Improving Customer Relationship Management through Integrated Mining of Heterogeneous Data

I. T. Fatudimu, C. O. Uwadia, and C. K. Ayo

***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. 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 which is the main focus of this paper.**

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.

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 integration approach is also compared with a similar approach which uses just syntactic relevance in its information extraction process to reveal a significant reduction in the large itemsets and execution time. This leads to reduction in rules generated to more interesting ones due to the semantic clustering of XML documents introduced into the improved integrated mining technique.

***Index Terms*—Association rule mining, customer relationship management, integrated mining, structured data, unstructured data.**

I. Introduction

Integrated mining can be defined as creating one platform for mining structured and unstructured data. The structured environment is made up of data that has fields, columns, tables, rows and indexes, while the unstructured environment has no particular order to it. It consists of text found in medical reports, warranties, contracts, emails and spreadsheets and so on [1]. Customer Relationship Management (CRM) on the other hand can be defined as 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 [2]. Analytical CRM is the branch of CRM that deals with data analysis modeling, campaign management, and long-term decisions on customer development strategies [3]. Presently, there exist a problem in analytic CRM which has to do with having an holistic view to the structured and unstructured CRM data [4], which this paper is focused at solving. A typical scenario in which analytical CRM could benefit from integrated mining includes the following; Products defects and warranty claims result in heavy costs to manufacturers, therefore companies can 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 being diagnosed.

II. Literature Review

Data mining (in the structured data environment) is a powerful technology for recognizing and tracking patterns within data [5]. It has been applied in CRM to solve various problems, which include the following just to mention a few; using data mining to predict from web-based E-Commerce store [5], combining knowledge management and data mining for marketing [6], using data mining to analysis and model for marketing based on attributes of customer relationship [7] and predicting customer loyalty using internal transactional database [8]. Unfortunately, this type of mining is limited due to the fact that 85% of the available information accessible to a company is mostly unstructured [9].

Text Mining is typically defined as a process of extracting useful information from document collections through the identification and exploration of interesting patterns [10]. It has been applied in the field of CRM, for example, it has been used to improve customer complaints management by

Manuscript received May 27, 2012; revised June 28, 2012.

Fatudimu Ibukun Tolulope is with Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. Her research interest is in the field of Data Mining (e-mail: ibkfat@yahoo.co.uk)

Uwadia C. O. is with Computer Science in the University of Lagos, Nigeria, Africa (e-mail: couwadia@yahoo.com).

Charles K. Ayo is with Computer Science and the Head of Computer and Information Sciences Department of Covenant University, Ota, Ogun state, Nigeria, Africa (e-mail: ckayome@yahoo.com).

automatic email classification using linguistic style features as predictors [11]. Also, there exist some commercial text mining tools for customer relationship management such as DiscoTEX (Discovery from Text EXtraction) [12] and TAKMI (Text Analysis and Knowledge MIning) [13]. In addition, there exist tools such as IBM Content Analyzer (ICA) built on UIMA3 as the text processing and mining engine [14] and SAS Text Miner, which extracts and automatically classifies textual documents. The outputs of the SAS Text Miner subsystem allow user to view term statistics, identify similar documents, and view term and concept relationships in a graphic display [15]. The above text mining approaches are still limited due to the fact that there is a need to maximize the richness of mining heterogenous information sources [15].

The following are the few approaches that have attempted to mine from structured and unstructured data; Sukumaran and Sureka [16] 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. For the unstructured data sources, the tagging and annotation platform extracts information based on domain ontology into an XML database. The main component of the system which converts unstructured to semi structured (XML) is based on natural language techniques and is therefore subject to the generational problems of information extraction such as high error rates thereby producing unreliable results.

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 [15]. The following are existing approaches to integrated structured and unstructured data [17-20]. In [21-23], information retrieval related features such as ranking and relevance-oriented search has been proposed to be integrated with XML query languages.

Finally, in [24] a system was proposed to query and analyze seamlessly across structured and unstructured data. It uses TAE (Text Anaysis Engine) to extract annotations from text which is automatically ingested into a structured data store. The major challenge with this approach is data uncertainty, which stems from natural language processing. However, our work is focused on providing a solution to the existing problem in CRM described in section 1 above, by integrating 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.

III. Methodology

There are basically two major phases; the data preprocessing phase and the knowledge distillation phase. In the data preprocessing phase, the structured component of this integration is selected based on the resulting keywords from the information retrieval process.

A. The Data Preprocessing Phase

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

Filtration: A word is selected as a keyword if it does not appear in a pre-defined stop-words list. The stop-words list consists of articles, pronouns, determinants, prepositions and conjunctions, common adverbs and non-informative verbs.

