DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4628

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PREDICTIVE ANALYTICS CONCEPTS IN BIG DATA- A SURVEY

S. Banumathi
Assistant Professor, Department of Computer Science, Holy Cross College, Trichy-2,
Tamil Nadu, India

A. Aloysius
Assistant Professor, Department of Computer Science, St. Joseph's College, Trichy-2.
Tamilnadu, India

Abstract: Nowadays the increase of data variety considered very dispute problem for analysis. So innovative methods are mandatory for analytics especially in big data where the data in characteristic very complex and unstructured. The analytics is the process of analysis to predict concealed pattern and association among data. The main objective of this survey paper is to provide the exhaustive view of different predictive analytics applications and approaches. Analytics methods focused with dissimilar perspectives based on applications and data variety. Some of the application discussed is big data in hotel governance, higher education, health care, data e-governance, consumer orientations. This paper present different predictive approaches adapted for different application with challenges and suggestions.

Keywords: Big data, Predictive analytics, Big data Applications, Predictive approaches, Challenges.

I. INTRODUCTION

Data at present refers excessively big and fast. Therefore conventional database approaches cannot process them. Therefore a method to capture, store, distribute, manage and analyze diverse larger data is big data. Main features of big data are volume, velocity, veracity, variety, validity, volatility [28].

A. Features of Big Data

- **Volume** - Big data implies enormous volumes of data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive.
- **Velocity** - Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc.
- **Veracity** - Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Inderal feel veracity in data analysis is the biggest challenge when compares to things like volume and velocity.
- **Variety** - Variety refers to the lot of sources and types of data both structured and unstructured. It used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc.
- **Validity** - Like big data veracity is the issue of validating meaning is the data correct and accurate for the intended use. Clearly validate data is key to making the right decisions.
- **Volatility -** Big data volatility refers to how long is data valid and how long it be stored. In this world of real time data you need to determine at what point is data no longer relevant to the current analysis.

B. Big Data analytics

Big data analytics is the process of collecting, organizing and analyzing large sets of data called big data to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the knowledge that comes from analyzing the data.

There are four types of big data analytics: Prescriptive, Predictive, Diagnostic and Descriptive. The prescriptive analytics reveals what actions should be taken. The predictive analytics is an analysis of likely scenarios of what might happen. The diagnostic analytics is a look at past performance to determine what happened and why. The descriptive analytics is typically use a real-time dashboards or reports.

II. PREDICTIVE ANALYTICS

Predictive analytics is the use of data, statistical algorithm and machine-learning techniques to identify the likelihood of future outcomes based on historical data. Predictive models use know results to develop a model that can be used to predict values for different or new data. The modelling results in predictions that represent a probability values for different or new data. The predictive analytics used to predict trends, improve performance, drive decision making, and predict the behavior.

III. PREDICTIVE ANALYTICS IN DIFFERENT APPLICATIONS

A. Prediction in health care

YiChuanWang[1] et al., Illustrated that big data is a result of controlling huge volume of digital information. Data form various sources are structured in the first layer and are acquired and transformed using transformation engines and stored. In the analytics it is mapped and processed. Cases from various regions under different parameters are collected which are then subjected to content analysis. According to the specific criteria map reduced algorithms by Apache Hadoop does the analytical process.

This process unlike conventional methods can identify and analyses semi-structured and unstructured data. Analytical capability can help discover similarities found in a patience massive health record and thereby create a balance between capacity and cost.

Predictive capability is following different methods in statistical analysis modelling machine learning and data mining. This helps in cross deciding current data to future events. The amount of information with regard to health care has been astronomical that conventional storage device find it hard to store, process, queue and retrieve as and when needed. Added to this understanding of patterns, trends within the data require Big Data Analytics, which has the potential to improve health care, save lives and lower cost. It also helps in detecting diseases early and treating them. In Research and Development, predictive modelling helps produce leaner and faster flow of drugs and devices. In Public health, Big Data helps analyze disease pattern and track outbreaks of diseases early for a speedy response. Efficient care could be given to individuals by combining structured and unstructured data. Gene sequencing could also be done efficiently and cost - effectively. Traditionally health care records have been static, be it x-rays, scans or test results. Big Data analytics replaces them. Then the methodology should be derived at including variable selection, data collection, analytic techniques, association and result. Final stage is the deployment part which include evaluation, validation and testing.

B. Prediction in higher education

Higher education today operated in a complex and competition environment and therefore has to face its challenges accordingly. More over different stack holders in higher education has paved way for big data to pay a crucial role. Vast data that keeps coming every day can be utilized only through big data [1]. Big data benefits have not be used its fullest possible extent in health care. In terms of business value health care has not reached out of big data [2] [3]. Readily available messages from social media and consumer generated content in the internet can be used to solve real life problems using Big Data analytics that will eventually reshape our understanding of the field and decision making [4]. The challenges of storing, accessing it in real time, analysis, obstacles and security become paramount [5] [6]. This paper discusses as to how using predictive analysis in big data could be used to progress decision making in different applications.

