SCRUTINIZE MESS OF ONLINE DATA TO EXTRACT THE FIRM INFORMATION USING LATENT DIRICHLET ALLOCATION

K.VijiyaKumar[1], S.Chandni[2], T.Esther[3], B.Uma Maheswari[4]

[1]Assitant professor, Departement of Information Technology,vijiya.kumar@gmail.com
[2, 3, 4] B.TECH, 4th year student, Departement of Information Technology, Manakula Vinayagar Institute of Technology,Pondicherry,India.

---***---

Abstract - *Big data refers to data that can provide for computational manipulations and analysis to reveal the trends and associations between various variables of organizational operation.Big data is the information that organizations use to determine most aspects of consumer and stakeholder behaviour.Space and Privacy plays a vital role in maintaining big data. Thinking ability of the customers in social media is monitored by the managers using monitoring concept.The maintenance of customer reviews is a biggest challenge in big data. This project deals with the maintenance of the spaces and focuses on the customers' desire. Our proposed system aims at providing desire for customers' reviews while benefiting the authorities of social media by using the LDA algorithm for analyzing customers reviews.*

Key Words: Big data, thinking ability, maintenance, challenges, customer, Latent Dirichlet Algorithm, reviews

1.INTRODUCTION

Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks.

Data is being generated about the activities of people and inanimate objects on a massive and increasing scale. We examine how much data is involved, how much might be useful, what tools and techniques are available to analyze it, and whether businesses are actually getting to grips with big data.

Big data is certainly one of the biggest buzz phrases in IT today. Combined with virtualization and cloud computing, big data is a technological capability that will force data centers to significantly transform and evolve within the next five years. Similar to virtualization, big data infrastructure is unique and can create an architectural upheaval in the way systems, storage, and software infrastructure are connected and managed. Unlike previous business analytics solutions, the real-time capability of new big data solutions can provide mission critical business intelligence that can change the shape and speed of enterprise decision making forever. However, given that the data and its structures are fundamentally different, it is increasingly evident that the infrastructure, tools, and architectures to support real-time analysis and insight from this data also must be different. As an IT solution, big data mirrors the growth in both content and data source, as well as the pervasiveness of technology in our everyday lives

Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in 2011, and in every ten minutes in 2013. This rate is still growing enormously. Though all this information produced is meaningful and can be useful when processed, it is being neglected[2].

Big Data is not a technology, but rather a phenomenon resulting from the vast amount of raw information generated across society, and collected by commercial and government organizations. It is the management and exploitation of large or complex data sets. Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

The characteristics such as large volume of data, high speed, variety of data, values are collectively termed as the Big Data. The traditional data processing are inadequate to manage the large and complex data and it also includes the challenges such as analysis, data curation, search, sharing, storage transfer, visualization querying, and information privacy[2]. To overcome these challenges the term big data is introduced. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.".

Big data is a voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity the term is often used when speaking about petabytes and exabytes of data etc. But today Big data can describe with 3Vs: the extreme Volume of data, the wide Variety of types of data and the Velocity at which data is traversing. As Big data takes too much time and costs too much money to load into a traditional relational database for analysis. So, new approaches to storing and analyzing data have been emerged which rely less on data schema and data quality.

Hadoop is a processing engine that is designed to handle extremely high volumes of data in any structure. The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between. HDFS is a reliable distributed file system that provides high through put access to data. The MapReduce programming paradigm which is meant for managing applications on multiple distributed servers.It is a framework for performing high performance distributed data processing and is based on divide and aggregate paradigm. It introduce about the project concept and given the overview idea about the project. It also consists of objective of the Extraction and Analysis of the Online data from the massive data Centers with the help of need for study. we selected papers for literature survey with help of these paper we came to know what the Big data can do and how it help for Analysis of data to provide services to the customers.we listed the Hardware requirements and Software Requirements of our project and also it include that what type of software used in it. Software like Hadoop is used for creating the application.we explained about the existing system the problem definition of the existing system, then finally disadvantages of existing system[1][8].we discuss about the project domain and the detailed description of existing systems by analysis the literature survey of the existing techniques. We also then presented about the techniques and methods of our proposed methods. we concluded that by the use of Hadoop tool in big data, we get the more accurate result from the data store and revival in data center with the help of evolution process.

