Bibliometrics analysis on big data research (2000- 2021) in the Scopus and Science Direct Database

Arun Dabas, *UIET, MDU*
Dr. Chhavi Rana, *UIET, MDU*

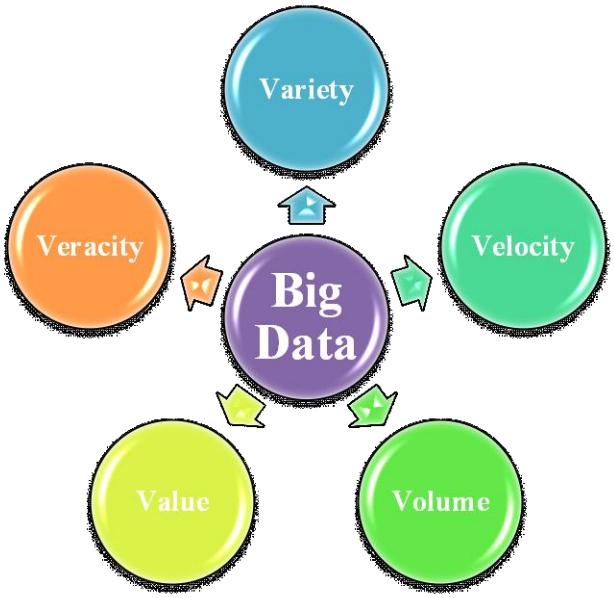
Abstract

In today's era, science and technology is spreading rapidly and mostly work is done with the help of technology. After the arrival of concept of big data, human life face drastic changes and mostly work is completed in less time and without much effort. This paper use bibliometric analysis and data analysis methods for analyze the big data publication. It analyze the most used keywords, journals, citation, influential counter research etc. of the database of Scopus and Science direct.

Introduction

Big Data is a term that is in very use to describe lots of data of any field or resource. Now days there are lots of data which is generating every second in every area of work and research or in normal life. Pandemic is a great example of this because due to pandemic situation mostly human being was self quarantine and working from home even than school/ college classes was online. So this can be assume that every second there was millions of data stored or downloaded by us to work or study without any complication and hurdle. To justify the amount of data, a term is used "Big Data". This "Big Data" term is not new but this is used frequently from last five to ten years. To define the data amount or volume or variety in simple form, term- Big Data is used. In 2005, firstly Roger Mougalas from O'Reilly Media think up the term "Big Data" [1]. It cited to a large amount of data that is very hard to process and cope using traditional methods and tools. But, the term "Big" can't justify that how large the data is. As much world is connecting the amount of big data is boosting rapidly. New developments of advanced facilities for communication like phone, social media sites or other online websites and other has producing big data very vastly. Big data mechanization narrates a new era of technologies and architectures which help to analysis volume, velocity and verity etc.

Big data's characteristics are value, volume, variety, veracity and velocity. To analyze the opportunities and challenges of data's growth mainly three characteristics are used which are known as 3 Vs which are volume, velocity and variety.



Vs of Big Data

Literature Review

Sr. No.	Name of Paper	Review	Year of Publication	Name of Authors
1	Theme Mapping and Bibliometrics Analysis of one Decade of Big Data Research in the Scopus Database	This paper presents an overview of big data research analysis of decade of journal publication from the Scopus database. Both bibliometric and text mining analysis were conducted on database.	2020	Anne Parlina, Kalamullah Ramli and Hendri Mufri
2	Global Trends on food Security Research: A Bibliometrics Analysis	This paper shows issue of food security in international community and to revel the research situation about food security. This paper utilizes bibliometrics techniques like VOSviewer, Citspace and bibliometrix to explore the research and development trends in the area of food security	2021	Hualin Xie, Yuyang Wen, Yongrok Choi and Xinmin Zhang
	Software tools for conducting	This paper aims to provide an up to date review of various tools available for	2020	Jose A. Moral-Munoz, Enrique

