MIS Review Vol. 20, No. 1, September (2014), pp. 31-43
DOI: 10.6131/MISR.2014.2001.02

Data Management Issues and Data Mining of Real Time System Application for Environment Monitoring

Dinesh Kumar Saini[1,2], Sanad Al Maskari[3]
[1]*Faculty of Computing and Information Technology, Sohar University, Oman*
[2]*Research Fellow and Adjunct Faculty, School of ITEE, University of Queensland, Australia*
[3]*School of ITEE, University of Queensland, Australia*

ABSTRACT: *Environment pollution monitoring and control is very big problem for the whole world. Taking decision in the environment is becoming more challenging. The aim of this paper is to present the challenges surrounding environmental data sets and to address these in order to develop solutions. Environmental data sets present a number of data management challenges including data collection, integration, quality and data mining. Environment data sets are also very dynamic and this presents additional challenges ranging from data gathering to data integration, particularly as these data sets are normally very large and expanding continuously. Statistical methods are very effective and economical way to analyze small, static data sets but they are not applicable for dynamic, real-time and large data sets. The use of data mining methods to discover hidden knowledge in large datasets therefore presents great potential to improve environmental management decisions. A representative environmental data set from quantitative air quality monitoring instruments has been assessed and will be used to demonstrate some of the issues in applying data mining approaches to poor data quality.*

KEYWORDS: *Data Management, Real Time Systems, Data Mining, Environment Monitoring Systems.*

1. Introduction

Huge amount of data are generated by environmental sensors every day across the globe. There is tremendous need for data analysis systems which are able to mine massive and continuous stream of real world data applications such as temperature monitoring, air pollution, stock market, network security, etc. Data generated by environmental sensors are recorded at time intervals of seconds through to minutes and over time these sensors will create datasets that need to be mined in real time in a way that takes into considerations the dynamic nature s of the real world changes that are being measured. Without appropriate analysis methods that allow inferences to be derived based on patterns observed within these data sets it will not be possible to lead to new knowledge discoveries (Fayyad, Piatetsky-Shapiro & Smyth, 2006) defines knowledge Discovery in

Databases as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." In this research we are attempting to identify best possible data mining solutions for environmental data using various algorithms, focusing on streaming data mining methods as these types of data are increasingly common now as a result of rapidly emerging technology developments. In order to do this we have developed an environmental data mining model (EDMM) which covers all critical stages needed to produce high quality data mining results. A key tenant of the model is that data analysis methods must take into consideration the dynamic behaviors of the incoming data streams. This means that statistical methods are not always appropriate for the large, dynamic data sets characterizing many environmental variables. The dynamic nature of such data means that their behaviors change over time and can lead to concept drift. The occurrence of concept drift (Helmbold & Long, 1994) will hinder the accuracy of the original data model thus make it increasingly inaccurate over time. In our model we will take into consideration the effects of concept drift to reflect provide greater data model accuracy.

Air pollution, marine pollution, and sand storms are three major problems that affect agriculture and fishing, which need to be addressed and treated proactively (Al-Maskari, Saini & Omar, 2010). Air pollution is a very critical problem that can lead to catastrophic outcomes if not managed and monitored proactively.

2. Research problem

We took as case of Oman at Sohar Port for the current study of our research. The new born industrial zone introduced a mixture of chemical and petrochemical plants that leads to odorous emissions. Such emissions in a populated area cause nuisance discomfort to citizens, and health issues. The government has installed a network of sensors mobile and static to monitor odorous emissions in the industrial zone. The electronic sensors are provided by Common Invent, Delft, The Netherlands All sensors communicate by wireless or telephone links with a central database management system maintained by Common Invent which handles over 2 million new data entries per day. The data base management system reads and stores the incoming raw data, as well as providing some interpretation of the data and coordinates automated event handling for clients (see http://www.comon-invent.com).

