NOU J. Sc. & Tech. Vol. 3 & 4 (I) : 57-61, 2015

Big Data and Natural Language Processing

Hima Bindu Maringanti

P.G. Department of Computer Science & Applications,
North Orissa University, Baripada-757003 Odisha
Corresponding Author: E-mail : profhbnou2012@gmail.com

Abstract

Natural Language Processing (NLP) is quite old compared to Big Data, an explosive bomb in the field of technology, encompassing each and every domain of science and arts. NLP is a specialized domain of Artificial Intelligence, wherein the natural language understanding and generation capabilities of humans are attempted to be accomplished by a Computer System. Though research in this direction is on since a couple of decades, not substantial has been achieved so far, albeit some IVR systems and text to speech transcription in medical domains. Big Data is the availability of vast data in the cyber space and its mining and extraction of useful data for specific needs. The present paper is an attempt in integrating NLP and Big Data and correlating the principles behind each of them and the technological requirements, tools and architectures for making it possible, so that a better interpretation of the huge data is achieved.

Keywords: NLP, Big data, AI, Architecture, Cloud

Introduction

Big Data is the huge quantity of data available on the Internet; you could search for anything on this earth on the Google Search Engine and get answers. Never (rarely) does a query give a blank answer. What does it mean? It means you have a variety of data and information, not all of which is useful to you. Also, there may be some information not available explicitly, but hidden, which needs extra inference & conclusions by the seeker. The various features of big data are the 7 Vs: volume, veracity, value, variability, velocity, variety, visualization. Putting them together, big data is defined as voluminous variety of ever changing, valuable and viewable data. It could simply be the data of an organization, including its features and used to meet the organizational goals and objectives. Rijmenam , the father of this concept, the big data, predicts the doubling of data , every two years. Managing, including storing this unstructured data, pre-processing viz.,, combing, smoothing and cleaning, filtering, mining and understanding are the issues with this growing data. The different components of big data, w.r.t. a business model are a) Data Sources: operational and functional systems, machine logs and sensors, Web and Social Networks b) Data Platforms, Warehouses and Discovery Platforms:

that enable the capture and management of data, and then its critical conversion to suit to customer insights , which are finally put into action c) Big Data Analytics Tools and Apps: the "front end" used by executives, analysts, managers and others to access customer insights, models, scenarios and manage the business. The evident role of big data in the present scenario is the available tens of TeraBytes of marketing database containing customer preferences and habits.

Natural Language Processing is a specialized domain of Artificial Intelligence; an ability of natural language understanding (NLU) and generation (NLG) by a Computer System. The various phases of NLP include Lexical, Syntactic, Semantic and sometimes Pragmatic Analyses to understand the spoken or written language. While it is relatively easier to understand written language, spoken language processing includes complexities like prosody, intonation and other signal based features. The general limitations on full-fledged NLP constitute intentions, goals, emotions, incomplete sentences/phrases, ellipsis, inter-relationship between the pair involved in conversation- the speaker and listener etc. Most of the times, the time of utterance also plays a major role, for conversations are in chunks, mostly connected/correlated to the previous-time utterance; commonly called as Discourse and its resolution as discourse analysis, one of which is Anaphora resolution- pronoun resolution.

This paper is organized as the current tools and technologies in big data and NLP, followed by the author's contributory philosophy, a journey towards the correlation between these two independent fields, one is about data and its management and the other is an understanding mechanism, that is analytical. Finally, the author concludes and discusses this upcoming domain and ends with references.

State of the Art

In [1], the author states how big data is becoming vital in realizing the overall business goals and strategic objectives of major business stalwarts like Amazon & Walmart, even NASA & US government. He also sees no reason as to why big data cannot be used by small and medium size organizations for their own benefit.

In [2], the author coins, "The three V's" - volume, velocity and variety, to refer to the challenge of data management. He briefly puts big data as s a lot of data produced very quickly in many different forms. This could involve customer transactional histories, production databases, web traffic logs, online videos, social media interactions, to name a few.

In the blog post [3], the author added "veracity, variability, visualization, and value" to the definition, broadening the realm even further. He stated that, 90% of all data ever created, was created in the past two years and predicted that the amount of data in the world will double every two years.

In the article [4], the author highlights the advantages of big data as competitive, data on the board's agenda and driving innovative products and startups. These features show how

technology can drive the business, and vice versa.

In Internet basics [5], the author puts across statistics about the growth of data or data explosion as follows: In a way, big data is exactly what it sounds like — a lot of data. It's been estimated that in all the time leading up to the year 2003, only 5 exabytes of data were generated — that's equal to 5 billion gigabytes. But from 2003 to 2012, the amount reached around 2.7 zettabytes (or 2,700 exabytes, or 2.7 trillion gigabytes) [sources: Intel, Lund]. According to Berkeley researchers, roughly 5 quintillion bytes (or around 4.3 exabytes) of data every two days [source: Romanov] is being produced. Defining the term 'big data' as massive, rapidly expanding, varied and often unstructured sets of digitized data that is difficult to maintain using traditional databases, the author reminds that you are producing data every time you do anything online, leaving a digital trail that others can come along and mine for useful information. Now, the next section deals with bringing a correlation between the so called Big Data and Natural Language Processing.

Novelty and Contribution

Initially, when you start defining Big Data, if parsing is done to identify the Part of Speech (POS) of this double word phrase, there arises a conflict or you agree to tag it with both a Noun and a Verb, a Noun by its very nature and intuitive; a verb by its functional necessity- without which, just a huge , unstructured data would be useless. So, the big data when combed, smoothed and or cleaned is only valid and hence a Verb.

