Feedback Analysis of unstructured data from Social Networking a Big Data Analytics Approach

Ishwarappa Kalbandi[1], Dr. J Anuradha[2]
[1, 2] Department of Computer Engineering
[1] Dr. D. Y Patil Institute of Engineering,Management & Research Akurdi, Pune, India
[2] Vellore Institute of Technology, Vellore, Tamilnadu, India

Abstract- *With the vast proliferation of online applications and cloud resources, there is a tremendous demand in usage of social networking platforms. Start from entrepreneur, players, politicians, students, or anyone are highly depending on social networking sites for example Face book, Twitter which generates lots of massive data that are not only challenging but also highly impossible to understand. The primary reason behind this is big data is massive in size and they are greatly unstructured. Because of this fact, the proposed system attempts to understand the possibility of performing knowledge discovery process from Big Data using conventional data mining algorithms. Designed in Java considering huge number of online educational data from social networking sites, the proposed system evaluates the usefulness of performing data clustering and data mining for Big data.*

Keywords- component; Big Data, data mining, clustering

I. INTRODUCTION

With the advancement of networking as well as communication technologies, social networking is increasing its pace [1] among global user. Normally, a social network (SN) consists of various individuals from various part of the world connected to each other on the basis of certain common interest. Social Network Analysis (SNA) is a specialized field of engineering that deals with investigation of various unique patterns and behavioral study of various users connected with each other in SN. The phenomenon's of social relationship among the users are considered in SNA using networking principles that comprises of nodes depicting users of SN and connectivity existing among them. A specialized social network diagram is used for such study and investigation purposes. Various researchers have started emphasizing on the social network for extracting more relevant and latent information that could be of any use of commercial improvement, business benefits, or improvement of user's experience. There could be various reason behind the utility of such massive data being captivated by the social networking applications. The service provider of various social networking applications therefore witnesses a tremendous rise of customer base leading to accumulation of various transactional as well as behavioral data on their database. Therefore, emergence of data mining technique evolves is one crucial elements to provide certain value added services to such massively growing data of social network. There are various data mining approaches already in use by social networking applications e.g. classification algorithms, clustering algorithms, association rule mining etc. Not only this, but certain standard techniques of data mining e.g. reconstruction based techniques, heuristic based techniques, cryptographic based techniques etc have already being used in the extensive study of social network from past decade. With the evolution of cloud platform, the issues of storage of such massive and bulky data of social networking application could be mitigated. However, design a new data analytical tool to extract more relevant information as knowledge discovery process is still a challenging problem of SNA. One of the biggest issues of deploying conventional data mining techniques on cloud environment is security and privacy vulnerabilities. Hence, data mining techniques have to be deployed in a very efficient manner so that without any security issues; it supports in identifying the patterns as well as trends from massive data sets from SNA. The proposed system discusses about the massively generated data from the educational sector. The justification of the topic in done in this manner-at present day, the adoption of various social networking applications are constantly on rise among the student community who shares various sorts of information using SN. Such educational information are highly diverse in nature, where the file types could be word file, excel file, presentation file, executable files, and various other formats. Such educational sharing process was also found to be adopted by various teaching community, thereby redefining the global education system. Hence, the usage of such social networking application frequently populates the massive data termed as Big Data over the cloud, which is required to be further studied for excavating the precise knowledge that will required by policymakers to take certain business decision.

This paper presents such a cost effective model that performs data mining over educational data captured from social networking application. Section 2 discusses about the related work followed by discussion about Big Data in Section 3. Section 4 discusses about the problem identification while Section 5 discusses about the proposed model that uses data mining over the educational data on cloud environment. and finally Section 6 summarizes the paper.

