Analysis of Data Mining Techniques and its Applications

Fathimath Zuha Maksood
L4 BSCO
Caledonian College of Engineering
Sultanate of Oman

Geetha Achuthan
Senior Lecturer
Caledonian College of Engineering

ABSTRACT

The exponential increase in data over the recent years has urged for techniques to log, process and analyze these records. Heavy data repositories with a bulk of unprocessed content can lead to wastage of storage space as well as loss of hidden information. Since the late 90s, efforts have been taken to refine the concept of Knowledge Discovery in Databases and data mining. Organizations have started incorporating this approach to market their promotions as well as predict the buyers' choices. This paper is aimed at providing a detailed introduction to data mining, review of real world applications pertaining to the concept, big data and data mining techniques, as well as an integrated overview of the recent studies related to smart cities in the field of traffic prediction and forecasting energy consumption, especially in Oman.

General Terms

Data Mining, Applications and Techniques, Literature Survey

Keywords

Data Mining, Big Data, Smart City, Clustering, Classification, Regression

1. INTRODUCTION

Data mining generally refers to the process of extracting interesting hidden information from available chunks of data, which could otherwise be manually impossible. Even though, this description provides a rather raw image of data mining, the concept has been defined in various formats in the past. These varying definitions can be attributed to the introduction of the phrase "Knowledge Discovery in Databases" in the first Knowledge Discovery in Databases (KDD) Workshop (1989). Since then, researchers and authors have related KD to Data Mining, with some claiming both to mean the same, mainly due to the fact that knowledge is the product obtained from mining data. One of the first attempts to precisely explain knowledge discovery was made in [1], where it was defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Although, the definition was explicitly meant for knowledge discovery, it was merged as a synonym for data mining and incorporated in [2], further clarifying that, it encompassed a number of technical approaches such as clustering, classification, analyzing changes and detecting anomalies. Initially, several approaches highlighted data mining from the knowledge discovery perspective, but, researchers have since then, performed analysis and focused on data mining and techniques as seen in [3]. A clear definition of data mining [4] states that, it is a search for the relationships and global patterns that exist in large databases which are hidden due to immensity of data, such as relationship between patients and their medical diagnosis. Even though, popular approval sided with the synonymous usage of the terms 'Data Mining' and 'Knowledge Discovery', several data analysts and researchers have debated on the clarity of this trend. The first distinct explanation for the two terms was provided in [5], referring to Knowledge Discovery as the entire process of discovering new information from data sources, of which, Data Mining is a step wherein particular rules and algorithms are applied. These algorithms form the basis of knowledge discovery so that interesting patterns can be extracted from the given data. Apart from Data Mining, KD involves a series of activities such as data preparation, data selection, data cleaning and preprocessing, mining for interesting relationships and presenting and visualizing the acquired patterns.

The above definitions and descriptions of Knowledge Discovery and Data Mining are theoretically formulated from the need for it in today's world. The development of technology has gradually resulted in the replacement of manual logs by machine. Data have since been accumulated, using the traditional file processing systems in an unordered way. This led to the mismanagement of data which was later replaced by databases. The ease of use of related databases have allowed most organizations to adopt the technology for storing their transactions and related information. The number of databases was approximated to be around five million in early 90s; 20 years later, one can just imagine the volume of data accumulated in various technological sites. Hence, data acquisition can be recognized as a double-edged sword. Even though it implements an ease-of-use data storage strategy for organizations, data can pile up at an immense rate creating raw and unprocessed records. Despite this perception, the large amount of accumulated records can be utilized in an advantageous manner if it can be processed using appropriate means. It is logical to predict the existence of interesting relationships or hidden patterns in records aggregated over the years. Often, this information can be used to describe the records stored, find patterns in a users' transaction which was previously unknown, predict forthcoming data and most importantly use these details to create an information rich smart system [6]. If a user's behavior can be predicted by an organization using machine learning methods on existing dataset, it can prove to be an advantage as they now have to culminate only a known required amount of resources for the particular customer. This is one example of how data mining can be used to create a smarter environment. The next section further illustrates certain real time applications of data mining.

This paper aims at providing a detailed analysis of data mining, its development and implications in the real world. Several examples have been chosen to indicate the usage of data mining in retail, medicine and health care as well as in the educational arena. The progress of this reviving technology to satisfy the big data culture and the emerging concept of smart cities are further analyzed in this paper. .

The paper structure is as follows. Section 2 provides an insight to various real time applications of data mining. Popular techniques adopted for these applications are discussed in Section 3. The emergence of big data and inherent technologies are stated in Section 4. The concept of Smart

City and the possibility of its successful implementation in Oman is explained through Section 5 and 6 respectively.

2. REAL TIME APPLICATIONS OF DATA MINING

Applications of data mining are now increasingly visible in day-to-day life. The various methods in which these mining techniques are implemented to extract interesting patterns, and benefit trade, health and medicine, as well as educational fields are examined in this section.

2.1 Retail and Services

Trade, business and entrepreneurship form a very important sector of development. Most of the data dealing with business related transactions are stored in data warehouses and never again accessed for cleansing or analysis purposes [7]. If processed and initiated appropriately by data analysts, this data could no doubt render useful relationships, predict forthcoming transactions as well as make it easier for the business owners to regulate their customers' purchases. The most basic example of a successful data mining application in this field is the largest retailer in US, Walmart. With its attempt to revive the failure in marketing online and e-commerce, Walmart began using creative, yet logical methods to reach out to their customers in contrast to normal centers in the retail industry [8]. The supply chain had been collecting and storing a large amount of information continuously over a countable period of time, which is now used as leverage. It can claim to house details of more than 145 million Americans at present, linking their information to their personal websites, pages, and their activities across the internet. In other words, Walmart acquires data from their consumers using digital interfaces and links them to their personal accounts, any other chunk of information available online, or from a raw source. This information is then aggregated with existing or new algorithmic rules, in order to provide specific details about the customer behavior. Customer behavior can refer to the prediction of their next purchase, times at which they visit the shop, comparison to global consumption strategies and prediction of diseases and other biological impacts. Walmart has successfully implemented a complicated procedure which some claim to have violated privacy of their consumers [9]. On the contrary, it is otherwise impossible to maintain a discreet image in this digitized world where every step we take logs some record into a database, and the accurate manipulation of these data is the reason that lead Walmart to achieve great success.

