Big Data: Information Retrieval, Extraction and Mining

Mehta Krupa , Dr. Devarshi Mehta , Dr. Vishal Dahiya

[1]GLS University, Ahmedabad, Gujarat, India [2]Indus University, Gujarat, India

ABSTRACT

The Internet has a huge collection of information. With the advent of big data, it is becoming easy to store any amount of information. As the name says "big data" is used to store the information with three characteristics i.e. volume, variety and velocity. The stored information is useful only when it is extracted properly at the item of requirement with accuracy. To extract the information from data source(s) is known as data mining. Accurately mined data produces the results that are directly applicable and useful to implement in real scenario. Such extracted information empowers the analytical capability which facilitates decision making process by highlighting the minute and crucial issues. The intended information is retrieved from various available data sources. This allows preparing a data source having related information and this data source becomes the base for further extraction of the information.

INTRODUCTION

Looking at current scenario of WWW, a large amount of information is generated every day. According to a survey, in 2012, Google received over 2 million search queries per minute. Fast forward to 2014 and that number has more than doubled. In January 2014, Google receives over 4 million search queries per minute from the 2.4 billion strong global internet population [7]. Considering usage of smart phones and latest technology, this number is going to increase enormously. We are living in the era where huge information is available but still we are striving for knowledge. To store such huge amount of information and to retrieve this information is becoming very crucial challenge. The information is of no use, if there is no mechanism to access relevant information with limited time span. To fulfil the need of extracting only relevant information, Information Extraction techniques are used. This paper presents an overview of Information Extraction from Big Data. To make the flow of topic logical, the paper covers overview of Information Extraction and Information Retrieval, How Data mining can be helpful to extract information, overview of big data, mining a big data, challenges in mining big data.

INFORMATION RETRIEVAL

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Information retrieval is about returning the information that is relevant for a specific query or field of interest. Note that this information could also be in the form of general documents, sure enough search engines are a notable example of such task. The most important entities recognizable for information retrieval are the initial set of documents/information and the query that specify "what to search for". IR is used to select from a collection of textual documents a subset which is relevant to a particular query. It generally returns ranked list of documents. IE and IR techniques complement each other. Information Extraction requires to filter required documents from huge collection [10]. After retrieval of related documents, required information is extracted from that available data source.

DATA MINING FOR INFORMATION EXTRACTION

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. Information extraction is more about extracting (or inferring) general knowledge (or relations) from a set of documents or information.

IE is to identify a predefined set of concepts in a specific domain, ignoring other irrelevant information, where a domain consists of a corpus of texts together with a clearly specified information need. It is deriving structured factual information from unstructured test [10]. IE is typically seen as a one-time process for the extraction of a particular kind of relationships of interest from a document collection [8].

Data mining is core and most challenging step in information extraction. Typically, data mining uncovers interesting patterns and relationships hidden in a large volume of raw data, and the results tapped out may help make valuable predictions or future observations in the real world [1].

As the name says data mining techniques are used to mine the data from a bulk of data. Data mining techniques examines the data and tries to find out some pattern or relation among them, a pattern is identified, using this pattern required information is extracted in a particular format by developing an algorithm. Looking at current scenario of information generation, data mining is becoming most important. Nowadays millions of MB of data is generated on internet. And with the availability of mobile internet and latest technology, the number is going to constantly increasing. Moreover, the information generated is unstructured. The need to store such large amount of data, initiated the word "Big Data".

HANDLING BIG DATA AND ITS ISSUES

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress". Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [4]. However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold[4]. As the name says the term big data represents the tremendous amount of information generated currently on internet.

The data generated are of two types [4]:

1. Structured data
2. Unstructured data

Structured data includes numbers and word that can be easily extracted and analysed. Such kind of data can be generated by smart phones, electric gadgets, etc. Such data are easy to capture and process.

Unstructured data includes reviews, feedbacks, videos, audio clips, photographs, E-mails etc. Such data are difficult to categorize and analyse. Unstructured scattered data collections are full of information.

