Unstructured Data Mining for Customer Relationship Management: A Survey

Benymol Jose[1], Sajimon Abraham[2]

[1] *School of Computer Sciences, Mahatma Gandhi University, Kottayam, India*

[2] *School of Management and Business Studies, Mahatma Gandhi University, Kottayam, India*

Abstract

These days, the data shortage issue has been supplanted by the data overflow issue. Advertisers and Customer Relationship Management (CRM) specialists have access to rich data on consumer behaviour. The majority of data kept in organizations are in unstructured arrangement including text, sound, video etc. The present test is effective use of these data in CRM procedures and determination of suitable data analytics methods. Data analytics strategies help in finding unseen patterns in data. Many of the requirements for performing these analytics depend on the storage efficiency of the gathered unstructured data. Alternative technologies such as NoSQL have been developed as the answer to the ever-growing data storage requirements of the corporate world. Text mining and Natural Language Processing (NLP) are the two methods with their techniques for information retrieval from textual context in documents. Given the research interest on unstructured data in CRM, the aim of this paper is to perform a survey on the different analysis methods used. Also, a proposal for a new analytics method for unstructured data with audio is included.

Keywords- *CRM, NLP, NoSQL, Text Mining, Unstructured Data*

1. Introduction

Data analytics research has its inceptions during 1970s. Nonetheless, it has encountered a sudden explosion of publications since 2008, essentially, because of enhancement of processing advances. The data analytics literature has been growing over the past few years, drawing in a constant stream of research and journal publications. Today numerous organizations view themselves as market driven as yet sorted out around their items. In the period of quickly changing, globalised economies, and exceptionally competitive markets, change from a product-centric focus to a more customer-centric view is required. Customers expect customised items and administration since they realize that companies have data about them and the open door exists to provide customisation. These days, the capacity to create helpful information from data is fundamental for CRM masters in business intelligence [1].

In CRM space the most widespread areas are related to the management of the contents of clients' email messages and voice messages. This sort of investigation regularly aims at naturally rerouting explicit requests to the suitable administration providers or at supplying prompt answers to the most frequently asked queries [2]. Services research has emerged as a green field area for application of advances in computer science and IT. CRM practices have emerged as hotbeds for utilization of innovations in the areas of knowledge

management, analytics, and data mining. Client reactions in organizations as email and voice messages (transcribed voice) created from an assortment of sources have exploded as far as the sheer volume generated. Organizations are progressively looking to comprehend and analyse this data to infer operational and business insights. The customer, the end shopper of products and services, is getting expanded attention. Analytics and Business Intelligence (BI) applications spinning around the customer has prompted the rise of areas like customer experience management, customer relationship management, and customer service quality. These are getting to be the basic to developmental growth, and sometimes even, survival. Applications with such customer centering are most apparent in services companies, particularly in CRM practices and contact centres.

The main threats in handling this huge volume of unstructured data is the selection of an efficient storage and the development of efficient methods for processing this data. In contrast to Relational Database Management Systems, the focus is on linking the unstructured data with NoSQL databases. The challenges in the storage of the big unstructured data can be managed efficiently with NoSQL databases. Also, text mining techniques along with Natural Language Processing methods can be incorporated to perform the analytics of this unstructured data. This survey gives a perspective on the significance of NoSQL databases, text mining techniques along with natural language processing in the processing of the unstructured data generated in business in CRM context.

The remainder of this paper is organized as follows: Section 2 reviews related work, and Section 3 presents about unstructured data and also the various challenges in performing data analytics with unstructured data is discussed. In section 4, a proposal of a new mining method for unstructured data is made and in the last section, the paper is concluded with findings of the survey and also with the new proposal.

2. Related Work

In one of the past studies, a text mining solution in services industry settings, specifically in contact centres was proposed. The Voice of Customer (VoC) and Customer Satisfaction (C-Sat) analysis settings are done and outlined several unique research challenges brought about by text mining and industrial services research [3].

One of the current studies explores the characteristics of data analytics as the integrated tool in CRM for sales managers. The paper aims at analysing some of the different analytics methods and tools which can be used for continuous improvement of CRM processes. A systematic literature of articles classified by the CRM dimension is conducted to achieve this goal [4].

Another study has been conducted considering the invaluable role played by academic journals, which are in turn unstructured in nature, using text analytic approach to extract relevant information from academic journals to build a structured database which can further be analysed to support decision making. The system uses text analytic approach to bridge the gap between structured and unstructured data and capable of extracting set of entities of interest from unstructured academic journals and store them in a RDBMS for further analysis. Academic journals are a major medium through which research findings are published [5].

In another work, an attempt is made by the author's team using Apache SparkTM, Natural Language Processing (NLP), and text mining, to analyse enterprise Customer Relationship Management (CRM) data to predict and combat premature sales attrition. The customer data are collected from electronic forms (email, web) as well as from transcribed voice conversations. NLP techniques were employed to identify critical patterns in

unstructured CRM data fields for business intelligence. [6].

In another paper, a new methodology and natural language processing approach are used to retrieve data from documents. The method consists in providing connections based on supervised retrieval of domain-specific expressions [7].

