IMPROVED INTEGRATED MINING OF HETEROGENEOUS DATA IN DECISION SUPPORT SYSTEMS

BY

FATUDIMU, Ibukun Tolulope
(* CUGP050145*)
B.Sc (Hons) Engineering Physics (Obafemi Awolowo University, Ile-Ife)
M.Sc.Computer Science (University of Ibadan, Ibadan)

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES TO THE SCHOOL OF POSTGRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY OF COVENANT UNIVERSITY OTA, OGUN STATE, NIGERIA

March 2012

CERTIFICATION

This is to certify that this thesis is an original research work undertaken by

Fatudimu Ibukun Tolulope and approved by:

1. Name: **Prof. Charles Onuwa Uwadia**

   Supervisor

13/3/2012

Signature …… Date:……

2. Name: **Prof. Charles Korede Ayo**

   Co-Supervisor

Signature……. Date:……..

3. Name: **Prof. Charles Korede Ayo**

   Head of Department

Signature ……. Date:……..

4. Name:

   External Examiner

Signature …… Date:……...

DEDICATION

This project is dedicated to God almighty for he has been my strong anchor during the course of this programme. I would not have made it this far if not because of your constant company and your sustenance. All because of you, the journey has been wonderful.

ACKNOWLEDGEMENT

I express my sincere gratitude to the Chancellor of Covenant University. I don't think I would have completed this project without God using you to intervene in my health. Indeed I am one of the living testimonies of the mandate which rescued me from destruction. Thanks for being what God intended you to be.

My sincere appreciation also goes to my supervisors, Professor C.O. Uwadia for his assistance, and critical supervision during this study. I also want to appreciate his kind and gentle ways of correcting and guiding me to achieve the aim of my project work. Professor C.K. Ayo, I can only pray that your children will never lack help. You have been a guide to me not only academically but have impacted my life through your fatherly guidance. From you, I learnt fulfillment through helping others, indeed you are one in a million.

Thanks to Professor E.A. Adebiyi for insisting on high standard which helped improve the quality of my research work. I cannot forget to acknowledge Dr J.O. Daramola, Mrs Oladipupo and Femi Afolabi who have been a source of help and encouragement throughout the course of this project.

ABSTRACT

The volume of information available on the Internet and corporate intranets continues to increase along with the corresponding increase in the data (structured and unstructured) stored by man organizations. In customer relationship management, information is the raw material for decision making. For this to be effective there is need to discover knowledge from the seamless integration of structured and unstructured data for completeness and comprehensiveness.

This study addresses two unique challenges experienced in business decision support systems, the first one is how to transform and analyze unstructured data alongside structured data. Secondly, the need to improve result obtained from the integrated mining system in order to reduce decision failure. There is also a necessity to solve the challenge of Customer Relationship Management, in terms of the ability to differentiate useful information from chatter or even disinformation. There is also further need to have a holistic view to mining from structured and unstructured information sources towards a better Customer Relationship Management.

Improved Integrated Mining Architecture (IIMA), our approach to solving the above challenges, consists of three major phases; the first phase is the Extraction and Integration phase. This phase is aimed at optimizing the performance of the knowledge mining phase. It consists of unstructured data (text) preprocessing which includes lexical analysis, stemming, application of weighing schemes and finally transforming the documents to an XML format. In the integration

process, the structured component is selected based on the resulting keywords from the unstructured text preprocessed. The second phase is the Knowledge distillation phase. In this phase, knowledge is distilled using the modified Generating Association Rules based on Weighting scheme (GARW) algorithm. In the third phase, generated rules are also interpreted for making efficient business decisions. To implement the above described methodology, the following programming design tools were used: Microsoft Visual Studio 2008 (C# in particular), Microsoft SQL Server 2008 and Extensible Markup Language (XML). For the data preprocessing phase, WordNet lexical database was referenced through Proxem Antelope.

Experiments carried out revealed that the extracted association rules contain important features which form a worthy platform for making effective decisions as regards Customer Relationship Management. The performance of the IIMA approach is also compared with an integrated mining approach which uses just syntactic relevance in its information extraction process. The result revealed a significant reduction in the large itemsets and execution time. Also, in the novelty evaluation, IIMA produced a 17% increase in novel rules generated when compared to the existing integrated mining approach.

CONTENTS

CHAPTER FOUR: APPLICATION OF IIMA TO CRM

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

DSS	Decision Support Systems
XML	Extensible Markup Language
CRM	Customer Relationship Management
GARW	Generating Association Rules based on Weighting scheme
BI	Business Intelligence
CI	Competitive intelligence
ESTEST	Experimental Software To Extract Structure from Text
Disco-TEX	Discovery from Text Extraction
TM	Text Mining
TAKMI	Text Analysis and Knowledge MIning
IE	Information Extraction
IR	Information Retrieval
NLP	Natural Language Processing
TF-IDF	Term Frequency, Inverse Document Frequency
IIMA	Improved Integrated Mining Architecture

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND INFORMATION

The volume of information available on the Internet and corporate intranets continues to increase along with the corresponding increase in the data (structured and unstructured) stored by many organizations. Over the past years, data mining has explored the large volume of data (structured) in order to discover knowledge, often in form of decision support systems. A critical component for the success of the modern enterprise is its ability to take advantage of all available information. This challenge becomes more difficult with the constantly increasing volume of information, both internal and external to an enterprise. According to this rate of increase, more data will be generated in the next three years than in all of recorded history. This explosion of information presents an exciting cross-industry business opportunity. Enterprises that can quickly extract critical nuggets of information from the sea of accessible data and transform them into valuable business assets are in a strong position to dominate their markets.

Decision support systems are interactive computer based systems that aid users in judgment and choice of activities. These systems have gained popularity in various domains such as business, engineering, military and medicine and are most valuable in situations where the amount of information is too large for human decision makers to use optimally and with precision (Druzdzel & Flynn, 2002).

The structured environment is made up of data that has fields, columns, tables, rows and indexes. It centers on transactions and has reports, audits and definitions of words. There is high degree of predictability associated with the structured environment (Inmon, 2007). Mining in this environment involves an analytic process designed to explore the structured data in search of consistent patterns and/or systematic relationship between variables, and then to validate the findings by applying the detected patterns to new subsets of data (http://www.statsoft.com/textbook/stdatmin.html #mining). This type of mining is limited due to the fact that the available information accessible to a company is mostly unstructured (Unitas Corporation, 2002; Blumberg, 2003).

The unstructured environment has no particular order to it. It consists of text found in medical reports, warranties, contracts, email and spreadsheets. The text has no rules governing its creation or usage. With text, there are no keys, no indexes, no columns or attributes (Inmon, 2007). Unstructured data can take formats such as, excel files, web blogs and so on. Closely related to this is the semi-structured environment which is an intermediate between structured and unstructured data. Semi structured data usually has some form of meta-data attached to it unlike unstructured that has no metadata at all (Ukelson, 2006). Examples of semi-structured data include XML (Extensible Markup Language) data storage. Mining in the unstructured environment is known as text mining. Text mining is the process of extracting interesting and non trivial patterns of knowledge from unstructured text documents. It can also be expressed as knowledge discovery from unstructured databases (Ah-Hwee, 2006). Mining unstructured data is important due to the following reasons:

- In today's era of information, OLTP (Online Transaction Processing) and data warehousing systems take increasing proportion of their data from applications and automated systems rather than users, and those applications feed data that has become predominantly semi-structured or unstructured. Statistics revealed that, as much as 85% of today's OLTP and data warehouse data are unstructured (Kernochan, 2006; Sukumaran & Sureka, 2007; Unitas Corporation, 2002; Blumberg, 2003).
- The rapid growth of the Internet has led to increase in the amount of information generated and shared by organizations in almost every industry and sector. This increase has led to the creation of huge, but largely unmet need for tools that can be used to manage what we call unstructured data (Blumberg, 2003). "In the case of web alone, more than 2 billion new web pages have been created since 1995, with additional 200 million new pages being added every month, according to market research firm IDC". (Blumberg, 2003).
- Finally some experiments produce a mix of unstructured and structured data.

Integrated mining in this context therefore can be defined as creating a platform for mining structured and unstructured data. The outcome of such integrated mining will solve the problem of having a holistic view to mining from structured and unstructured data sources which is currently a major challenge in the field of customer relationship management. Integrated mining will be beneficial to this application area as it combines the unique attributes of both structured and unstructured data format in order to provide greater efficiency to the organization, especially by minimizing the popular practice of

handling structured and unstructured as distinct information entities which often results in decision management failure (Sukumaran & Sureka, 2007).

1.2 STATEMENT OF THE PROBLEM

In business decision support system, information is the raw material for decision making. Therefore, effective decision making is based on sound information. There is a need to provide data that reduces the level of uncertainty in decision making (Solomon & & Paul, 2003). In order to do this, the inefficiency of the existing data integration systems needs to be reduced by minimizing the problem of uncertainty of extracted features leading to unreliable reports. Presently in the field of business decision support system, (Zhu et. al, 2005; Kernochan, 2006; Arnold, 2010); has been able to analyze both text and data using methods such as OLAP-style interaction model, pattern recognition technology, statistical models and text analytic engines. In all the attempts above, semantic analysis of the integrated data was not involved. Also, due to the fact that most of the systems reviewed just stop at the point of integration, there is also a need to develop a system that combines both integration and mining of data based on the most contributing extracted features.

There is also a need to solve the challenge with Customer Relationship Management, which is not lack of information (Solomon & & Paul, 2003), but the ability to differentiate useful information from chatter or even disinformation Currently, having a holistic view to mining from structured and unstructured information sources towards a better Customer Relationship Management is the problem of analytical Customer Relationship Management (CRM) system (Cody et al., 2002). Developing a

system that can solve this problem will therefore be a great contribution to the world of business intelligence.

In this thesis therefore, we intend to address the challenge of improving the efficiency of integrated mining system. The following are the research questions which the thesis tries to address:

- Can we source, transform and analyze unstructured data alongside with structured data?
- How can we integrate structured and unstructured data for effective decision making?
- How do we improve result obtained from the integrated mining system in order to reduce decision failure?

1.3 AIM AND OBJECTIVES OF THE STUDY

The aim of the research is to develop an improved integrated mining system for the purpose of making effective business decision. To achieve this aim, the following objectives were formulated;

- To develop a system that mines from the integration of structured and unstructured data.
- To improve the efficiency of the developed system by introducing a semantic preprocessing of unstructured data.
- To apply the developed system to improve the organizational profit of manufacturing and production companies through an effective Customer Relationship Management.

- To evaluate the system by comparing its results with an existing one.

1.4 METHODOLOGY

In order to evolve an improved integrated mining architecture, a review of the state of the art literature on integrated data management was done which revealed that the area of concern in integrated data management is in the preprocessing of the unstructured data. This led to a rigorous review of the existing text mining systems after which an improved system was proposed which is coined Improved Integrated Mining Architecture (IIMA). The system consists of three major phases; the first phase is the Extraction and Integration phase. This phase is aimed at optimizing the performance of the knowledge mining phase. It consists of unstructured data (text) preprocessing which includes lexical analysis, stemming, application of weighing schemes and finally transforming the documents to an XML format. In the integration process, the structured component is selected based on the resulting keywords from the unstructured text preprocessed. The second phase is the Knowledge distillation phase. In this phase, knowledge is distilled using the modified GARW (Generating Association Rules based on Weighting scheme) algorithm (Hany et al., 2007). The third phase is the rules visualization phase whereby the generated rules are interpreted for making efficient business decisions. Association rules are easy to understand and to interpret for an analyst or a normal user. However, it should be mentioned that the association rule extraction is of exponential growth and a very large number of rules can be produced. To implement the above described methodology, the following programming design tools were used, Microsoft Visual Studio 2008 (C# in particular), Microsoft SQL Server 2008 and Extensible Markup Language (XML). For the data preprocessing phase, WordNet lexical database is

referenced through Proxem Antelope. Based on the proposed architecture, the customer relationship management of the Nigerian mobile phone industry was investigated in order to compare and evaluate the efficiency of the developed system over the existing one. The mobile phone industry was selected because it provides a real life application to justify the research work. In this project, IIMA system is evaluated using a suitable method of estimating the novelty of rules discovered by data-mining methods (Basu et al., 2001).

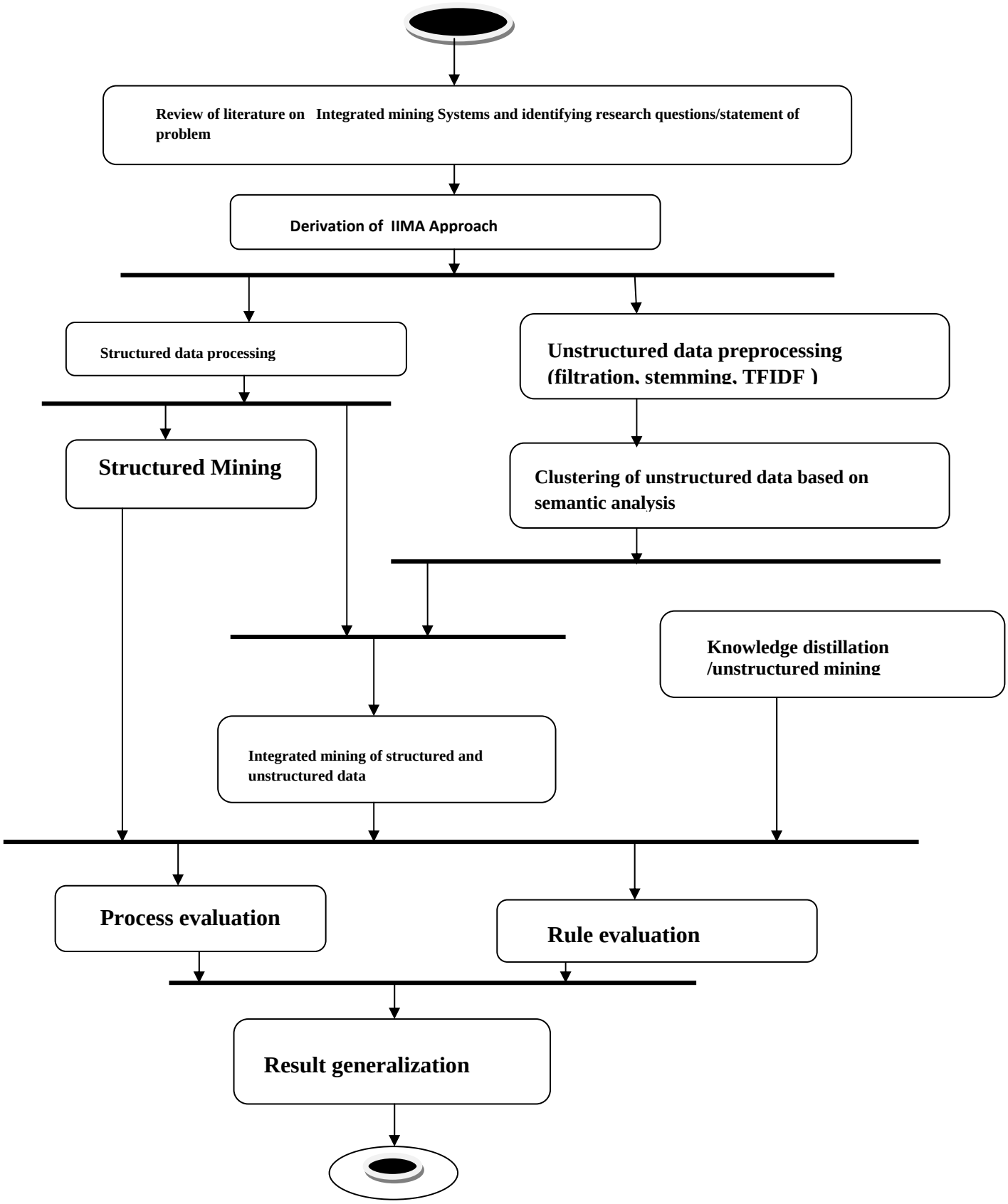


Figure 1.1 Model conceptualization of the methodology of this thesis.

1.5 SIGNIFICANCE OF THE STUDY

The following are the significance of this research;

- The research will minimize the popular practice of handling structured and unstructured data as distinct information entities which often results in decision management failure (Sukumaran & Sureka, 2007). For example, in the area of records management whereby invoices, statements and other operational documents need to be tied to customer data or supplier data.
- The research will help to find out what consumers need before production and better service after purchase, thereby improving the efficiency of customer relationship management.
- This is an information age and the level of sophistication in terms of knowledge, competition, taste and technology has increased. Therefore, Marketers need to procure and process accurate information, there is therefore a need to come up with better ways of information processing and gathering which this study aims at.
- The research will solve the challenge in CRM, which is not lack of information, but the ability to differentiate useful information from chatter or even disinformation.

1.6 MOTIVATION FOR THE STUDY

The desire to extend the capabilities of business intelligence applications to include textual information has existed for quite some time. The major inhibitors have included the separation of the data on different data management systems, typically across different organizations, and the immaturity of automated text analysis techniques for deriving business value from large amounts of text. For example, to understand sales

effectiveness, a telemarketing revenue data cube can help identify products who generate the most sales, and customers who are the most receptive to this sales approach. Unfortunately, the particular sales techniques used by these successful sales representatives in various situations are not captured by quantitative measures in the OLAP cube. However, these sales conversations are now frequently recorded and converted to text. The text of conversations associated with high-revenue sales representatives and high-yield customers can be analyzed by various language processing or pattern detection techniques to find patterns in the use of phrases or phrase sequences.

Secondly, there is a need, not just to develop a system to solve the above problem, but to also provide the most efficient solution. This thesis seeks to create a state-of-the-art solution and efficiency of integrated mining systems in the business decision support systems. The need described above informed our decision to pick business intelligence as our area of application in this thesis.

1.7 CONTRIBUTION TO KNOWLEDGE

The specific contributions of this research are both in the application area which is in business decision support system and also in the computer science field. Firstly, in the world of business decision support system (business intelligence), the task of integrating various data sources have been the burden of the enterprise application developer (Roth et al., 2002). A lot of commercial systems together with academic projects have addressed comprehensive information integration platform. Many of these approaches start “from scratch,” and build a special-purpose system to optimize a particular use. According to (Roth et al., 2002), products such as Tamino

(http://www.softwareag.com/tamino/) and Ipedo (http://www.ipedo.com) promise to deliver data stores optimized for XML documents. Also, data federation has a solid research foundation and several commercial implementations. TSIMMIS, (Garcia-Molina et al., 1995), DISCO,(Tomasic et al., 1997) HERMES,(Adali et al., 1996) and the Information Manifold (Levy et al., 1996), are products that are built specially to explore various aspects of federated database technology which includes, compensation, mediation and scalability. Also, Nimble, (http://www.nimble.com) Callixa,(http://www.callixa.com) and InfoShark (http://www.infoshark.com) federate data outside the database engine. In addition, Garlic (http://www3.ibm.com/solutions/lifesciences/discoverylink.html) and DB2* (Database 2*) Relational Connect (http://www3.ibm.com/software/data/db2/relconnect) extend a traditional relational database engine with federated capabilities. DiscoveryLink* (http://www3.ibm.com/solutions/lifesciences/discoverylink.html) for example, is a commercial application that is built on technology tailored to the life sciences community. But none of the above systems uses one platform to integrate and mine (with specific mining algorithm such as association rule) data at the same time. That is why we believe that the development of our integrated mining will be of great benefit to the business intelligence world. The following is a diagram showing the contribution of the system to business decision support system.

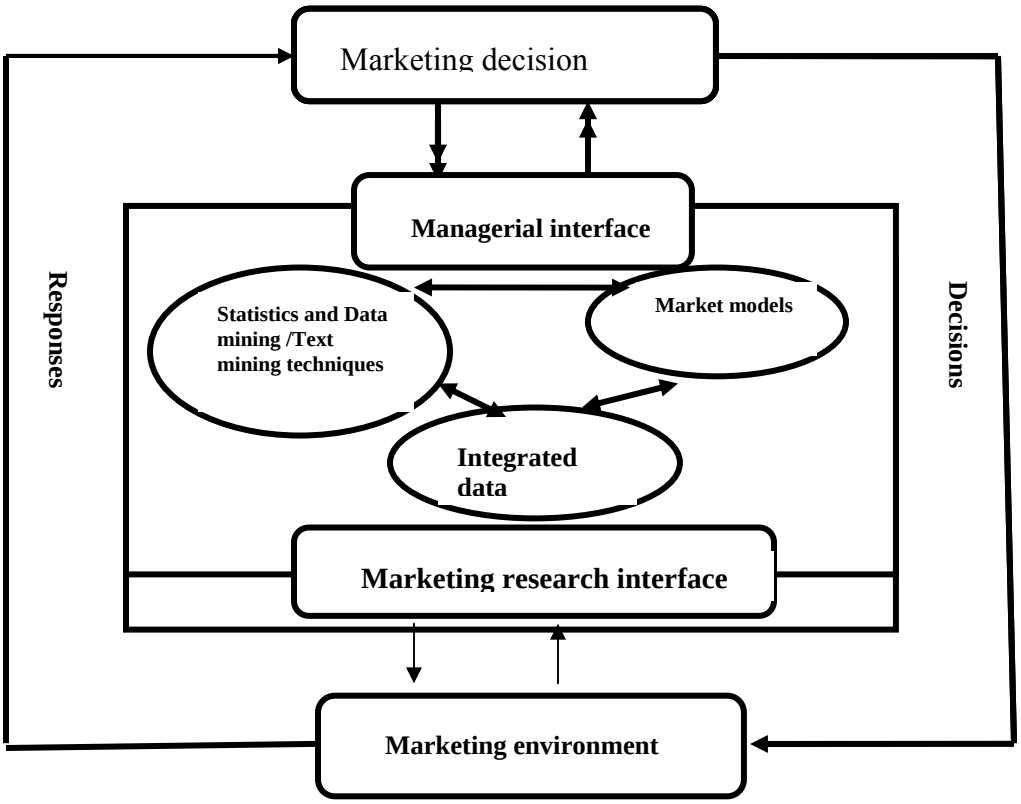


Figure 1.2 Marketing Decision Support system (Strauss et at., 2006)

The difference in the architecture in Figure 1.2, as regards the original marketing decision support system is that data mining & Text mining technique has been added to the statistical techniques component and the raw data component has been replaced by an integrated data warehouse which is obtained as a result of the data preprocessing. By applying the IIMA on the CRM data gathered for the purpose of experimental validation in this research, the result revealed novel CRM inferences which are an advantage of our modified market decision support system.

According to (Frieder et. al., 2000; Roth et. al., 2000; Dean & Alexandra, 2004; Aravindan, 2005; Robert, 2006; Prem, 2007; Sukumaran & Sureka, 2007), data

integration has been approached mainly from the information extraction perspective based majorly on syntactic analysis which is void of semantic analysis. Though semantic analysis was introduced by Oracle Database 11g (An Oracle White paper, 2007), it relies on domain specific ontologies which in turn rely on an application developer, this makes it limited in flexibility. The second contribution to knowledge of this research therefore is to address the novel problem of classifying semantically related XML documents. To do this Improved Integrated Mining Architecture (IIMA) was produced. In IIMA, semantic relatedness of XML documents is investigated by analyzing the content information in order to generate XML features with the support of lexical ontology knowledge. Content analysis applies to textual elements and combines methods that are conceived to compute term relevance from both syntactic and semantic viewpoints. Syntactic relevance takes into account the structural context of term occurrence, whereas semantic relevance depends on the degree of term polysemy.

This thesis therefore introduces efficiency into the integrated mining process. This is done through semantic data preprocessing in order to reduce the corpus that will move to the key phrase extraction stage to the most relevant documents based on the problem to be solved. Efficiency in this context refers to producing patterns (rules) which is interesting both objectively and subjectively, that is, rules that are interesting based on the underlying data collection, they are unexpected and the user can act on them because it contributes directly to the solution of the problem defined. Applying the IIMA on a mobile phone industry case study generated inferences towards competitive advantage through effective customer relationship management. This is due to the fact that semantically related keywords extracted from the combination of structured and

unstructured component of the data mostly generated the novel rules.

1.8 DELIMITATIONS OF THE SCOPE OF THE STUDY

Even though unstructured data also includes video and sound, in the context of the research it is limited to text-based information, and does not include video and sound. Also, Business Intelligence systems include components such as GIS, OLAP, EIS but this research is limited to the CRM.

1.9 THESIS ORGANIZATION

Chapter One of this thesis presents a general introduction, highlighting the motivation for the research, the aim and objectives of the work and its contributions.

Chapter Two gives a description of the area of application, a critical review of existing integrated mining systems and also a detailed review of existing text preprocessing approaches. The chapter presents a review of related work and defines the context of the research undertaken in this work by identifying the gaps that exist in literature. The chapter concludes with the proposal of a novel approach to integrated mining of heterogeneous data.

Chapter Three is a description of the Improved Integrated Mining Architecture.

Chapter Four presents the application of the system to the case study of CRM that was undertaken to validate the IIMA approach.

In Chapter Five, the details of the evaluation procedure for IIMA are discussed.

Finally, in Chapter Six, we give the summary, conclusion and a discussion of the future research outlook of this thesis.

CHAPTER TWO

LITERATURE REVIEW OF INTEGRATED MINING SYSTEMS

2.1 INTRODUCTION

The current explosion of information presents an exciting cross-industry business opportunity. Enterprises that can quickly extract critical nuggets of information from the sea of accessible data and transform them into valuable business assets are in a strong position to dominate their markets. The challenge for enterprise software applications today is information integration.

The integrated data can be differentiated into structured data (e.g., accountings which are computer process able) and unstructured data (e.g., text documents which are not computer process able). The name “unstructured data” is based on the circumstances that the structure of a text document is not clear to an information system because text semantics cannot be understood by machines. According to (Blumberg & Atre, 2003), A common problem facing many organizations today is that of multiple or disparate information sources and repositories, including databases, object stores, knowledge bases, digital libraries, information retrieval systems and electronic mail systems. Decision makers often need information from multiple sources, but are unable to get and fuse the required information in a timely fashion due to the difficulties of accessing the different systems and due to the fact that the information obtained can be inconsistent and contradictory. Until now there have only been a few published approaches which tackle this problem domain and give approved solutions for the coupling of internal and external data. This means the already structured data from inside the company and unstructured data from the Internet.

2.2 DECISION SUPPORT SYSTEMS

Decision support systems are interactive computer based systems that aid users in the judgment and choice of activities (Druzdzel & Flynn, 2002). DSS are specific class of computerized information system that supports decision making activities. In general DSS are interactive computer based systems and subsystems intended to help decision makers use communication technologies, data, documents, knowledge and /or model to identify and solve problems and make decisions (Daniel & Shashidhar, 2000).

There are three fundamental components of DSSs (Druzdzel & Flynn, 2002; Andrew, 1991):

- **Database management system (DBMS):** The purpose of a DBMS is to serve as a data bank for the DSS. The DBMS therefore, stores a large quantity of data that is relevant to the class of problems for which the DSS has been designed. It also provides logical data structures with which the users interact. A DBMS separates the users from the physical aspects of the database structure and processing. DBMS should also be capable of informing the user of the types of data that are available and how to gain access to them.
- **Model-base management system (MBMS:** The role of MBMS is similar to that of a DBMS. It basically provides independence between specific models that are used in a DSS from the applications that use them. MBMS is used to transform data from the DBMS into information that is useful in decision making. Since many problems that the user of a DSS will cope with may be unstructured, the MBMS should also be capable of assisting the user in model building.

- **Dialog generation and management system (DGMS):** The users of a DSS are often managers who are not necessarily computer trained. The main product of an interaction with a DSS is insight. Decision Support Systems therefore needs to be equipped with intuitive and easy-to-use interfaces. These interfaces aid in model building, but also in interaction with the model, such as gaining insight and recommendations from it. The DGMS's primary responsibility is to enhance the ability of the system user to utilize and benefit from the DSS. In the remainder of this research work, we will use the broader term user interface rather than DGMS.

Though a variety of DSSs exists, the above three components can be found in most DSS architectures. These three components play a prominent role in the structure of a DSS. Essentially, the user interacts with the DSS through the user interface (DGMS). This communicates with the DBMS and MBMS, which screen the user and the user interface from the physical details of the model base and database implementation (Andrew, 1991).

