IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol-6, Special Issue-8, Oct 2018 E-ISSN: 2347-2693

An Investigation on Social Media Issues Using Big Data Analytics

[1*]K. Yemunarane, [2]D. Hemavathi

[1,2]Kongunadu Arts and Science College, Coimbatore, India.

Available online at: www.ijcseonline.org

Abstract— This paper describes how big data technologies are converging to offer a cost-effective delivery model social media based big data analytics. Social Media is a powerful technology to perform massive-scale and complex computing. It eliminates the need to maintain expensive computing hardware, dedicated space, and software. Massive growth in the scale of data or big data generated through social media has been observed. Addressing big data is a challenging and time demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. In this paper the relationship between big data and social media, the classification of big data and the scope of big data analytics are discussed.

Keywords- Big Data, Social Media, Techniques, Big Data Analytics, Clustering.

I. INTRODUCTION

Data Mining is a database, with arrangements to extract hidden data from a large database. Sequencing lots of information packages and using complex calculations to select important data. Expect future methods and procedures for information mining equipment, allowing interactive companies to perform approved operating options. More information is collected by the amount of information being multiplied every year, and the Information Mining Motion has become indispensable to change this information into data [9]. Information technology was developed in a long analysis of experiments and improvements in the products.

Data mining analytics works with data and the best techniques look at data greatly and use data collected as much as possible to obtain reliable results and results. Analysis process begins with a set of data, which uses a method to generate optimal representation of the structure of data acquired timely. Once acquired knowledge, a large data set can be extended to larger packages of data that can be assumed to have an assumption of pattern. Is again similar to a mining operation, where large amounts of low-sized goods are wiped out by the means of finding the value.

The following diagram summarizes certain stages processes to identify data mining and knowledge innovation

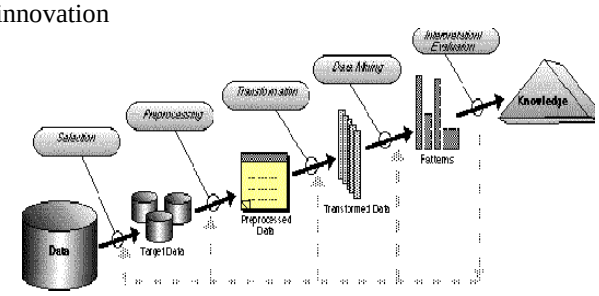

Fig 1. KDD Process

The phases depicted start with the raw data and finish with the extracted knowledge which was acquired as a result of the following stages:

Selection: Selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined.

Pre-processing: This is the data cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries for example unnecessary to note the sex of a patient when studying pregnancy. Also the data is reconfigured to ensure a consistent format as there is a possibility of inconsistent formats because the data is drawn from several sources e.g. sex may recorded as f or m and also as 1 or 0.

Transformation: The data is not merely transferred across but transformed in that overlays may added such as the demographic overlays commonly used in market research. The data is made useable and navigable.

Data mining: This stage is concerned with the extraction of patterns from the data. A pattern can be defined as given a set of facts(data) *F*, a language *L*, and some measure of certainty *C* a pattern is a statement *S* in *L* that describes relationships among subset *Fs* of *F* with a certainty *c* such that *S* is simpler in some sense than the enumeration of all the facts in *Fs*.

Interpretation and evaluation: The patterns identified by the system are interpreted into knowledge which can then be used to support human decision-making e.g. prediction and classification tasks, summarizing the contents of a database or explaining observed phenomena.

II. DATA MINING TECHNIQUES

Data mining provides meaningful data from the database. It includes various technological approaches. Data mining

creates a descriptive model or a forecast model. An explanatory model describes the general characteristics of the data in the database. A prediction model makes the current data reliable to predict. The predictive and descriptive model goal is to achieve a variety of data mining techniques as shown in fig 2.

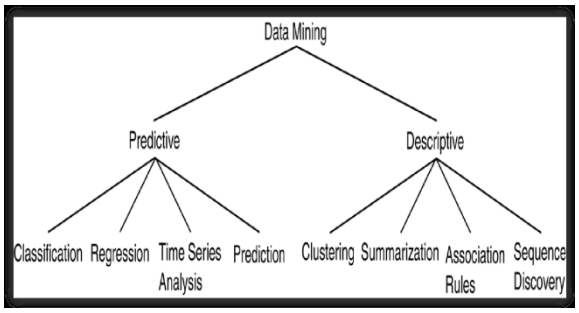

Fig 2. Data Mining Models

Classification: Classification based on assorted values (i.e., isolated, unordered). This technique based on supervised learning (that is, the output required for input is known) cannot predict the following values. This is derived from the existing data and values (the class label). Conclusion The tree can be classified using the neural network. Mathematical formulas and classification rule (IF- Then).

