MULTIMEDIA BIG DATA MINING

Rafael Marques Braga

University of Miami
rxm403@miami.edu

ABSTRACT

In today's digital era, the world has created more data than ever and has stepped into a multimedia big data age. It has become crucial to be able to find patterns and extract knowledge from the billions of multimedia bytes that are created every day. This paper will look at how big data is mined in order to efficiently extract information from it. It will also examine the challenges, applications and tools used for big data mining when it comes to multimedia.

Index Terms— *Big data, data mining, multimedia, multimedia storage, machine learning*

1. INTRODUCTION

In the past couple of decades data has become one of the most important goods for both individuals and corporations. It is something that is consumed every day in the order of quintillion bytes of data [1]. It is estimated that the world had an estimated 44 zettabytes of data by 2020 [1], with the number growing at a faster rate every day. With over 4 billion daily online users, massive amounts of data need to be processed every day at an increasingly faster rate in order to become accessible for every individual on the internet, for companies to find patterns in their sales, or for new scientific discoveries.

As we process these large amounts of data, or big data, new data mining techniques and processes need to be created so the data can be analyzed efficiently and used to deliver new business intelligence, relevant information and patterns between data systems and processes as well as aim to give summarized, targeted, and relevant information about a specific, large dataset [2]. Big data is a term used today when the data is too large and complex to be handled by conventional data processing tools and applications [3].

Most of the data that is created over today's digital world is not usually organized in sets of numbers with rows and columns (relational database), in a time series data file, or just in plain text with a structure (flat files). Instead, most data are created in a form of multimedia. This includes the data that is consisted by audio, video, images, and text media, either combined or separate. Data in these formats are harder to analyze because these files are typically more abstract and cannot be easily processed by the usual data mining algorithms. A main reason that multimedia is not compatible with the usual ML algorithms is that these files are much heavier than usual text files, contain quite a lot more information, and are existent in different formats.

This is the reason why the entire field of multimedia big data mining was created. This field focuses on the discovery of extracting interesting and unknown patterns from multimedia databases that store and manage large collections of multimedia files, including image, video, audio, and sequence data [3]. Multimedia data mining includes multi-interdisciplinary fields, such as image processing, computer vision, data mining and pattern recognition [3].

This paper will discuss the techniques used in multimedia big data mining, its storage and management methods, challenges, and solutions for multimedia data mining, as well as today's strengths and limitations for the existing methods being used for today's data mining of multimedia.

2. MULTIMEDIA DATABASES AND DATA MANAGEMENT

In today's digital world, storing so much data is a huge problem for large corporations like Google, Facebook and Microsoft. While multimedia is naturally large as files, when it comes to big data, corporations face multiple challenges to ensure that the computational time and storage capacity is reduced, while maintaining the results as accurate as smaller datasets. The databases made to store and manage multimedia are specialized and an essential part of maintaining the efficiency of the database. These are called Multimedia Database Management Systems (MMDBMS) [5]. These are characterized by a number of requirements, including storage, spatial and temporal constraints as well as retrieval and presentation [5].

The requirements for the MMDBMS are the following:

1. **Traditional Database Management Systems (DBMS) capabilities:** MMDBMS should provide the same set of functionalities as traditional DBMS. These include the integration, privacy, recovery

and data independence as well as the integrity and concurrency control among others [5].

2. **Multimedia Data Modeling**: Data modeling holds the objective of allowing the automatic retrieval of targeted information and representing it formally based on the related domain. This is different from usual DBMS because although there are multiple data modeling tools in DBMS, few are specialized for multimedia and for its different types of media data. In MMDBMS, the data modeling requirements are specific for multimedia and its data types [5].
3. **Huge Capacity Storage Managements:** Since multimedia is inherently large in size and comes in high volumes, MMDBMS are specific for even larger files and are able to move and store such files in a quick and efficient manner [5].
4. **Query support and information retrieval:** Images, videos and other types of multimedia files are usually harder to query and retrieve information from. MMDMBS have a focus on supporting multiple forms of queries such as keywords, content, and context based. Instead of returning exact matches like traditional DBMS, MMDMBS returns a list of results ranked by similarities to the query [5].
5. **Multimedia Interface and Interactivity:** MMDMBS are capable of dealing, viewing, and intuitively interacting with the different multimedia files. This includes specific interfaces in order to support and handle these datatypes [5].
6. **Media integration, composition, and presentation:** It is essential for these specific database systems to effectively integrate and compose different forms of data for better presentation. This includes the support and creation of an efficient design for the presentation of various media types, especially after converted to structured format [5].

There are also other specific requirements for MMDBMS which were not mentioned. These include performance, efficiency, and reliability of the systems to work specific with large multimedia files, as well as managing the duplication of files (temporal constraints).

