Indonesian Journal of Electrical Engineering and Computer Science
Vol. 10, No. 3, June 2018, pp. 1234~1243
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v10.i3.pp1234-1243

A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics

Jesmeen M. Z. H.[1], J. Hossen[2], S. Sayeed[3], C. K. Ho[4], Tawsif K.[5], Armanur Rahman[6], E. M. H. Arif[7]
[1,2,5,6,7]Faculty of Engineering and Technology, Multimedia University, Melaka, 75450, Malaysia
[3]Faculty of Information Science & Technology, Multimedia University, Melaka, 75450, Malaysia
[4]Faculty of Computing and Informatics, Multimedia University, Melaka, 75450, Malaysia

Article Info

Article history:

Received Jan 15, 2018
Revised Mar 11, 2018
Accepted Mar 24, 2018

Keywords:

Big data
Big data analytics
Data cleaning
Dirty data
Machine learning

ABSTRACT

Recently Big Data has become one of the important new factors in the business field. This needs to have strategies to manage large volumes of structured, unstructured and semi-structured data. It's challenging to analyze such large scale of data to extract data meaning and handling uncertain outcomes. Almost all big data sets are dirty, i.e. the set may contain inaccuracies, missing data, miscoding and other issues that influence the strength of big data analytics. One of the biggest challenges in big data analytics is to discover and repair dirty data; failure to do this can lead to inaccurate analytics and unpredictable conclusions. Data cleaning is an essential part of managing and analyzing data. In this survey paper, data quality troubles which may occur in big data processing to understand clearly why an organization requires data cleaning are examined, followed by data quality criteria (dimensions used to indicate data quality). Then, cleaning tools available in market are summarized. Also challenges faced in cleaning big data due to nature of data are discussed. Machine learning algorithms can be used to analyze data and make predictions and finally clean data automatically.

Corresponding Author:

Jesmeen M. Z. H. & Dr. Jakir Hossen
Faculty of Engineering and Technology,
Multimedia University,
Melaka, 75450, Malaysia.
Email: jesmeen.online@gmail.com, jakir.hossen@mmu.edu.my

1. INTRODUCTION

In 2016, IBM estimated that in last two years only, around 2.5 quintillion bytes' data have been produced each day, which is currently 90% of total data [1]. This big data is usually created using devices like sensors and new technologies evolving in today's era, even more the data evolution amount will possibly accelerate. Whereas, Cisco forecasted by 2020, the volume of worldwide traffic will cross the Internet with IP WAN networks may reach to 2.3ZB each year [2].

The bulky and heterogeneous nature of big data requires investigation using Big data analytics. Big data analytics helps to discover concealed patterns, anonymous relationships, trends of current market situation, consumer preferences and other aspects of data that can assist institutes and companies to make up-to-date, faster and better decision for business.

By now, most well-known companies realized the demand of implementing big data analytics into their system for better products and services. Using big data capabilities any company can improve their products and services outcomes and grow productivity by obtaining meaningful visions to advance their work forward. There are different tools available in market to handle the big data but these tools concernts with few issues [3]. These tools are not usually integrated with data quality managment, therefore, in market the

tools for data quality estimated by 2022 to reach 1,376.7 Million from USD 610.2 Million in 2017, where the Compound Annual Growth Rate measured is 17.7%. The base year considered for this report is 2016 and the forecast period is 2017–2022 [4]. It's not only use of big data capabilities an organization required to collect values without mistakes, incomplete values besides errors but it is very often negated too. This kind of data is usually known to dirty data, and to clean this data can be challenging for companies who want to get better results. Cleaning data manually requires experience and often human tent to make mistake. Currently, machine learning is adopted in different area for process the tasks automatically, such as [5, 6] . Therefore, as machine learning can help any task to complete automatically it is possible to clean dirty data by training classification models.

2. BIG DATA ANALYTICS

The general procedure for obtaining visions from Big Data can be break down into five main stages [7]–[9] as shown in Figure 1.