Regression: Regression is used to map a data item to a real valued prediction variable. In other words, regression can be adapted for prediction. The target value is known in this technique.

Time Series Analysis: The statistical techniques are used in time series analysis and gives detail about data points which is dependent on time series. Time series forecasting is used to generate predictions of future events depend on past events.

Prediction: Prediction discovers the relationship between independent variables and dependent variables. It gives continuous value or ordered value (between some ranges).

Clustering: Clustering is a set of similar data object. The steering object is another clustering. It analyzes data objects without a subject label. For example, it involves a lack of supervision, a retailer can build various committees based on a customer base such as weekly purchase and regular purchase.

Summarization: Summarization is abstraction of data. It is set of relevant task. For example, long distance calls can be summarized total minutes, seconds and total cost of the call.

*Association Rule***:** It is a data mining techniques which give a set of items and a huge collection of transaction in frequent item set. Association strives to discover patterns in data which are based upon relationships between items in the same transaction. Association rule is used in the market based analysis to identify a set, or sets of products that consumers often purchase at the same time.

Sequence Discovery: It is used among data for uncovers relationship. It is set of object each associated with its own timeline of events.

III. BIG DATA – AN INTRODUCTION

The continuous increase in the volume and detail of data captured by organizations, such as the rise of social media, Internet of Things (IoT), and multimedia, has produced an overwhelming flow of data in either structured or unstructured format. Data creation is occurring at a record rate [1], referred to herein as big data, and has emerged as a widely recognized trend. Big data is eliciting attention from the academia, government, and industry. Big data are characterized by three aspects: (a) data are numerous, (b) data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed rapidly. Moreover, big data is transforming healthcare, science, engineering, finance, business, and eventually, the society.

Social media is one of the most significant shifts in modern ICT and service for enterprise applications and has become a powerful architecture to perform large-scale and complex computing. The advantages of social media include virtualized resources, parallel processing, security, and data service integration with scalable data storage. Social media can not only minimize the cost and restriction for automation and computerization by individuals and enterprises but can also provide reduced infrastructure maintenance cost, efficient management, and user access [2]. As a result of the said advantages, a number of applications that leverage various cloud platforms have been developed and resulted in a tremendous increase in the scale of data generated and consumed by such applications.

3.1. What is BIG DATA?

Recently, the term of Big Data has been coined referring to those challenges and advantages derived from collecting and processing vast amounts of data. The sources of huge quantity of information are those applications that gather data from click streams, transaction histories, sensors, and elsewhere. However, the first problem for the correct definition of 'Big Data' is the name itself, as we might think that it is just related to the data volume. The heterogeneous structure, diverse dimensionality, and variety of the data representation, also have significance in the big data.

A. *Classification of Big Data*

Big data are classified into different categories to better understand their characteristics. Fig. 1 shows the numerous categories of big data. The classification is important because of large-scale data in the cloud. The classification is based on five aspects: (i) data sources, (ii) content format, (iii) data stores, (iv) data staging, and (v) data processing.

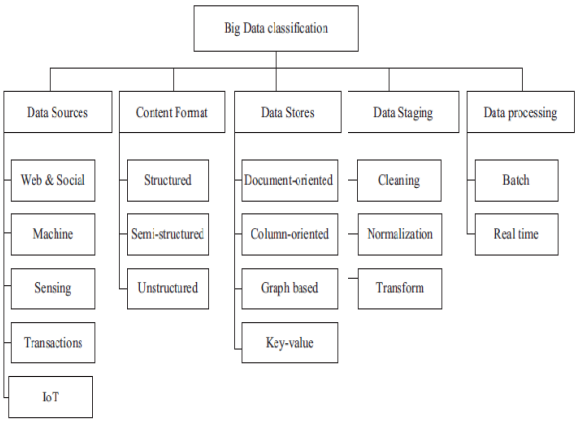

Figure 3: Classification of Big Data

IV. THE SCOPE OF BIG DATA ANALYTICS

Early interest in big data analytics focused primarily on business and social data sources, such as e-mail, videos, tweets, Facebook posts, reviews, and Web behavior. The scope of interest in big data analytics is growing to include data from intelligent systems, such as in-vehicle infotainment, kiosks, smart meters, and many others, and device sensors at the edge of networks—some of the largest-volume, fastest-streaming, and most complex big data. Ubiquitous connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information.

Interest in applying big data analytics to data from sensors and intelligent systems continues to increase as businesses seek to gain faster, richer insight more cost-effectively than in the past, enhance machine-based decision making, and personalize customer experiences.

