UNSTRUCTURED DATA MINING AND ITS APPLICATIONS

[1]Jagruti Jangal Wagh, [2]Jidnyasa Dharmik Gondane, [3]Ashvini Tulshiram Dukare.
[1,2]Student, Department of Computer Engineering, Cummins College of Engineering for Women
Pune, Savitribai Phule Pune University
Email: [1]wagh.jagruti89@gmail.com, [2]jidnyasagondane@gmail.com, [3]ashvinidukare31@gmail.com

Abstract

Information is generated in various forms. About 90% of the data in the world has been generated over the last two years. If this data is left unmanaged, then it becomes overwhelming, making it difficult to get information from it whenever it is needed. Structured data is information, usually text files, displayed in titled columns and rows which can be easily processed and ordered by data mining tools.

Information system may differ in its application and form but they serve a common purpose that is to convert data into some useful form through which knowledge can be obtained. Data is comprised of the basic, unfiltered and generally unrefined information. Information is much more refined data that is being useful for some form of analysis. Knowledge resides in the user and comes up only when human insights and experience is applied to information and data. But most of the today's data generated is unstructured. Unstructured data is that which has no identifiable internal structure. A leading industry analyst on the confluence of unstructured data and structured data sources, published an article that stated, "80% of business-related information originates in unstructured form basically text". So there is a great need to convert this unstructured data into some usable form.

Data mining is the analysis step of "Knowledge Discovery in databases" or KDD process which is the computational process of discovering patterns in large data that is involving methods at the intersection of artificial intelligence, statistics, machine learning and databases. The basic goal of a data mining process is the extraction of information from a large dataset and convert or transform it into understandable form for future use. Thus, this paper highlights on unstructured data-mining and its applications.

Index Terms: Data Mining, Information, Structured data, unstructured data.

I. INTRODUCTION

Everyday data is generated, collected in huge amount but many-a-times it remains unutilized without drawing useful information and meaningful insights. These insights are vital in strategic and operational decision making process like-marketing, customer engagement, branding, etc. Data generated by various channels like marketing, distribution, customer engagement, social channels and web contents is in different forms structured as well as unstructured and available in multiple systems. This unstructured data needs to be converted into something a bit more useful form, it requires finding approaches to convert the free form, unstructured data to some form of structured or semi-structured data and analyze it to get meaningful insights to address business problems and helps business decision making.

Based upon the type of input data, reports can be generated in the form of charts, bar-graphs etc. Business Intelligence dashboards can also be experimented to achieve effective visualization mechanism.

Structured data is the data which is in the organized form that is in the form of rows and columns and can be easily used by a computer program. Relationships exist between entities of data such as classes and their objects.

Unstructured data is the one that does not conform to a data model or is not in the form that can be used easily by a computer. Various examples include memos, chat-rooms, videos, images and researches, body of an email etc. Gartner estimates that almost 80% of the data that is generated in any enterprise today is Unstructured data. Roughly, around 10% of data is in the structured and semi-structured category [1].

II. DATA MINING

Data mining can be viewed as the result of the natural evolution of information technology. The databases and data management industry evolved in the development of several critical functionalities like the data collection and database creation, data management including the storage and retrieval of data. Data mining is the analysis step of the KDD or the ″Knowledge Discovery in Databases” process, an interdisciplinary subfield of computer science, is the process of discovering patterns in large datasets.