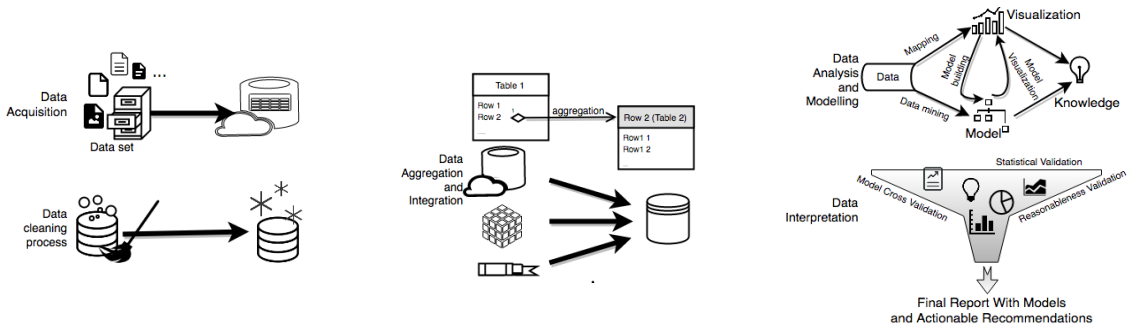


Figure 1. Processes for extracting insights from big data

Data Acquisition: Timeliness is one of the important requirement while data loading [10]. The fundamental characteristics of Big Data with its exponential rate of growing demands improve exceptional issue in Big Data engineering such as data acquisition and storing [7].

Data Mining and Cleansing: The most essential stage of processing big data is to implement a method to extract from loaded un-structured Big Data and mine-out the necessary data to able to coherent it in a typical and organized arrangement that will be easy to recognize. Data cleaning process is helps to clean dirty data.

Data Aggregation and Integration: The cleaned data obtained required to aggregate for processing these data by gathering and expressing into summary form [11], [12] following by integrating Data, to organize data from disparate sources by grouping of practical and business methods, and obtain meaningful and valued result [12].

Data Analysis and Modelling: From the viewpoint of Big Data, the goals are to produce business significance through the analysis of data which may fluctuate according to technique and data form. Construct and investigate meaningful reports to help the business for better and faster decision making.

Data Interpretation: Presenting data in understandable form for users, i.e. presenting data using analysis and modelling results to make decision by interpreting the outcomes and extracting knowledge. Data Interpretation queries are categorized together and indicate to the same table, diagram graph or other data demonstration options

3. DATA QUALITY PROBLEMS

The data cleaning process gets more complex when data comes from heterogeneous sources. Here, data quality problem has to be solved by data cleaning and data transformation. Despite of the various viewpoints on the effect of data quality, in the end, all have the probability to produce in economic expenses for groups. Some of survey of real case, involving economic costs due to dirty data, on a survey in 2014 its found that around $13.3 million dollars' annual costs in organizations and 3 trillion per year to US economy due to bad data. Another organization, the U.S. Postal Service, recognizes the cost of bad data, in 2013, an estimated amount of mail unsuccessful delivering to mentioned address was around 6.8 billion, which racks up to $1.5 billion in managing costs [13].

By some evaluations it is known that the in organizations and companies issue of dirty data already reached to epidemic amounts. The issue is equally prevalent and hypothetically equal beyond frightening in health care and other organization. [14]. For instance, in a telecommunication industry, dirty data has numerous costs. First and foremost, Experian approximates average 12% loses in business due to wrong records causing productivity reduction, resources wastage, and significantly, misused chances for marketing of cross-channel. The Experian investigation also focuses that approximately one-third of responders think that they waste almost 10% or more budget in marketing because of outcome obtained from inaccurate data. The Experian presents that 25% of survey participants in their research presently in their organization do not measures accuracy of data, where growths in telecoms and utilities companies to 33%, and in organizations like governments reaches to 36% [15].

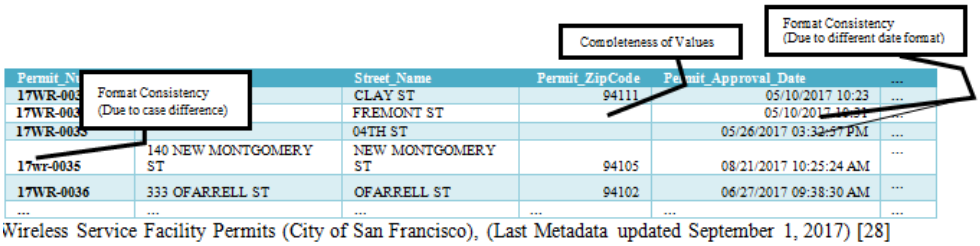
These measurements are within organizations, whereas observing external maters like marketing, marketers struggle with dirty data as well. Regarding to BizReport.com, "…marketers are generating a large portion of poor-quality leads, including those with improper formatting and even inaccuracies. Bad prospect information can have negative consequences, including wasted media investment, squandered resources, and poor customer experience, which marketers simply can't afford." [16]

In medical case, errors can able to kill patients or produce long lasting harm to heath of the patient. In 1999 an institute of Medicine reported [17] approximations, for instance, at least 44,000 to 98,000 people lost their lives each year for medical errors in hospitals only and which caused more $17 to $29 billion annually in healthcare costs. Other than heath issue, dirty data can also be involved in privacy issue for patients.