Target is yet another famous retail icon which employed data mining to target customers. Irrespective of the legitimacy of the source, [10] reported that the corporation had successfully predicted pregnancy in one of their customers by mining the combination of products bought. She was further provided with condiments and related promotion goods. Even though it sounds simple, mining of large complex databases is not an easy task and hence, recent discussions have led researchers to develop frameworks which could accurately predict consumer behavior [11, 12]. This can lead to rough estimation of the budgets required and sales made by different industries for a period of time, further allowing data outliers to be defined and their characteristics to be listed. Even though Walmart's data analytics schemes have allowed it to race way ahead, upcoming retail industries are also utilizing data mining on their large warehouses to provide consumers with what they require at any given time. Since shopping is not expected to zone out soon, data analysis and discovery in this field is a very sought-after procedure.

The telecommunication industry proudly own large data warehouses, wherein millions of call records are logged in every second. The bulk of data can be efficiently mined to benefit the companies involved and create a better customer service environment. Data mining and Business Intelligence applications in the telecom field faces four key challenges, also known as the 4 Cs: Consolidation, Commoditization, Customer service, and Competition [13]. Availability of data and algorithmic research has allowed better marking and customer relation management using Neural networks, Association rules, classification and clustering. High traffic related network faults can also be predicted using prior behaviors in data Telecommunication Alarm Sequence Analysis. Another area in which data mining has overcome existing challenges is in Subscription Fraud Detection using Deviation Detection and Anomaly detection [14].

2.2 Medicine and Health Care

Apart from successes in the business and retail arena, data mining has reflected its advantages in the field of medicine and health. The formulation of algorithms is still in its very basic stage due to the complexity of health care and slower rate of technology adoption [15]. Prediction algorithm is the main approach focused upon by medical informatics professionals. Researchers are aiming to aggregate patient data and relative response throughout consultation in order to predict the outcome of interest. Furthermore, it aims at using data mining to predict the effectiveness of certain surgical procedure, medical tests and medications [16]. This in turn, can help raise the standards of clinical decision making and thereby, contribute to the health and safety of people. In [17], the research used a simplified data set with twenty patient records to predict the patient's long term clinical status in physiotherapy. The dataset had only three attributes which stored the patient's health, the timing and complications of the operation and the outcome which was recorded two years after the successful treatment. After implementing data mining techniques (DMTs) such as naïve Bayesian classifier and decision trees on the available dataset, the research provided predictive results for patients undergoing the operation. This is one of the few approaches which can be undertaken to implement data mining in medical informatics.

Predictive analysis in the field of microbiology has been mainly related to mining genome related data. Researches were conducted on DNA microarrays consisting of thousands of genes, aiming at diagnosis of various diseases. In this approach, researchers target to answer biological questions by mining thousands of genomic datasets iteratively, potentially spanning various molecular activities, technological platforms and model organisms [18]. The most popular objective of genome related data mining is to revolutionize health care by intensifying our knowledge in the molecular level of the disease. Once data is mined at a core level, it will consequently be easier to determine the basis of various diseases, which can further be built upon using higher level research [19]. This approach is data intensive accommodating huge datasets which are highly heterogeneous. Hence, a lot of challenges are involved in this process, such as, the inability of conventional algorithms to scale well due to the large integrated dataset, as well as, data storage leading to wastage of unsustainable amounts of space for large repositories. Related researches in the present scenario have mainly tried to concentrate on solving these issues. One of the technical solutions involved usage of web applications instead of in-house strategies to aggregate datasets, and programmatic APIs with do-it-yourself solutions for computational queries [20].

Therefore, medical based data mining focused on predictive models to predict patient outcomes, surgery success rate, as well as the disease at a molecular level.

