International Journal of Library and Information Studies
Vol. 6(4) Oct-Dec, 2016 www.ijlis.org ISSN: 2231-4911

Big Data Search Strategies test on Google Search Engine And Big Data Applications in Libraries

Dr. J. Vivekavardhan
Asst. Professor in Digital Information Management
Department of Library and Information Science,
University College of Arts and Social Sciences,
Osmania University
Hyderabad.
e-mail: jvk_dn@yahoo.com

Dr. R.K. Pavan Kumar
Librarian (i/c)
Osmania University Library
Osmania University,
Hyderabad.

Abstract *- The paper explores and tests the 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. on Google Big Data. The paper describes on Big Data, Big Data Retrieval UseCase diagram and Characteristics of Big Data. The paper focuses on Google PageRank algorithm, with an example and Big Data applications in libraries. Paper finally presents findings of the study conclusion and suggestions for further research on Big Data.*

Keywords: Big Data, Search Strategies, PageRank Algorithm

Introduction

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 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. Effective use of big data is a challenging task for Library and Information Science Professionals. The main problem is displaying only important pages relevant to the keyword(s) typed by users. Each search engines has its own algorithm. The importance of a web page can be judged based on the content specified in it or based on link information. Searching for information on the World Wide Web (WWW) is done in much the same way that you look for information in a library, using an on-line catalog system (the updated version of the old index card system). The difference and the advantage is that you can get information from all over the world, instead of from a single library collection. Google search engine aims to use Big Data to its fullest extent, to judge search results, predict Internet traffic usage, and service customers, Big Data reaches deep with Google's own application called as PageRank Algorithm.

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".

Big Data Retrieval Use Case Diagram

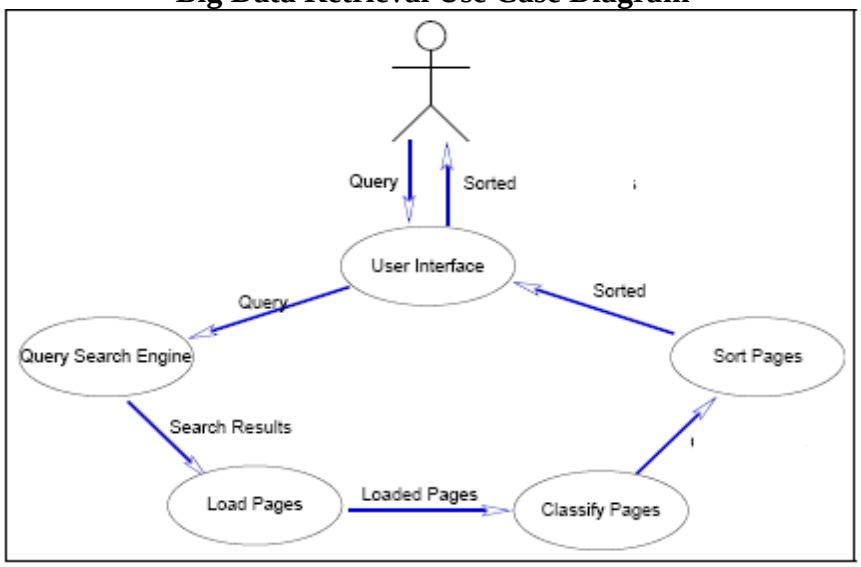

Figure (1): Big Data Retrieval Use Case Diagram (source: www.google.com)

The 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 Characteristics

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

Table (1) Big Data Characteristics

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

Objectives of the study

The main objective of the study:

- To test the different types of search strategies on Google Big Data
- To find out the various Big Data algorithms and the role of PageRank algorithm to search Big Data.
- To find out the Big Data analytics and applications in Libraries.

Statement of the Problem

The study has been undertaken to test the different types of search strategies. An individual cannot read Peta bytes of Big Data available on the web, so user need different search strategies to search the big data effectively and efficiently as per the need.

Significance of the study

Big Data have become an indispensable tool in our everyday life. When we seek information we often go to our favorite search engine and look at the returned pages. This study would help to assess the user what type of search strategies use while searching the Big Data to retrieve relevant and exact information from the web.

