International Journal on Emerging Technologies (Special Issue NCETST-2017) **8**(1): 570-572(2017)
(Published by Research Trend, Website: www.researchtrend.net*)*

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

An Analysis of Unstructured Textual Data for uniformly storage

***Rudransh Saxena*[1]*, Paras Chawla*[1] *and Mohit Giri Goswami*[2]**

[1]*Scholar, B. Tech, Computer Science and Engineering,*
Amrapali Institute of Technology and Sciences Haldwani, (U.K.), INDIA
[2]*Astt. Professor, Computer Science and Engineering,*
Amrapali Institute of Technology and Sciences Haldwani, (U.K.), INDIA

ABSTRACT: This is an overview of unstructured textual data and how it managed in companies. It also tell about how unstructured textual data is helpful for companies. We examined the totality of the world of unstructured textual data. There are many difficulties with unstructured textual data. It has many characteristics, depending on the type of unstructured textual data that is discussed. Different industries have different needs for unstructured textual data. Different functional areas in the organization have various needs for unstructured textual data. There are many different challenges when using unstructured textual data, such as challenges of terminology, volumes of data, and the cost of the infrastructure.

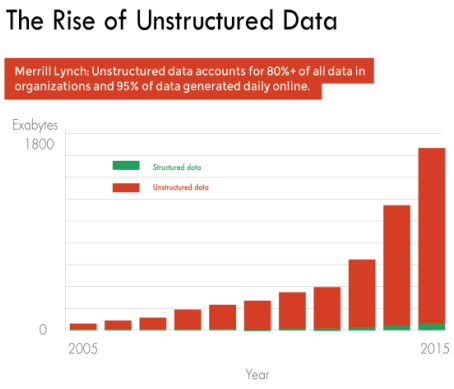
Keywords: UD, UDM and CUDSD

I. INTRODUCTION

More than 80% of information comes from unstructured data. Unstructured data is a generic label for describing any corporate information that is not in a data base that is information that either does not have a pre-defined data model or is not organized in pre-define manner.

Any data that has no identical structure like Power Point, excel file, word file, photos, videos etc.

Data is unstructured if its elements cannot be stored in rows and columns, which makes it difficult to query and retrieve by applications. For example, customer contacts that are stored in various forms such as sticky notes, e-mail messages, business cards, or even digital format files, such as .doc, .txt . Due to its unstructured nature, it is difficult to retrieve



II. HISTORY

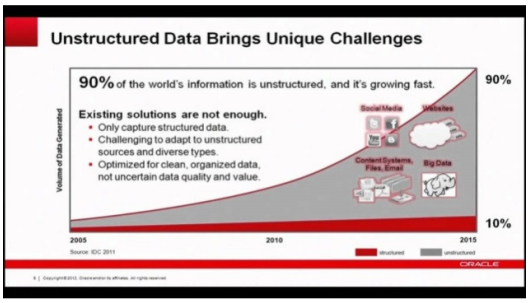
The earliest research into business intelligence focused in on unstructured textual data, rather than numerical data. As early as 1958, CS researchers like H.P. Luhn were particularly concerned with the extraction and classification of unstructured text. However, only since the turn of the century has the technology caught up with the research interest. In 2004, the SAS Institute developed the SAS Text Miner, which uses SVD to reduce a hyper-dimensional textual space into smaller dimensions for significantly more efficient machine-analysis. The mathematical and technological advances sparked by machine textual analysis prompted a number of business to research applications, leading to the development of fields like sentiment analysis, voice of the customer mining, and call centre optimization. The emergence of Big Data in the late 2000s led to a heightened interest in the applications of unstructured data analytics in contemporary fields such as predictive analytics and root cause analysis

III. CHALLENGES TO MANAGE UNSTRUCTURED DATA

a) A lack of tools that easily manage unstructured data. Tools need to provide efficient text parsing and analytics, taxonomy and metadata management.

b) Difficulty integrating unstructured data with existing information systems. The two are often seen as apples and oranges when it comes to analytics and decision making.

c) Shortage of skills in existing staff



IV. HOW UNSTRUCTURED TEXTUAL DATA WORKS IN COMPANIES

Most companies have two kinds of data structured data and unstructured textual data. Because structured data preceded unstructured data in the workplace, unstructured data is often best understood in contrast to structured data. Structured data is data that is represented by numbers, tables, rows, columns, attributes, and so forth. Most IT professionals have spent the better part of their professional lives with structured data. As its name implies, structured data is usually disciplined, well behaved, predictable, and repeatable.

Structured data usually is generated as by-products of doing a transaction. A check is cashed, an ATM activity is done, an insurance claim is made, a production run is completed, a car is sold—these are typical transactions that generate a lot of structured data about the activity that has been done. Although there is text in the structured environment, most text serves the purpose of identifying or describing some numeric data. The numeric data in the structured environment makes up the heart of the data that is found there and is heavily used for analytical purposes.

The other major category of data found in the corporation is unstructured data. There are several forms—textual unstructured data and non-textual unstructured data, which includes images, colours, sounds, and shapes.

Unstructured textual data is textual data found in emails, reports, documents, medical records, and spreadsheets. There is no format, structure, or repeatability to unstructured textual data.. In addition, there are other forms of text that occur well outside the email environs, such as contracts, warranties, spreadsheets, telephone books, advertisements, marketing materials, annual reports, and many more forms of textual information that are the fabric of the organization. In short, unstructured textual data occurs almost everywhere and represents both a challenge and an opportunity to the organization that wants to use it for decision-making purposes.