According to Berkeley researchers, we are now producing roughly 5 quintillion bytes (or around 4.3 exabytes) of data every two days [source: Romanov]. Not only is this data from Social Networking and Instant Messaging(IM) systems, but RFID tags, Wireless sensors, Surveillance Cameras, Wi-Fi, GPS, Traffic Sensors, Vital Statistics Sensors, Smartphones, uploading and downloading of photos, audio/video, multimedia sharing also are proliferating the Cyberspace to put in their data. Each of us when browse the Internet, leave a digital trail, giving a scope for analysis, enabling one to design profile-based display systems. Also there seems to be need for analysis of this massive, but unstructured data.

Mining and shuffling the huge data for possible new information, Patterns-known and unknown, Rules, Semantic Processing to understand and further, Pragmatic Analysis to get the intended meaning are some of the immediate analyses of big data. Sometimes, Emotions also play an important role while expressing during a conversation in Blogs, Facebook, WhatsApp, Twitter session. Clustering this huge unstructured data to mine patterns, to understand similarities and dissimilarities, correlations etc. is going to be a herculean task. As there are varied manners of conversation style, including emoticons/smileys and SMS-incomplete and mis-spelt words like dat for that, u for you, wud for would etc., which a human after some exposure can learn and understand, but if the system is expected to capture these literary nuances, it is going to be a challenging

task. Sometimes, to understand a simple sentence having partial meaning, a complete session of dialogue between people or even a past session may have to be considered. Not only that, the relationship between two people also influences the words and gestures used. The temperament and mood of a person is also conveyed in these immediate messaging systems, where once intentions and emotions that need to be quickly conveyed, are shared. In a customer feedback recorded system and its analysis, some key words or phrases will be used by the service provider's algorithm to assess the level of customer satisfaction and also predict the possibility of his/her retaining in future. The other major utility of NLP that surfaces is because of the need for Querying the huge data available and also a dire necessity of making it accessible by all, requiring a Natural Language Interface, where the queries are in English and not SQL; so no need of knowledge of SQL-like query language.

Tools and Technologies

The technological requirements for such a task are parallel architectures, distributed systems and centralized processing. Special, Cloud based frameworks have also come up, which need to be fine tuned and customized to individual requirements. The best example of natural language processing, on the market today is *Microsoft's Power Query,* which uses Excel Services and *PowerPivot along with Office 365* to allow users to just type in natural language a limited set of queries about data. For example, in a search box, the user can *type "nearest medical shop in Baripada"* and it would return a map of Baripada with the Chemist highlighted in the specific location. Real time big data processing requires integration of *Transactional, Analytical, Operational and Archival Systems*; so the need for an efficient *Parallel Architecture* with its varied facets and a robust integration mechanism. Efficient storage, metadata identification, indexing, archiving and linking are the preliminary requisites, in addition to effective Querying, resulting in acceptable *recall and precision measures.*

In Artificial Intelligence, the two commonly used reasoning mechanisms are forward and backward, but in Big data analysis, it is *Digital Reasoning*, which provides advanced semantic analysis and the *Hadoop ecosystem* is used to scale over hundreds of millions of documents and billions of entities, facts and relationships. *Oracle Big Data Appliance* is an enterprise-class engineered system to provide an optimized and complete solution for Big Data workloads. The appliance contains 18 Sun servers with a raw storage capacity of 648TB. Each server contains two 6-core Intel© Xeon© CPUs and 48GB of memory. Oracle Big Data Appliance runs Cloudera's Distribution including *Apache Hadoop* (CDH). CDH provides the #1 Hadoop-based distribution in commercial and non-commercial environments, akin to Means-Ends-Analysis in Heuristics. Hortonworks, MapR and Amazon, h HPCC and cloud-based Google BigQuery, NETFLIX/YouTube as examples for video streaming, Apache

Spark for Real time analytics, STORM, a Distributed & Scalable Architecture and Text to Audio Transcription, Speech processing for emotions, Prosody, Mood etc. are the future tools and technologies aiding Big Data analysis.

Applications

Various practical and research applications of Big data and NLP integration are Sentiment Analysis from Facebook, Twitter sessions, Opinion Mining from blogs, online rating etc., Profile Tracking / Web-Behaviour analysis, Remote Tracking and Diagnosis after medical transcription Audio data collection, Customer Satisfaction and Prediction of Client Behavior in future, to mention a few. Unlike the Closed World Assumption in AI, which assumes that whichever fact exists in the Knowledgebase, is true and that which is presently not in the base is false, the big data analysis includes mining facts that are not present and assumes as important latent/hidden/unknown patterns which it tries to discover; a kind of knowledge discovery.

Conclusion

As human language is heterogeneous, with varied nuances like emotions, intentions, sentiments, opinions, gestures, styles and in the Internet of Things, there is a flooding of massive, unstructured data, which needs combing, smoothing, clustering and finally mining of known and unknown/unexpected data, the two concepts are closely correlated for the larger benefit of the Business Enterprise. Thus the Natural Language Processing and Big Data are integrated with new storage, indexing, retrieval mechanisms and parallel architectures, which requires mathematical and statistical expertise, creative, communicative, problem-solving and business skills.

References

[1] Scott Matteson, **Big Data basic concepts and benefits explained, Big Data Analytics,** September 25, 2013.

[2] Doug Laney, 2001 PDF

[3] Mark van Rijmenam , Why The 3V's Are Not Sufficient To Describe Big Data, Blog Post, August, 2013.

[4] Parry Malm, "**Three reasons why Big Data is awesome", Econsultancy.com.**

[5] Bernadette Johnson, What is 'big data'?, Internet Basics