International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4 (2016) pp 2411-2416

A Survey on Machine Learning Approaches to Social Media Analytics

Magesh. S
Assistant Professor, Department of Information Technology,
SRM University, Kattankulathur, Kancheepuram, Tamil Nadu, India.

Nimala. K
Research Scholar, Department of Information Technology,
SRM University, Kattankulathur, Kancheepuram, Tamil Nadu, India.

Abstract
Social networks have greatly influenced the usage pattern of Internet by common man. It has paved way for the great proliferation of Internet into the life of even the ordinary people. Such an exponential rise in the use of the world wide web due to social networks started to fetch huge volume of data across diverse domains in short period of time. These characteristics by which the huge amounts of social network data are generated make them to categorize as Big Data. Since the use of big data analytical techniques in many domains have obtained remarkable improvements in the way various businesses operate, we consider that social network domain is not an exception to this perception. Though there are many literatures available to deal with big data and social networks separately, only few papers deal with the analytical techniques of big data algorithms in the field of social networks. Hence, this paper will serve to act as a survey of big data analytical techniques as applied to social networks and efforts are taken appropriately to present the pros and cons of each algorithm. We also present the suggestions for use of various algorithms to various classes of social network data

Keywords: Big Data, Hadoop, Map reduce, Decision tree, ID3, C4. 5, SVM, ANN, K-NN etc

Introduction
The onset of Big Data Analytics has created a revolutionary change in the way businesses and people perceive data. [1, 2] Big data has proved too costly for rival organizations in the way the decisions have been undertaken. There exists a common misconception among the society that any huge volume of data can be classified under big data. But the reality is that data generated by systems which fulfills the constraint imposed in the dimensions of Volume, Velocity, Variety and Veracity can be classified as big data[3] as shown in Figure 1 When any system or organization generates data at a higher rate with diverse attributes and that data which when used to make learned decisions will fall under the hood of big data. The processing of big data is different from the way how conventional data is handled. Big data processing requires certain special techniques and tools to accomplish effective analytical decisions. The proliferation of Internet in the nook and corner of the globe has made the presence of big data felt everywhere. This made the data scientists to use these big datasets towards provisioning informed decisions.
Organizations that generate big data, utilize the diverse tools and techniques to carry out sentiment analysis, predictive modeling, semantic web analysis to improve the Return on Investment (RoI). This massive buzzword "big data" is used by organizations to understand and target customers, optimize business processes and performance quantization in various domains such as health care, hospitality industry, education and learning systems, automobile, industrial automation systems, agriculture etc [4]. As a small step towards extending the footprint of the applications of big data, this paper tries to depict the machine learning techniques to perform Social network analytics that may provide a 360 degree insight into the social network data.

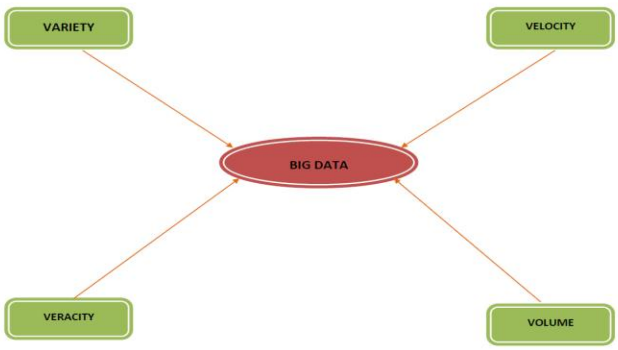

Figure 1: Attributes of Big Data

The term machine learning aptly denotes that, the system is made to learn by providing necessary inputs and carefully examining the obtained outputs. Machines can learn under different circumstances namely, Supervised, Unsupervised and Reinforcement [5]. Machine learning is a subfield of computer science that evolved from the computational learning theory in Artificial Intelligence. Machine learning algorithms help us to make effective predictions based on big data, upon which the operational convenience of a business can rely. Various machine learning tasks are categorized depending on the desired output of the machine learned system. They are Classification, regression, clustering, density estimation and dimensionality reduction. Based on the properties of the data obtained in the learning environment, machine learning algorithms perform predictions. The applications of machine learning are as diverse as the applications of big data. Adaptive websites, Bio informatics, Computational advertising, Information retrieval, credit card fraud detection, medical diagnosis, Natural language processing, stock market analysis are some areas where machine learning has found its use as shown in Figure 4.

