International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Data Mining from Unstructured Documents

Rajalakshmi Thiruthuraipondi Natarajan

rajalan11@gmail.com

Abstract

Data Mining is the process of identifying and extracting valuable data by scanning through large volumes of structured and unstructured data, which would form the base for further processing using data analytics tools for cleansing, categorization and organization, etc. This source data might not fit to a certain template and can be of any format ranging from plan test to media files and it is the responsibility of the mining process to understand the message, extract relevant information and finally convert to a standard format. Prior to its final stage, these data undergo several rounds to cleansing to eliminate irrelevant information and pick the right set of data intended by the organization with the best turnaround time possible. At each stage of the analysis, the data needs to gets cleaner and distinctive and provide a vision as to the areas it will be used.

This document provides insight on data mining and its potential impact in market. This explores the various sources and the type of data that might be associated with it and how to cleanse and various ways the information can be used for the development of a retail business. This also provides guidance on the patten recognition and the proper compartmentalization of the data so that it is readily available to the target groups for research and marketing.

Keywords: Data Mining, Data Analytics, Data Cleansing, Data sorting, Data Compartmentalization, Data Organization, Processing Raw and Unstructured Data, Pattern Recognition

Introduction

Data Mining is the process of identifying and extracting relevant data from a large volume of largely unstructured data fed through various streams for a pre-determined goal. In the old days, it was producers who, predominantly set the trend, and the people bought into it inspired or attracted by the marketing and the customer feedback was collected either in the form of company-initiated surveys or complaints to understand the mood and take the respective path. There were extremely limited sources where the general masses voiced their thoughts, and the organizations gathered their information. With social media's foray, the dynamics are quickly changing, where the customers express their desire and need, and the companies need to adopt accordingly. Unlike pre-social media era, customersfeedback are pouring in various forms and outlets, such as online reviews, customer ratings, social media influencers and the reactions are immediate. Such information is extremely valuable for any industry to understand the market and the customer pulse, and to steer in the right direction and the industry. With facilities such as online shopping and digital marketplace, it has become extremely easy to market the product making every home a potential competitor. To stay competitive, every industry have made significant investment in collecting the data from various sources and various format and use it to plan their next

course of action. With data available in various formats such as text, media and even emojis, there various in-house and third-party organizations, who have developed solutions using complex algorithms to read the data and process it, and the ability of the retail companies to monetize this is one of the key factors for its survival and growth.

Overview of Data Mining

There can be easily compared to the conventional mining for metals and various similarities can be found in each step of these areas. Similar to Gold mining, where the location of potential nerve is first identified, and then big chunks of rocks are broken down to smaller and smaller pieces until the precious metal is found. Once found, the nerve is tracked to find the flow and systematically drilled to retrieve the pieces, which is further cleansed and carried over for further refining and general purpose. Data Mining pretty much uses the same strategy in identifying, patterning and retrieve the needed information. Hence this can rightly be considered to be both a process by itself and an activity in a larger flow. Data Mining involves several steps, and the raw data gets more refined in each step before it can finally be used. As in every process, there are certain pre-requisites that the team need to consider and be prepared, before starting the data mining activity, such as,

Define the Goals

First and foremost, the reason for data mining need to be defined. The data, structured or unstructured, can have a lot of information within it. To find and extract the relevant data, one must know what is being looked for. These are usually business decisions taken by the leads of the respective teams. Considering the long-term and short-term goals, this can either be for a specific division or corporate wide. These decisions are either to venture in a new direction or analyze the existing status of the company so that educated, data-driven decisions can be made that will best suit their vision. These goals form the very basis of data mining. Depending on what is being looked for, the data miners can plug in conditions to the tools to look for data fitting this requirement and isolate it. It is important to note that in the same set of source data, there might be something for everyone and depending on who is looking and for what, the relevancy would change.

Identify the Source

With the goal defined, the next step is to identify, where should one go about looking for relevant information. Unlike pre-defined interfaces from third parties, which are structured, and interface points are well defined, the data source for data mining can be anything under the sun. Every customer or supplier and even onlooker can be a potential contributor to these data sources. Hence it is extremely important to choose where the miners are searching for the data. If incorrect sources are chosen, the source data might not be accurate or outright invalid, that could lead to the companies taking wrong decisions affecting their growth. Similarly, it is practically impossible for any system, no matter how big, to scan through every system or entry out there to gather information. It is an extremely expensive and time-consuming endeavor as a lot of information might be completely irrelevant and the organization would have wasted valuable time and resources in scanning through these. It can also be counterproductive, since infusing a lot of vague data might dilute the outcome resulting in inconclusive outcome.

