Review Paper for Types of Data in Big Data and Text Mining

Dr. Mehul P. Barot
Assistant Professor,
Computer Engineering dept.,
LDRP-ITR,

Mitul G. Prajapati,
Student,
Information Technology dept.,
LDRP-ITR,

Raj R. Patel
Student,
Information Technology dept.,
LDRP-ITR,

***Abstract*:- In recent world, big data is very popular term. Big data is generated from many sources such as social media, digital images or videos and so on. The data mining is very helpful for extracting useful information from big data. Mining of different types of data has different types of challenges in this IT era. Data mining tools are developed and improved day by day as the size of the data is growing tremendously at faster pace and it becomes more difficult to extract valuable information from such large amount of data of different type that is available these days for data mining or extraction of knowledge [1]. Various forms of data available in the digital world need different data models for their storage, processing and analysis. This paper discusses various kinds of data with their characteristics with examples, and also represents that the growing data is responsible for the numerous emerging data models and database evolution. In this paper, we will also discuss about Text Mining.**

Keywords:- Data mining; Data; Structured Data; Semi Structured Data; Unstructured Data; Text Mining.

INTRODUCTION

Data mining is a powerful tool which will facilitate to seek out hidden patterns and various relationship between the data. Data processing discovers hidden facts from massive databases. The overall objective of the data mining technique is to extract information from a huge data set and transform it into a comprehensible structure for more use.
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 [6].

BRIEF DESCRIPTION

There are three types of data in data mining. These are as below:

- Structured data
- Unstructured data
- Semi Structured data

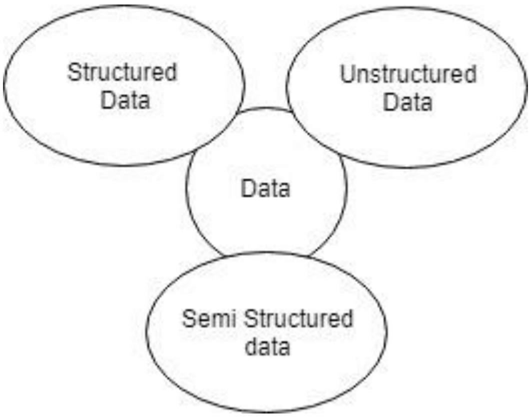

Figure.1 Kinds of Data

STRUCTURED DATA

Structured data are stored in structured form, In row and column structure. We can easily retrieve and analyze necessary data form structured data. For data mining structured data are very useful. So we can do mining and retrieve useful knowledge from them. ex. Excel, Database, Table, etc..

