different customers' baskets [33]. An example might be to analyse how many customers purchased a certain item against the number of transactions performed. Products that frequently appear in many customers' baskets might suggest products selling well and products with

low frequency might have low stock turnover. Product frequency analysis reveals products with low, medium and high sales frequencies. This might provide retail enterprise decision-makers with a base for decision-making and a platform for choosing the best marketing strategies, such as product combination sales and also promoting products that are not selling well by giving discounts on products with low and medium sales frequencies. This product frequency analysis can be done on each branch of a retail enterprise in order to recognise the buying habits of each branch. These strategies may improve the marketing strategies that can be implemented within each branch of a retail enterprise.

2.2.3 Next-product Sales Prediction Analysis

This is an analysis of the next item that a customer is likely to buy at any given time [34]. This allows decision-makers to predict the next items to be purchased by different customers and might help to improve promotional or marketing strategies. Next-product sale prediction analysis may help retail management to implement the best strategies, such as arranging products that customers are likely to buy closer to the shop entrance so that they may see the products as they enter. Products can also be sold as a combination sale, i.e. offering products that customers are likely to purchase next with those unlikely to be purchased next. This type of analysis might help decision-makers target the right customers with the right products at the right time. This might reduce marketing costs, as the company will not be promoting the wrong products. The decisions can be made at each branch. Centralised decisions might not reflect the true customer buying habits of each branch and the results might mislead the managers.

2.2.4 Product Quantity Analysis

Product quantity analysis is an analysis of the number of items purchased per product during a certain customer transaction [32]. An analysis will be done to determine if there is a relationship between the product and the quantity purchased, for example, customers tend to buy two battery cells most of the time. This analysis might help management to implement the appropriate marketing strategies, such as packing or arranging a certain number of products together or have promotions based on the quantity of products. An example might be to promote a certain brand e.g. customers are encourage to buy 2 * 2 kg of Sunlight washing powder at a certain discount. This encourages customers to buy higher quantities of a certain product, thus increasing the sales of the enterprise.

2.3 CHALLENGES OF SALES OPTIMISATION

- Retail enterprises face sales optimisation challenges because of lack of analytics-based models that can be used to make marketing decisions. Without these models it is difficult to understand customer preferences and behaviours.
- Inadequacy of transactional data is also a sales optimisation challenge. If a retail enterprise has inadequate transactional data, it becomes difficult to understand customers. This also hinders the enterprise from best serving customers in future in order to improve sales.
- Lack of analytical skills also hinders the implementation of analytics-based marketing programmes that might improve marketing strategies in a retail enterprise. This might create a poor base for decision-making processes.
- Unavailability of BI tools in other retail enterprises. These tools are used for retrieving, analysing, transforming and reporting BI analysis results generated from the transactional data. It is a challenge for enterprise management to make intelligent decisions without these BI tools.

2.4 ASSOCIATION RULES

2.4.1 Introduction to Association Rules

AR mining is an unsupervised data mining method, which was first introduced in [35] to find interesting associations in large sets of data items [36]. It extracts interesting associations, correlations and frequent patterns from transactional data generated from point-of-sale machines [37]. It was originally derived from point-of-sale data that describes which products are purchased simultaneously. AR discovers interesting associations that are often used by businesses such as retail enterprises for decision-making purposes, for example to find out which products are frequently purchased simultaneously by different customers [20]. It is one of the most common and widely used techniques in data mining, aimed at finding interesting relations [38], [39] or correlations between large data items [40]. AR provides decision-makers of retail enterprises with marketing insights for cross-selling by providing information about product associations [41]. AR can be applied in supermarkets, banks, telecommunication systems, risk management, stock control, etc. The main advantage of AR is its capability of revealing interesting associations among

different variables. In research conducted by [42], it was shown that AR mining plays a pivotal role in the data mining environment.

2.4.2 Problem Description

The general problem of mining ARs can be stated as follows [43]: given a transactional that the presence of some items in the customer's basket will imply the occurrence of the same items in other customers' baskets. The aim of AR will be to find all rules that satisfy the minimum support and minimum confidence thresholds set. The problem of AR mining is defined in [35] as follows:

Let $I = \{i_1, i_2, \dots, i_n\}$ be a list of items.

Let $D = \{t_1, t_2, \dots, t_n\}$ be a set of transactions.

A rule is defined as an implication of the form:

$A \Rightarrow B$

where $A, B \subseteq I$ and $A \cap B = \Phi$

An AR rule is made of antecedent (if) and consequent (then) for example $\{A \Rightarrow B\}$. The antecedent is an item found in the data while the consequent is the item found in combination with the antecedent. There are measures that can be used to quantify the interestingness of a rule, such as support, confidence, lift, laplace, conviction, etc.

Support value

Support determines how frequently a rule is contained in a given dataset. It is defined as the fraction of transactions that contains $A \cup B$ to the total number of transactions in the database [44] and this can be expressed as shown in equation (2.1):

$$Support(A \Rightarrow B) = P(A \cup B) = \frac{n(A \cup B)}{N}. \quad (2.1)$$

where $n(A \cup B)$ is the number of A and B occurring simultaneous, N is the number of all transactions performed, and $P(A \cup B)$ is the probability of event A and B occurring at the same time. If support $(A \Rightarrow B)$ is greater than or equal to the minimum support threshold (min_sup) then it is a frequent item set. An item set is frequent if $support(A \Rightarrow B) \geq min_sup()$.

Confidence value

Confidence is the ratio of the number of transactions containing A and B to the number of transactions containing A, and can be further expressed as shown in equation (2.2):

$$\text{Confidence}(A \Rightarrow B) = P(B / A) = \frac{n(A \cup B)}{n(A)}. \quad (2.2)$$

where $P(B/A)$ is the conditional probability of B given A. It is the probability of event B occurring given the knowledge that an event A has already occurred. If confidence $(A \Rightarrow B)$ is greater than or equal to the minimum confidence (min_con) then one is confident of the rule generated.

Furthermore, rules that satisfy both the minimum support threshold (min_sup) and the minimum confidence threshold (min_con) are called strong AR. A rule is **strong** if support $(A \Rightarrow B) \geq \text{min_sup}$ AND confidence $(A \Rightarrow B) \geq \text{min_con}$.

Lift value

Lift is a measure that categorises the AR's performance in order to improve the response. It takes baseline frequency into account, which is not considered when measuring confidence [45]. It is represented by equation (2.3):

$$\text{Lift}(A \Rightarrow B) = \frac{P(A \cap B)}{P(A) * P(B)}. \quad (2.3)$$

Conviction value

Conviction is the ratio of the expected frequency of A without the occurrence of B and can be expressed as shown in equation (2.4):

$$\text{Conviction}(A \Rightarrow B) = \frac{1 - P(A \cap B)}{1 - P(B / A)}. \quad (2.4)$$

It can be used to overcome the disadvantage of the confidence and the lift in that it attempts to measure the degree of implication of a rule [45].

Leverage value

Leverage is a measure that counts the number of cases obtained from the co-occurrence of the antecedent and consequent of the rule from the expected value [45]. It can be represented using equation (2.5):

$$Leverage(A \Rightarrow B) = P(A \cap B) - P(A) * P(B) . \quad (2.5)$$

Jaccard value

Jaccard is used to measure the degree of overlap between the antecedent and consequent of the AR [45] and is represented using equation (2.6):

$$Jaccard(A \Rightarrow B) = \frac{P(A \cap B)}{P(A) + P(B) - P(A \cap B)} . \quad (2.6)$$

There are two standard measures of interestingness of an AR in data mining, i.e. the support and the confidence values [46]. Section 2.4.3 shows an example of the application of these two measures in an AR environment in order to determine the interestingness of the AR rule.

2.4.3 Example of AR

In this Section an example of AR with two measures of interestingness is provided. Table 2.1 depicts an example of a transactional database.

Table 2.1: Example of Transactional Data in a Database

TransactionID	Items in the Basket of Customers
600	Bread, Sugar, Salt
601	Bread, Ice cream, Sugar, Salt
602	Bread, Lotion, Soup, Sugar, Ice cream

In Table 2.1 each transactionID represents a transaction performed by a customer. From Table 2.1 the following are the examples of sets that can be derived from the transactional data: {Bread => Sugar}, {Bread => Salt}, {Sugar => Salt}. The two measures of interestingness can be applied to the sets to determine if they can be accepted as strong associations or rejected as weak associations.

{Bread => Sugar}

$$\text{Support} = \frac{3}{3} = 100\%$$

$$\text{Confidence} = \frac{3}{3} = 100\%$$

Support of 100% means that 100% of all transactions performed in Table 2.1 show that bread and sugar were purchased together. There is 100% confidence that a customer who purchased bread also bought sugar.

{Sugar => Salt}

$$\text{Support} = \frac{2}{3} = 66.67\%$$

$$\text{Confidence} = \frac{2}{3} = 66.67\%$$

It can be seen that the product combination set {Sugar => Salt}, has support of 66.7%, implying that 66.67% of all transactions performed in Table 2.1 show that sugar and salt were purchased together. The second criterion also gives 66.67% confidence that a customer who purchased sugar also bought salt.

{Lotion => Soup}

$$\text{Support} = \frac{1}{3} = 33.33\%$$

$$\text{Confidence} = \frac{1}{1} = 100\%$$

Support of 33.33% means that 33.33% of all transactions performed in Table 2.1 show that lotion and soup were purchased together. This reflects 100% confidence that a customer who purchased lotion also bought soup.

Assuming that the minimum support is 50% and the minimum confidence is 60%, the following sets are accepted as strong: {Bread => Sugar} and {Sugar => Salt}, while {Lotion => Soup} is rejected.

2.4.4 Classification of AR

According to [31], the authors classified AR according to the following criteria:

A) Type of Values Handled

- Boolean AR: It is a Boolean AR if the rule involves associations with the presence or absence of items.
- Quantitative AR: It is a quantitative AR if the rule describes associations between quantitative items.

B) Number of Data Dimensions Involved in the Rule

- Single-dimension AR: It is a single-dimension AR if the items reference only one dimension, for example:

buys (Customer, "Sugar") => buys(Customers, "Bread")

The rule is a single-dimensional AR because it references only one dimension, which is 'buys'.

- Multidimensional AR: It is multidimensional AR if the items reference two or more dimensions, for example:

Age(Employee, "25...30") \wedge Salary(Employee, "15k...35k") => Buys(Employee, "Car")

The rule is a three-dimensional AR because it has three dimensions: 'age', 'salary' and 'buys'.

C) Kind of Rules to be Mined

This category can generate various kinds of rules and other interesting relationships, such as AR, strong gradient relationships and correlation rules. ARs are the most popular form of rules generated from frequent pattern analysis, for example:

{Laptop, Operating System Installation DVD} => {Office Installation DVD}

This indicates that if a customer purchases a laptop and an operating system there are high chances of purchasing an office installation disk.

D) Kinds of Patterns to be Mined

Data sets can be used to mine many different kinds of frequent patterns, such as frequent item-set mining, sequential pattern mining and structured pattern mining, for example:

(Colgate (Colgate, Toothbrush, Soap) (Colgate, Toothbrush) Drink (Colgate, Bread))

Colgate and toothbrush appears more than once in different baskets.

E) Levels of Abstraction Involved

Rules can be mined by some methods at different levels of abstraction, for example:

buys(Customer, "Laptop") => buys(Customer, "Windows installation DVD")

buys(Customer, "Netbook") => buys(Customer, "Windows installation DVD")

These items purchased are referenced at different levels of abstraction, e.g. a laptop is a higher abstraction of netbook.

2.4.5 Areas of Application

In [35] the AR algorithm was applied to a large database of customer transactions from a larger retailing company to test the effectiveness of the algorithm, which exhibited excellent performance. In [47], AR mining and fuzzy clustering were incorporated to assess the relationship between customer clusters and their preferences for products. Through this combination of techniques, they managed to categorise customer groups and the corresponding clusters in which they belonged. In research conducted by [48], they highlighted the importance of AR mining application in prediction of product sales' trends and customer behaviour. Other significant papers on ARs in retail enterprises are [49], [50], [51], [52], [41], [53] and [54].

AR was applied in [12] to a sports company with an issue regarding the arrangement of sports items in accordance with customer purchasing patterns. The retail company had no computerised mechanism for providing the best item arrangement. The study was performed to identify purchasing patterns that could be adopted by the retail enterprise. They analysed historical data to identify the associated patterns from transactional data. From the study, they found relationships between sports items purchased and the best ways of arranging items, either side by side or in the same retail area so that the items would

frequently be purchased together to yield high sales. In this study, AR was used for mining relationships between items purchased.

AR was applied in [55] to medical data containing combinations of categorical and numerical attributes to discover useful rules and from this experiment, useful and concise ARs were discovered for prediction purposes. In [56] the ARs were also used to predict the level of contraction in four arteries and risk factors. The experiment predicted accurate profiles of patients with localised heart problems, specific risk factors and the level of disease in one artery. Other significant papers on the application of ARs on medical data are as follows: [57] and [58].

In [59], the authors implemented a system for the discovery of ARs in web log usage data as an object-oriented application and they discovered excellent ARs in the data. They put forward “interestingness measures” as future work. A discussion on the goal of web association mining was presented by [60] and the authors highlighted that ARs can be revealed from web data. The following articles: [61], [62], [63], [64] and [65], also reported on research on the application of ARs to web data.

Other application areas of ARs are analysing accident data [66], identifying patterns of safety-related incidents in a steel plant [67], quality improvement of the production process [68], resource allocation in business process management [69], processing educational data [70], correlation analysis in telecommunication networks [71], [72], pattern mining in tourist attraction visits [73], processing accounting data [74], credit card fraud detection [75] and many more.

The AR model has been combined with other methods to improve its performance. In an article presented in [76], the authors combined AR and clustering methods in order to find links between rare attributes in a large data set. They applied clustering methods to reduce the AR errors and impractical number of rules. Variable clustering was used to reduce impractical rules. Combining AR with other methods might strengthen the result. In this research the AR model is combined with ANN so as to improve its performance in retail enterprises.

From the study conducted by [77], it was observed that AR is effective in revealing associations though it does not take into account special interests. A comprehensive survey was conducted by [78] on ARs on quantitative data in data mining. They examined it on

different parameters and concluded that direct application of ARs might produce a huge number of redundant rules. This is also supported in the article by [79]. In [80], the following drawbacks of AR were identified: discovery of non-interesting rules, a massive number of discovered rules and low algorithm performance. AR is a powerful data mining technique but the above problems have to be kept in mind. In [24], the author suggested the creation of intelligent hybrid systems that can be used to solve complex problems that cannot be solved by a single intelligent technique.

From the above literature, one can see that ARs can be very useful where there is a need for discovering associations between items. It reveals rules or associations between items. In retail enterprises, the rules might show the items that customers purchase together. This can help retailers develop marketing strategies by gaining insight into which items are frequently purchased by customers. Transactional data generated in retail enterprises reveals which products customers tend to purchase together, thus improving stocking, store layout strategies and promotions [81].

2.4.6 Apriori Algorithm

The apriori algorithm is presented in Table 2.2.

Table 2.2: The Apriori Algorithm. Adapted from [82]

Pseudo-code	
Steps	Input: Transactional data in database (D), Support Output: Candidates with min_support
	C_k : Candidate itemset of size k L_k : Frequent itemset of size k $L_1 = \{\text{frequent items}\};$ for (k=1; $L_k \neq \emptyset$; k++) do begin for each transaction t in database do increment the count of all candidates in C_{k+1} that are contained in t. $L_{k+1} = \text{Candidates in } C_{k+1} \text{ with mini_support}$ end return $\cup_k L_k$;

Example of the Implementation of the Apriori Algorithm

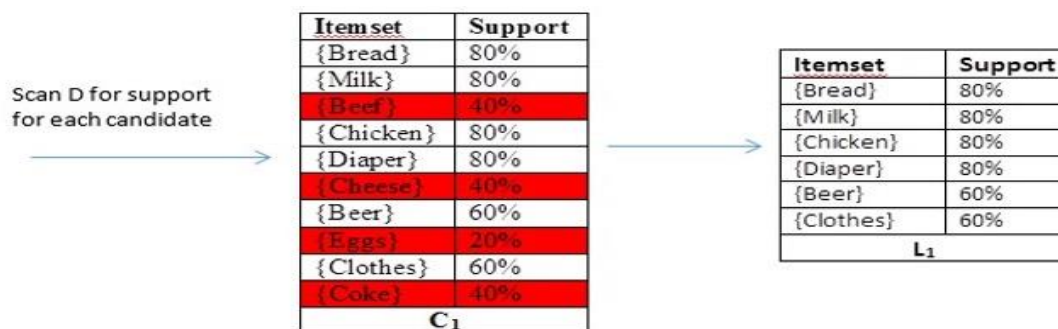
An example of the implementation of the Apriori algorithm is presented using the transactional data in Table 2.3. The dataset contains five transactions performed by different customers and each transaction has a unique identification. In this example, the minimum support (min_sup) was set to 50%, meaning that all item sets with support less than the min_sup are rejected.

Table 2.3: Transactional Data

TID	ITEMS BOUGHT
T500	Bread, Milk, Beef, Eggs
T501	Bread, Diapers, Cheese, Beer, Eggs
T502	Beef, Chicken, Milk, Clothes, Cheese, Diapers, Beer, Coke
T503	Bread, Milk, Chicken, Clothes, Diapers, Beer
T504	Bread, Milk, Diaper, Coke, Clothes, Chicken, Beef

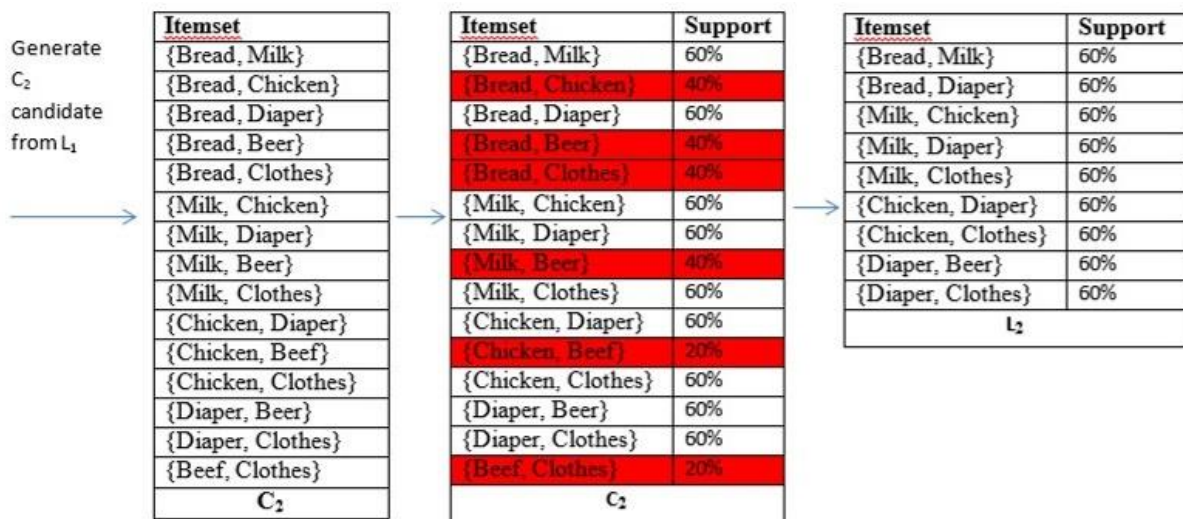
Step 1: Generating One-item Frequent Pattern

The database is scanned to generate the One-item set. The sets with support values below the min_sup set are rejected. All item sets with support values less than the min_sup are highlighted in red. Item set L_1 is accepted.