3	bibliometrics analysis in science: An up-to-date review	conducting bibliometric and scientometric analysis. And this paper's main conclusion is that every analyzed software tools have their advantages and disadvantages, So user needs to choose proper software for specific analysis.		Herrera-Viedma, Antonio Santisteban-Espejo, Manuel J. Cobo
4	A bibliometrics analysis on big data research(2009-2018)	This paper uses bibliometrics and visual analysis methods to proper study and analyze the big data publication included in the Science Citation Index and Social Science Citation Index databases. It analyzes most influential countries, journals and research institutions.	2019	Zeshui Xu, Dejian Yu
5	From conventional group decision making to large-scale group decision making: What are the challenges and how to meet them in big data era? A state-of-the-art survey	In this paper, bibliometric analysis is used to study short-out the challenges and present state of art survey of main achievements. This paper tells users can choose BD tool to perform their research in this field.	2019	Ming Tang, Huchang Liao
6	Discovering and forecasting interactions in big data research: A learning- enhanced bibliometric study	This paper uses a learning-enhanced bibliometric study to discover interactions in big data research by detecting and visualizing its evolutionary pathways. Concentrating on a set of 5840 articles derived from Web of Science covering the period between 2000 and 2015, text mining and bibliometric techniques are combined to profile the hotspots in big data research and its core constituents	2018	Yi Zhang, Ying Huang, Alan L. Porter, Guangquan Zhang, Jie Lu
7	A bibliometrics approach to tracking big data research trends	This papers main purpose was to find out the international research trends in big data field and which area is more relevant for research. this study's main result provides a formula to analysis	2017	Ali Kalantari, Amirrudin Kamsin

		citation from multi-regression analysis.		
8	Bibliometrics analysis of publications on wine tourism in the databases Scopus and WoS	Main purpose of this study is to show the current scientific state of research regarding wine tourism by comparing WoS and Scopus database with bibliometric anylsis.	2016	Amador Duran Sanchez, Maria de La Cruz Del Rio Rama, Jose Alvarez Garcia

Description

To analyze the database according to our choice and preference we need a tool which is helpful for more accuracy and to reduce the time work and effort and Bibliometric is the best tool regarding this. Because Bibliometric tool is used to analyze the data in systematic way and it tells the detailed information about the given database like citation, co-citation, author, published year, liked keywords, mostly used keywords etc. Now days, it becomes an important tool for evaluate and analyze the outputs of researchers, partnerships between universities, research centers. To use bibliometrics tool, a researcher only need a proper data and mind-set that which type of output is required. It is very helpful for general analysis, mapping analysis and meta analysis. Bibliometrics has ample set of ways to conduct analysis like VOSviewer, BibExcel, CitSpace, Netdraw, Pajek, Sci2, Pop and so on. Form them; VOSviewer is most popular now days, because this tool has remarkable visualization and has capacity to load or export as much as information researcher wants to use. To check the working of Vosviewer tool, user needs to mark the data for the output and find out the option which will be needed and fit for the aimed and accurate output.

In order to anticipate the ultra-modern availability of tools to perform bibliometric analysis, some process will be followed:

1. Download the data from a proper database for analysis
2. Download the bibliometric tool or software for performing data analysis and visualization
3. Perform the analysis and find out the desired output.

With the help of VOSviewer user can find the network visualization, overlay visualization, density visualization, clusters, keyword citation, co author citation etc. To clarify this, an example of Science Direct database is taken. Firstly 100 research papers data with keyword bigdata and machine learning is taken and export in RIS file. Then visualization and analysis process is performed with the help of VOSviewer. Output of this process is attached in figures.

