A FRAMEWORK FOR AN INTEGRATED MINING OF HETEROGENOUS DATA IN DECISION SUPPORT SYSTEMS

Fatudimu I.T., Uwadia C.O. (Prof.)* and Ayo C.K (PhD)
ibkfat@yahoo.co.uk, *couwadia@yahoo.com and ckayome@yahoo.com,
Department of Computer and Information Sciences, Covenant University, Ota, Nigeria
*Department of Computer Science, University of Lagos, Lagos Nigeria

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 many organizations. Over the past years, data mining techniques have been used to explore large volume of data (structured) in order to discover knowledge, often in form of a decision support system. For effective decision making, there is need to discover knowledge from both structured and unstructured data for completeness and comprehensiveness.
The aim of this paper is to present a framework to discover this kind of knowledge and to present a report on the work-in-progress on an on going research work. The proposed framework is composed of three basic phases: extraction and integration, data mining and finally the relevance of such a system to the business decision support system. In the first phase, both the structured and unstructured data are combined to form an XML database (combined data warehouse (CDW)). Efficiency is enhanced by clustering of unstructured data (documents) using SOM (Self Organized Maps) clustering algorithm, extracting keyphrases based on training and TF/IDF (Term Frequency/Inverse Document Frequency) by using the KEA (Keyphrases Extraction Algorithm) toolkit. In the second phase, association rule mining technique is applied to discover knowledge from the combined data warehouse. The final phase reflects the changes that such a system will bring about to the marketing decision support system.
The paper also describes a developed system which evaluates the association rules mined from structured data that forms the first phase of the research work.
The proposed system is expected to improve the quality of decisions, and this will be evaluated by using standard metrics for evaluating the interestingness of association rule which is based on statistical independence and correlation analysis.

Keywords: Structured data, unstructured data, Data Mining, Decision Support Systems, Data Integration, Competitive Intelligence, Customer Relationship Management

1.0 INTRODUCTION

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 situation where the amount of information is too large for human decision maker to use optimally and with precision [1].

The first era of information (before the late 1970s) typically involved small amount of primarily textual (semi-structured) static data, stored in flat files. The file handling systems of this era therefore focused on accessing flat-files data. The second Era involved systems that processed constant stream of dynamic, structured data such as airline reservations that was larger than the flat file data of the first era. In other words, this was the era when relational databases took over the job of information management. In the third era of information, data warehousing and business intelligence evolved to gain competitive advantage and insights via data mining. In today's fourth era of information, information processing is gradually moving towards semi-structured or unstructured data management [2].

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 [3]. Mining in this environment is 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 [5]. This type of mining is limited due to the fact that the available information accessible to a company is mostly unstructured [8, 9].

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 [3]. Unstructured data can take formats such as pdfs, 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 [4]. 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 be viewed as an extension of data mining or knowledge discovery from unstructured databases [6]. Unstructured mining 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 is unstructured [2, 7, 8, 9].
- 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. 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 [9].
- Finally some experiments produced a mix of unstructured and structured data.

Integrated mining in this context therefore can be defined as creating one platform for mining structured and unstructured data. The outcome of such integration becomes the fundamental infrastructure for strategic decision support and business intelligence.

This proposed system will be applied to the Customer Relationship Management and Competitive Intelligence (CI) component of Business intelligence (BI) in the manufacturing and production companies. These types of organizations are involved in the creation of goods and services and also making them available to consumers. The following are the types of decisions that are made in business organizations:

- Strategic decisions: These are long term decisions which define the organization's relationship with its relevant internal and external environments, notably in terms its products/services/ideas and market segments.
- Administrative decisions: These are decisions that arise from and subject to, the conflicting demands of strategic and operational organizational problems.
- Operating decisions: These are short-term decisions, which define issues such as output levels, pricing, and inventory levels etc.
- Basic decisions: These are long range in scope, for example, location of factory in a development area, or deciding what product/ service to make.
- Routine decisions: They are routine and repetitive decisions which have procedures set up to deal with them.
- Unprogrammed decisions: Non repetitive business decisions where risks involved is high and cannot be easily accessed in quantitative terms [29].

Integrated mining will contribute in this application area by combining the unique attributes of both structured and unstructured data format in order to provide greater benefits to

the organization, especially minimizing the popular practice of handling structured and unstructured as distinct information entities which often results in decision management failure [7]. For example, products defects and warranty claims results in heavy costs to manufacturers. Companies can therefore build early warning system that, by processing warranty data, helps in the early discovery of products and system failures. Warranty data is generated when a claim form is completed by a customer or a technician. These forms ask for the product code, model number, date, time, customer ID. This information falls into the category of structured data. Usually this form also contains comments section where customer or technician can provide detailed information about the problem. This unstructured data 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.