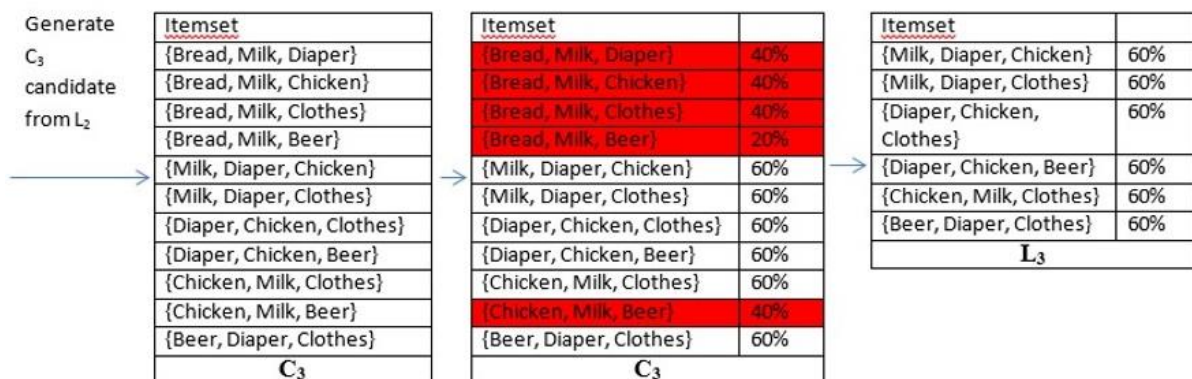
Step 2: Generating Two-item Set Frequent Pattern

Item set L_1 is used to generate the two-item set. All the item sets with support values less than the set min_sup are rejected and these are highlighted in red. All the item sets within the min_sup are accepted and form item set L_2 .



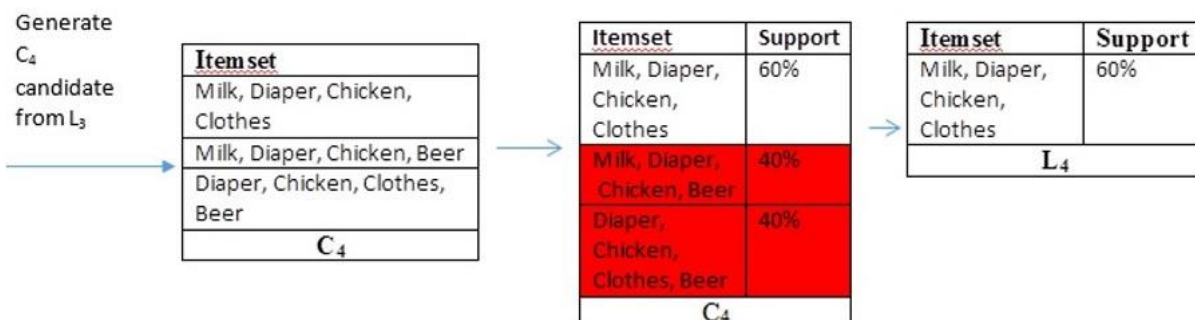
Step 3: Generating Three-item Set Frequent Pattern

Item set L_2 is scanned to generate a three-item set. At this stage, all item sets below the min_sup are rejected and all sets within the range are accepted.



Step 4: Generating Four-item Set Frequent Pattern

At this stage, L_3 is scanned to generate a C_4 item set. The same criteria are applied and only the valid range is accepted. The algorithm cannot generate further item sets so it terminates.



$C_5 = \emptyset$ and the algorithm terminates.

Step 5: Generating AR from frequent item sets

At this stage, possible combinations are extracted. The accepted combinations should have confidence values greater than or equal to the minimum confidence (min_con). The confidence value of 80% is used in this example. In the example, sets R1 and R4 are accepted before the confidence values are within the accepted range. On the other side, sets R2 and R3 are rejected because the confidences are less than the accepted range of 80%.

$L = \{\{\text{Bread}\}, \{\text{Milk}\}, \{\text{Chicken}\}, \{\text{Diaper}\}, \{\text{Beer}\}, \{\text{Clothes}\}, \{\text{Bread, Milk}\}, \{\text{Bread, Diapers}\}, \{\text{Milk, Chicken}\}, \{\text{Milk, Diapers}\}, \{\text{Milk, Clothes}\}, \{\text{Chicken, Diapers}\}, \{\text{Chicken, Clothes}\}, \{\text{Diapers, Beer}\}, \{\text{Diapers, Clothes}\}, \{\text{Milk, Diapers, Chicken}\}, \{\text{Milk, Diapers, Clothes}\}, \{\text{Diapers, Chicken, Clothes}\}, \{\text{Diapers, Chicken, Beer}\}, \{\text{Chicken, Milk, Clothes}\}, \{\text{Beer, Diapers, Clothes}\}, \{\text{Milk, Diapers, Chicken, Clothes}\}\}$

R1: {Milk, Diapers, Chicken} \Rightarrow {Clothes}

Confidence = 100%

R1 is **accepted**

R2: {Bread} \Rightarrow {Milk}

Confidence = 75%

R2 is **accepted**

R3: {Diapers, Clothes}

Confidence = 75%

R3 is **accepted**

R4: {Chicken, Clothes, Milk} \Rightarrow {Diapers}

Confidence = 100%

R4 is **accepted**

2.5 ARTIFICIAL NEURAL NETWORK

2.5.1 Introduction to Artificial Neural Networks

The ANN model started with the study of the functioning of biological neurons in the 1930s and 1940s. In 1943, Warren McCulloch and Walter Pitts proposed a simple model explaining the functionality of biological neurons and published their paper. This was implemented into digital computers by computer scientists. ANN simulates the behaviour of biological systems and is used to discover patterns and relationships. ANN is useful for studying complex relationships between input and output variables in a system [83]. The main advantage of an ANN is the ability to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques [84]. Research conducted by Nirkhi [85] showed that ANN is now commonly used to solve data mining problems because of the following advantages: robustness, self-organising adaptiveness, parallel processing, distributed storage and a high degree of fault tolerance.

2.5.2 Mathematical Basis

The ANN has a basic structure made up of three fundamental components, namely inputs, weights and the activation function. The first step is to insert the inputs into the ANN model. The second step is to determine the weights of the ANN model. The weights are determined during the learning phase of a neural network. Developing an ANN requires three phases. The process model is learned through experiments that correlate input examples and expected outputs [86]. In the recall phase, the network, when properly learned in the training phase, it generalises relationships for undemonstrated inputs and outputs [86]. The ANN sums the inputs x_i against corresponding weights w_i and compares the ANN output to the threshold value, Θ . The threshold is determined by the inputs used. These steps are presented in Figure 2.1.

Figure 2.1 shows the structure of a Feedforward multilayer perceptron network. It is a feedforward network because the information is flowing in one direction and it does not have loops. Feedforward network has a topology which has no closed paths and the input signals are applied in a forward direction through the network [87].

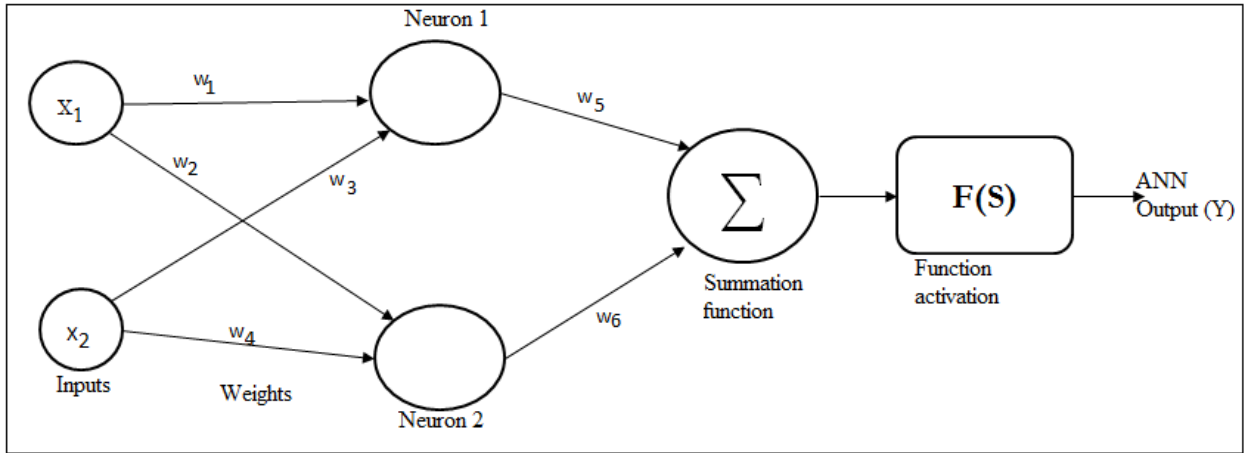


Figure 2.1: Feedforward Multilayer Perceptron Network

Let X be the net weighted input of the neuron as shown in equation (2.7). The selection of X is for discrete cases, since it takes only certain values.

$$X = \sum_{i=1}^n x_i w_i. \quad (2.7)$$

x_i is the input signal, w_i is the weight of the input and n is the number of neurons.

If the net input is less than the threshold, the neuron output is -1; if the net input is greater than or equal to the threshold then the neuron is activated and the output attains a value +1.

Let Y be the ANN output. The choice of Y is for continuous cases, since it can take any values in the range. The actual output of the neuron with the sigmoid activation function is expressed as shown in equation (2.8):

$$Y = \frac{1}{1 + e^{-x}}. \quad (2.8)$$

Other Activation Functions of a Neuron

The following are some common activation functions of a neuron: step function, sign function and linear function. The step and sign activation functions can also be called hard limit functions. These functions can be used in decision-making for classification and pattern recognition activities [88]. These functions can be represented as follows:

$$Step = \begin{cases} 1, & \text{if } X \geq 0 \\ 0, & \text{if } X < 0 \end{cases} \quad (2.9)$$

$$Sign = \begin{cases} +1, & \text{if } X \geq 0 \\ -1, & \text{if } X < 0 \end{cases} \quad (2.10)$$

$$\text{Linear} = X . \quad (2.11)$$

2.5.3 ANN Architecture

The basic ANN structure is made up of three components, namely input layer, hidden layer(s) and output layer. The data is entered into the ANN model through the input layer. The input layer represents the input nodes of the ANN. Feedforward networks consist of multiple layers of computational units, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. The output signal can form the final solution or can be the input of other neurons. The ANN can contain varying hidden layers, depending on the architecture chosen. Figure 2.2 presents a basic architecture of a multilayer ANN.

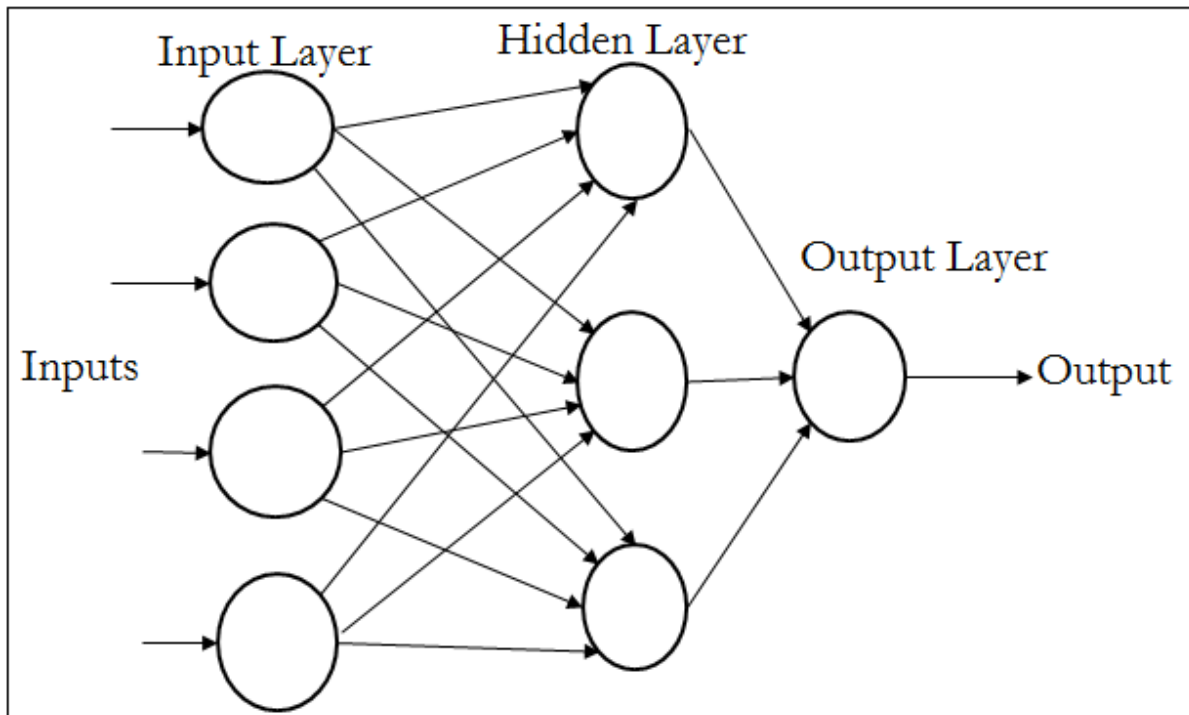


Figure 2.2: ANN Architecture

2.5.4 Areas of Application

ANN has been used in the past to search for patterns and predict future sales [89]. In their research, the authors referred to the sales information of 53 articles of a certain product group belonging to a supermarket. A feedforward multilayer perceptron network with one hidden layer with the back-propagation training method was used. It was shown that the model itself can learn to estimate the time series of sales in supermarkets.

In [90], the authors evaluated the predictive accuracy of ANN and logistic regression in marketing campaigns of a Portuguese banking institution. The algorithms indicated similar overall results. They further checked the runtime for the big data and results showed ANN to be more efficient and faster than logistic regression.

In research conducted by Boone and Roehm [91], ANN was applied for retail segmentation. The authors compared the ANN technique based on Hopfield networks against k-means and mixture model clustering algorithms. The results showed the usefulness of ANNs in retailing for segmenting markets. Hopfield network consists of a set of n interconnected neurons and all these are both input and output neurons. The state of the Hopfield output is maintained until the neuron is updated [92]. Another article [93] presented a unique approach to customer segmentation based on ANN and a clustering algorithm. This was used for identifying the behavioural characteristics of the customer. A satisfactory result was obtained and the experimental results showed the effectiveness of the method in customer classification.

Examples of business applications of ANN were presented in [94]. ANN can be used by financial institutions to develop superior ANN models of credit card risk and bankruptcy in order to improve management decisions. Trading companies also use it as a forecasting technique and trading strategy. Insurance companies use ANN to develop underwriters' models in order to manage risks. Manufacturing companies use ANN to develop predictive process control systems for improving product quality. The same paper also presented characteristics of ANN and its limitations. This was supported by [95], who presented similar ANN characteristics.

An article [96] reported the results of a survey on the application of ANN in forecasting financial market prices. The objective was to assess the potential of ANN in predicting financial systems. ANN yielded promising results in financial forecasting because of its ability to discover nonlinear relationships.

ANN was used in [97] to improve customer satisfaction on product colours using an expert system. The goal was to form appropriate rules for choosing the product's colours. The ANN technique was used to reveal hidden patterns of the customer's needs. The expert system had the capability of ranking the scores of the colours presented in the selection. It was shown that ANN identifies patterns within the data set.

In research by [98] ANN was used to analyse the impact of advertising and promotion on a business enterprise. The authors used 220 samples of daily data and 106 samples of weekly data to test the model. A back propagation algorithm was used for conducting the experiment. Their results showed the capability of ANN to capture the nonlinear aspects of complex relationships and the predictive quality of ANN was revealed. This depended on the frequency of the data used to observe the model's performance. Back propagation proved to be an efficient method for studying relationships.

In [99] ANN was applied to the Pima Indians' diabetes database and it generated rules with strong associations, thereby enhancing the decision-making process of doctors. In 2007 they further presented a method to overcome the disadvantage of ANN by attracting symbolic knowledge from a trained ANN using the tree approach. This was a way of boosting the ANN acceptance confidence as a data mining tool because it has strong generalisation ability. Many articles mentioned in [100] consider ANN to be a promising data mining tool. Other papers that support ANN in data mining include [101] and [102].

Other areas of ANN application include wind forecasting [103], social computing [104], engineering [105], medical and clinical decision-making [106]; [107], electric power industry [108]; [109], misuse detection [110], image recognition and classification of crop and weeds [111], as an analytical tool for groundwater salinity [112], pattern recognition [113], rainfall prediction [114], [115], stock market prediction of the values of indexes [116], etc.

Research reported in [117] described the possibilities of applying ANN to data mining and semantic integration in the field of machine learning. The authors concluded that ANN is suitable for the specified field. Kohonen self-organising networks and back propagation networks were tested. Kohonen network has a feed-forward structure with a single computational layer arranged in rows and columns. Each neuron is fully connected to all the source nodes in the input layer [118]. The research by Tiecheng [119], optimised the mining of AR based in ANN. Data mining techniques and ANN were used as a prediction model for financial distress [120]. Research was conducted by Chou, Li, Chen and Wua [121], for analysing customer navigation patterns through web mining and ANN in order to predict customers' future needs. A successful EC website with a dataset from December 2004 to December 2005 was used to carry out a computational intelligent analysis. Moreover, [122] supported ANN in data mining because of the ability to learn the target

concept better than data mining methods. However, the authors presented two limitations that make them poor data mining tools: excessive training times and incomprehensible learning.

In [123], the authors showed how data mining techniques and a neural network algorithm can be combined successfully in credit card fraud detection. A sample data set of 5 850 fraudulent transactions and 30 000 legal transactions was used to test the experiment. It was shown that the combination of these methods yielded very good results. In [124], they also developed an automatic diagnosis system, which can be used to detect breast cancer based on AR and ANN. The proposed model yielded a 95.6% correct classification rate. The proposed AR + neural network model proved to be better than AR and neural network models for obtaining efficient automatic diagnostic systems for other diseases.

In research conducted by Arockiaraj [125], it was shown that the combination of data mining methods and a neural network model can greatly improve the efficiency of data mining methods. The purpose of the research is to investigate the performance of the AR model complemented by the ANN model in retail enterprises. The research might strengthen the performance of the AR model by combining the strength of AR and ANN models in order to find the best product arrangement patterns that can be adopted in retail enterprises. This may be a better way of attracting customers to buy more products than expected.

2.6 DATA INTEGRATION

2.6.1 Concept of Data Integration

Data integration is the process of combining data stored at different sources to provide the organisational users with a single view of the data [126]. It refers to the process of providing seamless access to multiple sources of data that appear to the system's users as a single resource [127]. This process combines transactional data from different sources, such as different computers, applications (such as Excel and databases). In data integration, different tables are joined to form a single table. The integrated data can contain duplicate columns, missing values and unnecessary data values. It is crucial in centralised retail enterprises to bring data from multiple sources and different branches together. Data integration may lead to errors in the combined data, such as false positives (FP) and false negatives (FN). It reduces the size of data to obtain a smaller number of attributes, while

recording those that are informative. However, in distributed retail enterprises, transactional data does not need to be integrated.

2.6.2 Data Integration Techniques

There are three main data integration techniques: data consolidation, data propagation and data federation. These techniques are illustrated in Figure 2.3. The next subsections provide a discussion of the data integration techniques that can be used in centralised retail enterprises.

A) Data Consolidation

Data consolidation captures data from more than one data source (A, B and C) and integrates it into a single persistent data store (D). The data store can then be used as an operational data warehouse [128]. Data is extracted from different sources, transformed into the designed form and loaded into a target for analysis purposes. This technique provides the major benefit of updating the consolidated data store periodically without requiring additional storage on the original data stores. On the other hand, its main disadvantage is that the data consolidation process may cause a delay between the time of data update in the source system and the time of data update in the target system [128].

