Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 12, pp.43548-43553, December, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

BIG DATA ANALYTICS APPLICATIONS: A COMPREHENSIVE REVIEW

***Mukesh Shukla, Dr. P.K. Rai and Rinku Singh**

A.P.S. University, Rewa (MP) 486003, India

ARTICLE INFO

Article History:
Received 23rd September, 2016
Received in revised form
10th October, 2016
Accepted 29th November, 2016
Published online 30th December, 2016

Key words:

Big Data, Map Reduce,
Hadoop, Hive, Pig,
Spark, Data Mining.

ABSTRACT

Big data describes the collection of new information which must be made accessible to high numbers of users close to real time, based on enormous data inventories from multiple sources, with the goal of speeding up critical competitive decision-making processes. Enormous amounts of data have become available on hand to decision makers which is making analysis and decision making task much more challenging and tedious. Considering the sheer volume and variety of data, the analyses, predictive exploration of situations and business intelligence workloads are beyond the capabilities of traditional tools & solutions. In recent years a number of Big Data tools& solutions have arisen to handle these massive quantities of data. The objective of this paper is to analyze the scope of big data in different fields and it's ability to revolutionize individual area for enhancing the decision making process. This objective is considered via wide ranging review of literature.

Citation: Mukesh Shukla, Dr. P.K. Rai and Rinku Singh, 2016. "Big Data Analytics Applications: A Comprehensive Review", *International Journal of Current Research*, 8, (12), 43548-43553.

INTRODUCTION

Big Data refers to data that exceeds the typical processing, storage and computing capacity of traditional databases and techniques used for data analysis.One of the buzzwords in IT during the last few years is 'Big Data'. Organizations which had to handle the fast growth data, they initially shaped it for processing data resulting from scientific or business simulations, web data or data from other sources.Fundamental business model of some of those companies are rely on indexing and using this large amount of data. Google developed the the Google File System and MapReduce for handling the sheer volume data available on web.These technologies are available as open source software as Apache Hadoop and the Hadoop File System. All these efforts laid the foundation for technologies summarized today as 'big data'. Later big giants IBM Oracle, HP, Microsoft,SAS and SAP in information management field stepped in and invested to extend their business and build new products especially aimed at Big Data analysis. At the same time many start-ups like Cloudera entered the scene. Considering the trends analysts expectBig Data impact onto business and the praise they sing on 'big data', it was obvious for these big players to get part of the big data. As per the IDC predictions digital data created and consumed per year will grow up to 40.000 exabyte by 2020, from which a third 2 will potentially valuable to organizations if processed using big data technologies.

****Corresponding author: Mukesh Shukl,***
A.P.S. University, Rewa (MP)486003, India.

IDC also declared that in 2012 only 0.5% of potentially valuable data were examined, calling this the 'Big Data Gap'. McKinseyGlobal Institute also predicts 40% annual growth in global data per year. They describe big data trends in terms of monetary figures and see Big Data market of 300 billion $ in US health care sector and 250 billion in European public sector and a potential improvement of margins in the retail industry by 60%. As a resource, there is need of Big Data tools and methods that can be used to analyze and extract patterns from large-volume data. Increased computational processing power, increased data storage capabilities and availability of increased volumes of data are the major features of Big Data. The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such large amounts of data need to be properly analyzed, and pertaining information should be extracted. The contribution of this paper is to offer an analysis of the available literature on application of Big data analytics concepts in various sectors through available implementing tools. This study presents comprehensive survey of big data attributes with discussion of some of the various big data tools, methods, and technologies which can support future need of discovering knowledge from massive data. Application of Big data analytics and its impact on individual sectors are portrayed.

Big Data Analytics Applications

Academicians, Industry R&D experts and other prominent stakeholders positively agree that big data has become a big

game changer in most, if not all, types of application domain over the last few years. With this in mind, having a bird's eye view of big data and its application in different areas help us better appreciate what will be the direction and trends of future research across different domains. In this section, I discuss literatures which examine different verticals that are using big data and how big data solves domain specific challenges.

