International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Mining Social Media Data for Understanding Drugs Usage

Shadma Qureshi[1],Sonal Rai[2],Shiv Kumar[3]

[1] M.Tech Scholar, Department of Computer Science & Engineering, Lakshmi Narain College of Technology & Excellence Bhopal (M.P), India,
[2] Assistant Professor, Department of CSE, Lakshmi Narain College of Technology & Excellence, Bhopal (M.P), India
[3] Professor & Head, Department of CSE, Lakshmi Narain College of Technology & Excellence, Bhopal (M.P), india

---

Abstract: *This review/survey paper based on the research carried out in the area of data mining depends for managing bulk amount of data with mining in social media on using composite applications for performing more sophisticated analysis using cloud platform. Enhancement of social media may address this need. The objective of this paper is to introduce such type of tool which used in social network to characterised drug abuse. This paper outlined a structured approach to analyse social media in order to capture emerging trends in drug abuse by applying powerful methods like cloud computing and Map Reduce model. This paper describes how to fetch important data for analysis from social network as Twitter, Facebook, and Instagram. Then big data techniques to extract useful content for analysis are discussed.*

Keywords: *social media; data mining; Big data; illicit drug use; Hadoop; HDFS; clavier.*

1. INTRODUCTION

Social network (media) is one to extracting the information from the internet. Nowadays it is used for extracting the data of patient's to know the understanding of patient symptoms. Social media, classify from individual messaging to live for as, is providing immeasurable opportunities for patient to converse their experiences with drug and devices. Social media allows message contribution, gathering information and distribution in the health care space. Health care is one which contains the information of patients with their permission. It provides an effective social networking environment. The proper way of mining information and drift from the knowledge is cloud. Using network based analysis method it model the social media such as Facebook, Twitter, WebMD [1].Nowadays the scientific experiment often requires bulk amount of computation during simulation and data processing. Performance of super computer is increasing rapidly.It allows solving scientific problem by automatic computational through collection or array list which is emerged by set of sensors. Nowadays electronic mechanism is growing in recent scenario. [5]

1.1 Data Mining Techniques

The knowledge extracted allows predicting the behaviour and future behaviour. This allows the business owners to take positive knowledge drive decisions. Data mining is enforced in different domain like FMCG, economy, medical, education system etc. Knowledge is derived from the previously fact by applying pattern recognition, statistical, mathematical techniques those results in expertise form of facts, trends, association, patterns, anomalies and exceptions. There are some areas where data mining is applied.

Data Pre-processing: Data pre-processing make ready the natural data for mining process.

Data Mining: Mining is the way of separating some important patterns from a large amount of data.

Pattern Evaluation: This process evaluates the pattern that is generated by the data mining. The patterns are evaluated according to engagingness measure given by user or system.

***Knowledge presentation*:** Knowledge presentation uses visualization techniques that visualize the interesting patterns and help the user to notice and interpret the resultant patterns. [ ]

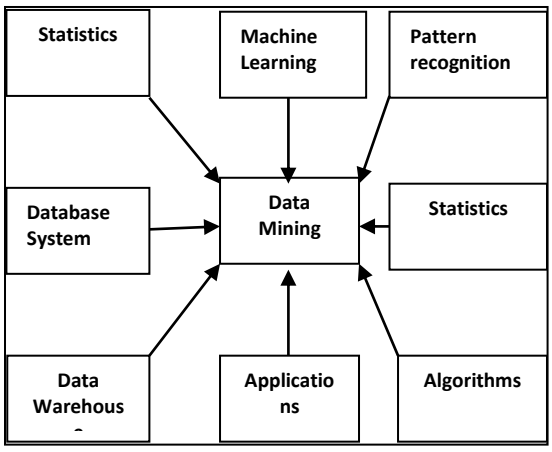

Figure 1: utilization of Data mining Industries [ ]