Ben K. Daniel [4] et al., suggested that today big data operated in a complex and competition environment and therefore has to face its challenges accordingly. More over different stack holders in higher education has paved way for big data to pay a crucial role. Vast data that keeps coming every day can be utilized only through big data. Data today is too big and too fast. Therefore conventional database cannot process them. Therefore a method to capture, store, distribute, manage and analyze diverse larger data is big data. Stored data can be properly explored using analytic techniques. System such as Apache Hadoop, Horton works, and map reduce and tableau are powerful software that could be used even without advanced technical knowledge. Big data in higher education can have an impeding effect in management decision making theory vast available administrative and operational data can be processed and assessed to predict future performance and identify potential areas in academic programming, research teaching and learning.

A large scale statistical techniques combined with predictive modelling helps improve decision making. Big data can have an important role in three data models: Descriptive, Prescriptive and Predictive. Descriptive analytics analysis the raw data received and predictive analysis tries to figuring out future probabilities based on predictive analysis, prescriptive information are given to students and stack holders. Value of big data will be based on creating governing structures and creating more progressive and better policies. Security and privacy are other challenges faced by big data.

Greenberg and Buxton’s [18] et al., stressed the need for “higher education to transform its own culture.” Information technology should be used to apply rigorous approaches to analytics in “supporting evidence-based decision-making and management”. In similar context, the online learning research community must bring transparency to effective practice of learning analytics to deter potentially wrongful uses of big data in online courses.

Kelderman[19] et al., reported that accreditors are attempting to keep pace with new federal regulations to provide tighter oversight on online programs, “requiring colleges to prove that students learn as much in distance courses as in face-to-face courses”. These requirements upsurge the pressure on educational institutions to respond to new rules and provide clear valuations of quality online education. Moreover, the trainer has a need to know what is happening in the online course; the use of learning analytics would produce information about student progress and the instructional process. Siemens [20] et al., insisted that the online learning community wants to guide the direction as to how knowledge analytics are used in defining and evaluating big data in online courses. This guidance includes the need for defining data, emerging learning analytics methodologies and tools, picturing and sharing the nature of education analytics output, and informing effective process and practice that leads to expressive decision-making about learner performance. Waltman [21] et al., claimed, needing cited Papers are not always suggestive of impactful research. However, as the authors further noted, on normal this idea does tend to hold true. As such, it is sensible to assume that high citation rates do imitate a certain level of excellence.

C. Prediction in hotel governance

Zheng Xiang[14] et al., suggested that big data generated through internet traffic, mobile transaction, user generated content, social media, sensor networks and other. This Big Data is crucial for business intelligence and intern help understand customers, competitors, market characteristic, products, business environment impact of technologies and so forth [10]. Unstructured human authored document can be put into sentiment analysis technologies. As far as this research paper is concerned, it studies customer satisfaction in a hotel as soon as the person receives the product or service given the complexity of hotel guest satisfaction measuring them is very challenging. Though customer reviews are found in many travel websites, expedia and travel velocity, allow only their customers to write reviews and share their experience. This prevents in authentic reviews. The main goal of this study is to understand the content and the structure of customer reviews and how their associated with Hotel guest satisfaction which pertains to overall rating. For this data were collected during the period of December 18-29 in 2007 using an automated web crawler. It collected 10,537 hotels resulting in 60,648 customer reviews. From it 6642 unique words were identified. Microsoft access with unique identifier were assigned to every hotel property and customer reviews. After this data analysis followed test analysis process which include, stemming, misspelling identification, removal of stop words, such as pronounce, adverbs, conjunction. Coding scheme was established to guide the domain identification process, removing generic nouns that lack specificity, generic verb, words with high ambiguity and finally hotel brand name. Based on the findings and using pivotal table in Excel sheet, 416 words were considered irrelevant. After this the findings were presented in two parts. (i)Basic description. (ii) Clean data. Interestingly top ten sites had 60% of total properties in clean data, while had only 34% original data set. Conventional method rely on set of predefined hypothesis justified using previously existing knowledge, big data let research understand a new pattern of reflective of customers evaluation there by generating and creating new knowledge. This study is not based on sentiment analysis which are subjective in nature, were as it is purely analytical in nature. Although this study evolves a new field of knowledge and understanding unlike conventional guest survey studies, it has much limitation and therefore the finding should be treated with caution, because customer reviews are basically a self-selection of bias. However it does not reduce the internal validity. Therefore future study applying method of triangulation to multiple sources of data to validate the semantic structure could be evolved using big data analytics. Authors might have improved their survey little deep. May be in future it may fulfil.