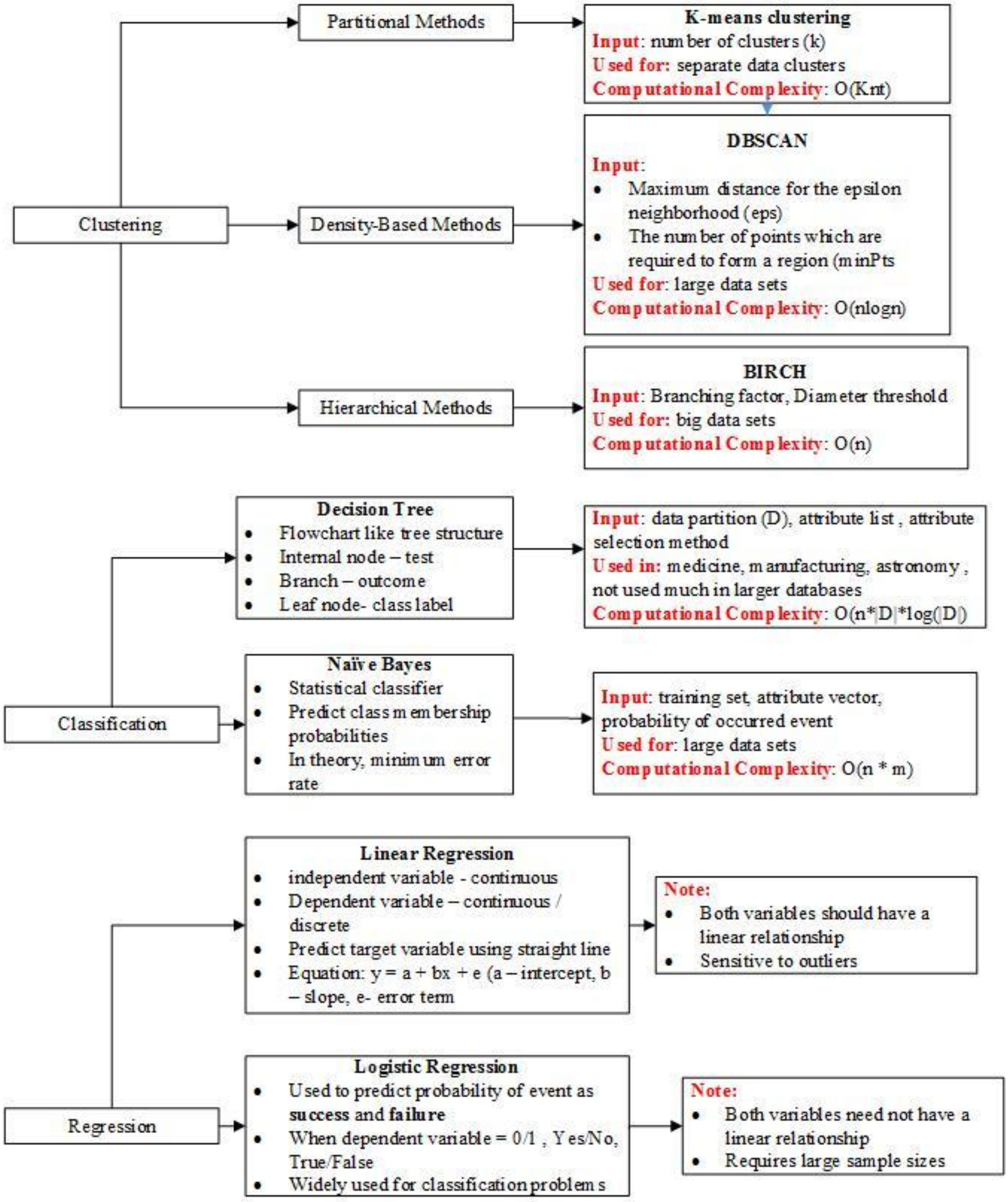
Clustering
Partitional Methods
K-means clustering
Input: number of clusters (k)
Used for: separate data clusters
Computational Complexity: O(Knt)
Density-Based Methods
DBSCAN
Input:
Maximum distance for the epsilon neighborhood (eps)
The number of points which are required to form a region (minPts
Used for: large data sets
Computational Complexity: O(nlogn)
Hierarchical Methods
BIRCH
Input: Branching factor, Diameter threshold
Used for: big data sets
Computational Complexity: O(n)
Classification
Decision Tree
Flowchart like tree structure
Internal node – test
Branch – outcome
Leaf node- class label
Input: data partition (D), attribute list , attribute selection method
Used in: medicine, manufacturing, astronomy , not used much in larger databases
Computational Complexity: O(n*|D|*log(|D|)
Naïve Bayes
Statistical classifier
Predict class membership probabilities
In theory, minimum error rate
Input: training set, attribute vector, probability of occurred event
Used for: large data sets
Computational Complexity: O(n * m)
Regression
Linear Regression
independent variable - continuous
Dependent variable – continuous / discrete
Predict target variable using straight line
Equation: y = a + bx + e (a – intercept, b – slope, e- error term
Note:
Both variables should have a linear relationship
Sensitive to outliers
Logistic Regression
Used to predict probability of event as success and failure
When dependent variable = 0/1 , Yes/No, True/False
Widely used for classification problems
Note:
Both variables need not have a linear relationship
Requires large sample sizes

Figure 1: Data Mining Techniques

2.3 Higher Education

An upcoming application of data mining is undertaken by higher educational institutions due to the gradual hike in the amount of data along the years. Data mining in this discipline is used to understand student behavior, such as the trends which would indicate student transfer, credit hours trend as well as the skill sets of various clusters of students and their redundant characteristics [21].

In [22], the authors have described different grounds on which mining technology can be applied, such as analysis and visualization of student scores using math and graphs, predicting student performance with regression techniques and fuzzy association rules, outlier detection from the lot by inheriting supervised, unsupervised, or semi-supervised learning, grouping students and managing classes accordingly by clustering such as k-means, model-based and hierarchical agglomerative, as well as planning and scheduling time tables and studying hours using regression, clustering and classification. Decision trees and back propagation cluster neural networks are further used to plan educational training courses at present.

Therefore, business and retail, medicine and health care, and higher education form three major sectors where the effects and advantages of employing data mining can be observed.

3. DATA MINING TECHNIQUES

Section 2 highlighted the areas where data mining is used in real life. This section will focus on the conventional techniques utilized by analysts to implement the data mining algorithms.

3.1 Classification

The most common technique used in mining is Classification [23]. Classification, as the name suggests, allows the user to classify large populous data into a model which sorts them into a predefined set of classes. Fraud detection, classifying patients from primary health care centers to specialists, and credit risk applications are a few ways in which classification is implemented [24]. Classification process often employs supervised learning and classification, and is mostly used for predictive modeling. Some of the popular algorithmic models employed in classification are decision trees, neural networks, Bayesian classification, Support Vector Machines (SVM) and classification based on association.

3.2 Clustering

Clustering is another DMT which has gained popularity among the mining community. It involves identifying clusters and grouping similar objects together in each cluster. While classification is mentioned to have employed supervised learning, clustering process mainly uses unsupervised learning method (some clustering models use both) [25]. Analyzing the similarity in organizational behavior, financial trends and clustering homes based on energy consumption are a few algorithms used in this technique. Even though researchers have mainly focused on evaluating and implementing Partitioned (K-means) algorithms [26, 27], other clustering methods include Hierarchical (CURE, BIRCH), Grid – based (STING, WaveCluster), Model-based (Cobweb) [28], and Density based (DBSCAN) [29].

