International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

Emerging Trends and Technologies in Big Data Analytics: Applications in Libraries

Dr. J.Vivekavardhan
Asst. Professor
Department of Library and Information Science,
University College of Arts and Social Sciences,
Osmania University, Hyderabad.

Abstract:

The paper explores on the concept of Big Data, Big Data Definition, Big Data Retrieval Use Case Diagram, Big Data Characteristics, Structure of Big Data, Different types of search strategies like keyword search, Boolean operator search, phrase search, proximity search, Truncation search, file format search, image search, domain search etc. are explained with relevant examples.

The paper also through some light on Big Data algorithms such as PageRank, K-means, Apriori, Expectation Maximization, AdaBoost, K-Nearest Neighbors, Naïve Bayes, Classification and Regression Trees, Support Vector Machines, Collaborative filtering, Recommendation Engine, segmentation, Gaussian processes, Logistic Regression, Linear Regression, Artificial Neural Networks, Dimensionality Reduction, RandomForest etc. Paper finally presents Big Data Analytics, Big Data Applications in Libraries, findings, conclusion and suggestions for further research on Big Data.

Keywords: Big Data, Search Strategies, Big Data Algorithms, Big Data Applications in Libraries

Introduction

Big data changed life in a number of ways. The greatest part about Big Data phenomenon is the rapid change in innovation and new discoveries. World Wide Web accommodates millions of websites, billions of web pages and tons of data, is growing exponentially. Big data generally refers to data is too big to fit in main memory. Big Data means bigger in size, just how big is big? Very large data, Extreme data, total data etc. Big Data is about data volume measured in terms of Peta bytes. Big data exceeds the typical storage, processing and computing capacity of conventional databases and data analysis techniques. Big data is the idea that computers can gather trillions of pieces of information about billions of different things and find useful patterns in that information.

According to smartinsights.com (2017) explores that around the world in every minute 29 Million WhatsApp messages are sending one another, 3.3 Million FaceBook Posts, 3.8 Million Google Searches, 65,972 Instagram Photos uploading, 500 hours video uploading in YouTube, 1,49,513 E-Mail messages, 4,48,800 Twitter Tweets are sharing around the world in every minute. Walmart handles more than 1 million customer transaction every hour, which are imported into databases estimated to contain more than 2.5 Petabytes (2560 terabytes) of data, this is equivalent of 167 times the information contained in all the books in the United States of Library of Congress. The concept of Big Data can be applied in all disciplines as an information resource. The emergence of Big Data the librarian's role is changing as data librarianship, data curators. There is a need of Librarian's to fit into the new reality of Big Data research.

Big Data Definition(s)

According to oxford dictionaries.com Big Data is "Extremely large data sets that may be analysed computationally to reveal patterns, trends and associations, especially relating to human behavior and interactions".

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

According to webopedia.com Big Data is a Massive volume of both structured and unstructured data that is too big it is difficult to process using traditional database and software techniques.

Big Data Analytics:

Big Data Analytics means it is the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Big Data Analytics is useful to find out the Intelligent Decisions to improve operations.

Objectives of the study

1. To find out the Big Data Trends and Technologies
2. To find out the different types of Big Data Search Strategies
3. To explore the different types of Big Data Algorithms
4. To find out the Big Data analytics and applications in Libraries**.**

Significance of the study

Big Data have become an indispensable tool in our everyday life. This paper helps to assess the user what type of search strategies use while searching the Big Data to retrieve relevant and exact information from the web. Big Data Algorithms, Big Data Analytics and Applications.

Methodology

The study is based on extensive review of literature available in the print journals, online journals on internet about Big Data technology, Big Data Analytics.

Limitations of the Study

Big Data available globally but the present paper is confined to the Big Data Characteristics, Big Data Use case diagram, different types of Big Data search strategies, Big Data Algorithms; paper concludes the Big Data Analytics and Applications in Library and Information science.

Big Data Characteristics

Big Data characteristics are Volume, Velocity, Value, Veracity, and Variety. The 5Vs of Big Data are as shown in table (1).