Hyun Jeong “Spring” Han [22] et al., indicated that the hotels are rated based on guest/customer satisfaction. The strategy resulted in negative comments having more weight than the positive comments. This uneven weighing, leads to guest’s bad feeling of poor service which will submerge the good service of positive feelings. In this study, text analytics using regression analysis to improve guest’s assessments and their ratings is done through big data analytics.

Zhen Xiang [23] et al., suggesting that the association between guest experience and satisfaction appears strong, suggesting that these two domains of consumer performance are characteristically connected. This study discloses that big data analytics can produce new visions into variables that have been widely studied in presenting hospitality literature. It even implications for theory and practice as well as directions for future research are discussed.

D. Big Data prediction in Data Governance

C. Mohanapriya[7] et al., suggested that Data Governance Incorporates, Data Confidentiality and Data quality and Privacy. It prevents unauthorised accessing of data. Data quality depends on data Privacy. Data governance is a total sum of usability, availability, integrity and security. Data governance is essential to get funds, increase confidence levels, increase speed in accessing data, fast decision making and precise trustworthy information. It has six steps namely, Data extraction, content analysis, data maintenance, process computing, secure delivery and fast delivery. Its benefits are six folds. They are heterogeneous data integrations, security and privacy, accounts deeper knowledge, Data validity, data protection and faster delivery. Henry C. Lucas [15] et al. Information technology has altered the traditional way of doing business by redefining business capabilities and entering into a new market space. There has been a marked change in the process of doing business [14] creating new organisation like Amazon, Facebook, Google; developing new relationships in terms of social media; creating new user experience [8]. creating new market like iTunes. Impacts of information technology have been different on individuals, firms and economy or society at large.

E. Big data prediction in Consumer orientations

Sunil Erevelles[15] et al., illustrated that the study of consumer analytics lies at the junction of big data and consumer behaviour. Data provide behavioral insights about consumers; marketers translate these insights into market advantage. Hidden insights means, predicting the possible activities that are unexploited by the consumers. Even though big data is the new form of capital in recent trend, it failed to exploit its benefits in many firms. To profit from this new form of capital, firms must allocate appropriate physical, human and organizational capital resources to big data. The conceptual frame work is introduced to illustrate the impact of big data, using this frame work, a firm can create a value and gain a competitive advantage. In today’s evolving technology the consumer data for any organization is generated incessantly of both transactional and behavioral data [9]. The persistent rapidity of data is constantly generated in 3 dimensions volume, velocity and variety. The volume of data is constantly increasing the consumer big data 1 zeta byte every two years. The velocity of data created is analyzed by evidence at given time. Comparing the census data and clothing retailer data i.e. what the consumers are posting on social networks about the retailer, gives the ability to make decisions. The variety of structured and unstructured data are been organized using various software that bring order to the unstructured data. The standard generalized mark-up language software enables the viewing of videos to determine common elements that an organization wants to capture. Resources such as physical capital resources, organizational capital resources. Enormous amount of data generated in context of resources [11]. In this hyper-competitive environment, organization must often update and reconfigure the resources with the changes in environment to sustain in competitive advantage. Both the dynamic and adaptive capability achieved through consumer insights got from Big data [16]. Ignorance is define to say that don’t know, in general researcher focus on what they know, similarly it’s important to focus on ignorance because it facilitates latitude and liberty for inspiring creativity within an organization. Inductive reasoning, one method of scientific review starts with observing a phenomenon before forming hypothesis.

Farshad Kooti[24] et al., Recently huge amount of population are spending large fraction of their economy in shopping and purchases. Consumer from affluent areas purchase more expensive item frequently, which results in more money spend on online shopping. Temporal patterns of consumer is identified to fine their finite budgets and they will wait for last purchase to buy it. It’s observed that shoppers who email each other purchase more similar items than socially unconnected shoppers. Using temporal patterns prediction is improved for consumer and when they will make online purchase again. Mostafa Sabbaghi [27] et al., indicating that the paper aims at providing astute statistical analysis of Electronic Waste(e-waste) dynamic nature by reviewing the effects of design features, brand and consumer type on the electronics tradition time and end of use time-in-storage.

IV. CONCLUSION

Complex analytics tasks have become commonplace for a wide range of users. This paper specifically identified main applications which depends thoroughly on big data predictive analytics solutions and already adopts themselves as one of the big data entities. However, instead of targeting the use cases and computing resources of the typical user, existing analytics frameworks are designed primarily for working with huge datasets in various applications. Therefore the future implications will be based on pattern predictions and different evolutionary techniques from various data.

V. REFERENCES

[1] YiChuan Wang, LeeAnn Kung, Chaochi Ting, “Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care”, publications on ResearchGate , 2015.