Electronic nose (or e-nose) is the name given to a wide class of instruments capable of measuring odor information in different environments (Saini & Yousif, 2013). The data from the e-noses provided by Comon Invent has been used to distinguish odorous petrochemical vapors, like fuel oil, naphtha, gasoline, jet fuel, from manure, from sewage gas or from VOCs like toluene, benzene, styrene etc. (Bootsma et al., 2012). In Sohar Port

the objective is to develop the e-nose data assessment system to be trained to associate digital fingerprints with various odors. Once trained, the e-nose database system can then recognize smells as they arise and inform the user as to their origin once this has been established. The current e-noses deployed at Sohar Port create a record every 3 minutes and stores the raw data in a remote server for the eight different sensors located in each unit. The system also records wind direction, speed, and air temperature, as well as GPS location and time. Currently a total of 7 static sensors and one mobile sensor are installed in the industrial zone with another two sensors installed in an adjacent urban community. Consequently data collected from these distributed sensors forms a large data source, which needs to be cleaned, mined and analyzed in different dimensions to create a usable prediction models and for consumption by different applications.

Data gathered from different environmental sensors needs to be analyzed and assessed appropriately. Any system that handles critical data that can reduce air quality impacts on the adjacent communities has to be credible and effective. Credibility demands that the system is accurate when dispatching a poor air quality alert.. The system will need to monitor and predict odors emissions in a distributed area taking into accounts external elements such as wind, direction and speed, rainfall and dust storms. The system has to be smart enough to issue an alarm based on area of effect and the likely degree of impact. In developing such a system the following are the main challenges faced in this research:

(1) Data integration and data quality issues need to be addressed before applying data mining techniques.

(2) Multiple data sources from different stake holders that need to be integrated.

(3) Odor sensors data sets have limited finger prints which make it hard to create an accurate prediction model.

(4) Current data mining methods used in environmental analysis do not take into considerations the surrounding environment changes and their dynamic behaviors.

(5) Algorithm evaluation in the scope of environmental monitoring is complicated by the complex nature of environment systems.

(6) Using statistical methods to analyze dynamic, real-time and large data sets are not appropriate (Arasu et al., 2003).

3. Research approach

The combination of emerging new semantic web technologies and web services to access, process and integrate data and models held within both centralized and distributed hydrological databases allows:

(1) Identification and prioritization of the key stakeholder user requirements, queries and datasets.

(2) Data quality and data cleansing processes.

(3) Development of the common data models and ontologies to integrate both static and real-time data streams, visual, spatial and temporal data, legacy databases and newly generated datasets.

(4) Prediction of the Air pollution: Using the captured data we intend to create models to identify the air pollution sources, delineate affected areas and estimate next areas to be polluted. A learning system will be employed to enable a better sensing of odors. Beside the data collected from sensors, feedback from experts, researchers, and community will be assessed to see if these improve the predictions. A community enabled sensor network will be considered to improve the odor prediction model.

(5) The data collected by the e-noses will be analyzed using various data mining and machine learning techniques. A prediction model for the air pollution effects will be the core focus of research. In the prediction model we will take into consideration the metrological variables (temperature, wind speed, wind direction, and rain). Clustering, Euclidean, Cosine similarities, ANN, CVFDT, Fuzzy logic and other techniques will be used and compared to achieve best prediction model.

(6) The output of the prediction model will be evaluated against criteria developed for an industry and community alert system.

4. Motivation

Information technology has to be proven to be cost effective and provide a reliable solution if it is to be deployed effectively in environmental management prevention and control (Abdelzaher et al., 2010). The deployment of distributed network sensor has grown rapidly in the past ten years (Hill et al., 2000). Many applications have been implemented using network sensors such as environment monitoring, intrusion detection (Wood & Stankovic, 2000), habitat monitoring (Crumiere, 1999), defense, transportation, and scientific exploration. These sensors produce very large volumes of data that need to be cleaned, stored, and retrieved for analysis and decision making (Li et al., 2000). The huge amount of data produced by these sensors need to be monitored and controlled to provide a meaningful and useful information and it can be used as a feedback loop to management and control systems.

Environment monitoring applications can be very complex because they involve many variables with different dimensions and different scales. Managing, accessing

and analyzing data generated from distributed environmental sensors remains a serious challenge for researchers and scientists. The complex data set created by these sensors present many challenges including visualization, data storage and retrieval, data quality and data integrity as well as data analysis and mining.

The ability of sensor network to deliver large amount of data in real-time create a data mining challenge. In our research we are attempting to cover these challenges and provide solutions to such issues. Most data mining algorithm doesn't take into considerations the dynamic behaviors of multi dimensions data generated by sensor networks especially for environmental data. Concept drift must be taken into considerations when analyzing environmental data in real time.