B) Data Propagation

This technique is applied to support the redistribution and sharing of master data back to the participating applications [129]. Data propagation copies data from one location to another on real-time or near real-time synchronisation of the target and master sources [128]. It is supported by enterprise application integration technology and it operates online, pushing the data to the target store [130]. Data propagation updates the replicated view on a near to real-time basis. It requires additional storage on the original data stores. It works mainly on small volumes of data and makes use of data on a cross-application basis [131].

C) Data Federation

Data federation integrates data from different sources into a virtual database for analysis purposes. It uses a virtual database (virtual view) with a mediated schema and wrapper commonly known as the adapter, which translates incoming queries and outgoing answers [128]. The integrated data is separated from the details of data sources. Data federation is

supported by enterprise information integration technology and the integrated table is used for analysis purposes [132]. The data source remains at the original place and only the results persist in the server when needed. The main advantage of data federation is that it provides real time on data and requires no additional data storage. Unlike data consolidation, the data federation technique provides data on a real-time basis though it does not require additional storage either.

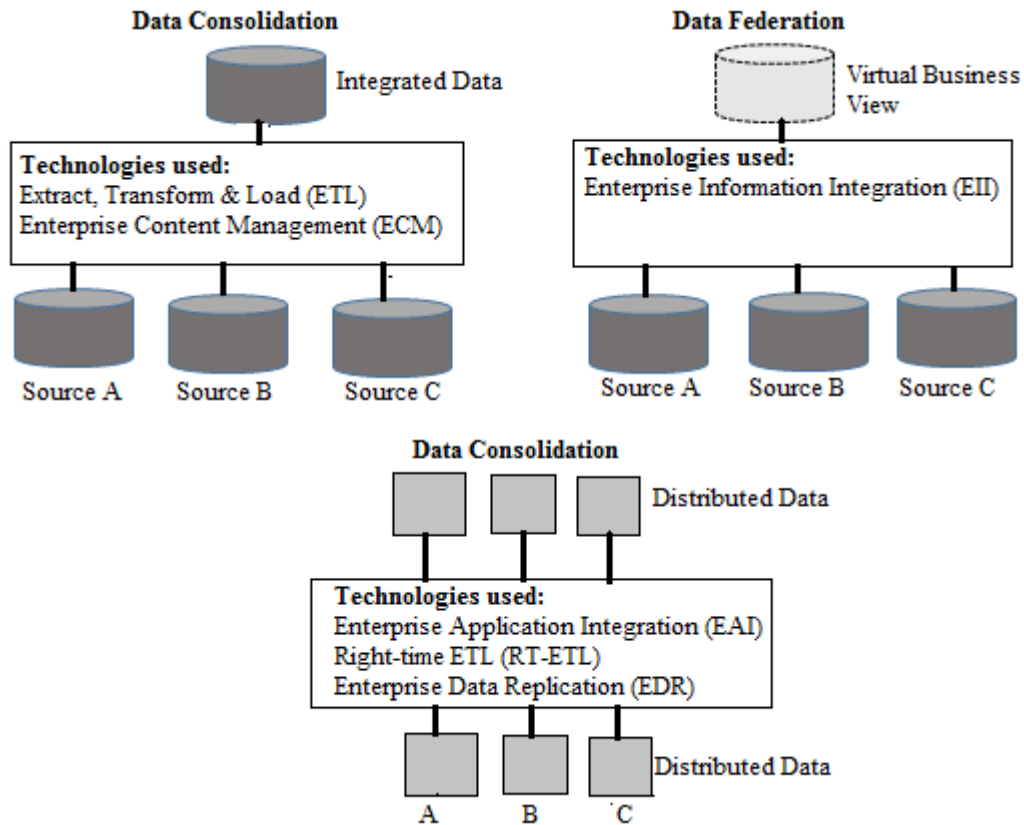


Figure 2.3: Data Integration Techniques. Adapted from [1]

2.7 CHAPTER SUMMARY

This chapter discussed the emergence of MBA and its benefits to retail enterprises. MBA can assist retail enterprises' decision-makers to make informed decisions. This in turn can improve sales and profit levels within the retail enterprises.

The study also pointed out some challenges of the current prevailing sales optimisation models in the retail enterprise environment. The research aims at improving the challenging models that hinder the performance of retail enterprises. MBA methods in the retail enterprises were identified and discussed. These methods help retail managers to

perform different analysis and make informed decisions that might improve the operations of an enterprise in the prevailing competitive environment.

The research also discussed and highlighted some examples of the two techniques that are used to build the ARANN model. The ARANN model is built on AR to be complemented with ANN in order to strengthen the results of both models. From the literature, the weaknesses and the strengths of the individual classical models were studied and a combination of data mining and ANN models was proposed. This research combines the two models in retail enterprises to observe the best product combination sets in order to improve sales levels. This might be a way of improving the current challenging sales optimisation models in retail enterprises. A cooperative intelligent analytics-based model combining AR and ANN models was proposed to complement each other in product arrangement sets' generation.

CHAPTER 3

RESEARCH METHODOLOGY AND PROPOSED SYSTEM MODEL

3.1 INTRODUCTION

This chapter explores the proposed ARANN model for business analytics for centralised and distributed retail enterprises. In centralised retail enterprises, data generated from different branches is integrated to form a single view of the organisation. The data is then pre-processed to remove any impurities such as empty columns and/or invalid data. The processed data feeds the proposed model for centralised retail enterprises. At the end, the ARANN model generates product arrangement sets that can be used across all retail branches in order to give uniformity and encourage customers to buy more than planned.

In distributed retail enterprises, data generated from different branches is collected in a data warehouse. The data for each branch is first pre-processed to remove data impurities. The processed data is entered into the ARANN model for generation of arrangement sets. Each dataset generates its own product arrangement sets, which are then applied per branch.

The claims of the current study are that the proposed ARANN model might improve the way in which products are arranged in centralised and distributed retail enterprises. This might in turn increase the sales levels of the retail enterprises, thereby increasing the profits.

3.2 DATA COLLECTION AND PREPARATION

3.2.1 Data Collection in Centralised Retail Enterprises

Real-life data was collected from a retail enterprise in South Africa from three different branch locations. The data was collected solely for research purposes and the identity of the retail enterprise is kept confidential. Each branch contains 66 records and 24 products. The data shows the transactions performed by different customers. Each row represents a customer transaction. Items purchased by a customer form a record or a row. Items that are often purchased by customers were identified and used to form the columns of the dataset. The datasets are stored in different folders. Table 3.1 shows a sample of real-life sample

data. “Yes” indicates that an item was purchased and “No” indicates that the customer did not include the item in the basket.

Table 3.1: Real-life Sample Data

Body lotion	Colgate	Rice	Bread	Salt	Beans	Beef	Chicken	...	Sugar
Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
No	No	Yes	Yes	No	No	Yes	Yes	No
....
Yes	No	No	Yes	No	Yes	Yes	Yes	Yes

The public data was downloaded from [3]. The dataset was designed for academic purposes. The data was modelled to work with the ARANN model. Table 3.2 shows the public sample data. The data contains 1000 records and includes seven products. In this research study, the dataset was randomly divided into five branches assumed to be folders, with about 200 records in each branch. This was integrated into a centralised database to observe the impact of the data integration process before generating optimal product arrangement on shelves. The integration time was nine seconds.

Table 3.2: Public Sample Data. Adapted from [3]

Cookies	Fish	Orange juice	Lemon tea	Red wine	Peanuts	Canned soup
Yes	Yes	Yes	No	No	No	Yes
.....
.
Yes	No	Yes	No	No	Yes	Yes

3.2.2 Importance of Data Integration into a Centralised Retail Enterprise

The following reasons reflect the importance of data integration into a centralised retail enterprise:

- It gives a single view of the data to the organisational data users.
- It ensures uniformity within a centralised retail enterprise. Results revealed by data that has been integrated is applied across the organisation.
- It enables the data to be analysed for patterns that can give meaningful information to decision-makers, such as customer preferences and trends.
- It enables better management of the data so that various sources of the data can be monitored effectively.
- It can ensure better data security, as the data can be stored in a centralised location.
- It makes centralised decisions easier across the branches of a retail enterprise.

3.2.3 Data Collection in Distributed Retail Enterprises

In distributed retail enterprises, real-life data was collected from a retail enterprise situated in South Africa with several branches nationwide. The data for the experiments was collected from only eight branches in different parts of a developing country. The retail enterprise has database servers at each branch for the storage of data. Real-life datasets consisting of 66 records were taken from each branch to be used for running experiments. In the experiment, the most frequently purchased products were considered. The data was then exported to a notepad application for storage. Each row in Tables 3.3–3.5 represents a transaction performed by the customer. Table 3.3 and Table 3.4 show samples of real-life data from different branches.

Table 3.3: Sample of Real-life Data for Branch 1

Body lotion	Colgate	Rice	Maize meal		
Meat	Rice	Roll on	Cooking oil	Body lotion	
.....
Drink	Roll on	Mince	Coke	Colgate	Perfume

Table 3.4: Sample of Real-life Data for Branch 2

Bread	Sugar	Rice	Meat	Salt	Cooking oil	Flour	Soup
.....
Fruits	Sugar	Meat	Cooking oil	Salt	Soap	Bread	

In the public dataset 1000 transactions were used. This dataset was randomly broken up into five datasets representing branches and the records for each branch contained 200 transactions. The data was saved in the .txt format. The public dataset in Table 3.5 is found in [3]. The data contains the following products: bread, beer, tea, wine, orange juice, chocolate milk and canned soup.

Table 3.5. Sample of Public Data. Adapted from [3]

Fish	Orange juice	Tea	Wine	Peanuts	Canned soup	Bread	Beer
.....
.....
Cookies	Fish	Orange juice	Tea	Wine	Peanuts	Canned soup	Chocolate milk

3.2.4 Data Preparation Stages

Data preparation involves all the activities performed on the generated raw data for knowledge discovery so that the data will be ready for modelling in data mining models [2]. It involves all the activities carried out to build the final dataset [133], as shown in Figure 3.1. It is a crucial stage in the knowledge discovery in database process. According to [134], the three important aspects of data preparation are:

- Real-world data is impure or dirty.
This is a database record with errors such as duplicating rows, an outdated database record, incorrect entries and blank columns or fields. This data may lack some attribute values or interesting attributes.
- High-performance mining systems require quality data.
- Quality data yields high-quality patterns.

The purpose of data preparation is to improve the quality of data for successful data mining [22]. In [135], data preparation was considered to be the first step of data pre-processing and data cleaning, data transformation and data selection were identified as the main steps in data preparation. In support [134] identified the following data preparation techniques: data collection, data integration, data transformation, data cleaning, data reduction and data discretisation. Data preparation is a crucial and repetitive step and the quality of data used determines the quality of patterns discovered [136]. Based on the above, it can be seen that data preparation is an important phase in ensuring quality data for modelling.

Data generated from different branches of a retail enterprise might have different column names and the product name may be written differently branch by branch. This data is regarded as dirty data, i.e. data with impurities. If this dirty data is used for modelling, misleading patterns may be revealed. Pre-processed data should be used in order to get correct values such as support and confidence. When dirty data is used, there are high chances of getting incorrect support and confidence values. Data preparation plays a crucial role in modelling retail enterprise data.

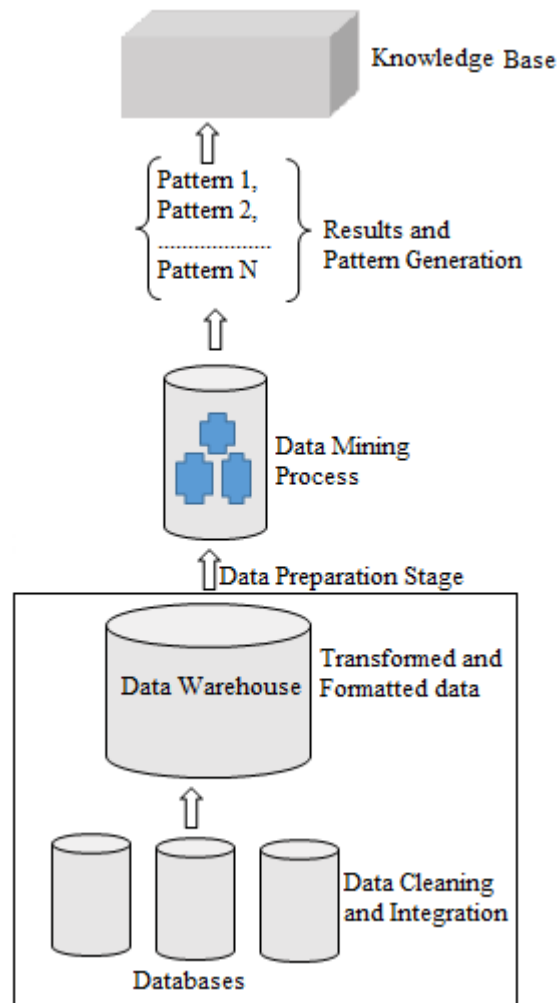


Figure 3.1: Stages and outputs of the knowledge discovery in database process. Adapted from [2]

This section discusses the following data preparation stages:

A) Data Collection and Integration

Data collection involves collecting transactional data from different branches of a retail enterprise. In a centralised retail enterprise, the data collected from different sources needs to be integrated. After data integration, the data becomes dirty due to integration problems. In distributed retail enterprises, there is no need for integrating data from different branches. In distributed retail enterprises data is processed per branch, while in centralised retail enterprises data is processed as a single dataset. The data collected from different sources is then cleaned.

B) Data Cleaning

Data cleaning detects and removes errors and data inconsistencies with the goal of improving the quality of data for modelling [137]. It checks and removes inconsistencies that occur during data entries, missing values after data integration and other invalid data in the databases. The following problems can occur in response to dirty data: data with missing values, non-existent data, incomplete data and unknown or null values. In data cleaning, empty rows, columns, unknown values and wrong entries are removed. Data cleaning is crucial to improve the quality of transactional data generated from different branches of a retail enterprise. The model results depends on the quality of data used for modelling. The current study claims that quality data can improve the performance of the ARANN model both in centralised and distributed retail enterprises. The cleaned data is then transformed.

C) Data Transformation

Data transformation involves changing the data into an appropriate form for modelling. It includes changing the data form to a format accepted by the ARANN model, such as .arff or .csv. This enables the ARANN model to process the data appropriately. In the current study, the pre-processed transactional data was formatted to a .txt file extension accepted by the ARANN model.

3.3 PROPOSED ARANN MODEL FOR RETAIL ENTERPRISES

The proposed ARANN model for retail enterprises is composed of the AR and ANN models shown in Figure 3.2. The transactional data is fed into the AR model and in turn outputs the support and confidence values. The AR model passes the support and confidence values to the ANN model. The ANN model uses those values as its input values. These inputs are multiplied to the corresponding weights and summed together. The ANN model outputs the DoB values for each generated product arrangement set.

Intelligent Model

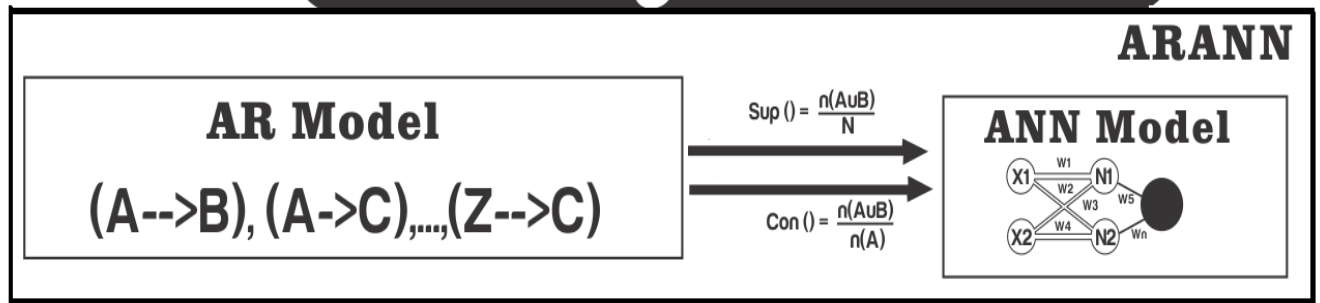


Figure 3.2: Intelligent Model for Retail Enterprises

3.4 ARANN SYSTEM MODEL FOR CENTRALISED ENTERPRISES

This section explores the proposed system model for business information analytics in centralised retail enterprises. The proposed ARANN model has three main subsystems: data integration, data preparation and the development of a cooperative intelligent analytics-based subsystem for product arrangement, as shown in Figure 3.3. The model collects data from sources and branches of retail enterprises. The data is generated at the point-of-sale machines. This data can be from different PCs or LANs or WANs. The collected data is integrated using data integration techniques. During data integration, the data encounters some integration errors, such as FNs and FPs. The integrated data is then prepared in order to remove some data impurities and to improve the quality of data to be used for data modelling. The steps discussed in Section 3.2.4 are carried out to pre-process the data. Those steps are repeated until the data is clean. The cleaned or pre-processed data can be stored in a data warehouse. The pre-processed data is then fed into the proposed ARANN model for the generation of product arrangement sets. The proposed model is built on ANN and AR, to complement each other. When the AR model generates product arrangement sets, the proposed model compares it with the results generated by the ANN model. The ARANN model accepts product arrangement sets where both models agree. The results generated determine how products are arranged on the shop shelves of each branch of a centralised retail enterprise. The accepted product arrangement sets will be applied across all the branches of a retail enterprise. This ensures uniformity to the customers and might improve the sales levels of the retail enterprise.

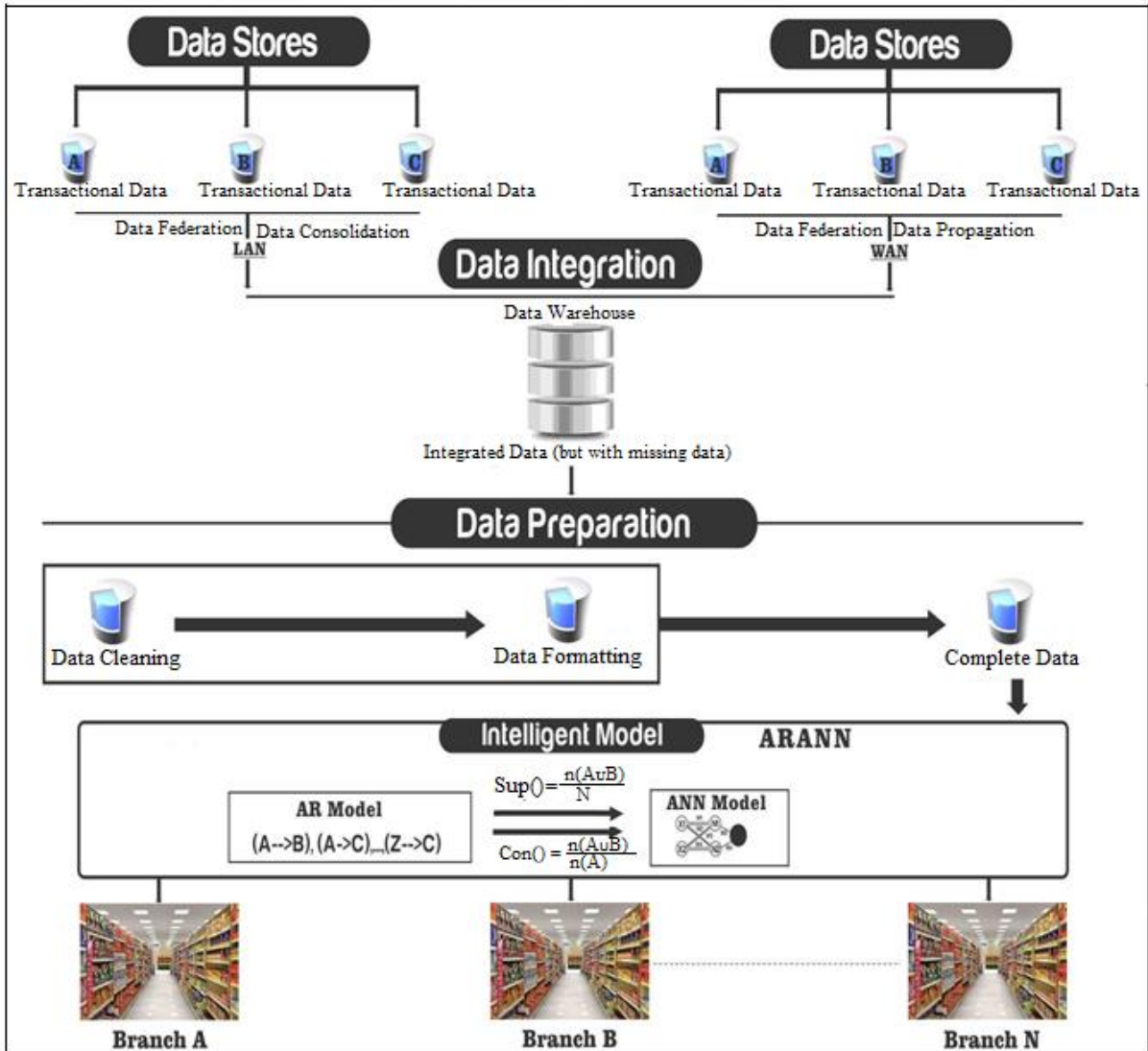


Figure 3.3: Proposed ARANN Framework for Centralised Analytics

The proposed framework's benefits are:

- Development of a cooperative intelligent analytics-based framework for centralised retail enterprise management.
- Development of a data integration subsystem for providing a single view to centralised enterprise data.
- Development of a data preparatory subsystem.
- Improvement of the basis for decision-making and confidence of professional managers in centralised retail enterprises.
- Reduction in poor data quality problems and losses.
- Improvement in the effectiveness of current product sales optimisation models.