The rest of the paper is organized as follows. Section 2.0 presents a review of related work, section 3.0 presents the objective, 4.0 describes the methodology and the proposed system architecture is presented in section 5.0. Report on work in progress is presented in section 6.0. Section 7.0 provides conclusion.

2.0 LITERATURE REVIEW

Over the years, systems have been developed in order to achieve the purpose of integrated mining. Frieder et al [10] developed a system 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, 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 2002, Roth et al [11], came up with an integration architecture consisting of three tiers: 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 Williams et al [13], 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 [12, 13]. ESTEST then takes the metadata in the global schema and uses this to suggest input into the information extraction process. GATE [14, 13], 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( ) store. The global query facilities of AutoMed are now available to the user in order to query the global schema [15, 16, 13]. 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 mining stops at information retrieval.

SQUAD (Storing and Querying Unstructured Data). SQUAD [17] 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 data in two steps. Introduce a new type of storage device called Intelligent Storage

Node (ISN) to store, manage and search unstructured data. Using ISNs as a building block, propose 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.

Sukumaran and Sureka[7] 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 [24]. PayGo aims at providing web scale, domain-spanning access to structured sources [25]. 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 missing from such search approaches is the post-processing of heterogenous search results [23].

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 in 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 [26]. Even though Oracle 11g has the facility to manage structured and unstructured data using ontologies, the responsibility of creating ontologies lies on the application developer which makes it not to be tailored directly towards business intelligence. Also, though it handles structured and unstructured data, it is not directly built for storing data towards integrated mining.

3.0 OBJECTIVES OF THE RESEARCH

The primary objectives of this paper are two fold:

1. To propose a framework for an integrated mining of heterogeneous data (a Ph.D research project).
2. To implement association rule mining algorithm and evaluate the interestingness of rules generated , which is the first phase of the project. This evaluation is reported under work-in- progress.

4.0 RESEARCH METHODOLOGY

For the proposed framework, the following tools would be used:

- SQL server 2008 wouldl be used as the database management system to store both the structured and unstructured data (inform of an XML database), this is because it possesses the ability to manipulate XML documents. This has the advantage of creating a single DBMS platform to store both structured and unstructured data. The SQL server 2008 also has embedded in it an ETL to which will be used to format, clean, and modify data.
- An interface which supports loading of both structured and unstructured data would be created using the C# 2008 programming environment.
- In order to have efficiency introduced through mining the most contributing data, the following would be done:
    - Clustering of unstructured data (documents) using SOM clustering algorithm [31]

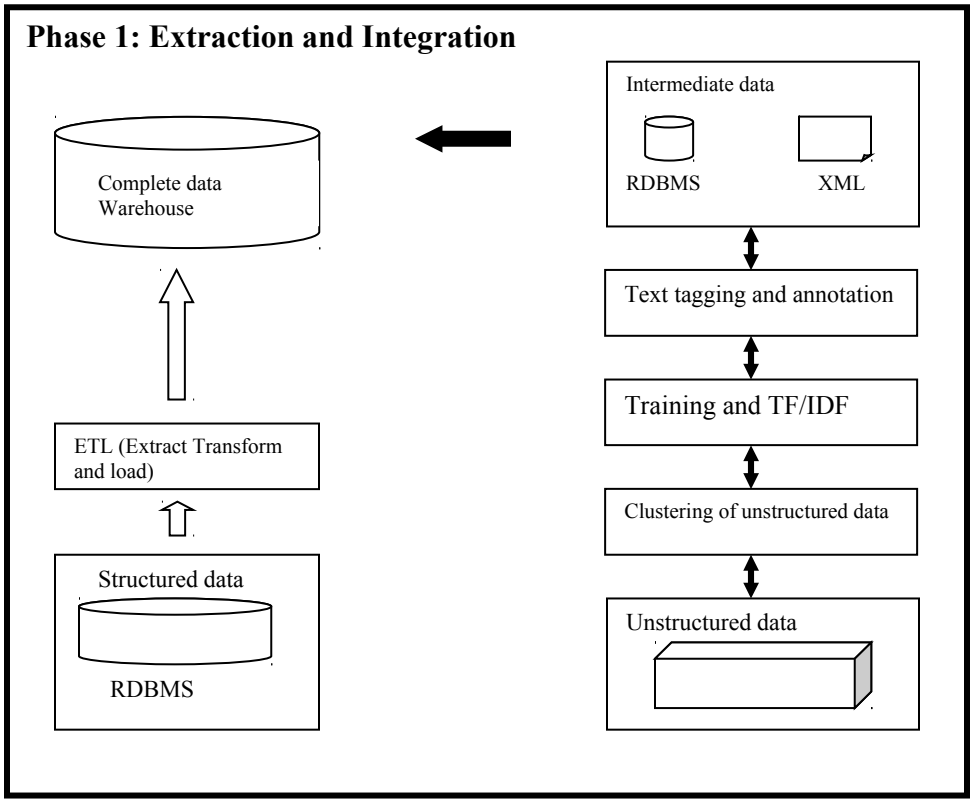
which was chosen because it is mostly used for document clustering [32] and can be used for visualizing high dimensional data on low dimensional views.