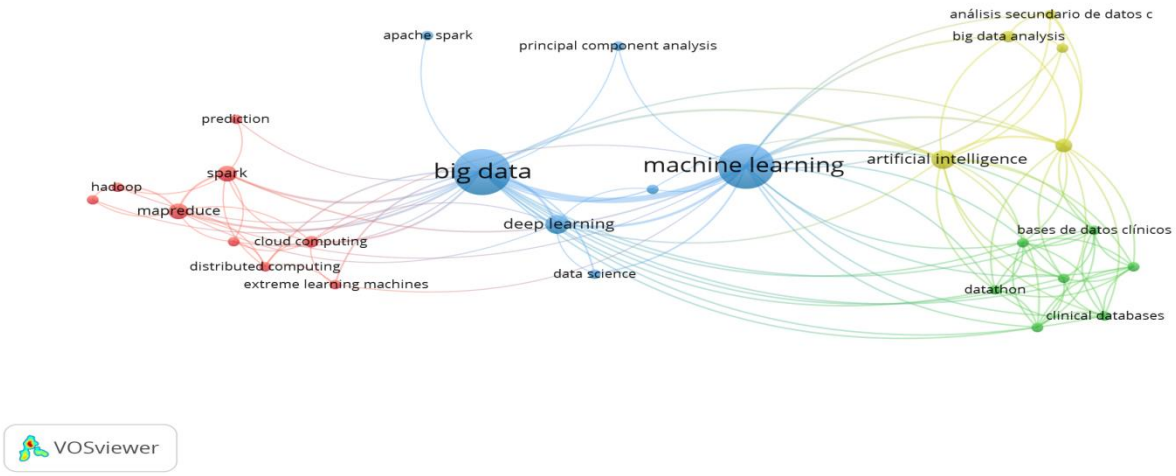


Fig- 1- Network Visualization of Co-occurrence of keywords of Database

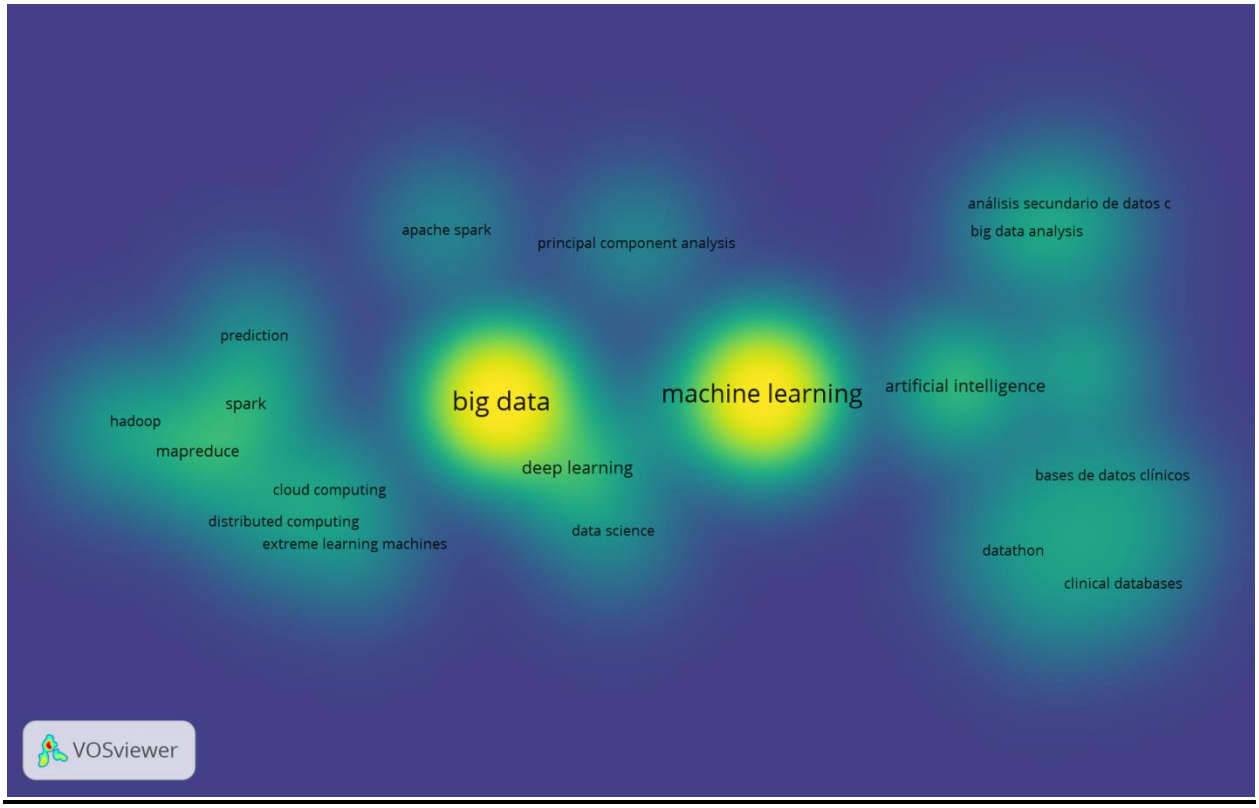


Fig -2 - Item Density Visualization of Co-occurrence of keywords of Given Database