Following are the steps involved in the knowledge discovery process –

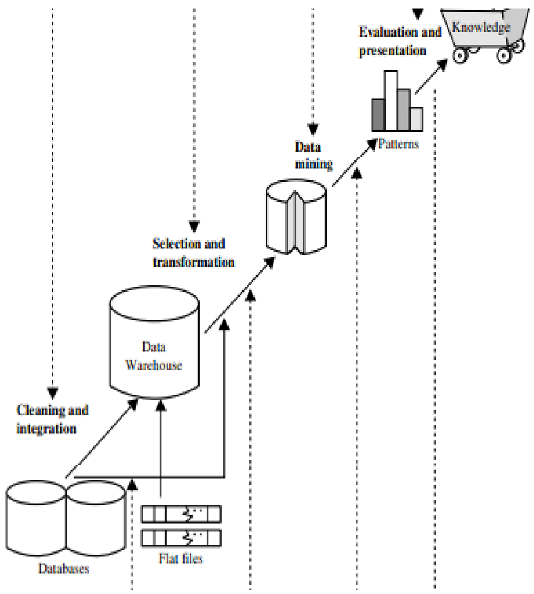

Fig. 1: Data mining which is a step in the process of knowledge discovery-[Han & Kamber & Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann]

- Data Cleaning − In this step the inconsistent data and noise is removed.
- Data Integration − In Data Integration multiple data sources are combined.
- Data Selection – Here, the relevant data to the analysis task are being retrieved from the database.
- Data Transformation − In transformation of data the data is transformed or consolidated into forms which are appropriate or valid for mining by performing various aggregation operations.
- Data Mining − In this step various intelligent methods are applied so as to extract the various data patterns.
- Pattern Evaluation − In this step the data patterns are being evaluated.
- Knowledge Presentation − In this step, knowledge is represented[1].

The goal of the data mining process is to extract the information from a data-set and convert or transform it into an understandable form for future use.

III. STRUCTURED VS UNSTRUCTURED DATA MINING

Structured data is data that can be organized easily, regardless of its simplicity many experts in data industry today estimate that structured data only for 20% of the available data. It is usually stored in databases and is analytical and clean. Some examples of structured data are:

•Machine Generated:

Sensory Data - manufacturing sensors, medical devices and GPS data

Point-of-Sale Data - Credit card information, product information, etc

Call Detail Records - Caller and recipient information, time for call.

Web Server Logs - Page requests, server activities

•Human Generated:

Input Data – The data that is given as a input to a computer: age, gender, code, etc.

Structured data has always had and will always be playing a crucial role in data analytics. It functions like backbone to the critical business insights. Without structured data, it would have been difficult to know where to find insights hiding in the unstructured data sets.

Unstructured data is the data that is not organized in a predefined manner and does not have a predefined model of data. It is obvious that social media plays a very important role in unstructured data. According to a research, 73% of online adults use social networking sites. One of the many ways in which businesses are utilizing this data is to gather and analyse the brand sentiments. In addition to social media there are many other forms of unstructured data which are common:

Word Doc's, Text Files and PDF's - Books, letters, audio and video files, other written documents
Audio Files -Customer service recordings, voicemails, phone calls
Presentations-SlideShares,PowerPoint's
Videos-YouTube uploads, personal videos
Images-Pictures, memos
Messaging- Text or instant messages

In all these examples the data can provide very compelling and useful insights. Using the right tools this unstructured data can add a depth to data analysis that could not be achieved otherwise very easily. The unstructured data enhances the ability of any business to derive greater insights from the datasets.

Unstructured data is the most important piece to the data pie of any business. Tools which are accessible widely can help businesses use this data to the greatest of its potential.

Text mining is the process of deriving the information that is of high quality from text. High-quality information is typically derived by the devising of trends and patterns through means such as statistical pattern learning.

Actually, text mining refers to the using data mining techniques for finding out or discovering useful patterns from texts. The primary difference is that unlike data mining, in text mining the data that is being used is unstructured.

IV. CHALLENGES IN UNSTRUCTURED DATA MINING

The challenges of unstructured data run the gamut from gathering to storing, to using it to make decisions:

- Usability: For unstructured data to be usable, businesses will have to come up with a way to locate, extract, organize, and store the data.

- Volume: Data Size is growing exponentially which is creating new challenges to deal in scale. The volume of unstructured data is growing at a rate of approximately 60% per year. For many businesses, that's more than they can keep up with. That presents challenges for both using and securing the data.

- Relevance: One way in which relevance comes into play is lack of insight into the previous story of certain pieces of data.