Stemming: After the filtration process the system does word stemming, a process that removes a word's prefixes and suffixes. Stemming is done by unifying word based on their dictionary meaning using the WordNet lexical database. WordNet is referenced through Proxem Antelope [25], 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.

Clustering of XML document: 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 equation (1) below [26], [27].

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

- 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.

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[28].

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

The combination of syntactic and semantic relevance gives the relevance weight of each term i.e. w_j as shown in equation (3) [28].

$$s - rarity(w) = \frac{1}{|o - terms(w)|} \left[\sum_{w_j eo - terms(w)} \ln(\frac{MAX - POLYSEMY + 1}{|sense(w_j)| + 1} \right] \tag{2}$$

$$relevance(w_j, u_i) = \frac{1 + s - rarity(w_j)}{|T(u_i)|} \sum\nolimits_{\tau eT(u_i)} ttf.itf(w_j, u_i / \tau) \tag{3}$$

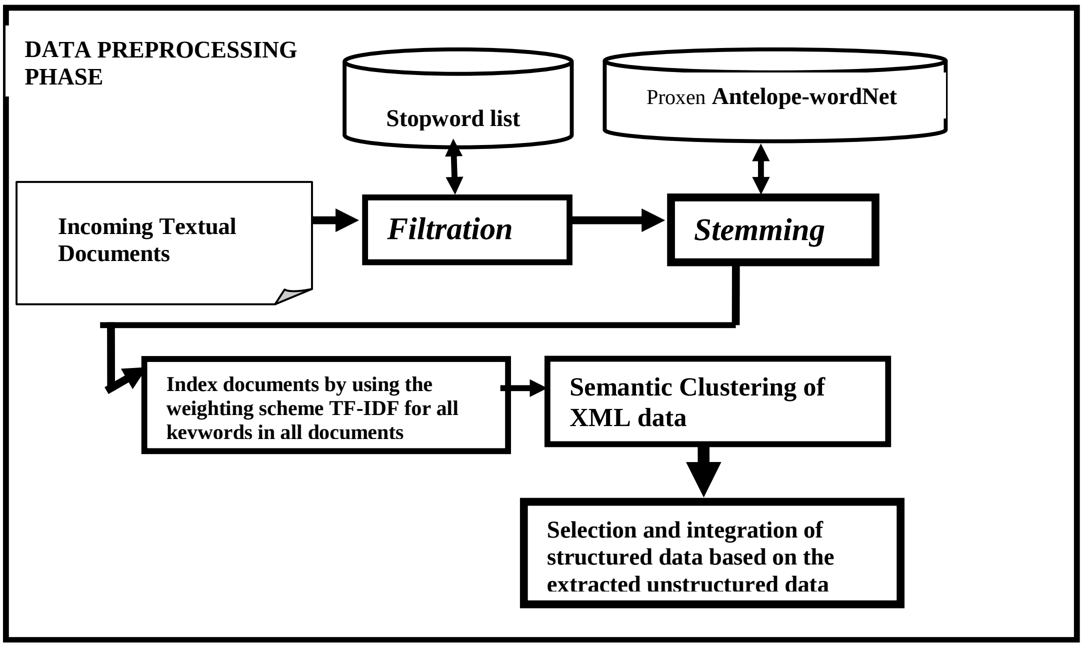
DATA PREPROCESSING PHASE
Stopword list
Proxen Antelope-wordNet
Incoming Textual Documents
Filtration
Stemming
Index documents by using the weighting scheme TF-IDF for all kevwords in all documents
Semantic Clustering of XML data
Selection and integration of structured data based on the extracted unstructured data

Fig. 1. Architechure of the data preprocessing phase.

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*)| -number of terms in T that their stem is *w* i.e. the particular term in question.
| senses(w_j)| -the number of meanings of w_j
relevance(w_j , u_i) - stores the reference value of term w_j in TCU
s-rarity(w_j) - gotten from semantic relevance
$\sum_{\tau\varepsilon T(u_i)} ttf.itf(w_j, u_i/\tau)$ -the TF-IDF weight
$T(u_i)$ - total number of TCUs or transactions.

The content similarity of 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.

$$sim(u_i, u_j) = \frac{u_i \bullet u_j}{\|u_i\| \times \|u_j\|} \quad (4)$$

A. Knowledge Distillation Phase

In the knowledge mining phase, association rules are generated based on the (GARW) Algorithm [26] which has been modified to accommodate content similarity of XML document based on the combination of syntactic and semantic relevance.