3.3 Regression

Regression is a technique used for predictive modeling. Often related to classification, regression technique also includes Support Vector Machines (SVM) is in the case of the former. The idea of regression analysis is to model a relationship between one or more attributes (independent and dependent variables) in the dataset, so that the change in one of the variables can be used to predict the values of the other. It can also be used to predict the advantages and disadvantages of the future market as well as the course of resource consumption in the coming years. Since, real world prediction requires the integration of many complex attributes, different models (as in the case of classification) have to be used to implement prediction. Classification and Regression Trees (CART) is a decision tree algorithm which uses classification trees to classify the dependent (response) variables and regression trees to predict the values of the response variables continuously. Different regression methods used are logistic regression, linear regression, multivariate linear regression, nonlinear regression and multivariate nonlinear regression.

Due to the emergence and invention of data mining algorithms and techniques at present, it should be noted that optimal DMTs, models and algorithms should be chosen as per the requirements of the project or research, and a considerable amount of data should be collected in order to achieve expected results and analyze them appropriately. A few common data mining techniques, examples and their characteristics are illustrated in Figure 1.

4. MINING BIG DATA: TECHNOLOGIES

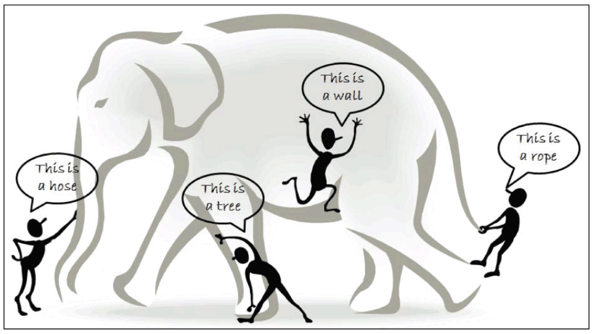

Figure 2: Big Data Perspective: Blind men and the elephant[30]

In the above figure, famously titled as “the blind men and the elephant”, four blind men are seen identifying the object from various perspectives, consequently logging in different data about the same object from four different views. Furthermore, if the object is expected to grow in size, and the four men discuss their identified characteristics, data obtained from them becomes even more complex. This is a simple analogy to explain the generation and enormity of big data [31]. The data mining concepts and applications discussed in this paper prior to this section were mostly related to big data. Normal data is characterized by small volumes of data, which have batched velocities and are structured in terms of variety [32]. These data can be stored into a simple computer’s main memory for mining. As the volume and the velocity of data flow increases, it reaches a stage where it can no longer be stored on a normal PC. In the bioinformatics research [33], authors state that Big Data approach is relatively new in Health Informatics and it can be used to predict high level patient information if a transitional approach is undertaken beginning from molecular level. Therefore, it led to various solutions to mine data at this volume, such as, introduction to Scalable parallel Classifier using parallel processors [34, 35] as well as use cloud architecture for storage [36].

The definition of big data was gradually standardized by IBM [37] to four Vs – namely Volume, Velocity, Veracity and Variety. Volume relates to the scale of stored data; since 2.5 quintillion bytes are created every day, and 6 billion people out of the existing 7 billion own cellphones, the data storage is expected to rise to at least 40 zettabytes at the end of 2020. Velocity analyses the rate of streaming data, i.e., it is determined by the speed at which data is logged into the system. A huge amount of data can be generated if the user data is stored in every second and these data cannot be stored into relational database as discussed earlier. Most common example in the present era for data velocity lies in the social media website Twitter , where people tweet about trending topics every second, creating a large amount of data in a very short time. In 2013, [38] cited that more than 140 million active users publish over 400 million tweets every day. Most automated data does not often present erroneous details, but the automated system could have flawed at some time and created incorrect information; this is the basis of veracity which indicates the existence of uncertain data. Many users often doubt the accuracy of data and this lead to further complexity of the huge dataset. Finally, the data logged into databases can be structured, unstructured, graphical, text or in any form, such as the different attributes which characterize health related levels as discussed in Section II. Even though these features describe Big Data, there have been contradictions to the standardization of this definition [39] stating that the definition of Big Data Analytics should be determined by the question “why” rather than “what”. The three perspectives that can define Big Data space includes ‘learning over knowing’, ‘extreme experimentation’ and the ‘new IP’, which could trigger people to think of it as an entity which occupies a lot of space, rather than a huge amount of data.

There exist a lot of technologies which can be used to implement Big Data Analysis. Hadoop is an open source software framework, which can be used to process huge datasets on a distributed file system. It is growing to be one of the popular technologies due its fault-tolerance, ability to withstand hardware failure, stream access to data sets and most importantly support large datasets [40]. Hadoop Distributed File System (HDFS) normally houses data in the order of gigabytes or terabytes. Since HDFS is mainly used to store raw files, it is often integrated with other software technologies during data mining. HBase is an open source, non-relational database which supports the distributed system and is used to store data during implementation. HIVE is also a storage model which incorporates relational database and SQL like techniques. More often than not, Hadoop is clustered with MapReduce, a software framework introduced by Google to process huge datasets [41]. Another, upcoming technology among statisticians is R, an open source statistical programming language widely used for statistical software development and data analysis. Recent researches have suggested the integration of R with Hadoop in three ways: R with Streaming, Rhipe and RHadoop, in order to merge analyzing techniques of the former with solutions to data storage and big data problems from the latter [42].

Big Data architecture in Hadoop is as shown in Figure 3. Various issues are expected to be faced by Big Data in the future. Data from secure agencies can be mined to generate important relationships and predict outcomes which can help national security and important decisions. But these benefits are often subdued by concerns over data protection and security [43]. Policies over privacy of data acquired, intellectual property and liability will need to be addressed. Furthermore, a talented workforce will be required to learn, analyze and understand big data. There is a tendency for organizations to face shortage of data analysts in the near future. It is also seen that companies will have to access data from multiples sources to create a reliable big data application, and this would not be an easy task. Therefore, solutions to these problems are to be formulated in order to overcome these challenges and implement analysis on big data.