The next section shows how the ARANN model can be implemented in centralised retail enterprises for the generation of product arrangement sets.

3.4.1 Pseudo-code of the ARANN Model for Centralised Analytics

The ARANN model for centralised analytics was implemented using the pseudo-code shown in Table 3.6.

Table 3.6: Pseudo-code for ARANN Model in Centralised Analytics

Pseudo-code	
Steps	Input: Transactional data from different branches Support () Confidence () Weights (W) = {w ₁ , w ₂ , w ₃ , ..., w _n } Output: Product arrangement sets
	Step 1: C _k : Candidate itemset of size k Step 2: L _k : Frequent itemset of size k Step 3 L ₁ = {frequent items}; Step 4: for (k=1; L _k !=∅; k++) do begin Step 5: for each transaction t in database do increment the count of all candidates in C _{k+1} that are contained in t. Step 6: L _{k+1} = Candidates in C _{k+1} Step 7: return Support () and Confidence () values for each set Step 8: Set all weights equal to the Support value (Heuristic 1) Step 9: Feed the Support () and Confidence () into Neuron 1 (N ₁) and Neuron 2 (N ₂) as inputs Step 10: Generate N ₁ by summing of the inputs with the corresponding weights and apply the output into the sigmoid function Step 11: Generate N ₂ by summing of the inputs with the corresponding weights and apply the output into the sigmoid function Step 12: Generate the summation of N ₁ and N ₂ after the sigmoid function and apply the output into the sigmoid function to obtain DoB by applying thresholding function (Heuristic 2) Step 13: Generate product arrangement sets where DoB ≥ ARANN activation } Display product arrangement sets within the range End }

3.4.2 Mathematical Description of the ARANN Model for Centralised Analytics

The ARANN model takes the integrated and prepared data. This processed data is fed into the AR model for the generation of the support (sup) and confidence (con) values using equation (2.1) and equation (2.2) respectively.

These sup and con values are fed into the ANN model as the inputs and multiplied to the corresponding weights, as shown in equation (3.1) and equation (3.2).

$$Neuron_1 = (Sup * W_1) + (Con * W_2) \quad (3.1)$$

$$Neuron_2 = (Sup * W_3) + (Con * W_4) \quad (3.2)$$

Equation (3.3) and (3.4) show the output of Neuron₁ and Neuron₂ after the sigmoid function.

$$NOutp_1 = \frac{1}{1 + e^{-Neuron_1}} \quad (3.3)$$

$$NOutp_2 = \frac{1}{1 + e^{-Neuron_2}} \quad (3.4)$$

Equation (3.5) shows the sum of the neurons.

$$NeuronOutput = NOutp_1 W_5 + NOutp_2 W_6 \quad (3.5)$$

Equation (3.6) shows the DoB.

$$DoB = \frac{1}{1 + e^{-NeuronOutput}} \quad (3.6)$$

$$\text{Product arrangement sets} = \begin{cases} \text{Accepted, if } DoB \geq \text{ARANN activation} \\ \text{Re jected, if } \textit{otherwise} \end{cases} \quad (3.7)$$

where W_1, W_2, W_3, W_4, W_5 and W_6 are the corresponding weights and ARANN activation is the threshold value set. The weights are all set to equal to the Support Value (Heuristic 1) and the Threshold Value Set are heuristic thresholds (Heuristic 2). It should be noted that the ANN cannot learn in this research.

3.4.3 Scenario: Arrangement of Products on Shelves for Centralised Retail Analytics

Figure 3.4 shows a scenario of how the ARANN model displays product arrangement results in centralised retail branches. Data is integrated from different branches of retail enterprises, which is then followed by data preparation. The data was pre-processed manually. The ARANN model runs on complete data to generate different patterns of arrangement.

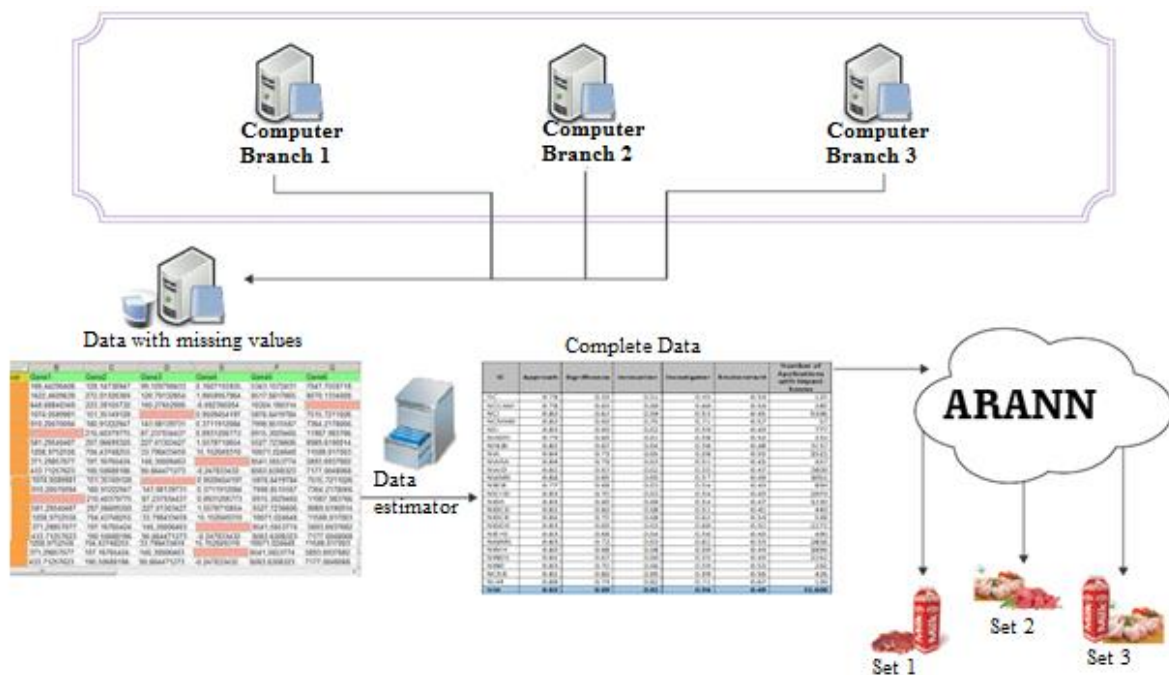


Figure 3.4: The ARANN Model for Centralised Retail Branches

The data in Table 3.7 was used for the generation of product arrangement sets presented in Figure 3.4. Transactional data presented in Table 3.7 was also used for the worked analysis of the ARANN model in centralised retail enterprises.

Table 3.7: Market Basket Transactional Data for All Centralised Branches

Market Basket Transaction Data – Data for All Branches	
TID	ITEMS
T500	Bread, Milk, Beef, Chicken
T501	Bread, Diapers, Cheese, Beer, Eggs
T502	Beef, Chicken, Milk, Clothes, Cheese, Diapers, Beer, Coke
T503	Bread, Milk, Chicken, Clothes, Diapers, Beer
T504	Bread, Milk, Diapers, Coke, Clothes, Chicken, Beef

To perform analysis on Table 3.7 data, even weights were applied to corresponding inputs to avoid bias on products. This was obtained through dividing the count of $a_{union}b$ over the number of records in the data set, where a and b are different products. The following ARANN activation was used:

≥ 0.75 strongly connected products (Strongly accepted)

≥ 0.65 moderately connected products (Accepted)

< 0.65 weakly connected products (Rejected).

Analysis of ARANN on Table 3.7

{Bread} => {Milk}

$$\text{Support} = \frac{n(A \cup B)}{N} = \frac{3}{5} = 0.6$$

$$\text{Confidence} = \frac{n(A \cup B)}{n(A)} = \frac{3}{4} = 0.75$$

$$N_1 = (\text{Sup} \times w_1) + (\text{Con} \times w_2)$$

$$N_2 = (\text{Sup} \times w_3) + (\text{Con} \times w_4)$$

$$= (0.6 \times 0.6) + (0.75 \times 0.6)$$

$$= (0.6 \times 0.6) + (0.75 \times 0.6)$$

$$= 0.81$$

$$= 0.81$$

$$O_1 = \frac{1}{1 + e^{-N_1}} = \frac{1}{1 + e^{-0.81}} = 0.69$$

$$O_2 = \frac{1}{1 + e^{-N_2}} = \frac{1}{1 + e^{-0.81}} = 0.69$$

$$F = w_5 O_1 + w_6 O_2$$

$$= (0.6 \times 0.69) + (0.6 \times 0.69) = 0.83$$

$$\text{DoB} = \frac{1}{1+e^{-F}} = \frac{1}{1+e^{-0.83}} = 0.70$$

Product pattern $\Rightarrow 0.70 \geq 0.65$

Therefore it is **moderately** connected and is **accepted**.

{Beef} \Rightarrow {Chicken}

$$\text{Support} = \frac{n(A \cup B)}{N} = \frac{3}{5} = 0.6$$

$$\text{Confidence} = \frac{n(A \cup B)}{n(A)} = \frac{3}{3} = 1.0$$

$$\begin{aligned} N_1 &= (\text{Sup} \times w_1) + (\text{Con} \times w_2) \\ &= (0.6 \times 0.6) + (1.0 \times 0.6) \\ &= 0.96 \end{aligned}$$

$$\begin{aligned} N_2 &= (\text{Sup} \times w_3) + (\text{Con} \times w_4) \\ &= (0.6 \times 0.6) + (1.0 \times 0.6) \\ &= 0.96 \end{aligned}$$

$$O_1 = \frac{1}{1+e^{-N_1}} = \frac{1}{1+e^{-0.96}} = 0.72$$

$$O_2 = \frac{1}{1+e^{-N_2}} = \frac{1}{1+e^{-0.96}} = 0.72$$

$$\begin{aligned} F &= w_5 O_1 + w_6 O_2 \\ &= (0.6 \times 0.72) + (0.6 \times 0.72) = 0.86 \end{aligned}$$

$$\text{DoB} = \frac{1}{1+e^{-F}} = \frac{1}{1+e^{-0.86}} = 0.70$$

Product pattern $\Rightarrow 0.70 \geq 0.65$

Therefore it is **moderately** connected and is **accepted**.

{Milk} \Rightarrow {Chicken}

$$\text{Support} = \frac{n(A \cup B)}{N} = \frac{4}{5} = 0.8$$

$$\text{Confidence} = \frac{n(A \cup B)}{n(A)} = \frac{4}{4} = 1.0$$

$$\begin{aligned} N_1 &= (\text{Sup} \times w_1) + (\text{Con} \times w_2) \\ &= (0.8 \times 0.8) + (1.0 \times 0.8) \\ &= 1.44 \end{aligned}$$

$$\begin{aligned} N_2 &= (\text{Sup} \times w_3) + (\text{Con} \times w_4) \\ &= (0.8 \times 0.8) + (1.0 \times 0.8) \\ &= 1.44 \end{aligned}$$

$$O_1 = \frac{1}{1+e^{-N_1}} = \frac{1}{1+e^{-1.44}} = 0.81$$

$$O_2 = \frac{1}{1+e^{-N_2}} = \frac{1}{1+e^{-1.44}} = 0.81$$

$$\begin{aligned} F &= w_5 O_1 + w_6 O_2 \\ &= (0.8 \times 0.81) + (0.8 \times 0.81) = 1.3 \end{aligned}$$

$$\text{DoB} = \frac{1}{1+e^{-F}} = \frac{1}{1+e^{-1.3}} = 0.79$$

Product pattern $\Rightarrow 0.79 \geq 0.75$

Therefore it is **strongly** connected and is **strongly accepted**.

3.5 ARANN SYSTEM MODEL FOR DISTRIBUTED ENTERPRISES

This section explores the proposed system model for BI analytics in distributed retail enterprises. The proposed model has three layers, namely data cleaning and formatting, intelligent subsystem and distributed product shops, as shown in Figure 3.5. The data cleaning and formatting layer is found at the bottom of the proposed ARANN model. In this proposed model, data is collected from transactional systems branch per branch. The data is cleaned and formatted to the appropriate file type accepted by the proposed model. Processed data is input using separate branches at the middle layer of the ARANN model. The processed data from the bottom layer is passed into the AR model and it outputs confidence and support values. These values are passed into the ANN model as inputs in order to get the DoB. The DoB of sets generated are compared to the ARANN activations set. The accepted sets generated are applied on the top layer of the proposed model. This proposed model is deployed to each branch and patterns are generated independently. The choice is left to every retail enterprise branch to adopt the best results, depending on the market competitiveness and profit levels.

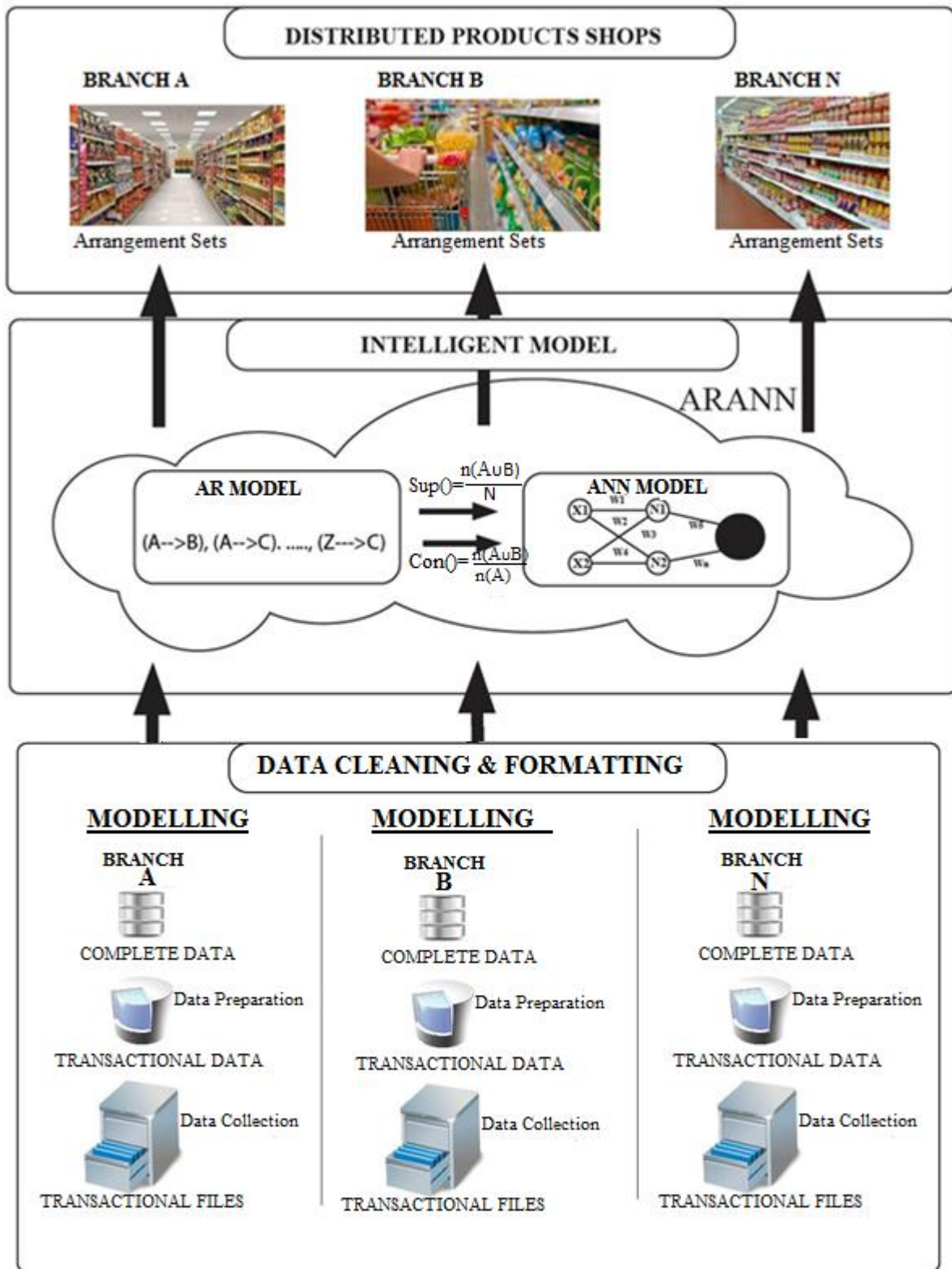


Figure 3.5. Proposed Intelligent Analytics-based Framework for Distributed Analytics

The proposed intelligent analytics-based framework has the following benefits:

- Reduction in risk of passing misleading results to all branches.
- No one point of failure.
- Consumption of fewer resources.
- Faster construction of distributed systems.
- No need for data integration.

3.5.1 Pseudo-code of ARANN to Distributed Analytics

This proposed intelligent analytics-based model can be implemented using the pseudo-code presented in Table 3.8. Table 3.8 shows how ARANN generates product arrangement sets that can be used by retail enterprise managers to arrange products on shop shelves so as to attract customers to purchase more products than planned.

Table 3.8: Pseudo-code for ARANN Model in Distributed Analytics

Pseudo-code		
Steps	Input:	Transactional data in database (D) = {t ₁ , t ₂ , t ₃ , ..., t _n } Support () Confidence () Weights (W) = {w ₁ , w ₂ , w ₃ , ..., w _n }
	Output:	Products pattern
		Step 1: D = {t ₁ , t ₂ , t ₃ , ..., t _n } //Transactions in the database Step 2: C _k = Candidate item set of size k Step 3: F _k = frequent item set of size k { for (k=1; F _k != ∅; k++) // F _k is not equal to empty set. { Scan the entire D to generate candidate sets C _k { Compare candidate support count from C _k with the minimum support count to generate F _k } } } Step 4: Generate Support () & Confidence () { Step 5: Set all weights equal to the Support Value. (Heuristic 1)

Step 6: Input Support () & Confidence () into Neuron 1 (N_1) and Neuron 2 (N_2) as inputs

Step 7: Generate N_1 by summing of the inputs with the corresponding weights and apply the output into sigmoid function

Step 8: Generate N_2 by summing of the inputs with the corresponding weights and apply the output into sigmoid function

Step 9: Generate the summation of N_1 & N_2 after the sigmoid function and apply the output into sigmoid function to obtain DoB by applying thresholding function (Heuristic 2)

Step 10: Display products pattern where **DoB** \geq **ARANN activation**

}

3.5.2 Mathematical Description of ARANN to Distributed Analytics

The ARANN model is implemented mathematically by determining the support (sup) value generated from equation (2.1) and the confidence (con) value generated from equation (2.2).