5. Environmental network sensors

The new developments in the area of electronics and wireless network have led to the creation of Environmental Network Sensors (ENS). ENS are large distributed systems communicate through wireless, telephone or satellite networks to stream the data to a central location for processing and analysis (Sohn et al., 2003). A fundamental aspect of environmental network sensors deployment is therefore data management and analysis of these data streams. To enable environmental organizations and authorities to be able to make constructive decisions regarding likely environmental impacts environmental informatics need to be in well shape and in place.

6. Why data mining

Traditional methods for analyzing data using statistical methods are limited to very small data sets and to single users dealing with them directly. Although statistical methods are very economical, simple and effective but they are complicated when trying to apply them to new applications (Friedman, n.d.). Unlike data mining methods, they are not meant to deal with huge, dynamic and real time data sets. Most statisticians will consider 1,000 point as a large data set but in data mining world millions of transactions can be analyzed using data mining algorithms (Roppel et al., 1998).

The evolution of data mining techniques started with the advent of the first computer devices within the university environment and more broadly with the subsequent development of personal computing devices (Pan & Yang, 2007). The improvements of data communication, networks, databases and the ability for users to ubiquitously access data and services in real time has revolutionized the data mining field. Each new technology and development of numerical techniques was based on the previous one. The following shows how we have developed through the data mining era:

(1) 1970s era: Data collection
In this era data was collected using computers, tapes and discs.

(2) 1980s & 1990s: Data Access
The introduction of Relational DBMS, Relational data was implemented, Data RDBMS, advanced data models (extended-relational, object oriented [OO], deductive, etc.) Application-oriented DBMS (spatial, scientific, engineering, etc.)

(3) 2000s: Data mining era
The introduction of the WWW and the maturity of the internet have created massive databases. Data mining algorithms started to emerge and advanced data mining algorithms has been introduced. New applications in multimedia, web mining, streaming data management and mining were deployed and still apply (Crumiere, 1999).

Data mining is mature enough to be applied in environmental applications and other large data applications due to:

(1) Massive data collected for instance large distributed sensors are used to collect large and complex data about the nature and pollution.

(2) We can now access powerful multiprocessors and data storage technologies at reasonable prices.

(3) Many data mining algorithms have been developed.

(4) Data mining is concerned with creating knowledge and information from dynamic and huge data sets. It is a blended field of statistics, machine learning and data bases. Data mining also referred to as Knowledge Discovery in Database (KDD). The definition of data mining varies between different authors based on their own background, experience and views. For example:

- Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.
- Data mining is the process of extracting previously unknown, comprehensible and actionable information from large database and using it to make crucial business decisions.
- Data mining is finding interesting structure (patterns, statistical models, relationships) in databases.
- Data mining is a decision support process where we look in large databases for unknown and unexpected patterns of information.

The use of data mining techniques in e-nose odor monitoring is still in an early stage. Few modern data mining techniques have been used in e-nose odor monitoring and analysis. Artificial Neural Networks have been applied to predict Sulphur dioxide concentration in Delhi (Green, Chan & Goubran, 2009), to process the signal from odor sensor arrays for near-real-time odor identification (Kahn, Katz & Pister, 1999), using electronic nose to identify the age of spoiled food based to predict piggery odor concentrations (Foster, 2002). In his paper (Aberer et al., 2010) describes the vision of community based sensing using a mobile geo-sensor network (Gehrke & Madden, 2004). The OpenSense project is aiming to investigate air pollution monitoring using community-driven sensing (European Commission, 2001).

7. Environmental data mining model

In this section we will describe our Environmental Data Mining Model (EDM) that we will use to create a prediction model using various data mining algorithms. The EDM has eight essential stages starting from data collection and ending up with decision making and result monitoring (refer to Figure 1). The use of this model will optimize the efficiency and accuracy of our data mining models.