4. DATA QUALITY CRITERIA

Data quality is generally described as the capability of data to satisfy stated and implied needs when used under specified conditions [18]. Data accuracy, completeness and consistency are most popular initiatives to address Data quality [19],[20], beside other dimensions like Accessibility, Consistent representation, timeliness, Understandability, Relevancy, etc. [19]. Moreover, data quality is combination of data content and form. Where data content must contain accurate information and data form essential be collected and visualized in an approach that creates data functioning. Content and form are significant consideration to reduce data mistakes, as they illuminate the task of repairing dirty data needs beyond simply providing correct data.

Likewise, while developing a scheme to improve data quality it is essential to identify the primary reasons of dirty data. The causes are categories into organized and unintentional errors. The basic sources of producing systematic errors include while programming, wrong definition for data types, rules not defined correctly, data collection's rules violation, badly defined rules, and trained poorly. The sources of random errors can be errors due to keying, unreadable script, data transcription complications, hardware failure or corruption, and errors or intentionally misrepresenting declarations on the portion of users specifying major data. Human role on data entry usually result error, this error can be typos, missing types, literal values, Heterogeneous ontologies (i.e. Different nature of data), Outdated values or Violations of integrity constraints. Similarly, see Figure 2. as an example, where few data quality problems can be identified in the Wireless Service Facility Permits (City of San Francisco) database.



Wireless Service Facility Permits (City of San Francisco), (Last Metadata updated September 1, 2017) [28]

Figure 2. Data quality problems identified in an open dataset

Therefore, the most common dimensions of dirty data including data duplication are:

Inaccurate data refers to any field contains wrong values. A right value of data will bring accurate and signified arrangement of consistency and unambiguous.

Incomplete data from missing data is produced by data sets basically missing values. These type of data considered concealed when the amount of values identified in a set, but the values themselves are unidentified, and it is also known to be condensed when there are values in a set that are eliminated.

Inconsistent data is data redundancy; i.e. same data value is stored in different files which may be in different formats.

Duplicate data is entries that have been added by a system user same data multiple times

5. CLEANING TOOLS

Different vendors provide data cleansing solutions, includes Tal presents the website link of the company. Where, the "like (s)" and "dislike (s)" are obtained from Customers comments obtained from different websites, like end, IBM, SAS, Oracle and Lavastorm Analytics. There are some free tools been work on data transformation [21] [22], such as, OpenRefine, plyr, and reshape2, although it is uncertain whether they can execute Big Data. Another well-known tool is ETL tools, which provides complex data conversion techniques by merging and repairing data [23]. A summarization of some available commercialized tools to manage Data Quality in presented in Table 1. Where the "Vendor" field mentions the company offering the tools and "Product" mentions the tool offered by the vendor for managing Data Quality. "Website" column [24], [25].

