ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Big Data Analytics In Social Media Using Data Mining Techniques: A Survey.

V Alli Rani*, and S Pradeepa.

School of Computing, SASTRA University, Thanjavur, India.

ABSTRACT

This rapid growth of data is not used in an efficient way as most of these data are not analyzed for their actual needs. However, these data can be so valuable than no one can ignore them. 2.9 billion Facebook status updates take place daily in the web all around the world. 4.3 billion tweets are generated in Twitter per day. 6.7 million photos are updated in the Instagram daily. These are a vast source of unstructured data yet to be tapped into useful information. It has become necessary to find out the insights from large data sets to predict the relationships, dependencies as well as the outcome and behavior of the consumer. Big Data has been characterized by 5 Vs – Volume, Velocity, Variety, Veracity and Value. This paper mainly addresses all these 5Vs, features, challenges, future of Big Data in social media arena using data mining algorithms, tools and Hadoop framework for overcoming challenges of Big Data.

Keywords: Big Data, Algorithms, Data mining, Hadoop, Data sets.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Recent technological advancements have been led into to a diversification of data from distinctive domains (e.g., health care, scientific sensors etc.,) over the decades. Big data not only encompasses a huge volume of data but it also exhibits unique characteristics which make it analytics friendly. For instance, a report from IDC [1] made a prediction that ,from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, representing a double growth every two years. This became true with the fast emergence of the social media platforms. A report from Mckinsey Global Institute [2] states that the value of global personal location data is estimated to be $100 billion in revenue to service providers which is more potential when analysed for insights and be as much as $700 billion in value to consumer and business end users. This level of huge potential harvesting characteristics of Big data has attracted more competitive players from various sectors for example, industry, government and research community. Facebook has 700 million active users which spend more than 12.3 billion hours a month, every month, more than 60 million items of content which includes photos, notes, blogs, posts, web links and news. You tube has 630 million visitors worldwide who spend approximately 4.3 billion hours each month and it upload 24 hours of video every minute, forming the enormous outstanding data aggregator [16],[17]. Social media hones large amount of user data which can be analyzed with various data mining techniques which can give more insightful predictions about a user's online behavior[5],[6]. To analyze a user's behavior towards making an opinion in an online platform, text analytics play a key role and the depth of the behavioral changes can be attributed by sentimental analytics [7]·When we apply big data analytics to the data mining techniques, we can extract information in both predictive and descriptive nature. Predictive nature can be attributed to many features like Classification, Regression and Deviation Detection. The Descriptive nature can be attributed to Clustering, Association Learning and Sequential Patterns [8]. The Data mining techniques are classified into three major types viz Unsupervised, Semi-Supervised and Supervised [9].

The enormous amount of data that is generated in the social media segment is posing a big challenge for researchers and scientists all over the globe. It is giving a new perspective of exploration with a more insightful approach towards data handling. Further, Big Data Analytics integrated with the Data mining can provide a better insightful information.[10][11]

MOTIVATION

Social media platforms generate massive amount of data on a daily basis. These data vary in form, size and time of input. Categorizing these massive datasets and analyzing them for potential insights is the need of the hour as the predictions may lead to a new level of predictive information insight-fulness. Computing the data from Social media platform introduces new challenges for analysis as the data flow and amount varies with the time and the user requirements. Data mining is the process of analyzing data from different perspectives and getting useful information. These information can be used to increase the revenue, cut costs or both [8].

The ultimate goal of Big Data Analytics in Social Media using Data Mining techniques is to maximize the profit for social platform analytics company by giving a better platform for the users and to minimize the hustle faced by the users while using the social media platforms.

Survey And Related Work

Rahul Sharan Renu et al. [18] proposed a system which uses some knowledge discovery and data mining (KDD) algorithms through the Waikato Environment for Knowledge Analysis (WEKA) interface. By using this interface the historical datasets obtained is analyzed automatically. The output of this automatic analysis is a important part as the analysis of large datasets of historical data was a challenging task while creating decision support systems.

Table I is used to describe the different techniques used for clustering algorithms. Table II is used to describe the techniques and drawbacks of clustering algorithms.

Fahad et al. [19] has given a framework which gives a detailed classification of clustering techniques. The author has taken five criteria's to categorize the clustering algorithms viz Partitioning, Hierarchical, Density, Model and Grid. The following table (Table 1) defines the types of clustering algorithms in a clear way.