To extract information, a pattern is required. Based on the derived pattern the data can be categorized and analysed. But if the data doesn't follow any pattern it is hard to extract information from it. To extract some information from news paper, a word or set of words can be identified. Later on to extract information, the document can be verified again identified word(s). But to extract information from a photograph of the same news paper is not that much easy. *For Example: If a news item explaining flood situation is to be analysed from the news paper, some keywords can be coined along with the news writing pattern. By this analysing some key parameters from that news story like: location, amount of rainfall, number of persons affected, etc is possible but through the photograph explaining flood situation, it is hard to extract such key parameters.*

The data generated by internet is mainly falls into second category of data i.e. Unstructured data. Big data deals with this large volume, heterogeneous data. Every individual is contributing his/her bit in this large collection without any central control over the production of data. This decentralization of data makes it more complex to handle as it is not possible to maintain any fix pattern which can later on help in identifying pattern and analysing. Looking at the large collection of data and its complexity, three V's are associated with the Big Data:

- Volume: The amount of data. It is core characteristics defined by the word big data itself. It refers to the quantity of data which is of prime focus.?
- Variety: This characteristic refers to the difference in nature of data and data sources. It is the major reason behind the complexity of the big data. As all the data from various sources are stored in big data, there is no particular pattern is followed which can help in mining process.?
- Velocity: It refers to the speed at which information is generated and added to the big data.?

From the above discussion it is clear that we are living in flood of information, but still we strive for relevant information. Big data is used to store a large amount of data from different sources and of different nature which is constantly increasing at high rate. Day by day it is becoming very important to extract some useful relevant information from the big data. As discussed, data mining techniques are used to extract information. Big data is a tool to store large amount of information and data mining is a technique to generate meaningful data from it. A combination of both the techniques enhances the user experience and helps to manage information properly which can be accessed as and when needed in no time. Some issues with big data mining are:

- Too much of data, how to relate it??

  Big data storage is going beyond zettabyte. To find relation between such huge amounts of information is very critical task. To make information access specific and fast, such kind of relation generation is very important. To find relevance between information, relation generation is key step.

- Does not follow any specific pattern:?
  Information generation is decentralized; there is no single control over the information generation. It means that information generated from various sources follows different process which makes the task if pattern generation very complex. It may be possible that generated pattern works for some information while some important information may leave out.
- Which Information to extract:?
  The decision of which information to be extracted is important. When dealing with large amount of data, it becomes very important to decide that which information is of vital importance and which one is not. It is also important to carry out the work for garbage collection, to keep the information up to date and remove unnecessary information.
- Pace at which information is generated?
  Day by day, the rate at which information stored in big data is increasing rapidly. It is difficult to manage this flood of information. It requires high scalability of mining techniques.
- Privacy?

Privacy is considered as supreme issue since data mining has begun. With the use of social media, lot many personal information can be mined like location, activity, friendship. The use of online payment is also becoming very common; mining applied to such data may cause serious issues, if not handled properly. As the Internet user's increases, their personal information is also stored somewhere around, to maintain such private information secret also becomes very important.

REFERENCES

Dunren Che , Mejdl Safran , and Zhiyong Peng. Mining Big Data: Current Status, and Forecast to the Future

Andrew Kehler, Jerry R. Hobbs, Douglas Appelt, John Bear, Matthew Caywood, David Israel,

Megumi Kameyama, David Martin, and Claire Monteleoni. Information Extraction Research And Applications: Current Progress And Future Directions,

Wei Fan,Albert Bifet. Mining Big Data: Current Status, and Forecast to the Future

Bharti Thakur, Manish Mann. International Journal of Advanced Research in Computer Science and Software Engineering, 2014

Ah-Hwee Tan. Text Mining: The state of the art and the challenges

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding. Data Mining with Big Data. 2014

http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/

Luis Tari, Phan Huy Tu, Jorg Hakenberg, Yi Chen,Tran Cao Son, Graciela Gonzalez, and Chitta Baral. Incremental Information Extraction Using Relational Databases, 2012.

Wei Fan, Albert Bifet. Mining Big Data: Current Status, and Forecast to the Future.

J. Piskorski, R. Yangarber. Information Extraction: Past, Present and Future.

❑