Table 1. Comparison of Commercialized Data Quality Management Tools

Vender	Product	Website	Like (s)	Dislike (s)
Trifacta	• Trifacta Data Wrangler	trifacta.com	intelligently recognizes imported data file and provides prescriptive methods	formula based
Informatica IDQ	• Data Quality Standard Edition • Address Validation Services and StrikeIron • Data Quality Advanced Edition • Data Quality Governance Edition	informatica.com	Easy interact with provided interface to identify the functions, Ease of Data Migration, Completely on cloud	It requires SQL knowledge
SAP	• Information Steward • Data Quality Management • SAP Data Services	go.sap.com	Ability to recognise organization's needs	No option to control Source code and integrate
Melissa Data	• Data Quality Components for SSIS • Personator • Global Data Quality Suite • Global MatchUp	melissadata.com	API's are easy and straightforward Able to use phoenics for address corrections	No better documentation, Performance is slow for real time queries.
BDNA	• Melissa Listware • BDNA Technopedia • BDNA Normalize • Technopedia	bdna.com	Append contacts, Standardize Address, Simple interface good coverage of vendors and products, Proactive in keeping their packs up to date, Adopt maturing technologies with manageable risk	No Mac OS integration Need to point out stale data, it will not refresh for months to years.
SAS	• Data Management • Data Quality Desktop	sas.com	The learning curve is manageable.	Needs training and education to use, no command window
Experian	• Capture, Clean and Enhance data quality tools • Experian Pandora • Experian Data Quality Platform	experian.com	Low cost and flexibility of use with various file formats.	
Pitney Bowes	• Spectrum Technology Platform • Code-1 Plus	pitneybowes.com	User interface is quite friendly and attractive, Can create APIs without programming	Hard to integrate and handle large amounts of data
CRMfusion	• DemandTools • CRMfusion PeopleImport	crmfusion.com	Manage large scale data. in real time Standardize, cleanse and overall manipulate data	Unable to enter a custom SOQL (e.g. with a subquery) as the basis for the data pulled down
Oracle	• Oracle Enterprise Data Quality	oracle.com	Profiling customers easily Great for batch-oriented processing	emphases in ETL instead of data context and management, not good for real-time processing
IBM	• Infosphere QualityStage • Infosphere Information Analyzer • InfoSphere Information Server	ibm.com	The lineage integrates metadata from Cognos, Datastage, Quality Stage, and Oracle Metadata.	
Addressy	• Addressy	addressy.com	Only a simple JavaScript snippet is required on the page the rest of the configuration can be done via the control panel	

6. BIG DATA ANALYTICS DATA CLEANING CHALLENGES

Generally, the data gathered will not be in a ready form for analyzing. For instance, consider data obtained from Telecommunication stored system, consisting of feedback obtained from different agents and structured data from routers. It is challenging to analyze such types of unstructured data. Requirement of extraction procedure that recovers necessary data from various sources and demonstrates it in a structured arrangement appropriate for analysis is compulsory. Data cleaning is an essential portion of data analysis and challenging too [26]. Researcher from data base research community offered few challenges to obtain useful data from big data [27], [28]. This is challenging through every data analysis, but after involving the variety and voluminous big data, it transforms even beyond pronounced. The data quality required to assured for accurate and correct data visualization. To deal this issue, organization require to overcome some common challenges:

6.1. Scalability

Cleaning techniques required scaling data capacities as quickly increasing data size of Big data, which is quite challenging. Existing procedures involve jamming data for identical data detection [29], [30], identification and linkage for data cleaning [30], clean data using sampling [31], and distributed data cleaning [32].

6.2. Semi Structured and Unstructured Data

Big data is usually set of variety of data, which may be populated with semi structured layout data e.g. in XML/JSON and unstructured format data e.g. in word-processing files, in e-mail besides in text fields in databases. Semi structured and unstructured data remain mostly unfamiliar for Data quality problems [28, 33].

6.3. User Engagement

While much research work was involved humans to execute deduplication process in data set. For instance, through active learning, including human expert in other to clean data [30], like getting user response to determine rules for data quality, is still to be discovered.

6.4. Raising Privacy and Security Interests

While cleaning data the most common task is to observe and examine complete set of raw data value which may be restricted by some domain is a significant challenges [9], like telecommunication, medicine and finance. For example, telecommunication data, such as the Internet connection login sessions log collected over an extensive period of time can reveal an individual's location and behavior, as shown in Figure 3.

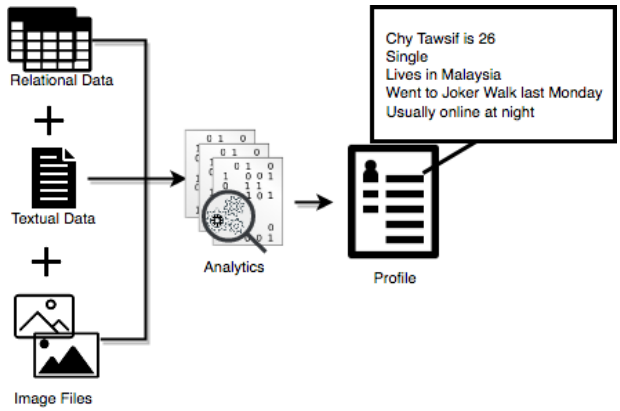


Figure 3. Information gathered from running analytics on data and files to create Tawsif's profile

6.5. Computational Complication for Data Streaming

Huge data collection from variety of sensors and user devices is always an interesting issue. Gartner, Inc. forecasted in 2017, that 8.4 Billion devices will be linked things and used in global in 2017, up 31%

from 2016, and will reach 20.4 billion by 2020 [34]. This is the reason data cleansing actions may engage huge processing power.