7.1 Data collection

Data collection can be done manually or automatically using sensors. The Sohar Environment Unit (SEU) uses Mobile Air Quality Monitoring Station (MAQMS) which

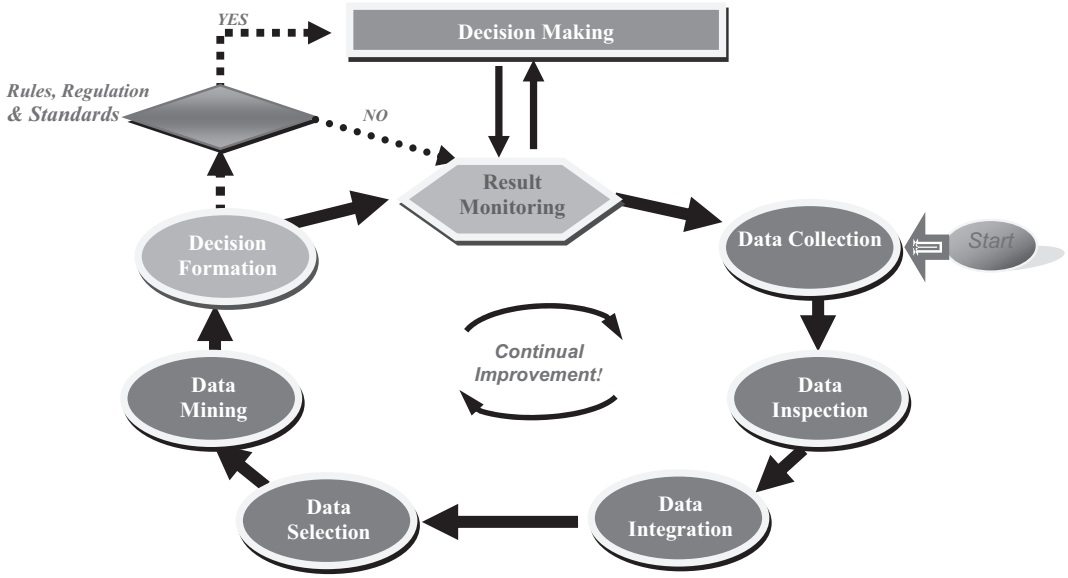

Figure 1 EDDM Model

record data hourly. MAQMS record PM10, O3, CO, SO2, H2S, NO, NOx concentrations with the corresponding meteorological data. The new distributed network sensors (e-Nose) records data every 3 minutes. The E-Nose data are designed to monitor odors around Sohar Industrial Port (SIP). The reading from the E-Nose sensors doesn't refer to any gas instead it writes some values based on the sensor reaction (refer to Table 1).

7.2 Data inspection

Current data collected by SEU are not hosted in centralized warehouses rather it's saved in Excel spreadsheets. This present an issue with data cleansing and integration between distributed sensors. Raw environmental data i is usually not clean, incomplete (empty values), noisy (contains error) and may have inconsistencies (Al-Maskari et al., 2010). The following table illustrates the multi source problems identified in SEU environmental datasets:

The employees of the system have all the rights to expect that the data they are dealing with are indeed correct. Otherwise wrong decisions will be made and the consequences of such decisions can be significant to them and to the environment

Table 1 Data Quality and Cleaning Issues

	Problem	Dirty Data	Remarks
Attribute	Missing Values	O3 = "No Data"	Value unavailable during data gathering (null value)
	Different values with same meaning	O3 = Zero, O3 = 0	Both refer to the same value
Record	Violated data type	O3 = Zero, O3 = Span	The value should be numeric
Record Type	Duplicate records	Sensor 1 (Date = 1/7/08, time = 11:00, dust = 0.025) Sensor 1 (Date = 1/7/08, time = 11:00, dust = 0.025)	Duplicate records from different data sources
	uniqueness violation	Sensor 1 (Date = 1/7/08, time = 12:00, dust = 0.016) Sensor 1 (Date = 1/7/08, time = 12:00, dust = 0.025)	The primary key used in both sensors are date and time
	Different time format	Sensor 1 (Date = 29/02/08, time = 0:00) Sensor 2 (Date = 02/29/08, time = 24:00)	Different date and time format is used by different sensors
	Different unit formats	Sensor 1 Dust unit ug/m^3 Sensor 2 Dust unit mg/m^3	Different units are used by different sensors
	Noisy	e-nose S1 = 0, -0.12	Errors and outliers

Poor data quality management of environmental data can lead to the following:

(1) Increase cost as more time will be spent correcting errors rather than performing critical operations.