Sl No	Big Data Characteristics	Description
1	Volume	Quantity of Data: Sheer amount of data being created. Terabytes, Records, Transactions, Tables, Files etc.
2	Velocity	Speed of processing Data: Speed with which data is being created. Near time, Real Time, Streams, Batches, etc.
3	Value	Data Value: Statistical, correlations, hypothetical, events, etc
4	Veracity	Trustworthiness of the data : Authenticity, origin, reputation, accountability, etc.
5	Variety	Categorizing the Data: Different types of data gathered. Structured, Semi-structured, Unstructured, Mixed data, etc. Text, Audio, Video etc.

Table (1) Big Data Characteristics

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

Big Data Retrieval Use Case Diagram

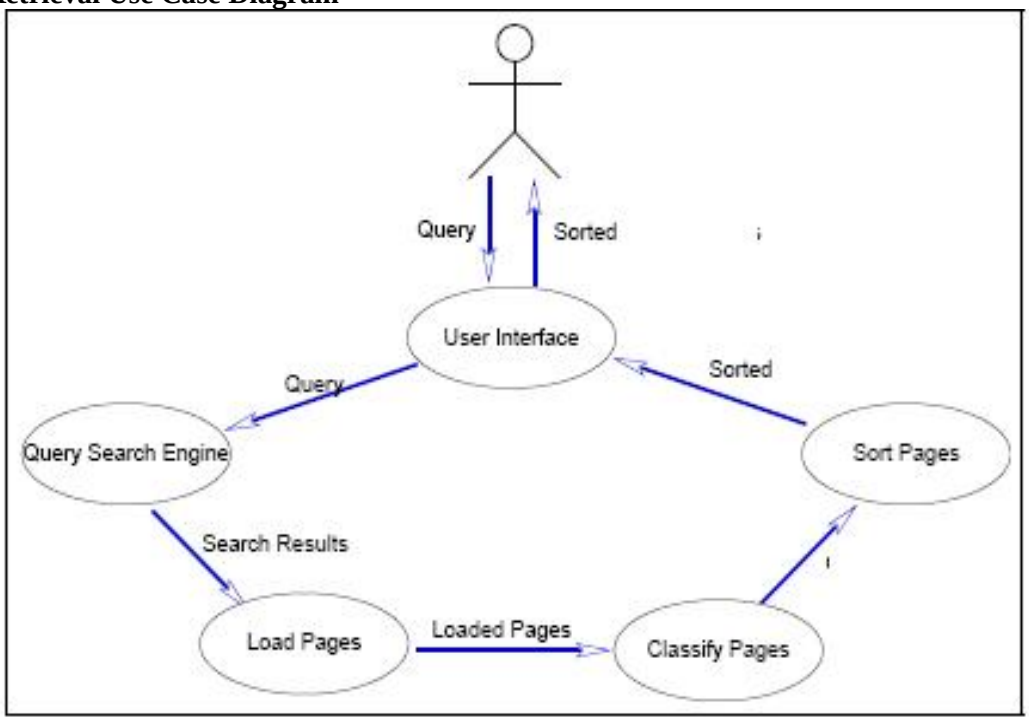

Figure (1): Big Data Retrieval Use Case Diagram (source: www.google.com)

The above figure Big Data Retrieval Use Case Diagram explains that user sends the query to the Big Data index servers. Web page contains the keywords that match the query. The query travels to the doc servers, which actually retrieve the stored documents. User gives query; it searches the search results, Load pages, Classify pages, sort pages. The search results are returned to the user in a fraction of a second. Big Data automatically generated by machines.

Structure of Big Data:

The structure of Big Data is three types, namely structure, semi-structured, and un-structured as follows:

Sl. No	Structure of Big Data	Explanation/Example
1	Structured	Most Traditional Data Sources: Ex. Data Ware Housing, Demand Forecasting in Manufacturing, Predictive Maintenance in Aerospace, Fraud Detection, etc. .
2	Semi-structured	Many sources of Big Data; Text Mining, Social Media, Sentiment Analysis.
3	Unstructured	This data cannot be easily indexed. Video Data, Audio Data, Image Data etc.