6.6. Machine Learning and Other Algorithms

Lastly, it known that big data analytics is still in its initial periods of development as a technical discipline. Hence many Machine Learning algorithms usable to scale big data sets or unable to tolerate the noises and gaps produced by real world [35]-[38]. There is still further research going to to improve these algorithms that will be more suitable with real world conditions which may contain millions and trillions of components for data cleaning.

6.7. Manually

Currently, after benefit of histograms, conversation tables and rules with algorithms individual interference is nevertheless compulsory to recognize and repair the data [30], [39].

7. MACHINE LEARNING PARADIGMS FOR BIG DATA CLEANING

Currently there are different types of learning paradigms available in machine learning; but, not all types applicable to all field. For instance, [40] presented a cleaning approach using Data mining and SVM (a machine Learning Paradigm). Machine Learning techniques can be used to teach the system and complete the task my minimum human interaction. It may reduce the time and resources required to analyze and transform dirty data to usable clean data. Machine Learning techniques are used to make system intelligent by learning capability. Data can be classified by three ways, un-supervised, supervised and semi supervised methods. Selection of algorithms must be dependent on the size, quality, and nature of the data. Some common learning algorithms can be used to clean data are shown in Figure 4.

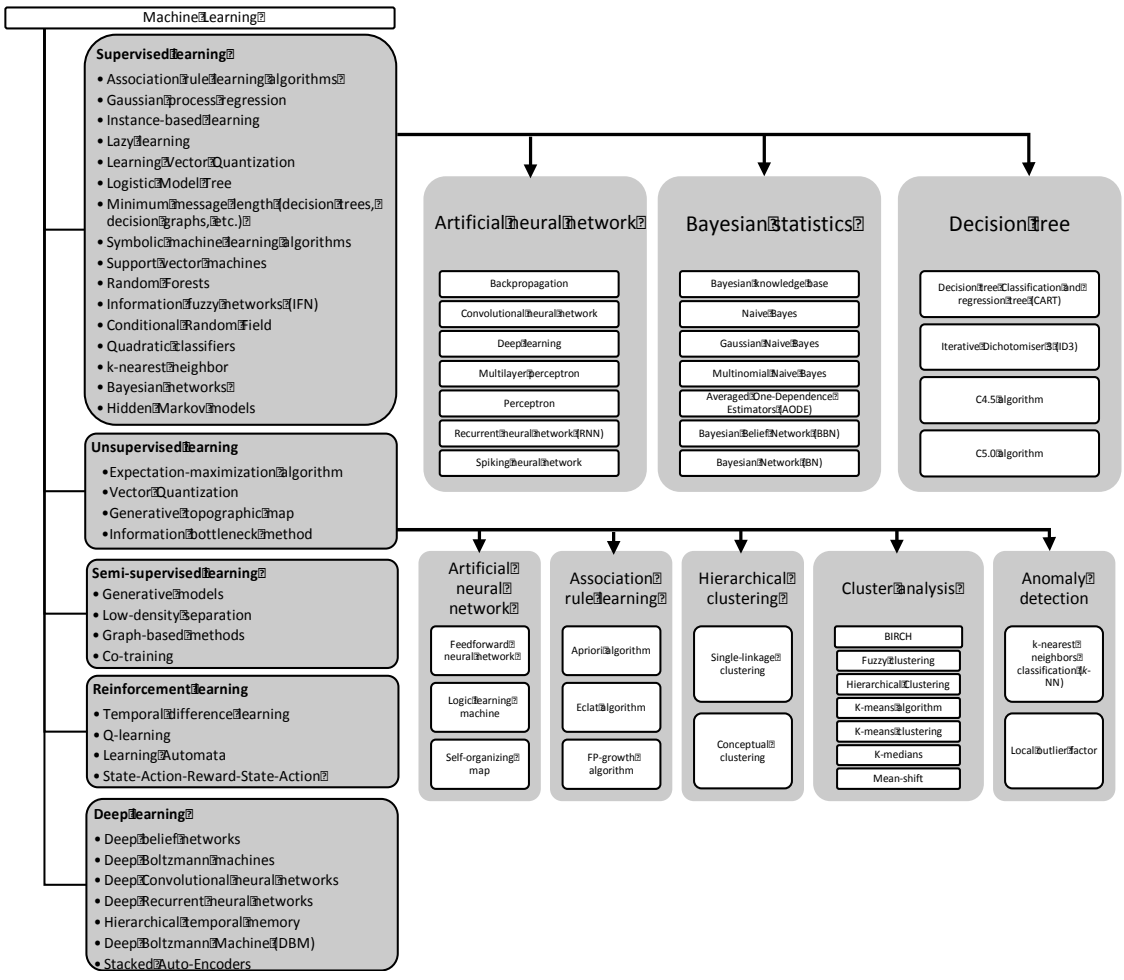


Figure 4. Machine learning algorithms

7.1. Deep Learning

This technique is widely used by data representation, rather than data features to execute data cleaning. Deep Learning Algorithms transforms data into abstract representations that allows learning features. Hence, there is no requirement for feature extraction as the features are learned right from the data. Due to nature of Big data, the capability to ignore feature extraction step is great deal.

7.2. Naïve Bayes Classifier Algorithm

This algorithm provides classification parameter and attributes to label the occurrences must be conditionally independent, if the instance contains several attributes. This algorithm is suitable for moderate or large training data set.

7.3. K-Means Clustering Machine Learning Algorithm

K-Means produces stronger clusters than hierarchical clustering in case of globular clusters. And for large number of variable K-Means clustering executes speedier than hierarchical clustering.

7.4. Apriori Algorithm

Apriori Algorithm is easy to implement and can be parallelized easily. Which uses large item set properties to implement.

7.5. Random Forest Machine Learning Algorithms