II. RELATED WORK

There have been certain volumes of research papers witnessed in past 5 years focusing on investigating the issues on social network analysis. Xie and Szymanski [2] have presented a rule and propagation factor to enhance the data mining techniques with respect to computational efficiency as well as detection of online communities. Same authors have presented a modified version of their prior techniques for large scale network [3]. Similar problems have also been discussed by Xie and Boleslaw et al. [4] who have particularly emphasized on the overlapping nodes in social network analysis using dynamic interaction rules. The problem of overlapping communities in social networking analysis was also studied by Gregory [5] exclusively on large network structure. Hu et al. [6] discussed a technique about identifying the essential communities using signal processing concepts in complex networking principle. Furthermore, Wakita et al. [7] have presented a setoff attributes for process control of online community analysis evaluated on 5.5 millions of users. The community structure and its explicit measurement of strength were studied by Newman et al. [8]. Blondel et al. [9] discussed about a heuristic technique for extracting the structure of the community for large network based on modularity organization. Work in similar direction of identifying community was also discusses about Capocci et al. [10] using spectral method that emphasize on weights and link orientation. Studies towards optimization techniques on spectral relaxation problems in social networking analysis were put forward by White and Symth [11]. An extended study towards community discovery was conducted by Newman and Girvan [12] using graph theory. Studies toward issues evolving from merging communities were carried out by Schuetz and Caflisch [13] who have introduced a multi-step greedy approach for the purpose of resisting premature condensation of large communities. Fortunato [14] have carried out an extended experiments to prove efficiency of clustering techniques on social network on real-time environment. Exclusive study using graphical tool was carried out by Leskovec et al. [15] who have presented a novel graph generator to extract the latent charecteristics of each node in social networks. Investigation on issues pertaining to real-time identification of online communities was performed by Leung et al. [16], who have also presented a new algorithm based on epidemic detection of community. Adoption of centrality metric for identifying community was seen in the study of Lu et al. [17]. Advance mathematical concepts were also seen to be attempted towards normalizing the studies towards social network analysis by Tang [18] and Darmon et al. [19]. However, none of the studies were found to discuss about the explicit effectiveness or contribution of data mining with respect to cloud and massively growing data exhibiting as a big trade-off for the domain of SNA.

III. BIG DATA

With the constant rise of advancement in the internet technology, ubiquitous computing is also growing exponentially and reaching the end customers day by day. The information is now highly distributed and gain extreme higher degree of mobility due to ubiquitous computing. Hence, it becomes easier for the user to access their resources from multiple computing devices, even on the move. Such communication pattern is also bidirectional, which means that user not only accesses the resources but also share their own resources on the other end. Hence, it can be said that on every seconds, millions of data are being generated and stored in server. This is how the data are growing in every seconds of life, which is now technically called as ‘Big data.’ However, the question arises is how far successful we are in retrieving some significant information (knowledge discovery) from the big data. As the information carried by big data is highly valuable for business goals and it gives a cut edge prediction capability to the organization by excavating some more secrets about their data which cannot be explored by conventional datamining or data warehousing techniques.

- **Educational Sector:**

The trend of higher adoption of technical classes, archival of documents for every session, lecture notes, feedbacks generated by students, instructors, as well as by critics are constantly on the rise to meet up the quality standards of education system for any country. Such big data is emphasized not only from storage viewpoint but also from processing viewpoint. As educational industry has revolutionized along with advancement in technology, so is the growth of big data, which is doubling every year. An authenticated survey report given by McKinsey [20] highlights that educational sector is the next evolving sector (after communication and government) that generates higher quantity of Big data. Google, Microsoft, Amazon etc are the prominent existing cloud service provider that caters up the storage requirement of Big Data being generated by educational sector.

- **Social Network:**

Online social networks are a part of daily life for billions of people worldwide endowing a sense of 'always being connected' to each other. This has generated a deluge of social and behavioral data that constitutes a significant part of the Big Data. Systematic investigations are warranted to achieve a better understanding of social interactions, complex behaviors, contextual patterns, and associated outcomes of interest. Through this paper, we intend to create a platform bringing together researchers and practitioners from different disciplines such as, computer science, electronics, sociology, economics, psychology to share, exchange, learn, and develop new concepts, ideas, principles, and methodologies