Section II provides information on various big data analytics framework that are available to perform efficient social network analysis

As mentioned above, the various approaches to machine learning algorithms in the context of usability in social network analytics domain will be dealt with, in detail, in section-III of this paper.

Section IV deals with the applicability of machine learning techniques in different big data frameworks by providing a ready reckoner. Section V and section VI focuses on Conclusion and Future Work and References respectively.

Big Data & Social Network Analytics

Social networks such as Face book, Twitter, LinkedIn etc pave way for generation of huge amount of diverse data in short period of time. Such social media data require the application of big data analytics to produce meaningful information to both information consumers and data generators. The impact of different big data analytics tools and techniques over processing social network data will be discussed in detail in this section of this paper.

Big data analytics techniques and tool types include all of the following such as Predictive analytics, data mining, statistical analysis, complex SQL, data visualization, artificial intelligence and natural language processing. The analysis of structured and unstructured data from social networks leads to social network analytics[6]. Even blogs, micro blogs and wikis contribute to social network analytics data sets. Though there are various sources of information available in social media, we are largely concerned about the user generated contents such as sentiments, images, videos and bookmarks and interactive relationships between people, organizations and products. These two classes of information is utilized in various big data analytics tool such as Hadoop and Map Reduce Framework, Apache Pig, Apache Hive, Jaql, NoSQL etc. When user posted information is used in the analytics approach, it is called as content based analytics and when relationships between entities is used for analytics, it is known as structure based analytics. Social networks consist of millions of connected objects and analysis of data from those objects is computationally intensive and expensive. Hence there are two different approaches that shall be followed. They are parallelization approach and Graph databases approach [7] as shown in Figure 2. In parallelization approach, focus is towards dividing a huge data set into smaller sub sets and utilize the computational power through cloud computing to process the data in a parallel manner. Map Reduce and Pregel from google are pioneer in this approach. However, lots of open source initiatives in the form of Hadoop are gaining popularity in the social network analytics. Spark and Hama are also registering their market share in the research of social network data. [8]

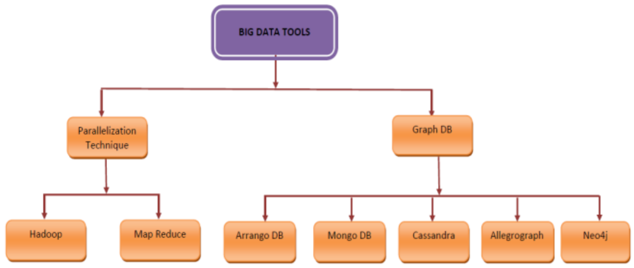

Figure 2: Categories of Big Data Tools

Map reduce framework consists of Map phase and Reduce phase which uses Key/Value pairs and Key-Value List pairs respectively. Any mapreduce application contains various hotspots such as Input Reader, Map, Partition, Compare, Reduce, Output Writer. Application of Map reduce is considered to enable the scalability of social networks, for the determination of graph based metrics. This application is used to determine the betweenness centrality. The chaining of Map reduce jobs in social network analytics is carried out for the estimation of shortest paths in a graph. [9, 10]. Blocking mechanism is an important part of Map Reduce that deals with machine failures in the application of social network data.

Apache Hadoop, as known for its fault tolerance is extensively used in social network analysis [11]. Hadoop is massively scalable without the need to change the data format, fault tolerant, cost effective and flexible. Hadoop also enables distributed processing of large social network data sets across clusters of commodity servers. The application of hadoop consists of implementing the map reduce paradigm on top of Hadoop distributed File System (HDFS). Researchers have utilized hadoop to analyze social network data along with machine learning techniques of MAHOUT and GIRAPH. [12]

Apache PIG[13] is utilized to process large scale social media datasets. Pig platform contains a high level language to describe data analysis programs coupled with infrastructure for program evaluation. The two components inside Pig is PigLatin and runtime environment for Pig Latin. The demonstrated use of PIG has been proven with twitter feeds and the tweet based Pig program can be embedded inside a script or java program and the program can be run from Grunt, Pig based command line[14].