Layer and Prioritize

Though every source might have something to contribute to the goal, it is important to understand that not every source can have the same weightage. Everyone might have their own opinion and suggestion of how to move forward, however, most of these inputs might be based on guess work or from a very narrow vision. For instance, in stock market discussion forums, there will be several investors who share their thoughts, ideas and predictions on how the stock would behave, but most of these comments are either guess work or based off of few sets of factors. However, there are certain users or forums, who have a deep understand of the stock and the general market and share their thoughts, which are more likely to happen. Similarly, even after the relevant information is gathered, it needs to be ordered based on the importance to the goal. Every piece of the product can be reviewed and commented, however, the data miners should not lose focus by considering all information with the same priority. For example, while reviewing a product, the comments can be provided for the product, such as, the look and feel, ease of use, availability, reliability, to relatively trivial factors such as the packaging, instruction manual, etc. The positive and negative feedback given for each of these areas can be given the same score, i.e., though all these comments are relevant to the product, the feedback on the color and material of the packaging cannot be provided the same relevance as product flaws.Though it need not be completely discarded, such data need to be categorized in a different bucket.

Data Mining Tools

Choosing the right Data Mining tool is another critical piece in the preparation process. There are several tools in the market to perform data mining activity with various capacities. Based on the organization's requirement, and the source from where the unstructured documents are scanned, the right tool needs to be chosen for the mining activity. Choosing an incorrect tool might lead to the inability of the tool to read through and understand the source data, thereby ignoring the valuable information available within the source or worse, incorrect the translate the information. The team deciding the tool should have a detailed knowledge of the source, goal along with other requirements, and decide if the data mining solution is compatible with the source format and can generate output in the desired structure. The cost factor and the reliability of the tools need to be considered too, since there needs to be a balance to ensure that the benefits outweigh the cost.

Data Mining Techniques and Steps

As mentioned earlier, Data Mining is both an activity and a process, where the unstructured data goes through several steps before the final outcome. Each steps takes the data one step closer to the goal of converting them into a structured format that can be easily for further research and reporting activities. It is critical that each step perform its task correctly, since the subsequent steps rely on the output of the earlier stage to perform their task and any rouge logic or uncertainty in any stage can lead to incorrect results.

Common Format Conversion

The first step that the unstructured documents undergo is the conversion of the source data into a common format. Based on the selected sources, the data within these sources can be of any format, such as simple text, pdf files, web pages, or media files, such as, images, video and audio files each with a

different format. These different data formats needs to be transformed into common format that is a *lingua franca*for all systems that needs to use the data. Though recommended, the target format need not be a single data format but should be one of the formats that the next stages can use. In addition to the file formats, the language used can also vary from person to person, like some like to provide long feedback, others, especially the new generation, communicate heavily using abbreviations, jargons, emojis, stickers, etc. Using the social knowledge, there needs to a library of such details maintained to translate it into the target format.

Data Cleansing

Though the source data has been converted to a readable format by the earlier step, it is still unstructured. It is now just heaps of data with no apparent pattern and it is in this stage that the data is rid of all the extra attachments which were needed for it work in its original format but have no data relevance. These add-ons are removed from the heap leaving behind only the raw data.

For the cleaning process to differentiate the actual data and add-ons, it is important to know the source format from which it was converted. For example, if the converted data is from a web page, which was converted to a HTML or XML text, then anything within the tag symbols “<>” and “</>” are the add-ons and entry between these two tags can be the actual data. Additionally, there are header and footer tags that aid the format rather than provide any information, however, if it was converted from any other format, these symbols might contribute to the information looked for. Using this knowledge, such parts can be safely eliminated there by finally providing with just the raw data. This acts as a common pool which would be used by various teams to mine the data based on their needs and goals,

