EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Intricacies of Unstructured Data

Rajeshwari K Rai[1]

[1]BITS Pilani

Abstract

This research paper is an attempt to explore unstructured data that has taken the world of data science by storm. Handling the unprecedented growth of unstructured data is a big challenge faced by most data scientists. This paper discusses about the types and sources of unstructured data and effective ways of processing and analysing this data.

Keywords: Big data, Unstructured data, Audio/video data, Hadoop, Data analysis. Machine learning

Received on 29 July 2017, accepted on 09 September 2017, published on 25 September 2017

doi: 10.4108/eai.25-9-2017.153151

1. Introduction

The boom in the usage of social media and e-commerce has led to massive collection of data specially unstructured data. Research has shown that 80% of all stored organizational data is unstructured (1). IDC and EMC project that data will grow to 40 zettabytes by 2020 of which unstructured data will probably occupy a major chunk (2).

Analysing this big data is a herculean task as it does not have a pre-defined structure or is not organized in a pre-defined manner. Effective mechanisms has to be explored to clean and extract meaningful insights from this massive data collection.

Comprehensive extraction of useful insights from unstructured data is not being done. Most data mining techniques can only handle structured or semi structured data. Sometimes during data pre-processing many valuable information is lost. Due to the complexity involved in segregating and analysing unstructured data some portion of the data is discarded. There are lot of useful information hidden in unstructured data however most people are clueless about the process of extracting the rich information from these unstructured data sources. This paper will review some of the effective ways of mining various types of unstructured data. The focus will be on image processing, text mining, and speech recognition as these data types have lot of significance in domains such as healthcare, education and environmental study.

Corresponding author. Email: krrai77@gmail.com

2. Unstructured Data

Big data comprises of structured, semi structured and unstructured data. However unstructured data dominates and its growth has been substantial over the last 10yrs (Fig.2). Unstructured data includes data generated from social media, images, videos, and audio files. Data from social media such as Facebook posts, tweets, LinkedIn feeds not only contain textual information but are also heavy in survey outputs (graphs and charts), links to other posts, and numerical. This combination of irregular data is not easy to put in a pre-set framework and analyse like structured data. Unstructured data is ambiguous and noisy. It has to be pre-processed to find useful information and recognize patterns. This is a major challenge due to the sheer size of data.

2.1 Image Data

Facebook users view 2.77 million videos every minute. There is a massive growth in video and photo data, where every minute up to 300 hours of video are uploaded to YouTube alone. By end of 2017, nearly 80% of photos will be taken on smart phones. By 2020, it is estimated that we will have over 6.1 billion smartphone users globally. The most astounding finding is that only less than 0.5% of all this data is analysed and used (14).

Due to the substantial growth in image data over the years (Fig. 1) importance of image analytics has taken precedence. There is increasing demand in the healthcare field for image processing. Machines are trained to predict outcomes based on visual reports. Millions of medical reports such as CT Scans are fed to image processing algorithms to learn, and recognize various medical conditions.

Image processing involves the following steps:

- Pre-processing of images where image features such as size and clarity are analysed.
- Cleaning images that might involve correcting the image visibility, selecting areas of interest.
- Extract data from image.
- Segment image to separate objects from the image background.
- Image recognition and interpretation.

Some of the commonly used tools for image processing include packages in Python (scipy, scikit-image) and R (imager), and Matlab.

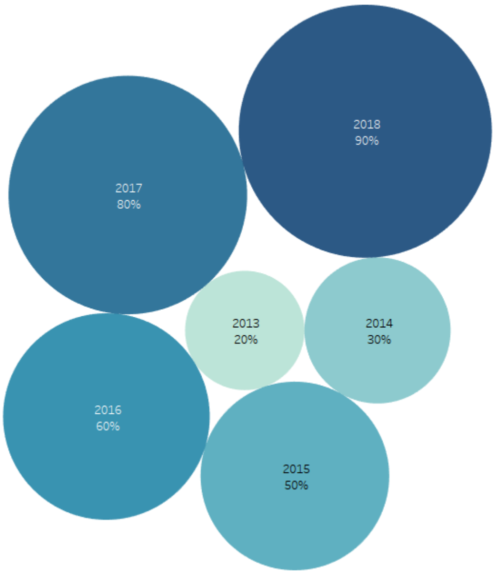


Figure 1. Percentage increase in image data

2.2 Textual Data

Digital footprints are closely watched like a spy by retailers, social media giants like Facebook to get up and close with the users. User’s data collected online is mined for information such as their interests, dislikes, etc. These information are very aptly used by vendors, search engines, social media to broadcast news feeds and ads matching user’s likes and dislikes.