HIVE [15]supports large datasets stored in HDFS and Amazon S3 systems. HIVEQL assists in converting the queries into Map/Reduce jobs, Apache Tez and Spark jobs. Apache Hive stores metadata, obtained from data processing in the embedded database known as Apache Derby and if needed, MySQL databases can also be used. Many features of HIVE as given below make it suitable for social network analytics along with Map reduce framework[16]

a. Faster processing based on indexing
b. Diverse file format support such as plaintext, ORC, RCFile and Hbase etc.
c. Compressed data processing using algorithms such as DEFLATE, Snappy, BWT.

Besides Hadoop and Map Reduce, there are other data analysis techniques that lend their hand to big data processing. Graph databases are considered to be well suited for networks like social networks, transportation networks, where the data structure directly represents the key aspects of the problem. Every element of a graph database has a direct pointer to its adjacent element and no index lookups are necessary. Node relationship is highly useful in dealing with highly interconnected social network data[17]. Graph databases are relatively faster compared to relational databases when dealing with associative datasets. Graph databases scalability to large data sets is high as they do not require expensive join operations [18]. These properties make graph databases, an excellent choice when dealing with social network analytics. Wide range of graph databases such as Allegrograph, ArrangoDB, MongoDB, Cassandra, Neo4j, Mapgraph, Oracle NoSQL database are used in social network analytics[19].
Among the available graph databases, MongoDB [20]has been used in social network analytics, since it is a cross platform document oriented database and classified as NoSQL database. The features of MongoDB such as document oriented, ad hoc queries, indexing, replication, load balancing, file storage made it an appropriate tool for social network analytics. MongoDB, considered as a " Cool Vendor in Information Infrastructure and Big Data" helps huge interconnected networks like twitter, LinkedIn, face book to be more scalable and agile. The practical application of MongoDB towards social media trend analytics is proven with a social network called Untappd[21]. Untappd, is a social network for beer lovers. MongoDB helped Untappd address problems with social feed and sustenance of high performance with 5000 to 6000 queries per second with the use of location indexes.
Easy scale outs, high write throughputs and low costs were the key for use of Cassandra [22] in social data analytics. Works with the help of Cassandra and Hadoop were carried out to prevent suicide by Scale Unlimited. When scalability, beyond a point is considered for a business, Cassandra overrides MongoDB due to its feature of high write throughputs. Cassandra also performs auto balancing as the data can be spread across nodes. A publisher analytics product using Cassandra recently helped organizations to understand which social media channels its website readers were using to share content. Cassandra's use in social network analytics increased because of its ability to aggregate most statistics as they were written rather than aggregating them with map/reduce frameworks. eBay uses Cassandra for managing time series data. There are proven test results which are available with Marked UP, to prove the efficiency of Cassandra in social network analytics compared to MongoDB. The test results provided the researchers with necessary configurations and use cases which shall be used by them in tuning the performance of Cassandra further for social network analytics. Most of the social network analytics tools which used MongoDB ran out of gas even before Cassandra could do so.
Java based open source graph database, that has witnessed wide spread applications is Neo4j[23]. In Neo4j, everything is considered either as a node, edge or an attribute. Every node and edge can have number of attributes and shall be labeled. A data model for time varying social network with the help of Neo4j graph database is provided in [reference to be provided]. The biggest con of Neo4j is that it cannot work directly with HDFS or HBase. Though, Neo4j has been used in HR analytics [43], its use in social network analytics is gaining popularity because of its pure graph like data structure support.

