An Introduction to Clustering Algorithms in Big Data

Rajit Nair

https://orcid.org/0000-0002-4564-0920

Jagran Lakecity University, India

Amit Bhagat

Maulana Azad National Institute of India, India

INTRODUCTION

In Big Data, clustering is the process through which analysis is performed. Since the data is big, so it is very difficult to perform clustering approach. Big data is mainly termed as petabytes and zeta bytes of data and high computation cost is needed for the implementation of clusters. In this chapter we will show how clustering can be performed on Big Data and what are the different types of clustering approach. The challenge during clustering approach is to find observations within the time limit. Chapter also covers the possible future path for more advanced clustering algorithms. The chapter will cover single machine clustering and multiple machines clustering which also includes parallel clustering.

Today, data is increasing rapidly which actually forms a big data due to it high velocity, huge volume and different varieties of data (Shobana & Kumar, 2015). Few years before we are dealing with data collection challenges, but now in this era we are more concerned with processing this huge amount of data or big data (Nair & Bhagat, 2018a). Big data is analyzed for prediction and analysis purpose (Tsai, Lai, Chao, & Vasilakos, 2015). Scientists and researchers believe that today one of the most important topics in computer science is Big Data, whether the data collected through social networking websites, ecommerce websites or any other websites which is having accessibility to billions of users. You tube has more than 1 billion user which is producing almost 200 hours of video in each hour and its content ID service scans over 400 years of video every day. Main reason behind storing these types of data is further used for knowledge discovery (Chu, 2013). Data mining is the method which is mainly used for knowledge discovery and clustering is the process which is used for dividing the data into groups that contain objects of similar patterns (Chu, 2013). Clustering is vastly used in many areas such as bio-informatics (Y. Q. Zhang & Rajapakse, 2008), machine learning (Kubat, 2017), pattern recognition (Neal, 2007), networking (Sucasas et al., 2016) and lot of research work is done in this area. A decade before clustering algorithm was applied on small data in order to handle their complexity and computational cost, this also increase the speed with scalability. Due to occurrence of Big data in recent years more challenges are added to perform clustering.

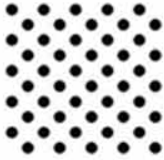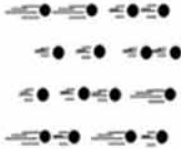## Big Data Challenges With its Important Characteristics

Big data is the phenomena which mainly work on 5 v's (Saporito, 2013), which is volume, velocity, variety, value and veracity, but it is not limited to 5 v's it can be extended further also. Normally Big data is unstructured, this means data cannot be easily transformed in the form of rows and columns and this

cannot be processed by traditional databases. So to process this Big Data, NoSql database (Lourenço, Cabral, Carreiro, Vieira, & Bernardino, 2015) are used. Let's discuss these V's one by one.

- Volume – This term is directly related to space occupied during data storage, whenever there is talk about big data, volume is the main aspect. Today most of the application are based on big data like data collected by Facebook which is having more than 2 billion users, data collected by traffic sensors, data collected by aero plane during transfer travel from one place to another, data collected by ecommerce website, data collected by Youtube having one billion users etc,. The amount of data collected is more than terabytes can be considered as big data but there is no limit defined to which we can say the data is big data.
- Velocity – This term is directly related to speed of incoming data which means when we are uploading the data at some specific server, in that case it may happen that many people are accessing that server at a time, so it experience a high speed. In case of facebook, 900 million photos are uploaded, 500 million tweets posted on twitter, more than 3.5 billion searches performed in google, 0.4 million hours videos are uploaded on Youtube. The incoming speed is just like nuclear explosion, so Big Data processing help us to deal with this high velocity of incoming data.
- Variety – Many of the website applications has to process and analyze different varieties of data like text, image, video, etc. and it is not an easy task to process different varieties of data. Facebook is the best example where we deal with different varieties because different users used to post textual data, images and videos at the same time, so facebook has to handle these varieties of data and later on processing should be done. To process this Big Data plays an important role.
- Veracity – It relates to trustworthiness of the data means the data which is collected that is reliable or not even it can be said that it is accurate or not. Veracity is very much needed during processing of big data because if we process the inaccurate data, in that case the analysis will also be inaccurate.
- Value – Finally the last V i.e. value which actually transforms the data in the form of value. In the complete chapter we will discuss the challenges, methods and applications of clustering in detail with respect to big data. It's mainly related to cost or profit included in the system in terms of value.

Figure 1. 5 V's of Big data

Volume	Velocity	Variety	Veracity	Value
Stationary data	**Moveable data**	**Varieties of Data**	**Suspicious data**	**Transformation**
It's a stationary data ranged from	Streaming of data that requires	Consist of structured and	Uncertainty occurs due to incomplete,	Data is converted into money for the

Related Content

A Protocol for Evaluating Mobile Applications
Clare Martin, Derek Floodand Rachel Harrison (2013). *Information Systems Research and Exploring Social Artifacts: Approaches and Methodologies (pp. 398-414).*
www.irma-international.org/chapter/protocol-evaluating-mobile-applications/70726

Collaborative Design: An SSM-enabled Organizational Learning Approach
Anita Mirijamdotterand Mary M. Somerville (2009). *International Journal of Information Technologies and Systems Approach (pp. 48-69).*
www.irma-international.org/article/collaborative-design-ssm-enabled-organizational/2546

Meta Data based Conceptualization and Temporal Semantics in Hybrid Recommender
M. Venu Gopalachariand Porika Sammulal (2017). *International Journal of Rough Sets and Data Analysis (pp. 48-65).*
www.irma-international.org/article/meta-data-based-conceptualization-and-temporal-semantics-in-hybrid-recommender/186858

Exploring the Integration of User-Generated Content in Media Organizations Through Participatory Journalism
Theodora A. Saridouand Andreas Veglis (2021). *Encyclopedia of Information Science and Technology, Fifth Edition (pp. 1152-1163).*
www.irma-international.org/chapter/exploring-the-integration-of-user-generated-content-in-media-organizations-through-participatory-journalism/260257

Philosophical Framing and Its Impact on Research
Eileen M. Trauthand Lee B. Erickson (2012). *Research Methodologies, Innovations and Philosophies in Software Systems Engineering and Information Systems (pp. 1-17).*
www.irma-international.org/chapter/philosophical-framing-its-impact-research/63255