Text mining is very popular and also complex. Search engine giant Google had to deal with crawling through millions of webpages for a typical search query. Google uses web crawlers to organize information from webpages in the Search Index. Powerful search algorithms are used to sort through billions of webpages to return relevant results to the users in a fraction of a second. Increased usage of mobile devices due to high speed internet access at affordable rates has changed the components of unstructured data and increased its volume. Apart from textual data it also contains video/audio content, images, open-ended surveys among other forms. Always the right ad or news feed may not be delivered. Computers may not be able to interpret homonyms (words with same spelling and pronunciation used in different contexts). This is one of the challenges faced in text classification given the mammoth amount of data that is generated every minute. It is estimated that by 2020, about 1.7 megabytes of new information will be created every second for every human being on the planet (14). Facebook's FastText library was developed with an objective of quick text classification and efficient word representation learning.

Text classification involves hypothesis generation, data pre-processing, data exploration, and finally building a model. Some of the commonly used techniques for text mining are bag of words, removing stop words, and lemmatization. Many tools are available for text mining such as SAS, Orange, and KNIME (open source). Even Python and R can be used to write robust programs for all the processes involved in text classification.

Text classification that is part of Natural Language Processing (NLP) is one of the most widely used methods to analyse huge volumes of unstructured digital content. As this is a time consuming and expensive task one has to find means of doing this process faster and in an effective manner. Google supported open source tool Word2Vec provides an efficient implementation of bag-of-words and skip-gram architectures for computer vector representation of words (3).

2.3 Audio Data

Audio data is another form of unstructured data that is growing in popularity. Most BPO and call centres record customers calls to extract useful information such as customer's positive and negative feedback, queries, employee's ability to handle customers, etc. Transcription of audio data is now outdated. Innovative methods of audio analysis are used to capture information from audio files.

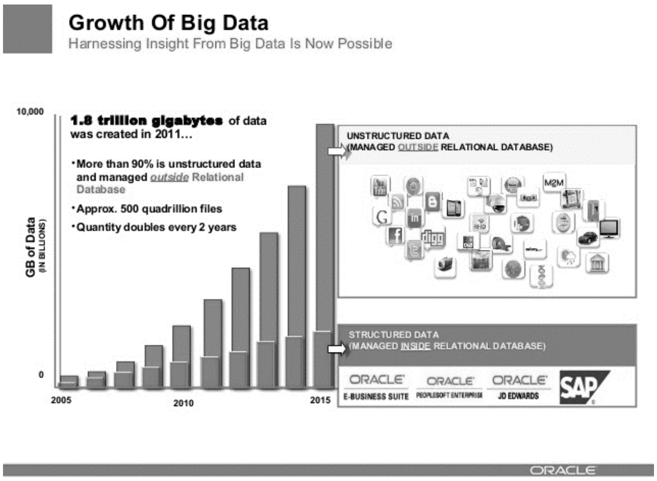
In few English testing exams such Pearson's, speech recognition technology is used to assess candidate’s speaking skill. The test focuses on oral fluency, intonation and pronunciation rather than the content. The model is optimized for non-native speech. Software interpret these sounds and represent it in computer readable formats such as mp3 and wav.

Audio analysis also involves pre-processing of the audio data where audio files are converted to different domain of data representation such as frequency domain. Next useful features are extracted from the audio representations and analysed. Finally models like deep learning models are executed on the data to get predict the output. Some of the tools used for audio analysis are SoundRuler, Marsyas, LibROSA library in Python, and tuneR in R.

2.4 Sensor Data

Another important type of unstructured data is the machine-generated real-time data. Sources for this include weather data from satellite images, scientific data, radar data like vehicle movements, and data collected from GPS. This kind of data requires dynamic analysis based on real-time data. The analysis has to happen constantly and reports/results generated continuously. This is a challenging task as information provided has to be accurate and relevant.

Data generated from sensors is massive. Every minute real time data is captured from these sensors and sent to data servers. These sensors can be GPS that provide location information to users or it can be sensors installed at traffic signals to record the vehicular density and capture other statistic data. There are varied Internet of Thing (IOT) sensors installed for different purposes like monitoring air quality, keeping tab on deforestation, tracing lost items such as mobile phones, etc. Most of the sensor data recorded by these devices are unstructured and has to be transformed to pre-processed, structured data for recognizing patterns and predicting trends.



Growth Of Big Data

Harnessing Insight From Big Data Is Now Possible

1.8 trillion gigabytes of data was created in 2011...