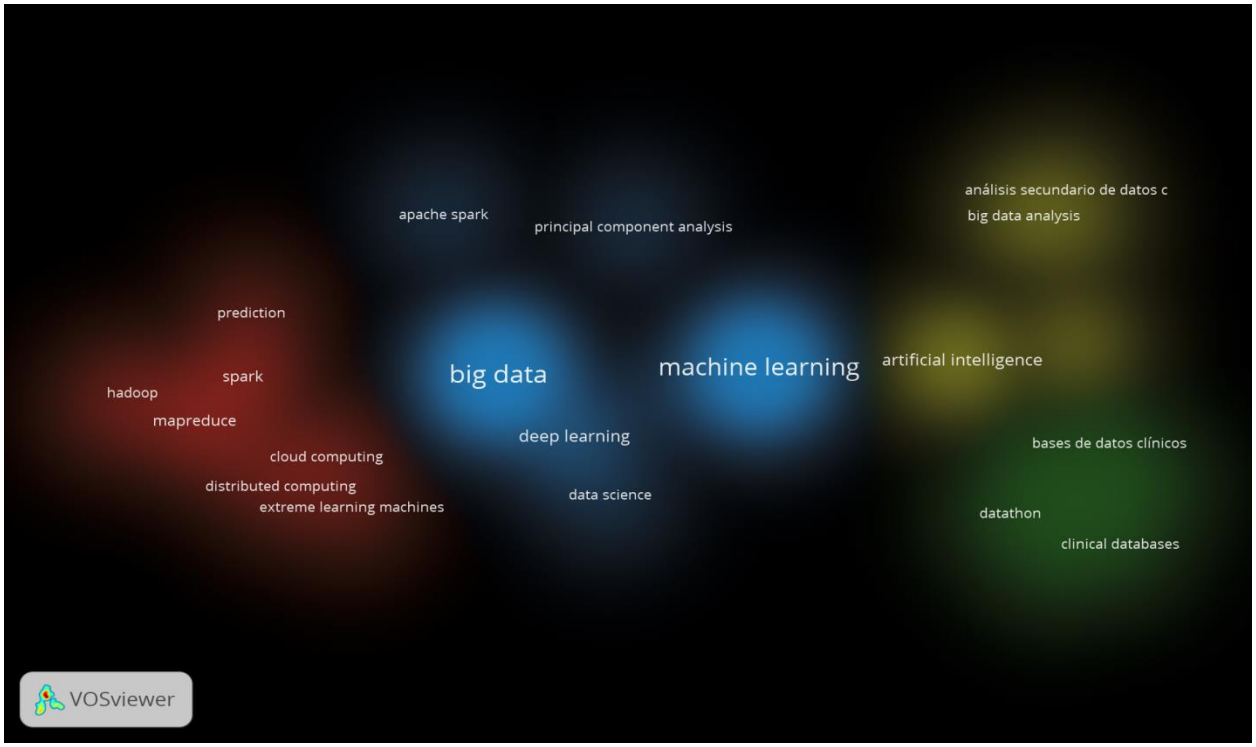


Fig -3 - Cluster Density Visualization of Co-occurrence of keywords of Given Database

