ISSN:0975-9646

Analysis of Unstructured Data

Rohith G Murali, Rekha Sunny T
Department of MCA, SCMS School of Technology and Management,
Mahatma Gandhi University Cochin, India

Abstract—The wide popularity and usage of Internet resulted in the proliferation of textual data particularly in an unstructured form. More than 80% of data generated through blogs, tweets, customer reviews, emails etc. are unstructured. The rapid increase of unstructured data in turn made the process of analyzing and making better business decisions more challenging. This paper describes about unstructured data, its role in leveraging information and the various techniques to analyze unstructured data.

Keywords—Big Data; Unstructured Data; Analytics

I. INTRODUCTION

The concept of big data has grabbed the attention of everyone since the time it was introduced and continues to be of high value in various aspects. According to TechAmerica Foundation the definition of big data is as "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information"[1].

Big Data consists high volume of data which grows exponentially every minute. It may consist of a wide variety of data starting from simple texts to videos. Storing and managing this information get cumbersome if we continue to use the traditional methods of data storage such as Relational DataBase since it only covers the storage of structured data. The big data is a collection of heterogeneous mixture of data containing structured, semi-structured and unstructured data. Over 90% percentage of data was generated over last 3 years. A leading industrialist Merrill Lynch has published, out of which "80% of business-related information originates in unstructured form basically text" [2]. Recent study reveals that 85% of the Fortune 500 organization may fall behind the other 15% of the organization because they failed on leveraging the information by exploiting the unstructured data [3]. Over 95% of big data is in the form of unstructured data, exploitation of this resource may provide various insights to the organization responsible and may help in the process of decision making very effective[1].

Data analytics helps us to find the hidden information in unstructured data through the computational process of Knowledge Discovery on Database (KDD). But firstly, the unstructured data has to be refined from its crude form to more structural one which is done through various analytical procedures. Various challenges are faced while combining the structured and unstructured data. The resultant data will be helpful in the predictive analytics [4].

II. LITERATURE SURVEY

A. Data Analysis

Data analysis involves data inspection along with cleaning and transformation techniques are applied. Later the data is modelled in such a way that the goal of analysis which is to recover some valid knowledge and help in decision making. Knowledge discovery methods such as data mining, business intelligence which deals with business data in decision making, Explorative Data Analysis which is used to find new features of data, Confirmatory Data Analysis which is used to check the validity of hypothesis, Predictive analytics used to forecast some event or scenarios are all part of Data Analysis.

In simpler words, data analysis is about intake of crude and raw data to process and later be modelled into some knowledge which can be used in decision making process of various users. Therefore data analysis can be closely related to data modelling and visualization for easier understanding.

B. Structured Vs Unstructured Data

Structured data refers to the data which has a high degree of organization. Using relational database to store these data can be very easy since it can be modelled into a tabular form. Simple straightforward algorithms are enough to search these data and can be directly applied without any additional requirements.

Unstructured data can be complete opposite from the concept of structured data. Unstructured data's level of organization is very less hence leveraging information from these data is not directly possible through searching algorithms. These noisy unstructured data has to be cleaned and modelled into an acceptable form such that from which the knowledge can be discovered. Therefore unstructured data has to undergo a hefty compilation and time consuming process which can become expensive. Hence all strata of business or any other fields handling immense amount of data find it beneficial for them to find least expensive methods to analyses the unstructured data.

III. ROLE IN LEVERAGING INFORMATION

Unstructured data are data which are not organized and not confined to any framework to which direct queries based on search engine algorithms can be executed to

retrieve the desired data. These data has to be refined and to be integrated with the structured data to retrieve desired information. Unstructured data can be of various type, such as:

A. Text
B. Audio
C. Images
D. Video
E. Social Media content…etc.

The ambiguity in these data makes these data less approachable by the traditional softwares. The major challenge of analyzing unstructured data is the level of noise that is present within the data. Therefore there is very need to cleanse the data and integrate these refined data along with the structured data before its analyzed. Integrating the unstructured data to structured data is another challenging task.