Machine Learning & Social Network Analytics

Having discussed about the big data applications in social media data in the previous section of this paper, this section will concentrate on provisioning the details of machine learning techniques as appropriately applied to social media.
Machine learning techniques, as implied by the term, is the process of inculcating knowledge to any machine like, PC, laptop or mobile devices to learn about a system with a set of input /dependent variables and the desired output. Any machine can perform learning under three modes. They are Supervised, Unsupervised and Reinforcement learning techniques [24]. Normally, machine learning techniques are employed in any system to carryout and produce results as part of predictive analytics and forecasting methods. Any machine learning techniques will be classified under the categories of Decision tree based, linear and logistic regression based and neural network based. Many organizations have kick started to utilize the impact of social media data in the decision making process. When social media data is utilized for such a critical decision making, it becomes necessary to process the huge datasets obtained from social networks using machine learning techniques. This will help organizations to foresee certain situations and decide based on the output of the social media analytics. The key aspect of any machine learning technique is iteration. This iterative aspect will make the system to independently adapt to new sets of input as they will be continuously subjected to variety of datasets. The advent of new computing technologies like big data have created a revolution in the machine learning domain, that complex mathematical calculation can be applied to heterogeneous huge datasets. [25]
Machine learning algorithms that have played a major role in social media analysis include Decision tree learning, Naïve Bayes, Nearest Neighbor classifier, Maximum Entropy method, Support vector machine(SVM), Dynamic Language Model classifier, linear regression and logistic regression, Simple logistic classifier, Bayes Net and Multilayer Perceptron as shown in Figure 3

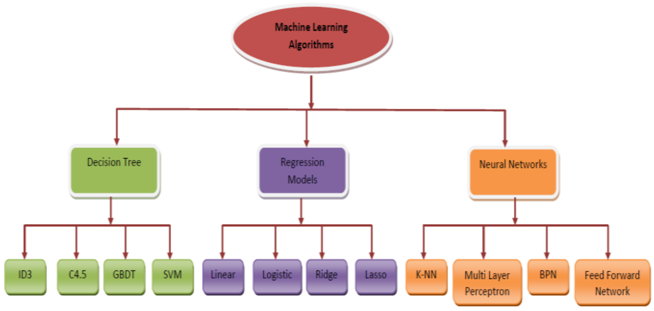

Figure 3: Algorithmic Approaches to Machine Learning

Upon carrying out literature research, it becomes quite evident that considerable amount of work has been carried out in the social network analytics field [26, 27, 44] utilizing the decision tree learning mechanisms. Decision tree learning uses decision trees to predict the values of a target variable and relate the same to the observations of that variable. Two types of trees can be built using a decision tree learning mechanism namely, classification trees and regression trees. Classification trees provide finite set of values to the target variables and regression trees provide continuous values to the target variable. In social network analysis [28], decision tree learning has been utilized to profile users based on their relationship with other users, and depending upon the decision tree obtained, clustering of users can take place. Two important algorithms that employ top down, greedy search through the space of decision trees are ID3 [29] and C4. 5 [30]. The working principle of ID3 algorithm is that it learns decision trees by constructing them top down and starts at the top of the tree and then decides on the attribute to be tested. C4. 5 is an extension of ID3 algorithm and it builds decision trees based on the concept of information entropy and a set of training data. Decision tree has been used to obtain the rules that govern the relationships among users in the online social network. These decision trees are also used to discover interesting patterns among the users. [31]. Gradient Boosted Decision Trees (GBDT) is used in classification of users based on certain attributes in social networks. GBDT is proved to provide much smaller decision trees and reduced decoding compared to Support Vector Machines (SVMs). [32].

Distributed Decision Tree Learning for mining big data streams is an important work carried out in this field. Mahout, a distributed decision tree learning framework utilized the volume and variety dimensions of big data while Massive Online Analysis (MOA), a streaming decision tree learning framework, took care of the velocity and variety dimensions. [33, 34] Corresponding references to be provided]. Scalable Advanced Massive Online Analysis (SAMOA) is designed to address all the three dimensions of big data. Flexible APIs are provided by SAMOA using latest ML algorithms to take care of variety aspect. Storm, a state of the art Stream Processing Engine (SPE) is integrated with SAMOA to address the volume and velocity issues.

Naïve Bayes classifier is a simple probabilistic classifier because of the simple mathematics involved. It is primarily used for text classification lies towards email spam detection, language detection, sentiment detection and sexually explicit content detection. A preliminary Research Project has been carried out in Cairo University where a proof of concept system is developed. In that system the input is given in Arabic language data stream and the tasks of topic extraction, subjectivity and sentiment analysis were carried out. In this project, for classification tasks, Naïve Bayes and Support Vector Machines were alternatively used and results show that SVMs outperform Naïve Bayes Techniques. [35]. Sentiment analysis and topic detection in Natural Language Processing witness an increasing trend of usage in social network analytics. Social media sentiment analysis and topic detection for Singapore English, [36] utilized Naïve Bayes Classifier along with Support Vector Machines, Labeled Latent Dirichlet Allocation (LDA) and Maximum Entropy. The comparative results shown that for supervised sentiment analysis, Naïve Bayes classifier delivered good accuracy compared to others.