2. CHARACTERISTICS OF BIG DATA

"Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" Gartner 2012 as shown in Fig-2.1. Big data can be described by the following characteristics:

2.1 Primary characteristics (3V's)

Volume Big data management software enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The quantity of data that is generated is very important in this context. Which data is generated by machines, networks and human interaction on systems like social media is very massive. The name 'Big Data' itself contains a term which is related to characterized size that describe it[2].

Velocity Today Data is generated too fast and also need to be processed fast. Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers only if you are able to handle the velocity[2][8]. Velocity is applied to data in motion. There are various information streams and the increase in sensor network deployment has led to a constant flow of data at a pace that has made it impossible for traditional systems to handle. Initially analysis of data is done by using a batch process. With the new sources of data such as social and mobile applications, the batch process breaks down. The data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

Variety Variety describes different formats of data. These include a long list of data such as documents, emails, social media text messages, video, still images, audio, graphs, and the output from all types of machine-generated data from sensors, cell phone GPS signals, DNA analysis devices, and more. This type of data is characterized as unstructured or semi-structured[2][7]. This variety of unstructured data creates problems for storage, mining and analyzing it. Unstructured data is growing much more rapidly than structured data. It is estimated that unstructured data doubles every three months and offers the example that there are seven million web pages added each day. The representation of 3V's are shown in Fig2.2.

There are two primary challenges regarding variety of data. First, storing and retrieving these data types quickly and cost efficiently. Second, during analysis, blending or aligning data types from different sources so that all types of data describing a single event can be extracted and analyzed together.

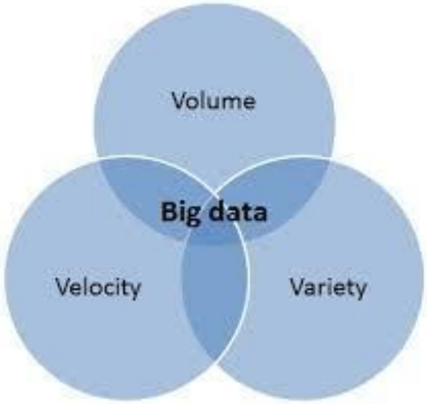

Fig-2.1: Characteristics of Big Data

Fig-2.2: Representation of Big Data

2.2 Additional Characteristics

Veracity Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Veracity in data analysis is the biggest challenge as compared to volume and velocity[1].In scoping our big data strategy ones need to have his own team and partners together to work to clean the data in the system to avoid 'dirty data' from accumulating in his systems.

Validity Like big data veracity, the issue of validity deals with the accuracy of data. So that it can be properly use from decision making and fir future conclusions[1].

Volatility Big data volatility refers to how long is data valid and how long should it be stored. In this world of real time data one need to determine at what point is data no longer relevant to the current analysis.

Big data management clearly deals with issues beyond volume, variety and velocity to other concerns like veracity, validity and volatility to hear about other big data trends and presentation and uses[2][6].

3. LITERATURE SURVEY

3.1 Model Trees With Topic Model Preprocessing

Text mining tools and topic models were used to analyze written text from the WikiLeaks war diary automatically by assigning overarching themes to the single documents[5]. This allowed to get a view on the data which is hard to obtain by manual processing and that may even discover connections between documents which may not be at all obvious. The assignment of topics to the single documents offered the opportunity to use those topics as splitting variables in further data analysis. One has to bear in mind, however, that the assignment of documents to topics is by far not absolute and that it can be difficult to interpret the meaning of latent topics, especially if they are to be named (as is often the case with unsupervised techniques)[5]. At any rate, we saw that split candidate variables generated by pre-processing with LDA proved to be very important in the subsequent analysis, whereas the variables that were already available played a minor role. Hence, discarding the information stored in the report summaries would have led to completely different segmentation, description and interpretation.