Nearly 95% of data are unstructured data of which major part lies as videos and other data retrieved from various sensors that has setup all over the world. Therefore there is a need for cost effective techniques and tools for exploiting these huge volume of high velocity growing data.

Most of the organizations even do not know how they could take advantage of these unstructured data. It is estimated that nearly 85% of the Fortune 500 companies do not exploit the unstructured data while lagging behind the other 15% of companies in competitive advantage [3]. Most of the critical information are stacked up as files in the organizations, organizations who fails to exploit this large repository of knowledge will be left with no competitive advantage over other companies and may even fails to survive in near future.

While exploiting unstructured data only we get a clear idea of what the factor that states the successfulness of the various organizations. For example, nowadays the hospital is considered to be running well if they have large financial support and highest facilities that can be offered, but an institution such as a hospital must be evaluated not on the basis of its financial status or the facilities provided by them even though they lead to a successful institution. It is the number of medical cases that were successfully solved by the institution that has to be taken while judging its proper running. These data may only be present in the papers and are not analyzed. Hence the need for data analytics on unstructured data.

IV. VARIOUS TECHNIQUES TO ANALYZE UNSTRUCURED DATA.

There are various techniques to analyses the unstructured data according to the type content that the data has, they are

A. Text Analytics

In text analytics, from various sources of text data such as feeds, documents, emails, advertisements, blogs, news content, logs, website contents, social media content…the information or insights are extracted or retrieved involving techniques such as statistical analysis, Computational linguistics and machine learning. These meaningful insights benefits the organizations in better decision making processes.

Information Extraction or IE techniques are used to retrieve structured data from the unstructured data. It particularly involves two subtasks which are 'Entity Recognition' and 'Relation Extraction' [5]. In the process of entity recognition various entities such as person, organizations and things are found out and are classified into separate classes. In the relation extraction phase of information extraction various semantic relations that lies between the classified entities are found out. For example, from a medical prescription various entities such as the person, the hospital or organization, drug dosage information can be found out and classified in entity recognition phase. While the relation extraction leaves you with a clear idea of the various relations between the entities that are found out.

Another method of text analytics is the 'Summarization'. Summarization involves the techniques that creates a summary from the text content of the information provided. The summarization can be of two types which are 'Extractive' or 'Abstractive' [6]. The extractive summarization involves creating a summary of information by adopting the actual original units form the original content. While the abstractive summarization involves the technique of analyzing the contents of information and extracting the summary by learning the semantic relations in data. Hence the abstractive method demands the use techniques such as Natural Language Processing or NLP to parse the text and produce summary. Abstractive summary may need not include the original units from the contents. However for big data extractive summarization is better since it is easier to adopt.

Another way of text analytics involves the Question and Answer method. It is mostly adopted in academic and healthcare sectors. There are basically three types of QA methods, they are Information Retrieval (IR), Knowledge-based approach and Hybrid approach. In information retrieval method the question and answer way of analytics are done in three phases. First phase consists of question processing from which a query is made. In second phase the pre-written documents are analyzed and the relevant information is extracted. The third phase involves the answer processing where the answer is matched to the previously extracted information and are ranked. The answer which ranks top are given out as solution. Knowledge-based approach is adapted in fields where absence of large volume of pre-written contents are present. It can be effectively adopted in the restricted domains such as medicine or tourism. This approach make use of semantic information of the question and then it is used for querying. In Hybrid approach, both IR and Knowledge-based approaches are adapted. The querying is

based on knowledge based approach and the answers are taken by information retrieval approach.

B. Sentiment Analytics

Sentiment analytics or Opinion mining or Emotion AI consists of natural language processing, computational processing, text mining and biometrics to identify, retrieve , visualize and to learn various affective states and information on the emphasized subject. Sentiment analysis can be applied to reviews as well as the voice of the customers, responses to surveys, social medias, costumer services and even targeted marketing.