In another research work, the unstructured data is structured and processed by using MapReduce technique and the automatic prediction of user's taste is done through collaborative filtering. Map Reduce is the most efficient technique for processing large volume of data and the application of collaborative filtering and sentiment analysis provides recommendation for huge volume of data provided as input [8].

A table showing the comparison of the methods followed by each author is shown below in Table1. Different research works have been tried to apply different data analytics techniques on unstructured data and the details available are summarized.

Table 1: A Survey of Analytics Methods Used in Different Research Papers

Sl No.	Authors	Year of Publishing	Method used for Analytics
1	Shantanu Godbole, Shourya Roy	2008	Text mining/Classification
2	Subramaniyaswamy V, Vijayakumar V, Logesh R and Indragandhi V	2015	Map Reduce
3	Orobor Anderson Ise	2016	Text analytics
4	Matthieu Quantina, Benjamin Hervya, Florent Larochea, Alain Bernarda	2016	Natural Language Processing
5	Pāvels Gončarovs	2017	Classification/Clustering
6	Keith Gutfreund	2017	Natural language processing/Text mining

3. Unstructured Data

Most of the earlier studies of data analytics have concentrated on structured data, for example, relational, transactional, and data warehouse. Because of its variability and unidentifiable internal structure, unstructured data cannot be analysed by the conventional technologies.

Unstructured data come in various organizations and sizes. Extensively, the textual data, sound, video, images, webpages, logs, emails and so forth are considered to be grouped as unstructured data. In some situations, even a bundle of numeric data could be collectively considered as unstructured as in the case of health records of a patient.

Structured data refers to the kind of data that is organized and displayed in a database with rows and columns, making it straightforward to work with. Examples of this include sales figures, names, phone numbers, and pretty much anything that can be categorized. The most commonly used universal type of structured data such as SQL and Access are considered as data sources. The content of structured data can be organized according to the data types and data is searchable [9]. Unstructured data refers to usually computerized

information that either does not have a data model nor has one that is not easily usable by a computer program [10]. Unstructured data distinguishes such information from data stored in fielded form in databases or annotated in documents [9].

Because of the expansion in the use of internet for social media channels, new analytics tools and processes were developed to understand and extract value from this huge volume of unstructured data. The value of unstructured data originates from the patterns and the meanings that can be obtained from it often incorporates recognizing issues, market patterns, or generally customer notion towards a brand.

3.1. Vital Issues in Processing Unstructured Data

With the review led through different research papers, it is clear that there exist several conceptual points that should be understood by the organizations to execute the technology adequately. The diverse issues identified with the attributes of this data are volume, velocity, variety and complexity. The storage and transport issues happen mostly because of two reasons. The quantity of data has exploded and also the data generated is in unstructured form. The information generated by the web is not in the structured format [11] and cannot be formed into the two-dimensional structure which is followed by the relational databases.

The data management issues**,** maybe, the most troublesome problem to address with big data. Settling issues of access, usage, updating, administration, and reference (in publications) have turned out to be major stumbling blocks. The sources of the data are varied - by size, by arrangement, and by strategy of collection. People contribute computerised data in mediums agreeable to them like documents, drawings, pictures, sound and video recordings, models, programming practices, user interface designs and so on, with or without sufficient metadata depicting what, when, where, who, why and how it was gathered and its provenance.

The genuine processing issues can be understood by thinking about an exabyte of data, which should be handled in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, as 1 Exabyte = 1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end to- end processing would be 20 nanoseconds. To process 1K petabytes would require an aggregate processing time of roughly 635 years. In this way, execution of Exabyte of data will require broad parallel processing and new analytics algorithms so as to give provide timely and actionable results [12].

Companies are processing the emails and voice messages of customers to make predictions about a product or the overall status of business. Even though several studies and research works have happened in this area with Text Mining and Natural Language Processing as mentioned in the related studies, the main issue faced by companies in nowadays is the storage of this unstructured data. In most of the recent research, Apache Spark, Hadoop and map reduce techniques are often used to manage it.

4. Proposed System

In Customer Relations Management, NLP techniques with Text Mining could be used to perform data mining and analytics with the big unstructured data and also interesting and useful patterns can be discovered for Business Intelligence. Text Mining is about extracting useful information from the available data. Information could be patterns in text or matching structure but the semantics in the text is not considered. The goal is not about making the system understand what does the text conveys, rather about providing information to the user based on a certain step by step process. Natural language is what humans use for communication. Processing such a data is NLP. The data could be speech or text. Thus, the main goal is towards understanding what is the semantic meaning conveyed in it. As defined Natural Language Processing concerns to the

automatic processing and analysis of unstructured textual information. Natural languages have lot of complexities as a text extracted from different sources don't have identical words or abbreviation. There is a need to detect such issues and make rules for their uniform identification [13].

Also, the proposal uses the NoSQL databases to handle the storage issues related with unstructured data. NoSQL databases were designed to handle ambiguity [14]. The NoSQL databases are schema free unlike relational databases which have pre-requirement to form the schema before storing the data in database. Instead NoSQL databases can store unstructured data making them appropriate tool for storage of big data. They can replicate and partition data over many servers with a simple call level interface. Most NoSQL data stores do not enforce any structural constraints on the data. The different types of NoSQL databases in this category typically use one of four basic types of data models and the choice of database will ultimately depend on the data being stored as well as the demands of the project for which it is being used. The four types of NoSQL databases include key value stores, document stores, column oriented and graph based.