1.1.1 Audit of data prospecting

Data prospecting is great deal of attention in the informative media as a whole in recent years, due to availability of bulk data and imminent need for turning such data into useful information and knowledge. Data prospecting is the way of looking meaningful fact and patterns. The clue gained can be utilizing during applications analysis such as fraud detection, and customer retention, to production flow and science exploration. Data prospecting can be viewed as a result of the natural expansion of information technology. Data prospecting is iterative process. The basic steps are:

1) ***Data cleaning*** (It is a procedure of removing duplicate data)

2) Data integration (In this step data from multiple sources are combined)

3) ***Data selection*** *(*In this step data relevant for mining task is selected)

4) ***Data transformation*** *(*In this step data will be transformed into form that is appropriate for mining)

5) ***Data mining*** *(*In this step some resourceful ways are applied for extracting data patterns)

6) ***Pattern evaluation*** *(*In this step really meaningful patterns representing knowledge based on some criteria.)

7) ***Knowledge presentation*** (In this step perception and proficiency representation techniques are used to present the mined knowledge to the user)

A social network is a platform where a number of people gather at one place. In social chain the main emphasis is on the relationships among people and organization.This paper present survey of work done in the field of social media on the basis of some parameters. [ ]

1.2 Architecture of CLAVIER

Backgrounds and construction of CLAVIRE is depends on basis of IPSE (Intelligent Problem Solving Environment) concept [3] which spread PSE approach [4] with the help of reproduction. In IPSE concept allows hiding the technical details of the used infrastructure. It is desirable to communicate with the model using specific languages which are itself converted into the conglomerate operation performed using the services applicable within e-Science infrastructure. This approach integrates different resources using domain specific description of their usage. [4]

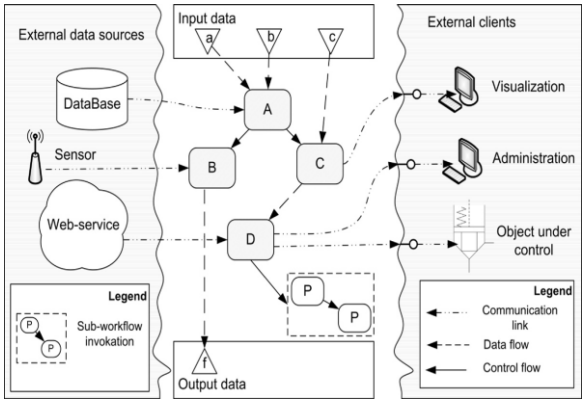

Figure 2: Working infra of CLAVIRE [4]

1.3 Data Management (HDFS)

Management of data from social chain which provides unified approach for solving scientific tasks. Because sociable chain contains abundant of data Author used Big Data paradigms to mine and analyse it. Firstly, data from sociable chain is mined using the crawler or engine [2] which saves it into Hadoop bunch/cluster. Secondly, big volume of mined data is filtered and aggregated to get comparatively tiny datasets of information that is applicable to the solving task. Finally collected data is used as an input for complex applications which perform final and sophisticated data analysis. To manage computational way of the complex application Author used AaaS (Application as a Service) model which is implemented in this environment for distributed computing-based cloud platform.[ ]

2. Literature survey

From Noemie Elhadad, et al[5], sociable chains are a major source for client generated feedback on nearly all products and services. Users frequently believe on social chain to disclose sometimes real life incidents rather than visiting social communication channels. This important, actionable, client created facts, if extracted truly and robustly from the social chain, has the potential to have the positive impact on critical applications related to social health and safety, and beyond. Unfortunately, the production of information from social chain where the output of the extraction process is used to take concrete actions in the actual world are not well supported by existing technology. Traditional information production approaches do not work well over the highly informal and ungrammatical syntax in social chain. They do not manage the production and aggregation of rare content. In our ongoing collective project between Columbia University and the New York City Department of Health and Mental Hygiene (DOHMH), this paper aim to address these difference in research and technology for one important public health.