IV. PROBLEM IDENTIFICATION

Social media has gained immense popularity with marketing teams and Twitter is an effective tool for a company to get people excited about its products. Twitter makes it easy to engage users and communicate directly with them, and in turn, users can provide word-of-mouth marketing for companies by discussing the products. The main obstacle regarding social network analysis of telecommunication networks is the vastness of the dataset. Analysis of a network consisting of millions of connected objects is usually very computationally costly and may take quite some time to perform. As an answer to the problems of very large data sets, as well as related problems, the development is directed towards parallel processes. Examples of this can be found in commodity computing and cloud-services. The idea is to divide a problem into parts and let several machines processes it at the same time, separately. This solution is generally preferred before one single strong computer, perhaps with multiple processors, as it is economically cheaper. Google is, as of now, the pioneer of this approach with their software solutions MapReduce and Pregel. However, open-source alternatives, most prominently Hadoop, are making its way into the industry. A somewhat different approach is that of using graph databases. Instead of the strict order found in relational databases, such as SQL, a graph database does not require any given pattern in which entries are defined. An example is Neo4j . A Neo4j database consists of a set of nodes and relations, which both has properties related to them. Nodes and relations are regarded as equally important, and this makes traversals of the graph faster than for RDBMSs. The task at hand regards analysis of large social networks that is a representation of a telecommunication network. The approaches mentioned above are tested in order to determine their ability to solve this problem. The analysis involves measuring the importance of each subscriber within the network. For this, metrics thought to be related to the ability of a subscriber to spread information within a network are calculated. To do this satisfactory the whole graph must be available. This can cause problems for large graph if the storage size is greater than the size of the hard drive. Additionally, solving the problem in parallel using MapReduce or Hadoop implies storing different parts of the data on separate hard drives. Each process then accesses only the locally stored part of this data, thus only a confined portion of the graph representing the network. Clever algorithms, taking into account the issues of parallelization, are a solution to this. In graph databases, such as Neo4j, without parallelization the whole graph will be available at all times, but the issue of graphs larger than disk space still exists. Using the technology of the Hadoop framework and making use of knowledge form the field of machine learning, a prototype for detecting subscribers of particular interest is developed.

V. PROPOSED SYSTEM

The prime aim of the proposed system is create a framework that can perform evaluation of conventional datamining algorithm for performing knowledge discovery of Big data in educational sector and social networking. The schematic architecture of the proposed study is as shown in Fig.1. The architecture shows a scenario where we possible exhibit the educational data generated from the students using social networking applications giving birth to larger size of files. Social networking applications are used by various students from various domain expertise and hence different types of data are captured. Obviously, such data are highly unstructured in size which is almost impossible to perform any sorts of analysis on it. In our previous study, we have already stated that performing conventional datamining techniques over big data is highly computational challenging task. Hence in this paper, we try to built a computational model using conventional datamining algorithm.

The above architecture shows a scenario where we possible exhibit the educational data generated from the students using social networking applications giving birth to larger size of files. Social networking applications are used by various students from various domain expertises and hence different types of data are captured. Obviously, such data are highly unstructured in size which is almost impossible to perform any sorts of analysis on it. In our previous study, we have already stated that performing conventional datamining techniques over big data is highly computational challenging task. Hence, in this paper, we try to built a computational model using conventional datamining algorithm Document clustering is an enabling technique for many other machine learning applications, such as information classification, filtering, routing, topic tracking, and new event detection. Today, dynamic data stream clustering poses significant challenges to traditional methods. Typically, clustering algorithms use the Vector Space Model (VSM) to encode documents. The VSM relates terms to documents, and since different terms have different importance in a given document, a term weight is associated with every term. These term weights are often derived from the frequency of a term within a document or set of documents. Much term weighting schemes have been proposed. Most of these existing methods work under the assumption that the whole data set is available and static. For instance, in order to use the popular Term Frequency – Inverse Document Frequency (TF-IDF) approach and its variants, one needs to know the number of documents in which a term occurred at least once (document frequency). This requires a priori knowledge of the data, and that the data set does not change during the calculation of term weights. The need for knowledge of the entire data set significantly limits the use of these schemes in applications where continuous data streams must be analyzed in real-time. For each new document, this limitation leads to the update of the document frequency of many terms and therefore, all previously generated term weights needs recalibration. For N documents in a data stream the computational complexity is O(N 2), assuming that the term space M per document is much less than the number of documents. Otherwise, the computational complexity is O(N2 MlogM), where O(MlogM) computations are needed to update a document. The proposed system considers that different online users gives feeds related to educational topic from multiple social networking sites. In order to consider the challenges, the study considers all the social networking sites which are on cloud. As student’s feedback will differ highly from one to another, so proposed system is considered to have high number of missing data, noisy data, or unambiguous data, which are preprocessed by cleaning operation in conventional datamining technique.