- More than 90% is unstructured data and managed *outside* Relational Database
- Approx. 500 quadrillion files
- Quantity doubles every 2 years

UNSTRUCTURED DATA (MANAGED OUTSIDE RELATIONAL DATABASE)

STRUCTURED DATA (MANAGED INSIDE RELATIONAL DATABASE)

Figure 2. Growth of big data

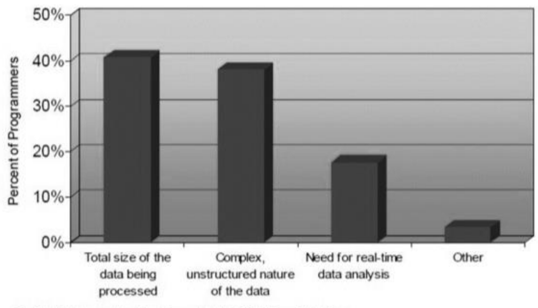
3. Analysis of Unstructured Data

Multimedia content are powerful sources of data. Many tools such as NVivo and MAXQDA are available for comprehensive analysis of images. Qualitative data analysis packages assist in the analysis of multimedia data. There are tools that directly code media files without transcribing them. Transana is a video analysis tool that helps in analysing and comparing multiple videos simultaneously (4).

There are significant amount of operational data available that provides vital information about business processes, events and operations as they are taking place. With the rise in online transactions, consumer expectations have also increased and hence outcomes based on most current data is crucial for better customer relationships. Analysis of such data requires event driven techniques that may use Complex Event Processing, Event Stream Processing and Mashup (web application hybrid) techniques to enable events to be analysed in-memory rather than being first stored in a database and then analysed. This reduces data latency and in turn analysis latency (5).

As per the recently published report, Big Data and Advanced Analytics Survey 2015, Volume I by Evans Data Corporation total size of the data being processed is 40.8% (Fig 3.). Out of this complex, unstructured nature of the data is 38.1% and the need for real-time data analysis is 17.7%. These are the top three factors driving big data adoption over traditional database solutions.

Hadoop is extensively used to handle the volume and complexity of big data. It contains several open source tools for processing, managing and analysing huge data. Since its open-source many use it for distributed storage and processing of big data. Hadoop environment can also be configured to process real time data. Hadoop scales well for growing volume of data from different sources (6).



Big Data & Advanced Analytics Survey, Vol. 1 ©2015 Evans Data Corp.

Figure 3. Top three factors driving big data adoption.

4. Managing Unstructured Data

The growth of unstructured data is phenomenal and it is expected to grow higher. Managing this complex volume of data in terms of storage, processing and retrieval is a challenge. Some of the popular tools used for storage and querying/analysis of this data are Apache Hadoop, Microsoft HDInsight, NoSQL, and Hive. Hadoop splits big data and distributes across many nodes in a cluster. NoSQL (Not Only SQL) is suitable for storing massive amount of unstructured data. Hive is used for data summarization, query and analysis.

Most organization's expenditure on storing data has steadily risen and expected to double in the coming years. This specially holds true for the fast-growing unstructured data. Usually this massive data is spread across multiple storage repositories such as distributed systems and cloud. This leads to difficulty in tracking the exact location of specific data and one has to scan through all the metadata across different storage systems. There has been ways to search and retrieve data from different repositories through highly scalable file analysis methods. These methods identifies data by its characteristics and then classifies it based on predefined rules (7).

5. Machine Learning and Unstructured Data

Unstructured raw data is useless without extracting its value. Analysing such enormous data manually is an impossible task. The only way to quickly and efficiently get useful insights, is to use machine learning algorithms such as text-mining, pattern/classification, and Natural Language Processing (NLP). Tools such as Adobe Analytics are used by publishers and content managers to get key information from the data pool such as how many people view the online content each day and what specific part of the content particularly engages the users (8).

Table 1. Data mining algorithms (10)

Tasks	Descriptions	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known data set.	Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbours.	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known data set.	Linear regression, logistic Regression.	Predicting unemployment rate for next year Estimating insurance premium.
Anomaly Detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, density based, local outlier factor (LOF).	Fraud transaction detection in credit cards Network intrusion detection.
Time Series	Predict the value of the target variable for a future time frame based on historical values.	Exponential smoothing, autoregressive integrated moving average (ARIMA), regression.	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	k-means, density-based clustering (e.g., density based spatial clustering of applications with noise [DBSCAN]).	Finding customer segments in a company based on transaction, web, and customer call data.
Association analysis	Identify relationships within an item set based on transaction data.	Frequent Pattern Growth (FP-Growth) algorithm, Apriori algorithm.	Find cross-selling opportunities for a retailer based on transaction purchase history.