It is true that many forms of unstructured data are not text-based. There are X-rays showing bones and breaks, real-estate listings with pictures, engineering change control documents mapping the structural changes made to complex edifices, MRIs that show detailed aspects of the human body, and scientific photos that help mankind unlock the secrets of the universe. But the most basic form of unstructured data is in the form of text. The focus in this book is on text, which presents its own set of challenges.

V. NON-UNIFORM CHARACTERSTICS OF UNSTRUCTURED TEXTUAL DATA

Emails—Emails are usually short relative to other documents. Emails usually contain a combination of business-related and non-business-related information. There are usually a lot of emails, many of which are informal. Emails can be in any language (English, Spanish, and German). In fact, emails can be in more than one language at the same time. Emails are normally identified by an email address and the time the email was sent.

Medical records—Medical records can become quite voluminous and are full of medical jargon and terminology. Medical records are often quite large, some containing volumes of text. The reason why most medical records are text-based is that doctors prefer it that way. Most doctors prefer to write notes in text when dealing with a patient. In most environments, there are almost always fewer medical records than there are emails. Medical records are somewhat formalized, in the sense that doctors have a certain protocol that is used when writing a medical record. These records usually contain only information relevant to health and medicine.

Contracts—Contracts are usually full of legal jargon and might contain business related jargon as well. The text found in contracts usually has nothing but information relating to the contract. Contracts by and large vary greatly in size—some contracts are short, and some contracts are lengthy. Contracts are written in a legal style, according to the conventions of law and lawyers. There sometimes are a fair number of contracts in an organization, but never the number of contracts that there are emails. Contracts are almost always in a single language–for example, English, Spanish, or French.

Unstructured Textual Data and Organizational Functions.

To start to understand unstructured textual data, consider that there are different kinds of unstructured textual data associated with different functions within the organization.

Corporate Functions—Unstructured

The corporate function of accounting typically has spreadsheets, Word documents, audit reports, and audit

trails associated with its activities. Call centres typically have recorded or transcribed conversations, replies, follow-up activities, and other notes associated with their activities. The engineering department typically has unstructured textual data associated with the bill of materials, engineering changes that have been made, production archives, and design specifications.

Industries and Unstructured Data

Unstructured textual data is not just endemic to different departments of the organization. Unstructured textual data appears in different forms and in different measures in different industries. Some industries have a lot of unstructured textual data while others have little.

Table 1 shows that different industries have different mixes of transaction processing data and unstructured textual data. For example, banks are rich with transaction processing, checks, ATM activities, and other banking activities. Whereas banks certainly do have unstructured textual data, they do not have (relatively speaking) nearly as much unstructured textual data as they have structured data.

On the other hand, medical environments are rich in unstructured textual data. The unstructured data is so ingrained in healthcare that it is part of the fabric of medicine and healthcare. Doctors write and take notes in textual fashion hospitals take notes textually, and so on. The world of medicine is rich in text, and certainly transactions occur in the medical environment. Patients get billed on a regular basis, for example. But—relatively speaking— the medical environment is heavily ingrained with textual data.

Table 1.

Corporate type	Type of transaction process	Amount of unstructured data.
Banking	Heavy transaction (rx) processing	Light, scattered
Government	Tx processing	Heavy concentration
Medicine	Light tx processing	In the fabric

The Challenges of Unstructured Textual Data

- Physically accessing unstructured textual data
- Terminology
- Languages
- Volume of data
- Differing priorities of unstructured textual data
- Search ability of unstructured textual data
- The economics of the infrastructure
- Security

This is just the short list of challenges that await the organization that attempts to come to grips with the unstructured textual data environment.

The Opportunities of Unstructured Textual Data

■ Customer feedback—Do customers like or hate a new product? A new service ?

■ Warranties—Is there a pattern to the warranties that are presented to a company?

■ Medical information—What medical conditions correlate with other medical conditions? If a subpopulation of subjects is created, how does that change the correlations to medical conditions?

■ Compliance— How has the corporation met reporting obligations?

■ Security — Are employees saying and doing things that are not proper according to the guidelines for corporate conduct?

■ Marketing "buzz " —What is said in the customer community about new products? new services the Company and its activities

■ Competition—what is known about new products new promotions, new services?

■ Human resources—Is the company actually living up to its obligations for hiring practices?

VI. CONCLUSION

With the challenges, there is the promise of opening up an entirely different world of processing and opportunity.

Unstructured textual data by itself contains much useful and interesting information. But unstructured textual data coupled with structured data opens up even more opportunities for unlocking the promise of unstructured textual data.

As powerful as the opportunity for unstructured data in the organization and business is, the technical environment that holds unstructured data must be understood, because many of the opportunities for exploiting unstructured data are shaped by, or certainly influenced by, the technical environment.

REFRENCES

[1]. https://i.ytimg.com/vi/gv5VqYdaWEA/maxresdefault.jpg

[2]. http://1.bp.blogspot.com/-KALuGWZYOrs/UnSHL2kMccI/AAAAAAAAAN8/6H62g6K9boU/s1600/unst.pNG

[3]. http://searchbusinessanalytics.techtarget.com

[4]. https://en.wikipedia.org/wiki/Unstructured_data

[5]. https://www.quora.com