The sup and con values feed the N_1 as the inputs and are multiplied with the corresponding weights.

$$N_1 = SupW_1 + ConW_3 \quad (3.8)$$

The output of N_1 after the sigmoid function:

$$O_1 = \frac{1}{1 + e^{-N_1}} \quad (3.9)$$

The sup and con values feed the N_2 as the inputs and are multiplied by the corresponding weights.

$$N_2 = ConW_4 + SupW_2 \quad (3.10)$$

The output of N_2 after the sigmoid function

$$O_2 = \frac{1}{1 + e^{-N_2}} \quad (3.11)$$

$$F = W_5 O_1 + W_6 O_2 \quad (3.12)$$

$$= \frac{W_5}{1 + e^{-N_1}} + \frac{W_6}{1 + e^{-N_2}} \quad (3.13)$$

$$DoB = \frac{1}{1 + e^{-F}} \quad (3.14)$$

$$\text{Product Patterns} = \begin{cases} \text{Accepted, if } DoB \geq \text{ARANN activation} \\ \text{Re jected, if } \textit{otherwise} \end{cases} \quad (3.15)$$

where N_1 and N_2 are Neuron 1 and 2 respectively; W_1, W_2, W_3, W_4, W_5 and W_6 are the corresponding weights; O_1 is Neuron 1 output after sigmoid function; O_2 is Neuron 2

output after sigmoid function, F is input to final neuron and ARANN activation is the threshold value set. The weights are all set to equal to the Support Value (Heuristic 1) and the Threshold Value Set are heuristic thresholds (Heuristic 2). It should be noted that the ANN used cannot learn.

3.5.3 Scenario: Arrangement of Products on Shelves for Distributed Retail Branches

Figure 3.6 shows how the cooperative model displays placement results in distributed branches. Transactional data from each retail branch was loaded into the ARANN model to determine the arrangement sets. The data was loaded branch by branch in order to determine the arrangement sets for each branch. The data for each branch was pre-processed first before being fed into the ARANN model. The results from each branch can be used to arrange products on the shelves of a particular branch.

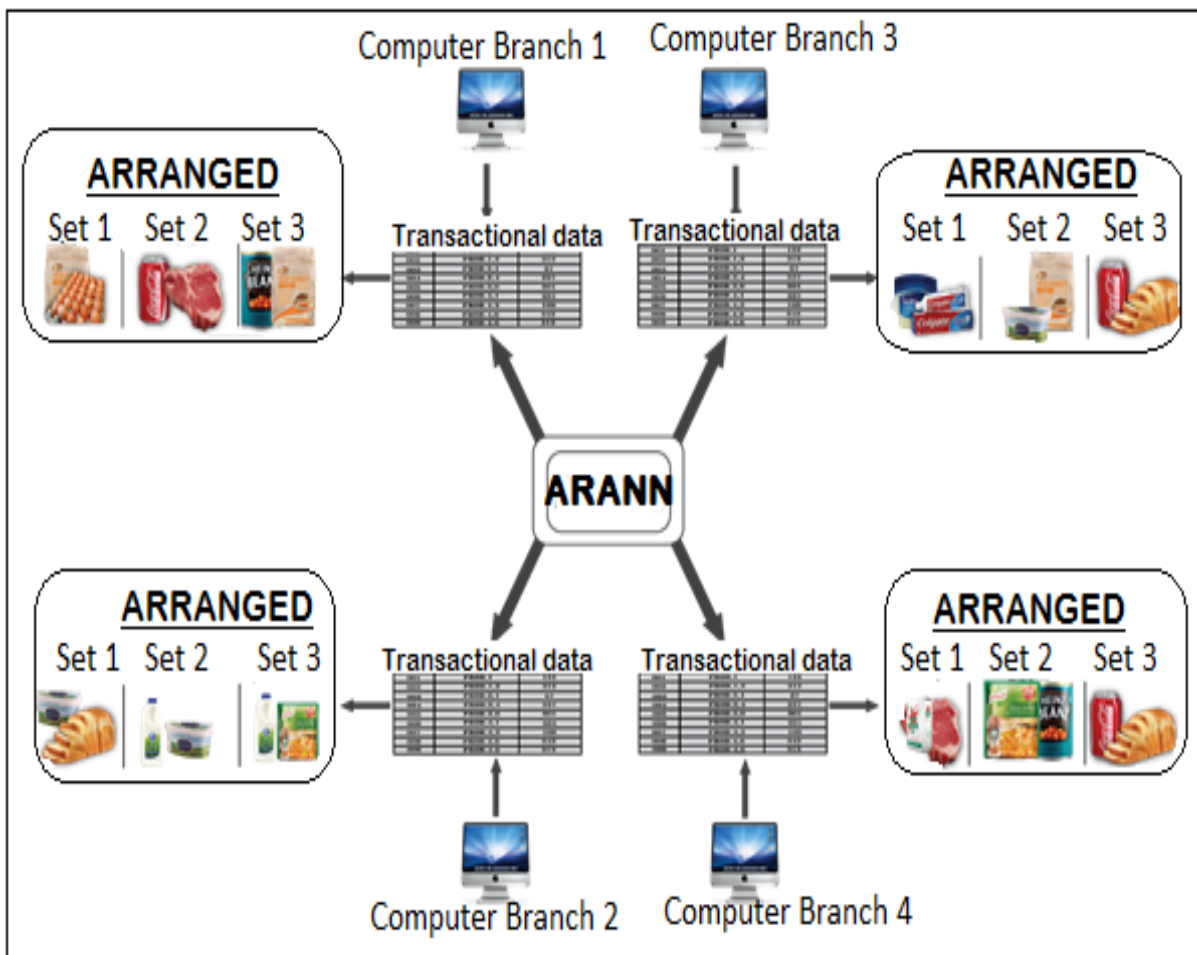


Figure 3.6: Intelligent Analytics-based Model for Four Branches

The products in the distributed retail enterprise are arranged according to the arrangement sets generated in Figure 3.6. For example, in branch 2 the following sets are generated;

{bread, margarine}, {milk, margarine} and {milk, soup}. These sets generated determine how the products are arranged branch by branch.

Table 3.9 and Table 3.10 show the transactional data used in branches 3 and 4 of a distributed enterprise. An analysis of the ARANN model is presented for each table.

Table 3.9: Market Basket Transactional Data for Branch 3 of a Retail Enterprise

Market-basket Transaction Data – Branch 3	
TID	ITEMS
T300	Colgate, Vaseline, Geisha, Margarine, Bread
T301	Margarine, Bread, Coke, Colgate, Vaseline
T302	Coke, Colgate, Chocolate, Bread, Sweets, Margarine
T303	Geisha, Colgate, Chocolate, Towel, Vaseline, Sweets
T304	Colgate, Vaseline, Sweets, Chocolate, Bread, Margarine, Coke

Even weights were applied to each corresponding input to avoid bias on products. This was obtained by dividing the count of a_union_b to the number of records within the data set, where *a*, and *b* are different products. The following ARANN activation was used:

>= 0.75 strongly connected products (Strongly accepted)

>= 0.65 moderately connected products (Accepted)

< 0.65 weakly connected products (Rejected).

Analysis of ARANN on Table 3.9

{Colgate, Vaseline} => {Bread}

$$\text{Support} = \frac{n(A \cup B)}{N} = \frac{3}{5} = 0.6$$

$$\text{Confidence} = \frac{n(A \cup B)}{n(A)} = \frac{3}{4} = 0.75$$

$$\begin{aligned} N_1 &= \text{Sup}w_1 + \text{Con}w_3 \\ &= (0.6 \times 0.6) + (0.75 \times 0.6) \\ &= 0.81 \end{aligned}$$

$$\begin{aligned} N_2 &= \text{Con}w_4 + \text{Sup}w_2 \\ &= (0.75 \times 0.6) + (0.6 \times 0.6) \\ &= 0.81 \end{aligned}$$

$$O_1 = \frac{1}{1 + e^{-N_1}} = \frac{1}{1 + e^{-0.81}} = 0.69$$

$$O_2 = \frac{1}{1 + e^{-N_2}} = \frac{1}{1 + e^{-0.81}} = 0.69$$

$$F = w5O_1 + w6O_2$$

$$= (0.6 \times 0.69) + (0.6 \times 0.69) = 0.83$$

$$\text{DoB} = \frac{1}{1+e^{-F}} = \frac{1}{1+e^{-0.83}} = 0.70$$

Product pattern => 0.70 >= 0.65

Therefore it is **moderately** connected and is **accepted**.

{Coke} => {Bread}

$$\text{Support} = \frac{3}{5} = 0.6$$

$$\text{Confidence} = \frac{3}{3} = 1.0$$

$$N1 = (0.6 \times 0.6) + (1.0 \times 0.6)$$

$$0.6)$$

$$= 0.96$$

$$N2 = (1.0 \times 0.6) + (0.6 \times$$

$$0.6)$$

$$= 0.96$$

$$O1 = \frac{1}{1+e^{-0.96}} = 0.72$$

$$O2 = \frac{1}{1+e^{-0.96}} = 0.72$$

$$F = w5O_1 + w6O_2$$

$$= (0.6 \times 0.72) + (0.6 \times 0.72) = 0.86$$

$$\text{DoB} = \frac{1}{1+e^{-0.86}} = 0.70$$

Product pattern => 0.70 >= 0.65

Therefore it is **moderately** connected and is **accepted**.

Table 3.10: Market Basket Transactional Data for Branch 4 of a Retail Enterprise

Market-basket Transaction Data – Branch 4	
TID	ITEMS
T400	Maize meal, Beef, Fish, Cooking oil, Soups, Bread, Coke
T401	Cooking oil, Beans, Beef, Soups, Maize meal
T402	Rice, Fish, Soups, Cooking oil, Bread
T403	Fruit, Coke, Bread, Milk, Chocolate, Soups
T404	Bread, Beef, Fruit, Coke, Sweets, Maize meal

Analysis of ARANN on Table 3.10

{Maize meal} => {Beef}

$$\text{Support} = \frac{3}{5} = 0.6$$

$$\text{Confidence} = \frac{3}{3} = 1.0$$

$$\begin{aligned} N1 &= (0.6 \times 0.6) + (1.0 \times 0.6) \\ &= 0.96 \end{aligned}$$

$$\begin{aligned} N2 &= (1.0 \times 0.6) + (0.6 \times 0.6) \\ &= 0.96 \end{aligned}$$

$$O1 = \frac{1}{1 + e^{-0.96}} = 0.72$$

$$O2 = \frac{1}{1 + e^{-0.4}} = 0.72$$

$$\begin{aligned} F &= w5O1 + w6O2 \\ &= (0.6 \times 0.72) + (0.6 \times 0.72) = 0.86 \end{aligned}$$

$$\text{DoB} = \frac{1}{1 + e^{-0.86}} = 0.70$$

Product pattern => 0.70 >= 0.65

Therefore it is **moderately** connected and is **accepted**.

{Chocolate} => {Soup}

$$\text{Support} = \frac{1}{5} = 0.20$$

$$\text{Confidence} = \frac{1}{1} = 1$$

$$\begin{aligned} N1 &= (0.20 \times 0.20) + (1 \times 0.20) \\ &= 0.24 \end{aligned}$$

$$\begin{aligned} N2 &= (1 \times 0.20) + (0.20 \times 0.20) \\ &= 0.24 \end{aligned}$$

$$O1 = \frac{1}{1 + e^{-0.24}} = 0.56$$

$$O2 = \frac{1}{1 + e^{-0.24}} = 0.56$$

$$F = (0.2 \times 0.56) + (0.2 \times 0.56) = 0.224$$

$$\text{DoB} = \frac{1}{1 + e^{-0.224}} = 0.56$$

Product pattern => 0.56 < 0.65

Therefore it is **weakly** connected and is **rejected**.

3.6 EVALUATION MECHANISM

The purpose of model evaluation is to assess the performance of the models so as to identify the best-performing model. To test the performance of the models, three sets were used. The confusion matrix shown in Table 3.11 was used to represent actual values and predictions.

Table 3.11: Confusion Matrix. Adapted from [4]

		Predicted	
		True	False
Actual	True	a	b
	False	c	d

$$\text{Error Rate} = \frac{b+c}{a+b+c+d} \quad (3.16)$$

where a is the number of sets predicted true when they are true, b is the number of sets predicted false when they are true, c is the number of sets predicted true when they are false and d is the number of sets predicted false when they are false. The error rate is then defined as shown in equation (3.16).

TPs are the number of patterns correctly detected, which is calculated using the following formula:

$$TP = \frac{d}{c+d}. \quad (3.17)$$

TNs are the non-matching patterns that were correctly rejected and are expressed as follows:

$$TN = \frac{a}{a+b}. \quad (3.18)$$

FPs are the proposed pattern matches that are incorrect and are expressed as follows:

$$FP = \frac{b}{a+b}. \quad (3.19)$$

FNs are the proposed patterns that were not detected correctly and are expressed as follows:

$$FP = \frac{c}{c+d}. \quad (3.20)$$

Thus, the accuracy (Acc %) is given as;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100. \quad (3.21)$$

3.7 CHAPTER SUMMARY

The chapter described the data collection process of the datasets for both the centralised and distributed retail enterprises. The data for the centralised retail enterprise was integrated from different sources into a single view. The data was integrated using data integration techniques such as data propagation, data federation and data consolidation. The integrated data was pre-processed following data preparation stages. The data for the distributed retail enterprise was also prepared and then transformed to the format accepted by the ARANN model.

The prepared data feeds the ARANN model in centralised and distributed retail enterprises. The ARANN model was built on AR complemented by ANN to improve the results of both methods. The pseudo-codes and the mathematical descriptions of the ARANN model were presented for the centralised and distributed retail enterprise. The chapter also presented scenarios for product arrangements on the shelves of centralised and distributed retail branches.

CHAPTER 4

EXPERIMENTAL EVALUATIONS AND RESULTS

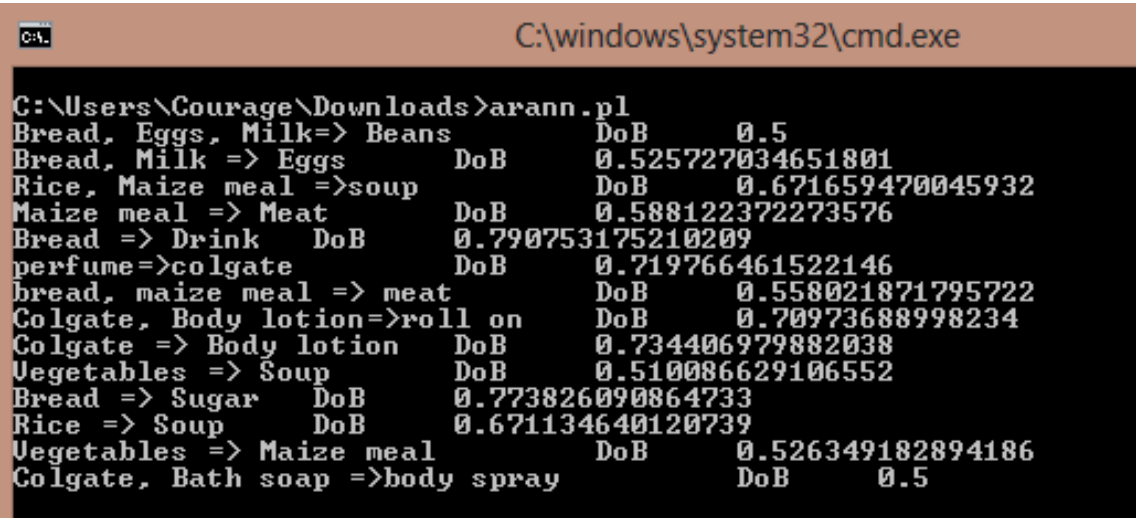
4.1 OVERALL EXPERIMENTAL SETUP

The ARANN model for both centralised and distributed retail enterprises was developed using the Perl programming language. The Perl programming language was used to implement the ARANN model using retail enterprises' transactional data observed under different ARANN activations. In distributed enterprises, the results from each branch showed how products were arranged per branch. On the other hand, in centralised enterprises the results were applied to all the branches of the enterprise.

4.2 EXPERIMENTAL EVALUATIONS FOR CENTRALISED ANALYTICS

4.2.1 ARANN Experimental Setup in Centralised Analytics

The datasets from different branches in different demographic groups of a centralised retail enterprise were collected from different computers. The data was stored in folders. A script was used to integrate the data from these different folders and it was completed in four seconds. After integrating the data, processes such as removing additional columns, rows and all unnecessary data items were performed in order to clean the data. This pre-processed data was fed into the ARANN model for the generation of product arrangement sets. The sample results using real-life and public datasets are presented in Figure 4.1 and Figure 4.2 respectively.



```
C:\windows\system32\cmd.exe

C:\Users\Courage\Downloads>arann.pl
Bread, Eggs, Milk=> Beans          DoB      0.5
Bread, Milk => Eggs                DoB      0.525727034651801
Rice, Maize meal => soup           DoB      0.671659470045932
Maize meal => Meat                 DoB      0.588122372273576
Bread => Drink                    DoB      0.790753175210209
perfume=>colgate                   DoB      0.719766461522146
bread, maize meal => meat          DoB      0.558021871795722
Colgate, Body lotion=>roll on     DoB      0.70973688998234
Colgate => Body lotion             DoB      0.734406979882038
Vegetables => Soup                DoB      0.510086629106552
Bread => Sugar                    DoB      0.773826090864733
Rice => Soup                      DoB      0.671134640120739
Vegetables => Maize meal          DoB      0.526349182894186
Colgate, Bath soap =>body spray  DoB      0.5
```

Figure 4.1: ARANN Rules on Real-life Data in Centralised Analytics

```

C:\windows\system32\cmd.exe

C:\Users\Courage\Downloads>arann.pl
Fish, Canned soup => Wine          DoB      0.641667569753749
Fish => Canned soup          DoB      0.735379217783688
Tea, Cookies =>Peanuts    DoB      0.607423172527644
Bread => Chocolate milk    DoB      0.725920215106361
Bread, Chocolate milk =>Tea      DoB      0.649979143433429
Bread => Chocolate milk    DoB      0.725920215106361
Beer => Tea                DoB      0.668689645578768
Beer => Chocolate milk    DoB      0.679430581365509
Wine => Beer              DoB      0.691360276196102
Canned soup => Bread      DoB      0.789840891751696
Orange juice => Bread    DoB      0.730205608690155
Peanuts, Bread => Canned soup  DoB      0.67837054417796
Tea, Bread => Orange juice  DoB      0.658936783975978

```

Figure 4.2: ARANN Rules on Public Data in Centralised Analytics

4.2.2 Experiment 1: Observations of ARANN with Varying Activations in Centralised Analytics

In this experiment, the following ARANN activations were used: $DoB < 60\%$, $60\% \geq DoB < 70\%$ and $DoB \geq 70\%$. The ARANN model rejects product arrangement sets where the DoB is less than 60%, accepts product arrangement sets where the DoB is between 60% and 69% and strongly accepts product arrangement sets where the DoB is above 69%. Equation (3.9) was used to determine the decision criteria to be applied to Table 4.1 and Table 4.2 of the ARANN model. In order to make the decision, the ARANN model compares the DoB value generated against the ARANN activations. Managers use the decision to determine how products are to be arranged across all the branches of a centralised enterprise.