Future Healthcare and Biology

The healthcare sector has access to massive amounts of data but has been plagued by failures in consuming the data to control the growing cost of healthcare and by inefficient systems that stifle faster and better healthcare benefits across the board. Main reason behind this is electronic data is unavailable, inadequate, or unusable. Additionally, the healthcare databases that hold health-related information have introduced complexity to link data that can show patterns useful in the medical field. Agarwal *et al.* 2016 made an attempt to create a patient wellness score which integrates many lifestyle components and a holistic patient prospective. He proposed models which are built combining both medical professional input and machine learning algorithms using a comprehensive survey conducted by the Centers for Disease Control and Prevention. Models comparisons result display that 8 out of 9 models are shown to have a statistically significant ($p = 0.05$) increase in area under the receiver operating characteristic when using the hybrid approach when compared to expert-only models. Patient friendly output was achieved by aggregation and linear transformation of models. The resulting predictive models provide a comprehensive numerical assessment of a patient's health, which may be used by healthcare providers to track patient wellness to help maintain or improve their current condition. Etani *et al.* 2015 proposed that two processes of "Big data analytics" and "Implementation of data modeling" should be collaborated with Model-driven architecture (MDA). Data modeling with those two processes in Model-driven architecture should be repeated in order to validate theData model and to find a new data resource for a service. They predicted side effect of drug which is one of screening methods in drug discovery. Their prediction model is constructed with data mining methods at the intersection of statistics, machine learning and database system. They confirmed that the prediction model and the data model for drug discovery are implemented as a prototype system to verify those models and their practicality.

Herland *et al.* 2014 presented recent research using Big Data tools and approaches for the analysis of Health Informatics data collected at multiple levels, including the tissue, molecular, patient, and population levels. In addition to collecting data at multiple levels, multiple levels of queries are addressed: human-scale biology, clinical-scale, and epidemic-scale. Siuly *et al.* 2016 explored the challenges of medical big data handing and also introduced the concept of the CAD system how it works. Their paper also provides a survey of developed CAD methods in the area of neurological diseases diagnosis. Their study helpful for the experts to have some idea and understanding how the CAD system can assist them in this point. Toga *et al.* 2015 provided a framework for developing practical and reasonable data sharing policies that incorporate the sociological, financial, technical and scientific requirements of a sustainable Big Data dependent scientific community. In their study they found many biomedical and healthcare studies may be considerably impacted by using large, heterogeneous and incongruent datasets; however there are significant social, technical, regulatory, and institutional obstacles that need to be overcome to ensure the power of Big Data overcomes these detrimental factors. Kumar *et al.* 2014 presented the database-centric molecular simulation (DCMS) system. The main idea behind this system is to store MS (Molecular Simulation) data in a relational DBMS to take advantage of the SQL, data access methods, query processing, and optimization mechanisms of modern databases. A unique challenge is to handle the analytical queries that are often compute-intensive. For that, they developed novel indexing and query processing strategies as integrated components of the DBMS. As a result, researchers can upload and analyze their data using efficient functions implemented inside the database. Index structures are generated to store analysis results that may be thought-provoking to other users, so that the results are readily available without duplicating the analysis. They developed a prototype of this system based on the PostgreSQL system and experiments using real MSdata and workload show that it significantly outperforms existing MS software systems. They also used it as a platform to test other data management issues such as security and compression.

Transportation

Discovering novel ways to manage and analyze big data to create value would increase the accuracy of predictions, improve the management and security of transportation infrastructure and enable informed decision-making. It is these challenges that may drive new insights and opportunities and transform the way we perceive transportation and traffic engineering phenomena. Discovering innovative ways to manage and analyzebig data to create value would increase the accuracy of predictions, increase the management and security of transportationinfrastructure and enable informed decision-making. These are the new challenges that drive new insights and opportunitiesand transform the way we perceive transportation and traffic engineering phenomena. Big Data has been speedily growing into the transportation arena. However, the methods, models and algorithms which are used today in our domain to mine and explore data think of estimation, prediction, validation of traffic and transportation theories and models – may not scale and/or perform well under these new conditions. Big data may play more critical role in transportation planning, traffic operations and safety.