The different steps involved in the proposed study are given below.

4.1. Steps in the Proposed Analytics Method

i). Consider a CRM dataset consisting of customer feedback regarding a product.

ii). Search this data for patterns like "I am having a problem with feature X" and then enhanced to it "I am having difficulty with feature X".

iii). For text mining, similarity search has to be performed and for this a software regular expression can be employed for pattern matching.

iv). After applying these regular expressions against the text field in all records, we can count the occurrences of the different values of X and display the most frequently occurring values.

v). For similarly search with audio, audio analytics is to be done.

vi). The search in the email text can be made more efficient by making the text shorter and by selecting the only related emails by looking at the subject part.

vii). Also, prepare a dictionary of related words and consider the relevant part of the email to reduce response time.

viii). For storing this unstructured big data, a NoSQL database can be used.

ix). With these points, using an interface, develop a new procedure with text mining algorithms and natural language techniques to extract the relevant parts of the text in less time.

In Customer relations management, Natural language Processing Techniques with text mining could be used with NoSQL storages to mine the unstructured textual data and discover interesting and useful patterns for business intelligence.

Conclusion

Businesses are anticipated to become digitized generally. Each digital practice produces data. The tremendous volume of unstructured data delivered should be analysed to extract important information for better business execution. Text mining strategies along with NLP techniques are utilized to investigate the interesting and relevant information successfully and effectively from this extensive measure of data. This paper presents a survey of different analytics methods used in different researches in the past with unstructured data. Choice and utilization of right strategies and tools according to the domain help to make the analytics procedure simple and

productive. Domain knowledge integration, varying concepts granularity, multilingual text refinement, and natural language processing ambiguity are the real issues and challenges that emerge during this procedure of analysing unstructured data. Additionally, a proposal for another new mining method with text mining techniques, NLP and NoSQL databases is also included.

References

[1] Econsultancy and Adobe, Digital Trends in the Financial Services and Insurance Sector, 2016.

[2] Bolasco S., Canzonetti A., Capo F.M., Della Ratta-Rinaldi F., Singh B.K., Understanding Text Mining: A Pragmatic Approach, In: Sirmakessis S. (eds) Knowledge Mining. Studies in Fuzziness and Soft Computing, vol 185, 2005, 31-50.

[3] Shantanu Godbole, Shourya Roy, Text to Intelligence: Building and Deploying a Text Mining Solution in the Services Industry for Customer Satisfaction Analysis, IEEE International Conference on Services Computing, USA, 2008, 441-448.

[4] Pāvels Gončarovs, Data Analytics in CRM Processes: A Literature Review, *Information Technology and Management Science*, (*20*), 2017, 103-108.

[5] Orobor Anderson Ise, Integration and Analysis of Unstructured Data for Decision Making: Text Analytics Approach, *International Journal of Open Information Technologies,4(10),*2016,82-88.

[6] Keith Gutfreund, Big Data Techniques for Predictive Business Intelligence, *Journal of Advanced Management Science, 5(2),* 2017, 158-163.

[7] Matthieu Quantin, Benjamin Hervya, Florent Larochea, Alain Bernarda, Supervised Process of Un-structured Data Analysis for Knowledge Chaining, Procedia CIRP, 26th CIRP Design Conference, France, 2016, 1-6.

[8] Subramaniyaswamy V, Vijayakumar V, Logesh R, Indragandhi V, Unstructured Data Analysis on Big Data using Map Reduce, Procedia Computer Science 50, 2nd International Symposium on Big Data and Cloud Computing, India, 2015, 456-465.

[9] Y. Shiqun, W. Gang, Q. Yuhui and Z. Weiqun, Research and Implement of Classification Algorithm on Web Text Mining, IEEE Explore, Third International Conference on Semantics Knowledge and Grid, China, 2007, 446-449.

[10] Y. Shiqun, Q. Yuhui, G. Jike and W. Fang,2008, A Chinese Text Classification Approach Based on Semantic Web, Fourth International Conference on Semantics Knowledge and Grid, 2008, 497-498.

[11] S.S Nyati, S. Pawar, R. Ingle, Performance evaluation of unstructured NoSQL data over distributed framework, IEEE International Conference Advances in computing communications and informatics, Mysore, India, 2013, 1623-1627.

[12]. Lenka Venkata Satyanarayana, A Survey on Challenges and Advantages in Big Data, *International Journal of Computer Science and Technology, 6(2),* 2015 ,115-119.

[13] Calvillo, E. A., Padilla, A., Munoz, J., Ponce, J., Fernandez, J. T., Searching research papers using clustering and text mining, International Conference on Electronics, Communications and Computing, IEEE, Choluva, 2013, 78-81.

[14] R. Cattell, Scalable SQL and NoSQL Data Stores, *ACM SIGMOD Record*, *39(4),* 2011, 12-27