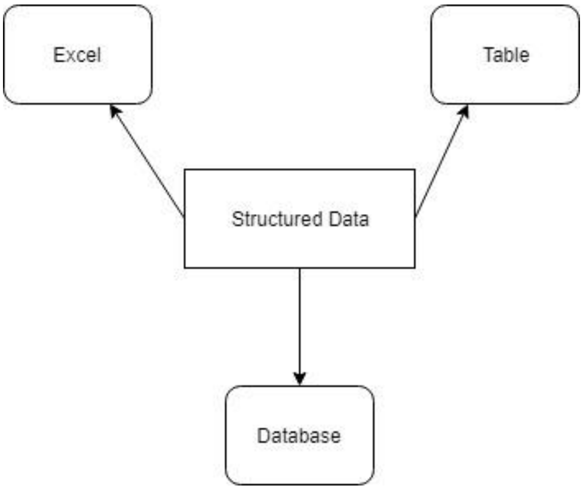

Figure.2 Structured Data

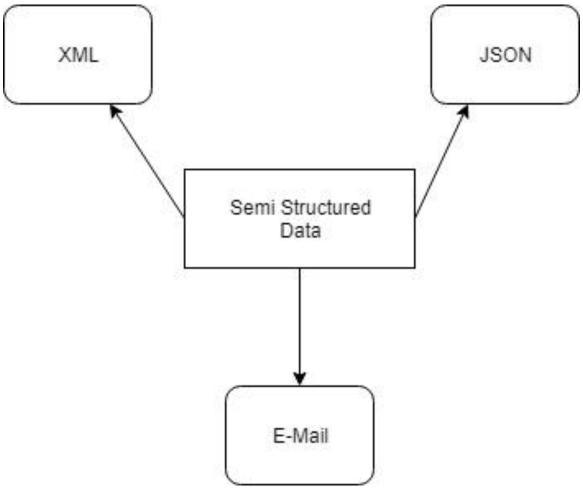

Figure.3 Semi Structured Data

Most of the content of the web pages are in the XML forms. These contents are included in structured data, companies like Google uses structured data to find on the web to understand the content of the page [7]. This way most of the Google search is done with the help of structured data.

CHARACTERISTICS OF STRUCTURED DATA

- Easily interaction with data
- Handle through SQL (Structured Query Language)
- Different data types (date, name, number)
- Structured data are dependent on schema, it is a schema based [3]

SEMI STRUCTURED DATA

Semi-structured data is a form of structured data that does not obey the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure.

In semi structured data, the entities belonging to the same class may have different attributes even though they are grouped together, and the attributes order is not important.

Semi structured data are increasingly occurring since the advent of the Internet where full text documents and databases are not the only forms of data anymore, and different applications need a medium for exchanging information. In object-oriented databases, one often finds semi-structured data [8].

XML and JSON database are example of semi structured data. In this data are stored in tree format so we can use data from them. Semi structured data are intermediate of structured and unstructured data.

CHARACTERISTICS OF SEMI STRUCTURED DATA

- It is not based on Schema
- It is represented through label and edges
- It is generated from various web pages
- It has multiple attributes [3]

UNSTRUCTURED DATA

Unstructured data are data which are not in structured format. All unformat data are known as unstructured data. Worlds 90% data are unstructured data. All PDF, Word, Audio, Video, Image file contain unstructured data.

For any analyses we have to convert over unstructured data into structured data. These things are done by different tools.

TOOLS/TECHNIQUES TO HANDLE UNSTRUCTURED DATA

- The different techniques used to search analyze and deliver unstructured data (4) are
  1. Content management system
  2. Relational Database
  3. Data Mining
  4. Text Analytics. Federal search or enterprise search data base
  5. Non-relational database
  6. Real time data visualization tools
  7. E-discovery application

- The new technologies for unstructured data era (4) are
  1. Log monitoring and reporting tools
  2. In-memory databases
  3. NOSQL databases
  4. Hadoop
  5. MPP data warehouses

These technologies bring high value information in real time instead waiting to store and perform operations like traditional methods.

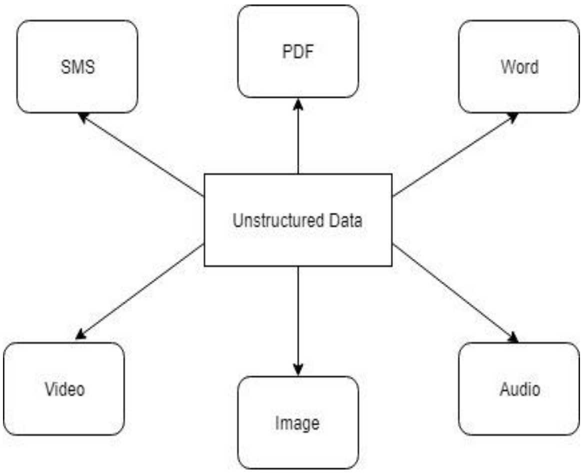

Figure.4 Unstructured Data

CHARACTERISTICS OF UNSTRUCTURED DATA

- It is not based on Schema
- It is not suitable for relational database
- 90% of unstructured data is growing today
- It includes digital media files, Word doc., pdf files
- It is stored in NoSQL database [3]

CHALLENGES IN UNSTRUCTURED DATA MINING [2]

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: Now a day's volume of data is increase day by day and most data in unstructured format so handle this data is very big challenge for us.
- Relevance: One way in which relevance comes into play is lack of insight into the previous story of certain pieces of data.
- Heterogeneity: There are many different resources from we get unstructured data. So, pattern and properties are varying as resource.
- Incompleteness: For do any analysis we have to make sure our data are complete. There are no missing data and noisy data are in data.
- Quality: Big amount of data require to get more secure analysis and get best result from that.
- Requirements for real-time data analysis has been predominant like for weather predictions, stock trading, time series, etc.