Data Extraction

Now that the source data is converted into a common data format and cleansed of any part that is not actually data, it is time to scan through this raw data to identify the relevant data and extract the same. This is the core piece of the entire process where the source data is loosely fir into a certain structure. Using the common pool of data generated after data cleansing, the relevant data is retrieved based on each team’s requirement. As stated earlier, with in the raw data, the relevancy can differ between team and the goals. There are several techniques using with the extract is done. Some of the techniques are as follows -

Keyword Search is one of the common and simple ways to search for relevant data. Based on the requirement, a set of key words which relate or represent these goals identified along with the commonly used grammar or language surrounding it. The raw data is scanned to look for these key words and track it from there. Using the common language and grammar patterns, the data surrounding these key words are identified and compose it to a draft structure.

Another technique is the pattern recognition. The first way is identifying the pattern by regression scanning, where the system picks up a certain combination and looks for similar combination across the raw data and recognizes it as a pattern and looks for similar combinations to identify the relevancy. The other way is the knowledge of general standard and historical documents based on which the pattern is pre-assumed. While this loosely might sound to be structured, there are significant caveats that still would make it unstructured. For example, the quarterly financial statements published by the public companies usually have a certain format that almost all companies use. Also, each company will

have its own personalization to these statements which they tend to stick on to. Using this knowledge, patterns can be picked up fed into future raw data to indicate where to look specifically, thereby easing and speeding up the process.

Challenges and Limitations

While the above section covers some of the key steps and techniques of data mining, there are several nuances that make data mining a niche skill. Since there is no control as to how the unstructured data can be, there are several challenges that the miners need to over come or avoid to accurately retrieve relevant data.

The main challenge is the data by itself. Since unstructured data can be provide by anyone, it is a huge challenge to actually recognize the validity of the data. Though eventually a pattern might emerge, the truthfulness or the critically is hard to recognize. With users getting influenced by the online influencers and other theorists, it is really hard to know if the relevant information is in general or form few sources which other have carried over. Also, there could be more emphasis on a certain area which might produce whole different pattern focusing on non-trivial areas.

Though the technology has come a long way, with AI and machine learning assisting in data mining and related activities significantly, there are still some significant gaps that needs to be taken care of. With the tone and way of expression constantly changing it makes it extremely difficult for data mining tools to accurately recognize the relevant information or the attributes surrounding it. Also, processing such large volumes of data would require high-capacity servers and resources to browse through and get the information which can be a costly affair.

Conclusion

The need to collect data and use them for one own benefit has been growing at an exponential rate. It has crossed the lines of organizational scope and has expanded to analyze global and political trends to plan one's business. With the growth of AI and ML, data analytics has seen capabilities that was never possible before. Yet, with the constant shift in the way the information is produced, such as new jargons, vernaculars or slangs, the algorithms, tools, and the methods need to continuously adopt with the new trends. Though the data driven decision making is the right way, it still has not eliminated the guess work and expert's knowledge, since the current technical capability cannot capture the factors such as sentiment and customer emotions. However, with constant research and redefining, the data mining can be a great boon for critical decision making.

References

[1].Whatisdataanalytics?,https://aws.amazon.com/what-is/data-analytics/#:~:text=Data%20analytics%20converts%20raw%20data,making%2C%20and%20foster%20business%20growth.
[2]. Unstructured Text in Data Mining: Unlocking Insights in Document Processing, https://www.shaip.com/blog/data-mining-unstructured-text-for-insights-in-document-processing/, Sep-2023
[3]. Bella Williams, How to Extract Valuable Insights from Unstructured Data with Text Data Mining, https://insight7.io/how-to-extract-valuable-insights-from-unstructured-data-with-text-data-mining/

[4]. The Basics Of Data Mining, https://www.datasource.ai/en/data-science-articles/the-role-of-ai-in-unstructured-data-mining-challenges-and-opportunities
[5]. New Mining Techniques Effortlessly Provide Greater Insight for Unstructured Text Data, https://provost.gatech.edu/news/new-mining-techniques-effortlessly-provide-greater-insight-unstructured-text-data, Aug-2019
[6]. What is text mining?https://www.ibm.com/think/topics/text-mining
[7]. Mining Unstructured Text – Database, https://docs.oracle.com/en/database/oracle/oracle-database/18/dmapi/mining-unstructured-text.html