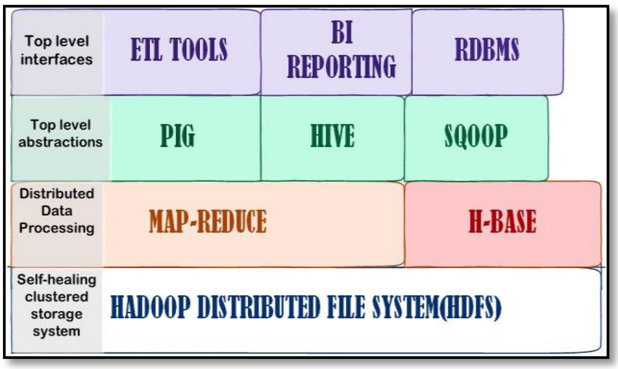
Top level interfaces
ETL TOOLS
BI REPORTING
RDBMS
Top level abstractions
PIG
HIVE
SQOOP
Distributed Data Processing
MAP-REDUCE
H-BASE
Self-healing clustered storage system
HADOOP DISTRIBUTED FILE SYSTEM(HDFS)

Figure 3: Hadoop Big Data Analytics Infrastructure [44]

5. SMART CITY: THE EMERGING CONCEPT

Converging technologies have led to the rise of smart cities [45]. The gradual progress from the invention of smart phones, tablets, meters, cars, homes have finally headed to the concept of Smart City. With its widespread citation in written works and steady growth, authors have attempted to describe the concept in various forms; but there still isn’t an accepted international definition [46]. [47] defines a six-function typology for the creation of smarter cities, which include smart economy(competitiveness), smart people (social and human capital), smart governance (participation), smart transport (transport and Information and Communication technology), smart environment (natural resources), and smart living (quality of life). It is said that the Focus Group on Smart Sustainable Cities (FG-SSC) formulated their definition of Smart Sustainable City after regarding about hundred prior definitions with, “A smart sustainable city is an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social and environmental aspects [48]. Authors [49] have also linked the derivation of this concept to researches based on virtual cities, ubiquitous cities and computable cities from the late 90s. Certain indicators were also mentioned which could rate the smartness level of cities such as, adoption of OpenData and OCG Standard, free WiFi, project implementation of augmented reality for tourism, crowdfunding initiatives, decisions taken by crowdsourcing, implementation of INSPIRE Directive and quantity of public services achievable through applications. These definitions can be related to the technological layer overlaid on the already existing cities. One has to break down this phrase so that common man can understand the concept of Smart cities, its importance and contribute in its growth and development. It is often misinterpreted that smart cities have much to do with smart electronics, devices and objects. Although these are some of the important components in the smart network, cities can be described as being “smart” only if it utilizes ICT to create an environment where the pillars of the city (people, economy, can work , interact and consequently improve the quality of living.

Human population is expected to rise to about 9 billion in the next fifty years, out of which, three in every five would live in urban areas [50]. Urbanization has always promised a higher standard of living; but, it has also led to problems which increasingly obstruct the way of life. Heavy traffic is an issue which is pondered upon in urban areas, as the number of personal vehicles has been increasing over time. This has led to traffic congestion during commute and delay in reaching workplace and destinations. A very high scope for traffic prediction is seen at present, as solutions to this issue can lead to a tremendous improvement in the urban way of life. Various approaches in traffic prediction include, prediction of traffic inducing nodes in spatially correlated clusters using Affinity Propagation and Neural Networks [51]. Simulation of real time data showed that this method could help predict the future traffic conditions in each of the cross roads. Similarly, Auto Regressive Integrated Modeling Algorithm and its enhanced versions are utilized for short term and long term traffic predictions, achieving expected results as seen in [52]. Energy demand has also faced a steep rise in urban areas due to the increase in the production and usage of smart devices. It is said that more than half the world population lives in cities, and utilizes about 75% of the worlds' energy production. Hence, urban life has led to the overall consumption of the produced energy with a demand for more. Energy consumption with degrading resources is a liability and hence, data mining with forecasting models have been employed to predict the future energy consumptions. Attempts at predicting energy consumption at city, building, and appliance levels can be reviewed in [53, 54, 55]. At the city level, k-means clustering followed by time series forecasting was seen as a good approach, leading to the reinvention of the energy bill for most cities of United States. Government and authorities can use this processed information to regulate and visualize the energy consumption in cities through comparisons, graphs and tables. Similarly, outlier detection which can predict faults in building energy consumption is also reviewed, with CART integrated with GESD as the best approach for fault diagnosis. The discussed aspect would create a smart economy, smart participation as well as smart transportation in cities as traffic and energy consumption can now be forecasted using Data Mining Techniques. The major feature about proposing smart cities in the present world is the utilization of ICTs to create a sustainable environment; a concept which has failed to kick-start in many countries even with water management strategies, green environment schemes, etc. [56, 57, 58]. Therefore, a fine proposal has to be stated in order to integrate the various aspects of urban life and initiate the development of smart cities around the world.

6. INTRODUCTION TO SMART CITIES IN OMAN*

Sultanate of Oman, a country which lies in the Arabian Peninsula is well known for its solar energy received throughout the year, but, electricity production of the nation is fully attributed to the abundant oil, gas and coal reserves [59]. Due to the increase in population and economic growth in the recent years as well as development of the industrial sector, demand for electrical power in the region has been growing rapidly. Despite efforts of securing energy resources, the demand has increased to about 8-10% in the recent years [60]. Sustainability and energy efficiency concepts in order to satisfy the demands of the public as well as conserve the depleting resources have pushed the electricity authorities to turn to renewable energy sources and make effective changes in the current electricity production management. They have also concluded that energy efficiency could almost halve the amount of gas consumed, thereby helping to sustain resources [61]. Similarly, development of transportation which involves road traffic and accidents is another dimension which can be delved into, in order to promote smart cities in the country. Royal Oman Police statistics indicate that more than five hundred people die due to road accidents in Oman every year, which estimates the nation to be one of the highest in traffic accident rate [62].