Berengueres *et al.* 2014 performed a case study on an airline's miles program resource optimization. The airline had a large miles loyalty program but was not taking benefit of advanced data mining concepts. For example, to predict if in the coming month(s), a new passenger would become a privileged frequent flyer or not, a linear extrapolation of the miles earned during the last months was used. This data was then used in CRM interactions between the airline and the passenger. Extrapolation correlation with whether a user would attain a privileged miles status was 39% when 1 month of data was used for a prediction. In contrast, when GBM and other blending techniques were used, a correlation of 70% was realized. This matched to a prediction accuracy of 87% with less than 3%false positives. The accuracy reached 97% if 3 months of data instead of one were used. An application that ranks users according to their probability to become part of privileged miles-tier was suggested. The application performs

real time allocation of limited resources such as vacant upgrades on a given flight. Now those resources can be allocated to high potential passengers without extra cost which is increasing the perceived value of the program. Kumar *et al.* 2015 proposed a framework for analyzing accident patterns for different types of accidents on the road which makes use of K modes clustering and association rule mining algorithm. Their study uses 11,574 accidents as sample that have occurred on Dehradun district road network between 2009-2014. Attributes accident type, road type, lightning on road and road feature used in K modes clustering to determine six clusters. Association rule mining has been applied on each cluster as well as on EDS to generate rules. Strong rules with high lift values are taken for the analysis. Rules for each cluster expose the circumstances linked with the accidents within that cluster compared with the rules generated for the EDS and comparison shows that association rules for EDS does not expose proper information that can be linked with an accident. More information can be identified if more feature are available that is associated with an accident. They also performed trend analysis of all clusters and EDS on monthly and hourly basis. Their methodology is supported by trend analysis result that performing clustering prior to analysis advantages in identifies improved and beneficial results that we unable to obtain without using cluster analysis.

Kumar *et al.* 2016 have proposed a framework to analyze road accident time series data that takes 39 time series data of 39 districts of Gujrat and Uttarakhand state of India. Proposed framework parts the time series data into multiple clusters. They suggested a time series merging algorithm to find the representative time series for each cluster. This RTS algorithm is also used for trend analysis of multiple clusters. The result discloses that road accident trend is going to surge in certain clusters and these districts should be the key concern to take preventive measure to overcome the road accidents. Kumar *et al.* 2016 have proposed in his study a method to analyze hourly road accident data using Cophenetic correlation coefficient from Gujarat state in India. The purpose of their study is to provide an effective and efficient way to pick the best suitable distance metric to cluster the series of counts data that provide an improved clustering result. The result displays that the suggested method is efficiently group the different districts with similar road accident patterns into single cluster or group which can be further used for trend analysis or similar tasks.

Communications, Social media and Sentiment Analysis

Social sentiment analysis is one of the hottest solution area for Big Data Analytics. These Big data solutions are widely being used in marketing and sales to gain insights into what customers feel about their product and services. Unstructured communication available on social media can be used to get business valuable insight through Big Data Analytics. Zhan *et al.* 2016 provided a technique to identify the top-K communities, based on the average Katz centrality of all the communities in a network of communities and the distinctive nature of the communities. These top-K communities can be used to spread information efficiently into the network, as these communities are capable of influencing neighboring communities and thus spreading the information into the network effectively and efficiently. Yang *et al.* 2016 provided a complex framework for Big Data processing that cannot be achieved with a single-machine utilizing average hardware and software. They introduced three modules that are capable of crawling raw online records, producing key features to characterize original samples in suitable and useful ways, and then running an association rule-mining algorithm on clouds for added content mining. The suggested framework is implemented on Hadoop platform, which is used as the fundamental tool for storing and processing harvested data sets. Thirteen high-level features are produced from three categories film details, reviews, and user profiles using aggregation functions, and the data is further quantified using the description set. More importantly, they provided an improved parallel Apriori algorithm for discovering significant correlations among these thirteen key features, with a view to expanding the analytical methods to a larger data set.Suggested framework offers efficient applicability and a flexible capability to process the large amounts of social media data that in turn can be fed back to producers and distributors of both commercial and user-generated digital media contents.

Chopade *et al.* 2015 discussed community dynamics and reviewed complex network structural parameters. They emphasized the importance of network centrality or degree centrality and network robustness for community detection. Centrality is correlated with degree. They discussed network or degree centrality (weighted Laplacian centrality) based on altered Laplacian, weighted micro-community centrality. They also suggested and introduced algorithm for k-clique sub-community for weighted modularity optimization and overlapping community discovery based on degree and weighted micro community centrality. These new matrices and algorithms are useful in detecting hidden level susceptibilities. They examined real-world extensive complex networks and carried out evaluation of different community detection algorithms. Their results showed certain relationship between degree centrality and modularity optimization. Network centrality and robustness will benefit for supervised community detection in overlapping communities. Recommended algorithms are useful for finding communities of densely linked vertices in network data. Sentiment analysis or opinion mining is a field of study that examines people's sentiments, attitudes, or emotions towards certain entities. Fang *et al.* 2015 presented a paper to tackle a major problem of sentiment analysis, sentiment polarity labeling. Online product reviews from Amazon.com are picked as data used for this study. A sentiment polarity categorization procedure has been suggested along with detailed descriptions of each step. Experiments for both sentence-level classification and review-level classification have been accomplished.