- Extracting keyphrases would be done based on training and TF/IDF. The tool to be used is the KEA (keyphrases extraction algorithm) toolkit [34]. The training data could either be structured or unstructured gotten from the problem definition phase of the methodology.

- Both the extracted keyphrases and the structured data will be put in the XML database format (Combined Data Warehouse), which will be mined using Association rule mining algorithm.
- Evaluation/validation of the system: Since the end product of the system is in form of association rules, therefore the interestingness of the rules that evolve in the following implementations would be compared with each other using a standard measure for evaluating interestingness of association rule which is a measure based on statistical independence and correlation analysis [30]. The percentage of positively correlated rule would be compared for each of the following combinations.
  - Text mining without the proposed contribution AND Text mining with the proposed contribution.
  - Text mining with the proposed contribution AND combination of structured and unstructured Mining with the proposed contribution.
  - Structured mining AND combination of structured and unstructured Mining with the proposed contribution.

  The above evaluation process would be carried out based on the same data input.
- Analysis and interpretation: Two types of visualizations, which include two dimensional matrix and directed graph would be created in order to analyze and interpret the association rules.

5.0 THE PROPOSED SYSTEM FRAMEWORK



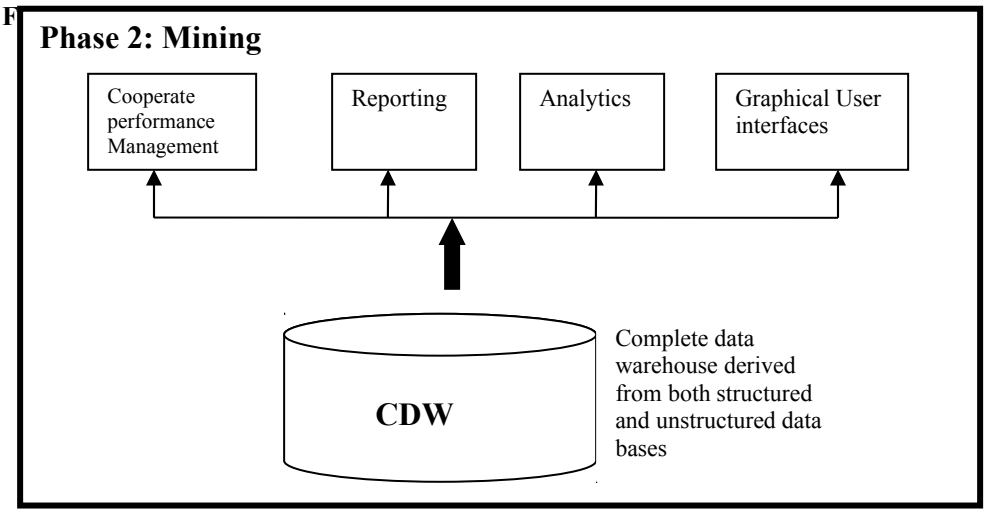


Figure 2: The second phase of the proposed integration architecture.