Random Forest is very less robust to noise, which makes it more efficient and versatile for classification and regression jobs. It is easy to define which parameters to use, since it's not delicate to the parameters required to run. This algorithm can be grown in parallel and efficient for large database with higher classification accuracy.

8. CONCLUSION

In recent years, probably big data processing brought the greatest revolution in computing. The data cleaning of massive sizes of data lies at the heart of big data analytics processing for all purpose of domains for better data investigation.

In this paper, an overview is initiated to identify the potential of data cleaning in big data analytics in the process of gathering, arranging and processing information. It is important to understand data quality criteria of dirty data to able to clean data sets without failure. A comparison of commercialized tools is presented by obtaining comments from different customers. Most of the tools mostly concerns to organize data sets and clean messy data and very methods uses machine learning. But they didn't give much importance to big data characteristics, which may lead to big challenge while cleaning data. There are many available data repairing algorithms, still it required human expert to take intelligent decision if the cleaning process is correct or not. Machine learning algorithms will probably replace most jobs in the world, with the fast evolution of big data and accessibility of programming tools like Python and R , machine learning is increasing mainstream existence for data scientists. Machine learning applications are highly automated and self-modifying which continue to improve over time with minimal human intervention as they learn with more data.

This survey has prompted us to conduct additional real-world evaluations and develop a modified framework of big data analytics by changing structure of cleaning phase to get more clear visions of data. It is expected to produce a new plan regarding the structure of data quality techniques which can be more efficient in big data analytics.

REFERENCES

[1] IBM: "10 Key Marketing Trends for 2017 Customer Expectations". (2017).
[2] Cisco: White Paper, Cisco Global Cloud Index : Forecast and Methodology, 2015–2020. (2016).
[3] Khan, N., Yaqoob, I., Abaker, I., Hashem, T., Inayat, Z., Kamaleldin, W., Ali, M., Alam, M., Shiraz, M., Gani, A., "Big Data : Survey , Technologies , Opportunities , and Challenges". (2014).
[4] Marketsandmarkets.com: Data Quality Tools Market. (2017).
[5] Xu, H., Zhang, R., "Research on Data Integration of the Semantic Web Based on Ontology Learning Technology". *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 2014; 12: 167–178.
[6] Khan, M., Pradeepini, G., Machine Learning Based Automotive Forensic Analysis for Mobile Applications Using Data Mining. *TELKOMNIKA (Telecommunication, Computing, Electronics and Control).* 2015; 16(2): 350–354.