3.2 Probabilistic Topic Models : Surveying a suite of algorithms that offer a solution to managing large document archives

As the collective knowledge continues to be digitized and stored—in the form of news, blogs, Web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult[8] to find and discover what we are looking for. We need new computational tools to help organize, search, and understand these vast amounts of information. Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents[8]. This is a powerful way of interacting with our online archive, but something is missing. Imagine searching and exploring documents based on the themes that run through them. We might "zoom in" and "zoom out" to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other[8]. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

3.3 Latent Dirichlet Allocation for Text, Images, and Music

In this report, they showed that LDA is not only useful in the text domain, but also in the image and music domain. In particular, we discuss algorithms that extend LDA to accomplish tasks like document classification for text, object localization for images, and automatic harmonic analysis for music[7]. For each domain, we also emphasize approaches that go beyond LDA's traditional bag-of-words representation to achieve more realistic models that incorporate order information[7].

Latent Dirichlet Allocation (LDA) is an algorithm that specifically aims to find these short descriptions for members in a data collection. Originally proposed in the context of text document modeling, LDA posits that one way of summarizing the content of a document quickly is to look at the set of words it uses. Because words carry very strong semantic information, documents that contain similar content will most likely use a similar set of words[7][10]. As such, mining an entire corpus of text documents can expose sets of words that frequently co-occur within documents. These sets of words may be intuitively interpreted as topics and act as the building blocks of the short descriptions. More formally, LDA is a probabilistic, generative model for discovering latent semantic topics in large collections of text data. Each discovered topic is characterized by its own particular distribution over words. Each document is then characterized as a random mixture of topics indicating the proportion of time the document spends on each topic[7].

This random mixture of topics is essentially our "short description": It not only expresses the semantic content of a document in a concise manner, but also gives us a principled harmonic analysis.

4. RELATED WORK

The emotional attachments of the customer to a brand name are at topic of interest in recent years in the marketing literature. It is defined as the degree of passion that a customer feels for the brand[1]. One of the main reasons for examining emotional brand attachment is that an emotionally attached person is more probable to be loyal and pay for a product or service. In this, the emotional term ranking system based on inference networks, for a specific brand in a given time window, where the aspects of the brand are determined dynamically. As a result, a marketer can have a detailed estimation on the bond between the users and the brand using objective data.

The rapid development of web technologies and social networks have resulted in the creation of a big volume of content. Opinions, emotions, ideas and thoughts can be extracted in order to explain or predict human decision making. Concerning consumers, a decision to buy a product or not can be influenced by the overall emotional attachment to the brand name that is expressed in the network through consumers' posts[9]. From these posts, there must be an adequate extraction of emotional posts for utilizing an observable conclusion for each one of the emotions represented as emotional terms. The purpose of this extraction is the ranking of specific terms and the latter discernment of those that are notable and can provide insights to the marketer experts. This multitude of emotions creates a challenge for both the consumer and the ranking system.

The approach utilized for extracting the aspects of this level, is based on the LDA topic modelling algorithm. This LDA algorithm is used her for the extraction of the data[10]. More precisely, this final level has been employed so as to play the role of the query layer in the traditional inference network model and its presence signifies that we are interested in modeling specific rankings for some preferable, based on users queries, aspects[14][5].

5. EXISTING SYSTEM

Traditional listening techniques (e.g., focus groups, surveys) can be very useful but are typically expensive, are limited in scope, and require great skill to run effectively[14]. Because these exercises are formally scheduled, the voice of the market tends to only emerge in short bursts and infrequently.

Furthermore, many of the key insights that managers wish to uncover are about potential customers[1]. These consumers aren't currently purchasing from your firm but could be enticed to do so. When approaching potential customers, your firm doesn't have any proprietary insight into their needs.

No crunching of internal data will allow you to better understand what they want. In such a world, insight comes from being able to look outside your organization for information.

If the market researcher doesn't ask the right questions, a firm may not uncover what matters to consumers[9]. Finally, there are also significant problems if the consumer finds it hard to fully verbalize the answers when put on the spot. The recent proliferation of user-generated content such as product reviews, tweets, and blogs has provided numerous ways for consumers to share their opinions[13].

In earlier online service the producers just monitors what the users are searching for and update them in their sites

.But most of the time the data's are not so accurate[5]. The producers do not consider the user reviews since it is time consuming for the producer, so the user like product will not be updated.