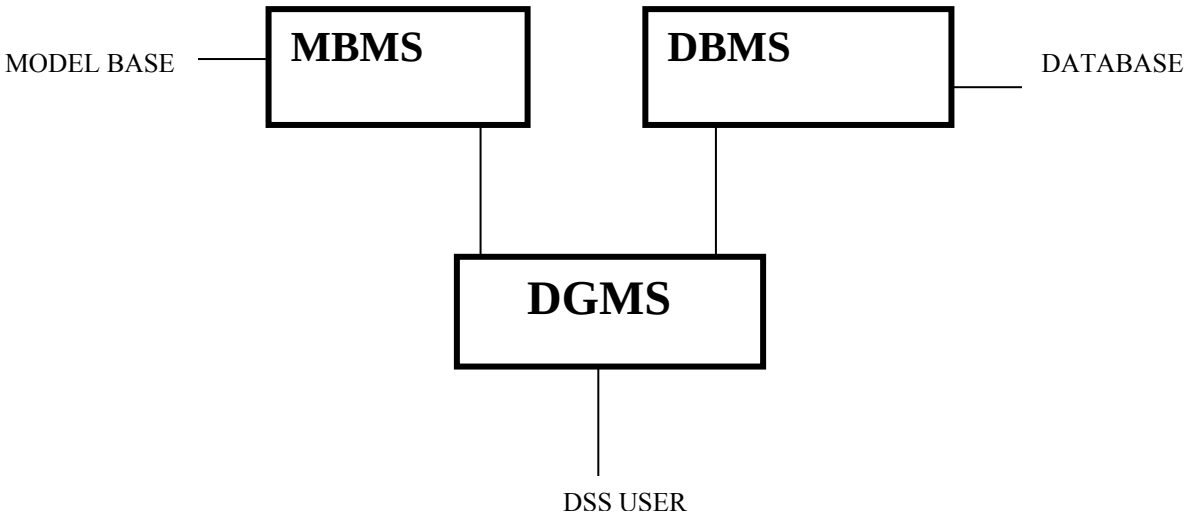


Figure 2.1: The architecture of a DSSs (Andrew, 1991).

A communications-driven DSS supports more than one person working on a shared task, it supports communication, collaboration, and coordination. Document-driven DSS manage, retrieve, summarize and manipulate unstructured information in a variety of electronic format. Knowledge-driven DSS have specialized problem solving expertise stored as facts, rules and procedures or similar structures. The "expertise" consists of knowledge about a particular domain, understanding of problems within that domain, and "skill" at solving some specific problems. A model-driven DSS emphasizes access to and manipulation of statistical, financial, optimization or simulation models. Model-driven DSS uses data and parameters provided by decision makers to aid them in analyzing a solution. Data-driven DSS emphasizes access to and manipulation of a time-series of internal company data and sometimes, external data (Daniel & Shashidhar, 2000).

Data driven DSSs include online analytical processing (OLAP) applications and data mining applications. OLAP applications allow users to think of a database as having multiple dimensions and to query those dimensions in various combinations. The queries can be used by the users to analyze relationships among the various dimensions and their associated data elements, aggregate data over time periods, and present data in multiple formats (e.g., graphical formats). Specific and unknown patterns in databases and data warehouses can be identified by data mining applications. These patterns cannot be revealed by typical queries. These applications can include one or more algorithms, including neural network algorithms, tree induction algorithms, and/or clustering algorithms. Users apply the algorithms to identify hidden patterns in the data (Lauría & Peter, 2004).

Decision support systems management consists of an ongoing, inseparable process of designing, implementing, and evaluating DSS. Decision support systems assist managers in their decision making processes to improve the effectiveness of decision making rather than its efficiency (Sean, 2004). Decision support systems have gained popularity in various domains such as business, engineering, military and medicine and are most valuable in situations where the amount of information is too large for the human decision maker to use optimally and with precision (Druzdzel & Flynn, 2004). In this research, we are particularly interested in business DSS, also known as business intelligent systems. This research will be applied to the Customer Relationship Management (CRM) component of Business Intelligence in the manufacturing and production companies.

2.2.1 Business Intelligence

Business intelligence (BI) is a data-driven Decision Support Systems (DSS) that combines data gathering, data storage, and knowledge management with analysis to provide input to the decision process. The term originated in 1989. Prior to that, many of its characteristics were part of executive information systems. Business intelligence emphasizes analysis of lager volume of data about the firm and its operations. It includes competitive intelligence (monitoring competitors) as a subset (Solomon & Paul, 2008).

Business Intelligence (BI) applies the functionality, scalability, and reliability of modern database management systems to build ever-larger data warehouses, and to utilize data mining techniques to extract competitive business advantage from the vast amount of available enterprise data (Mika & Virpi, 2002). BI systems combine data gathering, data storage and knowledge management with analytical tools to present complex and

competitive information to planners and decision makers (Solomon & Paul, 2003). Business Intelligence was defined in (Strauss et al., 2006) as the activity of gathering secondary data and primary information about competitors, markets, customers, and more. From the above definitions, it can be inferred that business intelligence is a combination of the following terms: ERP (Enterprise resource planning, EIS (Enterprise Information System), KM (Knowledge Management), CRM (Customer Relationship Management), DSS (Decision Support System), DM (Data Mining), GIS (Geographical Information System), OLAP (Online Analytical Processing) and Data Warehousing (DW). Business Intelligence is aimed at achieving the following goals: improving the timeliness and quality of input to the decision process; improving performance management; optimizing customer relations; monitoring business activity and traditional decision support; packaging standalone BI applications for specific operations or strategies; providing actionable knowledge delivered at the right time (Blumberg, 2003); offering better quality information (Mika & Virpi , 2002); offering better observation of threats and opportunities; improving the growth of the knowledge base; increasing sharing of information; improving efficiency; offering easier information acquisition and analysis; and offering cost savings (Mika & Virpi , 2002).

In BI, information is the raw material for decision making (Graham, 2004). Effective market decisions are therefore based on sound information and the decisions are not better than the information on which they are based. Information is therefore the lubricant of Business Intelligence. The more information a firm has, the better the value it can provide to each customer and the better the prospects in terms of accuracy, timely and

relevant offerings (Strauss et. al., 2006). A complete marketing knowledge database includes all data about customers, prospects, and competitor, the analyses and outputs based on the data and access to marketing experts, all available 24/7 through a number of digital receiving appliances. The Internet and other technologies facilitate marketing data collection (Strauss et. al., 2006).

A marketing information system interacts with managers to access their information needs, develops the needed information from internal company records and marketing intelligence (competitive intelligence) activities and the marketing research process (Linus, 2006).

BIDSS (Business Intelligence Decision Support System) is divided into three layers: i) the bottom layer, which is the path of source data collected and the repository of data storage; ii) the middle layer which consists of a wide variety of intelligence software or tools; iii) the top layer, the user interface, in which the final report is viewed and delivered to users. A Business Intelligence Decision Support System must be deployed within an infrastructure or platform with the capabilities to implement the BIDSS process to support the range of applications best suited to every business requirement (Kreulen, 2002). Figure 2.2 is a diagram of the current BIDSS framework as described above.

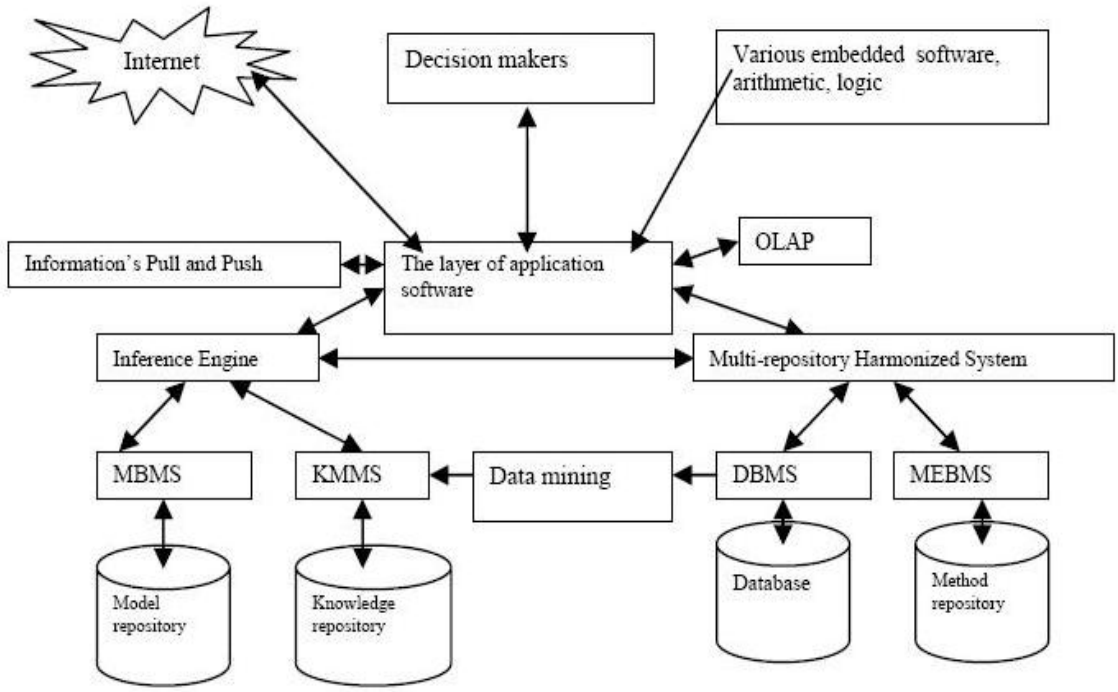


Figure 2.2 Current BI/DSS model (Monica, 2008).

The Figure 2.3 shows a framework that integrates the structured and semi structured data required for Business Intelligence. One implication of the BI framework is that semi-structured data are equally important, if not more, as structured data for taking action by planners and decision makers. A second implication is that the process of acquisition, cleanup, and integration applies for both structured and semi structured data (Solomon, 2004).

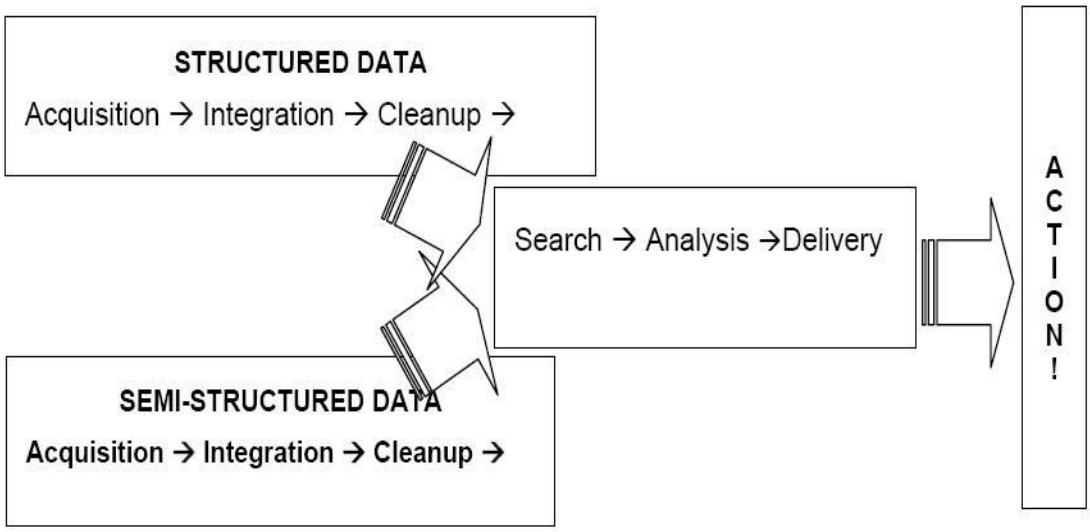


Figure 2.3 Business Intelligence data framework (Solomon, 2004).

2.2.2 Customer Relationship Management (CRM)

Customer relationship management is the component of Business Intelligence that deals with managing all customer interactions. Relationship marketing is the current paradigm in marketing that deals with attracting, maintaining and enhancing customer relationships. Relationship marketing depends solely on market research, which is concerned with the provision of information that can be used to reduce the level of uncertainty in decision making (Graham, 2004).

CRM is a strategic management system that manages all interactions and businesses with customers. It encompasses the capabilities, methodologies and technologies that are used to create and maintain lasting relationships with customers. It includes all the tenets of relationship marketing, grounded in customer data, and facilitated by technology (Strauss et. al., 2006). A typical CRM system is composed of three components: Operational CRM, Analytical CRM and Collaborative CRM. CRM is important because natural

customers' loyalty is a thing of the past (Linus, 2006) and companies are seeking to gain competitive advantage in today's stormy economy. Customer relationship management includes the following building blocks: CRM Vision, CRM strategy, Customer value experiences, Organizational collaboration, CRM process, CRM information, CRM Technology and CRM metrics. Customer service permeates every stage of the customer acquisition, retention and development practices. Report on one study showing customer service channels used by 60 firms revealed that information is stored most times in unstructured form (Strauss et. al., 2006). Since the CRM information component is the focus of this research project, more emphasis will therefore be placed on it. Data mining has changed the sales target of CRM systems from products to customers. How to classify customers? How to find out the common character of customers from database? How to dig up the potential customers? How to find out the most valuable customers? These kinds of questions become the most popular data mining applications in marketing (Xiaoshan, 2006). Current research in CRM includes; CRM portal design, development and maintenance, and updating to facilitate decision making (Asoo, 2002).

The grand knowledge discovery challenges in CRM also include;

• **Non-trivial results almost always need a combination of DM techniques.**

Analyzing CRM data requires exploring the data from different angles and looking at its different aspects. This should require application of different types of DM techniques.

• **There is a strong requirement for data integration before data mining.**

Data comes from multiple sources, for example in CRM, data needed may come from different departments of an organization. Since many interesting patterns span multiple data sources, there is a need to integrate these data before an actual data mining exploration can start.

• **Dealing with diverse data types when they are encountered.** There is need for the integrated mining of diverse and heterogeneous data.

• **Real-world validation of results is essential for acceptance.**

There is need to treat discovered patterns as hypothesis and test them on new data using rigorous statistical tests for the actual acceptance of results in DM applications. This is even more so for taking or recommending actions, especially in such high-risk applications as in the financial and medical domains.

• **Acquiring data for deeper understanding in a non-intrusive, low-cost, high accuracy manner.**

Data collection for CRM is still a problem in many industrial settings. Some methods are intrusive and costly. Datasets collected could be very noisy and in different formats and reside in different departments of an organization. Solving these pre-requisite problems is essential for data mining applications.

Deeper models of customer behavior:

In CRM, understanding customers is one of the key issues. Current models of customers mainly built based on their purchase patterns and click patterns at web sites. These type of models are very shallow and might not have a deep understanding of customers and their individual circumstances. Thus, many predictions and actions about customers are wrong.

• **Managing the "cold start" problem.**

Little is known at the beginning of the customer life cycle, but the list of customers and the amount of information known for each customer increases over time. Most times, a minimum amount of information is usually required for achieving acceptable results.

There is therefore a lot of challenge associated with being able to deal with cases where less than this required minimum is known.

• Highly and unavoidably noisy data must be dealt with.

In CRM, weblog data is associated with a lot of "noise". This noise is due to crawlers and missed hits because of the caching problem and so on.

• Legal considerations influence what data is available for mining and what actions are permissible.

There are some countries where it is not allowed to combine data from different sources or to use it for purposes different from those for which they have been collected.

• Evaluation framework for distinguishing between correct/incorrect customer understanding.

Asides from the difficulty of building customer models, evaluating them is also a major task. Satisfactory metric still does not exist to tell whether one model is better than another and whether a model really reflects customer behaviors. There also exist some metrics for measuring quality of customer models. Example of such includes metrics for measuring the quality of recommendations. These are quite rudimentary, and there is a strong need to work on better measures. Specifically, the recommender systems community has explored this area (Jaideep, 2010).

Finally, it was revealed in (Kernochan, 2006; Jaideep, 2010), that text integrated with business Data (structured) can provide valuable insights for improving the quality of business decisions. Also, in CRM systems identifying new customers is a critical task for any sales-oriented company. Of particular interest are companies that sell to other businesses, for which there is a wealth of structured information available through

financial and firmographic databases (Prem, 2007). In this research, (Prem, 2007) demonstrated that the content of company web sites can often be a richer source of information in identifying particular business alignments. They were able to establish that supervised learning can be used to build effective predictive models on unstructured web content as well as on structured firmographic data. This led to the establishment of the fact that there is a need to explore methods to leverage the strengths of both sources by combining these data sources. Apart from specific for-purchase marketing databases, there are several sources of data relevant to this task. These include:

1. Extensive financial information for publicly-traded companies (e.g. Standard and Poor's (http://www.standardandpoors.com))
2. Firmographic data (e.g. location, industry, estimated company revenue and number of employees) for a large number of companies (e.g. D&B (http://www.dnb.com))
3. News feeds (e.g. Reuters (http://www.reuters.com))
4. Content extracted from the websites of a universe of potential customers.

Any of these sources of data can be joined with the seller's historical transactions as a basis for building probability-to-purchase models (e.g. (Rosset & Lawrence, 2006). For example, D&B firmographic information can be joined with past transactions to build customer targeting models (Lawrence et al., 2007) that estimate purchase probabilities based a labeled set of positive examples, i.e. previous purchasers of a specific product. This scenario has introduced some very interesting machine learning issues in the emerging area of analyzing combined structured and unstructured data.

2.2.3 Competitive Intelligence (CI)

Competitive intelligence (CI) is a specialized branch of Business Intelligence. It is a systematic and ethical program for gathering, analyzing and managing external information that can affect the organization's plans, decisions and operations. CRM targets markets (customers) while competitive intelligence targets markets (customers) through industrial opportunities. Studies also show that CRM value strategies (operational excellence, innovation process and customer intimacy) and key dimensions of CRM harvesting reflect an overall organizational competitive advantage (Amin, 2008). The field of competitive intelligence has grown over the past two decades to become an integral part of most large organizations (John, 1999; Kahaner, 1996; McKinnon & Burns, 1992; Goshal & Westney, 1991). Global competition, the emphasis on quality management, and the realization by managers that actionable intelligence can be a key competitive advantage have spurred this growth (Prescott & P. Gibbons, 1993). Currently, the stage of development in competitive intelligence can be characterized as "Competitive Intelligence for Strategic Decision Making." The future rests on developing CI as a source of competitive advantage and is labeled "Competitive Intelligence as a Core Capability."

The sources of CI data include: government sources, online databases, interviews or surveys, special interest groups (such as academics, trade association and customer groups), private sector sources (such as competitors, suppliers, distributors, customers) and media (journals, wire services, newspapers, and financial reports). The collected data is transformed into intelligence through analysis. Analysis permits the CI professional to draw conclusions from information. Those conclusions then need to be interpreted in light of the original request leading to the production of implications and

recommendations. Unfortunately for many CI professionals, however, proficiency in analytical tools is often one of their weakest areas. Action-oriented CI is the result of producing implications and recommendations for managers (John, 1999).

The challenge with CI is not lack of information; it is the ability to differentiate useful CI from chatter or even misinformation and also maximize the richness of these heterogeneous information sources (Ukelson, 2006). Since one of the problems of competitive intelligence is the validity of the data used, data integration is proposed to be used to minimize this problem. For example, the source of competitive intelligence information is company's own propaganda, and this public relations activity can add texture to back ground statistical information through integrated mining.

Currently, mining from both unstructured data and XML are not naturally handled by the current generation of BI and integration tools (Ukelson, 2006). Companies are often seeking to associate unstructured content with structured data for example in the area of records management whereby invoices, statements and others (operational documents) need to be tied to customer data or supplier data. Review of the current problems faced in CI includes real time data warehousing, automated anomaly and exception detection, automatic learning and refinement and data visualization (Solomon & Paul, 2003) but the ability to differentiate useful CI information from chatter is most related to our research work.

2.3 INTEGRATED DATA MINING

2.3.1 Integrated Mining Problem Scenario

According to (Sukumaran & Sureka, 2007), integrated mining will be beneficial to this application area as it combines the unique attributes of both structured and unstructured data format in order to provide greater efficiency to the organization, especially by minimizing the popular practice of handling structured and unstructured as distinct information entities which often results in decision management failure. For example, products defects and warranty claims result in heavy costs to manufacturers. They suggested that companies can build early warning system that, by processing warranty data, helps in the early discovery of products and system failures. These warranty data is generated when a claim form is completed by a customer or a technician. This forms request for the following; product code, model number, date, time and customer ID. This information falls into the category of structured data. Most times, these forms also contain comments section where customer or technician can provide detailed information about the problem. The unstructured data part is the key to diagnosing and understanding the problem. An integrated analysis across the two forms of data (structured and text) might provide discoveries such as the trends of problems or faults exhibited by a particular model. It is clear that the concept of the model being complained about is not derivable from the unstructured data and at the same time, the structured data alone cannot tell us about the nature of the fault been diagnosed.

2.3.2 Review of Existing Integrated Mining Systems

Over the years, systems have been developed in order to achieve the purpose of integrated mining. A system was developed in (Frieder et. al., 2000) called SIRE (Scalable Information Retrieval Engine)**.** It is a relational information retrieval system that uses relations to model an inverted index. It stores full text in a relational environment and integrates the search of unstructured data with the traditional structured data search of relational database management systems. The drawbacks of this system include: (1) The problem of uncertainty of extracted features still persists due to the fact that semantic analysis is not involved in the retrieval process. (2) Mining in SIRE is limited to only information retrieval using SQL and not extended to data mining algorithms for the purpose of decision support. (3) SIRE is still prone to the generational information retrieval problems which includes high error rate, thereby producing unreliable reports.

In Roth et. al., (2000), an integrated architecture consisting of three tier was proposed: application, integration and foundation tiers. The application tier provides interfaces that allow applications to access and manipulate data and services provided by the foundation layer and integration tiers. The integration tier involves text search, combined text and parametric search and mining. The foundation tier offers a set of services to store and retrieve heterogeneous data. The limitations of the system include: (1) The mining algorithm is limited to the ones built into the foundation tier, which includes feature extraction, summarization and classification. Other mining algorithms could be used to mine the integrated data. (2) The architecture reveals a level of individual search of

structured and unstructured which is a disadvantage to the specific application area of this research work.

ESTEST (Experimental Software To Extract Structure from Text), developed in 2004 by (Dean & Alexandra, 2004), is a data integration approach that combines information extraction and data integration techniques (various sources) to better exploit text data. The data sources are first identified and integrated into a single global schema. This is done using AutoMed (http://www.doc.ic.ac.uk/automed; Dean & Alexandra, 2004). ESTEST then takes the metadata in the global schema and uses this to suggest input into the information extraction process. GATE (Cunningham et. al. 2002; Dean & Alexandra, 2004), IE (Information Extraction) architecture is used to build the ESTEST IE processor. The templates filled by the IE process will then be used to add to the extent of concept in global schema. Extracted annotations which match objects in the global schema will be extracted and put in the HDM (low-level graph-based data model) store. The global query facilities of AutoMed are now available to the user in order to query the global schema (Poulovassailis, 2001; Jasper, 2002; Dean & Alexandra, 2004). The drawbacks include: (1) it is not geared specifically to integrate structured and unstructured, but uses the combination of different structured sources to maximize extraction from text; (2) it is not detailed as regards data mining algorithms that is, the system stops at information retrieval.

SQUAD (Storing and Querying Unstructured Data) (Aravindan, 2005) is a unified framework for storing and querying unstructured and structured data. It aims at solving the problem of storing and querying unstructured and structured in two steps. It

introduces a new type of storage device called Intelligent Storage Node (ISN) to store, manage and search unstructured data. Using ISNs as a building block, proposed a new framework called SQUAD to seamlessly integrate structured and unstructured data. The limitations are that it performs exhaustive search which could be long running and I/ O intensive and the system stops at querying the database, no data mining algorithm was implemented.

The TSIMMIS system provides integrated access to heterogeneous information, stored in conventional databases, the Web, and legacy systems. The focus of this project is on semi-structured and/or unstructured information. This is information that may not conform to a rigid schema, and is frequently found, for instance, in the World-Wide-Web, SGML documents, semi-structured repositories. To represent such data, the authors use a "schema-less" object-oriented model, called Object Exchange Model (OEM) (Papakonstantinou et. al., 1995).

In their approach, Robert (2006) provided a means of browsing an adaptive database system with both structured and unstructured data through simplifying data structures and using subject carrying indexing information to order the data. The indexing terms or metadata are assigned probabilities, costs, and benefits, and a system that adapts its internal organization and its output to these user-based probabilities or costs, as well as to other metaphors (e.g., information theoretic) for user needs and interests is created. Their system integrates the information in a structure that can be optimally ordered for browsing, regardless of the type of source and the type of data, e.g., structured, semi-structured, or unstructured. Facts retrieved and presented are consistent with minimizing

the dissimilarity between adjacent facts, as well as the degree to which the facts match with the query.

In Prem (2007), the authors show how supervised learning can be used to build effective predictive models on unstructured web content as well as on structured firmographic data. Text are preprocessed by removing stop words, stemming the words into inflected forms (e.g. from the plural form to the singular form and from the past tense to the original form), and selecting features using the scores, which is shown to be the best feature-selection method in previous empirical studies (Yang & Pedersen, 1997). These processes result in a collection with a vocabulary of around 6000 words, which we convert into vectors using the bag-of-word representation with TF-IDF term weighting (Buckley et. al., 1994).

Sukumaran and Sureka, proposed an architecture, in 2007, that uses natural language processing and machine learning based techniques (text tagging and annotation) as a preprocessing step toward integrating structured and unstructured data. In the case of structured data sources, an ETL (Extract, Transform and Load) process executes the required formatting, cleansing and modification before moving data from transactional systems to the CDW (Combine Data Warehouse). For the unstructured data sources, the tagging and annotation platform extracts information based on domain ontology into an XML database. Extraction of data from an XML database into the CDW is accomplished with an ETL tool. This then materializes the unified data creation into the CDW. This architecture is not reported to have been implemented and the main component of the system which converts unstructured to semi structured (XML) is based on natural

language techniques and therefore still subject to the generational problems of information extraction such as high error rates thereby producing unreliable results.

Several approaches are being investigated to provide better integrated access to both unstructured and structured web sources with good scalability. For example, MetaQuerier provides unified entity search interfaces over many structured web sources of the hidden web (http://www.reuters.com). PayGo aims at providing web scale, domain-spanning access to structured sources (Madhavan, 2007). It tries to cluster related schemas together and to improve search results by transforming keyword search queries into structured queries on relevant sources. One aspect that is missing from such search approaches is the post-processing of heterogeneous search results (Erhard et al., 2007).

Finally, Oracle Database 11g incorporates native RDF (Resource Description Framework)/RDFS/OWL (Web Ontology Language) support in its ETL component, this makes for semantic data management. Individual application can be mapped to a standard information model order to make the meaning of the concepts in different application specific data schema explicit and relate them to each other. In order for Oracle 11g to handle data integration (that is, from various databases and also combination of structured and unstructured) the RDF and OWL models are integrated directly into the corporate DBMS, along with existing organizational data, XML and spatial information, and text documents (An Oracle White paper, 2007). Even though Oracle 11g has the facility to manage structured and unstructured using ontologies, the responsibility of creating ontologies lies on the application developer which makes it not tailored directly towards business intelligence. Also, though it handles structured and unstructured it is not directly built for storing data towards integrated mining.

The following is yet still another approach to integrating structured and unstructured data; a Loosely-Coupled Integration of a Text Retrieval System and an Object-Oriented Database System; this integration approach performs complex object retrieval using a probabilistic inference net model, and an implementation of this approach uses a loose coupling of an object oriented database system (IRIS) and a text retrieval system based on inference nets (IN QUERY). The resulting system is used to store long, structured documents and can retrieve document components (sections, figures, etc.) based on their text contents or the contents of related components (Croft et al., 1992).

2.3.3 Current state of Integrated Mining in CRM.

In the business decision support system, the CRM component to be specific, there are few integrated mining attempts. (Kernochan, 2006), was able to analyze both text and data using a particular approach based on an OLAP (on-line analytical processing) model enhanced with text analysis. They describe two tools that we have been developed to explore this approach—eClassifier performs text analysis, and Sapient integrates data and text through an OLAP-style interaction model. In this approach, work is still ongoing in the aspect of integrating text information into this OLAP system. Also another area of inadequacy in the system is that it promises to integrate (not yet integrated) ontologies into the taxonomy generation and dimension publishing portions of the BIKM (Business Intelligence and Knowledge Mining) Architecture. It is intended that these ontologies will provide a level of semantics that is not currently addressed, allowing improved taxonomies and reasoning about the data and text.

The SAS Text miner uses an integrated interface for analyzing text (unstructured data) in conjunction with multiple related database (structured) fields but it relies primarily upon pattern recognition technology instead of a linguistics-centric or dictionary-based approach (Arnold, 2010).