Methodology

The study is focused on to test the Big Data Search Strategies with select key words on Google search engine. The study is based on extensive review of literature available in the print journals, online journals on internet to examine about Big Data technology, and how search engine retrieve the Big Data from their servers and databases.

Limitations of the Study

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

Big Data Google Search Strategies

World Wide Web has become an indispensable source of information for any one, it needs to understand how people search and retrieve Big Data. Search engines have been playing an important role in finding the required information from ever growing internet. Big Data Search Strategies are as follows in the table 2.

Table 2: Big Data Google Search Strategies

Sl No	Big Data Search Strategies	Description / Use of Search Strategies	Example
1	AND (+ plus sign)	It Narrows search.	Joomla AND Drupal
2	OR	Broadens search.	Joomla OR Drupal
3	NOT (- sign)	Contain one keyword exclude the other keyword.	Joomla NOT Drupal
4	Nesting ( ) Parentheses	Utilizes parentheses to clarify relationships between search terms.	(Big OR Data) AND (Libraries)
5	Proximity Search	Search for two or more words that occur within a specified number of words of each other in the database.	Big Data Analytics 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
18	Weather	"weather" along with city and country.	weather Hyderabad, India
19	Calculator	Evaluate Mathematical Expressions	$(18+19)^9$
20	Images	Relevant images	Big Data Analytics
21	News Headlines	Latest News and stories	67th Republic day chief guest
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: Big Data
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)

(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.

PageRank Algorithm

PageRank is a link analysis algorithm used by the Google search engine. 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.

PageRank Calculation

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

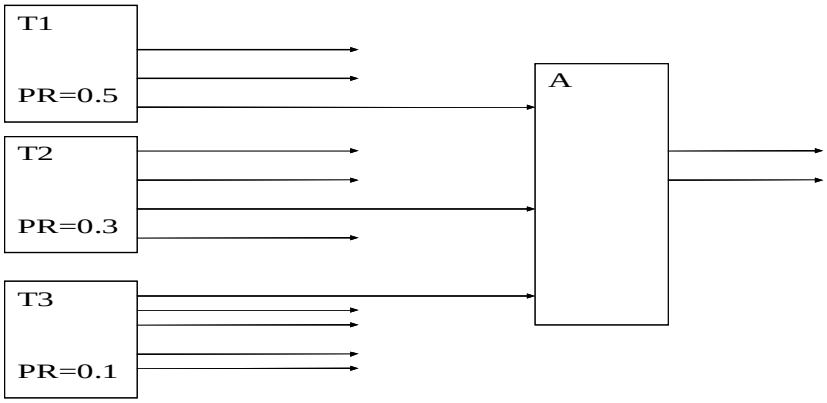

Figure: PageRank calculation

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

Results and Discussion

Big Data Search Strategies conducted during a particular time period on 27 & 28-11-2016 (mentioned in table 3), so that the data yielded would be valid as configuration, search results and retrieval time of search is mentioned. There will not be any additions, deletions in the Big Data of the Google Search Engine. This would give an idea as how these search strategies are facilitating the users in the process of Big Data retrieval. Data were tested by the researchers personally by using a set of search queries by using keywords to test how each and every search strategy would yield the search results.

Microprocessor configuration

The Microprocessor used to test the Big Data Search Strategies is intel core i5 Processor, 500 GB Hard Disk, with 4GB RAM and 100 mbps speed of internet connectivity.

Big Data Search Strategies test on Google Search Engine

Big Data search system, provides the users with keyword search facility. A search is conducted with select different search strategies by using Google search engine. The search strategies tested are single keyword, two keywords, Boolean operators search (AND, OR, NOT), Phrase search, Nesting, Proximity search, *Truncation, File Format search, Image search, Domain search. The following is the data yielded by the Google search engine.