Table I: An Overview Of Clustering Taxonomy [19]

Clustering Algorithms				
Model	**Hierarchical**	**Density**	**Grid**	**Partitioning**
EM Model	**BIRCH** **CURE** **ROCK** **Chameleon** **Echidna**	**DBSCAN** **OPTICS** **DBCLASD** **DENCLUE**	**Wave** **Cluster** **STING** **CLIQUE** **Optigrid**	**K-means** **K-medoids** **K-modes** **PAM** **CLARA** **CLARANS** **FCM**

Table 2: Summary Of Data Mining Techniques/Algorithms For Big Data Analytics

Author	**Technique/ Algorithm**	**Drawbacks**	Discussion in terms of Big Data Analytics in Social Media using Data Mining techniques (can be made by creating / availed by / integrated to / can take this model or algorithm or procedure)
Zhang et al.[22]	BIRCH (Balance Iterative Reducing and Clustering using Hierarchies)	Cannot work with no-spherical clusters.	Time and Space efficient algorithm that is used in this technique can be used
Bezdek et al.[30]	FCM (FuzzyCMeans)	More efficient in handling large datasets but comes with more noise.	Support can be made by integrating the large datasets but K-means problems must be minimized
Hinneburg et al. [31]	DENCLUE(DENsity CLUstering)	Input parameters must be carefully noticed.	Technique can be used but requires more modification in defining the rule set for Data Input.
Cai et al. [27]	RMKMC(Robust Multi View k-means Clustering).	Non singular covariance matrix is required.	Can be used only with singular covariance matrix with certain modifications.
Woo et al. [28]	wTabular–algorithm	Sensitivity to selection of original parameters	Can be used to improvise the convergence speed of large datasets.
Sergej et al. [29]	P3C+ - MR algorithm	Limited only to MapReduce component liability rules	Can be applied in complete MapReduce dataset environment.

Jain et al. [20] proposed detailed overview of Clustering steps which would be useful for clustering in image segmentation, character recognition, information retrieval. The data mining technique employed by the author paves a new insight to the application of techniques in a constraint clustering environment.

Tidke et al. [21] proposed a novel two step clustering algorithm. It uses clustering ensemble which will give accurate analytical output even when the datasets is large and is composed of varying features. Subspace clustering is done using PROCULUS. Clusters are formed by using average Mahattan sequential distance. k-means algorithm is applied on every subspace. The clusters are further split and processed by based on distance between the master and slave clusters.

Zhang et al. [22] proposed a hierarchical clustering algorithm BIRCH(Balance Iterative Reducing and Clustering using Hierarchies). This algorithm applies time and space efficiency to reduce noise in large databases by concentrating on dense and spatial regions.

Strehl and Ghosh [23] proposed a cluster framework in which large sets are partitioned multiple times and are optimized using greedy approaches to select best solution automatically without accessing the original attributes. CSPA(Cluster-based Similarity Partitioning Algorithm), HGPA(HyperGraph Partitioning Algorithm) and MCLA(Meta- Clustering Algorithm) are used in transformation of cluster label vectors into hypergraph representation.

Bezdek et al [24] proposed a technique called FCM (Fuzzy C means) which reduces the intra cluster variance which arises when the processing of large datasets from varied clusters are employed.But it inherits some problems of K-Means.

Hinneburg et al [25] proposed a new technique called DENCLUE which employs clustering based on density of the datasets. These datasets must be carefully inserted into the clusters as the output is prone to come with more noise than the original attributes tend to give as the clustering is based on density of the clustering attributes.

T.Thamarai Selvan [26] proposed a new algorithm known as optimal pulse system measurement algorithm and made it venal into the programmatic application which is very efficient in transfiguring data to its core binary format with lose-less nature and plot graph which can be exported into any Big data applications for further analyzing. The algorithm used in this paper can be referenced for creating a clustering model running big data applications as it focuses on a schema where the data is processed in real time and processed.

Cai et al. [27] has proposed a method known as RMKMC(Robust Multi View k-means Clustering). This method uses Cluster Indicator Reformation and structured Sparsity-Inducing Norm which is used to solve the issues arising in graph construction in multi view clustering algorithm and single view clustering in k-means algorithm.

Woo et al. [28] proposed an algorithm called as the wTabular–algorithm which does the reduction of rules which are used less by considering them unimportant by assigning a weighted value to them. It also employs a second method called the Quine – Mccluskey for rule reduction thus improvising the performance

Sergej et al. [29] proposed a novel approach based on MapReduce based implementation with P3C+ - MR algorithm which has highest accuracy. This algorithm showed good results for large datasets of value.

CONCLUSION

The social media platforms are generating more and more datasets which when analyzed using correct techniques and methods will provide more insightful predictive information. When such techniques are employed along with data mining techniques, the potential is so huge that every single piece of data will turn into an insightful information aiding in the way leading to a better and efficient predicament in future trends.

REFERENCES

[1] J. Gantz and D. Reinsel, ''The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,'' in Proc. IDC iView, IDC Anal. Future, 2012.

[2] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1–137.

[3] K. Cukier, ''Data, data everywhere,'' Economist, vol. 394, no. 8671, pp. 3–16, 2010.

[4] T. economist. (2011, Nov.) Drowning in Numbers—Digital Data Will Flood the Planet- and Help us Understand it Better [Online]. Available:http://www.economist.com/blogs/dailychart/2011/11/bigdata

[5] Wei Fan and Albert Bifet. Mining Big Data: Current status, and Forecast to the Future. SIGKDD Explorations 14(2):1-5, 2012.