Table (2) Structure of Big Data

Big Data Search Strategies

World Wide Web has become an indispensable source of information for any one. Big Data Search Strategies are as follows in the table 2.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

Sl No	Big Data Search Strategies	Description / Use of Search Strategies	Example
1	AND (+ plus sign)	It Narrows search.	Software AND Hardware
2	OR	Broadens search.	Software OR Hardware
3	NOT (- sign)	Contain one keyword exclude the other keyword.	Software NOT Hardware
4	Nesting ( ) Parentheses	Utilizes parentheses to clarify relationships between search terms.	(Data OR Information) AND (Knowledge)
5	Proximity Search	Search for two or more words that occur within a specified number of words of each other in the database.	Cloud Computing Technology retrieves records containing three words immediately adjacent to one another and in the same order.
6	NEAR	Find words within 10 words of each other. Near is the same as within 10.	Knowledge near discovery retrieves records that contain knowledge and discovery in any order and within a 10 word radius of one other.
7	BEFORE	Find words in a relative order, specified with the before expression	Data before Mining
8	AFTER	Find words in a relative order specified with the after expression.	Information after science.
9	Phrase Search	Retrieve search terms next to each other in the order user typed.	“Big Data Analytics”
10	* Truncation.	Expands a search term to include all forms of a root word.	patent* retrieves patent, patents, patentable, patented, etc.
11	Multi Charcter Wild Card *	Multi-character wildcard for finding alternative spellings.	behavi*r retrieves behaviour or behavior
13	Stop words	Stop words are ignored	a, and, the
14	File Format Search	Users can limit their search to any specific file format.	MicrosoftWord (.doc), Adobe Pdf (.pdf), Microsoft Excel (.xls), Text Format(.txt) etc.
15	Site/Domain	Limit to domain search	.com /.gov/ .edu / .org
16	Language	Search can be limited by language.	English, Hindi
17	Spelling Check	Mistake in spelling then system asks ‘did you mean this’.	Libray Scince Did you mean Library Science

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

18	Weather	"weather" along with city and country.	weather Delhi, India
19	Calculator	Evaluate Mathematical Expressions	$(28+12)^9$
20	Images	Relevant images	National Library of India
21	News Headlines	Latest News and stories	IFLA Director 2017 Year
22	Time	current time & city name	Time Hyderabad
23	Sports scores	scores and schedules for sports teams	Cricket score
24	Numeric Ranges	using a double dot between range numbers	70..80
25	Dictionary Lookup	"define" followed by a colon and the word(s) to look up	Define: Operating System
26	Maps	related maps can be displayed	Hyderabad: Map
27	Patent numbers	"patent" followed by the patent number	Patent 5123123
28	Google Goggles	Google app	Google app photos
29	Similar terms.	Use the "~" symbol to return similar terms.	~plane, also searches for aircraft, flight, jet, etc.
30	Search web pages with a specific domain extension	Search by domain with in education sector websites (.edu), or Government (.gov), or information (.info), or commercial (.com) etc.	(.edu) (.gov) (.info) (.com)

Table 2: Big Data Search Strategies (source: www.google.com)

Big Data Search Algorithms

Big Data is driving radical changes in traditional data analysis platforms and algorithms. Big Data Algorithms are PageRank, K-means, Apriori, Expectation Maximization, AdaBoost, K-Nearest Neighbors, Naïve Bayes, Classification and Regression Trees, Support Vector Machines, Collaborative filtering, Recommendation Engine, segmentation, Gaussian processes, Logistic Regression, Linear Regression, Artificial Neural Networks, Dimensionality Reduction, RandomForest etc. Google search engine uses PageRank algorithm to search Big Data. Google's Web Search Algorithm was one of the first tools to show the possibilities offered by Big Data.