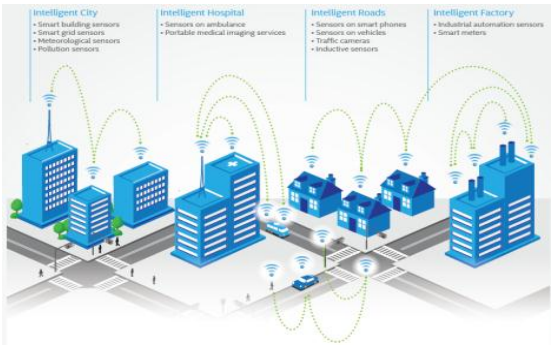

Figure 4: Big Data in Context: Smart City Example

V. CONCLUSION

The size of data at present is huge and continues to increase every day. The variety of data being generated is also expanding. The velocity of data generation and growth is increasing because of the proliferation of mobile devices and other device sensors connected to the Internet. These data provide opportunities that allow businesses across all industries to gain real-time business insights. The use of cloud services to store, process, and analyze data has been available for some time; it has changed the context of information technology and has turned the promises of the on-demand service model into reality. In this study, we presented a review on the rise of big data in social media. We proposed a classification for big data, a conceptual view of big data, and a cloud services model. In the future, significant challenges and issues must be addressed by the academia and industry. Researchers, practitioners, and social science scholars should collaborate to ensure the long-term success of data management in a social media environment and to collectively explore new territories.

REFERENCES

[1] D.Aruna Kumari , Dr.K.Rajasekhar rao, M.suman " Privacy preserving distributed data mining using steganography "In Procc. Of CNSA-2010, **Springer Libyary**

[2] T.Anuradha, suman M,Aruna Kumari D "Data obscuration in privacy preserving data mining in Procc International conference on web sciences ICWS 2009.

[3] Agrawal, R. & Srikant, R.(2000). Privacy Preserving Data Mining. In Proc. of ACM SIGMOD Conference on Management of Data (SIGMOD'00), Dallas, TX.

[4] Alexandre Evfimievski, Tyrone Grandison Privacy Preserving Data Mining. IBM Almaden Research Center 650 Harry Road, San Jose, California 95120, USA

[5] Agarwal Charu C., Yu Philip S., Privacy Preserving Data Mining: Models and Algorithms, New York, Springer, 2008.

[6] Oliveira S.R.M, Zaiane Osmar R., A Privacy-Preserving Clustering Approach Toward Secure and Effective Data Analysis for Business Collaboration, In Proceedings of the International Workshop on Privacy and Security Aspects of Data Mining in conjunction with ICDM 2004, Brighton, UK, November 2004.

[7] Flavius L. Gorgônio and José Alfredo F. Costa"Privacy-Preserving Clustering on Distributed Databases:A Review and Some Contributions

[8] D.Aruna Kumari, Dr.K.rajasekhar rao,M.Suman "Privacy preserving distributed data mining: a new approach for detecting network traffic using steganography" in international journal of systems and technology(IJST) june 2011.

[9] Binit kumar Sinha "Privacy preserving, and C. S. Yang, A Fast VQ Codebook Generation Algorithm via Pattern Reduction, *Pattern Recognition Letters*, vol. 30, pp. 653{660, 2009}

[10] C. W. Tsai, C. Y. Lee, M. C. Chiang Kurt Thearling, Information about data mining and analytic technologies http://www.thearling.com/

[11] K.Somasundaram, S.Vimala,"A Novel Codebook Initialization Technique for Generalized Lloyd Algorithm using Cluster Density", International Journal on Computer Science and Engineering, Vol. 2, No. 5, pp. 1807-1809, 2010.

[12] K.Somasundaram, S.Vimala, "Codebook Generation for Vector Quantization with Edge Features", CiiT International Journal of Digital Image Processing, Vol. 2, No.7, pp. 194-198, 2010.
[13] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino State-of-the-art in Privacy Preserving Data Mining in SIGMOD Record, Vol. 33, No. 1, March 2004.
[14] Quantization: A Review", IEEE Transactions on Communications, Vol. 36, No. 8, August 1988.
[15] Berger T, "Rate Distortion Theory", Englewood Cliffs, Prentice-Hall,NJ, 1971.
[16] A.Gersho and V.Cuperman, "Vector Quantization: APattern Matching Technique for Speech Coding", IEEE Communications, Mag., pp 15-21, 1983.
[17] "Privacy Preserving Data Mining - IBM Research: Almaden: San Jose
[18] D.Aruna Kumari, Dr.K.Rajasekhara rao, M.suman "Privacy Preserving Clustering in DDM using Cryptography"in TJ-RJCSE-IJ-06.