TEXT MINING

Most of people focus on structured data for any analysis. From data warehouses and relational and transactional database but most data are stored in textual format in unstructured or semi structured data.

So, we have to do text mining for convert that data in structured data and get data from textual format. Below figure shows the process of text mining from any document:

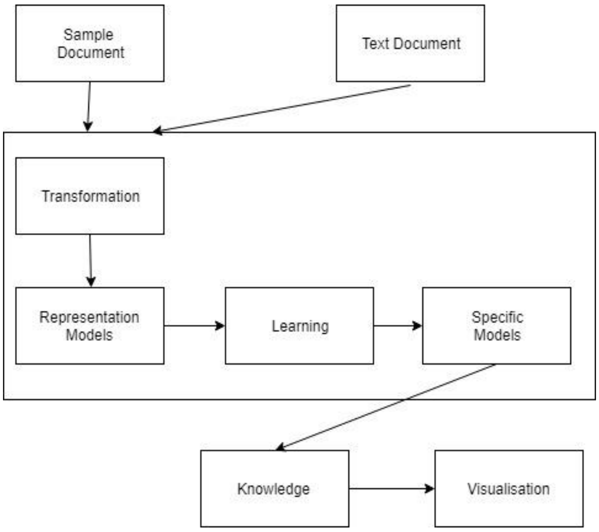

Figure.5 Text Mining

Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining [1].

Some application in which Text Mining are used:

- Security Application
- Biomedical Application
- Software Application
- Online media Application
- Sentiment Analysis

Some techniques for Text Mining are:

- Information Extraction
- Information Retrieval
- Categorization
- Clustering
- Summarization

Technique	**Characteristics**	**Tools**
Retrieval	Retrieval valuable information from unstructured text	Intelligent Miner, Text Analyst
Extraction	Extract information from structured database	Text Finder, Clear Forest Text
Summarization	Reduce length by keeping its main points and overall meaning as it is	Tropic Tracking Tool, Sentence Ext Tool
Categorization	Document based categorization	Intelligent Miner
Cluster	Cluster collection of documents, Clustering, Classification and analysis of text document	Carrot, Rapid Miner

Table.1 Techniques and Tools for Text Mining

CONCLUSION

Data mining is very important for analyze different types of data and extraction different pattern and knowledge from them.

In this paper we learn that there are three types of data. which are structured, unstructured and semi structured. They have different characteristic and applications.

World 10 % data are structured and 90% data are unstructured. We need structured data for analysis so we get easily useful knowledge. We have to convert unstructured data to structured data by different tools for analysis. By using these tools, we get interesting pattern and knowledge from data.

In this paper we also discuss about text mining. How we can convert text in structured format by different tools and techniques and use that data for data mining.

REFERENCES

[1] Review of Types of Data Used for Data Mining - Vipan, Robin Kumar

[2] UNSTRUCTURED DATA MINING AND ITS APPLICATIONS - Jagruti Jangal Wagh, Jidnyasa Dharmik Gondane, Ashvini Tulshiram Dukare.

[3] Influence of Structured, SemiStructured, Unstructured data on various data models - Shagufta Praveen, Umesh Chandra

[4] Unstructured Data Analysis-A Survey - K.V.Kanimozhi, Dr.M.Venkatesan

[5] J. McKendrick. Survey on unstructured data, produced by Unisphere Research 2011;

[6] Available from: http://www.ciosummits.com/media/pdf/solution_spotlight/ Marklogic_ 20 11- survey.pdf, 2011.

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

[8] https://developers.google.com/search/docs/guides/ intro-structured-data

[9] https://en.wikipedia.org/wiki/Semi-structured_data