Finally, in (Zhu et. al, 2005) a system was proposed to query and analyze seamlessly across structured and unstructured data. It proposes an information system in which text analytics bridges the structured–unstructured divide. Annotations extracted by text analytic engines (TAE), with associated uncertainty, are automatically loaded into a structured data store. The interface is capable of supporting rich queries over this hybrid data. Uncertainty associated with the extracted information is addressed by building statistical models. It also shows that different classes of statistical models can be built to address issues such as ranking and OLAP style reporting. There is a prototype system called AVATAR that utilizes an existing commercial relational DBMS system as the underlying storage engine. The major limitation of this approach is data uncertainty. This is due to the fact that the particular algorithm underlying a TAE is limited in its understanding of text.

.

The above review of integrated mining systems reveals that the major area of contribution in such a system is in preprocessing of unstructured data. The result of the review is in agreement to the current problem in business decision support system which has to do with the fact that existing system have problem differentiating useful information from chatter. This further lead to reviewing existing and state of the heart preprocessing of unstructured data for mining.

2.4 UNSTRUCTURED DATA MINING (TEXT MINING)

Traditional data mining algorithms are generally applied on structured databases, but text mining algorithms try to discover knowledge from unstructured textual data (Basu, 2001). “Text mining” is used to describe the application of data mining techniques to automated discovery of useful or interesting knowledge from unstructured text (Han, 2000). Text mining is relatively a new research area at the intersection of natural language processing, machine learning and information retrieval. Several techniques have been proposed for text mining including conceptual structure, association rule mining, episode rule mining, decision trees, and rule induction methods. In addition, Information Retrieval (IR) techniques have widely used the “bag-of-words” model (Baeza-Yates, 1999; Raymond & Un Yong, 2005) for tasks such as document matching, ranking, and clustering. In order to reduce the work in handling huge amounts of textual data, various technologies have been developed. Information retrieval technology is probably the most common technology to use when we are faced with a very large number of documents. The term “text mining” (or “text data mining”) is sometimes used to indicate this technology because it detects and extracts documents that we want from mountains of documents, and it allows us to select data related to some specific topics that we are interested in so that the amount of data we have to handle is reduced without losing the information we want.

Generally, a text mining framework consists of the following components;

1. Concept extraction based on robust natural language processing
2. Data mining for discovering rules and patterns
3. Visualization and interactive analysis

Concept extraction for text mining; the term “concept” is a representation of the textual content in order to distinguish it from a simple keyword with the surface expression. There are certain issues in representing textual content. The first problem is because of the ambiguities in natural language, the same keyword may express entirely different meanings. For example, the word “Washington” may represent a person, place, or something else. The meaning of such polysemous words is normally determined according to their context. The inverse problem is that different expressions may refer to the same meaning, for instance, “car” and “automobile” or “H/W” and “hardware.” Even when the meaning may not be exactly the same, it may be necessary to treat these expressions as denoting the same meaning for text mining, especially when some of the synonyms are used infrequently, in order to avoid data sparseness, since a small number of appearances compared to others tend to be ignored in the final output.

Disco-TEX(Discovery from Text EXtraction) is an example of a text mining application that discovers prediction rules from natural language corpora using a combination of principles of information extraction and data mining. DiscoTEX (Taffet1, 2001; Joakim, 2000) in (Raymond & Un Yong, 2005) uses Information Extraction to obtain structured data from unstructured text and then use traditional KDD (Knowledge Discovery in Data Mining) tools to discover knowledge from this extracted data. It was applied to mine job postings and resumes posted to USENET newsgroups as well as Amazon book-description pages from the web. In DiscoTEX, IE plays the important role of preprocessing a corpus of text documents into a structured database suitable for mining. DiscoTEX uses two learning systems to build extractors, Rapier (Levy, 1996) in (Raymond & Un Yong, 2005) and BWI (Feldman & Dagan, 1995) in (Raymond & Un Yong, 2005). By training on a corpus of documents annotated with their filled templates,

these systems acquire pattern-matching rules that can be used to extract data from novel documents. Following the construction of an IE system that extracts the desired set of slots for a given application, a database is then constructed from a corpus of texts. This is done by applying the extractor to each document to create a collection of structured records. Standard KDD techniques is then applied to the resulting database to discover interesting relationships. DiscoTEX is used to induce rules for predicting each piece of information in each database field using all other information in a record. Prediction rules are discovered by treating each slot-value pair in the extracted database as a distinct binary feature, and learn rules for predicting each feature from all other features (Raymond & Un Yong, 2005).

TAKMI (Text Analysis and Knowledge MIning) is another text mining system that has been developed to acquire useful knowledge from large amounts of textual data such as internal reports, various technical documents, messages from various individuals, and so on (Nasukawa & Nagano, 2001) TAKMI analysis a large set of documents as a whole rather than focus on the specific information in each document. The most important issue for this text mining technology is how to represent the contents of textual data in order to apply statistical analysis. After the statistical analysis, it applies appropriate mining functions adapted to the representations of the original content of the text. Finally, since the content of the text varies greatly, it is essential to visualize the results and allow an interactive analysis to meet the requirements of analysts working from multiple points of view. The system was applied on some specific topics that were of interest. This technology is limited when we do not have a clear intention about what to search for and knowledge of what can be retrieved from the database we are searching. Moreover, even when we have some specific topics to search for and successfully make some queries, the

output we obtain is a list of documents that we still have to read to find the information, unless we are simply interested in such data as the number of documents that contain specific keywords or character strings (Nasukawa & Nagano, 2001).

In the proposed integrated mining system which is an improvement on the Sukumaran integration architecture above, the processes of mining from an intermediate data (combination of structured and unstructured in an XML data format) is reduced to text mining and the algorithm used for the mining is association rule mining algorithm. In the following section, we therefore intend to review the state-of-the-art as regards text mining using association rule mining algorithm.

Text mining systems consists basically of two major phases, the text preprocessing phase and the knowledge mining phase.

2.4.1 Preprocessing of unstructured data

This phase is aimed at optimizing the performance of the knowledge mining phase. There are many approaches to text preprocessing, one of which is mainly information retrieval. Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata that describe documents, or searching within databases, whether relational standalone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data. Information Retrieval (IR) deals with the representation, storage, organization of and access to information items. The models for text retrieval can be primarily divided into two categories: keyword oriented and matrix oriented (Jessup, 2001). Keyword

based models uses certain data structures and searching algorithms. Matrix oriented models changes the keyword representation of documents into a matrix format. Vector Space Model (VSM) is a conventional IR model, which represents a document collection by a term-document Matrix (Aswani & Srinivas, 2009).

Information extraction (IE) is a type of information retrieval whose goal is to automatically extract structured or semi structured information from unstructured machine-readable documents. A typical example is the extraction of information on corporate merger events, whereby instances of the relation ”MERGE (company1, company2, date)” are extracted from online news (“*Yesterday, New-York based Foo Inc. announced their acquisition of Bar Corp.*”).

The task of information extraction is therefore reduced to natural language processing, which can be defined as “the design and implementation of effective natural language input and output components for computational systems” (Joakim, 2000). The most important problems in Natural Language Processing is in relation to natural language input and output. The following are the few typical and uncontroversial examples of such problems:

- Part-of-speech tagging: Annotating natural language sentences or texts with parts-of-speech.
- Machine translation: Translating sentences or texts in a source language to sentences or texts in a target language.
- Natural language generation: Producing natural language sentences or texts from nonlinguistic representations.

Fundamentally, the steps involved in information extraction include the following (Taffet1, 2001):

- **Cleaning*:*** removes unwanted control characters, etc.
- **Tokenization:** adds spaces to separate text at boundary points between words and surrounding punctuation, or between different punctuation marks.
- **End-of-sentence detection:** identifies and marks sentence boundaries.
- **Part-of-speech tagging:** adds a tag indicating the part of speech for each token.
- **Phrase detection:** identifies and marks units that consist of multiple words - typically they are noun phrases of some type.
- **Entity detection:** identifies and marks entities, which usually consist of person names, place names, organization or company names and other proper nouns
- **Categorization:** identifies and marks what category something belongs to; typically categorization is used primarily for named entities (i.e. proper nouns)
- **Event detection: identifies** and marks events, which generally correspond to verbs.
- **Relation detection:** identifies and marks relations, which are connections between two or more entities or between entities and events.
- **XML or SGML:** applies the designated tagging scheme used to markup the document for tagging sentences, phrases, entities, categories, events, relations, etc.
- **Extraction:** the identified entities, events, relations, and any other identified concepts (such as dates) are extracted from the document and stored externally.

Literature has revealed that natural language processing systems employ statistical models and methods for processing (Joakim, 2000). Most examples of statistical application methods in the literature are methods that make use of a stochastic model, but where the algorithm applied to this model is entirely deterministic. Typically, the abstract model problem computed by the algorithm is an optimization problem which consists in

maximizing the probability of the output given the input. The following are some of the examples;

- Language modeling for automatic speech recognition using smoothed n-grams to find the most probable string of words out of a set of candidate strings compatible with the acoustic data (Jelinek, 1976; Bahl, 1983).
- Part-of-speech tagging using hidden Markov models to find the most probable tag sequence given a word sequence (Church, 1998; Cutting, 1992; Merialdo, 1994).
- Syntactic parsing using probabilistic grammars to find the most probable parse tree given a word sequence w_1; : : : ;w_n (or tag sequence t_1; : : : ; t_n) (Black, 1992, Stolcke, 1995; Charniak, 1997).
- Word sense disambiguation using Bayesian classifiers to find the most probable sense for word w in context c (Gale, 1992; Yarowsky, 1992).
- Machine translation using probabilistic models to find the most probable target language sentence t for a given source language sentence s (Brown, 1990; Brown, 1993).

TF*IDF (term frequency and inverse document frequency) is a commonly used weighting technique (a statistical technique) for information retrieval (Evans & Zhai, 1996). The weighing scheme is applied after POS tagging is used to attach a syntactic tag on a word, vector representation is done by selecting only words that are labeled with a noun tag, adjective tag or verb tag as features and finally, the words are stemmed.

Let w_i,d be a weight associated with a term t_i in a document page d. Then, the document vector d is defined as $d = \{w_1,d, w_2,d, \ldots,w_t,d\}$. Moreover, the weight of each attribute (vector) can be assigned either a boolean value or a *tf-idf* (*term frequency-inverse*

document frequency) value. Here, the *tf-idf* weight is a statistical measure used to evaluate how important a word is to a document in a collection; hence the learning algorithm could distinguish this relatively accurately. The *tf-idf* function assumes that the more frequently term t_i occurs in documents d_j, the more important it is for d_j, and furthermore the more documents d_j that term t_i occurs in, the smaller its contribution is in characterizing the semantics of a document in which it occurs. Weights computed by *tf-idf* techniques are often normalized so as to contrast the tendency of *tf-idf* to emphasize long documents. The type of *tf-idf* that will provide the normalized weights for data representation considered in our term classification is

$$tf - idf = tf(t_i, d_j).\log\frac{|D|}{\#_D(t_i)} \qquad \text{Eq. 2.1}$$

where the factor $tf(t_i, d_j)$ is called *term frequency*, the factor $\log\frac{|D|}{\#_D(t_i)}$ is called inverse document frequency, while $\#_D(t_i)$ denotes the number of documents in the document collection D in which term t_i occurs at least once and

$$tf(t_i, d_j) = \begin{Bmatrix} 1 + \log\#(t_i, d_j), if\ \#(t_i, d_j) > 0 \\ 0..otherwise \end{Bmatrix} \text{ Eq. 2.2}$$

where $\#(t_i,d_j)$ denotes the frequency of t_i occurs in d_j. Weights obtained by tf-idf function are then normalized by means of cosine normalization, finally yielding

$$w_{i,j} = \frac{tfidf(t_i, d_j)}{\sqrt{\sum_{k=1}^{|T|} tfidf(t_s, d_j)^2}} \qquad \text{Eq. 2.3}$$

In (Theobald & Weikum, 2002; Al-Khalifa & Jagadish, 2003; Fuhr & Grobjohann, 2001), information retrieval related features such as ranking and relevance-oriented search has been proposed to be integrated with XML query languages.

Currently, research is gradually moving from the statistical approach to natural language processing to semantic analysis (Evans & Zhai, 1996).

2.4.1.1 Text preprocessing with Information Extraction

Over the past two decades, significant efforts have focused on the problem of extracting structured information (e.g., researchers, publications, co-author and advising relationships, etc.) from such data. The information extracted is then exploited in search, browsing, querying, and mining. Recently, the explosion of unstructured data on the World Wide Web has generated significant further interests in the above extraction problem. This interest is the central research goal in the database, AI, data mining, IR, NLP, and Web communities' extraction (AnHai et. al., 2006).

According to (Andrew & David, 2003), finite state machines are the dominant model for information extraction both in industry and research. Hidden Markov models, which is a finite state machine whose parameters are set by machine learning, have parameters for state-to-state transition probabilities and per-state observation emission probabilities. This makes it easy to calculate the probability that the model would have generated a particular state sequence associated with a particular observation symbol sequence. When used for extraction, the emission symbols are with different extraction fields. Hidden Markov model for example, have two states, to extract person names, one for person-names, and one for other. The Viterbi algorithm is used to perform extraction on a particular word sequence, to find the state sequence most likely to have generated the

observed word sequence, and then designates as person names any words Viterbi that claims were generated while in the person-name state.

According to Andrew & David (2003), the disadvantage of hidden Markov models is that, being generative models of the observation sequence, they are limited in their ability to represent many non-independent, overlapping features of the sequence. Consequently, since the observations are generated by the model, the model must represent any correlations between features in order to faithfully reproduce them. If there are many correlated features, or complex dependencies among them, this modeling is prohibitively difficult, and in many cases impossible.

As a result of the above described disadvantage, Andrew & David (2003), proposed the use of unified, relational, undirected graphical models for information extraction and data mining, such that extraction decisions and data-mining decisions are made in the same probabilistic "currency," with a common inference procedure where each component is able to make up for the weaknesses of the other and therefore improving the performance of both. An example is that data mining run on a partially filled database can find patterns that provide "topdown" accuracy-improving constraints to information extraction. Information extraction can therefore provide a much richer set of "bottom-up" hypotheses to data mining if the mining is set up to handle additional uncertainty information from extraction.

Marios et. al., (2003) also proposed an approach that is based on using hierarchical hidden Markov models to represent the grammatical structure of the sentences being processed. Their approach uses a shallow parser to construct a multi-level representation of each sentence being processed. After which they trained hierarchical HMMs to capture

the regularities of the parses for both positive and negative sentences. They evaluated their method by inducing models to extract binary relations in three biomedical domains.

In their work, Ganesh et. al, (2006) approached information extraction using Inductive Logic Programming (ILP). Specifically, they demonstrated the use of ILP to define features for seven IE tasks using two disparate sources of information. Their findings reveal that the ILP system is able to identify efficiently large numbers of good features. Typically, the time taken to identify the features is comparable to the time taken to construct the predictive model. They also discovered that SVM (Support Vector Machines) models constructed with these ILP-features are better than the best reported to date that rely heavily on hand-crafted features. For the ILP practitioner, they also present evidence supporting the claim that, for IE tasks, using an ILP system to assist in constructing an extensional representation of text data (in the form of features and their values) is better than using it to construct intentional models for the tasks (in the form of rules for information extraction). This explanation is captured in Figure 2.4 below.

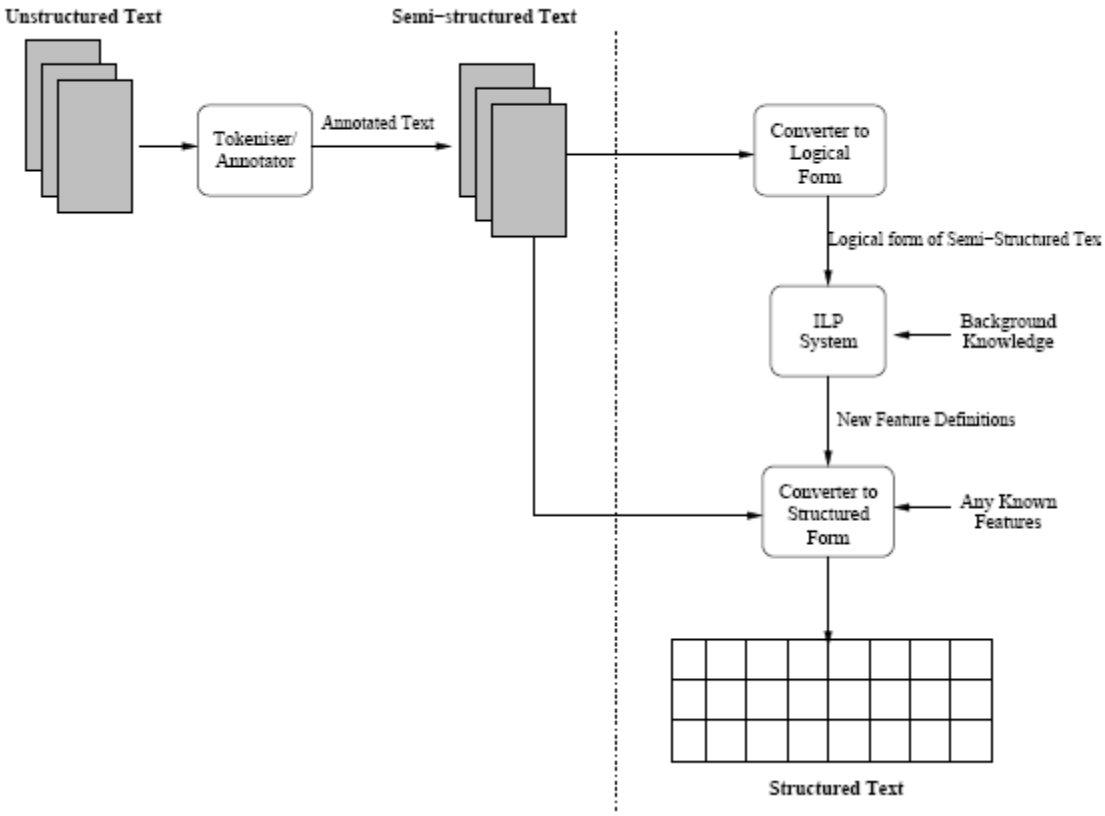


Figure 2.4 A simplified view of a role for ILP in information extraction (Ganesh et. al, 2006).

One type of IE, named entity recognition, involves identifying references to particular kinds of objects such as names of people, companies, and locations (Bikel & Weischede, 1999). In addition to recognizing entities, an important problem is extracting specific types of relations between entities. For example, in newspaper text, one can identify that an organization is located in a particular city or that a person is affiliated with a specific organization (Zelenko et. al., 2003). IE can also be used to extract fillers for a predetermined set of slots (roles) in a particular template (frame) relevant to the domain.

In their work, (Raymond & Razvan, 2005) considered the task of extracting a database from postings to the USENET newsgroup, Austin jobs. Another application of IE is extracting structured data from unstructured or semi-structured web pages. When applied

to semi-structured HTML, typically generated from an underlying database by a program on a web server, an IE system is typically called a wrapper (Kushmerick et al., 1997) and the process is sometimes referred to as screen scraping. A typical application is extracting data on commercial items from web stores for a comparison shopping agent (shopbot) (Doorenbos et al., 1997) such as MySimon (www.mysimon.com) or Froogle (froogle.google.com).

Text mining concerns looking for patterns in unstructured text, it is capitalizes on locating specific items in natural-language documents (which can be termed Information Extraction (IE)). A framework for text mining by (Raymond & Un Yong, 2005) called DISCOTEX, uses a learned information extraction system to transform text into more structured data which is then mined for interesting relationships. It integrates an IE module acquired by an IE learning system, and a standard rule induction module. Also added is the rules mined from a database extracted from a corpus of texts, used to predict additional information to extract from future documents, thereby improving the recall of the underlying extraction system. Figure 2.5 is the general overview of an Information Extraction based Text mining framework.

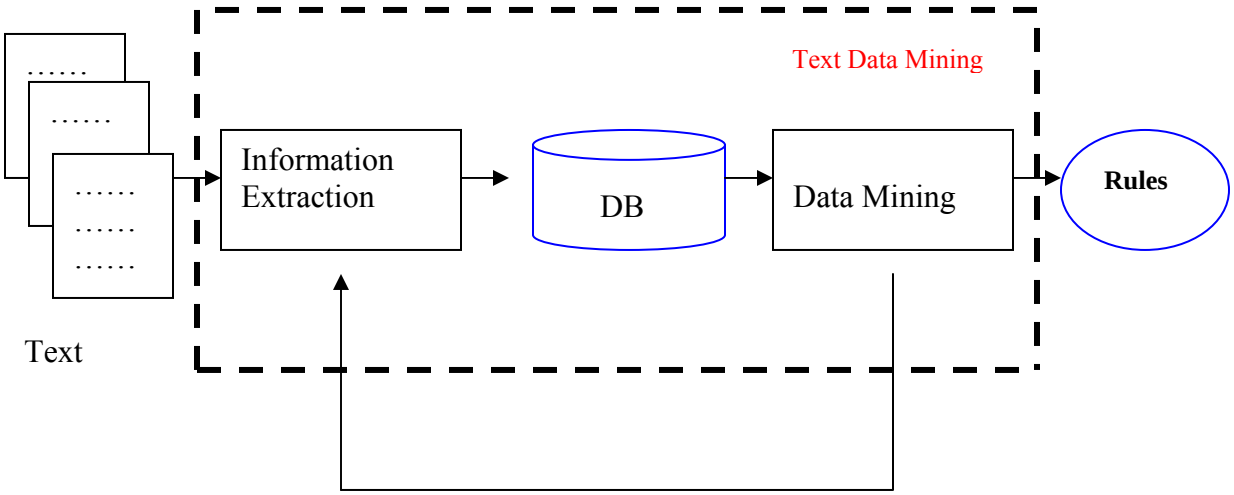


Figure 2.5: Overview of IE-based text mining framework (Raymond & Un Yong, 2005).

Furthermore, the following is a list of projects that addresses information extraction (AnHai et. al., 2006):

- Entity matching and approximate joins at AT&T Research, MSR and Stanford.
- Answering structured queries over text at Columbia and UCLA.
- Personal information management and intelligent email (PIM) at CMU, Massachusetts, MIT and Washington.
- Querying and Extracting semantic entities/relations at IIT Bombay, CMU, MSR and Washington.
- Data cleaning at MSR.
- Doing OLAP-style analysis using extracted information at IBM Almaden and Wisconsin.
- Standardization efforts at IBM Watson on interfaces for NLP extraction tools.
- Managing unstructured data in bioinformatics at Illinois and Michigan.
- Web-based community information management (CIM) at Illinois and Wisconsin.

2.4.1.2 Using ontology for semantic information extraction/ text preprocessing

Ontology is a branch of philosophy that attempts to model things as they exist in the world (Shanmugasundaram et al., 1999). It is particularly appropriate for modeling objects including their relationships and properties (EXCELON CORP, 2002). A domain ontology is a vocabulary of concepts and their relationships for that given domain, which defines the domain semantics (Viral, 2004). Other benefits that could be derived from the use of ontologies apart from knowledge sharing are reusability of domain knowledge and separation of domain knowledge from operational knowledge (Noy, 200). Ontologies are viewed as the most advanced knowledge representation model.

Efficiency gains in information extraction are realized by formalizing concepts and the relations of these concepts which are to be searched in documents. This formalization can be done through specification of a taxonomy of concepts, or more generally an ontology. Different domains require different categories of terms, phrases, and concepts (Goffman, 1992). Formalizing the coding schemes and organizing the knowledge extracted can be aided by the development of ontology for specifying and relating document characteristics and concepts of interest. Therefore, integration of an ontological backbone into information extraction tool will result in better coding consistency. This is because the target categories will be clearly defined and the ontology will be able to establish a common controlled vocabulary for concepts (Karin, 2004).

After clarifying the usage of the term ontology, a variety of methods have been used to extract information from text using ontologies (Sánchez & Moreno, 2005; Leveling & Hartrumpf, 2005). Taxonomies that are built automatically from web data are used by

Sánchez and Moreno in (Sánchez & Moreno, 2005) to group query results returned by a search engine. In this case, the user's behavior of accepting or rejecting the interface is the instance of judgment. Improving question answering by overcoming the shortfalls of the bag-of-words model is the objective of Leveling and Hartrumpf in (Leveling & Hartrumpf, 2005). Here, a semantic lexicon forms the background knowledge for semantic parsing, which yields a semantic representation much more precise than simply considering presence or absence of terms. Despite the immense efficiency introduced in information extraction by using domain specific ontologies, there is a need to promote ontologies with high coverage as applications are usually tested in a generic rather than in a domain-specific setting. Research reveals that using a semantic resources such as WordNet (Nasukawa T. & Nagano, 2001) as additional features, could bring about more gain though small in magnitude due to lack of subject coverage.

Also of not is that, the enormous amount of information existing in natural language form can be automatically processed and analyzed, if it is first be distilled into a more structured form, from which individual facts are accessed. Information extraction and wrapping technologies therefore offer the potential for selective information structuring: extracting selected data from documents and structuring such data in order to make it processable in enterprise applications. There exist recently, a variety of information extraction and wrapping applications for which XML is the preferred output format. (Andrea & Sergio, 2010).

Recently, there exist several approaches to XML data storage and management which fall into four main categories: flat files (Sahugue, 2000), relational database systems (Deutsch et al., 1999; Shanmugasundaram et al., 1999), object oriented database systems (EXCELON CORP, 2002; Lahiri, 1999; Runapongsa & Patel; 2002) and native XML

repositories (Jagadish et al. 2002; Fiebig, 2002). Also, ACM Transactions on Information Systems (TOIS) recently devoted a special issue to advances in XML retrieval (Baeza-Yates & Ribeiro-Neto, 1999). The presence of structure and content information in XML data enables us to devise various scenarios of data management and knowledge discovery for which it might be advisable to consider structure features alone, or content features alone, or even features of both kind. Early approaches to structural similarity detection are based on tree edit distances, which allow for computing a minimum cost sequence of edit operations to align a pattern document to a target document. In (Nierman & Jagadish, 2002) an XML-aware edit distance is exploited in a standard hierarchical clustering algorithm to evaluate how closely clustered documents correspond to their respective DTDs. In general, computing tree edit distances turns out to be unpractical, as it requires a quadratic number of comparisons between document elements. To address this issue, the level similarity measure is introduced in (Nayak. & Xu, 2006), to compute the structural match between elements according to the level of information of each object. Elements in different level positions are differently weighted, whereas hierarchical relationships are taken into account by counting occurrences of common elements sharing common ancestors. A different approach to similarity detection is proposed in (Flesca et al., 2003), where the structure of an XML document is represented as a time series in which each occurrence of a tag corresponds to an impulse, and the degree of similarity between documents is computed by analyzing the frequencies of the corresponding Fourier transforms. Mining XML data from a structure/content combination point of view has attracted significant attention in the last few years. The XML document mining track at INEX has been proposed as a major contest for researchers since 2005, with a special focus on clustering and categorization (Denoyer, L. and Gallinari, 2007). Early

attempts in XML document clustering by structure and content are given in (Guillaume & Murtagh, 2000); In (Guillaume & Murtagh, 2000), clustering of data-centric XML documents is seen as a partitioning problem based on a weighted graph-based representation, in which nodes are documents, edges are links between documents, and edge weights are computed by considering keywords in the documents. An alternative representation of XML data, called BitCube, is presented in (Yoon, 2001) as a 3-dimensional bitmap index of triplet document, XML-element path, word. BitCube indexes can be manipulated to partition a document set into clusters, by exploiting bit-wise distance and popularity measures. However, the approach suffers from the typical disadvantages of Boolean representation models, such as the lack of partial matching criteria and natural measures of document ranking. Vector-space models have been increasingly used to represent XML data, especially text-centric XML documents (Andrea & Sergio, 2010). It should be noted that none of these approaches take semantic aspects into account in handling information available from XML structure and content. To date, no approach is effective for the unsupervised semantic organization of XML data. In this respect, our work differs from existing ones in that it originally addresses the XML document clustering problem from the more complex perspective of semantic relatedness. Moreover, our approach uses a tree-tuple based decomposition of the documents that is particularly suitable to represent semantically cohesive portions of the individual documents.

2.4.2 Knowledge Mining

2.4.2.1 Association Rule Mining

The formal statement of the association rule problem is stated below: (Agrawal et. al., 1993), (Cheung et. al., 1996) and (Margaret et. al., 1999):

Definition 1: Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct attributes, also called *literals*. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$. An *association rule* is an implication of the form $X \Rightarrow Y$, where X, $Y \subset I$, are sets of items called *itemsets*, and $X \cap Y = \phi$. Here, X is called antecedent, and Y consequent.

The two important measures for association rules, support (s) and confidence (c), can be defined as follows:

Definition 2: The *support (s)* of an association rule is the ratio (in percent) of the records that contain $X \cup Y$ to the total number of records in the database.