IV. Implementation

The system was implemented using *C#* programming language and Visual Studio.Net 2005 as the programming environment. It loads both the structured and unstructured data through the SQL server, receives two thresholds from the user, runs the program and displays the generated association rules

A. Data Description

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 following are the samples of questions asked in order to gather data:

- What is the brand of your mobile phone?(structured)
- What is you gender? (structured)
- What is your age range? (structured)
- Is you mobile phone user friendly? (structured)
- Why do you change your phone? (structured)
- What do you like most about your mobile phone? (unstructured)
- Share your best mobile phones experience. (unstructured)
- Why did you decide to purchase that particular brand of mobile phone? (unstructured)
- What improvements would you like to see, if any on your mobile phone. (unstructured)
- What type of problem do you usually encounter while using your mobile phone? (unstructured).

B. Observations and Argumentation of the Thresholds

It was observed that the nature of the questionnaires retrieved from the respondents was such that almost 40% of the unstructured part of the questionnaires were not filled by them. The negative effect of this on the rules generated was reduced greatly by clustering of the XML data gotten from the unstructured data. In order to have a fair representation of structured and unstructured data, a low threshold support of 20% was chosen and a higher threshold confidence value of 70% to make sure that the final rules gotten from the system are the most interesting ones.

Evaluation of the system: Another 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 our system 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.

To measure the performance of the our system, 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 perform on the same corpus.

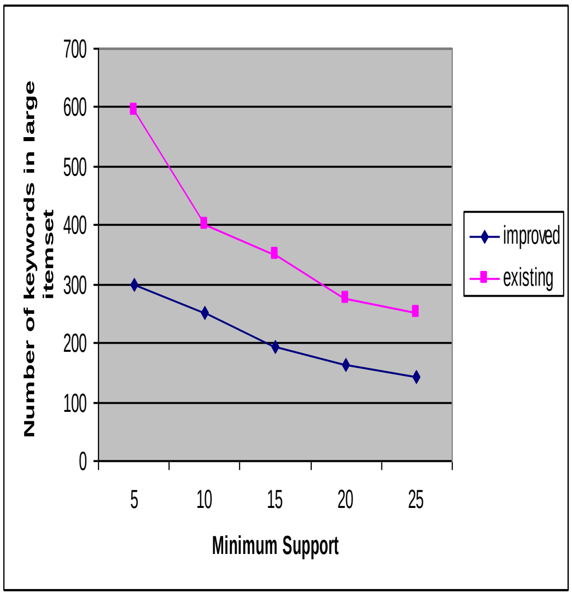

Fig. 2. Improved integrated mining system Vs existing system.

The experimental results displayed in the graph 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 Fig. 3 below.

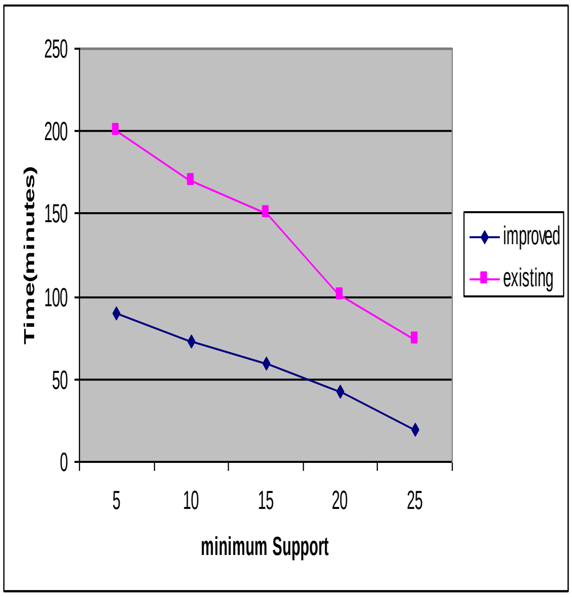

Fig. 3. Graph of execution time against support.

It can be seen that our system always outperforms the Existing system for all values of minimum support.

V. Conclusion and Future Work

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 extracted association rules 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, the results gotten from the experiments is traceable to the fact that, since the large itemset is responsible for the keyword combination stage (which accounts for most of the execution time) of the association rule mining algorithm, therefore the smaller the size of the large itemset, the faster the algorithm execution time and the more semantically relevant the keywords in the large itemset, the more interesting the rules will be.

Work is still on going so as to extend this system by evaluating the resulting rule gotten from the association rule mining phase by using an evaluation technique which reveals interestingness by evaluating the novelty of discovered knowledge.

References

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

[2] J. Strauss, A. El-Ansary, and R. Frost, "E- marketing," International Edition," *Published by Pearson Prentice Hall*, vol. 30, pp. 135-168, 2006

[3] T. Breur, "Integration of Analytical CRM in Business Processes: an Application," 2000.