Table 4.1: Real-life ARANN Results for All Centralised Branches

Dataset All Branches	Patterns Generated	DoB	ARANN Cooperative Decision with	
			$60 \geq DoB < 70$	$DoB \geq 70$
	Rice, Maize meal => Soup	67%	Accepted	N/A
	Bread => Drink	79%	N/A	Strongly Accepted
	Perfume => Colgate	72%	N/A	Strongly Accepted
	Colgate, Body lotion => Roll-on	71%	N/A	Strongly Accepted
	Colgate => Body lotion	73%	N/A	Strongly Accepted
	Bread => Sugar	77%	N/A	Strongly Accepted
	Rice => Soup	67%	Accepted	N/A

Table 4.1 shows the optimal product combination sets for all the branches of a centralised retail enterprise. As indicated in Table 4.1, the ARANN model has accepted two product arrangement sets and strongly accepted five product arrangement sets that met the model's requirements. The accepted sets are: ({Rice, Maize meal, Soup}, {Rice, Soup}). These product arrangement sets were accepted because the DoB values are greater than 59%. This means that there is a moderate connection between these product combinations. For instance, it makes logical sense for retail shops to arrange rice and maize meal next to each other, as these products highly supplement each other.

The strongly accepted sets are: ({Bread, Drink}, {Perfume, Colgate}, {Colgate, Body lotion, Roll on}, {Colgate, Body lotion}, {Bread, Sugar}). These product arrangement sets were strongly accepted because the DoB values are greater than 69%. This means that there is a strong connection between the product combinations. For instance, it makes sense to arrange toiletry products such as roll-on deodorant, combined with perfume and toothpaste; it makes sense for these products to be arranged on the same shelf, as they highly complement each other. It also makes logical sense for the retail shops to arrange bread, sugar and drinks next to each other, as these products highly complement each other.

The products in all the branches of a centralised retail enterprise are arranged on the shelves uniformly according to either accepted or strongly accepted sets. The decision is made at the central or head office.

Table 4.2: Public ARANN Results for All Centralised Branches

Dataset All Branches	Patterns Generated	DoB	ARANN Cooperative Decision with	
			60>=DoB<70	DoB>=70
	Fish, Canned soup => Wine	64%	Accepted	N/A
	Fish => Canned soup	74%	N/A	Strongly accepted
	Tea, Cookies => Peanuts	61%	Accepted	N/A
	Bread => Chocolate milk	73%	N/A	Strongly accepted
	Bread, Chocolate milk => Tea	65%	Accepted	N/A
	Beer => Tea	67%	Accepted	N/A
	Beer => Chocolate milk	68%	Accepted	N/A
	Wine => Beer	69%	Accepted	N/A
	Canned soup => Bread	79%	N/A	Strongly accepted
	Orange juice => Bread	73%	N/A	Strongly accepted
	Peanuts, Bread => Canned soup	68%	Accepted	N/A
	Tea, Bread => Orange juice	66%	Accepted	N/A

According to Table 4.2, the ARANN model accepted the following eight product arrangement sets: ({Fish, Canned soup, Wine}, {Tea, Cookies, Peanuts}, {Bread, Chocolate milk, Tea}, {Beer, Tea}, {Beer, Chocolate milk}, {Wine, Beer}, {Peanuts, Bread, Canned soup}, {Tea, Bread, Orange juice}). These product arrangement sets were accepted because the DoB values were greater than 59% and less than 70%. These are products that are moderately connected.

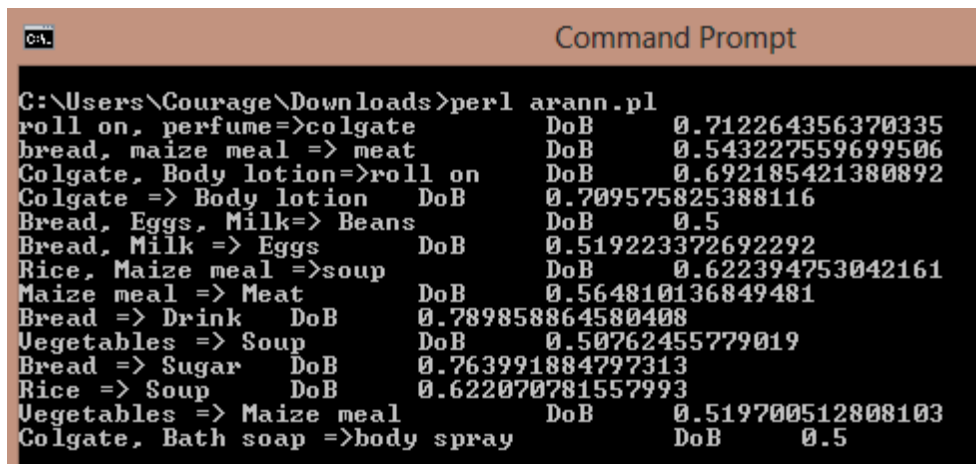
The strongly accepted product arrangement sets are: ({Canned soup, Bread}, {Orange juice, Bread}). These were strongly connected products and were strongly accepted because the DoB was greater than 69%. These results suggest a strong connection of bakery products with drinks and potages. This means that bakery products need to be kept closer to the drinks and potages to improve the sales of these products.

It remains the responsibility of the centralised retail enterprises' managers or decision-makers to make the decision on adopting either moderately connected or strongly connected products, depending on the market's competitiveness and profit levels.

4.3 EXPERIMENTAL EVALUATIONS FOR DISTRIBUTED ANALYTICS

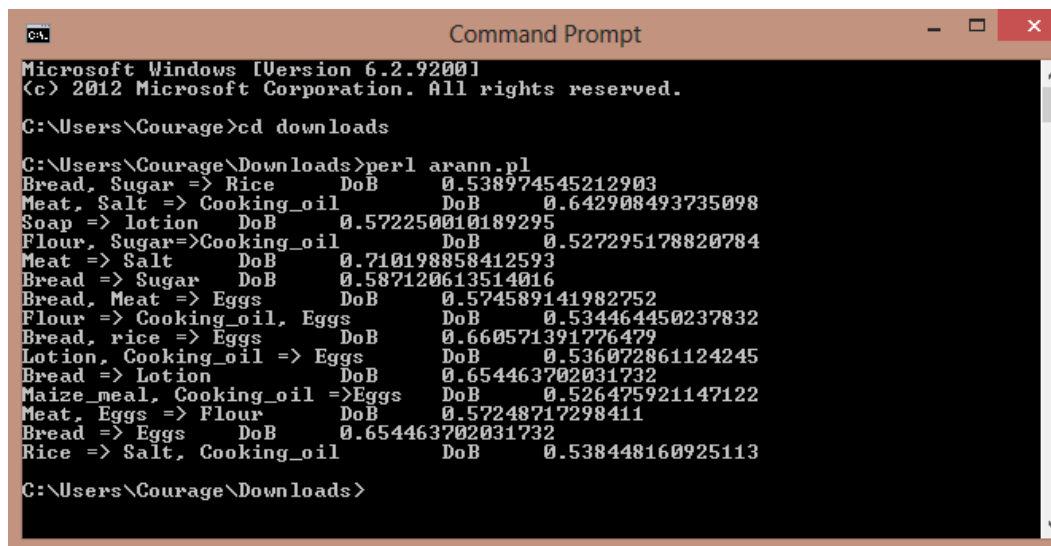
4.3.1 ARANN Experimental Setup in Distributed Analytics

In this experiment, Perl programming language was used to develop and implement the ARANN model. The processed transactional data was then fed into the ARANN model branch by branch. Notepad was used as the text editor and results were displayed through the command prompt. Figure 4.3 to Figure 4.8 show sample product combination sets generated by the ARANN model using both real-life and public datasets.



```
Command Prompt
C:\Users\Courage\Downloads>perl arann.pl
roll on, perfume=>colgate      DoB      0.712264356370335
bread, maize meal => meat     DoB      0.543227559699506
Colgate, Body lotion=>roll on DoB      0.692185421380892
Colgate => Body lotion      DoB      0.709575825388116
Bread, Eggs, Milk=> Beans    DoB      0.5
Bread, Milk => Eggs         DoB      0.519223372692292
Rice, Maize meal => soup     DoB      0.622394753042161
Maize meal => Meat          DoB      0.564810136849481
Bread => Drink              DoB      0.789858864580408
Vegetables => Soup          DoB      0.50762455779019
Bread => Sugar              DoB      0.763991884797313
Rice => Soup                DoB      0.622070781557993
Vegetables => Maize meal    DoB      0.519700512808103
Colgate, Bath soap =>body spray DoB      0.5
```

Figure 4.3: ARANN Rules on Real-life Data for Branch A



```
Command Prompt
Microsoft Windows [Version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.
C:\Users\Courage>cd downloads
C:\Users\Courage\Downloads>perl arann.pl
Bread, Sugar => Rice          DoB      0.538974545212903
Meat, Salt => Cooking_oil    DoB      0.642908493735098
Soap => lotion               DoB      0.572250010189295
Flour, Sugar=>Cooking_oil    DoB      0.527295178820784
Meat => Salt                 DoB      0.710198858412593
Bread => Sugar               DoB      0.587120613514016
Bread, Meat => Eggs          DoB      0.574589141982752
Flour => Cooking_oil, Eggs   DoB      0.534464450237832
Bread, rice => Eggs          DoB      0.660571391776479
Lotion, Cooking_oil => Eggs  DoB      0.536072861124245
Bread => Lotion              DoB      0.654463702031732
Maize_meal, Cooking_oil =>Eggs DoB      0.526475921147122
Meat, Eggs => Flour          DoB      0.57248717298411
Bread => Eggs                DoB      0.654463702031732
Rice => Salt, Cooking_oil    DoB      0.538448160925113
C:\Users\Courage\Downloads>
```

Figure 4.4: ARANN Rules on Real-life Data for Branch B


```

C:\Users\Courage\Downloads>perl arann.pl
Fish, Canned soup => Wine      DoB      0.641667569753749
Fish => Canned soup      DoB      0.735026920277072
Tea, Cookies =>Peanuts    DoB      0.609803423913402
Bread => Chocolate milk  DoB      0.729239472797766
Bread, Chocolate milk =>Tea    DoB      0.643679847969827
Bread => Chocolate milk  DoB      0.729239472797766
Beer => Tea              DoB      0.667485188255028
Beer => Chocolate milk  DoB      0.686898451603118
Wine => Beer             DoB      0.686284674024776
Canned soup => Bread     DoB      0.786985620245376
Orange juice => Bread    DoB      0.728435777511132
Peanuts, Bread => Canned soup DoB      0.674095060864138
Tea, Bread => Orange juice DoB      0.650143937955844

```

Figure 4.5: ARANN Rules on Public Dataset for Branch C

```

C:\Users\Courage\Downloads>perl arann.pl
Fish, Canned soup => Wine      DoB      0.642522737666967
Fish => Canned soup      DoB      0.737118094340529
Tea, Cookies =>Peanuts    DoB      0.605752517642561
Bread => Chocolate milk  DoB      0.723933904731428
Bread, Chocolate milk =>Tea    DoB      0.657363780809137
Bread => Chocolate milk  DoB      0.723933904731428
Beer => Tea              DoB      0.670938314096633
Beer => Chocolate milk  DoB      0.673090039546977
Wine => Beer             DoB      0.697645354712109
Canned soup => Bread     DoB      0.79423499982228
Orange juice => Bread    DoB      0.73335094248875
Peanuts, Bread => Canned soup DoB      0.683780516307824
Tea, Bread => Orange juice DoB      0.668865796107564

C:\Users\Courage\Downloads>

```

Figure 4.6: ARANN Rules on Public Dataset for Branch D

4.3.2 Experiment 2: Observations of ARANN with Varying Activations in Distributed Analytics

In this experiment, equation (3.15) was used to determine the decision criteria to be applied to Table 4.3 through Table 4.6 of the ARANN model. The ARANN model accepts product combination sets defined in equation (3.15) and uses the following ARANN activations: $DoB < 60\%$, $60\% \geq DoB < 70\%$ and $DoB \geq 70\%$. The ARANN model rejects product arrangement sets where the DoB is less than 60% and accepts product arrangement sets between 60% and 69%, while those with a DoB greater than or equal to 70% are strongly accepted. To make the decision, the ARANN model compares the DoB value generated against the ARANN activations. Managers use the decision to determine how products are to be arranged in each branch.

Table 4.3: Real-life ARANN Results for Branch 1 in Demographic Group A

Dataset	Patterns Generated	DoB	ARANN Cooperative Decision with	
Branch 1			60>=DoB<70	DoB>=70
	Roll-on, perfume=>Colgate	71%	N/A	Strongly accepted
	Colgate, Body lotion=>roll on	69%	Accepted	N/A
	Colgate => Body lotion	71%	N/A	Strongly accepted
	Bread, Milk => Eggs	70%	N/A	Strongly accepted
	Rice, Maize meal =>soup	62%	Accepted	N/A
	Bread => Drink	79%	N/A	Strongly accepted
	Bread => Sugar	76%	N/A	Strongly accepted

As shown in Table 4.3, five product combination sets were strongly accepted using the ARANN activation of $DoB \geq 70$; these include: {Roll-on, Perfume => Colgate}, {Colgate => Body lotion}, {Bread, Milk => Eggs}, {Bread => Drink} and {Bread => Sugar}. This means that there is a strong connection between the product combinations. For instance, it makes logical sense for the retail shops to arrange bread, milk and eggs next to each other, as these products highly complement each other. The same applies to toiletry products such as roll-on deodorant, combined with perfume and toothpaste; it makes sense for these products to be arranged on the same shelf, as they highly complement each other.

Figure 4.3 also shows that only two product combinations were accepted using the ARANN activation of $60 \geq DoB < 70$; these are: {Colgate, Body lotion=>Roll on} and {Rice, Maize meal => Soup}. These findings suggest that these two set of accepted product combinations have a moderate connection with each other. For instance, it is possible for a customer to buy soup in the same basket as maize meal and rice, depending on the needs of each customer. Based on the above results, the choice is thus left to every retail enterprise to adopt either moderately or strongly connected products, depending on their market competitiveness and profit levels.

Table 4.4: Real-life ARANN Results for Branch 2 in Demographic Group B

Dataset	Patterns Generated	DoB	ARANN Cooperative Decision with	
			$60 \geq \text{DoB} < 70$	$\text{DoB} \geq 70$
Branch 2				
	Meat, Salt => Cooking_oil	0.64	Accepted	N/A
	Meat => Salt	0.71	N/A	Strongly accepted
	Bread, rice => Eggs	0.66	Accepted	N/A
	Bread => Lotion	0.65	Accepted	N/A
	Bread => Eggs	0.65	Accepted	N/A

Applying the ARANN activation of $\text{DoB} \geq 70$, only one product combination set {Meat => Salt} was strongly accepted, as shown in Table 4.4. These results suggest a strong connection between meat and salt. This means that salt needs to be shelved closer to the meat refrigerators to improve the sales of both products. According to Table 4.4, four product combination sets were accepted using the ARANN activation of $60 \geq \text{DoB} < 70$; these are: {Meat, Salt => Cooking oil}, {Bread, Rice => Eggs}, {Bread => Lotion}; and {Bread => Eggs}. The results of the accepted product combination sets suggest that these products are moderately connected. For instance, by chance a customer might buy bread and lotion in the same basket, and there is a fair chance that a customer will buy cooking oil in the same basket with meat and salt. In light of the above findings it is therefore up to the retail enterprise's decision-makers to adopt either moderately or strongly connected products, depending on the market competitiveness and profit levels.

Table 4.5: Public Data ARANN Results for Branch 3 in Demographic Group C

Dataset	Patterns Generated	DoB	ARANN Cooperative Decision with	
			$60 \geq \text{DoB} < 70$	$\text{DoB} \geq 70$
Branch 3	Fish, Canned soup => Wine	0.64	Accepted	N/A
	Fish => Canned soup	0.74	N/A	Strongly accepted
	Tea, Cookies =>Peanuts	0.61	Accepted	N/A
	Bread => Chocolate milk	0.73	N/A	Strongly accepted
	Bread, Chocolate milk =>Tea	0.64	Accepted	N/A
	Beer => Tea	0.67	Accepted	N/A
	Beer => Chocolate milk	0.69	Accepted	N/A
	Wine => Beer	0.69	Accepted	N/A
	Canned soup => Bread	0.79	N/A	Strongly accepted
	Orange juice => Bread	0.73	N/A	Strongly accepted
	Peanuts, Bread => Canned soup	0.67	Accepted	N/A
	Tea, Bread => Orange juice	0.65	Accepted	N/A

Table 4.5 shows that only four product combination sets were strongly accepted using the ARANN activation of $\text{DoB} \geq 70$; these are: {Fish =>Canned soup}, {Bread => Chocolate milk}, {Canned soup => Bread} and {Orange juice => Bread}. These product combination sets are considered to be strongly connected. Table 4.5 also shows that eight product arrangement sets were accepted using the ARANN activation of $60 \geq \text{DoB} < 70$. The following are examples of the accepted product arrangement sets: {Fish, Canned soup => Wine}, {Tea, Cookies =>Peanuts} and {Wine => Beer}. These product arrangement sets comprise moderately connected products. Every retail enterprise is left with the choice to adopt either moderately or strongly connected products, depending on the market competitiveness and profit levels.

Table 4.6: Public data ARANN results for Branch 4 in Demographic Group D

Dataset	Patterns Generated	DoB	ARANN Cooperative Decision with	
			$60 \geq \text{DoB} < 70$	$\text{DoB} \geq 70$
Branch 4	Fish, Canned soup => Wine	0.64	Accepted	N/A
	Fish => Canned soup	0.74	N/A	Strongly accepted
	Tea, Cookies =>Peanuts	0.61	Accepted	N/A
	Bread => Chocolate milk	0.72	N/A	Strongly accepted
	Bread, Chocolate milk =>Tea	0.66	Accepted	N/A
	Beer => Tea	0.67	Accepted	N/A
	Beer => Chocolate milk	0.67	Accepted	N/A
	Wine => Beer	0.70	N/A	Strongly accepted
	Canned soup => Bread	0.80	N/A	Strongly accepted
	Orange juice => Bread	0.73	N/A	Strongly accepted
	Peanuts, Bread => Canned soup	0.68	Accepted	N/A
	Tea, Bread => Orange juice	0.67	Accepted	N/A

Table 4.6 shows that only five product combination sets were strongly accepted using the ARANN activation of $\text{DoB} \geq 70$. Some of the examples of the strongly accepted sets are: {Bread => Chocolate milk} and {Fish => Canned soup}, and the results suggest that the product combinations are strongly connected. Table 4.6 also indicates that seven product combination sets were accepted using the ARANN activation of $60 \geq \text{DoB} < 70$. Some of the examples of the accepted product combination sets are: {Fish, Canned soup => Wine}, {Orange juice => Bread} and {Tea, Bread => Orange juice}. These findings suggest that the product combinations are moderately connected. The decision-makers of retail enterprises are left with the choice to adopt either moderately or strongly connected products, depending on their market competitiveness and profit levels.