Definition 3: For a given number of records, *confidence (c)* is the ratio (in percent) of the number of records that contain $X \cup Y$ to the number of records that contain X. Mining of association rules from a database consists of finding all rules that meet the user-specified threshold support and confidence.

Apriori

(Agrawal & Srikant, 1994) developed the Apriori algorithm. It is a great achievement in the history of mining association rules according to (Cheung et. al., 1996). It is by far the most well-known association rule algorithm. Apriori technique uses the property that any subset of a large itemset must be a large itemset. Also, it is assumed that items within an itemset are kept in lexicographic order. These common itemsets are extended with other individual items in the transaction to generate candidate itemsets. However, those individual items may not be large. A superset of one large itemset and a small itemset will result in a small itemset, these techniques generate too many candidate itemsets which turn out to be small. The Apriori algorithm therefore addresses the issue just

mentioned. The Apriori generates the candidate itemsets by joining the large itemsets of the previous pass and deleting those subsets which are small in the previous pass without considering the transactions in the database. By only considering large itemsets of the previous pass, the number of candidate large itemsets is significantly reduced.

In the first pass, the itemsets with only one item are counted. The discovered large itemsets of the first pass are used to generate the candidate sets of the second pass using the apriori_gen() function. Once the candidate itemsets are found, their supports are counted to discover the large itemsets of size two by scanning the database. In the third pass, the large itemsets of the second pass are considered as the candidate sets to discover large itemsets of this pass. This iterative process terminates when no new large itemsets are found. Each pass *i* of the algorithm scans the database once and determines large itemsets of size *i*. L_i denotes large itemsets of size *i*, while C_i is candidates of size *i*.

The apriori_gen() function as described in (Agrawal & Srikant, 1994) in (Margaret et. al., 1999) has two steps:

Step 1: L_{k-1} is joined with itself to obtain C_k.

Step 2: apriori_gen() deletes all itemsets from the join result, which have some (k-1)–subset that is not in L_{k-1}. Then, it returns the remaining large k-itemsets.

Method: apriori_gen() (Agrawal & Srikant, 1994)

Input: set of all large (k-1)-itemsets L_{k-1}

Output: A superset of the set of all large k-itemsets

//Join step

I_i = Items i

insert into Ck

Select $p.I_1$, $p.I_2$, ……. , $p.I_{k-1}$, q .I_{k-1}

From L_{k-1} is p, L_{k-1} is q

Where $p.I_1 = q.I_1$ and …… and $p.I_{k-2} = q.I_{k-2}$ and $p.I_{k-1} < q.I_{k-1}$.

//pruning step

forall itemsets c ϵ C_k do

forall (k-1)-subsets s of c do

If ($s \notin L_{k-1}$) then

delete c from C_k

Figure 2.6 appriori_gen() (Margaret et. al., 1999)

In Table 2.1, large itemsets after the third pass are shown in the first column. Suppose a transaction contains {Plantain, Bagel, Chicken, Eggs, Fanta}. After joining L_3 with itself, C_4 will be {{ Plantain, Bagel, Chicken, Fanta }, { Plantain, Chicken, Fanta, Eggs}. The prune step deletes the itemset { Plantain, Chicken, Fanta, Eggs} because its subset with 3 items { Plantain, Fanta, Eggs} is not in L_3.

Table 2.1 Appriori Example

Large Itemsets in the third pass (L_3)	Join (L_3, L_3)	Candidate sets of the fourth pass (C_4 after pruning)
{{ Plantain , Bagel, Chicken}, { Plantain , Bagel, Fanta }, { Plantain , Chicken, Fanta }, { Plantain , Chicken, Eggs}, {Bagel, Chicken, Fanta }}	{{ Plantain , Bagel, Chicken, Fanta }, { Plantain , Chicken, Fanta , Eggs}}	{{ Plantain , Bagel, Chicken, Fanta }}

Table 2.2 Transaction Table

Transaction ID	Items
T1	Bread,Butter,Eggs
T2	Butter, Eggs, Milk
T3	Butter
T4	Bread,Butter

```
Function count(C: a set of itemsets, D: database)
begin
        for each transaction T ϵ D=∪D^i do begin
                forall subsets x ⊆ T do
                        if x ϵ C then
                                x.count++;
        end
end
```

Figure 2.7 function count () (Margaret et. al., 1999)

Apriori (Agrawal & Srikant, 1994)

Input:

I, D, s

Output:

L

Algorithm:

//Apriori Algorithm proposed by Agrawal R., Srikant, R. (Agrawal & Srikant, 1994) in (Margaret et. al., 1999

//procedure LargeItemsets

1) C 1: = I; //Candidate 1-itemsets

2) Generate L1 by traversing database and counting each occurrence of an attribute in a transaction;

3) **for** (k = 2; Lk-1; k++) **do begin**

//Candidate Itemset generation

//New k-candidate itemsets are generated from (k-1)-large itemsets

4) Ck = apriori-gen(Lk-1);

//Counting support of Ck

5) Count (Ck, D)

6) Lk = {c ϵ Ck | c.count ε minsup}

7) **end**

9) L:= ∪kLk

Figure 2.8 Apriori Algorithm (Margaret et. al., 1999)

The illustration of how the Apriori algorithm works on the following example is shown in Figure 2.9;

Consider a small database with four items I={Bread, Butter, Eggs, Milk} and four Transactions as shown in Table 2.2. Suppose that the minimum support and minimum confidence of an association rule are 40% and 60%, respectively. There are several potential association rules. At first, we have to find out whether all sets of items are large. Secondly, we have to verify whether a rule has a confidence of at least 60%. If the above conditions are satisfied for a rule, we can say that there is enough evidence to conclude that the rule holds with a confidence of 60%. Itemsets associated with the aforementioned rules are: {Bread, Butter}, and {Butter, Eggs}. The support of each individual itemset is at least 40%. Therefore, all of these itemsets are large. It is evident that the first rule (Bread $\Rightarrow$ Butter) holds. However, the second rule (Butter $\Rightarrow$ Eggs) does not hold because its confidence is less than 60%.

Initially, each item of the itemset is considered as a 1-item candidate itemset. Therefore, C_1 has four 1-item candidate sets which are {Bread}, {Butter}, {Eggs}, and {Milk}. L_1 consists of those 1-itemsets from C_1 with support greater than or equal to 0.4. C_2 is formed by joining L_1 with itself, and deleting any itemsets which have subsets not in L_1. This way, we obtain C_2 as {{Bread Butter}, {Bread Eggs}, {Butter Eggs}}. Counting support of C_2, L_2 is found to be {{Bread Butter}, {Butter Eggs}}.

Using apriori_gen(), we do not get any candidate itemsets for the third round. This is because the conditions for joining L_2 with itself are not satisfied.

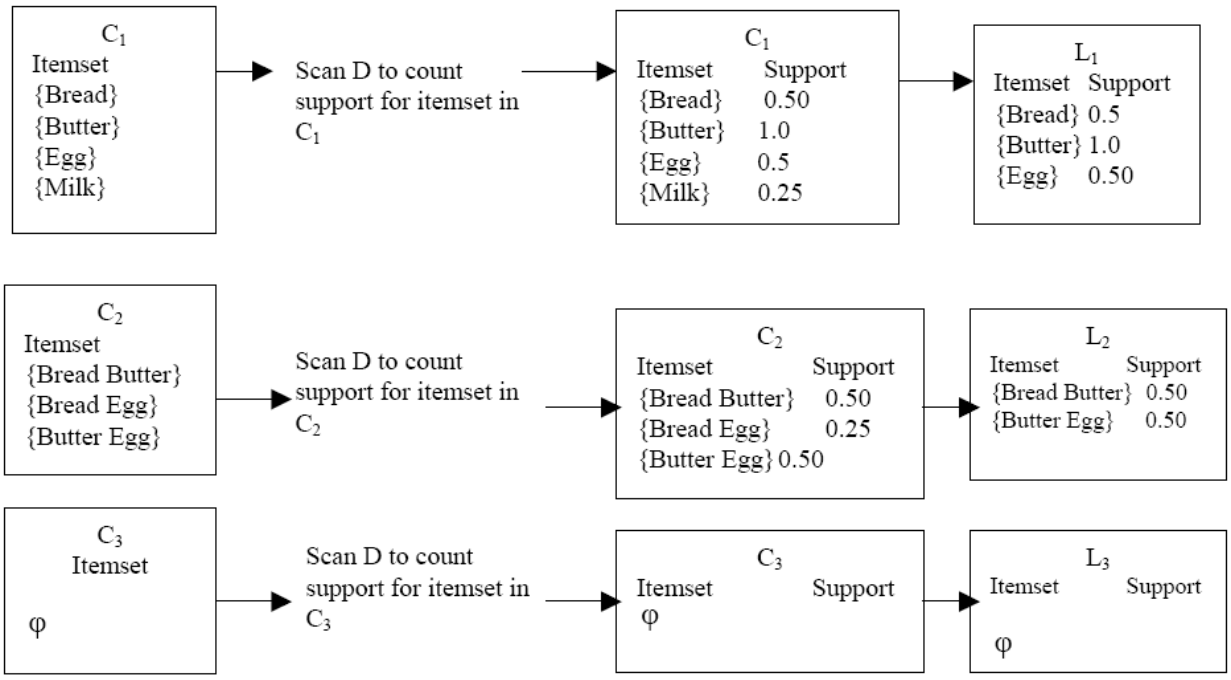


Figure 2.9 Discovering Large Itemsets using the Apriori Algorithm (Margaret et. al., 1999)

2.4.2.2. Knowledge Mining from XML data

Organizations are increasingly employing machines capable of generating semi-structured (XML-like) text data (for example, projections by IBM in that corporation's Global Technology Outlook for 2003 suggests that by 2010, nearly 75% of the data stored in an organization may be of this type). This trend has led to a substantial industrial need to develop automated methods for extracting information of potential commercial interest from such data (Ganesh et. al, 2006).

Since XML allows the definition of semantic markup that is, customized tags describing the data enclosed by them. The increase in volume and heterogeneity of XML-based application scenarios has made data sources exhibit both different structures and contents

and different ways to semantically annotate the data. As a result of this, differently annotated XML documents may refer to similar concepts, thereby being semantically related to a certain degree. Dealing with XML documents can be simplified if they are explicitly associated with a schema. This schema can be used to specify the content models of document elements and their relationships. For example, XML documents having different element values but similar schemas could be put together in the same structural class. Since most real world XML sources do not provide schema, exploiting schemas for organizing XML data may not be always feasible in practice. Organizing XML data from a large collection is central in the context of XML data management and knowledge discovery. Most research on XML clustering centers around the ever increasing interest for developing solutions to the document clustering problem, which has been studied intensively because of its wide applicability. There exists a difficulty of devising suitable notions of semantic features and semantic relatedness among XML documents. This difficulty is caused by the weak support offered by traditional models for representing and understanding XML documents. Structural models of XML documents are based on tree or graph paradigms, whereas content models usually refer to the vector-space model. Another issue is related to the generation of XML features which are able not only to bundle structure and content information together, but also to handle the semantics of structure and content (Andrea & Sergio, 2010).

Currently, knowledge has been mined from XML data using different data mining techniques, such as SOM (Self Organized Map) and k-means clustering. Of note is the XK-means (k-means algorithm adapted for xml data) algorithm which is developed in two main phases. In the first phase, it works as k -Means to compute k + 1 clusters: starts choosing k objects as the initial cluster centroids, then iteratively reassigns each remaining

object to the closest cluster and recomputes the centroid of clusters, until all the cluster centroids do not change. The (k + 1)-th cluster, the trash cluster, is created to contain unclustered object i.e. objects not assigned to any of the early k clusters. The second phase recursively splits the trash cluster into a small number of clusters (Andrea & Sergio, 2010). In this project, association rule mining technique is used for knowledge discovery. According to (Feldman & Dagan, 1995) and (Feldman & Hirsh, 1996), association rules have been mined from manually assigned keywords. This method has the following disadvantages: it is time consuming, subjected to discrepancy and the textual sources are constrained to those that have the keywords predetermined. In (Rajman & Besancon, 1997), two examples of text mining task were presented, association extraction and prototypical document extraction, along with several related NLP techniques. In the case of association extraction task, association rules were extracted from a collection of indexed documents. In their work on mining association rule form biomedical text, they performed entity extraction using BioTeKS which aims at both identifying the location of an entity in a text and categorize it according to the standard MeSH (Medical Subject Headings) taxonomy (Berardi et. al., 2005). This makes the application restricted to only the medical domain. In their work on the survey of basic concepts in the area of text data mining and some of the methods used in order to elicit useful knowledge from collections of textual data, the authors suggested that there has been some minor attempts to use (partially or fully) structured textual documents such as HTML or XML documents in order to develop text mining systems (Jan P. & Peter, 1999). There have also been some approaches to text mining with information extraction as the text preprocessing phase. This approaches inherits the generational problems of information extraction systems which includes uncertainty of extracted features. Some of the approaches include that of

(Shenzhi et al., 2004) where an algorithm that learns rules and extracts entities from unstructured textual data was developed. In (http://www.reuters.com) semi-automatic ontology based text annotation (OnTeA) tool is used to analyze document or text using regular expression patterns and detects equivalent semantics elements according to the defined domain ontology which can then be used as input to a text mining system, again, this approach is domain specific.

In addition to the above, knowledge has also been mined from XML data using association rule mining. This can be done by mapping the XML documents to relational data model and to storing them in a relational database. This allows standard tools to be used to perform rule mining from relational databases. Though this approach makes use of existing technology, it is often time consuming and involves manual intervention because of the mapping process. The above stated reasons makes it not quite suitable for XML data streams (Ding & Gnanasekaran, 2007).

Recently World Wide Web consortium introduced an XML query language called XQuery (Brundage, 2004). This query language addresses the need to intelligently query XML data sources. The query language is also flexible enough to query a broad spectrum of XML information sources, including both databases and documents. Naturally this led to the use of XQuery to perform the association rule mining directly from XML documents. Since XQuery is designed to be a general purpose XML query language, it is often very difficult to implement complicated algorithms. So far only the Apriori algorithm has been implemented by using XQuery (Wan & Dobbie, 2003). It has been raised as an open question in (Wan & Dobbie, 2003), whether or not FP-Growth

algorithm can be implemented by using XQuery, and there is no such implementation available at this point.

The other approach is to use programs written in a high level programming language for this task. Most of such implementations require the input to be in a custom text format and do not work with XML documents directly. In order to adopt this approach to XML rule mining, it requires an additional step to convert the XML documents into the custom text files and apply these tools.

Ding & Gnanasekaran (2007) looked at the various approaches for association rule mining from XML data. They presented a Java-based implementation of the Apriori and the FP-Growth algorithms for this task and compared their performances. They also compared the performance of their implementation with an XQuery-based implementation. Their findings revealed that the Java based approach proposed by them performed very well against the one that we compared.

2.5 EVALUATION

Evaluation of DSS is concerned with analyzing costs and benefits of decision support systems before and after DSS development and implementation. DSS evaluation is often a difficult problem though some DSS provides substantial cost saving and profit increase. This difficulty is due to the fact that quantification of the positive impacts of improved decision process is difficult (Keen & Scott, 1978). Evaluating DSS is concerned with determining the value of DSS. The value of DSS can be measured by a smorgasbord of eight methodologies: (1) decision outputs; (2) changes in the decision process; (3) changes in managers' concepts of the decision situation; (4) procedural changes; (5)

classical cost/benefit analysis; (6) service measures; (7) managers' assessment of the system's value; and (8) anecdotal evidence (Keen & Scott, 1978). In this project, we have chosen to measure the value of our DSS using the decision output since this is directly related to the result of our system which is a set of rules upon which decisions can be made.

Data mining tools tend to produce a huge number of patterns which makes it difficult for users to find interesting and useful ones quickly and easily (Xin & Yi-Fangm, 2006). According to (Xin & Yi-Fangm, 2006): most of the rules generated from data mining systems are not useful, and those "that come out at the top, are things that are obvious". This problem is even compounded in text mining because of large number of documents and the high dimensionality of textual data. In evaluating the interestingness of association rules, there are objective and subjective methods, which have been proposed in (Liu et al., 2001; Padmanabhan & Tuzhilin, 1999; Piatesky-Shapiro & Matheus, 1994; Silberschatz & Tuzhilin A., 1996).

Objective methods are insufficient because they rely only on the characteristics (surface features) of the patterns and the underlying data collection without considering users' knowledge and interests. This is a big disadvantage because, a large number of rules can be generated that are interesting "objectively" but of little interest to the user (Klemettinen, 1994). Subjective measures, such as unexpectedness (a pattern is interesting if it is "surprising" to the user) and actionability (a pattern is interesting if the user can act on it to his/her benefit), Silberschatz & Tuzhilin (1996), assess the interestingness of patterns from the users' perspective, but they require explicit expressions of users' subjective opinions (expectation/unexpectation) in order to perform the comparison. This is difficult or even nearly impossible in practice for users to do,

especially before the discovered patterns are presented to them. In this project, we will be evaluating IIMA (Improved Integrated Mining Architecture) system using a new method of estimating the novelty of rules discovered by data-mining methods using WordNet (Basu et al., 2001). WordNet is a lexical knowledge-base of English words. We will assess the novelty of a rule by the average semantic distance in a knowledge hierarchy between the words in the antecedent and the consequent of the rule - the more the average distance, more is the novelty of the rule. This method of evaluation is used because according to (Basu et al., 2001), by computing correlation coefficients between pairs of human ratings and between human and automatic ratings, this method was found to correlate with human judgments as well as human judgments correlate with one another (Basu et al., 2001).

2.6 THE CONTEXT OF THIS RESEARCH

From the foregoing issues, a number of gaps exist in literature which defines the context of this research. The first is the need for the generation of more dependable decision making rules which have not been adequately addressed by existing integrated data environments examined. This gap becomes the premise for the central research question being investigated in this thesis, which is:

How do we improve result gotten from the integrated mining system in order to reduce decision failure?

2.7 SUMMARY

The chapter presents the issues that define the research context of this thesis. It started by introducing decision support systems in the field of customer relationship management which is intended to expose the current state of the application area to reveal the need for an improved integrated mining framework in such a domain. Furthermore it gives the current state of integrated data issues generally as regards customer relationship management. This is followed by a detailed review of the state-of-the-art as regards information extraction which is the area we intend to introduce the contribution of thesis research work. Finally, we introduce evaluation method to be employed in evaluating such a contribution.

CHAPTER THREE

IMPROVED INTEGRATED MINING ARCHITECHURE (IIMA) APPROACH

3.1 INTRODUCTION

The Improved Integrated Mining Architecture is the proposed solution to the three research questions posed in this thesis. The chapter presents an overview of IIMA (Improved Integrated Mining Architecture) as a novel hybrid of Association Rule Mining and Information Retrieval technique based on content similarity of XML (Extensible Markup Language) document. It gives insight into

its strategy and underlining assumptions, its process architecture, and its main sub-processes. In addition the modalities for the validation of the IIMA approach are discussed. The chapter closes with a summary and discussion on expected results.

3.2 OVERVIEW OF THE EXISTING INTEGRATED MINING SYSTEM

The existing integrated mining architecture as proposed by Sukumaran and Sureka, in 2007 is based on natural language processing and machine learning based techniques (text tagging and annotation). These two approaches form the preprocessing step toward integrating structured and unstructured data. For the unstructured data preprocessing, text tagging and annotation platform extracts information (basically using syntactic analysis) into an XML database. Apart from the fact that this architecture has not been implemented, the main component of the system which converts unstructured to semi structured (XML) is based on natural language techniques and therefore still subject to the generational problems of information extraction such as high error rates thereby producing unreliable results. The Figure 3.1 below represents the Sukumaran architecture.

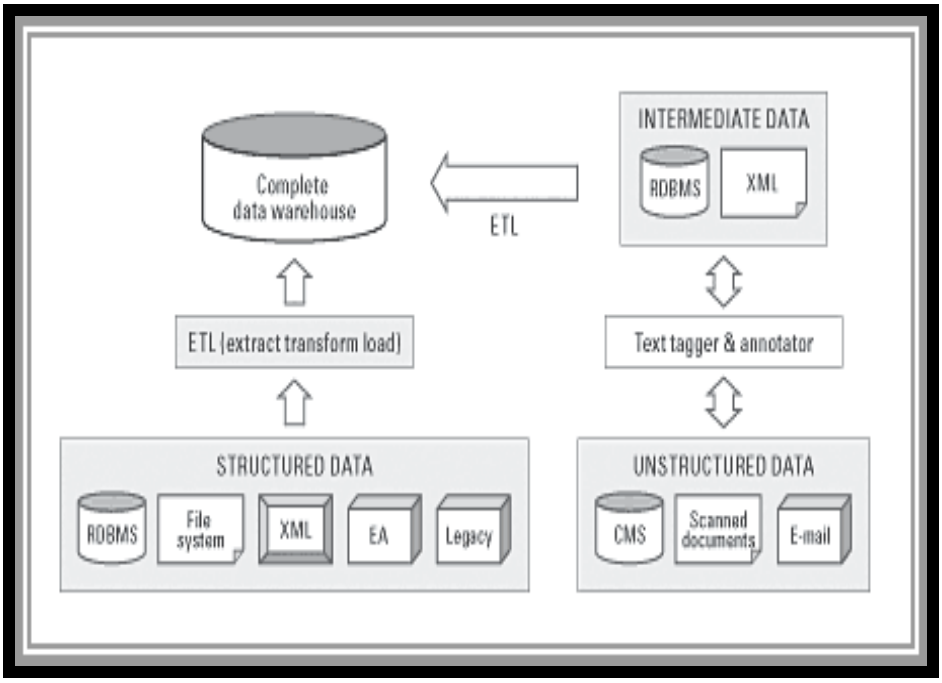


Figure 3.1 Data integration architechure (Sukumaran & Sureka, 2007).

Due to the inefficiencies of the system described above the IIMA approach is introduced as an improvement on this existing system by introducing a domain independent semantic analysis in the data preprocessing stage.

3.3 IIMA DESIGN CRITERIA

The following are the basic design criteria required to be fulfilled by the IIMA.

- IIMA should be able to preprocess unstructured data syntactically and semantically and integrate it with structured data into an XML database.
- IIMA should be flexible and applicable to any domain of interest.
- Knowledge distillation in the IIMA approach should be based on a dataset which is a basis for novel rules.

- Rules generated as a result of the application of the IIMA approach which is in the textual format should be easy to understand.

3.4 OVERVIEW OF THE PROPOSED SOLUTION: IIMA APPROACH

The IIMA approach originated from the generic data integration architecture presented in section 2.3.2, which is an approach to integration that is based on information extraction technique. The proposed approach is domain independent, so it is flexible and can be applied on different domains without having to build a domain specific stemming dictionary. The approach is focused on providing a solution to the existing problem of data uncertainty, which stems from natural language processing. It integrates structured and unstructured data seamlessly for association rule mining. The unstructured component of the integration is based on information retrieval technique which combines syntactic and semantic relevance-oriented search with XML technology.

3.4.1 IIMA Process architecture

The IIMA process architecture consists basically of two major phases: the data preprocessing phase and the knowledge distillation phase. In the data preprocessing phase, the structured component of the integration is selected based on the resulting keywords from the information retrieval process.

3.4.2 The Data Preprocessing Phase

This phase is aimed at optimizing the performance of the knowledge mining phase. It consists of text filtration, stemming and clustering of XML document generated using semantic content similarity.

3.4.2.1 Filtration

In this process, the textual documents are filtered by removing the unimportant words from documents content. Such unimportant words include: articles, pronouns, determiners, prepositions and conjunctions, common adverbs and non-informative verbs. As a result of this process, more important or highly relevant words are single out. To achieve this, we build a list of unimportant words called stop words, where the system checks the documents content and eliminate these unimportant words from it. The system also replaces special characters, parentheses, commas, etc., with a space between words in the documents.

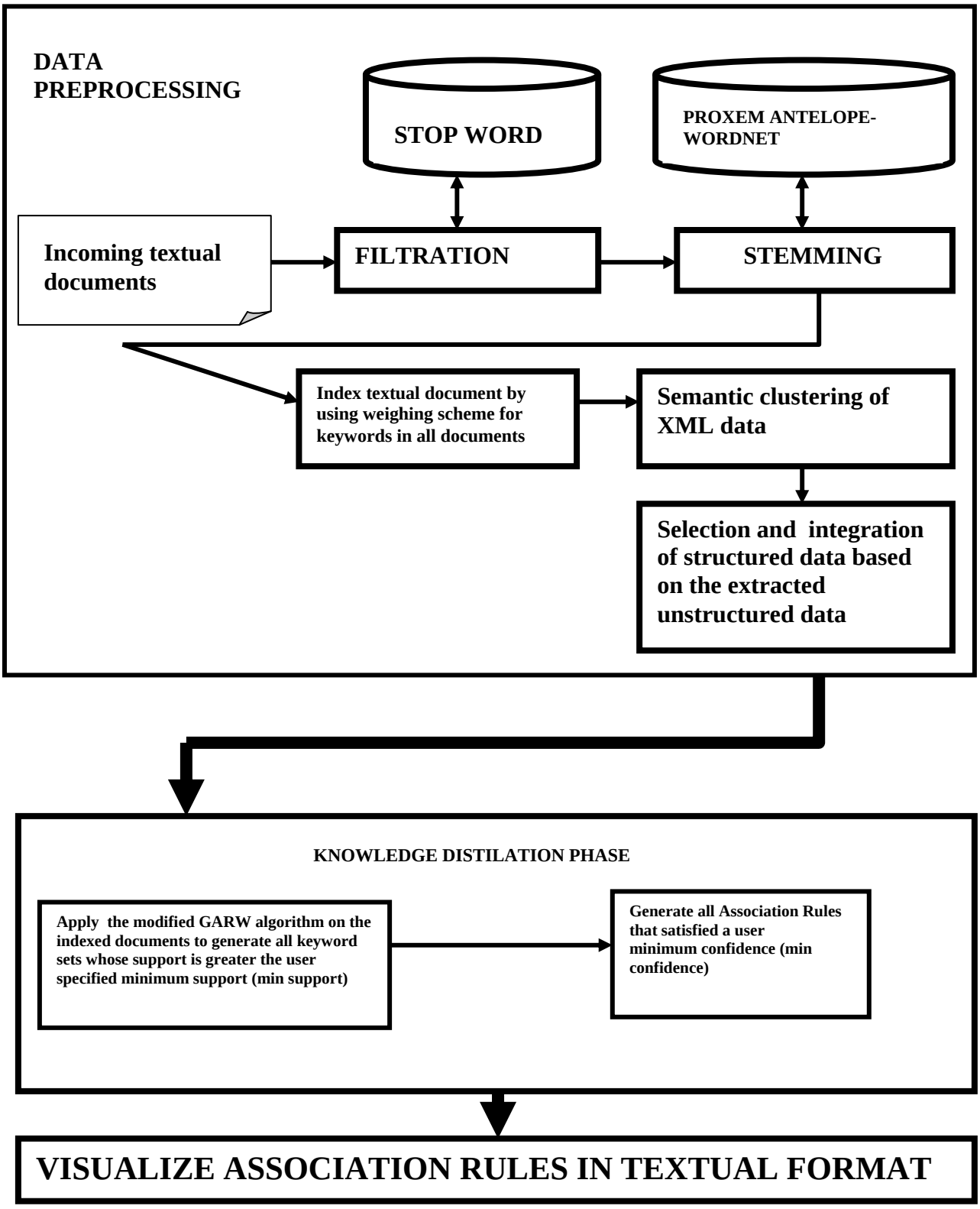


Figure 3.2 The IIMA Process Architecture

3.4.2.2 Stemming

After the filtration process the system does word stemming, a process that removes a word's prefixes and suffixes (such as unifying both infection and infections to infection). Stemming is done by unifying word based on their dictionary meaning using the WordNet lexical database. WordNet is referenced through Proxem Antelope (http://www.proxem.com/Default.aspx?tabid=55), which is a framework that makes the development of Natural Language Processing software easy to use. Proxem Antelope is designed to load WordNet files into the memory so as to make searches amazingly fast.

The Antelope is fully object-oriented. It supports an interface-based programming model. Each module (lexicon, tagger, parser, etc.) defines standard interfaces and many components can implement these interfaces. Antelope is designed for the Microsoft .NET framework (version 2.0 and above). Therefore, you can use it with C#, Visual Basic.NET, Delphi.NET and many other .NET compliant languages (even COBOL.NET!). Antelope consists in the following assemblies (stored in the bin subdirectory). This diagram shows the dependencies between them.