[4] W. F. Cody, J. T. Kreulen, V. Krishna, and W.S. Spangler, "The integration of business intelligence and knowledge management," *IBM Systems Journal*, vol. 41, no. 4, 2000, pp. 697-713.

[5] Z. Jidi, S. T. Huizhang, L. Duo, and D. Lei, "Predictive Data Mining on Web-Based E-Commerce Store," *IACIS*, 2002, pp. 687-692

[6] J. S. Michael, S. Chandrasekar, W. T. Gek, and E. W. Michael, "Knowledge management and data mining for marketing," *Decision Support Systems*, vol. 31, issue 1, pp. 127–137, 2001,

[7] X. Du, "Data Mining Analysis and Modeling for Marketing Based on Attributes of Customer Relationship," *Reports from MSI*, School of Mathematics and Systems Engineering, Vaxjo University, Report 06129, 2006.

[8] B. Wouter, V. Geert, and V. P. Dirk, "Predicting Customer Loyalty Using the Internal Transactional Database," Working Paper, Hoveniersberg 24 B-9000 GENT, 2005.

[9] R. Blumberg and S. Atre, "The Problem with Unstructured Data," *DM Review* vol. 13, no. 4, 2003.

[10] R. Feldman and J. Sanger, "The text mining handbook: Advanced approaches in analyzing unstructured data." Cambridge: Cambridge University Press, 2007.

[11] K. Coussement and V. P. Dirk, "Improving Customer Complaint Management by Automatic Email Classification Using Linguistic Style Features as Predictors" Working Paper, Hoveniersberg 24 B-9000 GENT www.crm.UGent.be, 2007.

[12] U. Y. Nahm and J. M. Raymond, "Using Information Extraction to Aid the Discovery of Prediction Rules from Text," in *Proceedings of the KDD (Knowledge Discovery in Databases)*, 2000 Workshop on Text Mining, pp. 51-58, Boston, MA, August 2000

[13] T. Nasukawa and T. Nagano, "Text analysis and knowledge mining system," *IBM Systems Journal*, vol. 40, no. 4, 2001, pp. 967-984

[14] B. Indrajit, G. Shantanu, and G. Ajay, "Enabling Analysts in Managed Services for CRM Analytics," *KDD'09*, June 28–July 1, 2009, Paris, France.

[15] S. E. Arnold, "Beyond Search-and-Retrieval: Enterprise Text Mining with SAS®," [Online]. Available:

http://www.sas.com/technologies/analytics/datamining/textminer,http:/www.cxoamerica.com/pastissue/printarticle.asp?art=25408, 2010

[16] S. Sukumaran and A. Sureka, “Integrating Structured and Unstructured Data Using Text Tagging and Annotation,” *Business Intelligence Best Practises SM*. Bi-bestpractices. [Online]. Available: http://www.bi-bestpractices.com/view-articles/4735, 2007.

[17] D. A. Grossman, O. Frieder, D. O. Holmes, and D. C. Roberts, “Integrating structured data and text: A relational approach,” *Journal of the American Society for Information Sciences*, vol. 48, no. 2, pp. 122– 132, 1997.

[18] W. B. Croft, 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*, pp. 223–232, June 1992

[19] S. Agrawal, S. Chaudhuri, and G. Das, “DBXplorer: A system for keyword-based search over relational databases,” in *Proc. of ICDE*, 2002.

[20] V. Hristidis and Y. Papakonstantinou, “Discover: keyword search in relational databases,” in *VLDB*, 2002.

[21] A. Theobald and G. Weikum, “The index-based XXL search engine for querying XML data with relevance ranking,” in *EDBT*, 2002.

[22] S. Al-Khalifa, C. Yu, and H. V. Jagadish. “Querying structured text in an xml database,” in *SIGMOD*, 2003.

[23] N. Fuhr and K. Grobjohann, “XIRQL: A language for information retrieval in XML documents,” in *Proc. of SIGIR*, 2001.

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

[25] Proxem. [Online]. Available: http://www.proxem.com/

[26] M. Hany, R. Dietmar, I. Nabil, and T. Fawzy, “A Text Mining Technique Using Association Rules Extraction,” *International Journal of Computational Intelligence*, vol. 4, no. 1, pp. 1304-2386, 2007.

[27] F. Teng-Kai and C. Chia-Hui, “Exploring Evolutionary Technical Trends From Academic Research Papers,” *Journal Of Information Science And Engineering* vol. 26, pp. 97-117, 2010.

[28] T. Andrea and G. Sergio, “Semantic Clustering of XML documents,” *ACM Trans. Inform. Syst.* vol. 28, no. 1, pp. 3:1-3:55, 2010.

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 nclude 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.