In other words, sentiment analysis involves identifying the attitude of a writer, speaker or any other actors on any topic or even the overall polarity of the context or sentiment in the context. The view of the actor can be evaluation, judgement or intended emotional communication or even the affected state of actor. Sentiment analytics can be categorized into three types they are document level, sentence level and feature/aspect based where these level indicates whether the opinion is expressed within a sentence, document or based on aspects such as 'negative', 'positive' or 'neutral' opinion about an entity. Apart from these polarity based opinion mining, going beyond polarity can lead to mining of emotional states such as 'sad', 'angry' or 'happy'. Using sentiment analysis at the document level to identify the polarity of various media such as reviews can provide us with the predictive quality of the entity. However, use of aspects based sentiment analysis can provide us with more detailed study as well as with more insights to the entity since it covers the entity from various aspects than from just analysing the polarity of the document.

In sentiment analysis normally the neutral words are ignored. But in a different method of sentiment analysis, there uses a scaling system where the words denoting positive, neutral and negative words are given within a scale from +10 to -10. This eliminates the possibility of ignoring the neutral words nearly to an extent. Later the whole text is rated using the score obtained during the analysis. Since such a sophisticated and accurate method is used to evaluate the document in its textual level than in document level, this will ensure the correctness of the system to an extent.

The sentence level analysis involves classification of the text to classes containing subjective as well as objective. This can be difficult when compared to document level analysis. The subjective sentences may sometimes depend on the context and objective texts may contain subjective texts within them such as quotes containing opinions.

The feature/aspect based analysis involves identifying and extracting opinions about various features of an entity. The various features of an entity can be found out by topic modelling or deep learning. Sentimental expressions on different aspects of an entity helps getting the insight of in what it excels and what feature it has to develop more.

C. Audio Analytics

Audio analytics deals with analyzing, extracting or retrieving data from unstructured audio content. When applied to human speech, the audio analytics can be called as speech analytics [8, 9]. This kind of information is very useful in fields which consists of lot of information in the form of audio. This method can be successfully implemented in fields such as call centers and healthcare sectors. Audio analytics can even be applied in analyzing the emotional conditions of a person [8].

There are basically two approaches in audio analytics which are transcript based and phonetic based. Transcript based analytics can also be called as Large Vocabulary Continuous Speech Recognition (LVCSR) which is further divided into two sub process: Indexing and Searching. In indexing phase, Audio Speech Recognition (ASR) algorithms are used to match sounds with words in pre-defined words in dictionary. If failed to do so, the most similar word is returned. The output file contains sequences of words which were identified during the analysis. In second stage, simple text based operations are applied to search terms. In phonetic-based analytics, the analysis is done based on phonemes instead of words. It also involves two phases which are phonemic indexing and phonemic searching.

D. Video Analytics

Process of monitoring, analyzing and gaining meaningful insights from video streams also called as Video Content Analysis [10]. This type of analysis can be used in both real-time as well as pre-recorded videos. These techniques are still at its initial stages. The main source of information is video streaming sites such as YouTube and are Closed Circuit TVs (CCTVs). There can be two approach while parsing the information which are Server-based and Edge based [11]. In Server based approach the information captured are sent to a central server at which further analysis of the data occur. The server based approach will be more efficient if the network bandwidth is higher and server processing capacity is higher. If the bandwidth of network low, the data which has to be sent will meet a need to be compressed hence making analysis of data less reliable. The edge-based approach involves analysis of data locally hence reducing the use of bandwidth requirement since no chance of data loss. However the edge-based approach may have lesser processing capacity and costs more due to the fact that the processing is done locally.

Implementing video analytics can become very crucial in near future. For example, a store CCTV can be used to monitor the customer and their purchase pattern, items bought together, pattern of searching items…etc.