[7] Wang, Y., Victor J., W., "Big Data Analytics on the Characteristic Equilibrium of Collective Opinions in Social Networks Big Data Analytics on the Characteristic Equilibrium of Collective Opinions in Social Networks". *Int. J. Cogn. Informatics Nat. Intell.* (2014).
[8] Ext, A.P.F.O.R.T., Udio, A., Ideo, V.: Big Data Analytics : Challenges And. 5, 41–51 (2016).
[9] Erl, T., Khattak, W., Buhler, and P.: Big Data Fundamentals:Concepts, Drivers & Techniques: Book.
[10] He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: *RCFile : A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems.* In: ICDE Conference 2011. pp. 1199–1208 (2011).
[11] Munshi, A.A., Mohamed, Y.A.I.: Big data framework for analytics in smart grids. *Electr. Power Syst. Res.* 151, 369–380 (2017).
[12] Zhou, X., Jin, Q., Wu, B., Wang, W., Organic Streams, "Data Aggregation and Integration Based on Individual Needs". In *2013 International Joint Conference on Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA)*. pp. 535–541. IEEE, Aizu-Wakamatsu, Japan (2014).
[13] Bernardino, J., Laranjeiro, N., Soydemir, S.N., Bernardino, J., A Survey on Data Quality : Classifying Poor Data A Survey on Data Quality : Classifying Poor Data. In: The *21st IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2015)*, (2015).
[14] VHAInc: The Cost of Dirty Data. (2012).
[15] Experian: Unlocking the power of data : the cost of dirty data and how to improve its accuracy Foreword. (2011).
[16] Kristina Knight: Study: Dirty Data a problem for marketers, http://www.bizreport.com/2015/01/study-dirty-data-a-problem-for-marketers.html, (2015).
[17] MEDICINE, I.O., TO ERR IS HUMAN: BUILDING A SAFER HEALTH SYSTEM. (1999).
[18] Sidi, F., Hassany, P., Panahy, S., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A., "Data Quality : A Survey of Data Quality Dimensions". In: *2012 International Conference on Information Retrieval & Knowledge Management (CAMP)*. pp. 300–304. IEEE, Kuala Lumpur, Malaysia, Malaysia (2012).
[19] Juddoo, S., "Overview of data quality challenges in the context of Big Data". In *2015 International Conference on Computing, Communication and Security (ICCCS). IEEE*, Pamplemousses, Mauritius (2015).
[20] Taleb, I., Kassabi, H.T. El, Serhani, M.A., Dssouli, R., Bouhaddioui, C., "Big Data Quality : A Quality Dimensions Evaluation". In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*. pp. 759–765. IEEE (2016).
[21] Beeharry, Y., Fowdur, T.P., Hurbungs, V., Bassoo, V., Ramnarain-Seetohul, V., "Analysing transportation data with open source big data analytic tools". *International Journal on Electrical Engineering and Informatics (IJEEI).* 2017; 5(2): 174–184 .
[22] Naumann, F., Herschel, M.: Data Fusion in Three Steps : Resolving Inconsistencies at Schema , Tuple- , and Value-level Data Fusion in Three Steps : Resolving Inconsistencies at Schema- , Tuple- , and Value-level. (2006).
[23] Michel, P., Dmitriyev, V., Abilov, M., Marx, J., "ELTA : New Approach in Designing Business Intelligence Solutions in Era of Big Data". 16, 667–674 (2014).
[24] gartner.com: Home page gartner.com, www.gartner.com.
[25] G2crowd: Home page of g2crowd, www.g2crowd.com.
[26] Jin, X., Wah, B.W., Cheng, X., Wang, Y., "Significance and Challenges of Big Data Research". *Big Data Res. 2*, 59–64 (2015).
[27] Garg, N., Singla, S., Jangra, S., "Challenges and Techniques for Testing of Big Data". *Procedia - Procedia Comput. Sci*. 85, 940–948 (2016).
[28] Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V., "Critical analysis of Big Data challenges and analytical methods". *J. Bus. Res*. (2016).
[29] Dupare, J.M., Sambhe, N.U., "A Novel Data Cleaning Algorithm Using RFID and WSN Integration 1". (2015).
[30] Liu, H., Tk, A.K., Thomas, J.P., "Cleaning Framework for Big Data — Object Identification and Linkage". (2015).
[31] Wang, J., Krishnan, S., Franklin, M.J., Goldberg, K., Milo, T., Kraska, T., Berkeley, U.C., "A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data". In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. pp. 469–480. , Snowbird, Utah, USA (2014).
[32] Khayyat, Z., Ilyas, I.F., Madden, S., "BigDansing : A System for Big Data Cleansing". In *SIGMOD'15*. pp. 1–16. , Melbourne, Victoria, Australia (2015).
[33] Labrinidis, A., Jagadish, H. V: "Challenges and Opportunities with Big Data". 2032–2033.
[34] Egham: Gartner Says 8.4 Billion Connected "Things" Will Be in Use in 2017, Up 31 Percent From 2016. , U.K. (2017).
[35] Pruengkarn R., Wong K.W., F.C.C., "Data Cleaning Using Complementary Fuzzy Support Vector Machine Technique". In *Neural Information Processing. ICONIP 2016* (2016).
[36] Procedures, P., Liu, G., Yu, H.S.D., "Prediction of Protein – Protein Interaction Sites with Machine". In *The Journal of Membrane Biology. Springer US* (2015).
[37] Chen, Y., He, W., Hua, Y., Wang, W., "CompoundEyes : Near-duplicate Detection in Large Scale Online Video Systems in the Cloud". In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications CompoundEyes*: pp. 1–9. IEEE (2016).
[38] Wang, Y., Li, Q., "Review on Studies and Advances of Machine Learning Approaches". *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)* 12, 1487–1494 (2014).
[39] Høverstad, B.A., Tidemann, A., Langseth, H., "Effects of Data Cleansing on Load Prediction Algorithms". In *2013 IEEE Computational Intelligence Applications in Smart Grid (CIASG)*. pp. 93–100. *IEEE*, Singapore (2013).