From Erwan Le Martelot et al [6], today everywhere network is available. The community disclosure received an increasing attention as a way to bring to light the formation of networks and connected internally than superficially. Yet most of the effective methods available do not consider the possible levels of organization, or scales, a network may encompass and are therefore limited. In this paper Author said about compatible with global and local criteria that enables fast multi-scale community finding. The method is to explain with two algorithms, one for each type of criterion, and executed with 6 known criterion. Discovery communities at various level is a computationally luxurious task. Therefore, this job puts a strong attention on the reduction of computational complexity. Several heuristics are commenced for speed up purpose. Experiment exhibit the competency and exact of our way with respect to individual algorithm and criterion by testing them against large out- comes in multi-scale network. This work also offers an assessment between criteria and between the global and local approaches.

From Hari Kumar and Dr. P. Uma Maheshwari [7] Big data is the term that characterized by its increasing volume, velocity, variety and veracity. All these characteristics make processing on this big data a complex task. So, for processing such data Author need to do it differently like Map Reduce Framework. When an organization exchanges data for mining useful information from this Big Data then privacy of the data becomes an important problem in the previous years, several privacy preserving models have been given.Anonymizing the dataset can be done on many operations like generalization, suppression and specialization. These algorithms are all suitable for dataset that does not have the characteristics of the Big Data.To perpetuate the privacy of dataset an algorithm was proposed recently.An author represents how the growth of big Data characteristics, Map Reduce framework for privacy preserving in future of our research.

From E. Srimathi, K. A. Apoorva [8], In recent days many internet services require clients to share their confidential electronic health records for research analysis or data mining, which leads to security issues.The scale of data in cloud infrastructure rises in terms of nature of Big Data, thereby creating it a conflict for traditional software tools to process such bulk data within an endurable lapsed time.As a consequence,it is a conflict for current anonymization techniques to preserve privacy on confidential extensible data sets due to their inadequacy of scalability.An Author represents an extensible two-phase approach to Anonymizing scalable datasets using dynamic Map Reduce framework and LKC privacy model.

3.Problem Identification

Author have learned several things from this study (work). First, define the programming paradigm makes it simple to correlate and distribute computations for fault-tolerant. Second, network channel is a limited resource. A number of surge in this model targeted at reducing the amount of data sent among number of client's node. Third, redundant execution may use to reduce the impact of slow machines, and to handle machine failures and data loss.