E. Social Media Analytics

It is the analysis of structured and unstructured content of social media channels such as social networks, social news, blogs, micro blogs, media sharing, wikis, social bookmarking, question-and-answer sites and review sites [12]. Two sources of information that can be found in social Medias are content and relation between various

entities. Therefore social media analytics can be classified into two type Content-based and Structure-based analytics [13]. In content based analytics, various contents published by the users of social Medias are analyzed. These analysis consists of text, audio, video…etc.

Structure-based approach of social media analytics involves identifying relations between various entities within the social media, analyzing these relations and extraction or retrieval of information. A social media can be compared to graphs where each node is an entity and the edge connecting them are the activities or relations between them. This type of graph can be called as an Activity graph and provides a reliable way of analyzing the social media.

There are several methods of social media analytics one of them is Community detection or discovery. Communities can be considered as sub-networks of the system where the entities communicates with each other more frequently than with the other entities. Community detection or discovery is the process of finding implicit community in a network. This method is useful in marketing, if the common interests of the community can be satisfied by the organization [14].

Social Influence Analysis is another method which analyze the social media to find out that the actions done by a participant which is likely to be an entity could affect others. This method is helpful in giving insights on strength of connections and patterns of influences in social media.

Link Prediction another method of analytics in social media which is used to predict the possible linkages of entities in future. The goal of link prediction is to understand and predict the possible interactions, collaboration or influence among participants done in each social network. It is by using the link prediction method popular sites offer suggestions such as “People You May Know” or “Recommended for You” [13].

REFERENCES

[1] Amir Gandomi and Murtaza Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *International Journal of Information Management*, vol. 35, Issue 2, pp 137-144, April 2015

[2] Christopher C. Shilakes and Julie Tylman, “Enterprise Information Portals”, Merrill Lynch, 16,November 1998 (*references*)

[3] Stephen Prentice, “From Data to Decision: Delivering Value from ‘Big Data,’” Gartner Inc., March 28, 2012.

[4] Mona Tanwar, Reena Duggar and Sunil Kumar Khatri “Understanding Unstructured Data: A Wealth of Information in Big Data” Amity Institute of Information Technology, Amity University Uttar Pradesh, Noida, India, pp. 2, 2015. (*references*)

[5] J. Jiang, “Information extraction from text,” in C. C. Aggarwal, & C. Zhai (Eds.), Mining text data, Springer, pp. 11–41, 2012.

[6] U. Hahn and I. Mani, “The challenges of automatic summarization,” Computer, vol. 33(11), pp. 29–36, 2000.

[7] B. Liu, “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, vol. 5(1), pp. 1– 167, 2012.

[8] J. Hirschberg, A. Hjalmarsson and N. Elhadad, “You’re as sick as you sound: Using computational approaches for modeling speaker state tgauge illness and recovery,” A. Neustein (Ed.), Advances in speech recognition, Springer, pp. 305–322, 2010

[9] H. A. Patil, “Cry baby: Using spectrographic analysis to assess neonatal health status from an infant’s cry,” in A. Neustein (Ed.), Advances in speech recognition, Springer, pp. 323–348, 2010.

[10] B. K. Panigrahi, A. Abraham and S. Das, “Computational intelligence in power engineering,” Studies in Computational Intelligence, Springer, vol. 302, 2010.

[11] Agent Comparative Analysis. [Online]. Available: http://www.agentvi.com/images/Video_Analytics_Architectures_Comparative_Analysis.pdf

[12] G. Barbier, and H. Liu, “Data mining in social media,” C. C. Aggarwal (Ed.), Social network data analytics, Springer, pp. 327–352, 2011.

[13] Charu C. Aggarwal, “An Introduction to Social Network Data Analysis,” Social Network Data Analytics, Springer, 2011.

[14] C. C. Aggarwal, “An introduction to social network data analytics,” C. C. Aggarwal (Ed.), Social network data analytics, Springer, pp. 1–15, 2011.