(2) Poor data quality may lead to poor decision making which lead to incorrect estimate and predictions. E-environment is so sensitive that we should not tolerate any compromise when it comes to decision making or we will endanger or people and planet earth.

(3) More difficult to set strategy and execute it. Environmental strategic decision requires data gathering from various data sources with some uncertain quality. This makes it harder to develop a sound strategic decision. Executing the strategy becomes difficult as inaccurate results become evident.

To overcome the above concerns the following operations will be conducted to the data sets:

(1) Fill in missing values.

(2) Identify outliers and smooth out noisy data.

(3) Correct inconsistent data.

(4) Duplicate identification.

7.3 Data integration

Environmental data is heterogonous by nature therefore combining data from different sources can be a very challenging task. Legacy data, heterogonous data, time synchronization can be a source of problems when integrating data from different sources. Before integrating multiple data sources they must pass data quality checks, otherwise data quality problems will be inherited from the source databases. By looking at Sohar industrial region environmental data we can observe the multi-source cleansing problem (refer to Tables 1 and 2).

Table 2 Data Collection for Various Gases on the Same Day at Different Time

Date	Time	Dust	O3	CO	nCH4	SO2
		mg/m^3	ppb	ppm	ppm	PPB
01/07/2008	10:00	0.37	S <	S <	RS232	3.608
01/07/2008	11:00	0.025	36.75	0.86	RS232	1.249
01/08/2008	0:00	No Data	15.82	0	RS232	0.125

7.4 Data selection

Data inspection and integration is considered to be the most difficult and longest stages in EDMM. After creating integrated clean data sets it's necessary to define application domain for the data mining algorithm. Meta data information, prior knowledge and application goals must be defined. Once they are defined a target data set will be generated. One of our objectives is to predict odors in the SIP area. We will use E-Nose data combined with social network feedback dataset.

7.5 Data mining

In this stage an appropriate data mining approach will be selected. Various data mining approaches exist including association rule mining, clustering, induction and streaming data mining. Once the appropriate approach is selected then a suitable implementation will be applied. Data mining can focus on a variety of areas.

Typical areas to examine are:

(1) habits and behavior,

(2) demographics,

(3) time,

(4) product characteristics.

8. Conclusion

In this paper we introduced our Environmental Data Mining Model which addressed all aspects surrounding environmental datasets. The EDM introduced eight essential stages necessary to create an accurate data mining results. EDM is a continuous improvement model aiming to improve the process of environmental data mining. EDM will help in the improvement of decision making and environment pollution control.

Acknowledgements

The authors of the paper are thankful for the University Research Council and Sohar University, Oman. Authors are thankful for the University of Queensland, school of ITEE. Thanks to Prof Lance Bode and Prof Tony for the Valuable Input.

References

Abdelzaher, T., Blum, B., Cao, Q., Evans, D., George, J., George, S., et al. (2010), 'EnviroTrack: towards an environmental computing paradigm for distributed sensor networks', *Proceedings of the 24th International Conference on Distributed Computing Systems*, Tokyo, Japan, pp. 582-589.

Aberer, K., Sathe, S., Chakraborty, D., Martinoli, A., Barrenetxea, G., Faltings, B., et al. (2010), 'OpenSense: open community driven sensing of environment', *Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming*, San Jose, CA, pp. 39-42.

Al-Maskari, S.S., Saini, D.K. and Omar, W.M. (2010), 'Cyber infrastructure and data quality for environmental pollution control in Oman', *Proceeding of International Conference on Data Analysis*, *Data Quality & Metadata Management*, Mandarin Orchard, Singapore, doi: 10.5176/978-981-08-6308-1_D-038.

Arasu, A., Babcock, B., Babu, S., Datar, M., Ito, K., Nishizawa, I., et al. (2003), 'STREAM: the Stanford stream data manager', *IEEE Data Engineering Bulletin*, Vol. 26, No. 1, pp. 19-26.

Bootsma, R.J., Marteniuk, R.G., Mackenzie, C.L. and Zaal, F.T.J.M. (1994), 'The speed-accuracy trade-off in manual prehension: effects of movement amplitude, object size and object width on kinematic characteristics', *Experimental Brain Research*, Vol. 98, No. 3, pp. 535-541.

Crumiere, M. (1999), 'Artificial neural network prediction of ground-level ozone concentration in Palm Beach City', Unpublished master thesis, Folrida Atlantic University, Boca Raton, FL.