PageRank Algorithm

PageRank is a link analysis algorithm used by the Google search engine to retrieve Big Data. PageRank is a vote by all other Web Pages.

$$PR(A)= (1-d) + d (PR(T1)/C(T1)+…….+ PR (Tn)/C(Tn)$$

PageRank (PR) for a page A, assume there are pages T1 to Tn that link to page A. PR is the PageRank of any given page and C is the count of outgoing links. The PageRank of every page linking to page A is divided by the number of outgoing links. The PageRank agent works by acting as a Random Surfer that follows links around the internet randomly. This is to stop the agent getting stuck in a group of web pages as it jumps to a random page. This is done using d, the dampening factor always set around (0.85) in the PageRank algorithm which defines how often this happens. It also helps stop pages trying to trick the agent by leading it around, as it will jump randomly sometimes, instead of following the links.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

PageRank Calculation

The importance of a web page can be judged by the number of hyperlinks pointing to it from other web pages.

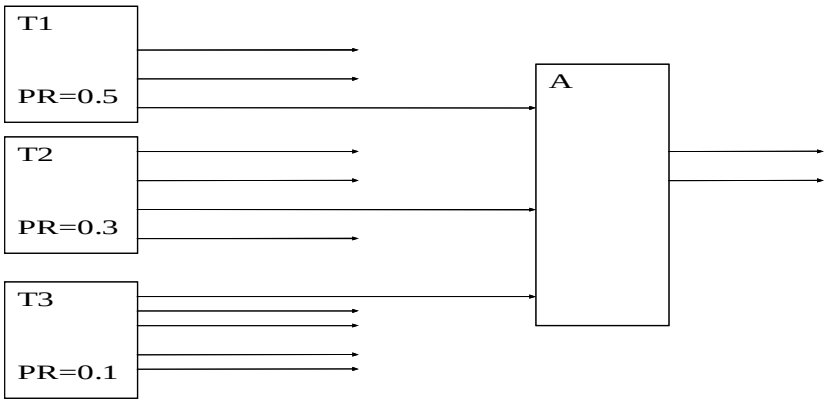

Figure: PageRank calculation

PR (A) = (1-d) + d (PR (T1)/C(T1) + PR(T2)/C(T2) PR(T3)/C(T3)

PR (A) = (1- 0.85) +0.85(0.5/3 + 0.3/4+ 0.1/5)
PR (A) = 0.15+ 0.85 (0.166+0.075+0.02)
PR (A) = 0.15+ 0.85 (0.261)
PR (A) =0.15 + 0.22185
PR (A) = 0.37185

Big Data Analytics

Big Data Analytics means the process of examining large amount of data. Big Data Analytics has been occurred in every domain such as Better Business Decision, Retail Banking, Real Estate, Effective Marketing, oil and gas, Travel and Transport Sector, Retail Industry, Identification of Hidden Patterns, Customer Satisfaction, Traffic Control, Smarter Health Care, Search Quality, Trading Analytics, Manufacturing, Smarter Health Care, Multichannel Sales, Telecom, Social Media and online services, Education and Research, Law enforcement and defence industry etc. In 2012, the Obama administration announced the Big Data Research and Development Initiative, to explore how big data could be used to address important problems faced by the government. Big Data analysis was in parts, responsible for the BJP and its allies to win a highly successful Indian General Election 2014. Big Data is the ability to make better decisions and take meaningful action at the right time.

Big Data Techniques and Technologies:

The Techniques associated with Big Data are:

- Association rule learning
- Data Mining
- Cluster Analysis
- Correlation Analysis
- Statistical Analysis
- Regression Analysis
- Crowd Sourcing
- Machine Learning
- Text Analytics

The Technologies Associated with Big Data are

- Enterprise Data Warehouses

- Visualization Products
- MapReduce
- Hadoop
- NoSQL databases

Big Data Applications in Libraries

Libraries and Librarians are uniquely suited to working with Big Data. Libraries have a long tradition of being easily technology adopters and Big Data should be no exception. Big Data Applications in libraries offering more online services, predictive analysis of user reading habits, better understand the user needs and requirements. Big Data better forecasts for future library planning, Better usage of systems and resources, it fully supports in data management, Library of congress world cat, data federation technology, web archives, community management, open access, open data standards, digital archives, copyright acts, social media use like facebook, Linkdin, twitter, instagram, whatsapp, etc. for library support services. Big Data also supports for productivity gain with better decision making.