4. Methodology

4.1 Block Diagram

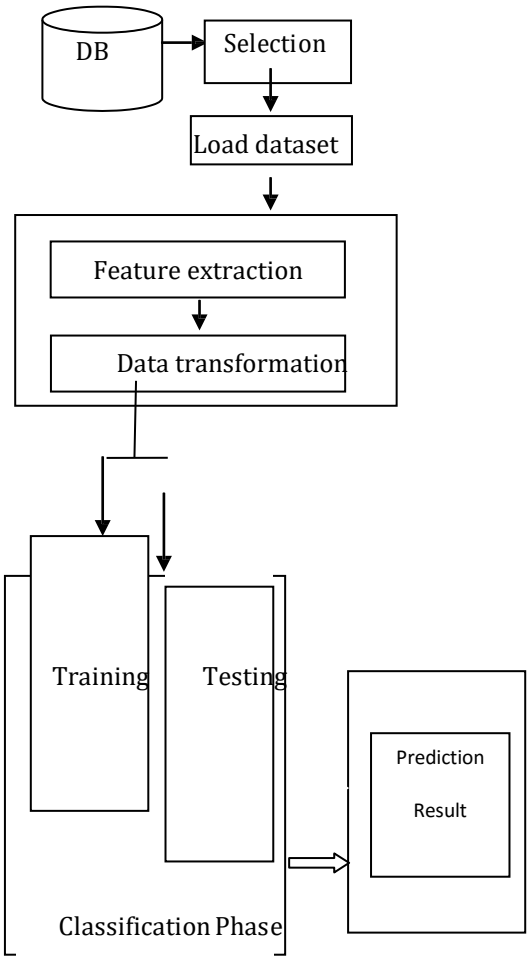

Figure 3: Block Diagram

4.2 Algorithm

Input / Load data set

Apply supervised learning for feature extraction
Received Extracted data as output
Generate Training and Testing data set (By applying techniques)
Apply Machine learning algorithm to training dataset (MLR)
Build the Classifier / model using the "training"
Dataset
Apply Classifier on testing data set

Perform / Obtain Prediction (classification) of the testing set.

Utilize the "test" set predictions to calculate all the performance metrics (Measure Accuracy and other parameters)

Note: All the steps done on data mining tools those are available on Azure.

5. CONCLUSION

This paper presented our approach for mining and managing data from social chain which depends upon combination of bulk amount of data from social networks which is based on combination of big data and infrastructure paradigms. Map Reduce model is used to mine, store and process bulk data through social network. Processing of mined data is also performed by Hadoop which simplifies development of new algorithms and provides high scalability and flexibility. The Map Reduce programming path has been successfully used by Google for many different purpose. Author attributes this success for many reasons. First, the model is accessible to use, even for programmer without any experience with parallel processing and distributed system, because it shields the details of parallelization, fault tolerance, and load balancing. Second, a large variety of problem is easily expressible as Map Reduce computation. For example, Map Reduce is used for the generalization of data for Google's production web search service for sorting, for data mining, for machine learning and many other systems. This paper presents development of an implementation of Map Reduce that extend to large chunks(storage) of machines comprising thousands of machines. The utilization makes efficient use of these machine resources is suitable for many large computational issue encountered at Google.

References

[1] The Fourth Paradigm, in: T. Hey, S. Tansley, K. Tolle (Eds.), "Data-Intensive Scientific Discovery," Microsoft, 2009.

[2] S.C. Glotzer, et, al, WTEC Panel Reprt "On international assessment of research and development in simulation based engineering and science", World Technology Evaluation Centre, Inc,2009.

[3] A.V. Boukhanovsky, S.V. Kovalchuk, S.V. Maryin, "Intelligent software platform for complex system computer simulation: conception, architecture, and implementation", Izvestiya, VuZov, Priborostroenie10 (2009) , 5-24, in Russian.

[4] J.R. Rice, R.F. Boisvert, "From scientific software to problem-solving environments", IEEE, Computational Science and Engineering 3 (3) (1996) 44-53.

[5] Noemie Elhadad, et al "Information extraction from social media for public health".

[6] Erwan Le Martelot, "Fast multi scale detection of relevant communities".

[7] Hari Kumar.R M.E (CSE), Dr. P. Uma Maheshwari , Ph.d, "Literature survey on big data in cloud," International Journal of Technical Research and Applications e-ISSN: 2320-8163.

[8] E.Srimathi, K.A. Apoorva "Preserving identity privacy of healthcare records in big data publishing using dynamic MR", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5, Issue 4, 2015.

[9] V. Borkar. M.J. Carey, C. Li., "Inside big data management: ogres, onions, or parfaits?", Prof. 15th Int. Conf. Extending Database Technol., 3-14, 2012.

[10] X. Sun , B. Gao, Y. Zhang, W. An, H. Cao, C. Guo, et al, "Towards delivering analytical solution in cloud: business models and technical challenges", E-Bus. Eng. (ICEBE), 2011 IEEE 8th Int. Conf.

[11] D.J. Abadi, "Data management in cloud: limitation and opportunities", IEEE Data Eng. Bull, 32: -12, 2009.

[12] Matthew Kerland et al, "A review of data mining using big data in health informatics".

[13] Matthew Herland et al, "A review of data mining using big data in health informatics".