The unstructured data being collected is subjected to open source APIs for extracting the knowledge from unstructured data. The anticipated issues in the proposed system are highly likely to occur as the data is massive and highly unstructured. Moreover, the study eases the computation by not considering other file format and only considered text file with unstructured data. The framework captures the data from one row and check for noisy data ending up performing data cleaning process. The open source API is designed using java that performs extraction of the term frequency as well as inverse document frequency along with computation of simulation time. Also, it should be noted that the data are highly distributed type, where the system is developed focusing on faster processing of the datamining algorithms. The outcome of the results highlights that proposed system is found with increasing simulation time with the increase of dataset, and less linearity is found in the simulation time.

VI. CONCLUSION

The proposed system discusses about the framework that evaluates the extent of effectiveness of conventional data mining algorithms on Big Data captured from education data in multiple social networking sites. The outcome of the study shows higher simulation time, more overhead, and inaccuracy in knowledge discovery process. Therefore, we are successfully exhibiting the fact that conventional data mining algorithms cannot be directly applicable to Big Data for performing knowledge discovered process. Our future work will be in the direction of extending the same framework using Hadoop and Map Reduce.

REFERENCES

[1] J.A. Hendricks, Social Media: Usage and Impact, Lexington Books,2012.

[2] J.Xie and B. K. Szymanski. Community detection using a neighborhood strength driven label propagation algorithm. In IEEE Network Science Workshop 2011, pages 188-195, 2011.

[3] [3] J Xie and B. K. Szymanski. Towards linear time overlapping community detection in social networks. In PAKDD, pages 25-36, 2012.

[4] J. Xie, B. K. Szymanski and X. Liu. SLPA: Uncovering Overlapping Communities in Social Networks via A Speaker-listener Interaction Dynamic Process. In Proc. of ICDM 2011 Workshop, 2011.

[5] S. Gregory. Finding overlapping communities in networks by label propagation. New J. Phys., 12:103018, 2010.

[6] Y. Hu, M. Li, P. Zhang, Y. Fan and Z. Di. Community detection by signaling on complex networks. Phys. Rev. E., 78:1, pp. 016115, 2008.

[7] K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks. In WWW Conference, pp. 1275-1276, 2007.

[8] M. E. J. Newman and M. Girvan. Finding and Evaluating Community Structure in Networks. Phys. Rev. E, 69, pp. 026113, 2004.

[9] V. Blondel, J. Guillaume, R. Lambiotte and E. Lefebvre.

Fast Unfolding of Communities in Large Networks. J. Stat. Mech., 2008.

[10] A. Capocci, V.D.P. Servedio, G. Caldarelli and F. Colaiori. Detecting communities in large networks.Physica A, 352, pp. 669-676, 2005.

[11] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. Proc. of SIAM International Conference on Data Mining, pp. 76-84, 2005.

[12] M.E.J. Newman and M. Girvan. Finding and Evaluating Community Structure in Networks. Phys. Rev. E, 69, pp. 026113, 2004.

[13] P.Schuetz and A. Caflisch. Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. Phys. Rev. E, 77, pp. 046112, 2008.

[14] S. Fortunato. Community detection in graphs. Physics Reports, 486:75174, 2010.

[15] J Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In SIGKDD, pages 177-187, 2005.

[16] I. Leung, P. Hui, P. Lio, and J. Crowcroft. Towards real-time community detection in large networks. Phys. Rev. E, 79:066107, 2009.

[17] Z. Lu, Y. Wen, and G. Cao, Community Detection in Weighted Networks: Algorithm and Applications, PerComm, pp.179-184, 2013

[18] J. Tang, X. Wang, H. Liu, Integrating Social media data for community Detection, Springer, 2012.

[19] D. Darmon, E. Omodei, C.O. Flores, Detection Communities Using Information Flow in Social Network, Proceedings of the Complex Systems Summer School. Santa Fe Institute, 2013.