Table 3: Testing Big Data Search Strategies of Google Search Engine

Sl. No	*Big Data Search Strategies*	*Keyword Search*	*Big Data Google Search Results*	*Retrieval Date 27& 28th January 2016 and Time*	*Retrieval Time Taken in seconds*
1	Single keyword	Big	4,03,00,00,000	10.49pm	0.63
2	Single keyword	Data	5,72,00,00,000	10.50pm	0.91
3	Two keywords	Big Data	83,60,00,000	11.06pm	0.92
4	Boolean Operators Search	Big AND Data	1,72,00,00,000	11.13pm	0.52
5		Big OR Data	11,31,00,00,000	11.15pm	0.77
6		Big NOT Data	1,46,00,00,000	11.16pm	0.53
7	Phrase Search	"Big Data"	5,47,00,000	11.23pm	0.54
8	Nesting ( ) Parentheses	(Big OR Data) AND (Libraries)	23,70,00,000	11.28pm	0.48
9	Proximity Search	Big Data Applications	20,90,00,000	11.35pm	0.47
10	Proximity Search	International conference on big data and knowledge discovery, Bangalore, 2016	39,400	11.54pm	0.49
11	*Truncation	Knowledg*	4,05,000	12.11am	0.43
12	File Format Search	Big Data.doc	1,60,000	12.13am	0.42
13		Big Data .pdf	36,70,00,000	12.18am	0.43
14		Big Data .xls	87,500	12.19am	0.42
15		Big Data .txt	2,25,000	12.21am	0.44
16	Image search	Big Data Analytics Images	10,20,00,000	12.29am	0.48

17	Domain search	Big Data .com	79,30,00,000	12.32am	0.50
18		Big Data .gov	1,04,000	12.34am	0.56
19		Big Data .edu	83,90,00,000	12.35am	0.44
20		Big Data .org	1,76,000	12.37am	0.44

Retrieved on 27-11-2016 at 10.49pm to 11.54pm and 28-11-2016 at 12am to 12.37am (source: www.google.com)

Findings

The above Big Data (table 3) shows unpredictable search results with the Google search engine. The search results as shown in table (3) searches with single keyword is more than that of two keywords. Huge number of hits is the major problem in use of Big Data. Boolean operators AND restricts search results, OR increases search results and NOT restricts search results. Phrase search minimises recall value. Proximity search or sentence search results are restricts search results. File Format search .xls results are less comparatively .pdf search results. Domain name .gov recall value is less comparatively to .edu search results. The search results with any extension (i.e phrase, format, domain, etc.) should be less than those without any extension. This would project the efficiency of the Big Data search. Any extension should restrict the scope of the search.

Big Data Analytics and Applications in Libraries

Big Data Analytics means the process of examining large amount of data.Big Data Analytics has been occurred in every domain such as search quality, trading analytics, manufacturing, traffic control, smarter health care, multichannel sales, telecom, social media, etc. Big Data is the ability to make better decisions and take meaningful action at the right time. Big Data Applications in libraries offering more online services, predictive analysis of user reading habits, better understand the user needs and requirements, 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.

Conclusion and suggestions for further research

Big Data isn't Big if you know how to use it. The functioning of Big Data Search Strategies is observed to be inconsistent from time to time and search to search. Google search engine have extraordinary speed as hundreds of crores of entries are searched within fractions of seconds. This indicates that the search engines' speed is unimaginably high and they function at the level of microseconds and nanoseconds. Around the world in every minute 204 million emails are sending one another, 18 million face book likes, 278 thousand tweets, 200,000 photos uploading to face book, 100 hours video uploaded in You Tube, 570 new web sites are generating. Wall Mart handles more than million customers every hour. Facebook generates 10TB daily and Twitter 7TB daily.

The search techniques developed and followed by library and information science professionals since the inception of information retrieval search may be utilised, so that the search engines which prove to have more semantic value in the search results. Thus the wisdom of library and information science professionals should be applied Big Data search strategies of the Google search engine. Librarians use emerging search tools to collect more online data.
The Big Data developers should apply the criteria of precision, relevance, and recall while developing the search engines for efficient retrieval of Big Data.

Acknowledgments

The authors would like to thank professors Dr.Chandrashekar Rao, Dr.Sudarshan 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 29-09-2016 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 26-11-2015 at 3pm.

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.