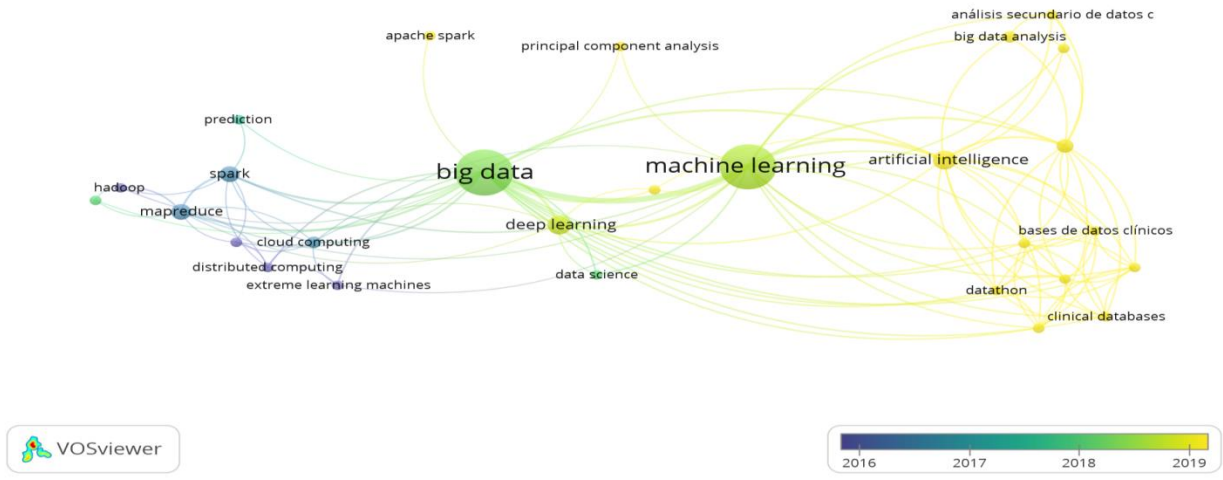


Fig- 4 - Overlay Visualization of Co-occurrence of keywords of Given Database

Cluster Table

Clusters of Keywords	
Cluster-1	Artificial Neural Network
	Cloud Computing
	Distributed computing
	Extreme learning machine
	Extreme learning machines
	hadoop
	map reduce
	prediction
	Spark
Cluster-2	Bases de datos clinicos
	Clinical databases
	Collaborative work
	datathon
	Mimic iii
	Trabajo calaborativo
Cluster-3	Apache Spark
	Big data
	Data science
	Deep learning
	Healthcare
	Machine learning
	Principal component analysis
Cluster-4	análisis secundario de datos clínicas electrónica
	Artificial Intelligence
	Big data analysis
	Inteligencia artificial
	Secondary electronic helath record data analysis

Critical Analysis

As in above figures (1,2,3 and 4), it is clearly shown that bibliometric analysis is very easy on big data with the help of VOSViewer. Because with the help of vosviewer, user can find out the co-occurrence of keywords and co-authorship of authors. User can analysis in any form like density visualization, overlay visualization and network visualization. Item and cluster formation

is very easy and user can check which keyword is linked with which keyword and how many times a paper was citied and in which year how many paper was published on the give database. User can check which keyword is popular in which year and mostly research is done on that keyword or not. Even then user can find out the year to year analysis with the given data base that in which year how many paper was published and in which year's paper was most cited then other etc.

Conclusion

This paper builds a concerned framework meld with data analysis and bibliometrics to find out the big data research analysis from 2000 to 2021. Specially the keywords, journals, citation etc. Scopus and Science direct database is used to extend the amount of data for research and analysis. This study tells about the research on big data is influenced by big data analytics like methods, application and supportive infrastructure. It is very difficult to recognize that big data always congregate with analytics.