[40] Natarajan, K., Li, J., Koronios, A., Science, I., Mining, D., "Cleaning, D.: Data mining techniques for data cleaning". In *Proceedings of the 4th World Congress on Engineering Asset Management*. pp. 796–804 (2009).

BIOGRAPHIES OF AUTHORS

Siti Nadiah Che Azmi is currently doing her Master of Philosphy degree in Electrical Engineering at Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Malaysia. She has completed her Bachelor of Applied Science (Electronics and Instrumentation Physics) from Universiti Malaysia Terengganu, Malaysia at 2013.

Dr. Jakir Hossen is graduated in Mechanical Engineering from the Dhaka University of Engineering and Technology (1997), Masters in Communication and Network Engineering from Universiti Putra Malaysia (2003) and PhD in Smart Technology and Robotic Engineering from Universiti Putra Malaysia (2012). He is currently a Senior Lecturer at the Faculty of Engineering and Technology, Multimedia University, Malaysia. His research interests are in the area of Artificial Intelligence (Fuzzy Logic, Neural Network), Inference Systems, Pattern Classification, Mobile Robot Navigation and Intelligent Control.

Dr. Md Shohel Sayeed obtained the B. Sc. Ag. (Hons) from Bangladesh Agricultural University. He completed his M.Sc. (IT) from Universiti Kebangsaan Malaysia (UKM) and Ph.D. in Engineering from Multimedia University, Malaysia. At present, he is holding a position of Associate Professor at the Faculty Information Science and Technology, Multimedia University, Malaysia. His main research interests are Biometrics, Pattern Recognition, Signal and Image Processing, Big Data and Data Mining.

Dr. Ho Chin Kuan obtained the B. Sc. (Hons) in Computer Science with Electronics Engineering from University College London, UK. Subsequently, he completed his M.Sc. (IT) and Ph.D. in Information Technology from Multimedia University, Malaysia. At present, he is a Professor and Dean at the Faculty of Computing and Informatics, Multimedia University, Malaysia. His main research interests are Natural Computing, Combinatorial Optimization and Data Mining.

Chy. Mohammed Tawsif K. currently a postgraduate student in Engineering Program and researching with Artificial Intelligence, Event Processing and Big data from Multimedia University (MMU). He pursued bachelor's degree in Computer Science and Engineering from International Islamic University Chittagong, Bangladesh.

Md. Armanur Rahman received the B.Sc. degree in computer science and engineering from Asian University of Bangladesh (AUB) in 2010. He is currently working toward the MEngSc degree at the Multimedia University (MMU), Malaysia. His research interest includes performance optimization of big data system, data mining, machine learning and image processing.

Md Arif Hossain currently a postgraduate student in Engineering specializing in Solar Energy Technology from Multimedia University (MMU) in Malaysia. He has completed bachelor‟s degree in Electrical & Electronic Engineering from Bangladesh University, Dhaka, Bangladesh. His Research Interest is Electrical Engineering, Renewable Energy, Wind Energy, Controller, Fuzzy logic, Neural Network and Intelligent System.