[2] Baker, R. S. J. D. “Learning, schooling, and data analytics”. Handbook on innovations in learning for states,districts, and schools, Philadelphia, PA: Center on Innovations in Learning , 2013, pp. 179–190.

[3] BasU.A, “Five pillars of prescriptive analytics success”s. Analytics-magazine.org, 2013, pp. 8–12.

[4] Ben K. Daniel, “Big Data and analytics in higher education: Opportunities and challenges”, British journal of educational technology. September , 2015.

[5] Raghupathi, W, “Big data analytics in healthcare: promise and potential. Health Information Science and System”s, volume2, 2014.

[6] Bharadwaj, A, El Sawy, O.A. Palou, P.A. and Venkatraman, “Digital Business Strategy: Toward A Next Generation of Insights”, MIS Quarterly, 2013.

[7] C. Mohanapriya, “A Trusted Data Governance Model For Big Data Analytics”, Volume 1, Issue 7, ISSN (online): 2349-6010, Dec 2014.

[8] Aiden, E., Michel, “The Predictive Power of Big Data. News week”. April 2014.

[9] Sunil Erevelles, Nobuyuki Fukawa, Linda Swayne, “Big Data consumer analytics and the transformation of marketing”, Journal of Business Research, JBR-08469, July 2015.

[10] Mithas. S, Lee, M. R, Earley, Murugesan, “Leveraging big data and business analytics”, IT Professional, IEEE Explore, 2013.

[11] X. Wang, D.E. Brown, M.S. Gerber, “Spatio-Temporal Modelling of Criminal Incidents Using Geographic, Demographic, and Twitter-Derived Information,” Proc. IEEE International Conference Intelligence and Security Informatics (ISI), 2012.

[12] R.Y. Lau, C. Li, and S.S. Liao, “Social Analytics: Learning Fuzzy Product Ontologies for Aspect-Oriented Sentiment Analysis”, Decision Support Systems, vol. 65, 2014.

[13] C.C. Yang, J. Yen, and J. Liu, “Social Intelligence and Technology,” IEEE Intelligent Systems, vol. 29, no. 2, 2014.

[14] Yichuan Wang, Chaochi Ting, Leeann Kung, Terry Byrd, “Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care”, Hawaii International Conference on System Sciences, 2015.

[15] Sunil Erevelles, Nobuyuki Fukawa, Linda Swayne, “Big Data consumer analytics and the transformation of marketing”, Journal of Business Research, 2015.

[16] Boying Lia, Eugene Ch’ngb, Alain Yee-Loong Chongc, Haijun Baod,” Predicting online e-marketplace sales performances: A big data approach”, Elsevier Ltd, 2016.

[17] Henry C. Lucas, Jr, Eric K. Clemons, Omar A. El Sawy, Bruce Weber, “Impactful Research on Transformational Information Technology: An opportunity to Inform New Audiences”, MIS Quarterly, Vol. 37 No. 2, 2013, pp. 371-382.

[18] Greenberg, S., and Buxton, B., “Usability evaluation considered harmful (some of the time)”, Florence, Italy: ACM Press, 2008, pp. 111-120.

[19] Kelderman, E. “Online programs face new demands from accreditors. The Chronicle of Higher Education”, 2011.

[20] Siemens, G. “Learning and knowledge analytics”, http://www.Learning analytics. net, 2011.

[21] Waltman, L., Van Eck, N.J., and Wouters, P., “Counting publications and citations: Is more always better? “, Journal of Informatics, 2013.

[22] Hyun Jeong “spring” Han, Shawn Mankad, Nagesh Gavirneni, Rohit Verma, “What Guests Really Think of Your Hotel: Text Analytics of Online Customer Reviews”, Cornell Hospitality Report, 2016.

[23] Zheng Xiang, Zvi Schwartz, John H. Gerdes Jr, Muzaffer Uysal, “What can big data and text analytics tell us about hotel guest experience and satisfaction?”, International Journal of Hospitality Management, 2016.

[24] Farshad Kooti,Kristina Lerman, Luca Maria Aiello, Mihajlo Grbovic, Nemanja Djuric,Vladan Radosavljevic,” Portrait of an Online Shopper: Understanding and Predicting Consumer Behavior” , 2015.

[25] Mostafa Sabbaghi, Behzad Esmaeilian, Ardeshir Raihanian Mashhadi, Sara Behdad,” An investigation of used electronics return flows: A data-driven approach to capture and predict consumer’s storage and utilization behavior ”, Elsevier Ltd, 2014.

[26] S.Banumathi, A. Aloysius , “ Big data prediction using an evolutionary techniques”, International journal of Emerging technologies and Research, Volume 3, Issue 9, (ISSN-2349-5162),2016 .