Nearest Neighbor Rule, a supervised statistical pattern recognition method, achieves high consistency without apriori assumptions about the distributions from which training samples are drawn. The approaches to improve the speed and performance of nearest neighbor classification technique are to pre sort the training sets based on voronoi cells, and to choose a subset of training data such that 1-NN approximates to Bayesian classifiers. The biggest problem of NN rule is huge computational intensity. K-NN techniques are selected in social media analytics when there is less knowledge about the distribution of data. Classification accuracy of three algorithms namely, SVM, K-NN, and Naïve Bayes were compared with the help of DBLP data, largest bibliographic dataset of Computer Science Publications using a set of features like sum of papers, weighted sum of neighbors, weighted sum of secondary neighbors and weighted shortest distance. [37].

Support Vector Machine, a machine learning algorithm works with labeled training data and output the results to a hyper plane. SVM is usually when the features of input data are less. When the features grow exponentially, it is not optimum to use SVM but to adapt logistic regression. SVM has witnessed applications in the area of medical imaging, text analytics and image recognition. The work of Rittermann et al. [38], was first of its kind to use twitter data to analyze whether an infectious disease would become pandemic based on support vector machine regression. The mentioned work utilized the data from HudDub, an online prediction market.

Language model classifiers are used to sort test data into two categories. An intern at Lingpipe, has worked on Language model classifiers based on twitter data. [39]. LM classifiers need to be trained on annotated data before they can perform the exact classification. Evaluation of customized and generic language classifiers were carried out using training data from twitter based on the Corpus training section.

Linear and logistic regressions are approaches predominantly utilized in predictive analytics. Linear regression deals with obtaining a perfect measure of fit for dependent variable or output based on the linear function of the input parameters. Logistic regression deals with estimating a regular multivariate model in which transformation of information about a dependent variable into an unbounded continuous variable take place. Proven work has taken place with respect to predictive analytics for fraud detection (Fraud analytics). One of the early works in the field of social network analytics utilized logit models and logistic regression techniques based on graph theory. [40]. Logistic and linear regression models are easily understood by the businesses as stated by a Manager of PwC, unlike the neural network based models which requires some amount of technical knowledge. Logistic regressions are applied to problem domains where normally the target variable exhibits binomial nature. These regression models are used to determine which independent variable from the input is influencing the dependent variable/output to a greater extent. On a broader note, logistic and linear regression models are used by organizations to decide on the

parameter that maximizes their profit or Return on Investment (RoI) [41].
Multilayer perceptron is a class of Artificial Neural Network (ANN), which works on the principle of feed forwarding and it creates a model of input and output variables used for machine learning. Every layer in MLP is connected to the next layer in a complete manner using directed graphs. Though MLP is considered to be a feed forward network, it utilizes the principle of back propagation algorithm towards training the network. The application of multilayer perceptron can be well found in the context of social network analytics in clustering the social community based on social media data. A multilayer perceptron approach is not utilized in creating customer intelligence in social media data since they could not provide a meaningful training data for the network to learn [42]
Though there exist a number of machine learning techniques, only few are applied in the context of social network analytics. Also, certain techniques that were applied in social analytics need some modifications either to the environment or to the learning technique to improve the efficiency of the output. Hence, this section will surely serve to be an eye opener for researchers concentrating in the domain of social network analytics.