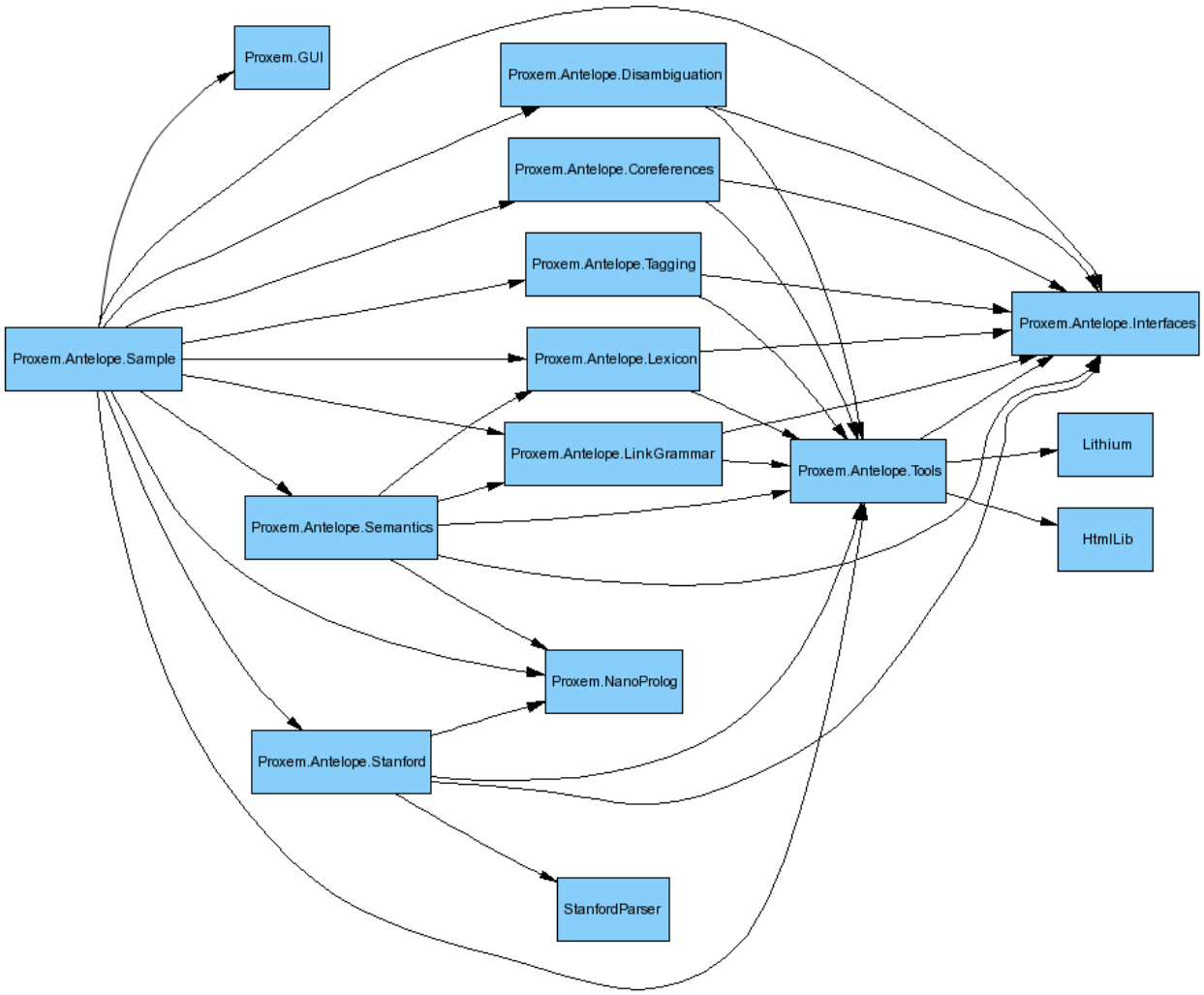


Figure 3.3 An Overview of the Proxem Antelope

(http://www.proxem.com/Default.aspx?tabid=55)

3.4.2.3 Clustering of XML document

In this stage, the weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) is combined with semantic relevance weight to give a combined relevance weight as stated below.

The TF-IDF is used to assign higher weights to syntactically distinguished terms in a document, and it is the most widely used weighting scheme which is defined as (Feldman et al., 1996; Hany et al., 2007; Teng-Kai et al., 2010; Fang-Yie et al., 2010).

$$w(i,j) = tfidf(d_i, t_j) = \left\{ \begin{array}{cc} Nd_{i,}t_j * \log_2 \frac{|c|}{Nt_j} & if \quad Nd_i, t_j \geq 1 \\ 0 & if \quad Nd_i, t_j = 0 \end{array} \right\} \quad \text{Eq. 3.1}$$

- w(i,j) is known as the weighting scheme and could be greater than 0.
- Nd_i,t_j is the number of times the term t_j occurs in the document d_i.
- Nt_j is the number of documents in the collection C in which the term t_j occurs at least once.
- |C| is the number of documents in the collection C.

In general, this weighting scheme includes the intuitive presumption that the more often a term occurs in a document, the more it is representative of the document (term frequency) and the more the documents the term occurs in, the less discriminating it is (inverse document frequency). The system sorts the keywords based on their scores and selects them based on the given weight chosen as threshold.

The semantic relevance is gotten by exploiting the degree of polysemy of terms i.e. we want to weigh the semantic relevance of a term with respect to a notion of semantic rarity, in such a way that the higher the number of meanings of the term, the lower its rarity, thus, its relevance. Since we assume that the terms in the XML data have been

reduced to their stems, then, the semantic relevance is calculated based on the polysemy of its variant terms that originally appear in the text (Andrea et al., 2010).

$$s-rarity(w)=\frac{1}{|o-terms(w)|}\left[\sum_{w_j eo-terms(w)}\ln\left(\frac{MAX-POLYSEMY+1}{|sense(w_j)|+1}\right)\right] \quad \text{Eq. 3.2}$$

T - collection of XML tree tuples i.e. a set of transactions

w - index term i.e we pick each term one by one

o-terms(w) - set of original terms in T having w as the common stem

| o-terms(w)| - absolute number of terms in T that their stem is w i.e. the particular term in question.

| senses(w_j)| - absolute number of meanings of w_j

MAX-POLYSEMY - a constant denoting the number of meanings of the most polysenous term in the reference lexical knowledge base.

Note: the MAX-POLYSEMY depends on the part of speech of the selected terms e.g. its 32 for nouns in WordNet 2.0 (Andrea et al., 2010).

Combination of syntactic and semantic relevance to get the relevance weight of each term i.e. w_j (Andrea et al., 2010).

$$relevance(w_j,u_i)=\frac{1+s-rarity(w_j)}{|T(u_i)|}\sum_{\tau \epsilon T(u_i)} ttf.itf(w_j,u_i/\tau) \quad \text{Eq. 3.3}$$

relevance (w_j , u_i) - stores the reference value of term w_j in TCU

s-rarity(w_j) - gotten from semantic relevance

$\sum_{\tau \in T(u_i)} ttf.itf(w_j, u_i/\tau)$ -the TF-IDF weight

$T(u_i)$ - total number of TCUs or transactions.

The content similarity of the XML documents is then measured by calculating $sim(u_i,u_j)$ where u_i and u_j are vectors which represents xml documents.

Content similarity between any two tree tuple items is measured by comparing their respective TCUs. Given a collection of XML tree tuples T, any TCU, u_i , is modeled with a vector, u_i , whose j -th component corresponds to an index term, w_j , and contains the value relevance (w_j , u_i). The size of each TCU vector is equal to the size of the collection vocabulary: the set of index terms extracted from all TCUs in T. The well-known cosine similarity is then used to measure the similarity between TCU vectors (Andrea et al., 2010).

Let e_i and e_j be tree tuple items, and $\_u_i$ and $\_u_j$ their respective TCU vectors. The content similarity between e_i and e_j is defined as:

$$sim(u_i, u_j) = \frac{u_i \bullet u_j}{\|u_i\| \times \|u_j\|} \qquad \text{Eq. 3.4}$$

Since the combination of structure and content information characterizes an XML tree tuple item, there is need to take tolerance on computing similarity between XML tree tuple items. For this purpose, a similarity threshold is introduced to represent the minimum similarity value for considering two XML tree tuple items as similar (Andrea et al., 2010).

3.4.2.4 Data Integration

In the integration process, the structured component is selected based on the resulting keywords from the unstructured text preprocessing process, and association rules is generated based on the modified GARW (Generating Association Rules Based on Weighting Scheme) Algorithm. The main contribution of this technique is that the unstructured component of the integration is based on Information retrieval technique which is based on content similarity of XML (Extensible Markup Language) document. This similarity is based on the combination of syntactic and semantic relevance.

3.4.3 Knowledge Distillation Phase.

Knowledge is distilled using the GARW (Generating Association Rules based on Weighting scheme) algorithm described below (Hany et al., 2007): In this phase, association rules are generated based on the (GARW) Algorithm, which has been modified to accommodate content similarity of XML document based on the combination of syntactic and semantic relevance.

3.4.3.1 Generating Association Rules Based on Weighting Scheme (GARW) Algorithm

Given a set of terms

$A = \{w_1, w_2, \ldots\ldots\ldots\ldots w_n\}$ Eq. 3.5

A set of indexed documents $D = \{d_1, d_2, \ldots\ldots\ldots\ldots\ldots\ldots d_n\}$ Eq. 3.6

- $d_1 \ldots\ldots d_n$ are indexed documents that contains keywords.
- Those keywords are also members of A i.e. the general database of keywords.

Association Rule

Association rule is one of the most important techniques in Data Mining. The problem of association rule mining deals with how to discover association rules that have support and confidence greater than the user-specified minimum support and minimum confidence. It is intended to capture dependency among items in the database.

The support of an item set is the fraction of transactions in the database that contain all the items in the database

$$Support\ (W_i W_j) = \frac{SupportCou\ nt(W_i W_j)}{TotalNumbe\ rOfTransac\ tions} \qquad \text{Eq. 3.7}$$

The confidence of rule a (association rule) $W_i \rightarrow W_j$ can be defined as the proportion of those transactions containing W_i that also contain W_j.

$$Confidence\quad (W_i W_j) = \frac{Support\quad (W_i W_j)}{Support\quad (W_i)} \qquad \text{Eq. 3.8}$$

The algorithm for generating association rules based on the weighting scheme is given as follows:

1. Scan the file that contains all the keywords that satisfy the threshold weight value and their frequency in each document.
2. Let N denote the number of top keywords that satisfy the threshold weight value.
3. Store the top N keywords in index file along with their frequencies in all documents, their weight values relevanceWeight and documents ID in the following format: <doc-id><keyword>< keyword frequency>< relevanceWeight >

4. Scan the indexed file and find all keywords that satisfy the threshold minimum support. These keywords are called large frequency1-keywordSet L_1.
5. When K is greater than 2, (Note K is a keyword set having k-keywords sets). The candidate keywords C_k of size K are generated from large frequent (k-1) keywords sets, L_{k-1} that is generated in the last step.
6. Scan the index file, and compute the frequency of candidate keyword sets C_k that is generated in step 4.
7. Compare the frequencies of candidate keywords sets with minimum support.
8. Large frequent keyword sets L_k, which satisfy the minimum in support, is found from step 7 above.
9. For each frequent keyword set, find all the association that satisfies the threshold minimum confidence.

3.4.3.2 Rule Post Processing

The generated rules are refined by using parameters such as the support and confidence which in this case we already been included in the GARW algorithms above. One particular aspect of rule mining in text is that often a high support means the rule is too obvious and thus less interesting. Another technique that was used to remove unwanted rules is to specify stop rules i.e. rules that are common and can be removed automatically. Association rules are easy to understand and to interpret for an analyst or a normal user. However, it should be mentioned that the association rule extraction is of exponential growth and a very large number of rules can be produced.

3.4.3.3 Rule Visualization Phase

Even though association rules extracted from the above phases can be reviewed in textual format or tables, or in graphical format, in this work the system is designed to visualize the extracted association rules in textual format or tables.

3.5 TOOL FOR SUPPORT OF IIMA

The following are the essential tool-support for implementing the IIMA approach:

- Proxem Antelope (http://www.proxem.com/Default.aspx?tabid=55), which is a framework that makes the development of Natural Language Processing software easy to use.
- WordNet lexical database (http://wordnet.princeton.edu/wordnet/download/). WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.
- Software design: UML-based tools (Microsoft Visio, Rational Rose, ArgoUML etc.),

3.6. APPLICATION SCENARIOS

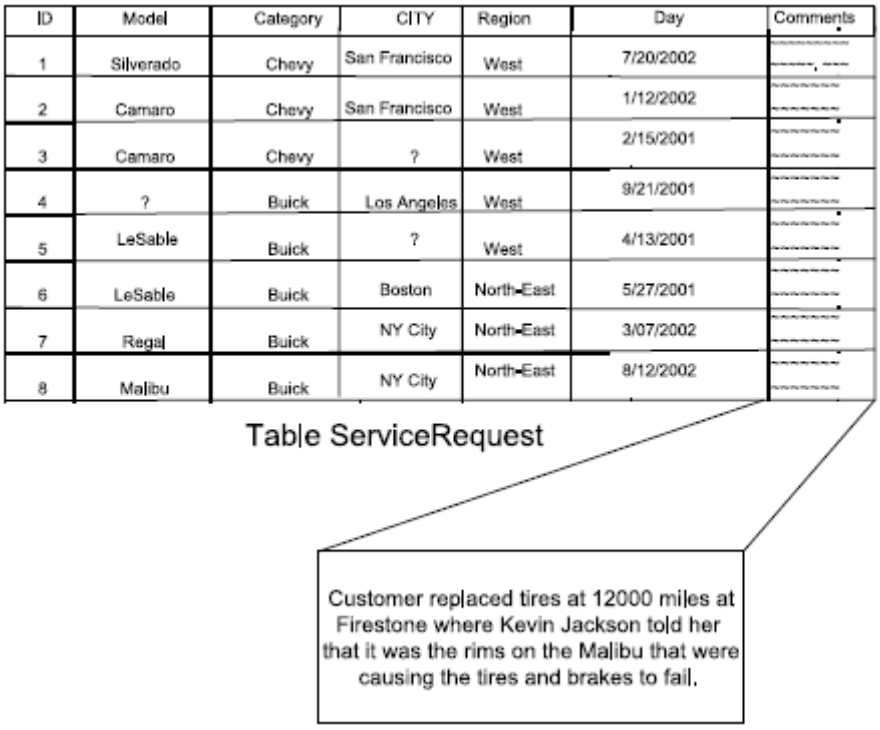
- Financial services: This scenario is for a financial services company which subscribes to several commercial research publications. These publications come with RIXML (Research Information Markup Language) formatted data. RIXML is an XML

vocabulary that combines investment research with a standard format to describe the report's meta-data.

A received report, which is archived in its native XML format, and the audio and visual clips in form of meta-data, such as company name, stock price, and earnings estimates, are extracted from the document and stored in relational tables. This information is used to detect and recommend changes in buy/sell/hold positions to equity and bond traders and key customers. Mining applications more thoroughly analyze the original document and its extracted metadata, looking for such keywords as "merger," "acquisition," or "bankruptcy" to categorize and summarize the content. The summarized information is combined with historical information and this made available to the company's market research and investment banking departments. After this, these departments combine the summarized information with financial information stored in spreadsheets and other documents. This is used to perform trend forecasting and identify merger and acquisition opportunities (Roth et al., 2002).

- Another scenario is that of an auto manufacturer employing an enterprise customer relationship management (CRM) application. This application is used to track and manage service requests across its worldwide dealer operations. Using the CRM application, the individual dealers file "customer service reports". The report includes both structured and unstructured part. The structured part consist of attributes such as "day", "customer ID", "make", "model", "dealer name", "vehicle identification number (VIN)", etc. The unstructured part consists of "comments" field where the personnel in charge of handling a service request can record additional information about the precise nature of the problem and how the issue was addressed. The Figure

3.4 below shows a simplified version of a "service reports" table and also highlights the text associated with one of the reports. Integrated mining based on the result of the query based on the available structured attributes combined with the appropriate text-index can be constructed (Zhu et. al, 2005).



ID	Model	Category	CITY	Region	Day	Comments
1	Silverado	Chevy	San Francisco	West	7/20/2002	~~~~~~~~~~
2	Camaro	Chevy	San Francisco	West	1/12/2002	~~~~~~~~~~
3	Camaro	Chevy	?	West	2/15/2001	~~~~~~~~~~
4	?	Buick	Los Angeles	West	9/21/2001	~~~~~~~~~~
5	LeSable	Buick	?	West	4/13/2001	~~~~~~~~~~
6	LeSable	Buick	Boston	North-East	5/27/2001	~~~~~~~~~~
7	Regal	Buick	NY City	North-East	3/07/2002	~~~~~~~~~~
8	Malibu	Buick	NY City	North-East	8/12/2002	~~~~~~~~~~

Figure 3.4 Integrated data storage (Zhu et. al, 2005).

3.7 VALIDATION APPROACH

In order to validate the plausibility of the proposed solution approach, a case study of evolving rules to form the bases of customer relationship management decisions making will be reported in chapter 4 to show the practical real-life application scenario of the IIMA approach. This is done to validate the hypothesis that: The IIMA approach provides a seamlessly analysis across structured and unstructured data. This is particularly interesting because, presently, there exist a problem in analytic CRM which has to do with having a holistic view to the structured and unstructured CRM data (Cody et al., 2000).

3.8 SUMMARY & DISCUSSION

In this chapter the concept of Improved Integrated Mining of Architecture (IIMA) approach has been presented as an integrated solution model for the two research questions posed in this thesis. The practical application of IIMA will be discussed in the subsequent chapters.

CHAPTER FOUR

APPLICATION OF IIMA TO CRM

4.1 INTRODUCTION

This chapter presents details of a case study of Customer Relationship Management scenario where the IIMA approach has been applied. The core motivation of this case study is that, presently, there exists a problem in analytic CRM which has to do with having an holistic view to the structured and unstructured CRM data which this thesis plans to address. This chapter therefore reports the practical application of the Improved Integrated Mining Architecture in solving the above described problem.

4.2 IMPLEMENTATION COMPONENTS AND TOOLS

1. Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It can be used to develop console and graphical user interface applications along with Windows Forms applications, web sites, web applications, and web services in both native code together with managed code for all platforms supported by Microsoft Windows. Visual Studio supports different programming languages by means of language services, these languages include C/C++ (via Visual C++), VB.NET (via Visual Basic .NET), C# (via Visual C#), and F# . Support for other languages such as M, Python, and Ruby among others is available via language services installed separately. It also supports XML/XSLT, HTML/XHTML, JavaScript and CSS.
2. **C#** is a multi-paradigm programming language encompassing imperative, declarative, functional, generic, object-oriented (class-based), and component-oriented

programming disciplines. It was developed by Microsoft within the .NET initiative and later approved as a standard by Ecma (ECMA-334) and ISO (ISO/IEC 23270). C# is one of the programming languages designed for the Common Language Infrastructure. C# is a simple, modern, general-purpose, object-oriented programming language. The language, and implementations provide support for software engineering principles such as strong type checking, array bounds checking, detection of attempts to use uninitialized variables, and automatic garbage collection. For C#, software robustness, durability, and programmer productivity are important. C# is used in developing software components suitable for deployment in distributed environments. It supports internationalization and is suitable for writing applications for both hosted and embedded systems, ranging from the very large that use sophisticated operating systems, down to the very small having dedicated functions.

3. Microsoft® SQL Server™ is a database management and analysis system for e-commerce, line-of-business, and data warehousing solutions. SQL Server 2008, the latest version is enhanced with XML support, integration of .NET Framework objects in databases, improved integration with Microsoft Visual Studio and the Microsoft Office System. It also consists of an improved analysis, reporting, and data integration services.

4. Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. Extensible Markup

Language (XML) is a set of rules for encoding documents in machine-readable form. It is defined in the XML 1.0 Specification produced by the W3C, and several other related specifications, all gratis open standards. XML's design goals emphasize simplicity, generality, and usability over the Internet. It is a textual data format with strong support via Unicode for the languages of the world. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services.

4.3 IMPROVING CUSTOMER RELATIONSHIP MANAGEMENT THROUGH INTEGRATED MINING OF HETEROGENEOUS DATA

4.3.1 Problem definition

No business is an island. For a business to succeed, it will need to deal with customers, suppliers, employees, and others. In almost all cases, there will also be other organizations that offer the same or similar products to similar customers. These organizations are known as the competitors. Not so long ago, the mobile phone was an amazing invention, which revolutionized communication between humans. It has now become a piece of technology, which is deeply ingrained in modern life. In less than twenty years, mobile phones have gone from being rare and expensive pieces of equipment used by businesses to a pervasive low-cost personal item. In Nigeria today, mobile phones outnumber land line telephones, with most adults and many children now owning mobile phones (Ayo et.al., 2007). Each new handset provides new exciting designs with MP3 players, cameras and other interactive goodies all of which fit into a pocket sized package, which with each evolution becomes lighter and thinner. With all the focus on the technology and the desire to squeeze more and

more exciting software on to mobile phones, the key function of the phone as a communication device seems to have been overlooked in some cases. The experiences of many mobile phone users are being overshadowed by the evolving technology, which in essence, might not be solving the problem of telephone industries, that is, the need to be bigger and brighter than the competitor.

4.3.2 Improving Organizational Profit of Mobile Phone Industry

To apply the system to the case study which has to do with improving organizational profit of manufacturing and production companies, the problem we aim to solve is as follows: to reveal which of the mobile phone technology help to improve the competitive advantage of mobile phone industry through analytical customer relationship management. This will also reveal if any, the weaknesses of various mobile phones in Nigeria.

4.3.3 Data requirements

The primary means of gathering data in our field of application, which is CRM is through the use of questionnaires. A questionnaire was therefore designed and administered to 2,215 respondents out of which 1,518 were returned valid. These questionnaires were designed with the goal of retrieving CRM information from mobile phone users towards effective customer relationship management in the mobile phone manufacturing industry. This questionnaire was justified through a pilot study and meeting with experts in CRM field. The questionnaire contained both structured and unstructured part. The sample of the questionnaire can be gotten in appendix A.

4.3.4 Expected outputs

It is expected that the study will reveal:

- Actionable rules that reveal strengths and weaknesses of specific mobile phones.
- Rules for decision making based on underlying data collection.
- The rules that are unexpected but yet if acted upon can give a mobile phones industries a competitive edge over its competitors.
- Rules that will solve mobile phone related problems in the society.
- Rules that can also give mobiles phone consumers the recommendations as to what type of phone to buy or what functionalities to look out for specific mobile phone brands.
- Customer recommended improvements on mobile phones so as to improve the customer relationship management of these mobile phone industries.
- Recommendations which competitors of mobile phone industry can use to make more profit.

4.3.5 Scope

The research is based on analyzing customer data through data mining and does not include issues as regarding implementing the inferences gotten form the system. In order wards, even though CRM includes acquisition, analysis and the use of knowledge about customers in order to sell more goods or services and to do it more efficiently, this research is focused on the only the analytical CRM aspect of it.

Each of the following systems described below is applied to the above described problem definition. This is done to be able to compare the results of structured mining, existing text mining approach, and existing integrated mining approach and IIMA approach.

- Structured mining: This is an implementation of association rule mining algorithm on structured data.

- Text mining (Unstructured mining): This is an implementation of the existing text mining technique described in (Hany et al., 2007).

- Existing Integrated Mining approach: This is based on the existing integrated mining approach proposed by (Sukumaran, S. and Sureka, 2007).

- Improved Integrated Mining Architecture (IIMA): This is the implementation of the architecture described in chapter 3 of this thesis.

4.4 STRUCTURED MINING

4.4.1 Data Input to the Structured Mining System

The data input to the structured mining system is the structured part of the questionnaire which includes the respondents' answers to questions such as;

- What Brand of mobile phone are you using now?

  ( ) Nokia ( ) Motorola ( ) Ericsson ( ) Samsung ( ) LG ( ) Philips ( ) Sagem ( ) Arcatel ( ) Sony ( ) Siemens Nokia Others…….

- Gender: ( ) Male ( ) Female

- What is your age range?15-20 ( ) 21-30 ( ) 31-40 ( ) 41-50 ( ) 51-60 ( ) Above 60( )
- Is your Education IT related? YES ( ) NO ( )
- What is the general assessment of your mobile phone user friendliness?

  ( ) Excellent ( ) Good ( ) Satisfactory ( ) Unsatisfactory ( ) Poor
- What is the reason for changing your phone? Phone got spoilt ( ) Stolen ( ) Gave it out ( ) Misplaced it Others specify……

The data gotten from the above questions are structured because they require a one word answer which is selected from the available options given.

For more of these questions, check the appendix A.

Table 4.1 distributions of users by mobile phone brand

TECNO	70
VodaPhone	14
STARCOMS	30
DORADO	4
SENDO	9
VisaPhone	4
Sony Ericson	83
Motorola	117
SAGEM	88
SAMSUNG	90
PANASONIC	9
LG	20
Nokia	965
Empty	15

Figure 4.1 represents the interface to the system that mines from purely structured data. It implements association rule algorithm and receives two thresholds which is minimum support and minimum confidence. The data to be mined is received from SQL database and converted to an XML database on which the association rule acts upon. The result generated by the system is displayed in the figure 4.2. Each rule is displayed against its respective support and confidence.

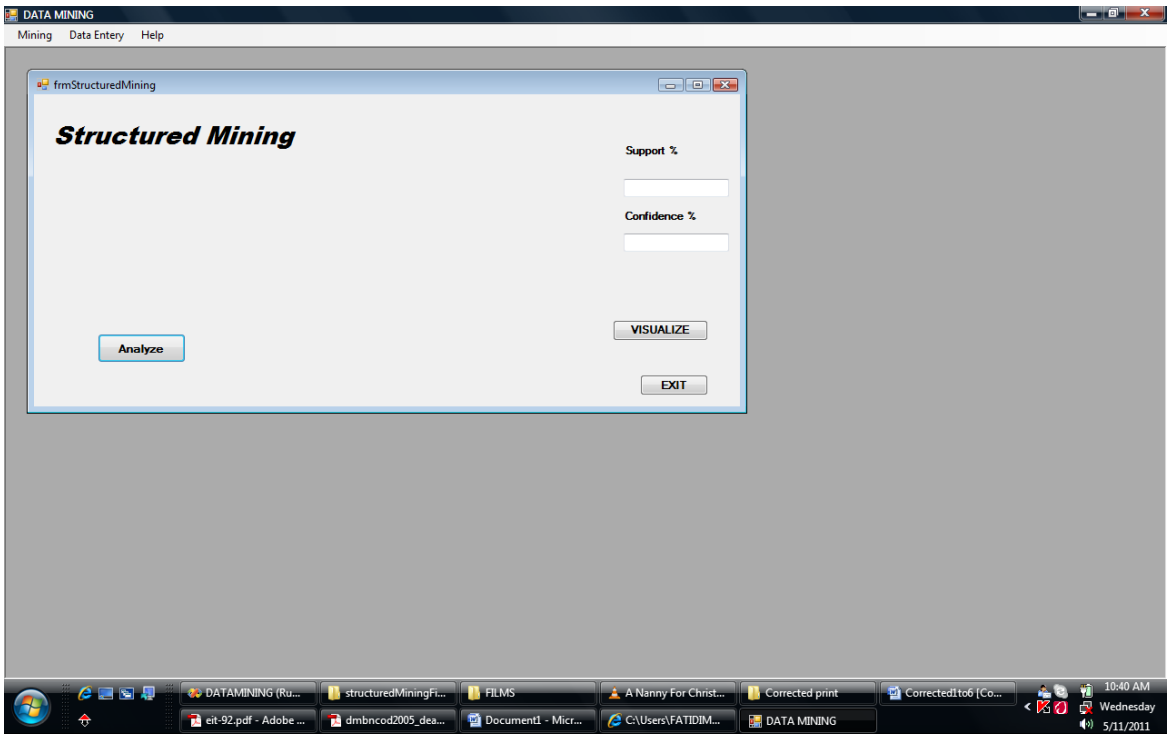


Figure 4.1 Structured Mining Interface

The Table 4.2 shows the number of rules generated for different support thresholds while keeping the confidence constant at 80%. These results were gotten by applying the structured system on the structured part of the data gotten from the application of the questionnaire.