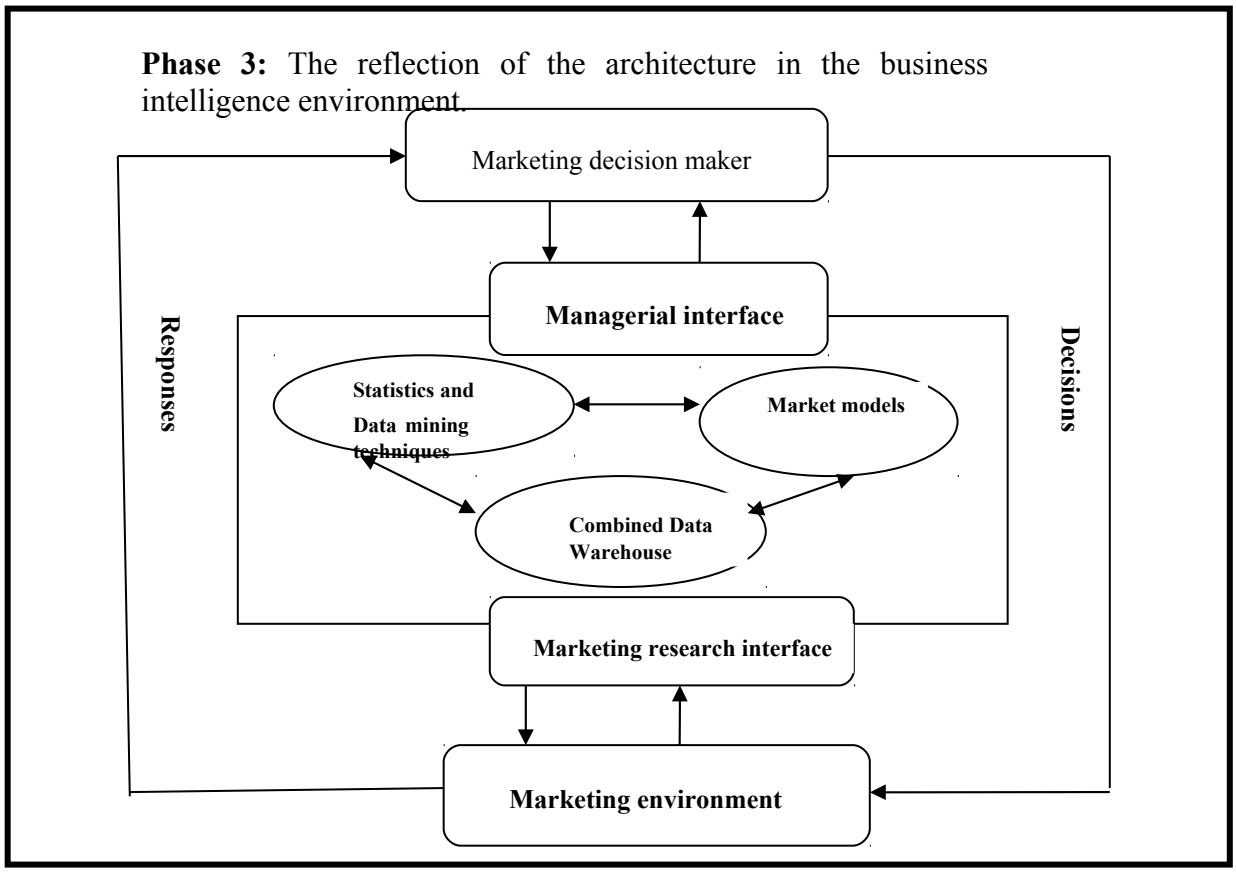


Figure 3: Marking Decision Support System.

Figure 1 reflects the actual process of integrating the structured and unstructured data as explained in the methodology. It also reflects the two important components that makes up the contributions of the proposed system which are (1) Clustering of unstructured data and (2) Training and TFIDF.

data. Figure 3 is the reflection of the proposed architecture in the business intelligence environment. The difference in the above architecture as regards the original marketing decision support system, according to Strauss et al [27] is that, the data mining technique component has been added to the statistical technique component and the raw data component has been replaced by a combined data warehouse which is gotten as a result of phase 1 above.

6.0 REPORT ON WORK IN PROGRESS

Presently, the association rule mining algorithm which is the main data mining technique to be used in order to mine from the combination of structured and unstructured has been implemented using the C# 2008 development platform. Also a module has been included in the developed system which evaluates the resulting rules based on statistical independence and correlation analysis. The algorithm presently runs basically on structured data.

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

$$Support(A,B) = \frac{SupportCount(A,B)}{TotalNumberOfTransactio}$$

(eq.1) (

The confidence of rule a (association rule) A → B can be defined as the proportion of those transactions containing A that also contain B

Figure 2 reveals the stage where mining is carried out on the combined data warehouse, which is made up of structured and unstructured

$$Confidence(A/B) = \frac{Support(AB)}{Support(A)} \quad \text{(eq. 2)}$$

Correlation analysis: The occurrence of itemset A is independent of the occurrence of itemset B if P(AUB) = P(A)P(B); otherwise itemsets A and B are dependent and correlated as events. i.e

$$Corr(AB) = \frac{P(AUB)}{P(A)P(B)} \quad \text{(eq. 3)}$$

If the resulting value of the equation is less than 1, then the occurrence of A is negatively correlated with (or discourages) the occurrence of B. If the resulting value is greater than 1, then A and B are positively correlated, meaning the occurrence of one determines the other. If the resulting value is equal to 1, the A and B are independent and there is no correlation between them.