4.4 PERFORMANCE EVALUATION OF DISTRIBUTED AND CENTRALISED ANALYTICS

4.4.1 Experiment 3: Comparison of ARANN in Terms of Memory and Time Usages

The current study compares the performance of the ARANN model in a distributed retail enterprise against a centralised retail enterprise. In the distributed retail enterprise, a computer was used to represent a branch and the time (wall clock times) taken by the ARANN model to generate product combination sets was observed. Figure 4.7a shows raw integration time. Figure 4.7b shows the time of response (ToR) taken by the ARANN model to integrate a number of records from various workstations. Figure 4.7c shows the ToR taken by the ARANN model to generate product combination sets in distributed and centralised retail enterprises. Figure 4.7d shows the ToR taken by the ARANN model to generate product combination sets across different data sizes.

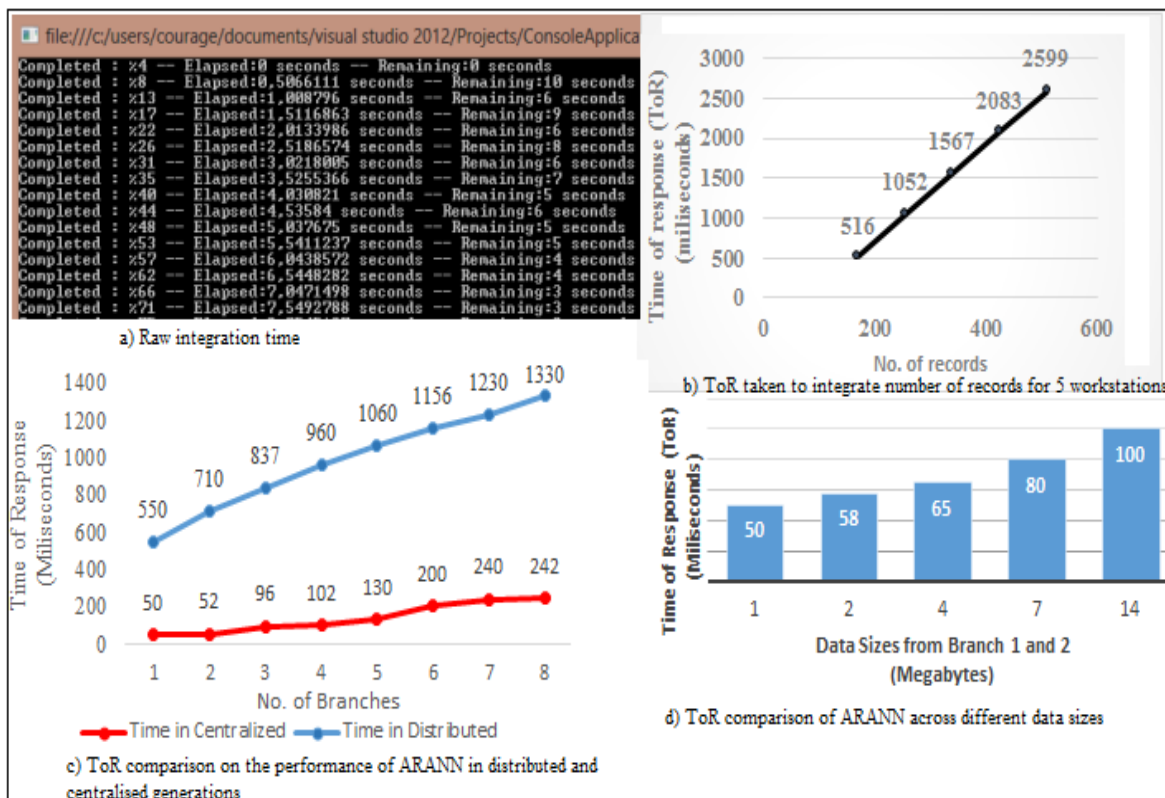


Figure 4.7: Comparison of the Performance of ARANN in Retail Enterprises

From the experiment conducted, it was observed that the ARANN model performs faster in distributed retail enterprises than in centralised retail enterprises (see Figure 4.7c). The ARANN model takes more time to generate product combination sets in a centralised retail enterprise than in a distributed retail enterprise. The ToR to integrate data depends on the

number of records being integrated. The more the records, the more time is needed to integrate those records. This was observed in Figure 4.7b. In addition, the performance time taken by the ARANN model depends on the size of the data set being used. The ARANN model’s performance is affected by the size of the data set, as shown in Figure 4.7d.

4.4.2 Experiment 4: Benchmarking ARANN with Classical Models

A) Benchmarking ARANN with Classical Models in Centralised Analytics

AR and ANN models were compared to the ARANN model. The AR model was implemented using Waikato Software for Knowledge Analysis (WEKA) version 3.7. The data was fed into the WEKA and the minimum support was adjusted. A simple ANN model which cannot learn was implemented. The weights of the ANN model were set by using heuristics thresholds. Table 4.7 shows the number of product combination sets generated by each model, correctly detected patterns, incorrectly detected patterns and the accuracy of detection under different transactional datasets. It indicates the accuracy of a model’s algorithm in generating patterns. The correctly detected patterns were calculated by finding the sum of equation (3.17) and equation (3.18), while the incorrectly detected patterns were calculated by finding the sum of equations (3.19) and equation (3.20). Equation (3.21) along with the confusion matrix was used to calculate and determine each model’s accuracy rate. From the results in Table 4.7, it can be observed that ARANN has a higher accuracy rate compared to the modelled classical models which cannot learn.

Table 4.7: Quantitative Comparison of Three Models on Retail Datasets

Dataset	Algorithms	No. of Patterns	Correctly Detected (TP, TN)	Incorrectly Detected (FP, FN)	Accuracy
Real life	AR	6	83%	17%	83%
	ANN	6	67%	33%	67%
	ARANN	5	83%	17%	83%
Public	AR	4	75%	25%	75%
	ANN	4	75%	25%	75%
	ARANN	3	100%	0%	100%

As shown in Table 4.7, the ARANN model has the highest accuracy rate compared to other classical models. The ARANN model has 83% and 100% accuracy rates, while the AR model has 83% and 75%, with the ANN being the least accurate (67% and 75%), in both the real-life and public datasets. Based on the accuracy results and correctly detected product combination sets, the ARANN model seems to be the most reliable and effective model when it comes to product combination arrangements in centralised retail shops for sales maximisation.

B) Benchmarking ARANN with Classical Models in Distributed Analytics

Table 4.8 shows the error rate of the individual AR and ANN models against the ARANN model. Equation (3.16) was used to determine the error rate of each model. The column “No. of patterns” indicates the number of product combination sets evaluated. The column “Correctly classified sets” is composed of product combination sets that the ARANN model predicted as true when they were actually true (a) and sets predicted as false when they were actually false (d), as shown in Table 3.10 (the confusion matrix). The column “Incorrectly classified sets” is composed of product arrangement sets that the ARANN model predicted as false when they were actually true (b) and sets predicted as true when they were false (c). Randomly generated product combination sets were used to evaluate the performance of the three models. For example, in Branch A (real-life dataset), of the 10 rules that were used in AR, five rules were predicted as true when they were actually true (a); two rules were predicted as false when actually false (d); three rules were predicted as true when actually false (c) and no rules were predicted as false when actually true (b).

Table 4.8: Quantitative Evaluations of the Cooperative Model in Distributed Branches

Dataset	Algorithms	No. of Patterns	Correctly Classified sets (a, d)	Incorrectly Classified sets (b, c)	Error Rate
Real life	AR	10	7	3	30%
Branch 1 (66 records)	ANN	10	6	4	40%
	ARANN	6	5	1	17%
Branch 2 (66 records)	AR	10	8	2	20%
	ANN	10	8	2	20%
	ARANN	7	6	1	14%
Public	AR	10	8	2	20%
Branch 3 (200 records)	ANN	10	6	4	40%
	ARANN	6	5	1	17%
Branch 4 (200 records)	AR	10	8	2	20%
	ANN	10	7	3	30%
	ARANN	8	6	2	25%

From the results displayed in Table 4.8, it is clear that the ARANN model has a lower error rate compared to the individual classical models.

4.5 CHAPTER SUMMARY

In this chapter different experimental evaluations were conducted on the ARANN model in the centralised and distributed environment. The Perl programming language was used to develop and implement the ARANN model. The following ARANN activations were used: $DoB < 60\%$, $60\% \leq DoB < 70\%$ and $DoB \geq 70\%$. The ARANN model rejects arrangement sets where the DoB is less than 60% and accepts arrangement sets between 60% and 69%, while those with a DoB greater than or equal to 70% are strongly accepted. The manager makes decisions on how the products can be arranged in each branch, depending on the results generated by each branch.

The experiments on ARANN were conducted using real-life and public datasets. The real-life datasets were collected from different branches in different demographics of a retail

enterprise. Datasets for centralised retail enterprises were integrated and cleaned before being fed into the ARANN model. In distributed retail enterprises, there was no need for data integration. The data was only cleaned and fed into the ARANN model for the generation of product arrangement sets. The generated product arrangement sets in the distributed retail enterprise were applied to individual branches, while in centralised retail enterprises, the product arrangement sets were applied uniformly across all the branches.

The ARANN model's performance was also compared in distributed and centralised analytics. In the experiment, a computer was used to represent a branch and the time taken by the ARANN to generate the patterns was noted. It was observed that the ARANN model performed faster in distributed and centralised retail enterprises.

CHAPTER 5

CONCLUDING REMARKS

5.1 CONCLUSION

In this study a cooperative ARANN model was developed and implemented for both centralised and distributed retail enterprises for the generation of product arrangement sets. This ARANN model was developed on the AR and ANN models to complement each other cooperatively. The idea was for the ARANN model to take the strengths of these individual models in order to improve product arrangement sets in centralised and distributed retail environments. The product arrangement sample sets generated from the ARANN model were presented in Figure 4.1 and Figure 4.2 in centralised retail enterprises and for distributed retail enterprises were presented in Figure 4.3 to Figure 4.6. This ARANN model for centralised enterprises was implemented using the algorithm shown in Table 3.6, while for distributed enterprises it was implemented using the algorithm shown in Table 3.8. In the centralised retail enterprise the data was integrated as shown in Figure 3.3. On the other hand, there was no need for data integration in distributed enterprises, as shown in Figure 3.5 of the ARANN model. The Perl programming language was used to develop and implement the ARANN model. In centralised retail enterprises, the program was programmed according to the mathematical descriptions presented in equation (2.1), equation (2.2), and equation (3.1) to equation (3.7), while in distributed retail enterprises, equation (2.1), equation (2.2) and equation (3.8) to equation (3.15) were used.

The ARANN model was tested using transaction data from the real-life and public environment. The real-life datasets were collected within different demographic groups in South Africa, while the public dataset was downloaded from the internet. The data used on the ARANN model was structured as shown in Table 3.1 to Table 3.5. The ARANN model's accuracy rate was tested in centralised retail enterprises and the results are shown in Table 4.7. The ARANN model was noted to have a higher accuracy rate than the individual created models. The ANN model used cannot learn. The error rates of the ARANN model in distributed retail enterprises was compared, as shown in Table 4.8, for both real-life and public datasets. From the table it can be seen that the ARANN model has a lower error rate compared to the modelled classical model.

Tables 4.1 to 4.6 documents experimental results of the cooperative ARANN model showing improved product arrangement sets. These product arrangement sets can be used to arrange products on shop shelves in order to improve sales within retail enterprises.

5.2 MANAGERIAL IMPLICATIONS

The research observed some improved product arrangement sets that can be discovered from retail enterprise transactional datasets. The findings have some managerial implications for retail enterprises' decision-makers that can improve retail sales. Retail enterprise management can implement different marketing strategies based on the results generated from the transactional data. The results obtained can assist retail management to know the levels of product associations, such as weak, moderate or strong. This can help retail managers implement strategies such as the best product arrangement on shop shelves, product promotions and next product sales prediction.

The results generated from transactional data can reduce company expenses such as advertising and minimise promoting the wrong product and poor product arrangement, which can influence customer buying habits.

There are some differences in how the ARANN model is developed and implemented in centralised and distributed retail enterprises. Managers should consider some observations of the ARANN model in distributed analytics:

- ◆ The proposed ARANN model retains complete control of product combination set generation.
- ◆ The product combination sets generated by the ARANN model show a lower error rate (Table 4.8).
- ◆ The ARANN model reveals the real buying habits of each branch.
- ◆ The model reduces the risk of passing misleading results to all branches (Table 4.3 to Table 4.6).
- ◆ There is no need for data integration (Figure 3.5).

On the other hand, observations of the ARANN model to be considered by managers are:

- ◆ This ARANN model exercises full control of product combination set generation.
- ◆ The same cooperative ARANN model results are used for all branches (Table 4.1).
- ◆ The implementation/software of this model runs in a single process.

- ◆ There is a single point of control of product placement patterns for all branches.
- ◆ However, a single point of failure to generate correct product combination sets could mislead all branches.

5.3 FUTURE WORK

In future, the author wishes to:

- ◆ Improve on ARANN performance by considering nature-inspired algorithms;
- ◆ Investigate a standard method of selecting the threshold;
- ◆ Use big data transactional datasets; and
- ◆ Integrate a sophisticated learning algorithm into ARANN and deploy it into wholesale enterprises.
- ◆ Explore the implementation of ARANN using open-source data mining frameworks.

The strategy and observations in the current study are, therefore, good for addressing challenges in the competitive environment.

REFERENCES

- [1] C. White, "Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise," TDWI Research Report, New York, 2005.
- [2] J. Lara, D. Lizcano, M. Martínez and J. Pazos, "Data Preparation for KDD Through Automatic Reasoning Based on Description Logic," *Information Systems*, vol. 44, pp. 54-72, 2014.
- [3] 2010. [Online]. Available: http://www.informatics.buu.ac.th/~ureerat/321641/Weka/Data%20Sets/supermarket/supermarket_basket_transactions_2005.arff. [Accessed 21 January 2014].
- [4] I. Witten, E. Frank and M. Hall, "Introduction to Weka," in *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann, 2011, pp. 403-440.
- [5] "Treasury.gpg.gov.za," 2012. [Online]. Available: <http://www.treasury.gpg.gov.za/Documents/QB1%20The%20Retail%20Industry%20on%20the%20Rise.pdf>. [Accessed 5 November 2015].
- [6] "Atkearney.com," [Online]. Available: <https://www.atkearney.com/documents/10192/6437503/Retail+in+Africa.pdf/b038891c-0e81-4379-89bb-b69fb9077425>. [Accessed 8 October 2015].
- [7] "South African Edition," 2012. [Online]. Available: <https://www.pwc.co.za/en/assets/pdf/retail-and-consumer-products-outlook-2012-2016.pdf>. [Accessed 12 November 2015].
- [8] R. Rolee, D. Woodward, A. Ligthelm and P. Guimaraes, "The Viability of Informal Micro-enterprise in South Africa," in *Entrepreneurship in Africa*, New York, 2010.
- [9] P. Trkman, K. McCormack, M. de Oliveira and M. Ladeira, "The Impact of Business Analytics on Supply Chain Performance," *Decision Support Systems*, vol. 49, no. 3, pp. 318-327, 2010.
- [10] R. Kohavi, N. Rothleder and E. Simoudis, "Emerging Trends in Business Analytics," *Communications of the ACM*, vol. 45, no. 8, pp. 45-48, 2002.
- [11] C. Velu, S. Madnick and M. van Alstyne, "Centralizing Data Management with Considerations of Uncertainty and Information-Based Flexibility," *Journal of Management Information Systems*, vol. 30, no. 3, pp. 179-212, 2013.

- [12] W. Abbas, N. Ahmad and N. Zaini, "Discovering Purchasing Pattern of Sport Items Using Market Basket Analysis," in *2013 International Conference on Advanced Computer Science Applications and Technologies*, Kuching, 2013.
- [13] A. Haug, F. Zachariassen and D. van Liempd, "The Costs of Poor Data Quality," *Journal of Industrial Engineering and Management*, vol. 4, no. 2, pp. 168-193, 2011.
- [14] "Hungary retail sales down in June," *Regional Today*, 2013.
- [15] "Gap Cost Key Categories Billions," *Furniture/Today*, vol. 26, no. 1, p. 14, 2001.
- [16] L. Briggs, "Case Study," *Business Intelligence Journal*, vol. 16, no. 4, pp. 39-41, 2011.
- [17] N. Stoodley, "Democratic Analytics: A Campaign to Bring Business Intelligence to the People," *Business Intelligence Journal*, vol. 17, no. 1, pp. 7-12, 2012.
- [18] R. Tront and S. Hoffman, "Pervasive Business Intelligence and the Realities of Excel," *Business Intelligence Journal*, vol. 16, no. 4, pp. 8-14, 2011.
- [19] S. Akbay, H. Fryman, D. Stodder and J. Taylor, "Experts' Perspective," *Business Intelligence Journal*, vol. 16, no. 4, pp. 19-26, 2011.
- [20] M. Berry and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, 2nd ed., Indiana: Wiley, 2004.
- [21] V. Sigurdsson, H. Saevarsson and G. Foxall, "Brand Placement and Consumer Choice: An In-store Experiment," *Journal of Applied Behavior Analysis*, vol. 42, no. 3, pp. 741-745, 2009.
- [22] D. Pyle, *Data Preparation for Data Mining*, San Francisco: Morgan Kaufmann Publishers, 1999.
- [23] I. K. El-Far and J. A. Whittaker, "Model-based Software Testing," *Encyclopedia on Software Engineering (edited by Marciniak)*, pp. 1-22, 2001.
- [24] M. Hambaba, "Intelligent Hybrid System for Data Mining," in *Computational Intelligence for Financial Engineering, Proceedings of the IEEE/IAFE 1996 Conference*, New York, 1996.
- [25] I. Song, *Data Warehouse*, US: Springer, 2009, pp. 657-658.
- [26] D. Resnik, "What is Ethics in Research & Why is it Important?," [Online]. Available: <http://www.niehs.nih.gov/research/resources/bioethics/whatis/>. [Accessed 26 October 2011].

- [27] "What is Research Ethics?," [Online]. Available: <http://research-ethics.net/introduction/what/>. [Accessed 26 October 2011].
- [28] L. Charlet, M. Annie and D. Kumar, "Market Basket Analysis for a Supermarket Based on Frequent Itemset Mining," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 5, 2012.
- [29] R. Blattberg, B. Kim and S. Neslin, "Database Marketing: Analyzing and Managing Customers," in *Marketing Basket Analysis*, New York, Springer, 2008, pp. 339-351.
- [30] M. Dhanabhakyaam and M. Punithavalli, "An Efficient Market Basket Analysis Based on Adaptive Association Rule Mining with Faster Rule Generation Algorithm," *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, vol. 1, no. 3, pp. 105-110, 2013.
- [31] K. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Amsterdam Boston: Morgan Kaufman, 2001.
- [32] M. Svetina and J. Zupančič, "How to Increase Sales in Retail with Market Basket Analysis," *Systems Integration*, pp. 418-428, 2005.
- [33] M. Redlon, "A SAS Market Basket Analysis Macro: The "Poor Man's Recommendation Engine"," in *SUGI 28*, Eden Prairie, Minnesota, 2003.
- [34] L. Cavique, "Next-Item Discovery in the Market Basket Analysis," in *Artificial Intelligence*, Covilha, 2005.
- [35] R. Agrawal, T. Imieliński and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, New York, 1993.
- [36] K. Cios, W. Pedrycz, R. Swiniarski and L. Kurgan, *Data Mining a Knowledge Discovery*, New York: Springer, 2007.
- [37] Q. Zhao and S. Bhowmick, "Association Rule Mining: A Survey," Technical Report, CAIS, Nanyang Technological University, Singapore, 2003.
- [38] H. Liu, B. Su and B. Zhang, "The Application of Association Rules in Retail Marketing Mix," in *Automation and Logistics*, Jinan, 2007.
- [39] M. Chen, A. Chiu and H. Chang, "Mining Changes in Customer Behavior in Retail Marketing," *Expert Systems with Applications*, vol. 28, no. 4, pp. 773-781, 2005.