European Commission. (2001), *IST 2001: Technologies Serving People*, European Commission, Rue de la Loi, Belgium.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (2006), 'From data mining to knowledge discovery: an overview', in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park, CA, pp. 1-34.

Foster, I. (2002), 'The grid: a new infrastructure for 21st century science', *Physics Today*, Vol. 55, No. 2, pp. 42-47.

Friedman, J.H. (n.d.), 'Data mining and statistics: what's the connection?', available at http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf (accessed 21 November 2014).

Gehrke, J. and Madden, S. (2004), 'Query processing in sensor networks', *IEEE Pervasive Computing*, Vol. 3, No. 11, pp. 46-55.

Green, G.C., Chan, A.D.C. and Goubran, R.A. (2009), 'Identification of food spoilage in the smart home based on neural and fuzzy processing of odour sensor responses', *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN, pp. 2625-2628.

Helmbold, D.P. and Long, P.M. (1994), 'Tracking drifting concepts by minimizing disagreements', *Machine Learning*, Vol. 14, pp. 27-45.

Hill, J., Szewczyk, R., Woo, A., Hollar, S., Culler, D. and Pister, K. (2000), 'System architecture directions for network sensors', *ASPLOS*, Vol. 35, No. 11, pp. 93-104.

Kahn, J.M., Katz, R.H. and Pister, K.S.J. (1999), 'Next century challenges: mobile networking for "smart dust"', *Proceedings of ACM/IEEE International Conference on Mobile Computing and Networking*, Seattle, WA, pp. 271-278.

Li, Y., Callahan, T., Darnell, E., Harr, R., Kurkure, U. and Stockwood, J. (2000), 'Hardware-software co-design of embedded reconfigurable architectures', *Proceedings of the Design Automation Conference*, Los Angeles, CA, pp. 507-512.

Pan, L. and Yang, S.Y. (2007), 'A new intelligent electronic nose system for measuring and analyzing livestock and poultry farm odours', *Environment Monitoring and Assessment*, Vol. 135, pp. 399-408.

Roppel, T.A., Padgetta, M.L., Waldemark, J. and Wilson, D. (1998), 'Feature-level signal processing for near-real-time odor identification', in Dubey, A.C. and Harvey, J.F. (Eds.), *The SPIE Conference on Detection and Remediation Technologies for Mines and Minelike Targets III*, Orlando, FL, pp. 13-17.

Saini, D.K. and Yousif, J.H. (2013), 'Environmental scrutinizing system based on soft computing technique', *International Journal of Computer Applications*, Vol. 62, No. 13, pp. 45-50.

Sohn, J.H., Smith, R., Yoong, E., Leis, J.W. and Galvin, G. (2003), 'Quantification of odours from piggery effluent ponds using an electronic nose and an artificial neural network', *Biosystems Engineering*, Vol. 86, No. 4, pp. 399-410.

Wood, A. and Stankovic, J.A. (2000), 'Denial of service in sensor networks', *IEEE Computer*, Vol. 35, No. 10, pp. 54-62.

About the authors

Dinesh Kumar Saini is working as Associate Professor in Faculty of Computing and Information Technology Sohar University which is affiliated with university of Queensland. He received his Ph.D. in software Systems, Prior to that he has done his Masters of Engineering in software systems. He is member of major professional bodies. He has been actively engaged in teaching and research since last 16 years. His main research interests are in Environmental Informatics, Learning Content Management Systems, Searching & Recommending Techniques, Mathematical Modeling, Simulation, Cyber Defense, Network Security, Computational Intelligent Techniques, Software Testing and Quality.

Corresponding author. Faculty of Computing and Information Technology, Sohar University, Oman. Research Fellow and Adjunct Faculty, School of ITEE, University of Queensland, Australia. P.O.Box.-44, PC-311, Sohar, Sultanate of Oma. Tel: (+968) 26720101. E-mail address: dinesh@soharuni.edu.om, dkssohar@gmail.com

Sanad Al Maskari is working as Lecturer in the Faculty of Computing and Information Technology Sohar University. Currently he is on study leave for pursuing his Ph.D. in the university of Queensland Brisbane Australia. E-mail address: sanad.almaskari@uqconnect.edu.au