Interfaces of the developed system:

TID	List of Items_IDs
T100	I1,I2,I5
T200	I1,I4
T300	I2,I3
T400	I1,12,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Table 1: Sample database

Table 1 above is the sample database which was use
to obtain the resulting rules displayed in figure 4.

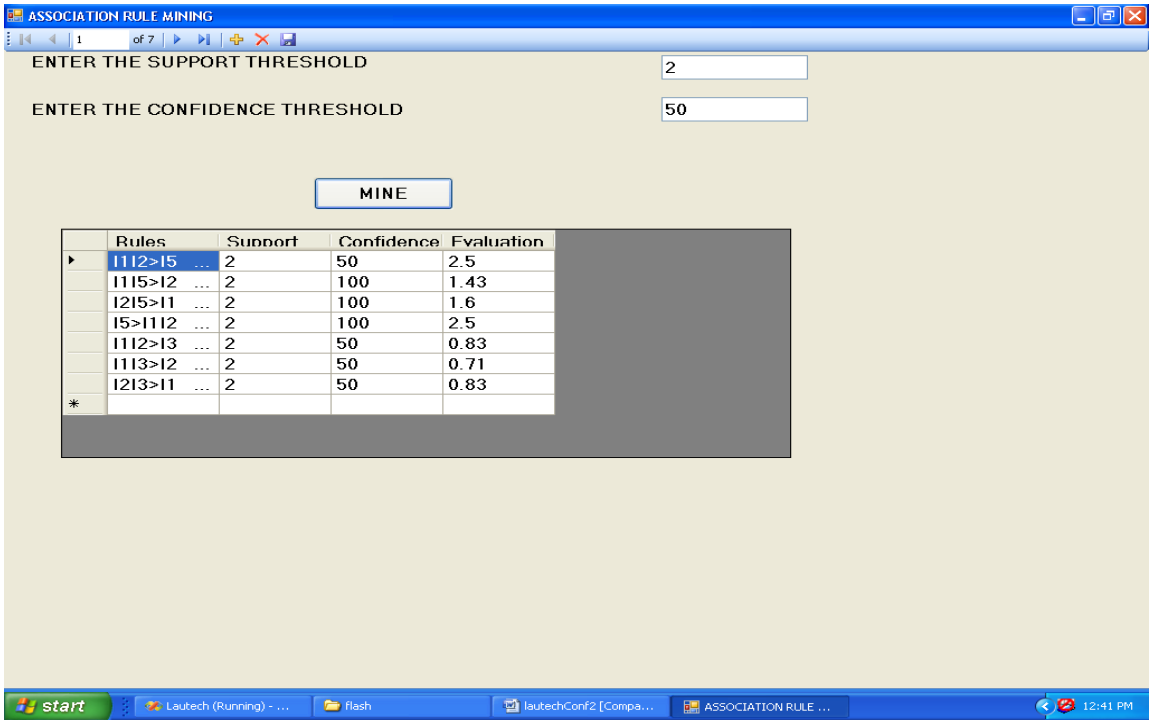


Figure 4: Interface of the Association Rule Mining System

In figure 4, entering a support threshold value of 2 and confidence threshold of 50% for example give the result displayed. The generated rules are displayed under the Rules Column and their respective support and confidence also displayed. Rules that do not meet the value 2(support) and 50% (Confidence) have been eliminated. Also the corresponding evaluation (Correlation analysis) is displayed under the evaluations column. Form this display, rules such as I1I2->I5 , I1I5->I2, I2I5->I1 and I5->I1I2 can be said to be interesting and retained for decision making but I1I2->I3, I1I3->I2 and I2I3>I1 are negatively correlated and therefore not interesting.

7.0 CONCLUSION

This paper 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 proposed system would place at the disposal of decision makers, complete knowledge that would be a platform for more robust, efficient, accurate and effective decision. Furthermore, the implemented module reveals a reduction in the set of rules to be used for decision making to the most interesting subset.