Table 4.2 Support and Confidence for Structured Mining

Support	**Confidence**	**Number of Rules Generated**
50%	80%	42
40%	80%	90
30%	80%	156
25%	80%	186
20%	80%	210

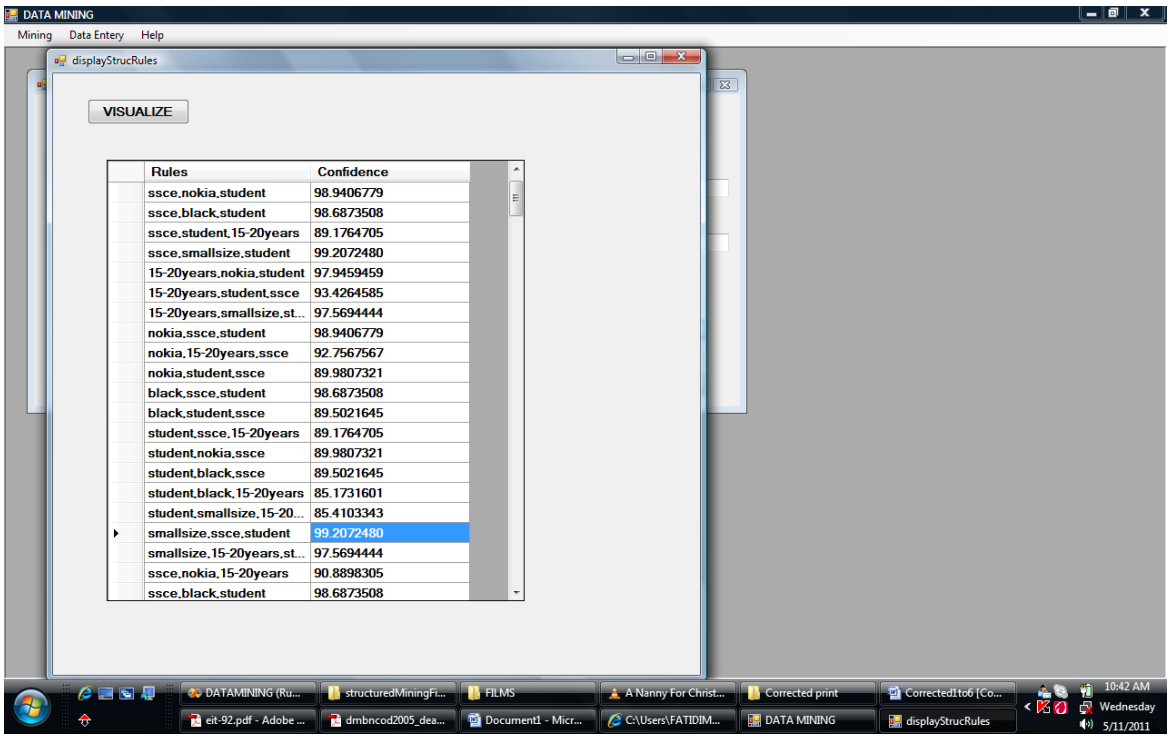


Figure 4.2 Output of structured Mining

4.4.2 Discussion

Some of the association rules that describe the relations between attributes in the database and are interpreted for competitive intelligence as follows;

- The rule *Nokia,SatisfactUserFriendly=>Above1year:* gives a clear inference that the users of Nokia products find it to be user friendly and most of them have for this reason been using it for over one year and are therefore, not disposed to change their phone often.

  Companies that produce other brands such as Samsung and Panasonic should therefore improve on their user-friendliness.

- The rule *Nokia,SatisfactUserFriendly=>21-30 years:* reveals a high population of Nigerian between the age of 21 and 30 who use Nokia phones because they find it to

be user friendly. Therefore marketing could be targeted towards this sector of the population.

- The rule *Nokia, SatisfactUserFriendly=>Available Purchase Good:* reveals that Nokia phones are easily and readily available to purchase.
- The rule *Nokia,Stolen=>Durable:* reveals that even though Nokia phones are durable, they are often changed by the user because they are stolen very often. Nokia company can therefore look for means of improving on security features. Also, because of the fascinating features (radio, TV etc) it is the target of most customers and thieves alike.
- The rule *GoodServices,Durable=>Nokia* implies that Nokia phone users were interested in buying it because they find it durable and easy to navigate through the services that are present on the phone.
- The rule *DifficultComposeRingTone,Satisfact UserFriendly=>Nokia* reveals that even though users of Nokia phones are satisfied with its user friendliness, they however have difficulty in a particular service that it offers which is "Composing ringing tones". Both the Nokia company and other companies competing with it such as Sony Ericson and so on, can therefore improve on this service so as to be able to win more customers.

The support (which is 50%) is chosen to be a little lower than the confidence (which is 80%) so as to have a fair representation of the important attributes which might not be included in the rule generation. Confidence was fixed at 80% so as to reduce the number of rules generated.

The rules described above are extracted rules at the support threshold of 50% and they give information on some interesting patterns that can be used for competitive business intelligence in the Nigerian mobile phone industry.

4.5 TEXT MINING (UNSTRUCTURED MINING)

The text mining system is such that automatically extract association rules from collections of textual documents. It discovers association rules from keyword features extracted from the documents. This implemented existing text mining (Hany et al., 2007) technique integrates XML technology, Information Retrieval Scheme, with machine readable dictionary (WordNet) for keyword/feature selection that automatically selects the most discriminative keywords for use in association rules generation.

4.5.1 Data Input to the Unstructured Mining System

The data input to the unstructured mining system is the unstructured part of the questionnaire which includes the respondents' answers to questions such as:

- What do you like most about your mobile phone?.....

  ……..
- What do you dislike most about WAP?.....

  ……..
- Share your best mobile phones experience……..

  ……..
- Why did you decide to purchase that particular brand of mobile phone?.....

  ……..
- What improvements would you like to see, if any on your mobile phone?.....

……..

- What type of problem do you usually encounter while using your mobile phone?

……..

The above response allows the respondents to freely express themselves and therefore provide useful information which might not have been previously known or thought about by the mobile phones manufacturing industry. This form of reply is largely unstructured because it is not a one word answer.

4.5.2 Argumentation of the thresholds

In text mining in general, a very large number of association rules are found. So the measures like support and confidence are important when creating keyword sets and selecting the final rules. However, the problem is that we may find the important keywords which have frequently appeared recently but not discovered because of the height of support and confidence threshold values. In order to have a fair representation of the important keywords in the corpus to be mined, we selected a TF-IDF threshold of 30%. This helped us to find informative keywords to extract rules from. Furthermore, a low threshold support of 5% was used so as to extract important keywords (such as durability, brand) that would not have appeared if we chose high support value, and these keywords happen to be very informative regarding customer relationship management as regards mobile phones. Lastly, we chose higher threshold confidence value 50% to make sure that the final rules gotten from the system are the most interesting ones.

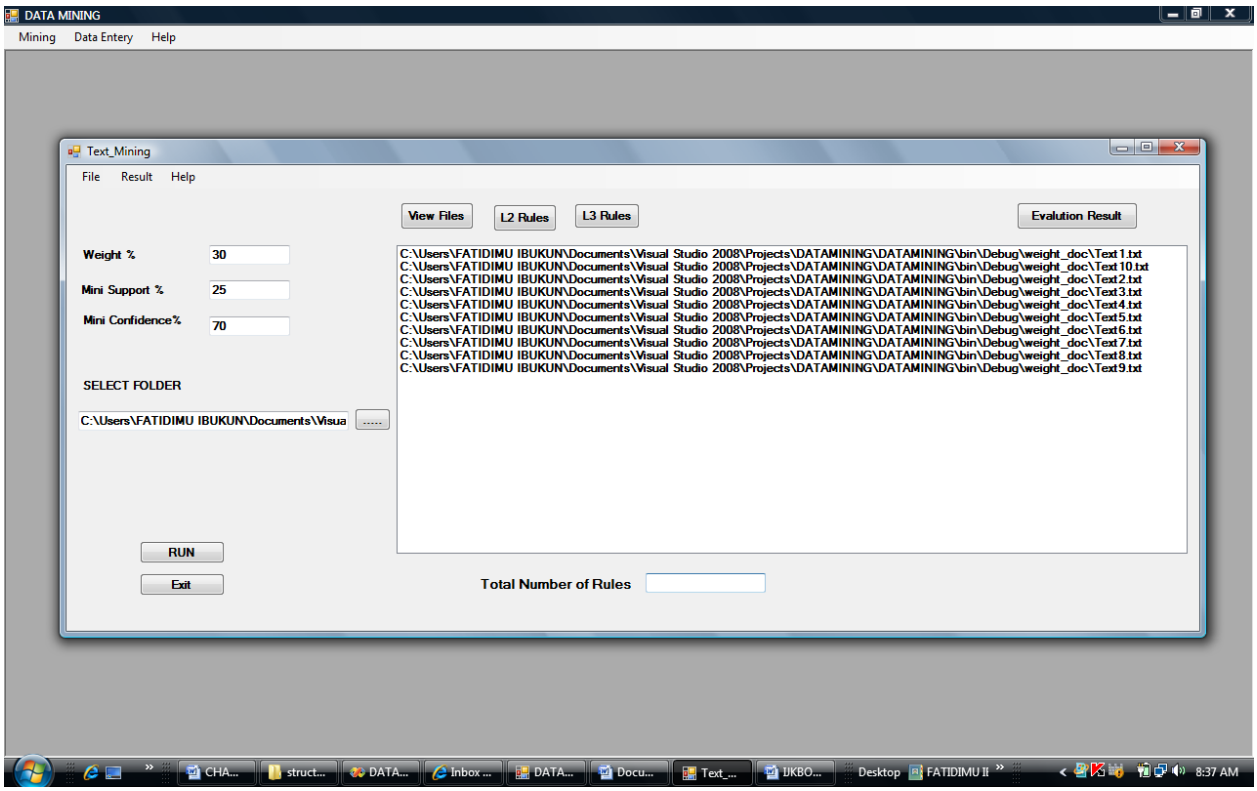


Figure 4.3 Unstructured Mining Interface

4.5.3 Discussion

Figure 4.3 is the interface to the unstructured data mining system. Some of the association rules that describe the relations between keywords in the documents are presented below. The rules give information on some interesting patterns that can be used for customer relationship management in the Nigerian mobile phone industry.

Samples of the generated rules are interpreted for customer relationship management as follows:

- The rule **internet, screen -> problem** shows that there is a problem with the screen of mobile phones while browsing the internet.

- The rule **Nokia, camera -> durability** gives the inference that there is a strong correspondence between Nokia phones, the fact that it is durable and users find it user friendly. Companies that produce this particular brand can improve on such features and competing companies should therefore improve on such qualities.
- The rule **camera, best->portable** reveals that what mobile phone user like best about their mobile phone is the fact that it is portable and also have a camera facility.
- The rule **long, battery->time** can also help to infer that the mobile phone users like the battery time of their phones to be long.

The proposed approach is domain-independent, so it is flexible and can be applied on different domains without having to build a domain specific stemming dictionary. Also, since identifying user requirements and understanding the user is a major part of contributing to the profit of the organization and this can be achieved through an effective customer relationship management. The generated rules therefore, give pointers to various characteristics of customers of mobile phone manufacturing industry which will help in identifying, attracting, developing and maintaining successful customer relationships over time in order to increase retention of profitable customers.

4.6 EXISTING INTEGRATED MINING APPROACH

This is based on the existing integrated mining approach proposed by (Sukumaran, S. and Sureka, 2007). This architecture uses natural language processing techniques (text tagging and annotation) as a preprocessing step toward integrating structured and unstructured data. For the unstructured data sources, the tagging and annotation platform extracts information

based on the integration of XML technology, Information Retrieval Scheme and with machine readable dictionary (WordNet) for keyword/feature selection that is automatically integrated with structured data.

4.6.1 Data input format for Existing Integrated Mining System

The data input to the existing integrated mining system contains both the structured and unstructured part of the questionnaire which includes the respondents' answers to questions such as:

- What Brand of mobile phone are you using now?

  ( ) Nokia ( ) Motorola ( ) Ericsson ( ) Samsung ( ) LG ( ) Philips ( ) Sagem ( ) Arcatel ( ) Sony ( ) Siemens Nokia Others…….
- Gender: ( ) Male ( ) Female
- What is your age range?15-20 ( ) 21-30 ( ) 31-40 ( ) 41-50 ( ) 51-60 ( ) Above 60( )
- Is your Education IT related? YES ( ) NO ( )
- What is the general assessment of your mobile phone user friendliness?

  ( )Excellent ( ) Good ( )Satisfactory ( )Unsatisfactory ( )Poor
- What is the reason for changing your phone? Phone got spoilt ( )Stolen ( )Gave it out ( )Misplaced it
- What do you like most about your mobile phone?.....

  ……..
- What do you dislike most about WAP?.....

  ……..
- Share your best mobile phones experience……..

……..

- Why did you decide to purchase that particular brand of mobile phone?.....

……..

- What improvements would you like to see, if any on your mobile phone?.....

……..

- What type of problem do you usually encounter while using your mobile phone?

……..

The response to the above is both structured (one word answer chosen from available options) and unstructured (answers that contain sentences).

See appendix A for more.

Figure 4.4 Existing Integrated Mining Interface

The existing integrated mining system interface in Figure 4.4, takes in three thresholds which includes: TFIDF Weight, minimum support and minimum confidence. It is fed in both structured and unstructured data gathered from the questionnaire into an XML database. The unstructured component is preprocessed and the result of this preprocessing is used to extract from structured data and seamlessly integrated together, ready to be used as input to the association rule mining system.

4.6.2 Discussion

Samples from the existing integrated mining system are listed against their confidence in the association rules text visualization interface in Figure 4.5. It was observed that the rules generated were too many. Though important ones were there, but have been overcrowded by a lot of redundant rules. Redundant in the sense that the keywords that come from the unstructured part are mostly not related to the problem at hand which is managing customer relationship of mobile phones. Also due to the nature of the questionnaire (some respondents did not really take time to fill in the unstructured part in detail) which was the primary means of data collection, the rules generated were mostly made up of structured data and not a fair representation of structured and unstructured data. This still defeats the purpose of integrated mining.

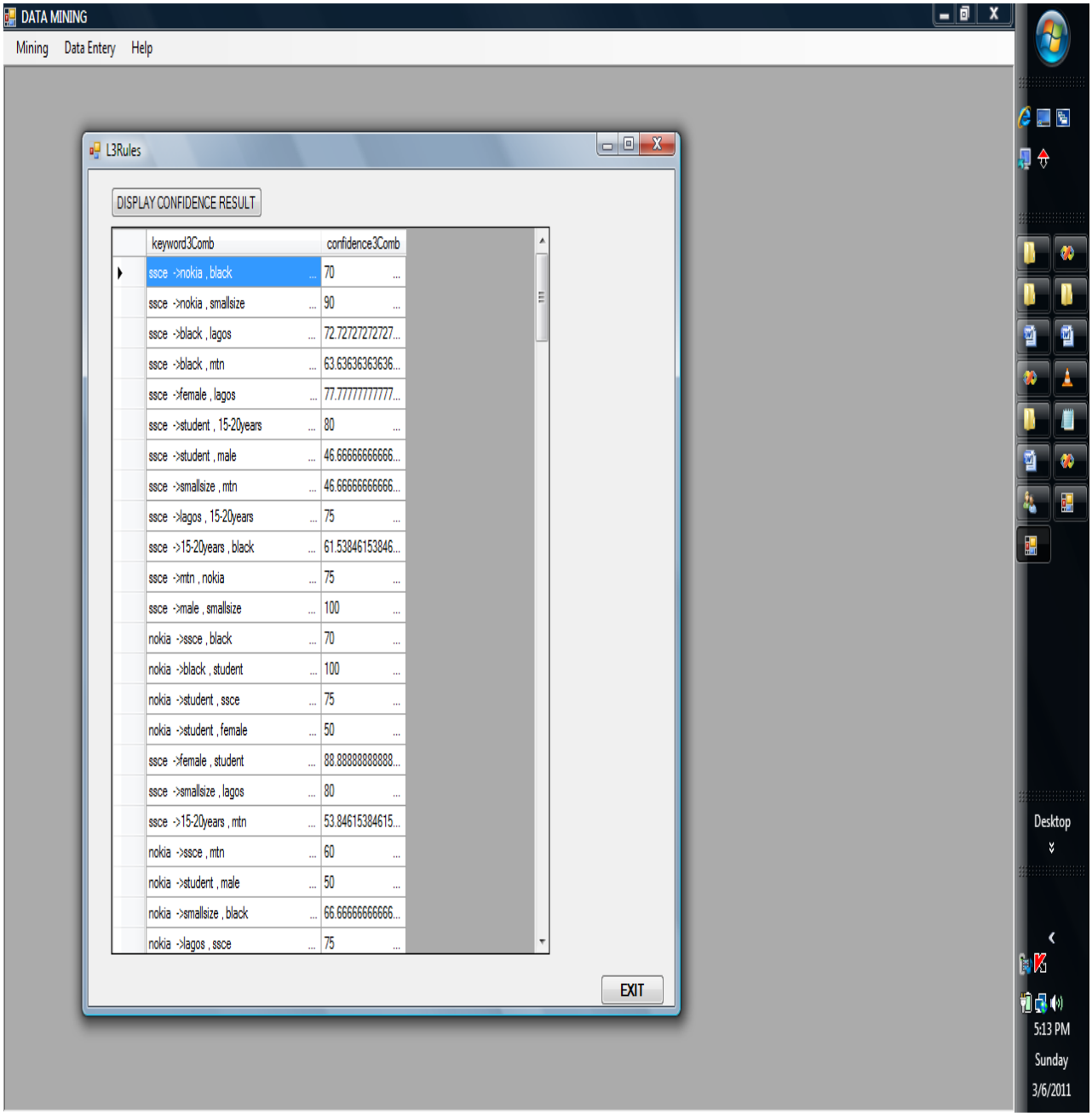


Figure 4.5 Existing Integrated Mining Rule Visualization Interface

4.7 IMPROVED INTEGRATED MINING ARCHITECHURE (IIMA)

4.7.1 Data input to the IIMA

The data input to the IIMA system is the same as that of the existing system described in section 4.6.1. This is, the answer to the questions require both one-word (eg. YES or NO) and also sentences (e.g I would like my mobile phone to have more memory and very high speed). This is to ensure an effective comparism so as to distinctly bring out the efficiency of the existing systems.

4.7.2 Input interface

The interface to the IIMA system is displayed in Figure 4.6. It receives three thresholds namely, relevance threshold, minimum support and minimum confidence. While the system is clustering the preprocessed data based on the relevance threshold, it receives a threshold range with which it groups documents. The result of this clustering is what is used for the association rule mining.

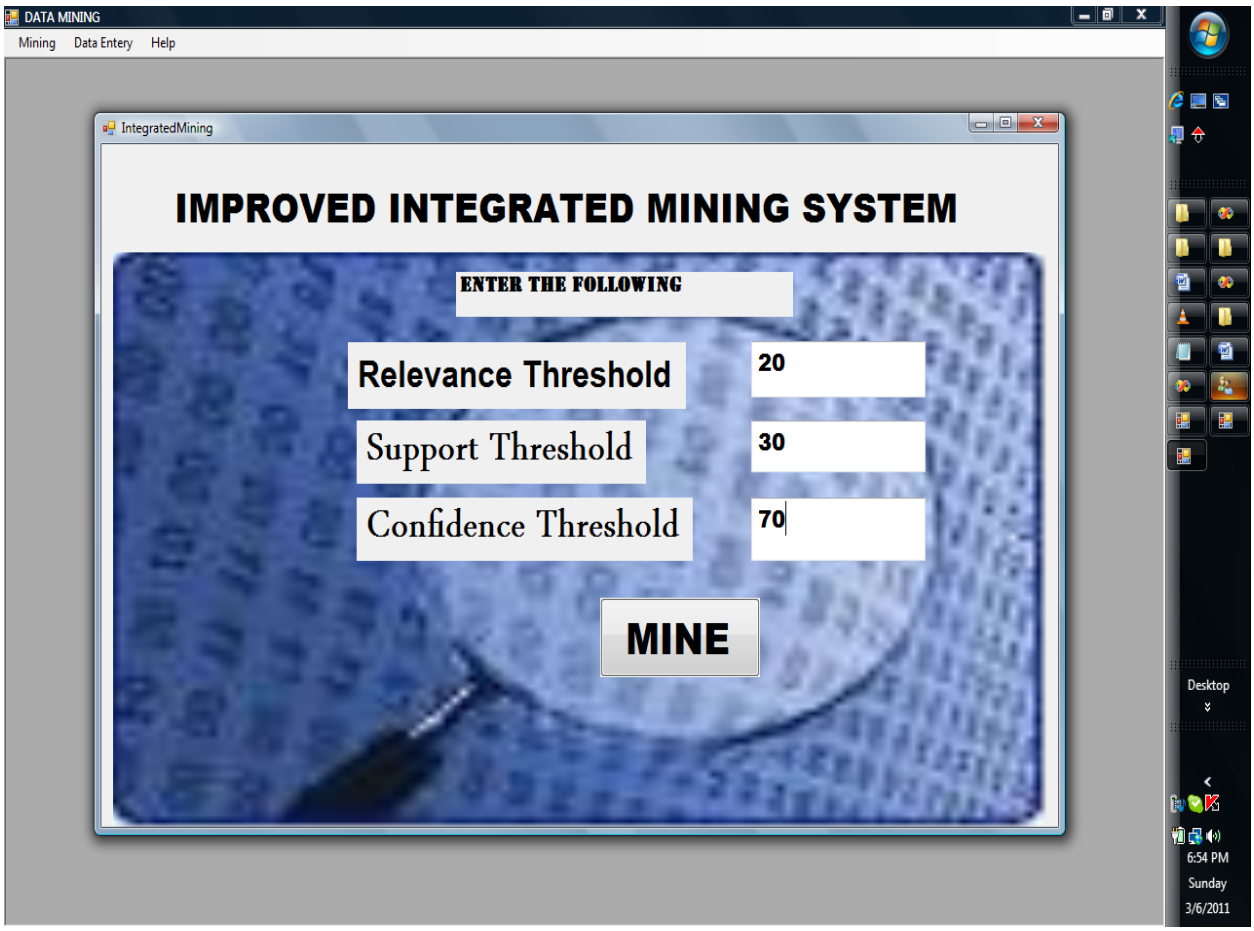


Figure 4.6 Snapshot of the integrated data in XML format.

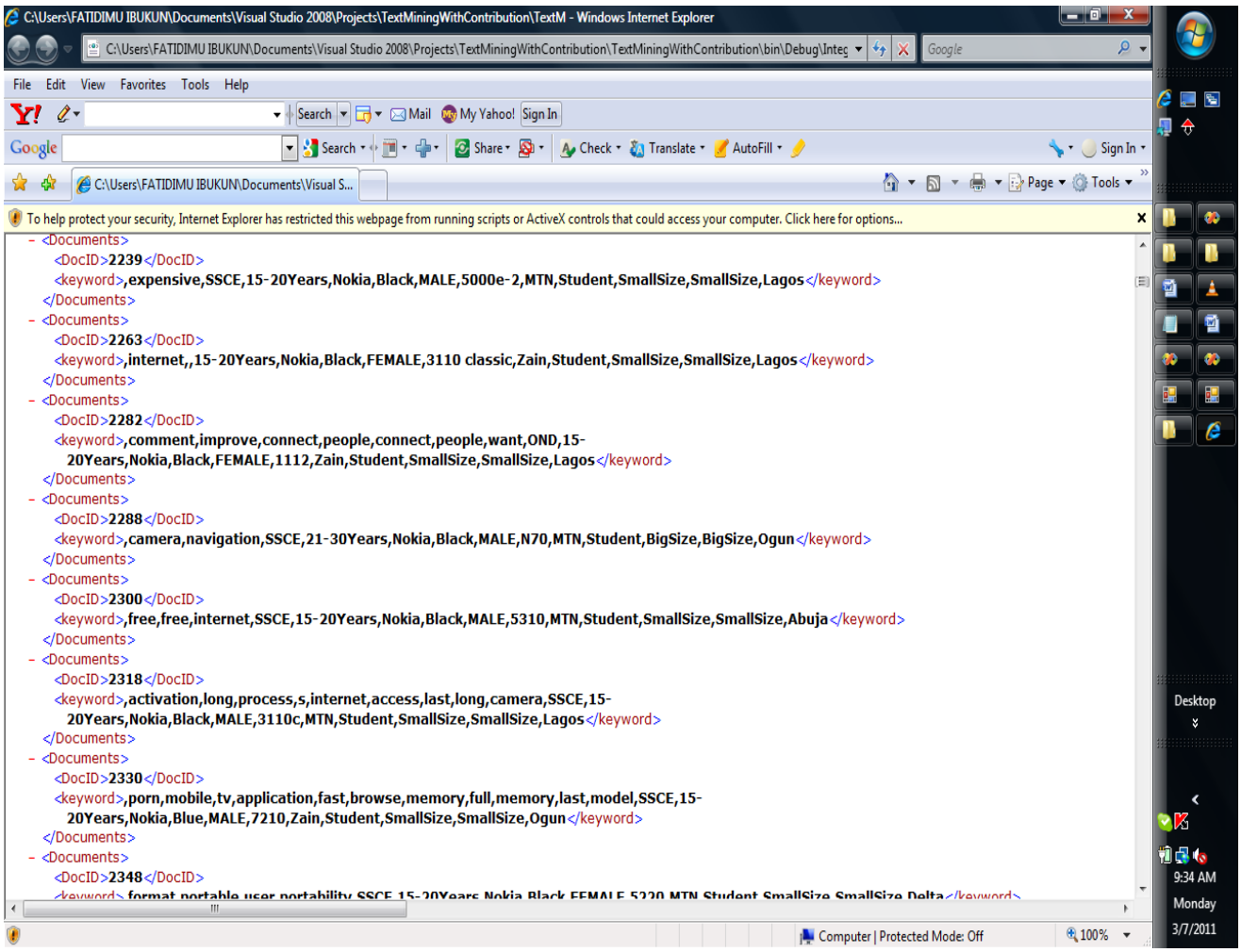


Figure 4.7: Selected Integrated data

The snapshot in Figure 4.7 represents the integrated data gotten from the resulting combination of the semantically clustered XML documents and their corresponding structured data. Each document is uniquely identified by their document ID. Figure 4.7 above therefore represents semantically related documents to be used for the association rule mining.

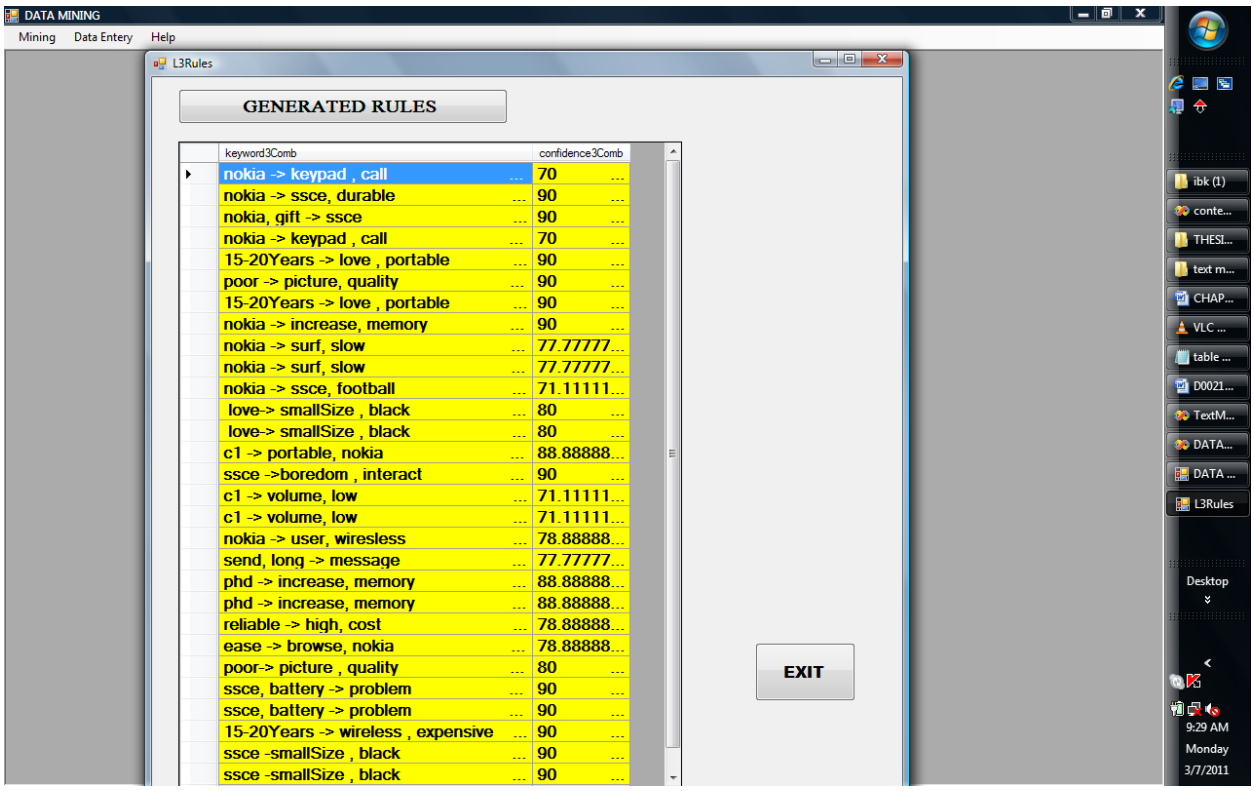


Figure 4.8 IIMA Rule Visualization Interface

4.7.3 Discussion

In order to have a fair representation of structured and unstructured data, a low relevance threshold of 20% was chosen, support of 5% and a higher threshold confidence value of 70% to make sure that the final rules gotten from the system are the most interesting ones. The following are samples of the resulting rules from the improved integrated mining system and they have been interpreted for effective customer relationship management.