Researchers in Oman have often overlooked the concept of smart cities, therefore, only a few documents pertaining to this issue can be observed from this region All these sources either explain the concepts of smart cities [63], or analyze potential areas and data without dealing with mining techniques [64, 65]. A data analysis report of residential electricity usage was published in [66] which provided results, but failed to mention the methodology used to mine and analyze the available data. The process of Knowledge Discovery in Databases is yet to be promoted and its advantages should be reviewed by mining existing datasets in the field of economy, transportation, energy and other modules which could lead to the formation of smart cities. Smart City initiatives can be bootstrapped through a self-sustainable model as presented in [67]. In this paper, the authors have mentioned three dimensions, namely political, technological, and financial so as to model smart cities in a country. Political dimension should involve the emergence of smart city departments, similar to IT departments, where in the administrative side of the technology is developed and managed. The technological dimension should facilitate the technological equipment which can store data and improve its availability as Open Datasets. A coherent self-sustainable business model should emerge and manage the source of finance for the entire setup. This model can be adopted in order to begin a smart urban culture in Oman.

7. CONCLUSION

This paper presented a review of data mining, different applications and techniques involved during data analysis, as well as the emergence of the concept of smart cities and its association with the data mining technology.

The paper clearly describes the evolution of the terminology of data mining and knowledge discovery of databases as well as the necessity and importance of its existence during this "bulk-data" era. It also denotes how data mining techniques are used to ease the commercial aspects in our daily life. Real Time applications of Data Mining are further described using three major areas, namely, Retail, Medicine and Healthcare as well as, Higher Education. It is seen that DMTs are more commonly applied in retail in order to market products and increase sale. Apart from shopping centers, health care units are emerging with surgery outcome prediction rates for advanced diseases, while, higher education which is relatively new to the field, are finding new ways to create a healthy study environment and redefine pupils' interests in their area of study. Classification, clustering, and regression, which are the most commonly used data mining techniques worldwide, are also described precisely in this paper.

Big Data technology has been a fast growing concept in the past few years due to the increase in production and storage of data over the years. This paper delves into the major notions of big data and the way it is perceived by the data mining community in the present era. Another feat achieved by DM is its ability to develop notions for smart cities, wherein various components such as energy, transport, economy, environment, and people intermingle to form a sustainable and hence, smart society. A few works based on smart cities are represented in the Table 1. This shows that development of this technology will be able to accurately predict various entities such as energy consumption, road traffic etc., thereby, allowing people to make quicker, smart decisions, devising into a much anticipated Smart City.

*This literature review is conducted as a pre-requisite for a project to suggest the concept of smart cities in Oman, wherein traffic and energy consumption data will be mined and forecasted to achieve a predictive model for the region.

Table 1 Different works based on the notion of Smart Cities

Author(s), Year	**Paper Title**	**DMTs used**	**Observations/ Areas focused**
Carlos Costa, Maribel Yasmina Santos, 2015	Improving Cities Sustainability through the use of Data Mining in a context of Big City Data	k-means clustering, time series forecasting	Validity of simulated dataset is not confirmed. Near accurate energy prediction and reinvention of the energy bill can open new doors to a sustainable environment.
Imran Khan, Alfonso Capozzolia,, Stefano Paolo Corgnati, Tania Cerquitelli, 2013	Fault Detection Analysis of Building Energy Consumption using Data Mining Techniques	CART, k-means clustering, DBSCAN	Faults could be detected more accurately using CART with GESD outlier detection than clustering
Alexandra Moraru, D Mladenić, 2012	Complex Event Processing and Data Mining for Smart Cities	Association rules using Weka machine learning toolkit	Results contradicted pre-assumptions; hence, larger volume of data should be mined for clarity of results.
Bowu Zhang, Kai Xing, Xiuzhen Cheng, Liusheng Huang, and Rongfang Bie, 2012	Traffic Clustering and Online Traffic Prediction in Vehicle Networks: A Social Influence Perspective	Affinity propagation, clustering, neural networks	Predicts traffic inducing nodes on road clusters. Real time traffic prediction yet to be performed.
Bei Pan, Ugur Demiryurek, Cyrus Shahabi, 2012	Utilizing Real World Transportation Data for Accurate traffic position	ARIMA, HAM, time series prediction, regression	Long term and short term prediction of traffic can be performed accurately

8. REFERENCES

[1] W. J. Frawley, G. Piatetsky-Shapiro and C. Matheus, "Knowledge Discovery in Databases: An Overview," AI Magazine, vol. 13, no. 3, pp. 57-70, 1992.

[2] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining Process," The Knowledge Engineering Review, vol. 21, no. 1, pp. 1-24, 2006.

[3] F. Weiping and W. Yuming, "The Development of Data Mining," International Journal of Business and Social Science, vol. 4, no. 16, pp. 157-162, 2013.

[4] T. Silwattananusarn and K. Tuamsuk, "Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012," International Journal of Data Mining & Knowledge Management Process, vol. 2, no. 5, pp. 13-24, 2012.

[5] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, no. 3, pp. 37-54, 1996.

[6]] Smita and P. Sharma, "Use of Data Mining in Various Field: A Survey Paper," IOSR Journal of Computer Engineering, vol. 16, no. 3, pp. 18-21, 2014.

[7] S. Adelman, "The Data Warehouse Database Explosion," Enterprise Information Management Institute, March 2008.