Yasseri *et al.* 2016 developed theoretically informed methods for election prediction based on information seeking behavior on Wikipedia, responding to existing critiques of predictions generated from new sources of socially generated data. They applied these approaches to a variety of different European countries in the context of 2 different European elections. They formed three main experiential findings. First, that the relative change in the number of page views to the general Wikipedia page on the election can offer a realistic estimation of the relative change in gathering for that election at the country level. This backs the idea that increases in online information looking for at election time are driven by voters who are seeing voting in the election. Second, that a theoretically learned model based on previous national results, Wikipedia page views, news media mentions, and basic information about the political party in question can offer a good prediction of the overall vote share of the party in question. However, the

Wikipedia variable itself was of relatively minor significance in this prediction. Third, they offered a model for predicting change in vote share (i.e., voters swinging towards and away from a party).They showed that Wikipedia page view data provided for an significant surge in predictive power in this context. We also showed, however, that this relationship is exaggerated in the case of newer parties.

Other: Finance, Telecom, Manufacturing, E-Governance, Learning

Big Data Analytics is making read difference in various other fields like finance, telecom, manufacturing, e-governance, learning etc. Application of Big Data Analytics in these sectors has gained lot of attention from researchers. Lu *et al.* 2014 proposed a virtualized web map service system, v-TerraFly, and its autonomic resource management in order to address this challenge. Virtualization helps the deployment of web map services and increases their resource utilization through encapsulation and consolidation. Autonomic resource management permits resources to be automatically provisioned to a map service and its internal tiers on demand. Specifically, this paper recommends new techniques to predict the demand of map workloads online and optimize resource allocations considering both response time and data newness as the QoS target. The offered v-TerraFly system is prototypedon TerraFly, a production web map service, and assessed using real TerraFly workloads. The results display that v-TerraFly can accurately predict the workloaddemands: 18.91% more precise and efficiently assign resources to meet the QoStarget. It shows improvement in the QoS by 26.19% and 20.83% resource usage are saved as compared to traditional peak-load-based resource allocation.

Zang *et al.* 2014 conducted comparative study between the incremental learning and ensemble learning methods. They first presented the concept of "concept drift", and suggested how to quantitatively measure it. Then, they evoke the history of incremental learning and ensemble learning, introducing milestones of their developments. In tests, they systematically compare and examine their performances w.r.t. accuracy and time efficiency, under various concept drift scenarios. Bughin J. (2016) in his study extended the past work on the effects of big data on corporate performance. His main innovation is to develop an approach to corporate performance and to the production function, that permits us to answer more directly questions such as the complementarily of big data capital and labour.He observed that the major performance effect of big data resides in the close complementarily between big data IT investment and labour skills. The performance impact is also slightly higher for application domains such as business intelligence and customer interface. Weiss *et al.* 2016 have formally defined transfer learning in their survey paper and presented information on current solutions, and reviewed applications applied to transfer learning. Their paper offers solutions from the literature indicating current trends in transfer learning. Cerchiello *et al.* 2016 presented a model for the estimation of systemic risk models using two different data sources: financial markets and financial tweets, and a proposal to combine them, using a Bayesian approach. In their research they have proved how big data and, specifically, tweet data, can be usefully employed in the field of financial systemic risk modeling. (Singh, 2014) focused in his paper to analyze the growing need of Big Data technology in financial domain, especially in Capital Markets. Why Capital market sector is keen to leverage this technology, what all benefits could be earned? It also highlights the existing modules where big data is already in use, areas where implementation is ongoing and also the difficulties in the path of implementation of the same. Barrachina *et al.* 2014 presented in a complete open source solution for processing and categorization of similar service calls within large technical supports data sets to allow for identification of similar calls with potential for faster resolution. The solution was examined using a subset of VMware technical support data with the output and accuracy of the five commonly employed clustering algorithms. Although, this paper presents the analysis of VMware support data in particular, the preposed techniques and procedures are generally applicable to other organizations providing similar services, thereby providing a proof of concept Industry framework.