[6] Junyu Xuan; Xiangfeng Luo; Jie Lu, "Mining Websites Preferences on Web Events in Big Data Environment," Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on , vol., no., pp.1043,1050, 3-5 Dec. 2013 doi: 10.1109/CSE.2013.152

[7] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. Vol. 2, No 1-2 (2008) 1–135.

[8] Sanjeev Pippal, Lakshay Batra, Akhila Krishna, Hina Gupta, Kunal Arora, Data mining in social networking sites : A social media mining approach to generate effective business strategies, International Journal of Innovations & Advancement in Computer Science(IJIACS), Vol 3, Issue 2, April 2014.

[9] G Nandi and A Das, A survey on using data mining techniques for online network analysis, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013.

[10] D.Heckerman, H.Mannila, D.Pregibon, and R.Uthurusamy, editors. Learning bayesian networks: the combination of knowledge and statistical data. AAAI Press,1994.

[11] D.Heckerman, H.Mannila, D.Pregibon, and R.Uthurusamy, editors.Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97). AAAI Press,1997.

[12] "Apache Hadoop" [Online]. Available:http://hadoop.apache.org. [Accessed: 16-Sept-20I6].

[13] Toshniwal, Ankit, et al., "Storm@ twitter," Proc. ACM SIGMOD. International conference on Management of data, July 2014, pp. 147-156.

[14] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari, "S4: Distributed stream computing platform," Proc. IEEE. International Conference on Data Mining, ICDM, Dec. 2010, pp. 170-177.

[15] Qian, Zhengping, et al., "Time stream: Reliable stream computation in the cloud," Proc. ACM. The 81h European Conference on Computer Systems, Apr. 2013, pp. 1-14.

[16] Xindong Wu; Xingquan Zhu; Gong-Qing Wu; Wei Ding, "Data mining with big data," Knowledge and Data Engineering, IEEE Transactions on, vol.26, no.1, pp.97,107, Jan. 2014 doi: 10.1109/TKDE.2013.109

[17] Xiao Cai, Feiping Nie, and Heng Huang. 2013. Multi-view K-means clustering on big data. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI'13), Francesca Rossi (Ed.). AAAI Press 2598-2604

[18] Rahul Sharan Renu, Gregory Mocko, Abhiram Koneru, Use of Big Data and Knowledge Discovery to Create Data Backbones for Decision Support Systems, Procedia Computer Science, Volume 20, 2013, Pages 446-453,ISSN1877-0509http://dx.doi.org/10.1016/j.procs.2013.09.301.

[19] FAHAD, A; Alshatri, N.; Tari, Z.; Alamri, A; Y.Zomaya, A; Khalil, I; Foufou, S.; Bouras, A, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis," Emerging Topics in Computing, IEEE Transactions on , vol.PP, no.99, pp.1,1 doi: 10.1109/TETC.2014.2330519.

[20] A K Jain, M N Murty, P. J. Flynn, 'Data Clustering : A Review', ACM COMPUTING SURVEYS , 1999

[21] B.A Tidke, R.G Mehta, D.P Rana, "A Novel Approach for High Dimensional Data Clustering" in International Journal of Engineering Science and Advanced Technology (IJESAT) ISSN 22503676 Vol.02(3) May-Jun 2012

[22] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: an efficient data clustering method for very large databases. SIGMOD Rec. 25, 2 (June 1996), 103-114. DOI=10.1145/235968.233324 http://doi.acm.org/10.1145/235968.233324

[23] Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3 (March 2003), 583-617. DOI=10.1162/153244303321897735 http://dx.doi.org/10.1162/153244303321897735

[24] J. C. Bezdek, R. Ehrlich, and W. Full. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2):191–203,1984.

[25] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. Proc. of the ACM SIGKDD Conference on Knowledge Discovery ad Data Mining (KDD), pp. 58– 65, 1998.
[26] T.Thamarai Selvan, Glidersoft, "Nadi Aridhal: A pulse based automated diagnostic system", Electronics Computer technology (ICECT), 2011 3rd International Conference, April 2011
[27] Xiao Cai, Feiping Nie, and Heng Huang. 2013. Multi-view K-means clustering on big data. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI'13), Francesca Rossi (Ed.). AAAI Press 2598-2604
[28] Woo Sik Seol; Hwi Woon Jeong; Byungjun Lee; Hee Yong Youn, "Reduction of Association Rules for Big Data Sets in Socially-Aware Computing," Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on , vol., no., pp.949,956, 3-5 Dec. 2013
[29] Sergej Fries, Stephan Wels, Thomas Seidl, Projected Clustering for Huge Data Sets in MapReduce, Proceedings of the 17th International Conference on Extending Database Technology (EDBT), pp.49-60 2014.