- **Nokia -> ssce, durable**

  In the above rule nokia and ssce comes from the structured part of the data and durable comes from the unstructured part. The above rule can be interpreted to mean the users of nokia phones who also have an ssce certificate as their highest

qualification believe that durability is one of the qualities of mobile phones; this in turn can inform mobile phones producers who are targeting this kind of audience to make this as a key point of their marketing campaign.

- **c1 -> portable, nokia**

  C1 and nokia is the structured part while portable originated from the unstructured part. This rule gives the inference that customers have noticed the portability of the particular model of nokia mobile phone named c1.

- **15-20Years -> wireless , expensive**

  15-20Years is the structured part while wireless and expensive comes from the unstructured part. This rule gives the inference that people with the age bracket of 15-20 yeears find the wireless facility on mobile phones quite expensive to use**.**

- **nokia -> love, football**

  Nokia is the structured part while football and love is the unstructured part. This rule gives the inference that nokia users love football game and so to market to this audience or retain them as customers of nokia phone a form of football facility could be included on mobile phones, this could be in form of game or a special facility that motivates watching live football matches.

- **poor-> picture, quality**

  This rule gives the inference that mobile phones users are experiencing a poor picture quality which even though is not particularly associated with any brand of mobile phone can be used for competitive advantage such that any mobile phone producers who is willing to improve on pictures quality will gain more customers.

- **ssce ->boredom, interact**

  ssce comes from structured while boredom and interact comes from unstructured. It reveals that fact that ssce holders are always bored and probably want to interact, therefore facilities that can be used for interaction can be used to market to this kind of audiences.

- **nokia, gift -> 15-20Years**

  nokia and 15-20Years come from the structured part while gift comes from unstructured. This rule gives the inference that for this particular age bracket, nokia phones is usually used as a gift, this can inform mobile industry to position such products in gift shop and not only in the main phone market.

- **smallsize, problem -> message**

  smallsize comes from the unstructured part while problem and message comes from the unstructured part. This rule gives the inference that there is a problem sending messages with phones with small sizes. This inference is quite useful in order to improve on the quality of sending messages in small size phone.

- **2023 -> keypad, problem**

  2023 comes from the structured part while keypad and problem comes from the unstructured part. This inference gives pointers to the fact that with this particular brand of mobile phones, users do have a keypad problem.

- **nokia -> increase, memory**

  nokia comes from the structured part while increase and problem comes from the unstructured part. This rule gives the inference that nokia phone users want an

increase in memory for their phones. As marketing strategy, such facility can be used to attract customers.

- **phd -> increase, memory**

  phd comes from structured part while increase and memeory comes from unstructured part, this rule gives pointers to the fact that PhD holders advocate for increase in memory such facility will definitely help in marketing to such audience.

- **c1 -> volume, low**

  c1 comes from structured while volume and low comes from unstructured. This rule clearly reveals the fault associated to c1 model which is that its volume is quite low.

- **love-> smallSize , black**

  love comes from the unstructured part while small size and black comes from the structured part. This rule gives the inference that mobile phone consumers love to purchase their phones because its black and also because it have a small size, smart producers of such phones will therefore gain more customers if they can combine this two unique features into one.

- **ssce, battery -> problem**

  From this rule, ssce holders are complaining about battery problem of their phones a smart marketing strategy will be such that will capitalize on the improvement of the battery life of their product.

- **nokia -> surf, slow**

  Nokia comes from the structured part while surf and slow comes from the unstructured part. This rule gives the inference that surfing (browsing) on nokia

phones is slow as experienced by the consumer. This can be working upon towards customer satisfaction.

- **ease -> browse, nokia**

  Though, according to the rule above that browsing is slow on nokia mobile phone, it is quite easy to browse on this type of phones.

4.8 SUMMARY AND DISCUSSION

In this chapter the full scope of the application of the IIMA has been discussed using a practical case study of Customer relationship Management. The components such of IIMA such as Data preprocessing, knowledge distillation and rule visualization were developed. The developed modules were applied on a mobile phone industry case study and inferences were generated towards obtaining competitive advantage through effective customer relationship management.

The experience and observations gained from the application of these three aspects of IIMA in a practical real-life scenario, demonstrates the potential viability of the IIMA approach.

CHAPTER FIVE

EVALUATION OF THE IIMA APPROACH

5.1 INTRODUCTION

This chapter reports evaluation of the IIMA approach. It reports the detailed evaluation using both the objective and subjective means of evaluating association rules gotten from data mining systems. It also presents a comparative evaluation of the scenario of integrated mining system with or without improvement proposed in this thesis.

5.2 EVALUATION OVERVIEW

To discover hidden correlations, association rule mining methods use two important constraints known as support and confidence. However, mining methods are often unable to find the best value for these constraints: large number of rules when these thresholds are low; very few rules when these thresholds are high. In addition, regardless of these above thresholds, mining methods produce many rules that have identical meaning or redundant rules. Indeed, such redundant rules seem as a main impediment to efficient utilization of discovered rules and should be removed.

Basically in evaluating rules generated from data mining systems could be either objective or subjective. Objective measures, rely on the characteristics (surface features) of the patterns and the underlying data collection. In addition to the above, the subjective measure also considers users knowledge and interest (Xin & Yi-Fangm, 2006).

5.2.1 Objective Evaluation

In this research objective evaluation is used to assay the performance of the IIMA system. This comparism is directed towards the execution time and extracted rules in order to reveal

the performance of both system depending on the number of keywordsets. It is also aimed at examining which of the system (existing integrated mining system or IIMA) generates frequent keywordsets from the most important keywords rather than both the important and unimportant keywords. Consequently, this leads to extract interesting and uninteresting rules. The result of this evaluation helps to determine which system extracts the more interesting rules at short time.

In order to carry out the above, another system (existing integrated mining system) was designed without including the semantic XML clustering module for the extraction of keywords from the unstructured data. It was only based on information exaction technique which uses the TFIDF weight, and we call this the Existing system. This system corresponds to IIMA in the following processes:

- Transformation of documents into XML format
- Filtration and stemming of the transformed documents
- Reduction of keywords using the TF-IDF weighing scheme.

In order to have a fair representation of structured and unstructured data, a low TFIDF weight of 20% was chosen for the existing system. The same 20% relevance threshold was chosen for the IIMA system.

To measure the performance of the existing system to IIMA, we compared the large itemsets (first step of the association rule mining phase) generated from our system for different support thresholds with that of the one generated by the Existing System. The experiment was performed on the same corpus.

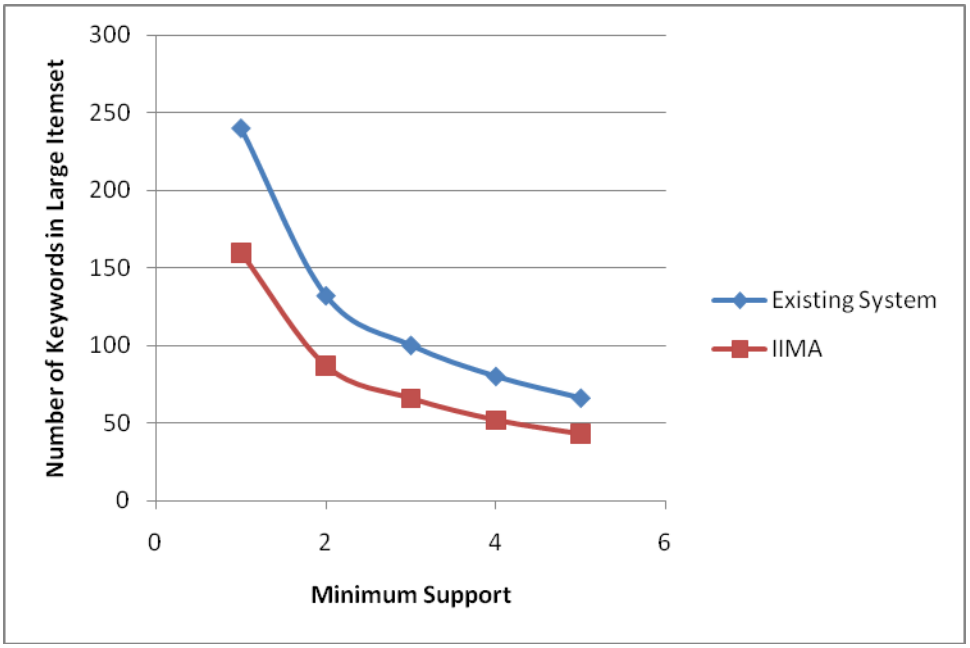


Figure 5.1 Improved Integrated Mining system Vs Existing system

The experimental results displayed in Figure 5.1 above reveals a reduction in the large itemset size generated from our system compared to the Existing system. Also, the execution time of our system was compared with the Existing system, to reveal the results displayed in Figure 5.2 below.

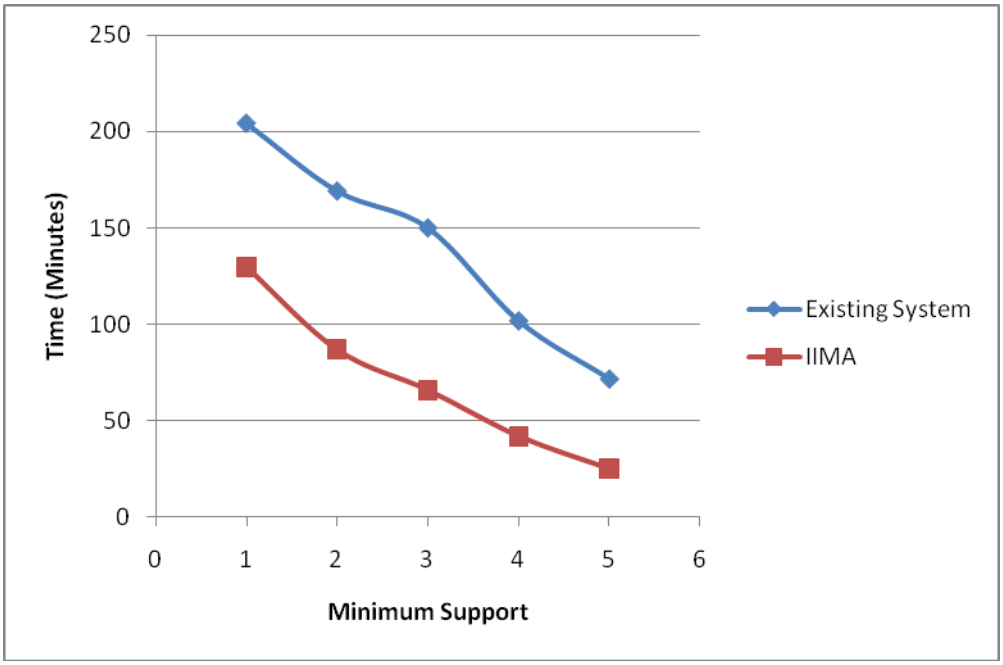


Figure 5.2 Graph of execution time against support.

The above shows that the IIMA system always outperforms the Existing system for all values of minimum support.

5.2.2 Motivation for Novelty Evaluation

A data mining system may discover a larger body of rules. However, relatively few of these may convey useful new knowledge to the user. Several metrics for evaluating the interestingness of mined rules have been proposed. These metrics can be used to filter out a larger percentage of automatically extracted less interesting rules, thus yielding a more manageable number of higher quality rules to be presented to the user. An important but less explored aspect of interestingness is novelty. Novelty refers to a rule representing an association that is currently unknown. If for example we discover rules from a computer science announcement posted to a news group such as SQL-> database. This kind of rule might be said to be uninteresting because it represents knowledge that is currently known. Evaluating the novelty of a rule requires comparing it to an existing body of knowledge the

user is assumed to already possess. For the purpose of this integrated mining which rules consists mostly of words in natural language, a relevant body of common knowledge is basic lexical semantics, i.e. the meanings of words and the semantic relationships between them. In this research, we have employed a method for measuring novelty of integrated-mined rules using wordNet.

5.2.3 Novelty Evaluation

The novelty evaluation used in this research is based on the work of (Basu et al., 2001). It falls under the subjective means of evaluating association rules.

This method is based on two basic steps;

1. Semantic Distance Measure

2. Rule Scoring algorithm

5.2.3.1 Semantic Distance Measure

We can define the semantic distance between two words w_i and w_j as:

$$d(w_i, w_j) = Dist(P(w_i, w_j)) + K \times Dir(P(w_{i,} w_j)) \qquad \text{Eq. 5.1}$$

Where $P(w_i,w_j)$ is a path between w_i and w_j, Dist(p) is the distance along path p according to our weighting scheme, Dir(p) is the number of direction changes of relations along path p, and K is a suitably chosen constant. The second part of the formula is derived from the (Hirst et. al., 1998), where the relations of WordNet are divided three direction classes-"up", "down" and "horizontal", depending on how the two words in the relation are lexically related.

Table 5.1 summarizes the direction information for the relation types used. The more the change in direction in the path from one word to another, the greater the semantic distance between the words. Changes of direction along the path reflect large changes in semantic context.

The path distance it is based on the semantic distance definition of sussna (1993). In this definition, the path distance is defined as the shortest weighted path between w_i and w_j. Every edge in the path is weighted according to the weight of the wordNet relation corresponding to that edge, and is normalized by the depth in the WordNet tree where edge occurs. There are 15 different relations between words in wordNet and these relations have been assigned different weights. The weight chosen for different relations are given in the table 5.1;

Table 5.1 Relation table (Basu et al., 2001)

Relation	**Direction**	**Weight**
Synonym. Attribute, Pertainym, Similar	Horizontal	0.5
Antonym	Horizontal	2.5
Hypernym,(Member/Part/Substance) Meronym	Up	1.5
Hyponym,(Member/Part/Substance) Holonym, Cause, Entailment	Down	1.5

5.2.3.2 Rule scoring Algorithm

The scoring algorithm of rules according to novelty is outlined in the algorithm below

For each rule in a rule file

Let A = set of antecedent words,

C = set of consequent words

For each word $w_i \in A$ and $w_j \in C$

If w_i and w_j are not a valid words in WordNet

Score (w_i, w_j) <- PathViaRoot ($d_{avg,}d_{avg}$)

Elseif w_j is not a valid word in WordNet

Score (wi, w_j) <- PathViaRoot(w_i,d_{avg})

Elseif w_i is not a valid word in WordNet

Score (w_i, w_j) <- PathViaRoot(d_{avg},w_j)

Elseif path not found between w_i and w_j (in user-specified time- limit)

Score (w_i, w_j) <- PathViaRoot(w_i,w_j)

Else

Score (w_i,w_j)<- d(w_i,w_j)

Score of rule = Average of all (w_i,w_j) scores

Sort scored rule in descending order

Figure 5.3 Rule Scoring Algorithm (Basu et al., 2001)

The noun hierarchy of the WordNet is disconnected, there are 11 trees with distinct root nodes. The verb hierarchy is also disconnected with 15 distinct root nodes. After introducing R_{nouns}, R_{verbs} and R_{top}, all words in the Wordnet are connected to each other. So in this composite hierarchy, any two words are connected by a path. The function PathViaRoot computes the distance of the default path. For nouns and verbs the ePathViaRoot function

calculates the distance of the path between the words as the sum of the path distances of each word to its root.

If one of the words is an adjective or an adverb, and the shortest path method does not terminate within the specified time-limit, then the algorithm finds the path from the adjective or adverb to the nearest noun, through relations like "pertainym", "attribute", etc. It then finds the default path up the noun hierarchy, and the PathViaRoot function incorporates the distance of the path into the path distance measurement.

If some of the words extracted from the rules are not valid in Wordnet e.g. abbreviations, names like Philip, domain specific terms like booknews, etc., they are assigned the average depth of a word in the wordNet hierarchy, which was estimated by sampling techniques to be about 6, and then estimated its path distance to the root of the combined hierarchy by using the PathViaRoot function.

5.3 SUBJECTIVE EVALUATION RESULTS

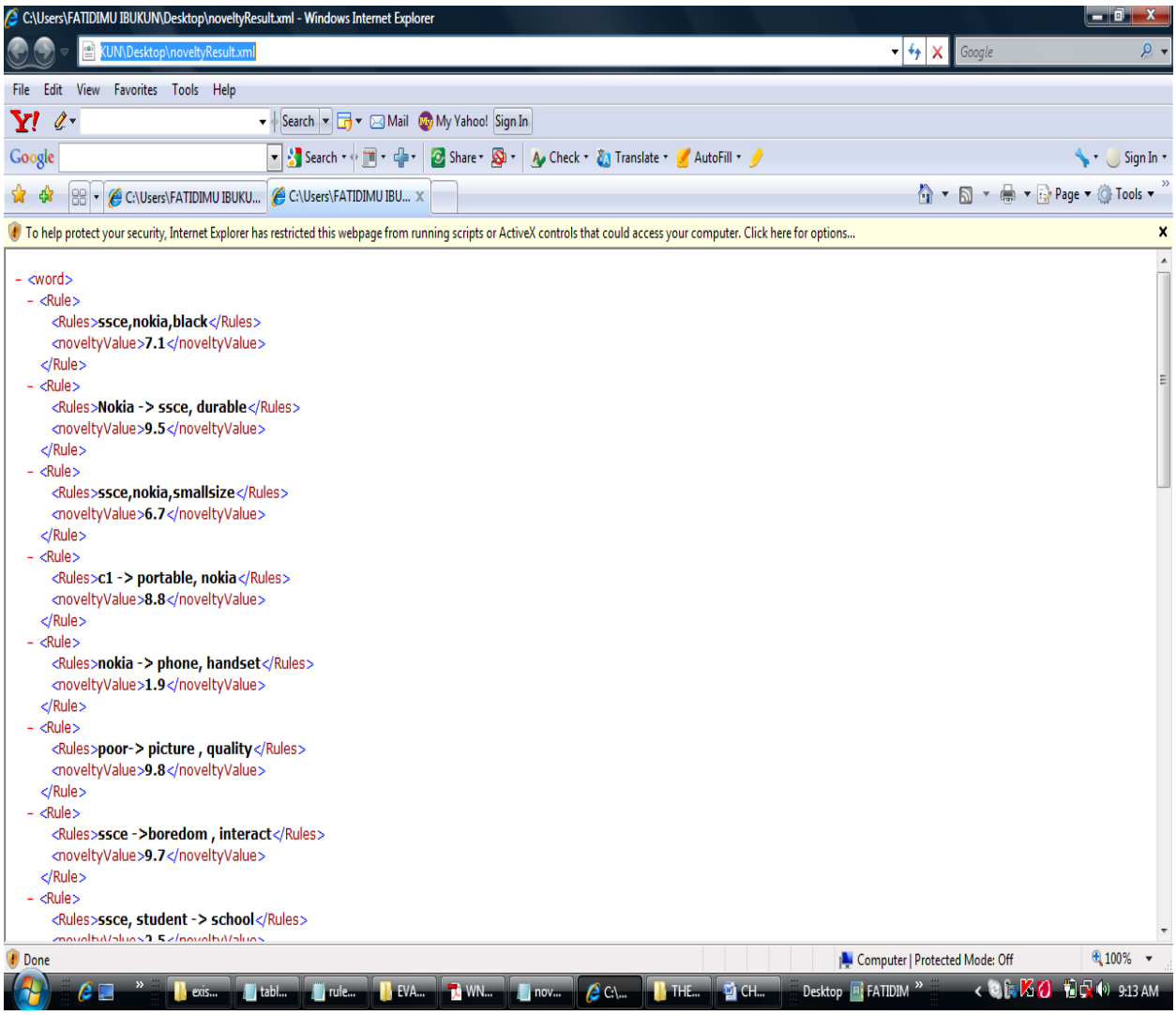


Figure 5.4 Evaluation rules in XML format for IIMA system

Each XML <Rule> </Rule> node in figure 5.4 represents each rule and its corresponding novelty value as calculated by the novelty methodology described above. This snapshot was taking before eliminating the rule termed uninteresting.

5.3.1 Discussion

The above evaluation algorithm was applied on the existing system and the IIMA system. In order to compare the percentage of rules that are novel for the existing system and the IIMA system, we evaluated rules gotten at the following thresholds for both systems.

Table 5.2 Result (1) displayed

Existing System	IIMA System
TFIDF Weight = 20%	Relevance weight = 20%
Support =5%	Support = 5%
Confidence=70%	Confidence= 70%

For each of the system, 50% of the highest evaluation result was taken as the medium evaluation score and the rules having their evaluation score below this value for each system was eliminated. Using the following formula, we calculated the percentage of the rules that are novel for each system to reveal the following.

$$\%novelty = \frac{\text{Total number of novel rules}}{\text{Total number of rules generated}} \times 100 \qquad \text{Eq. 5.2}$$

Table 5.3 Result (2) displayed

Existing System	IIMA System
Percentage novelty= 54%	Percentage novelty= 71%

5.4 POSSIBILITIES FOR GENERALIZATION OF RESULT

Having shown that IIMA produced a 17% increase in novel rules generated, in this case study in which it is applied, we therefore postulate that IIMA can indeed be applied to generate 71% novel rules in other domain of application due to the fact that the developed system is not domain dependent. IIMA is particularly applicable to domain whose terminologies are largely represented in a lexical database such as wordNet which was used for this experiment.

5.5 SUMMARY AND DISCUSSION

In this chapter a report of the procedure adopted for the evaluation of the IIMA approach and its results were clearly stated. It is shown that there was indeed an improvement in the IIMA system over the existing system variant. Furthermore, the case study scenario has demonstrated the applicability of IIMA in a real-life context and proved the viability of the IIMA approach.

This is because IIMA produced measurable reduction in time and numbers of rules generated, demonstrated the potential to improve customer relationship management decision based on highly reliable and novel rules. The case study therefore, successfully validates IIMA as platform for generating dependable rules to populated decision support systems.

CHAPTER SIX

SUMMARY OF FINDINGS, CONCLUSION AND FUTURE WORK

This Chapter summarizes and discusses the contributions of the thesis, and presents an outlook of the opportunities for future research work. The thesis has presented a platform to mine from integrated data.

6.1 SUMMARY OF FINDINGS

The thesis has shown that integrating structured and unstructured data is a critical component for the success of the modern enterprise because of its ability to take advantage of all available information in a holistic perspective.

However, the existing few integrated mining systems need to reduce the level of uncertainty in decision making based on the quality of rules on which these decisions are based. Also, there is a need to apply such discoveries to improve the field of customer relationship management.

The thesis intervened by introducing an approach which is a solution to the above concerns. This system coined IIMA (Improved Integrated Mining architecture) is a hybrid of Association Rule Mining and Information Extraction technique based on content similarity of XML (Extensible Markup Language) document.

IIMA approach consists of three main phases; 1) Extraction and Integration phase. This phase is aimed at optimizing the performance of the knowledge mining phase. 2) knowledge distillation phase which is concerned with generating association rules based on the result of the extraction and integration phase 3). The rules visualization phase whereby the generated rules are interpreted for making efficient business decisions.

IIMA is based on a set of assumptions which defines the condition for its optimal applicability. These are:

- That the domain knowledge is fairly represented in a lexical database or an ontology.
- High threshold value for confidence and low threshold value for support have to chosen to ensure a fair representation of the important variables for the rule generation.

The thesis provided a validation of the IIMA approach by using a case study of creating a competitive advantage for mobile phones industry through effective customer relationship management. This has demonstrated the applicability and viability of IIMA in a real-life context.

Based on the results obtained from mining customer relationship management integrated data in mobile phone industry, the thesis made some significant contributions. Firstly, in the world of business decision support system (business intelligence), the task of integrating various data types which has been the burden of the enterprise application developer (Roth et al., 2002; Garcia-Molina et al., 1995; Tomasic et al., 1997; Adali et al., 1996; Levy et al., 1996; http://www.infoshark.com), has been addressed. By applying IIMA on CRM data gathered for the purpose of experimental validation in this research, the result revealed novel CRM inferences which are an advantage of our modified market decision support system.

According to (Frieder et al., 2000; Roth et al., 2000; Dean & Alexandra, 2004; Aravindan, 2005; Robert, 2006; Prem, 2007; Sukumaran & Sureka, 2007; An Oracle White paper, 2007) data integration has not been approached from the semantic analysis and those that have been relies on domain specific ontologies making it limited in flexibility. The second contribution therefore is to address the novel problem of classifying semantically related XML

documents. Applying IIMA on mobile phone industry case study generated inferences towards competitive advantage through effective customer relationship management. This is due to the fact that semantically related keywords extracted from the combination of structured and unstructured component of the data mostly generated the novel rules.

6.2 CONCLUSION

This research has proposed a more efficient approach of integrating structure and unstructured data for integrated mining in decision support system by minimizing the practice of handling structured and unstructured data as distinct information entities, which often results in decision management failure. The research has provided a domain independent, flexible and efficient solution to discovering decision making rules. It has succeeded in extracting association rules which contain important features which form a worthy platform for making effective decisions as regards customer relationship management in the mobile phones manufacturing industry. This was made possible due to the efficient refinement of the data selected for mining from both the structured and unstructured platform. This refinement was brought about by the semantic clustering of unstructured data.

Also, identifying user requirements and understanding the user is a major part of contributing to the profit of the organization and this can be achieved through competitive intelligence. This research has helped to reduce the uncertainty and inaccuracy of rules from which decisions are based towards the competitive advantage of an organization. The generated rules therefore, give pointers to various characteristics of customers of mobile phone manufacturing industry which will help in identifying, attracting, developing and

maintaining successful customer relationships over time in order to increase retention of profitable customers.

6.3 FUTURE WORK

The thesis provides several opportunities for further research in the immediate future. The IIMA approach as implemented in this thesis has some limitations which has created some research possibilities to enhance the concept in the following areas:

1) Ontological knowledge

In future research, the role of automatically generated ontological knowledge in supporting the detection of semantic relatedness among XML data is very viable. Of particular interest is WSD (Word Sense Disambiguation). The WSD algorithm is based on a context-free grammar, to find structural semantic interconnections: structural specifications of the possible senses for each word in a context; the graph representation of word senses can be automatically built from several sources, including WordNet, annotated corpora, and glossaries. (Navigli & Velardi, 2005).

There are also other approaches such as direct approaches that require a preliminary qualitative/quantitative analysis of the characteristics of the data; a different approach may be devised by mapping the problem at hand to a problem of clustering ensembles (Strehl & Ghosh, 2002). In this way, multiple clustering solutions would be generated, each one according to a different setting of the f-γ parameter field (and other specific parameters of the clustering algorithm(s) being used), in order to form a clustering ensemble; then, from this ensemble, the consensus partition would be selected by employing a clustering ensembles method.

2) Increase in the unstructured components

In future studies, the algorithms can be improved by taking the attachments of the e-mails (pictures, text files, etc.) into consideration and extending the proposed system to multilingual context.

3) Interpretation of association rules

The integrated mining system can be extended to use the concept features to represent text and to extract the more useful association rules that have more meaning. Moreover, the system can be improved by visualizing the extracted association rules in graphical representation in two or three-dimension association networks.

REFERENCES

Adali S., Candan K., Papakonstantinou Y., and Subrahmanian V. S., "Query Caching and Optimization in Distributed Mediator Systems," Proceedings, ACM SIGMOD Conference on Management of Data, Montreal, Canada, pp. 137–148, 1996.

Ah-Hwee T., "Text mining: The state of the art and the challenges", 2006.

Al-Khalifa S., Yu C., and Jagadish H.V.. "Querying structured text in an xml database" In SIGMOD, 2003.

Amin A.Shaqrah "Using Knowledge Sharing Strategies as an External Structure to Improve CRM: An Empirical Investigation toward a Conceptual Frame work" A thesis submitted to the Arab Academy for Banking and Financial Sciences, Febuary 2008. See (http://www.aabfs.org/Ar/pdf/15.pdf).

Andrea T. and Sergio G., "Semantic Clustering of XML documents", ACM Trans. Inform. Syst. 28, 1, Article 3 DOI = 10.1145/1658377.1658380 http://doi.acm.org/10.1145/1658377.1658380, 2010.