O'Donovan *et al.* 2015 The main contributions of his research are a set of data and system requirements for implementing equipment maintenance applications in industrial environments, and an information system model that provides a scalable and fault tolerant big data pipeline for integrating, processing and analyzing industrial equipment data. These offerings are considered in the context of highly regulated large-scale manufacturing environments, where legacy and emerging instrumentation must be supported to ease initial smart manufacturing efforts. They addressed the main challenges and required characteristics linked with large-scale data integration and processing in industry, such as automating and simplifying data ingestion, embedding fault tolerant behavior in systems, promoting scalability to manage large quantities of data, supporting the extension and adaption of systems based on emerging requirements, and harmonizing data access for industrial analytics applications. The contributions and conclusions of this study are important for facilitating big data analytics research in large-scale industrial environments, where the requirements and demands of data management are significantly different to traditional information systems. O'Donovan *et al.* 2015 in their study focused on the systematic mapping of big data technologies in manufacturing. The research offered in this paper provided a breadth-first review of the research relating to big data in manufacturing to promote a better understanding of a new and pervasive area.

Bughin *et al.* 2016 have looked in his article at the returns generated by five big data use cases applied in the telecom industry. The article discovers evidence that big data projects generate positive contribution, but also with only a few providing additional risk adjusted, market value from big data. Using a probit model of big data adoption as well as a regression model returns to adoption, theyobserved that big data returns can be enhanced pointedly to the extent that companies obey to a few managerial and organization practices. Hurtado *et al.* 2016 have found using association analysis and ensemble forecasting to automatically discover topics from a set of text documents and forecast their evolving trend in a near future. In order to uncover meaningful topics, they collect publications from a particular research area, data mining and machine learning, as their data domain. An association analysis process is applied to the collected data to first detect a set of topics, followed by a temporal correlation analysis to help learn correlations between topics, and find a network of topics and communities. After that, an ensemble forecasting approach is suggested to predict the popularity of

research topics in the future. (Chandak, 2016) tried to propose a strategy based on string or pattern matching to handle data streams. The offered strategy can handle infinite-length, concept-evolution and concept-drift. It can also identify multiple novel classes occurring simultaneously. Crawford *et al.* 2016 survey the prominent machine learning techniques that have been proposed to solve the problem of review spam detection and the performance of different approaches for classification and detection of review spam. They provided a strong and comprehensive comparative study of current research on detecting review spam using numerous machine learning techniques and to devise methodology for conducting further exploration.

O'Donovan *et al.* 2015 in their study focused on the systematic mapping of big data technologies in manufacturing. The research presented in this paper offered a breadth-first review of the research relating to big data in manufacturing to endorse a better understanding of a new and pervasive area. The work presented in Hayes *et al.* 2015 paper describes a novel framework for anomaly detection in Big Data. Specifically, the framework utilizes a hierarchical approach to identifyanomalies in real-time, while also detecting a number of false positives. Then, a contextual anomaly detection algorithm is used to prune the anomalies detected by the content detector, but using the meta-information associated with the data points. To cope with the velocity and volume of Big Data, the anomaly detection algorithm relies on a fast, albeit less accurate, point anomaly detection algorithm to find anomalies in real-time from sensor streams. These anomalies can then be processed by a contextually aware, more computationally expensive, anomaly detection algorithm to determine whether the anomaly was contextually anomalous. This approach allows the algorithm to scale to Big Data requirements as the computationally more expensive algorithm is only needed on a very small set of the data, i.e. the already determined anomalies. The evaluation of the framework was also discussed based on the implementation details provided in this paper. The evaluation of the framework was performed using three sets of data; one for a set of HVAC electricity sensors, one for a set of temperature sensors, and a third set for a traffic system in California. Najafabadi *et al.* 2015 In their study, they explore how Deep Learning can be utilized for addressing some important problems in Big Data Analytics, including extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. We also investigate some aspects of Deep Learning research that need further exploration to incorporate specific challenges introduced by Big Data Analytics, including streaming data, high-dimensional data, scalability of models, and distributed computing. Agnihotri *et al.* 2015 focused on the impact of big data analysis for the E-governance. He suggested getting insight to the results of the predictive analysis can give huge benefits to the E-governance. Suresh *et al.* 2013 discussed the challenges that are imposed by E-governess on the modern and future infrastructure. His paper refers to map reducing algorithm to define the requirements on data management, access control and security. He suggested the map reducing architecture model that provides the basic for building interoperable data.