Conclusion and suggestions for further research

Big Data sources are users, Mobile Devices, Scanners, Social Media, Systems, etc. Big Data isn't Big if you know how to use it. The search techniques were developed and followed by library and information science professionals since the inception of information retrieval. The wisdom of library and information science professionals should be applied Big Data Search Strategies for effective information retrieval. Librarians use emerging search tools to collect more online data.

Librarians and Information Professionals have always worked with data in order to meet the information needs of their patrons.

The world has fallen in love with Social Media, readers are approaching online platforms to research, to purchase books and new library services. Big Data gives huge challenges of getting out through and keeping up-to-date Librarians. This gives Fantastic opportunities for Librarians to engage readers and encourage content sharing.

The future of Big Data is the ability to make well informed decisions quicker and easier than ever before. The Big Data developers should apply the criteria of precision, relevance, and recall for efficient retrieval of Big Data.

Acknowledgments

The author would like to thank professors Dr.Chandrashekar Rao, Dr.Sudarshan Rao, Dr. Laxman Rao, Dr.V.Vishwa Mohan, for their able guidance, encouragement, constant support and whole-hearted cooperation.

References

[1] Brin, S., Page, L. (1998) *"The anatomy of a large-scale hyper textual Web search engine"*. Computer Networks and Isdn Systems, Vol. 30, No. 1-7, pp. 107-117.

[2] Bernard Marr (2015*) Big Data: Using SMART Big Data, Analytics and metrics to make better decisions and improve performance,* Wiley Publishing.

[3] Bertino Elisa (2013) Big Data - Opportunities and challenges, IEEE 37th Annual Computer Software and Applications Conference, Computer Society, pp.479-482.

[4] Big Data definition retrieved from http://www.oxford dictionaries. com /definition/english /big-data, retrieved on 24-10-2016 at 8.41pm.

[5] Data Science Central: The online resource for Big Data practitioners, retrieved from www.datasciencecentral.com /profiles /blogs / to 10 machine learning algorithms on 28-07-2017 at 8.00am.

[6] Geanina Elena et al. (2012) Perspectives on Big Data and Big Data Analytics, *Database Systems Journal* Vol.III, no.4/2012, pp. 3-13.
[7] Najafabadi et al. (2015) Deep Learning applications and challenges in big data analytics, *journal of Big Data* 1-21.
[8] Ohlhorst, Frank, (2015) Big Data Analytics: Turning Big Data into Big Money (pp 12-90), Wiley Publishing.
[9] Prajapati, Vignesh (2014) *Big Data Analytics with R and Hadoop* (pp 1-11), Packt Publishing.
[10] Ravi Kumar Jain, B. (2007) *"Dynamics of Search Engines:* An Introduction ICFAI University press," Hyderabad.
[11] Sangeeta, K. Shivarajadhanavel, P. (2007) *"Google's Growth A Success Story"* ICFAI University Press, Hyderabad.
[12] Tavish Srivastava, PageRank explained in simple terms retrieved from www.analytics vidhya.com/blog/2015 04/ PageRank explained simple on 26-07-2017 at 3pm.
[13] Tanvi Ahlawat and Radha Krishna Rambola, (2016) "Literature review on Big Data" International Journal of Advancement in Engineering Technology, Management and Applied Science, Volume 3, Issue 5 May 2016, ISSN No. 2349-3224
[13] Very Short History of Big Data retrieved from http://www.forbes.com/sites /gilpress/2013 /05/09 / very-short-history-of-big-data/#d5b7ca855da9 retrieved on 24-01-2015 at 8.56pm.
[14] www.smartinsights.com.internet marketing statistics/what happens online in 60 seconds retrieved on 30-07-2017 at 8 pm.
[15] http://www.whitehouse.gov/blog/2012/03/29/bigdata/bigdeal. white house.
[16] http://www.livemint.com Are Indian companies making enough sense of Big Data?