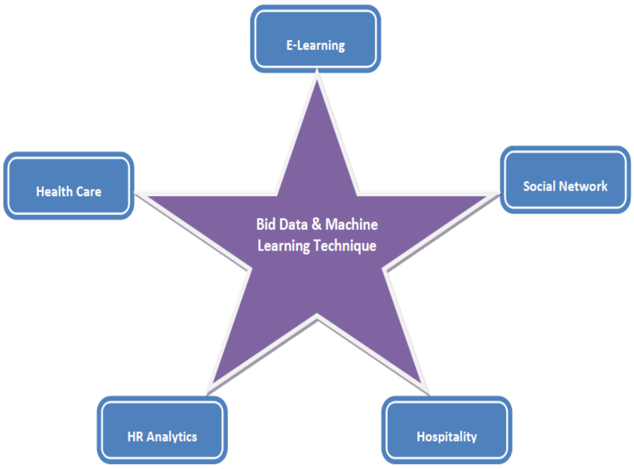

Figure 4: Application Domains of Big Data and Machine Learning Techniques

Conclusion and Future Work

Though the domains of big data and machine learning had witnessed huge attention from the research fraternity separately, there are only few works that had taken the above domains together. In this paper, necessary information on big data in social network analytics and machine learning in social network analytics has been surveyed. We know that this survey will definitely provide open areas where more attention is needed from researchers in these domains in a smaller magnitude, if not in a huge manner. This paper addressed the importance of big data and machine learning in the first section along with the growing importance to analyze growing social media data utilizing big data and machine learning algorithms. The next section discussed about the various parallelization frameworks in the form of Hadoop and Map Reduce along with different Graph databases in the context of social network analytics. Various machine learning algorithms applied in the field of social data analysis with pros and cons is dealt with in the third section of this paper.
With these discussions, the scope to carry out further work by integrating big data and machine learning techniques towards social network analytics has been widened. Various performance enhancement techniques by modifications to existing machine learning algorithms in configured scenarios and comparative performance optimizations for different classes of graph databases when utilized in social network analysis are considered to be the main stay of future work.

References

[1] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.
[2] Will Dobbie, Roland G. Fryer, Jr. "*Getting Beneath the Veil of Effective Schools: Evidence from New York City*. " NBER Working Paper No. 17632. Issued Dec. 2011.
[3] www. ibm**bigdata**hub. com/infographic/four-vs-**big-data**
[4] Tania Cerquitelli, Silvia Chiusano, MirkoKampf, "Special Session on Big Data: New Trends and Applications", BigData 2013.
[5] Lloyd Allison, "Types and classes of machine learning and data mining", Proceedings of the 26th Australasian Computer Science Conference, Vol. 16, Page No. 207-215, 2003.
[6] Jonathan Magnusson, "Social Network Analysis Using big Data technology", Uppsala Universitet, January 2012
[7] Charu. C. Aggarwal, "Social Network Data Analytics", IBM Watson Research Center, USA
[8] Ostrowski D. A, "Map Reduce Design Patterns for Social networking Analysis" Page nos. 316-319, IEEE International conference on Semantic Computing, 2014
[9] Guojin Liu, Ming zhang, Fei Yan, " Large Scale social network analysis based on MapReduce", International conference on computational aspects of social networks, Page Nos. 487-490, 2010
[10] Nan Gong, "Using Map-Reduce for large Analysis of Graph Based Data", M. S. Thesis, 2011.
[11] Pradeepa. A, Dr. Antony. S, "Significant Trends of Big Data Analytics in Social Network", In Journal of Advanced Research in Computer Science and Software Engineering, Vol-3, Issue 8, August 2013.
[12] Arinto Murdopo, "Distributed Decision Tree Learning for mining Big Data Streams", M. S. Thesis, University Politenica de Catalunya, Barcelona Tech, July 2013.
[13] Mirko Kampf, " Hadoop in Social Network Analysis-Overview on tools and some best practices", Gridka School, Karlstuhe, 2013.
[14] Kyar Nyo Aye, Ni Lar Thein "A Comparison of Big Data Analytics Approaches Based on Hadoop