Conclusions

In this paper, we have studied the pioneering topic of big data analytics, which has recently gained lots of interest due to its perceived exceptional opportunities and benefits. Industry influencers, academicians, and other prominent stakeholders undoubtedly agree that big data has become a big game changer in most, if not all, types of modern industries over the last few years. As big data continues to permeate our day-to-day lives, there has been a significant shift of focus from the hype surrounding it to finding real value in its use. Our study provides an analysis of the big data analytics concepts which are being explored, as well as their significance to decision making. Consequently, some of the big data analytics tools and methods in particular were examined. Thus, big data storage and management, as well as big data analytics processing were detailed. In addition, we have examined different verticals that are using big data, specific solutions for different sectors which big data provides for decision makers. Though, There are various research are available on application of Big data in various fields like Banking and Securities, Communications, media, entertainment, healthcare, education, manufacturing, e-governance, Transportation etc. We believe that big data analytics can provide unforeseen insights and benefits to decision makers in various other areas which are not yet explored. Capital market, Trade surveillance, Compliance, AML, election campaigns, energy & utilities, Insurance etc., are some of area we have identified which can be potentially benefited from big data analytics.

Trade surveillance is experiencing increased regulatory scrutiny and complexities due to the prevalence of multiple communication platforms, making it difficult for regulators to perform market oversight functions. Big Data technology will play a more important role in monitoring market participants' trading activity both at participants' and regulators' ends. This is done by ingesting enormous volumes of various types of data originating from different channels (such as social media messages, blogs, emails, phone call logs, bank statements) and consolidating this structured and unstructured data into a usable database that will allow advanced pattern-matching analytics to spot any anomalous behavior. We see application of big data in Trade surveillance as a potential field where more analysis and research are desired. In our study we also found that various open source tools and techniques are available for big data analytics framework implementations and existing data mining algorithms are being integrated within these frameworks. The future scope of research work is to explore most effective framework for addressing various issues like performance, quality, feasibility, scalability etc. Also scope of future research is to integration of leading data mining, ETL & Business Intelligence tools with big data analytics tools.

REFERENCES

AgarwalAnkur, Baechle Christopher, Behara Ravi S., RaoVinaya, 2016. Multi-method approach to wellness predictive modeling.Big Data 3:15, DOI: 10.1186/s40537-016-0049-0

AgnihotriNishant, Sharma Dr. AmanKumar, 2015. Big data analysis and its need for effective e-governance. *International Journal of Innovations and Advancement in Computer Science IJIACS*, ISSN 2347 – 8616, Volume 4, Special Issue March.

BarrachinaArantxa Duque and O'Driscoll Aisling, 2014. A big data methodology for categorizing technical support requests using Hadoop and Mahout.Big Data 1:1DOI: 10.1186/2196-1115-1-1.

BerengueresJose and EfimovDmitry, 2014. Airline new customer tier level forecasting for real-time resource allocation of a miles program. Big Data 1:3DOI: 10.1186/2196-1115-1-3.

BughinJacques, 2016. Big data, Big bang?. Big Data 3:2DOI: 10.1186/s40537-015-0014-3

BughinJacques, 2016. Reaping the benefits of big data in telecom.Big Data 3:14,DOI:10.1186/s40537-016-0048-1

CerchielloPaola, GiudiciPaolo, 2016. Big data analysis for financial risk management. Big Data 3:18,DOI: 10.1186/s40537-016-0053-4

Chandak M. B. 2016. Role of big-data in classification and novel class detection in data streams. Big Data 3:5 DOI 10.1186/s40537-016-0040-9

ChopadePravin *† and Zhan Justin, 2015. Structural and functional analytics for community detection in large-scale complex networks.Big Data 2:11 DOI 10.1186/s40537-015-0019-y.

Crawford Michael, KhoshgoftaarTaghi, M., Prusa Joseph, D., Richter Aaron, N. and NajadaHamzahAl, 2015. Survey of review spam detection using machine learning techniques. Big Data 2:23 DOI 10.1186/s40537-015-0029-9.

EtaniNoriko, 2015. Database application model and its service for drug discovery in Model-driven architecture. Big Data 2:16 DOI 10.1186/s40537-015-0024-1.