Integrating data held in your organization is an excellent way of improving knowledge of your customers but gives a limited picture[13]. Most of the information about your customers isn't held anywhere on the company servers; it is housed on various websites that are typically as visible to your competitors as they are to you.

6. PROBLEM DEFINITION

The main problem with the existing work is the users are monitor and it creates an insecure feeling for the Consumers[3]. There are many chances that the customer data cannot be safe. The another problem that the customer encounter is that the companies are not extracting the problems of the user when they post about the successes and failures of products, brands, and firms precisely because they feel that their views should be listened to[11]. Well-managed organizations agree and want to listen to what consumers have to say. Thechallenge in understanding the message from much user-generated content arises from the nature of unstructured data[12].

7. PROPOSED SYSTEM

This project deals with the maintenance of the spaces that create for customers and it focuses on the customers privacy and so that they feel secured whenever they access their desired sites[6]. Our proposed system is to providing beneficiary for both the producer and consumer by providing the abstract of the reviews for time saving.we believe the big data revolution can produce firms that better respond to consumers' wishes[7].

Firms have easy access to data regarding the performance of their products, what consumers really care about, and the strengths and weaknesses of competitors. Consumers are not shy about sharing their thoughts on any number of topics via public forums. This user-generated content contains incredible potential, but many firms don't know how to properly tap it. We suggest that firms consider Latent DirichletAllocation[14], a non-proprietary technique that can be applied by anyone with advanced statistical training. This allows analysts to extract what consumers are thinking about from user-generated content. This technique even allows a manager to understand which attributes consumers see as positives or negatives of his/her product and competitors' products[15] as shown in Fig-7.1. Such analysis can inform the firm's strategy to better serve consumers.

With the right tools, the message can be extracted from the mess of big data.

Steps involved

Step 1 : To fetch the data from the collection of reviews posted in the website that are stored in the database.

Step 2 : The fetched data is to be preprocessed to eliminate all the stop words for example is, that, was, like, help, much, put, etc.,

Step 3 : The algorithm go through each document, and randomly assign each word in the document to one of the K topics.

Step 4 : This assignment already gives us both topic representations of all the documents and word distributions of all the topics.

Step 5 : Go through each word w in d, and for each topic t, compute two things: 1) p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and 2) p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w. Reassign w a new topic, where you choose topic t with probability p(topic t | document d) * p(word w | topic t) (according to our generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current word's topic with this probability).

Step 6 : After repeating the previous step a large number of times, eventually reach a roughly steady state where our assignments are pretty good. So we use these assignments to estimate the topic mixtures of each document (by counting the proportion of words assigned to each topic within that document) and the words associated to each topic (by counting the proportion of words assigned to each topic overall)[10].

Step 7 : The final outcome is made as a line graph, which can be attained by the R Studio IDE.