Mapreduce", International Conference On Computer Applications, February 2013.
[15] Doctuts Read the docs. Org/en/latest/hive. html.
[16] Yin Huai, Ashutosh Challhan, Alan Gates, Gunther Hagleitner, Eric N. Hanson, Owen O' Malley, Jitendra Pandey, Yuan Yuan, Rubao hee, Xiaodong Zhang, " Major Technical Advancements in Apache Hive", Proceedings of ACM Sigmod, June, 2014.
[17] Anteneh Ayanso, " Harnessing the power of Social Media and web Analytics", IGI Global, February 2014.
[18] https://en. wikipedia. org/wiki/Graph –Database.
[19] William M. Campbell, Charlie K. Dugli, Clifford J. Weinstein, "Social Network Analysis with content and Graphs", Lincoln Laboratory Journal, Page Nos 62-81, Vol 20 No. 1, 2013.
[20] Yash Badiani, Nirav shah, " Social Media Analytics Enabled by Big Data portal using Life ray and MongoDB", Gignex Datamatics Open Source Work, September 2014.
[21] www. rachspace. com/blog/tapping-scalability-and -good-beer-with-objectrocket-and-mongodb/
[22] www. informationweek. com/strategic-cio/why-we-picked-cassandra-for-big-data/a/d-id/1318250.
[23] neo4j. com/developer/graph-database/
[24] K. L. Du, M. N. S. Swamy, "Neural Networks and statistical Learning, " Springer-Verlag London, 2014.
[25] Steve Oberlin, " Machine Learning Cognition and Big Data", CA Technologies, 2012.
[26] Y. Ku, Y. Chen, C. Chiu, " A proposed Data Mining Approach for Internet Fraud Detection" LNCS, Springer, Vol : 4430 pp 238-243, 2007.
[27] J. Cheng, D. Romero, B. Meeder, J. Kleinberg, " Predicting Reciprocity in social Networks".
[28] E. Trandafili, M. Biba, A. Xhuvani, " Profiling Social Network Users with Machine Learning", BCI, September 2013 pp. 14-20.
[29] https://en. wikipedia. org/wiki/ID3-algorithm.
[30] B. Hassina, A. Merbarha, H. EZZ, Kouri, M. ERRITALI, " A Comparative study of decision tree ID3 and C4. 5", International Journal of Advanced Computer Science and Applications special Issue on Advances in Vehicular Ad Hoc networking and Applications pp-13-19.
[31] R. D. Chiang, M. Y. Chang, H. C Keh, C. H. Chan, " Mining Unexpected Patterns by Decision Trees with Interestingness Measures", IEEE International Conference on Artificial Intelligence, Modelling Simulation, 2013, pp-104-109.
[32] D. Metzler, T. Kanungo, " Machine Learned Sentence Selection strategies for Query-Biased Summarization", SIGIR, L44IR Workshop, Singapore, 2008.
[33] B. R Prasad, S. Agarwal, " Handling Big Data Stream Analytics using SAMOA Framework-practical experience", International Journal of Database Theory and Application, Vol. 7 No. 4. pp. 197-208.
[34] D. Bhatacharrya, M. Mitra, " Analytics on Big Fast Data Using Real Time stream Data processing Architecture", EMC proven professional Knowledge sharing, 2013.
[35] Amir Atiya, Mohammed Aly, Mahmoud Nabil, "Social Media Analytics", PRP Project, Faculty of Engineering, Cairo University, 2015.
[36] Yee Ling Phua, "Social Media Sentiment Analysis and Topic detection for Singapore English", Thesis, Calhourn Postgraduate school, September 2013.
[37] Elena pupazan, " Social Networking Analytics", Bussiness Mathematics and Informatics paper, Vu University, Amsterdam.
[38] J. Ritterman, M. Osborne, E. Klein, " Using prediction markets and twitter to predict a swine flu pandemic", 1st International Workshop on Mining Social Analytics, 2010.
[39] lingpipe-blog. com/2011/07/25/language-id-said-the-ngram-boundary-dynamics-language-model-classifier/
[40] J. Holton Wilson, " An analytical Approach to detecting Insurance Fraud Using Logistics Regression", Journal of Finance and Accountancy.
[41] http://analyticstraining. com/2012/regressiuon-modelling/
[42] R. Silipo, P. Winters, K. Thiel, T. Kotter, "Creating Usable Customer Intelligent from Social Media Data by clustering the social Community", Whitepaper, KNIME. COM, 2012.
[43] Rik Van Bruggen, " Graphs for HR Analyticss", September 2014 (neo4j. com/blog/Using –neo4j-hr. analytics)
[44] H. Zhang, R. Dantu, " Predicting Social Ties in Mobile phone Networks" IEEE ISI, May 2010.