- [40] Y. Zhao, R and Data Mining: Examples and Case Studies, New York, NY: Elsevier, 2012.
- [41] K. Ahn, "Effective Product Assignment Based on Association Rule Mining in Retail," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12551-12556, 2012.
- [42] M. Devi and A. Babysarajini, "Applications of Association Rule Mining in Different Databases," *Journal of Global Research in Computer Science*, vol. 3, no. 8, pp. 30-34, 2012.
- [43] M. Chen, J. Han and P. Yu, "Data Mining: An Overview from a Database Perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866-883, 1996.
- [44] S. Kotsiantis and D. Kanellopoulos, "Association Rules Mining: A Recent Overview," *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 71-82, 2006.
- [45] C. Deora, S. Arora and Z. Makani, "Comparison of Interestingness Measures: Support-Confidence Framework versus Lift-irule Framework," *International Journal of Engineering Research and Applications (ISSN)*, vol. 3, no. 2, pp. 208-215, 2013.
- [46] A. Jiménez, F. Berzal and J. Cubero, "Interestingness Measures for Association Rules within Groups," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*, Heidelberg, Springer Berlin Heidelberg, 2010, pp. 298-307.
- [47] D. Bhanu and S. P. Madeshwari, "Retail market analysis in targeting sales based on consumer behaviour using fuzzy clustering – A rule based model," *Journal of computing*, vol. 1, no. 1, p. 92 – 99, 2009.
- [48] P. Prasad and L. Malik, "Using association rule mining for extracting product sales patterns in retail store transactions," *International journal on computer science and engineering (IJCSE)*, vol. 5, pp. 2177-2182, 2011.
- [49] H. Liu, B. Su and B. Zhang, "The Application of Association Rules in Retail Marketing Mix," in *Automation and Logistics, 2007 IEEE International Conference*, Jinan, 2007.
- [50] O. Cakir and M. Aras, "A Recommendation Engine by Using Association Rules," *Procedia - Social and Behavioral Sciences*, vol. 62, pp. 452-456, 2012.

- [51] P. Paranjape-Voditel and U. Deshpande, "A Stock Market Portfolio Recommender System Based on Association Rule Mining," *Applied Soft Computing*, vol. 13, no. 2, pp. 1055-1063, 2013.
- [52] S. Yen and A. Chen, "An efficient data mining technique for discovering interesting association rules," in *Database and Expert Systems Applications, 1997. Proceedings., Eighth International Workshop*, Toulouse, 1997.
- [53] H. Le, S. Arch-int, H. Nguyen and N. Arch-int, "Association Rule Hiding in Risk Management for Retail Supply Chain Collaboration," *Computers in Industry*, vol. 64, no. 7, pp. 776-784, 2013.
- [54] W. Chiang, "To Mine Association Rules of Customer Values via a Data Mining Procedure with Improved Model: An Empirical Case Study," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1716-1722, 2011.
- [55] B. Aldosari, G. Almodaifer, A. Hafez and H. Mathkour, "Constrained Association Rules for Medical Data," *Journal of Applied Sciences*, vol. 12, no. 17, pp. 1792-1800, 2012.
- [56] C. Ordonez, "Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction," *Information Technology in Biomedicine*, vol. 10, no. 2, p. 334–343, 2006.
- [57] J. Nahar, T. Imam, K. Tickle and Y. Chen, "Association Rule Mining to Detect Factors which Contribute to Heart Disease in Males and Females," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1086-1093, 2013.
- [58] R. Chaves, J. Ramírez, J. Górriz and C. Puntonet, "Association Rule-based Feature Selection Method for Alzheimer's Disease Diagnosis for the Alzheimer's Disease Neuroimaging Initiative," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11766-11774, 2012.
- [59] M. Dimitrijević, Z. Bošnjak and E. Cohen, "Web Usage Association Rule Mining System," *Interdisciplinary Journal of Information, Knowledge & Management*, vol. 6, pp. 137-150, 2011.
- [60] S. Madria, C. Raymond, S. Bhowmick and M. Mohania, "Association rules for Web data mining in WHOWEDA," in *Digital Libraries: Research and Practice, 2000 Kyoto, International Conference*, Kyoto, 2000.

- [61] V. Nebot and R. Berlanga, "Finding Association Rules in Semantic Web Data," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 51-62, 2012.
- [62] S. Shin and W. Lee, "Online Generation Association Rules over Data Streams," *Information and Software Technology*, vol. 50, no. 6, pp. 569-578, 2008.
- [63] M. Dimitrijevic and Z. Bosnjak, "Pruning Statistically Insignificant Association Rules in the Presence of High-confidence Rules in Web Usage Data," *Procedia Computer Science*, vol. 35, pp. 271-280, 2014.
- [64] Y. Kim and B. Yum, "Recommender System Based on Click Stream Data using Association Rule Mining," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13320-13327, 2011.
- [65] S. Matthews, M. Gongora, A. Hopgood and S. Ahmadi, "Web Usage Mining with Evolutionary Extraction of Temporal Fuzzy Association Rules," *Knowledge-Based Systems*, vol. 54, pp. 66-72, 2013.
- [66] A. Mirabadi and S. Sharifian, "Application of Association Rules in Iranian Railways (RAI) Accident Data Analysis," *Safety Science*, vol. 48, no. 10, pp. 1427-1435, 2010.
- [67] A. Verma, S. Khan, J. Maiti and O. Krishna, "Identifying Patterns of Safety related Incidents in a Steel Plant using Association Rule Mining of Incident Investigation Reports," *Safety Science*, vol. 70, pp. 89-98, 2014.
- [68] B. Kamsu-Foguem, F. Rigal and F. Mauget, "Mining Association Rules for the Quality Improvement of the Production Process," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1034-1045, 2013.
- [69] Z. Huang, X. Lu and H. Duan, "Mining Association Rules to Support Resource Allocation in Business Process Management," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9483-9490, 2011.
- [70] Z. Abdullah, T. Herawan, N. Ahmad and M. Deris, "Mining Significant Association Rules from Educational Data using Critical Relative Support Approach," *Procedia-Social and Behavioral Sciences*, vol. 28, pp. 97-101, 2011.
- [71] T. Li and X. Li, "Novel Alarm Correlation Analysis System Based on Association Rules Mining in Telecommunication Networks," *Information Sciences*, vol. 180, no. 16, pp. 2960-2978, 2010.

- [72] T. Li and X. Li, "Preprocessing Expert System for Mining Association Rules in Telecommunication Networks," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1709-1715, 2011.
- [73] M. Versichele, L. de Groote, M. Bouuaert, T. Neutens, I. Moerman and N. van de Weghe, "Pattern Mining in Tourist Attraction Visits Through Association Rule Learning on Bluetooth Tracking Data: A Case Study of Ghent, Belgium," *Tourism Management*, vol. 44, pp. 67-81, 2014.
- [74] P. Alpar and S. Winkelsträter, "Assessment of Data Quality in Accounting Data with Association Rules," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2259-2268, 2014.
- [75] D. Sánchez, M. Vila, L. Cerda and J. Serrano, "Association Rules Applied to Credit Card Fraud Detection," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3630-3640, 2009.
- [76] M. Plasse, N. Niang, G. Saporta, A. Villeminot and L. Leblond, "Combined use of Association Rules Mining and Clustering Methods to Find Relevant Links Between Binary Rare Attributes in a Large Data Set," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 596-613, 2007.
- [77] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A. I. Verkamo, "Finding Interesting Rules From Large Sets Of Discovered Association Rules.," in *CIKM '94: Proceedings of the Third International Conference on Information and Knowledge Management*, New York, 1994.
- [78] A. Gosain and M. Bhugra, "A comprehensive survey of association rules on quantitative data in data mining," in *Information & Communication Technologies (ICT), 2013 IEEE Conference*, JeJu Island, 2013.
- [79] Y. Xu and Y. Li, "Generating concise association rules," in *Proceeding CIKM '07 Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, New York, 2007.
- [80] M. Moreno, S. Segrera and V. López, "Association Rules: Problems, Solutions and New Applications," in *Actas deI III Taller Nacional de Minería de Datos y Aprendizaje*, Los autores, 2005.
- [81] C. Rygielski, "Data Mining techniques for Customer Relationship Management," *Technology in Society*, vol. 24, no. 4, pp. 483-502, 2002.

- [82] A. Wasilewska, "Apriori Algorithm," [Online]. Available: http://www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf. [Accessed 12 03 2015].
- [83] H. Poh and T. Jasic, "Forecasting and Analysis of Marketing Data Using Neural Networks: A Case of Advertising and Promotion Impact," in *Artificial Intelligence for Applications, 1995. Proceedings of the 11th Conference*, Los Angeles, 1995.
- [84] J. Mistry, F. Nelwamondo and T. Marwala, "Estimating Missing Data and Determining the Confidence of the Estimate Data," in *Machine Learning and Applications, ICMLA '08. Seventh International Conference*, San Diego, 2008.
- [85] S. Nirkhi, "Potential use of Artificial Neural Network in Data Mining," in *Computer and Automation Engineering (ICCAE)*, Singapore, 2010.
- [86] S. Tzafestas and H. B. Verbruggen, *Artificial Intelligence in Industrial Decision Making, Control and Automation*, Springer Science & Business Media, 2012.
- [87] D. Montana and L. Davis, "Training Feedforward Neural Networks Using Genetic Algorithms," in *Proc. 11th Int. Joint Conf. Artificial Intelligence*, 1989.
- [88] M. Negnevitsky, *Artificial Intelligence A Guide to Intelligent Systems*, London: Addison Esley, 2005.
- [89] O. Vornberger, F. Thiesing and U. Middleberg, "Short Term Prediction of Sales in Supermarkets.," in *Proceeding of Neural Networks*, Perth, 1995.
- [90] A. Koç and Ö. Yeniay, "A Comparative Study of Artificial Neural Networks and Logistic Regression for Classification of Marketing Campaign Results," *Mathematical and Computational Applications*, vol. 18, no. 3, pp. 392-398, 2013.
- [91] D. Boone and M. Roehm, "Retail Segmentation using Artificial Neural Networks," *International Journal of Research in Marketing*, vol. 19, no. 3, pp. 287-301, 2002.
- [92] N. Popoviciu and M. Boncut, "On the Hopfield Algorithm. Foundations and Examples," *General Mathematics*, vol. 13, no. 2, pp. 35-50, 2005.
- [93] X. Yan and Y. Li, "Customer Segmentation based on Neural Network with Clustering Technique," in *Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Databases*, Madrid, 2006.
- [94] E. Li, "Artificial Neural Networks and their Business Applications," *Information & Management*, vol. 27, pp. 303-313, 1994.

- [95] K. Jha, "Artificial Neural Networks and its Application," 2007. [Online]. Available: http://iasri.res.in/ebook/EBADAT/5-Modeling%20and%20Forecasting%20Techniques%20in%20Agriculture/5-ANN_GKJHA_2007.pdf. [Accessed 8 December 2014].
- [96] Y. Li and W. Ma, "Applications of Artificial Neural Networks in Financial Economics: A Survey," in *Computational Intelligence and Design (ISCID), 2010 International Symposium*, Hangzhou, 2010.
- [97] F. Qian and L. Xu, "Improving Customer Satisfaction by the Expert System using Artificial Neural Networks," *Intelligent Control and Automation*, p. 8303–8306, 2008.
- [98] H. Poh, J. Yao and T. Jasic, "Neural Networks for the Analysis and Forecasting of Advertising and Promotion Impact," *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 7, no. 4, pp. 1-17, 1998.
- [99] S. Anbananthen, G. Sainarayanan, A. Chekima and J. Teo, "Artificial Neural Network Tree Approach in Data Mining," *Malaysian Journal of Computer Science*, vol. 20, no. 1, pp. 51-62, 2007.
- [100] P. Cerny, "Data Mining and Neural Networks from a Commercial Perspective," *ORSNZ Conference Twenty Naught One*, 2001.
- [101] S. Nirakhi, "Potential Use of Artificial Neural Network in Data Mining," in *International Conference on Computer and Automation Engineering (ICCAE)*, Singapore, 2010.
- [102] S. Viademonte and F. Burstein, "An Intelligent Model for Decision Support Based on Neural Networks Components," in *Australasian Data Mining Workshop, Joint with The 15th Australian Joint Conference on Artificial Intelligence (2002). Workshop Proceedings*, Canberra, 2002.
- [103] S. Nogay, T. Akinci and M. Eidukeviciute, "Application of Artificial Neural Networks for Short Term Wind Speed Forecasting in Mardin - Turkey," *Journal of Energy in Southern Africa*, vol. 23, no. 4, pp. 2-7, 2012.
- [104] M. Bayzid, A. Iqbal, C. Hyder and M. Irfan, "Application of Artificial Neural Network in Social Computing in the Context of Third World Countries," in *5th International Conference on Electrical and Computer Engineering ICECE 2008*, Dhaka, 2008.
- [105] O. Awodele and O. Jegede, "Neural Networks and Its Application in Engineering," in *Proceedings of Informing Science & IT Education Conference (InSITE) 2009*, 2009.

- [106] I. Khan, P. Zope and S. Suralkar, "Importance of Artificial Neural Network in Medical Diagnosis Disease Like Acute Nephritis Disease and Heart Disease," *International Journal of Engineering Science and Innovative Technology (IJESIT)*, vol. 2, no. 2, pp. 210-217, 2013.
- [107] R. Price, E. Spitznagel, T. Downey, D. Meyer, N. Risk and O. El-Ghazzawy, "Applying Artificial Neural Network Models to Clinical Decision Making," *Psychological Assessment*, vol. 12, no. 1, pp. 40-51, 2000.
- [108] P. Nagendra, S. Halder nee Dey and T. Dutta, "Artificial Neural Network Application for Power Transfer Capability and Voltage Calculations in Multi-Area Power System," *Leonardo Electronic Journal of Practices and Technologies*, vol. 16, pp. 119-128., 2010.
- [109] R. Aggarwal and Y. Song, "Artificial Neural Networks in Power Systems. III. Examples of Applications in Power Systems," *Power Engineering Journal*, vol. 12, no. 6, pp. 279-287, 1998.
- [110] J. Cannady, "Artificial Neural Networks for Misuse Detection," in *Proceedings of the 1998 National Information Systems Security Conference (NISSC'98)*, Arlington, 1998.
- [111] C. Yang, S. Prasher, J. Landry, H. Ramaswamy and A. Ditommaso, "Application of Artificial Neural Networks in Image Recognition and Classification of Crop and Weeds," *Canadian Agricultural Engineering*, vol. 42, no. 3, pp. 147-152, 2000.
- [112] M. Seyam and Y. Mogheir, "Application of Artificial Neural Networks Model as Analytical Tool for Groundwater Salinity," *Journal of Environmental Protection*, vol. 2, pp. 56-71, 2011.
- [113] J. Basu, D. Bhattacharyya and T. Kim, "Use of Artificial Neural Network in Pattern Recognition," *International Journal of Software Engineering and its Applications*, vol. 4, no. 2, pp. 23-34, 2010.
- [114] A. Kumar, A. Kumar, R. Ranjan and S. Kumar, "A Rainfall Prediction Model using Artificial Neural Network," *Control and System Graduate Research Colloquium (ICSGRC)*, pp. 82-87, 2012.
- [115] H. Nagahamulla, U. Ratnayake and A. Ratnaweera, "An ensemble of Artificial Neural Networks in Rainfall Forecasting," in *Advances in ICT for Emerging Regions (ICTer), 2012 International Conference*, Colombo, 2012.

- [116] V. Shahpazov, V. Velev and L. Doukovska, "Design and Application of Artificial Neural Networks for Predicting the Values of Indexes on the Bulgarian Stock Market," in *Signal Processing Symposium (SPS)*, Serock, 2013.
- [117] C. Biryulev, Y. Yakymiv and A. Selemonavichus, "Research of Artificial Neural Networks Usage in Data Mining and Semantic Integration," in *Perspective Technologies and Methods in MEMS Design (MEMSTECH), 2010 Proceedings of Vith International Conference*, Lviv, 2010.
- [118] J. Bullinaria, "Introduction to Neural Networks : Lecture 16," 2004. [Online]. Available: <http://www.cs.bham.ac.uk/~jxb/NN/116.pdf>. [Accessed 23 January 2016].
- [119] L. Tiecheng, "Optimizing Mining Association Rules Based on Artificial Neural Network," *World Automation Congress (WAC)*, pp. 1-4, 2012.
- [120] W. Chen and Y. Du, "Using Neural Networks and Data Mining Techniques for the Financial Distress Prediction Model," *Expert Systems with Applications*, vol. 36, no. 2, pp. 4075-4086, 2009.
- [121] P. Chou, P. Li, K. Chen and M. Wua, "Integrating Web Mining and Neural Network for Personalized E-commerce," *Expert Systems with Applications*, vol. 37, no. 4, pp. 2898-2910, 2010.
- [122] M. W. Craven and J. W. Shavlik, "Using Neural Networks for Data Mining," *Future Generation Computer Systems*, vol. 13, no. 2-3, pp. 211-229, 1997.
- [123] R. Brause, T. Langsdorf and M. Hepp, "Neural Data Mining for Credit Card Fraud Detection, Tools with Artificial Intelligence," in *Proceedings. 11th IEEE International Conference*, Washington DC, 1999.
- [124] M. Karabatak and M. Ince, "An Expert System for Detection of Breast Cancer Based on Association Rules and Neural Network," *Expert Systems with Applications*, vol. 36, pp. 3465-3469, 2009.
- [125] C. Arockiaraj, "Applications of Neural Networks in Data Mining," *International Journal of Engineering and Science*, vol. 3, no. 1, pp. 8-11, 2013.
- [126] M. Lenzerini, "Data Integration: A Theoretical Perspective," in *Proceeding PODS '02 Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, New York, 2002.

- [127] P. Nadkarni and L. Marenco, "Data Integration: An Overview Methods in Biomedical Informatics," in *Methods in Biomedical Informatics A Pragmatic Approach*, Elsevier, 2014, pp. 15-47.
- [128] M. Varga and K. Curko, "Some Aspects of Information Systems Integration," in *MIIPRO 2012 Proceedings of the 35th International Convention*, Opatija, 2012.
- [129] D. Loshin, "Data Consolidation and Integration," in *Master Data Management*, New York, Morgan Kaufmann, 2009, pp. 177-199.
- [130] M. Prabhavathy and K. Sivasankari, "Federated Query Processing Service in Service Oriented Business Intelligence," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 3, no. 1, pp. 1266-1272, 2014.
- [131] R. Reinhard, T. Meisen, D. Schilberg and S. Jeschke, "A Framework Providing a Basis for Data Integration in Virtual Production," in *Numerical Simulations of Physical and Engineering Processes*, India, InTech, 2011, pp. 541-550.
- [132] M. Hema and S. Chandramathi, "Service Oriented Ontology based Data Federation for Heterogeneous Data Sources," *Journal of Theoretical and Applied Information Technology*, vol. 55, no. 1, pp. 126-136, 2013.
- [133] J. Jackson, "Data Mining: A Conceptual Overview," *Communications of the Association for Information Systems*, vol. 8, pp. 267-296, 2002.
- [134] S. Zhang, C. Zhang and Q. Yang, "Data Preparation for Data Mining," *Applied Artificial Intelligence*, vol. 17, pp. 371-381, 2003.
- [135] C. Zhang, Q. Yang and B. Liu, "Guest Editors' Introduction: Special Section on Intelligent Data Preparation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1163-1165, 2005.
- [136] V. Bogorny, P. Engel and L. Alvares, "Spatial Data Preparation for Knowledge Discovery," *IEEE Computer Graphics*, 2005.
- [137] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," *Bulletin of the Technical Committee on Data Engineering*, pp. 1-11, 2000.