References

1. *Dr. Mark Van Rijmenam, M. A Short History of Big Data. 2019. Available online: A Short History Of Big Data (datafloq.com) (accessed on 12 February 2021)*
2. *Anne Parlina, Kalamullah Ramli and Hendri Mufri . Theme Mapping and Bibliometrics Analysis of one Decade of Big Data Research in the Scopus Database. 2020.*
3. *Hualin Xie, Yuyang Wen, Yongrok Choi and Xinmin Zhang. Global Trends on food Security Research: A Bibliometrics Analysis. 2021*
4. *Jose A. Moral-Munoz, Enrique Herrera-Viedma, Antonio Santisteban-Espejo, Manuel J. Cobo. Software tools for conducting bibliometrics analysis in science: An up-to-date review. 2020.*
5. *Zeshui Xu, Dejian Yu. A bibliometrics analysis on big data research (2009-2018). 2019.*
6. *Ming Tang, Huchang Liao. From conventional group decision making to large-scale group decision making: What are the challenges and how to meet them in big data era? A state-of-the-art survey. 2019.*
7. *Yi Zhang, Ying Huang, Alan L. Porter, Guangquan Zhang, Jie Lu. Discovering and forecasting interactions in big data research: A learning- enhanced bibliometric study. 2018.*
8. *Ali Kalantari, Amirrudin Kamsin. A bibliometrics approach to tracking big data research trends. 2017.*
9. *Amador Duran Sanchez, Maria de La Cruz Del Rio Rama, Jose Alvarez Garcia. Bibliometrics analysis of publications on wine tourism in the databases Scopus and WoS. 2016.*
10. *Hofmann T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval; 1999.*

11. *Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003.*
12. *Blei DM, Lafferty JD. A correlated topic model of science. Ann Appl Stat 2007.*
13. *Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical dirichlet processes. J Am Stat Assoc 2006.*
14. *Lu K, Wolfram D. Measuring author research relatedness: a comparison of word-based, topic-based, and author cocitation approaches. J Am Soc Inf Sci Technol 2012.*
15. *Nichols LG. A topic model approach to measuring interdisciplinarity at the National Science Foundation. Scientometrics 2014.*
16. *Yau CK, Porter A, Newman N, Suominen A. Clustering scientific documents with topic modeling. Scientometrics 2014.*
17. *Hornik K, Grün B. Topicmodels: an R package for fitting topic models. J Stat Softw 2011*
18. *Meyer D, Hornik K, Feinerer I. Text mining infrastructure in R. J Stat Softw 2008.*
19. *Yao Q, Chen K, Yao L, Lyu PH, Yang TA, Luo F, Chen SQ, He ZY, Liu ZY. Scien-tometric trends and knowledge maps of global health systems research.Health Res Policy Syst 2014*
20. *Boyack KW, Klavans R, Börner K. Mapping the backbone of science. Sciento-metrics 2005*
21. *Takeda Y, Mae S, Kajikawa Y, Matsushima K. Nanobiotechnology as an emer-ging research domain from nanotechnology: a bibliometric approach. Scien-tometrics 2009.*
22. *Wang H, Liu M, Hong S, Zhuang Y. A historical review and bibliometric ana-lysis of GPS research from 1991–2010. Scientometrics 2013.*
23. *Kulkarni SS, Apte UM, Evangelopoulos NE. The use of latent semantic analysis in operations management research. Decis Sci 2014.*
24. *Adler MD, Johnson KB. Quantifying the literature of computer-aided instruc-tion in medical education. Acad Med 2000.*
25. *Hsieh WH, Chiu WT, Lee YS, Ho YS. Bibliometric analysis of patent ductus arteriosus treatments. Scientometrics 2004*
26. *Meneghini R, Packer AL. Is there science beyond English? EMBO Rep 2007.*
27. *Porter MF. An algorithm for suffix stripping. Program: Electron Libr Inf Syst 1980.*
28. *Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM 1975.*
29. *Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Doc 2004.*
30. *Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci USA 2004.*
31. *Mann HB. Nonparametric tests against trendc. Econom: J Econom Soc 1945.*
32. *Guney MS. Evaluation and measures to increase performance coefficient of hydrokinetic turbines. Renew Sustain Energy Rev 2011.*
33. *Coglianese C. Information technology and regulatory policy new directions for digital government research. Soc Sci Comput Rev 2004.*

34. *Shulman SW, Schlosberg D, Zavestoski S, Courard-Hauri D. Electronic rule-making a public participation research agenda for the social sciences. Soc Sci Comput Rev 2003.*
35. *Levy KEC, Franklin M. Driving regulation: using topic models to examine political contention in the U.S. trucking industry. Soc Sci Comput Rev 2013.*
36. *Bail CA. The cultural environment: measuring culture with big data. Theory Soc 2014.*
37. *Hofmann T. Unsupervised learning by probabilistic latent semantic analysis.Mach Learn 2001.*