Andrew McCallum & David Jensen, "A Note on the Unification of Information Extraction and Data Mining using Conditional-Probability, Relational Models" Center for Intelligent Information Retrieval, SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-02-1-8903, Advanced Research and Development Activity under contract number MDA904-01-C-0984, and the Knowledge Discovery Laboratory and DARPA contract F30602-01-2-0566, 2003

Andrew P. S.(1991) "Decision Support Systems Engineering". John Wiley & Sons, Inc., New York.

AnHai Doan, Raghu Ramakrishnan, Shivakumar Vaithyanathan" Managing Information Extraction" *SIGMOD 2006,* June 27–29, 2006, Chicago, Illinois, USA. ACM 1595932569/ 06/0006 ...$5.00.

An Oracle White paper , "Semantic Data Integration for the Enterprise", 2007.

Aravindan R., Meera J., Biploh D., David H.C. D., Mohamed M., "A unified framework for storing and Querying Unstructured And Structured Data.", 2005.

Arnold S.E.. "Beyond Search-and-Retrieval: Enterprise Text Mining with SAS®" SAS white paper www.sas.com/technologies/analytics/datamining/textminer,http:/www.cxoamerica.com/pastissue/printarticle.asp?art=25408, 2010.

Asoo J V., "e-business and supply chain management" Decision Sciences; Fall 2002; 33, 4; ABI/INFORM Global pp. 495-504, 2002.

Ayo C.K, Ekong U.O. and Fatudimu I.T., " The Prospects of m-Commerce Implementation: Issues and Trends", *Proceedings of the* 8th IBIMA Conference, 2007.

Baeza-Yates R. and Ribeiro-Neto B. "Modern Information Retrieval". ACM Press, New York, 1999.

Bahl, L. R., Jelinek, F. and Mercer, R. L. "A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence" 5(2), pp179–190, 1983.

Basu S., Mooney R. J., Pasupuleti K. V. and Ghosh J. "Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge". Proceedings of the Seventh ACM SIGKDD Conference, pp. 233-238, San Francisco, CA, 2001

Berardi M., Malerba D., Marinelli C., Leo P., Loglisci C., Scioscia G. "A text-mining application able to mine association rules from biomedical texts", 2005.

Bikel D.M., Schwartz R., and Weischedel R.M.. "An algorithm that learns what's in a name. Machine Learning, 34:211–232, 1999.

Black, E., Jelinek, F., Lafferty, J. D., Magerman, D. M., Mercer, R. L. and Roukos, S. "Towards History-Based Grammars: Using Richer Models for Probabilistic Parsing" In Proceedings DARPA Speech and Natural Language Workshop, Harriman, New York, pp.134–139. Los Altos, CA: Morgan Kaufman, 1992.

Blumberg, R., and Atre, S. "The Problem with Unstructured Data," DM Review(13:4), 2003. http://www.information-management.com/issues/20030201/6287-1.html

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R. and Rossin, P. "A Statistical Approach to Machine Translation". Computational Linguistics 16(2), pp79–85, 1990

Brown, P., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. "The Mathematics of Statistical Machine Translation: Parameter Estimation." Computational Linguistics 19(2), pp263–311, 1993.

Brundage M., XQuery: The XML Query Language, Addison-Wesley Professional, 2004

Buckley C., Salton G., and Allan J. "The effect of adding relevance information in a relevance feedback environment" In Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval. Springer-Verlag, 1994

Charniak, E. "Statistical Parsing with a Context-Free Grammar and Word Statistics". In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97). Menlo Park: AAAI Press, 1997.

Ch. Aswani K. and Srinivas S., **"**On the Performance of Latent Semantic Indexing-based Information Retrieval" Journal of Computing and Information Technology - CIT 17, 2009, 3, 259–264 doi:10.2498/cit.1001268, 2009

Church, K. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. Second Conference on Applied Natural Language Processing, ACL, 1988.

Cody W.F., Kreulen J.T., Krishna V. and Spangler W.S., "The integration of business intelligence and knowledge management" IBM SYSTEMS JOURNAL, VOL 41, NO 4, pp 697-713, 2002

Croft W. B. , L. A. Smith, and H. R. Turtle, "A loosely-coupled integration of a text retrieval system and an object-oriented database system". In Proc. of the 15th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 223–232, June 1992

Cunningham H., Maynard D., Bontcheva K., and Tablan V. "Experience with a language engineering architecture: Three years of GATE," Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02), 2002.

Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. "A Practical Part-of-speech Tagger". In Third Conference on Applied Natural Language Processing, ACL, pp133–140, 1992.

Daniel J.P. and Shashidhar K., "Building Web-based Decision Support System" Studies in Informatics and Control Vol. 11 No 4. pp 291 -302

David Wai-Lok Cheung, Jiawei Han, Vincent Ng, Ada Wai-Chee Fu, and Yongjian Fu, A Fast Distributed Algorithm for Mining Association Rules, *Proceedings of PDIS*, 1996.

Dean W. and Alexandra P. "An example of the ESTEST approach to combining unstructured text and structured data," In DEXA Workshops, pages 191-195. IEEE Computer Society, 2004.

Denoyer, L. and Gallinari, P. Report on the XML Mining Track at INEX 2005 and INEX 2006: Categorization and clustering of XML documents. *SIGIR Forum 41*, 1, 79–90., 2007.

Deutsch, A., Fernandez, M., and Suciu, D. "Storing semistructured data with STORED". In Proceedings of the of ACMSIGMOD International Conference on Management of Data (SIGMOD). pp431–442, 1999

Doorenbos R.. B, Etzioni O, and Weld D.S. A scalable comparison-shopping agent for the World-Wide Web. In Proceedings of the First International Conference on Autonomous Agents (Agents-97), pages 39–48, Marina del Rey, CA, Feb. 1997.

Druzdzel, M. J. and Flynn, R. R. " Decision Support Systems," to appear in Encyclopedia of Library and Information Science, Second Edition, Allen Kent (ed.), New York: Marcel Dekker, Inc, 2002

Erhard R., Andreas T., David A., "Dynamic Fusion of Web Data", 2007.

Evans D. A. and Zhai C., "Noun-phrase analysis in unrestricted text for information retrieval," in Proceedings of the 34th Annual Meetings of the Association for Computational Linguistics, pp. 17-24, 1996.

Fang-Yie, L& Chih-Chieh, K. (2010). An Automated Term Definition Extraction System Using the Web Corpus in the Chinese Language *Journal Of Information Science and Engineering*, 26, 505-525.

Feldman, R. & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). *In Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining,* 112 – 117

Feldman R. and Hirsh H., "Mining associations in text in the presence of background knowledge," in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, USA, 1996

Fiebig, T., Helmer, S., Kanne, C., Moerkotte, G., Neumann, J., Schiele, R., And Westmann, T. Anatomy of a native XML base management system. *VLDB J. 11*, 4, 292–314, 2002

Flesca, S., Furfaro, F., Greco, S., and Zumpano, E. Repairs and consistent answers for XML data with functional dependencies. In *Proceedings of the International XML Database Symposium (XSym)*. 238–253, 2003.

Frieder O., Chowdhury A., Grossman D., and McCabe M. C., " On the Integration of Structured Data and Text: A Review of the SIRE Architechure," Information Retrieval Labouratory, Illionis Institute of Technology, 2000.

Fuhr N. and Grobjohann K.. "XIRQL: A language for information retrieval in XML documents," In Proc. Of SIGIR, 2001.

Ganesh Ramakrishnan, Sachindra Joshi, Sreeram Balakrishnan and Ashwin Srinivasan "Using ILP to Construct Features for Information Extraction from Semi-Structured Text" 2006.

Gale, W. A., Church, K. W. and Yarowsky, D. "A Method for Disambiguating Word Senses in a Large Corpus". Computers and the Humanities 26, pp415–439, 1992.

Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Ullman J., and Widom J., "Integrating and Accessing Heterogeneous Information Sources in TSIMMIS,"

Proceedings, AAAI Symposium on Information Gathering, Stanford, CA, pp. 61–64, 1995.

Guillaume, D. and Murtagh, F. Clustering of XML documents. *Comput. Phys. Comm. 127*, 215–227, 2000.

Goffman, E. "Frame Analysis: An Essay on the Organization of Experience." London: Harper and Row, 1974

Goshal, S. and D. Westney. . "Organizing competitor analysis systems." Strategic Management Journal, 12: pp17-31.,1991

Graham H., John S., Nigel P., "Marketing strategy and Competitive Positioning" published by Pearson Education Limited, pp 180-232, 2004.

Han J. and Kamber M. "Data Mining: Concepts and Techniques." Morgan Kaufmann, San Francisco, 2000

Hany M.; Dietmar R.; Nabil I. and Fawzy T. "A Text Mining Technique Using Association Rules Extraction", International Journal of Computational Intelligence volume 4 number 1 2007 ISSN 1304-2386, 2007.

Hirst and D, St-Onge. Lexical chains as representations of context for the detection and correction of malapropism. In C. Fellbaum, editor, WordNet: An Electronic Lexical Database, chapter 13, pages 305-332. MIT Press. 1998.

Inmon, B. "Structured and Unstructured Data, Bridging the gap," in Business Intelligence Network's Bill Inmon Channel. http://www.b-eye-network.com/view/4955,2007

Jaideep Srivastava "Data Mining for Customer Relationship Management (CRM)"

JAGADISH, H., Al-Khalifa, S., Chapman, A., Lakshmanan, L., Nierman, A., Paparizos, S., Patel, J., Srivastava, D., Wiwatwattana, N., Wu, Y., And Yu, C. (2002) Timber: A native XML database. *VLDB J. 11*, 4, 274–291

Jan P. and Peter B. "Text Mining for Documents Annotation and Ontology Support" IST-1999-20364 Webocracy: "Web Technologies Supporting Direct Participation in Democratic Processes".

Jasper E., "Global query processing in the AutoMed heterogeneous database environment." In Proc. BNCOD02, LNCS 2405, pp46-49, 2002

Jelinek, F. "Continuous Speech Recognition by Statistical Methods". Proceedings of the IEEE 64(4), pp 532–557, 1976.

Jessup E. R., Martin J. H., Taking a New Look at the Latent Semantic Analysis Approach to Information Retrieval. Computational Information Retrieval, SIAM Publishers, pp121–144, 2001

Joakim N., "On Statistical Methods in Natural Language Processing" ,2000. Joakim N., "On Statistical Methods in Natural Language Processing" ,2000.

John E. P. "The evolution of Competitive Intelligence, designing a process for action" journal of the Association of Proposal Management Professionals. pp37-52, 1999.

Kahaner, L. "Competitive Intelligence: From Black Ops to Boardrooms How Businesses Gather, Analyze, and Use Information to Succeed in the Global Marketplace." New York: Simon and Schuster, 1996

Karin V., Antonio S., Mark E. and Ed MacKerrow, "Deploying Natural Language Processing for Social Science Analysis", 2004.

Keen, P.G.W., and Scott Morton, M.S. Decision Support Systems: An Organizational Perspective, Addison- Wesley, Reading, MA, 1978.

Kernochan, W. "XQuery and XML data: DB2 helps manage the era of unstructured data," Infostructure Associates, 2006. http://searchdatamanagement.techtarget.com/tip/XQuery-and-XML-data-DB2-helps-manage-the-era-of-unstructured-data

Klemettinen M., Mannila H., Ronkainen P., Toivonen H. and Verkamo A. I. "Finding Interesting Rules from Large Sets of Discovered Association Rules". Proceedings of the CIKM Conference, 1994.

Kreulen, J. T. The integration of business intelligence and knowledge management. *IBM Systems Journal*. 41(4), 697-713, 2002.

Kushmerick N., Weld D. S., and Doorenbos R. B. Wrapper induction for information extraction. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97), pages 729–735, Nagoya, Japan, 1997.

Lahiri, T., Abiteboul, S., And Widom, J. Ozone: Integrating structured and semistructured data. In *Proceedings of the International Conference on Database Programming Languages (DBPL)*. 297–323, 1999.

Lauría E.J. and Peter J. D., (2004) "A Bayesian Belief Network-driven Decision Support System for Client-Server Implementation" Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8.3 International Conference pp 447-456.

Lawrence R., Perlich C., Rosset S., Arroyo J., Callahan M., Collins M., Ershov A., Feinzig S., Khabibrakhmanov I., Mahatma S., Niemaszyk M., and Weiss S.. "Analytics-

driven solutions for customer targeting and sales force allocation." IBM Systems Journal, 2007.

Leveling, J. and S. Hartrumpf . "Indexing and translating concepts for the GIRT task." University of Hagen at CLEF 2004: pp. 271–282. Berlin: Springer, 2005

Levy A., Rajaraman A., and Ordille J. J., "Querying Heterogeneous Information Sources Using Source Descriptions," Proceedings, 22nd International Conference on Very Large Data Bases, Bombay, India, pp. 251–262, 1996.

Linus Osuagwu, "Small Business & Entrepreneurship Management", Second Edition, Published by Grey Resource Limited, pp 9-10, 2006.

Liu B., Ma Y. and Lee R. "Analyzing the interestingness of association rules from the temporal dimension". IEEE International Conference on Data Mining, Silicon Valley, CA, 2001.

Madhavan, J., Jeffery, S. R., Cohen, S., Dong, X., Ko, D., Yu, C., Halevy, A. "Web-scale Data Integration: You can only afford to Pay As You Go." Proc. CIDR, 2007

Margaret H. Dunham, Yongqiao Xiao, Le Gruenwald and Zahid Hossain, "A Survey of Association Rules" http look for it.

Marios Skounakis ,Mark Craven & Soumya Ray "Hierarchical Hidden Markov Models for Information Extraction", NIH grant 1R01 LM07050-01, NSF grant IIS-0093016, and a grant to the University of Wisconsin Medical School under the Howard Hughes Medical Institute Research Resources Program for Medical Schools, 2003

McKinnon, S. and Burns W. "The Information Mosaic" Boston: Harvard Business School Press. 1992.

Merialdo, B. Tagging English Text with a Probabilistic Model. Computational Linguistics 20(2), 155–172., 1994

Mika H. and Virpi P., "Investigating Business Information Management Practices in Large Finnish Companies," in FRONTIERS OF E-BUSINESS RESEARCH, pp 121-136, 2002.

Monica C. H., Roger L. H., and Jie Z. "Enchancing The Use Of Bidss/Ci Systems: A Proposed Framework" Issues in Information Systems, VOL IX, No. 2, 2008.

Navigli, R. & Velardi, P. "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites". In Computational Linguistics, Volume 30, Issue 2. June, 2004.

Nasukawa T. and Nagano T. "Text analysis and knowledge mining system" IBM SYSTEMS JOURNAL, VOL 40, NO 4, pp 967-984, 2001

Nayak, R. And Xu, S. XCLS: a fast and effective clustering algorithm for heterogeneous XML documents. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. 292–302, 2006

Nierman, A. And Jagadish, H. Evaluating structural similarity in XML documents. In *Proceedings of the ACM SIGMOD International Workshop on the Web and Databases (WebDB)*. 61–66, 2002.

Noy N. and McGuinness D. (), "Ontology Development Guide 101: A guide to creating your first ontology".

Padmanabhan B. and Tuzhilin A. Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems, 27:303-318, Elsevier Science, 1999.

Papakonstantinou Y., Garcia-Molina H., and Widom J.. Object exchange across heterogeneous information sources. In Proc. ICDE Conf., pp 251-260, 1995

Piatesky-Shapiro G. and Matheus C. The interestingness of deviations. KDD-94, 1994.

Prem M., Yan L., Richard L., Ildar K., Cezar P., Timothy B. "Finding New Customers Using Unstructured and Structured Data" KDD, San Jose, California, USA, 2007.

Poulovassailis A. "The AutoMed Intermediate Query Language Technical report," AutoMed Project, 2001.

Prescott, J. and P. Gibbons. (Eds.) "Global Perspectives on Competitive Intelligence Alexandria, VA: SCIP, 1993.

Rajman M. and Besancon R., "Text mining: natural language techniques and text mining applications", in Proc. 7th working conf. on database semantics (DS-7), Chapan &Hall IFIP Proc. Series. Leysin, Switzerland pp7-10, 1997

Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, Mining Association Rules Between Sets of Items in Large Databases, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207-216, Washington, D.C., May 1993.

Raymond J. M. and Un Yong N., "Text Mining with Information Extraction" Proceedings of the 4th International MIDP Colloquium, Van Schaik Pub., South Africa, pp.141-160, 2005

Robert M. L., "Browsing Mixed Structured and Unstructured Data" Information Processing & Management 42 (2), pp 440–452, 2006

Rosset S. and Lawrence R. "Data enhanced predictive modeling for sales targeting". In Proceedings of SIAM Conference On Data Mining, 2006.

Roth M.A., Wolfson D.C., Kleewein J.C., and Nelin C.J., "Information Integration: A New Generation of Information Technology," IBM Systems Journal 41, No. 4, pp563-577, 2002

SahugueT, A. Kweelt, the making-of: Mistakes made and lessons learned. Tech. Rep., Department of Computer and Information Science, University of Pennsylvania, 2000.

Sánchez, D. and A. Moreno . "Web-scale taxonomy learning". In Proceedings of the Workshop on Learning and Extending Lexical Ontologies by using Machine Learning (OntoML 05), Bonn, Germany, 2005

Shanmugasundaram, J., Tufte, K., Zhang,C.,He, G.,Dewitt, D., Andnaughton, J. Relational databases for querying XML documents: Limitations and opportunities. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*. 302–314, 1999.

Shenzhi L., Tianhao W. and William M.P. "Distributed Higher Order Association Rule Mining using Information extraction from Textual Data", SIGKDD Explorations, Volume 7, Issue 1-pp 26-35, 2004

Sean E. "The Changing Structure of Decision Support Systems Research: An Empirical Investigation through Author Cocitation Mapping" in Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8.3 International Conference pp 243-251, 2004.

Silberschatz and Tuzhilin A.. "What Makes Patterns Interesting in Knowledge Discovery Systems". IEEE Transactions on Knowledge and Data Engineering, 8 (6), 1996.

Solomon Negash, "Business Intelligence" Communications of the Association for Information Systems Volume13, 177-195, 2004.

Solomon N. and Paul G., "BUSINESS INTELLIGENCE" Ninth Americas Conference on Information Systems, pp 3190-3199, 2003

Solomon N. and Paul G., "Handbook on Decision Support Systems 2" Published by Springer Berlin Heidelberg, pp 175-193, 2008.

Stolcke, A. "An Efficient Probabilistic Context-Free Parsing Algorithm That Computes Prefix Probabilities". Computational Linguistics 21(2), pp165–202, 1995.

Strauss, J., El-Ansary, A., Frost, R., "E-marketing", International Edition, Published by Pearson Prentice Hall, pp30, 135-168, 2006.

Strehl A. and Ghosh J. "Cluster ensembles—a knowledge reuse framework for combining multiple partitions". J. Mach. Learn. Res. *3*, 583–617, 2002

Sukumaran, S. and Sureka, A., "Integrating Structured and Unstructured Data Using Text Tagging and Annotation," in Business Intelligence Best Practises [SM], 2007. http://www.bi-bestpractices.com/view-articles/4735, http://www.tdan.com/view-articles/4735

Tomasic A., Amouroux R., Bonnet P., Kapitskaia O., Naake H., and Raschid, L. "The Distributed Information Search Component (Disco) and the World Wide Web," Proceedings, ACM SIGMOD Conference, Tuscon, AZ pp. 546–548, 1997.

Taffet1 D. "Application of Natural Language Processing Techniques to Enhance Web-Based Retrieval of Genealogical Data", 2001

Teng-Kai, F. & Chia-Hui, C. (2010). Exploring Evolutionary Technical Trends From Academic Research Papers *Journal Of Information Science And Engineering* 26, 97-117

Theobald and G. Weikum. "The index-based XXL search engine for querying XML data"

Ukelson, J. "Combining Structured, Semi structured and Unstructured Data in Business applications," in DM Direct Newsletter, 2006. http://www.information-management.com/infodirect/20061201/1069202-1.html

Unitas Corporation, " A Single View: Integrating Structured and Unstructured Data/Information with the Enterprise," in unitas, the portal is the business ™,. http://lsdis.cs.uga.edu/GlobalInfoSys/Structured-and-Unstructured-for-EIPs.pdf, 2002.

Viral P., Jack G., and Tim F. (2004) "Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies" EXCELON CORP.(2002). eXcelon XML platform.

Wan J. W.W. and Dobbie G., "Extracting association rules from XML documents using XQuery," Proceedings of the Fifth ACM International Workshop on Web Information and Data Management, New Orleans, LA, USA, November 2003, pp 94–97

Xiaoshan D., **"**Data Mining Analysis and Modeling for Marketing Based on Attributes of Customer Relationship", Reports from MSI - Rapporter från MSI, Report 06129 Växjö University ISSN 1650-2647, 2006. pp 1-46 www.msi.vxu.se

Xin Chem and Yi-Fang Wu "Personalized Knowledge Discovery: Mining Novel Association Rules from Text", 2006

Yang Y. and Pedersen J. O. A comparative study on feature selection in text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning, pp 412–420. Morgan Kaufmann Publishers, San Francisco, US, 1997

Yarowsky, D. "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora". In Proceedings of the 14th International Conference on Computational Linguistics (COLING-14), pp. 454–460, 1992.

Yoon, J., Raghavan, V., Chakilam, V., And Kerschberg, L. BitCube: A three-dimensional bitmap indexing for XML documents. *J. Intell. Inform. Syst. 17*, 1, 241–252, 2001.

Zelenko D, Aone C., and Richardella A. Kernel methods for relation extraction. Journal of Machine Learning Research, 3:1083–1106, 2003.

Zhu H. , S. Raghavan, S. Vaithyanathan, J.S. Thathachar., R. Krishnamurthy, P. Deshpande, R. Gupta and Chitrapura K.P. "AVATAR: Using text analytics to bridge the structured–unstructured divide" http://www.almaden.ibm.com/cs/projects/avatar/techrep04.pdf, 2005.

http://www.nimble.com

Callixa, see http://www.callixa.com

Infoshark, see http://www.infoshark.com

http://www.softwareag.com/tamino/

DiscoveryLink,see http://www.ibm.com/solutions/lifesciences/discoverylink.html.

UDB Relational Connect, see http://www.ibm.com/software/data/db2/relconnect/

Standard & Poor's. http://www.standardandpoors.com

Dun and Bradstreet (D&B). http://www.dnb.com

Reuters. http://www.reuters.com.

http://www.statsoft.com/textbook/stdatmin.html #mining.

AutoMed Project. http://www.doc.ic.ac.uk/automed/

Disco-TEX

http://www.proxem.com/Default.aspx?tabid=55

Wikipedia

APPENDICES

A.1 QUESTIONNAIRE

An Empirical Analysis of Mobile Phone users for Customer Relationship Management

Introduction

This questionnaire aims at eliciting information from you in order to measure the analyze mobile phone user for the purpose of Customer Relationship management. Please answer the questions honestly by ticking or writing the answer that best express your view. We would like to assure you of the confidentiality of the information you provide. Thank you.

Statistical information

1. Gender: ( ) Male ( ) Female
2. What is your age range?15-20 ( ) 21-30 ( ) 31-40 ( ) 41-50 ( ) 51-60 ( ) Above 60( )
3. What is your profession? ……
4. What is your highest Academic qualification? ……
5. Are you a Nigerian? YES ( ) NO ( )
6. What State of Nigeria do you live in ……
7. Is your Education IT related? YES ( ) NO ( )

Core Product (CP) Questions

1. Your mobile phone is used for the following:

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Voice calls					

Data (as modem):					
Answering machine:					
SMS:					
WAP (Internet):					
Address book					
Calendar:					
Use GPRS function:					

Other (specify usage apart from above): ……

2. If you use WAP (internet), for what do you use it?

( ) News ( ) Sports ( ) Shopping ( ) Travel service ( ) Local information

( ) Cinema Other: ……

3. What network service do you have on your mobile phone……

Basic Product (BP) Questions

1. How much does a cell phone's name matter to you?

( ) Does not matter at all ( ) Not Much ( ) Very Much

2. How much do you spend on your mobile phone each month?

( ) less than N1,000 ( ) N1,000 - N5,000 ( ) N5,000 - N10,000

( ) N10,000 - N20,000 ( ) N20,000 - N50,000 ( ) Above N50 000

3. What Brand of mobile phone are you using now?

( ) Nokia( ) Motorola ( ) Ericsson ( ) Samsung ( ) LG ( ) Philips ( ) Sagem ( ) Arcatel ( ) Sony ( ) Siemens Nokia Others…….

4. Please write your model of phone ……

5. In what color do you prefer your mobile phone?.....

6. Select the average size that you prefer for your mobile phones.

( ) Very small ( ) Small ( ) Big ( )Very big

Expected Product (EP) Questions

1. Rate the customer service of your Cell Phone Carrier;

( ) Good ( ) Satisfactory ( ) Unsatisfactory ( ) Poor

2. For how long have you been using this particular phone?.....

3. How often do you change phone in a year?.....

4. What is the reason for changing your phone? Phone got spoilt ( )Stolen ( )Gave it out ( )

Misplaced it ( ) Others specify……

5. What is the general assessment of your mobile phone user friendliness?

( ) Excellent ( ) Good ( ) Satisfactory ( ) Unsatisfactory ( ) Poor

6. Rate how easy it is to navigate through the services you have on your mobile phone.

( ) Excellent ( ) Good ( ) Satisfactory ( ) Unsatisfactory ( ) Poor

Augmented Product (AP) Questions

1. What do you **like most** about your mobile phone? ……

……

……

…….

2. What do you dislike most about WAP?.....

……

……

……

3. Do you consider health hazard in the purchase of mobile phones?

( )Yes ( )No ( )No Comment ( )

4. If yes, how do you think you can avoid it when buying your phone?

……

……

…….

Potential Product (PP) Questions

1. Based on why you use your cell phone, do you feel you have a need for 4G higher-speed wireless (i.e. for faster video) rather than 3G wireless? **( )** Yes ( ) No ( ) No comment

2. Share your best mobile phones experience.....

….....

.....

.....

3. How satisfied are you with the overall performance your mobile phone provider?

( ) Excellent ( ) Good ( ) Satisfactory ( ) Unsatisfactory ( ) Poor

4. What phone accessories do you have (you can choose more than one option(s)

( ) Shell ( ) Hand free ( ) MP3 plug-in ( ) Computer ringing tone editor

Others (please specify)……

6. How would you rate the availability of purchasing your mobile phone?

( ) Excellent ( ) Good ( ) Satisfactory ( ) Unsatisfactory ( ) Poor

7. Why did you decide to purchase that particular brand of mobile phone…….

……. … ……..

8. What improvements would you like to see, if any on your mobile phone……

……

……..

9. What type of problem do you usually encounter while using your mobile phone?.....

……

……

10. Are there any other comment you would like to make regarding your mobile phone.

……

……

11. Which of the following is a priority when buying a mobile phone?

( ) Maintenance **( )** Repairs ( ) Warranty ( ) All of the above

A.2 LIST OF PUBLICATIONS FROM THE THESIS

Refereed Conference Proceedings:

- Fatudimu I.T.,Uwadia C.O. and Ayo C.K. " A framework for an Integrated Mining of Heterogenous data in decision support Systems" *in* Proceedings of First International Conference on Mobile-computing, Wireless Communication, E-Health, M-Health and TeleMedicine(FICMWiCOmTelHealth '08) pp 59-66

- Fatudimu I.T.,Uwadia C.O. and Ayo C.K. "Improving Customer Relationship Management through Integrated Mining of Heterogeneous Data" in Proceedings of International Conference on Database and Data Mining, Sanya, China, 2011

Journal Publication

- Fatudimu I.T., Uwadia C.O. and Ayo C.K. (2008) "An Emperical analysis of Mobile Phone data for Competitive Intelligence" i-manager's Journal on Management with special focus on Competitive Intelligence. Vol. 3 l No. 2 l September - November 2008 pp 70-75 www.imanagerpublications.com

- Fatudimu I.T., Uwadia C.O. (Prof.)* and Ayo C.K (PhD)" A Knowledge Mining approach for effective Customer Relationship Management" International Journal of Knowledge Based Organisations, IGI publishers , 2010*(Accepted for Publication)*

- Fatudimu I.T., Uwadia C.O. (Prof.)* and Ayo C.K (PhD) "Improving Customer Relationship Management through Integrated Mining of Heterogeneous Data" Procedia Engineering 00(2011) 000-000, Elsevier, GCSE 2011: 28-30 December 2011, Dubai, UAE (*Accepted for Publication)*