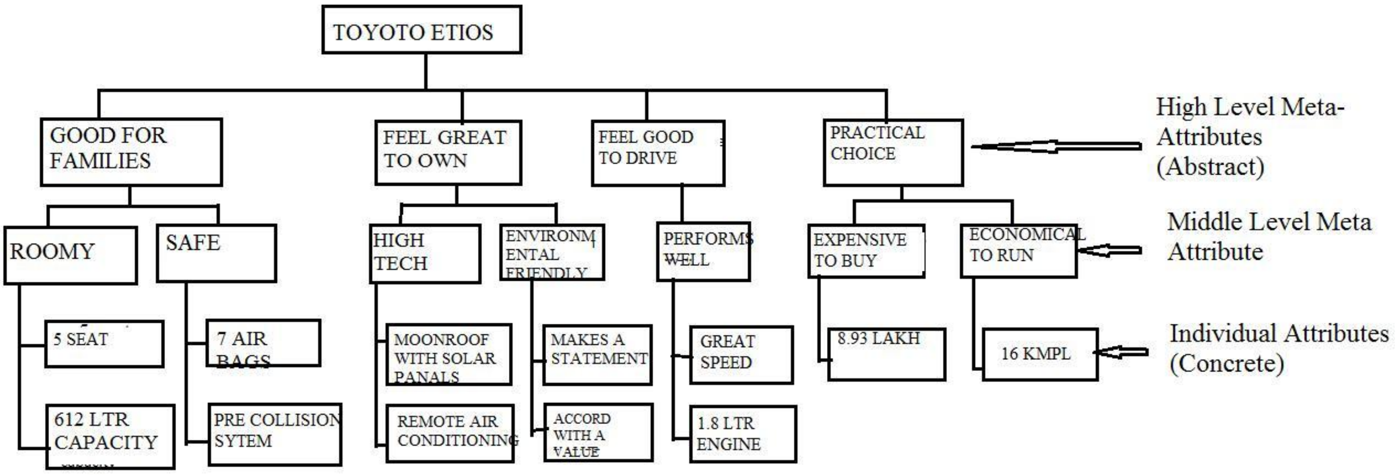

Fig-7.1: Example for hierarchy attributes

8. CONCLUSION

Firms have easy access to data regarding the performance of their products, what consumers really care about, and the strengths and weaknesses of competitors .Now a days the customers feel free to share their reviews in social medias.

We suggest that firms consider Latent Dirichlet Allocation, a non-proprietary technique that can be applied by anyone with advanced statistical training. This allows analysts to extract what consumers are thinking about from user-generated content. This technique makes it easier for the social media authorities to understand which attributes consumers see as positives or negatives of his/her product and competitors' products. Such analysis can inform the firm's strategy to better serve consumers. With the right tools, the message can be extracted from the mess of big data.

9. REFERENCES

[1] Neil T. Bendle, Xin (Shane) Wang, "Uncovering the message from the mess of big data", Business Horizons 2016, 59, 115—124

[2] VibhaBhardwaj, Rahul Johari , "Big Data Analysis: Issues and Challenges" 2015 IEEE.

[3] Andreas Kanavos, EleannaKafeza, Christos Makris "A Brand Love Ranking System for Emotional Terms" 2015 IEEE International Congress on Big Data.

[4] Anne Immonen, PekkaPääkkönen, And EilaOvaska "Evaluating the Quality of Social Media Data in Big Data Architecture" Date of publication October 16, 2015, date of current version November 5, 2015.

[5] Thomas Rusch and Paul Hofmarcher, Reinhold Hatzinger1 and Kurt Hornik "Model Trees with Topic Model Preprocessing: An Approach For Data Journalism Illustrated With The Wikileaks Afghanistan War Logs" The Annals of Applied Statistics 2013, Vol. 7, No. 2, 613–639 DOI: 10.1214/12-AOAS618.

[6] David M. Blei "Surveying a suite of algorithms that offer a solution to managing large document archives PROBABILISTIC TOPIC MODELS" .review articles april 2012.

[7] Diane J. Hu Department of Computer Science University of California, San Diego."LatentDirichlet Allocation for Text, Images, and Music".

[8] LeeAnn Kung, Hsiang-Jui Kung, Allison Jones-Farmer and YiChuan Wang, Combine Big data and other capabilities for performance "Managing Big Data for Firm Performance: a Configurational Approach" Twenty-first Americas Conference on Information Systems, Puerto Rico, 2015.

[9] Okal Christopher Otieno "Managing & Analyzing Large Volumes of Dynamic & Diverse Data" Department of Information Technology (IJCSIS) International Journal of Computer Science and Information Security, August 2015, Vol. 13 No. 8.

[10] David M. Blei, Andrew Y. Ng, Michael I. Jordan "Latent Dirichlet Allocation" Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03.

[11] Martin Ponweiser, Bettina Gr¨un and Kurt Hornik "Finding Scientific Topics Revisited".

[12] IBM "Technology and Trends for Smarter Business Analytics" Technology and Trends for Smarter Business Analytics Don Campbell Chief Technology Officer, Business Analytics.

[13] Simonson, I., & Rosen, E. (2014), "Absolute value: What really influences customers in the age of (nearly) perfect information" New York: Harper Business.

[14] Wang, X., Bendle, N. T., Mai, F., &Cotte, J. (2015). The journal of consumer research at forty: "A historical analysis". Journal of Consumer Research.

[15] Tirunillai, S., and Tellis, G. (2014). "Mining marketing meaning from chatter : Strategic brand analysis of big data using Latent Dirichlet Allocation". Journal of Marketing Research, 51(4), 464—479.