[8] Universidad San Pablo, "Case study: The Rise of Wal-Mart," 21 June 2012. [Online]. Available: http://biolab.uspceu.com/datamining/WalMart.pdf. [Accessed 28 November 2015].

[9] Ian Davey and Technolegis, "Consumers, Big Data, and Online Tracking in the Retail Industry: A CASE STUDY OF WALMART," 10 August 2014. [Online]. Available: https://saveballston.files.wordpress.com/2014/08/walmart_privacy_.pdf. [Accessed 29 November 2015].

[10] K. Hill, "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did," Forbes, 16 February 2012. [Online]. Available: http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/. [Accessed 29 November 2015].

[11] N. Baby and P. L.T, "Customer Classification And Prediction Based On Data Mining Technique," International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 12, pp. 314-318, 2012.

[12] G. Simonsen, "Retail Insights," Online Digital Publishing, [Online]. Available: http://onlinedigitalpublishing.com/article/Retail_Insights/549723/52404/article.html. [Accessed 30 November 2015].

[13] D. Pareek, Business Intelligence for Telecommunications, New York: Auerbach Publications, 2007.

[14] M. V. Joseph, "Data Mining and Business Intelligence Applications in Telecommunication Industry," International Journal of Engineering and Advanced Technology, vol. 2, no. 3, pp. 525-528, 2013.

[15] D. Crockett and B. Eliason, "What is Data Mining in Healthcare?," HealthCatalyst, [Online]. Available: https://www.healthcatalyst.com/data-mining-in-healthcare. [Accessed 30 November 2015].

[16] J. Jackson, "DATA MINING: A CONCEPTUAL OVERVIEW," Communications of the Association for Information Systems, vol. 8, pp. 267-296, 2002.

[17] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," International journal of medical informatics: Elsevier, vol. 77, pp. 81-97, 2008.

[18] C. Huttenhower and O. Hofmann, "A Quick Guide to Large Scale Genomic Data Mining," 3 April 2012. [Online]. Available: http://www.stat.harvard.edu/18ACCF14-7036-4F35-A3BD-A3D55AF66DE8/FinalDownload/DownloadId-4E694BAB09E8713F88915D8E891CF0D9/18ACCF14-7036-4F35-A3BD-A3D55AF66DE8/NESS10/HuttenhowerMarkowetz/A%20Quick%20Guide%20to%20Large%20Scale%20Genomic%20Data%20Mining.pd. [Accessed 30 November 2015].

[19] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy and P. Tarczy-Hornoch, "Data integration and genomic medicine," Journal of Biomedical Informatics, vol. 40, no. 1, pp. 5-16, 2006.

[20] P. Szolovits, "Mining Clinical Data to build Predictive Model," 2 May 2013. [Online]. Available: https://www.siam.org/meetings/sdm13/szolovits.pdf. [Accessed 30 November 2015].

[21] J. Luan, "Data Mining and Its Applications in Higher Education," Wiley Periodicals, pp. 17-36, 3 June 2002.

[22] M. Goyal and R. Vohra, "Applications of Data Mining in Higher Education," International Journal of Computer Science Issues, vol. 9, no. 2, pp. 113-120, 2012.

[23] B. M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," Indian Journal of Computer Science and Engineering, vol. 1, no. 4, pp. 301-305, 2011.

[24] R. Petre, "Data Mining Solutions for the Business Environment," Database Systems Journal, vol. 4, pp. 21-28, 2013.

[25] L. Rokach and O. Maimon, "Clustering Methods," in The Data Mining and Knowledge Discovery Handbook, New York, Springer US, 2006, pp. 321--352.

[26] [N. Sharma, A. Bajpai and R. Litoriya, "Comparison the various clustering algorithms of weka," International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 5, pp. 73-80, 2012.

[27] S. Kumar and N. , "K-Mean Evaluation in Weka Tool and Modifying It using Standard Score Method," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, no. 9, p. 2704 – 2706, 2014.

[28] J. Han and M. Kamber, Data Mining - Concepts and Techniques, 2nd Edition ed., San Fransisco: Elsevier, 2008.

[29] [29] P. Shrivastava and H. Gupta, "A Review of Density-Based clustering in Spatial Data," International Journal of Advanced Computer Research , vol. 2, no. 5, pp. 200-202, 2012.

[30] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, Blind Men and the elephant, 2014.

[31] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, 2014.

[32] "Taming Big Data: Small Data vs. Big Data," IBM. [Online]. [Accessed 14 November 2015].

[33] M. Herland, T. M. Khoshgoftaar and R. Wald, "A review of data mining using big data in health informatics," Springer Journal of Big Data, vol. 1, no. 2, pp. 1-35, 2014.

[34] J. Shafer, R. Agrawal and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining," in Proceedings of the 22nd VLDB Conference , Mumbai, 1996.

[35] D. Luo, C. Ding and H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining," in IEEE 12th International Conference on Data Mining , Brussels, 2012.

[36] Y. Li, L. Guo and Y. Guo, "An Efficient and Performance-Aware Big Data Storage System," in Cloud Computing and Services Science, New York, Springer International Publishing, 2013, pp. 102-116.

[37] "The Four V's of Big Data," IBM, 2015. [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data. [Accessed 30 November 2015].

[38] S. Kumar, F. Morstatter and H. Liu, Twitter Data Analytics, New York: Springer, 2013.

[39] S. Agrawal, "I hate the whole concept of describing Big Data as a lot of data: Mu Sigma’s Dhiraj Rajaram," Tech Circle, 2 September 2014. [Online]. Available: http://techcircle.vccircle.com/2014/09/02/i-hate-the-whole-concept-of-describing-big-data-as-a-lot-of-data-ipo-is-a-possibility-mu-sigmas-dhiraj-rajaram/. [Accessed 20 October 2015].