MMDBMS are essential for effective multimedia big data mining as they provide the efficiency and ease for preprocessing as well as the capabilities to provide fast data cleaning, integration and transformation of the specific files from multimedia.

3. TECHNIQUES USED IN MULTIMIDIA BIG DATA MINING

3.1. Pre-processing

The first step for big data mining in multimedia after the data is collected and properly stored is the pre-processing of the data. After data collection, this multimedia data is considered to be raw data which is unstructured and noisy, therefore useless for the data mining tools, especially once studied in large scale. For such data to be mined, it first needs to be pre-processed. Data pre-processing is the conversion of raw data to a set of clean and structured data prepared for further mining analysis [5]. This includes data cleaning, data transformation and integration and data reduction [5].

Data cleaning is the step in which the noise is reduced, outliers are identified, and the inconsistencies are eliminated [5]. This is one of the most important steps of data mining and usually takes up to 60% of a data scientist's time.

Data integration is where multiple data sources and the different data types from multimedia are combined along with their metadata [5]. In this step, data is also transformed. This is where the data goes through normalization and formatting as well as the reduction of data redundancy [5].

Lastly, comes data reduction, which looks to improve the quality and speed of the data. Different techniques are used for this step with one of the main ones being feature reductions, which removes features from the data (which are thought to be not so useful) either through metadata or through algorithms. This will improve the speed and accuracy which the mining and analysis of the data is done.

3.2. Data Mining and Analysis

After the data is pre-processed, it is now ready for the mining and analysis. Different techniques are applied for big data mining when it comes to multimedia, notably because of its different data file types, including text, audio, image and/or video. Sometimes the same techniques can be used for analyzing the different types of data; however, these need to be translated and formatted to be used for such data type. In the following paragraphs, some of the techniques used for the analysis and mining of each multimedia file type will be examined.

3.2.1. Text Analytics

Text is one of the most common forms of media today and one of the easiest to analyze since it doesn't take a lot of work to adapt the algorithms to look into textual data instead of flat or relational files. Text analysis mostly focuses on the extraction of meaningful and structured data from the unstructured, unlabeled, and unorganized data [4].

The information extraction from the text plays a huge role as this technique helps to structure the unstructured data. Using information extraction helps to understand the various entities. Information extraction is when the algorithm

classifies the data into entities; for example, to classify a sad story as such from the text [4]. This can be done via machine learning algorithms. These can be either supervised or unsupervised. Supervised learning is when you need to have a class label as an outcome while using labelled datasets and unsupervised learning is when you cluster the data into different sections using unlabeled datasets. Some examples of supervised learning are Support Vector Machine (SVM), Hidden Markov models, decision trees or K-nearest Neighbor (KNN). Some unsupervised learning models are clustering and association models [4].

After the data is mined using one of the algorithms above, there are two main methods which tell how correct the information extracted from the algorithm is: precision and recall [4].

3.2.2. Audio Analytics

Audio analytics, as the name suggests, is when data patterns are extracted from audio signals. This can be used in multiple manners. One example of audio analysis that can be mined is after analyzing the sound of an infant's voice, his/her health can be determined. The main techniques used for audio analysis are:

- **Approach based on phonetics:** This is when the system converts the input audio into a sequence of phonemes, then the system searches for the output based on the labeling of phonemes [4].
- **Large Vocabulary Continuous Speech Recognition (LVCSR):** This approach has two steps. The first is to match sound to words using algorithms like Automatic Speech Recognition (ASR). Using this output, an indexed file is maintained providing the sequence of the words. In the 2nd step, the system extracts the information from the indexed file from step one [4].

3.2.3. Image Analytics

This method deals with information extraction from images and therefore uses various image processing techniques. Some of the applications using image analytics include facial recognition and movement analysis. This is the type of data along with videos which is usually found to be the most unstructured, so it is very important to have effective techniques for image analysis. Some of the algorithms that can be used for this are recurrent or convolution neural networks or deep learning algorithms.

3.2.4. Video Analytics

Video Analytics extracts information from videos, making it the most challenging one to extract among the multimedia files [4]. This is mainly due to its size, as 2 seconds of an HD video is equal in size to 4000 pages of text [4]. This is also a huge part of big data mining, as large amounts of data in video format usually need to be analyzed, ranging from hundreds of thousands of hours from CCTV footage or YouTube videos. For this, adapted machine learning techniques are also used to extract information. This includes neural network and deep learning. These algorithms are usually adapted from image analytics to analyze multiple frames of a video.

4. FRAMEWORKS AND TOOLS

Today, there are numerous frameworks and tools that are open source and focus on the big data mining for multimedia. These have data mining and machine-learning tools and libraries that can be implemented in order to analyze different data file types. These tools can help to mine big data faster and more efficiently as well as visualize the data correctly. Some of these open-source tools are Apache Mahout, Moa, WEKA, SparkR, GraphLab and Hadoop [5].