8.0 REFERENCES

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

2. Kernochan, W. (2006) " XQuery and XML data: DB2 helps manage the era of unstructured data," *Infostructure Associates.*

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

4. Ukelson, J. (2006) "Combining Structured, Semistructured and Unstructured Data in Business applications," in *DM Direct Newsletter*.

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

6. Ah-Hwee Tan, (2006) "Text mining: The state of the art and the challenges,".

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

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

9. Blumberg, R., and Atre, S.(2003) "The Problem with Unstructured Data," DM Review(13:4).

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

11. M.A. Roth, D.C. Wolfson, J.C. Kleewein, and C.J Nelin, (2002) "Information Integration: A New Generation of Information Technology," IBM Systems Journal 41, No. 4, 563-577

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

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

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

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

16. E. Jasper. (2002) "Global query processing in the AutoMed heterogeneous database environment." In Proc. BNCOD02, LNCS 2405, pages 46-49.

17. Aravindan Raghuveer, Meera Jindal, Biploh Debnath, David H.C Du, Mohamed Mokbel, (2005) "A unified framework for storing and Querying Unstructured And Structured Data."

18. Hemamalini S (2002). "Customer Relationship Management, An opportunity for Competitive Advantage,".

19. Sigala Marianna, "Customer Relationship Management (CRM) Evaluation: Diffusing CRM Benefits into Business process".

20. Solomon N. and Paul G., (2003) "Business Intelligence," Ninth Americas Conference on Information Systems. pages 3190- 3199.

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

22. Maria MilosavIjevic, Claire Grover' and Louise Corti, (2006) "Smart Qualitative Data (SQUAD): Information Extraction in Large Document Archive."

23. Erhard Rahm, Andreas Thor, David Aumueller, (2007) "Dynamic Fusion of Web Data".

24. Chang K:, He, B: 1. Chang, K:, He, B:, Zhang; Z. (2005) "Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web." Proc. CIDR

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

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

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

28. Graham Hooley, John Saunders, Nigel Piercy, (2004) "Marketing strategy and Competitive Positioning" published by Pearson Education Limited, pp 180-232.

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

30. Jiawei H. and Micheline K., (2001) "Data Mining, Concepts and techniques" published by Morgan Kaufmann , pp 259-262

31. Richard Freeman, Hujun Yin and Nigel M. Allinson, (2002) "Self organizing Maps for Tree View Based Hierachical Document Clustering", in proceedings of the IEEE IJCNN'02, Honolulu, Hawaii, 12-17 May, Vol. 2, pp 1906-1911.

32. Jan Paralic and Peter Bednar, "Text Mining for Documents Annotation and Ontology Support", in Web Technologies Supporting Direct Participation in democratic Processes" IST-1999-20364 Webocracy.

33. Xin Chen and Yi-Fang Wu, "Personalized knowledge Discovery: Mining Novel Association Rules from Text", pp588-592

34. Jones, S. and Paynter, G.W. (2002) "Authomatic extraction of document keyphrases for use in digital libraries: evaluation and applications." Journal of American Society for Information Science and Technology (JASIST). http://www.nzdl.org/kea/.

Biographies of Authors

Fatudimu Ibukun Tolulope holds a B.Sc in Engineering Physics and M.Sc in Computer Science. She is currently a Ph.D student in the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. Her research interest is in the field of Data Mining. She is an Assistant Lecturer in the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. She enjoys reading and engages in creative arts.

Uwadia C.O. holds a B.Sc, M.Sc and Ph.D in Computer Science. His research interest include

Software Engineering. He is the present president of the Nigerian Computer Society (NCS), and Computer Professional Registration Council of Nigeria (CPN). He is currently a Professor of Computer Science in the University of Lagos, Nigeria, Africa.

Charles K. Ayo holds a B.Sc. M.Sc. and Ph.D in Computer Science. His research interests include: mobile computing, Internet programming, e-business and government, and object oriented design and development. He is a member of the Nigerian Computer Society (NCS), and Computer Professional Registration Council of Nigeria (CPN). He is currently an Associate Professor of Computer Science and the Head of Computer and Information Sciences Department of Covenant University, Ota, Ogun state, Nigeria, Africa. Dr. Ayo is a member of a number of international research bodies such as the Centre for Business Information, Organization and Process Management (BIOPoM), University of Westminster. http://www.wmin.ac.uk/wbs/page-744; the Review Committee of the European Conference on E-Government, http://www.academic-conferences.org/eceg/; and the Editorial Board, Journal of Information and communication Technology for Human Development.