[40] D. Borthakur, "HDFS Architecture Guide," 4 August 2013. [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf. [Accessed 30 November 2015].

[41] G. Yogaraj and A. A. Arun, "Mining High Dimensional Data Sets Using Big Data," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 2, pp. 970-974, 2015.

[42] B. Oancea and R. Dragoescu, "Integrating R and Hadoop for Big Data Analaysis," Revista Romana de Statistica, vol. 2, pp. 83-94, 2014.

[43] A. Chakravarthy, Components of Hadoop Architecture, Cisco, 2012.

[44] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburg and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," May 2011. [Online]. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. [Accessed 29 November 2015].

[45] J. Svetlik, "Rise of the smart city: The awesome and scary reality of future urban living," Wearable, 22 July 2015. [Online]. Available: http://www.wareable.com/internet-of-things/the-awesome-and-scary-future-of-our-cities-2025. [Accessed 30 November 2015].

[46] M. Batty, K. Axhausen, G. Fosca, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis and Y. Portugali,

"Smart Cities of the Future," Centre for Advanced Spatial Analysis, London, 2012.

[47] M. Romkey, " Smart cities…not just the sum of its parts," Deloitte, Dubai, 2015.

[48] "Focus Group on Smart Sustainable Cities," ITU, 2015. [Online]. Available: http://www.itu.int/en/ITU-T/focusgroups/ssc/Pages/default.aspx. [Accessed 30 November 2015].

[49] B. Murgante and G. Borruso, "Cities and Smartness: A Critical Analysis of Opportunities and Risks," in Computational Science and Its Applications – ICCSA 2013, New York, Springer Berlin Heidelberg, 2013, pp. 630-642.

[50] M. Tercek, "More People, More Problems: Future-Proofing our Cities," [Online]. Available: http://bigthink.com/experts-corner/more-people-more-problems-future-proofing-our-cities. [Accessed 20 December 2015].

[51] B. Zhang, K. Xing, X. Cheng, L. Huang and R. Bie, "Traffic Clustering and Online Traffic Prediction in Vehicle Networks: A Social Influence Perspective," in 2012 Proceedings IEEE INFOCOM, Orlanndo, 2012.

[52] B. Pan, D. Ugur and S. Cyrus, "Utilizing Real-World Transportation Data for Accurate Traffic Prediction," in 2012 IEEE 12th International Conference on Data Mining (ICDM), Brussels, 2012.

[53] C. Costa and M. Y. Santos, "Improving Cities Sustainability through the Use of Data Mining in a Context of Big City Data," in Proceedings of the World Congress on Engineering, London, 2015.

[54] I. Khan, A. Capozzoli, S. P. Corgnati and T. Cerquitelli, "Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques," in Energy Procedia: The Mediterranean Green Energy Forum 2013, Fez, 2013.

[55] N. Arghira, S. Ploix, I. Fagarasan and S. S. Iliescu, "Forecasting Energy Consumption in Dwellings," in Advances in Intelligent Control Systems and Computer Science, Berlin, Springer Berlin Heidelberg, 2013, pp. 251-264.

[56] V. Ginn, "TPPF: California’s failed green energy project lesson for Texas," Midland Reporter-Telegram, Texas, 2015.

[57] Sara, "FAIL: 20 Infamous ‘Green Innovations’ That Aren’t," WebEcoist, [Online]. Available: http://webecoist.momtastic.com/2008/10/20/failed-green-technologies-designs-and-innovations/. [Accessed 13th December 2015].

[58] T. Singh, "6 Ways in Which London 2012 has Failed to be ‘The Green Olympics’," inhabitat, 8 May 2012. [Online]. Available: http://inhabitat.com/6-ways-in-which-london-2012-has-failed-to-be-the-green-olympics/. [Accessed 13 December 2015].

[59] "Oman - Electricity production," Indexmundi, [Online]. Available: http://www.indexmundi.com/facts/oman/electricity-production. [Accessed 8 March 2016].

[60] "Renewable energy," Public Authority for Electricity and Water, [Online]. Available: https://www.paew.gov.om/Our-role-in-Oman/Renewable-energy. [Accessed 8 March 2016].

[61] C. Prabhu, "Energy efficiency can halve gas consumption in Oman," Oman Observer, 30 May 2015. [Online]. Available: http://omanobserver.om/energy-efficiency-can-halve-gas-consumption-expert/. [Accessed 8 March 2016].

[62] E. H. AlHarrasi, B. Jrew and M. Abojaradeh, "Development of Traffic Accident Models in Oman," in Seventh Traffic Safety Conference, Amman, 2015.

[63] International Organisation for Knowledge Economy and Enterprise Development, "Smart Data & Well-Being," 29 October 2014. [Online]. Available: http://iked.org/pdf/Proj%20GENERAL%20Pres%2017%20Nov.pdf. [Accessed 10 December 2015].

[64] Authority of Electricity Regulation, Oman, "Study on Renewable Energy Resources, Oman," May 2008. [Online]. Available: http://www.aer-oman.org/pdf/studyreport.pdf. [Accessed 10 December 2015].

[65] Y. H. Zurigat, N. M. Sawaqed, H. Al-Hinai and B. A. Jubran, "Analysis of Typical Meteorological Year for," International Journal of Low Carbon Technologies, vol. 2, no. 4, pp. 323-338, 2007.

[66] T. Sweetnam, "Residential Energy Use In Oman:A Scoping Study," 13 January 2014. [Online]. Available: http://discovery.ucl.ac.uk/1425280/1/Oman%20Final%20Report%20v0%208_revised.pdf. [Accessed 12 December 2015].

[67] I. Vilajosana, J. Llosa, B. Martinez, M. Domingo-Prieto, A. Angles and X. Vilajosana, "Bootstrapping Smart Cities through a Self-Sustainable Model Based on Big Data Flows," IEEE Communications Magazine, pp. 128-134, June 2013.