Frameworks like Apache Mahout provide a scalable environment for data mining applications which include a wide range of learning and mining algorithms like data stochastic analysis, classification, clustering and collaborative filtering. Using this framework, big data techniques from Hadoop, Spark H20 and Flink can also be used. The data mining algorithms can also all be applied to multimedia files. MOA, which is similar to WEKA, is an open-source data mining/machine-learning framework, but is very scalable, ideal to be used in big machine learning.

5. APPLICATIONS & CHALLENGES

With these frameworks providing scalability, the more computer power and resources are available, the further multimedia big data mining is possible. Even though the difficulty of dealing with a huge amount of large data files is high, the applications are endless. Some of the applications of multimedia big data mining as follows:

- Traffic and Pedestrian monitoring. One study was done for unsupervised feature extraction and clustering based on CNN deep learning conducted for pedestrian detection in video surveillance [5].
- Action/object detection in a video [5]
- Disaster events from web image datasets
- Goal detection in a soccer video
- Fake-news detector in social media
- Terrorist detection in social media [4]

However, even with the limitless applications for multimedia big data mining, it does not stop from having major challenges in order to become more efficient and affordable. In order to further study the challenges in multimedia big data mining, there are 5 Vs which are usually looked at specifically for this subject: Volume, Variety, Velocity, Veracity and Value. [4].

Volume is the first one. With over 100 hours of YouTube videos being uploaded a minute, all this data must be quickly sorted. The huge amount of data that exists today is exponentially growing, posing a great threat on how computer power and storage can keep up with the data creating rate.

Variety is also huge in multimedia, beginning from all the different file types as well as sources. Indexing and structuring such big variety of multimedia data has become a difficult task in today's world.

As for velocity, as the amount of data increases, the longer and more computer power it takes to process such data. Even though recent advancements using scalability, GPUs and FPGAs have greatly improved the velocity at which big data is mining, it still one of the major challenges of multimedia big data due to its huge data file sizes.

Veracity looks at the accuracy of the data as an issue and challenge. One example is how difficult it is to identify a corrupted source from one of thousands of security cameras [4]. Not all the data, especially in multimedia, can be trusted today. With many sources of multimedia being from social media, the accuracy of these sources is usually very limited and constrained.

Lastly, value. This looks at how useful it is to filter, sort and select essential information from all the data we have available today in order just to keep and use the most valuable datasets.

These sets of values must always be analyzed when looking at the different applications of big data mining when related to any type of text, image, audio, or video file. These are used to evaluate the challenges and the opportunities that accompany the multimedia big data topic.

6. CONCLUSION

To summarize, big data mining in multimedia is one of the biggest applications there are for data mining. In today's digital world and increasing amounts of data that need to be analyzed in a fast and efficient way, big data mining provides some efficient tools in order to store, preprocess and analyze this data in real time. MMDBMS are essential databases in order to store and manage data efficiently, while scalable frameworks such as Apache Mahout or SparkR give us the opportunity to use computers in parallel to analyze new patterns in big data using different and typical machine learning tools such as neural networks or deep learning.

Using these frameworks and tools, a simple KNN algorithm can be used to cluster, sort, and extract features from thousands of videos and hundreds of gigabytes of data in a matter of hours. Furthermore, with visualization tools from databases, multimedia can easily be visualized and interpreted, facilitating navigation and presentation of the data studied.

7. REFERENCES

[1] "How much data is created every day? [27 powerful stats]," SeedScientific, 07-Feb-2022. [Online]. Available: https://seedscientific.com/how-much-data-is-created-every-day/#:~:text=How%20much%20content%20is%20created,rate%20will%20become%20even%20greater. [Accessed: 28-Feb-2022].

[2] Techopedia, "What is Big Data Mining? - definition from Techopedia," Techopedia.com, 16-Apr-2014. [Online]. Available: https://www.techopedia.com/definition/30215/big-data-mining. [Accessed: 28-Feb-2022].

[3] J. Han, M. Kamber, and J. Pei, "13 - Data Mining Trends and Research Frontiers," in Data Mining: Concepts and Techniques, San Francisco (Calif.): Elsevier, 2012, p. 596.

[4] A. Kumar, S. R. Sangwan, and A. Nayyar, "Multimedia Social Big Data: Mining," Intelligent Systems Reference Library, pp. 289–321, 2019.

[5] S. Pouyanfar, Y. Yang, S.-C. Chen, M.-L. Shyu, and S. S. Iyengar, "Multimedia Big Data Analytics," ACM Computing Surveys, vol. 51, no. 1, pp. 1–34, 2019.