There are many data mining tools such as R, SAS Enterprise, and IBM SPSS that implements these algorithms easily. As Eric Siegel has quoted in his book, Predictive Analytics (Siegel, 2013) that, “if all the data in the world was equivalent to the water on earth, then textual data is like the ocean, making up a majority of the volume”, textual data is the key product of data retrieval and this text data is mainly unstructured. This unstructured data has to be converted to semi-structured data through text mining. Then analytic techniques can be applied to classify and predict outcomes from the data (9).

Oracle Data Mining provides comprehensive data mining functionality within Oracle Database that is especially useful for document classification. Some of the algorithms supported by this functionality are Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. Unstructured thousands of text features can be easily processed through these techniques. It is essential that analytic applications evaluate the structured data along with unstructured information. Oracle Data Mining offers this capability. It can be used to mine data sets that contain regular relational information (numeric and character columns), as well as one or more text columns (10).

5.1 Artificial Intelligence and Unstructured Data

Andrew Ng, the former chief scientist at Baidu makes a notable statement that, "Data is the new fuel that will power the digital generators of the AI revolution". Huge inflow of data does not imply that Artificial Intelligence (AI) can directly make use of it. The data has to be transformed to fuel the emerging AI inventions (11).

Artificial Intelligence provides ways to organize the massive volume of unstructured data. AI technologies helps to provide instant response to people’s queries by scanning the endless stream of data. Google's ML tool

TensorFlow allow speed recognition and conversion to text, image recognition and translation capabilities. Machine leaning technologies are used to create metadata from live events by focusing on the audience. This helps in gauging and analysing human reactions for different situations. This information helps providing recommendations and custom news feeds to the users (12).

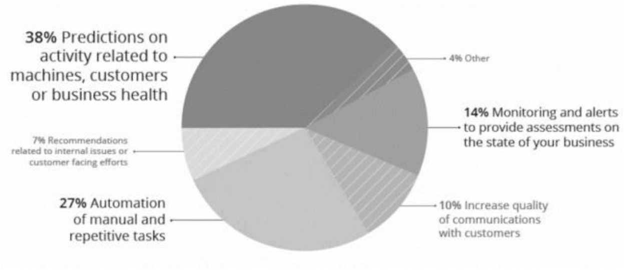


Figure 4. Outlook on AI-Powered solutions (13)

6. Conclusion

The era of big data will last for many more years to come. Its volume will grow in leaps and bounds and data from all genres has to be accommodated. Data from mobile devices will be the major contributor for this. Real time data analytics will gain more prominence and dynamic analysis of real time data will be a challenging task for data analysts.

Since majority of the data collected is unstructured from sources such as social media, mining of relevant information becomes the key factor. Search engines such as Google strive to provide the most relevant result for the search queries in the shortest time possible. Online retailers such as Amazon keep improving their recommender systems to provide best user experience for their customers. It is not enough to just retrieve, process, and explore unstructured data but one has to predict and forecast future patterns. Unstructured data has no value by itself unless it is thoroughly scanned for useful insights.

With growing volume of data one has to effectively use the available tools and techniques to meet the demand of predictive analytics. As data is constant, methods has to vary to suit its types and needs. For example, accurate face recognition through image processing has helped trace many missing people, and important forecasts of natural calamities has saved many lives.

There is still lots to unearth about big data and there is always scope to improve existing techniques of managing unstructured data. As data is the undisputed entity that fuels AI and ML technologies, it demands more resources and effort to manage efficiently and economically. This paper suggests to work on innovative ways to store, scale, analyse and secure data as it's an asset for posterity.

References

[1] C. White. Consolidating, Accessing, and Analysing Unstructured Data. 2005 Dec. Business Intelligence Network article. Powell Media. LLC.

[2] Christopher C. Shilakes and Julie Tylman, "Enterprise Information Portals", Merrill Lynch, 16 November 1998.

[3] https://code.google.com/archive/p/word2vec/.

[4] Digital Tools for Qualitative Research by Trena Paulus, Jessica Lester, Paul Dempster.

[5] The Evolution of Real-Time Business Intelligence, Gravic.com. Retrieved 2012-09-19.

[6] David Loshin, Knowledge Integrity Inc.

[7] http://www.infostor.com/storage-management/manage-unstructured-data-file-analysis-at-scale.html.

[8] Trevor Paulsen, March 28, 2016.

[9] Predictive Analytics and Data Mining, Vijay Kotu, Bala Deshpande, 2015 March.

[10] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/intro_concepts.htm#DMCON001.

[11] Jim Crowley, Forge.AI.

[12] Tom Coughlin, Forbes.

[13] Narrative Science and National Business Research Institute.

[14] Bernard Marr, 20 Mind-Boggling Big Data Facts Everyone Must Read.