Fang Xing and Zhan Justin, 2015. Sentiment analysis using product review data. Big Data 2:5 DOI 10.1186/s40537-015-0015-2.

Hayes Michael, A. and Capretz Miriam, A.M. 2015. Contextual anomaly detection framework for big sensor data. Big Data 2:2 DOI 10.1186/s40537-014-0011-y.

Herland Matthew, KhoshgoftaarTaghi, M. and Wald Randall, 2014. A review of data mining using big data in health informatics.Big Data 1:2DOI: 10.1186/2196-1115-1-2.

Hurtado Jose, L., Agarwal Ankur and Zhu Xingquan, 2016. Topic discovery and future trend forecasting for texts.Big Data 3:7 DOI 10.1186/s40537-016-0039-2

Kumar Anand, Grupcev Vladimir, BerradaMeryem, Fogarty Joseph, C., Yi Tu Cheng, 2014. Zhu Xingquan, PanditSagar A and Xia Yuni.:DCMS: A data analytics and management system for molecular simulation.Big Data2:9DOI: 10.1186/s40537-014-0009-5.

Kumar Sachin and ToshniwalDurga, 2015. A data mining framework to analyze road accident data. Big Data 2:26DOI: 10.1186/s40537-015-0035-y

Kumar Sachin, ToshniwalDurga, 2016. A novel framework to analyze road accident time series data. Big Data 3:8 DOI 10.1186/s40537-016-0044-5

Lu Yun, Zhao Ming, Wang Lixi and Rishe Naphtali, 2014. v-TerraFly: large scale distributed spatial data visualization with autonomic resource management.Big Data1:4DOI: 10.1186/2196-1115-1-4

NajafabadiMaryam, M., VillanustreFlavio, KhoshgoftaarTaghi M., SeliyaNaeem, Wald Randall and Muharemagic Edin, 2015. Deep learning applications and challenges in big data analytics.Journal of Big Data 2:1 DOI 10.1186/s40537-014-0007-7.

O'Donovan Peter, Leahy Kevin, Bruton Ken and O'Sullivan Dominic T. J. 2015. Big data in manufacturing: a systematic mapping study. Big Data 2:20DOI: 10.1186/s40537-015-0028-x.

O'Donovan, P., Leahy, K., Bruton, K. and O'Sullivan, D. T. J. 2015. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. Big Data 2:25DOI: 10.1186/ s40537-015-0034-z.

PääkkönenPekka, 2016. Feasibility analysis of AsterixDB and Spark streaming with Cassandra for stream-based processing. Big Data 3:6 DOI 10.1186/s40537-016-0041-8

Singh Manpreet, 2014. Big Data in Capital Markets.International Journal of Computer Applications (0975 – 8887)Volume 107 – No 5, December.

SiulySiuly, Zhang Yanchun, 2016. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis.Data Sci. Eng. 1(2):54–64 DOI 10.1007/s41019-016-0011-3.

Suresh, M., Parthasarathy, R., Prabakaran, M., Raja, S. 2013. Big Data Challenges for E-governess System in Distributing Systems. *International Journal of Soft Computing and Engineering (IJSCE),* ISSN: 2231-2307, Volume-3, Issue-5, November.

Toga Arthur, W. and DinovIvo, D. 2015. Sharing big biomedical data. Big Data 2:7 DOI 10.1186/s40537-015-0016-1.

Weiss Karl, KhoshgoftaarTaghi, M. and Wang DingDing, 2016. A survey of transfer learning. Big Data 3:9 DOI 10.1186/s40537-016-0043-6

Yang Jie, YeciesBrian, 2016. Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews. Big Data 3:3 DOI 10.1186/s40537-015-0037-9

YasseriTaha and Bright Jonathan, 2016. Wikipedia traffic data and electoral prediction: towards theoretically informed models.Yasseri and Bright EPJ Data Science, 5:22.

ZangWenyu, Zhang Peng, ZhouChuan and GuoLi, 2014. Comparative study between incremental and ensemble learning on data streams: Case study. Big Data1:5DOI: 10.1186/2196-1115-1-5.

Zhan Justin, GuidibandeVivek, Parsa Sai PhaniKrishna, 2016. Identification of top-K influential communities in big networks. Big Data3:16,DOI: 10.1186/s40537-016-0050-7

****