- Heterogeneity: The difficulties of big data analysis are caused because of its large scale and the presence of mixed data based on different rules or patterns in the collected and stored data. In the case of complicated heterogeneous mixture data, this data has several patterns and rules and the properties of their patterns may vary greatly.

- Incompleteness: Incomplete data creates uncertainties during its analysis and it must be managed during its analysis. Incomplete data refers to the missing of data field values for some of its samples.

- Quality: By nature, a large volume of unstructured data is being unverified. This presents serious challenges for consumers and enterprises. On a consumer level, people could be negatively impacted by companies that make decisions based on unstructured data. On an enterprise level, making business decisions based on inaccurate data could be extremely costly.

- Requirements for real-time data analysis has been predominant like for weather predictions, exstock tradings, time series, etc.

V. Applications of Unstructured data mining

- Diseases Cure and Bio-informatics: One of the most important application trends dealing with mining and interpretation of biological structures and sequences. Data mining tools are rapidly being used in finding genes regarding cure of diseases like Cancer and AIDS .
- Business Trends: Today's business environment is more dynamic, so businesses must be able to react quickly, must be profitable, and also offer high quality services than ever before. Here, data mining serves as a technology that is fundamental in enabling customer's transactions more accurately, faster and meaningfully. The Business analyst will get meaningful insights to address business problems and helps business decision making.
- Finance **:** Most financial institutions and banks offer a very large variety of banking services like saving checking and individual and business customer transactions, credit like mortgage, business and automobile loans and investment services like the mutual funds. Some also offer stock services. Financial data that is being collected in the banking and financial industry is often relatively reliable, high quality and complete which facilitates systematic data analysis. For example it can also help in detection of fraud by detecting a group of individuals who stage their accidents to collect on the insurance money.
- Area of forensics, security and intelligence: The manual monitoring of masses usually of unstructured data is not feasible. The ability of identifying the person's names, credit card, addresses and bank account numbers and other entities automatically is the key.

VI. Future scope and Conclusion

Considering the exponential growth of data, scope of the system can be enhanced to cloud storage and its integration with external sources of data. Implementation of the data mining techniques by the use of Cloud Computing will allow the users to get the meaningful information and insights from virtually integrated data warehouse that would eventually reduce the cost of storage and infrastructure.

Data mining in the Cloud offers tremendous potential for extracting and analyzing the useful information in different fields of human activities like business, advertising, economics, health care medicines, biology, heredity, pharmacy, etc. The use of this technology should allow that with just a very few clicks of mouse one would be able to reach the preferred information about clients regularity of purchases of certain items, behavior, wellbeing, purchasing power, spot and so on.

Seamless User Experience can be provided: Analytics solutions must integrate with other tools so as to become standard. Many companies use Sales force for CRM, for instance. But while they live inside Sales force, a lot of their metrics and data do not. As the analytics gets more and more integrated, you can use the Sales force for CRM or also use a tool like Tableau for analytics, bring the Sales force data into Tableau, then can also use the links into the analytics so as to jump right back into Sales force. For any end user or salesperson, the experience will always be seamless. Multimedia Data Mining is the mining and analysis of various types of data, including video, images , animation and audio. The main idea of mining the data which contains various different kinds of information is the main objective of multimedia data mining. As the multimedia data mining involves the areas of text mining and also hypertext/hypermedia mining, these fields are very closely related to each other. Much of the information which describes these other areas also applies to multimedia data mining. This field is also new, but it holds much promise for the future.

Thus, the Unstructured Data Mining would be of great help to all the Business Analysts to draw useful insights in a very short span of time and thus propose the decisions which would indeed be for the betterment of their firm.

VII. Acknowledgement

We would like to express our gratitude towards Neeta Maitre Mam for her kind co-operation and encouragement which helped us in completion of this paper.

REFERENCES

[1] Han & Kamber & Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann.

[2] Nurdatillah Hasim and Norhaidah Abu Haris. 2015. A study of open-source data mining tools for forecasting. ACM, New York, NY, USA, Article 79, 4 pages. DOI=10.1145/2701126.2701152