Volume No:2, Issue No:11 (April-2017) ISSN No : 2454-423X (Online)

International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Exploring Library Using Big Data

R. Rohith Sai Venkatpathi Raju
Student,
Dept of CSE,
Sphoorthy Engineering College.

K.Pavan Kumar
Assistant Professor,
Dept of CSE,
Sphoorthy Engineering College.

Mrs. J. Deepthi
HOD,
Dept of CSE,
Sphoorthy Engineering College.

ABSTRACT:

Libraries play an important role at the intersections of government, universities, research institutes, and the public since they are storing and managing digital assets. The large amount of data and those data in library need to be transformed into information or knowledge which then be used by researchers or users. Librarians might need to understand how to transform, analyze, and present data in order to facilitate knowledge creation. For example, they should know how to make big datasets more useful, visible and accessible. With new and powerful analytics of big data, such as information visualization tools, researchers/users can look at data in new ways and mine it for information they intend to have. In this work, we discussed the characteristics of datasets in library, conducted a review for the research work on library big data and then summarized the applications in this field. The issues associated with it were also discussed and explored.

1. INTRODUCTION:

Libraries have collected a large amount of data, such as books, research articles and reports, both in physical and electronic formats. The library collection was originally for researchers or public users to find necessary information they need . However, this data becomes so large and the format is so various which might affect the efficient use. Although a lot of library data has been digitalized, most of them have not been used for data mining or big data technology.

For example, the Geological Library of China for which one author works has 710,000 records of physical and electronic books and journals, recordings, maps, and field trip notes, however, most of these individual records remain isolated from the Web, requiring a detailed study how this data could be effectively exposed for use with current big data and other technologies. On the other hand, although some work has been done in the past on how to maintain those library collections in order to efficiently and effectively use , there is no much research on using meta data to organize digital assets so that the big data and cloud computing technology could be used. Big data is a popular term. The three common characteristics of big data include high-volume, velocity and/or variety information . Although a few researches have linked library data into big data, some researchers raise questions about that since there is no clear characteristics of velocity . In addition, based on the current terminology, it seems that the database management systems is enough for storing and processing library data, thus does not require big data technology such as distributed systems for analysis or processing. Therefore, it might worth to clarify this doubt. Moreover, it seems that there is no much general review work s on the research for library "big data".

2. RELATED WORK:

"Big data" describes innovative techniques and technologies to capture, store, distribute, manage and analyze datasets that traditional data management methods are unable to handle. The concept of "Big data" was first defined by Laney in his research note . According to the definition, big data is mainly characterized by three Vs: Volume, Velocity, and Variety . The first V, refers to the data volume. General speaking, the size of the data sets of big data is huge compared to regular data. However, itseems that there is no fixed definition for the size, i.e. how big of data could be classified as big data. Therefore, the size might vary based on the disciplines. Traditional software usually can handle megabyte and kilobyte sized data sets, while big data tools should be able to handle terabyte and peta-byte sized data sets. The second V, velocity, refers to the situation where data is created dynamically and fast. The data come in every second or so. The third V, refers to variety, which makes big data sets harder to organize and analyze. The regular type of data collected by researchers or business is strictly structured, such as data entered into a spreadsheet with specific rows and columns.

However, big data sets might have unstructured data and different types of data, such as email messages or notes. The hardware and software for storing and analyzing big data is cheaper and available for business and government now which makes the big data technique interesting to a lot of users including library. The important part is that the user could make prediction based on big data analysis. Work about big data in library could also be found because library data need to be transformed into information or knowledge which then be used by users. Bell tried to explore the issues and possibility of big data in library . Parry studied how colleges are using big data to help students chose classes, retain them, and provided necessary advising . The government initiatives on work of big data for libraries and the impact on the library collections have been discussed by Schwartz. Affelt described how traditional library skill sets could match up to the needs of data analysis and discussed big data technology for library and how librarians could use it. Pro-Quest tried to understand the behaviour of library users such as how to perform search, by using big data technology . They mentioned their work could help to develop some search services to better serve library users. Salo studied the characteristics of library data and summarized several emerging research challenges in this area.Reinhalter and Wittmann mentioned that librarians could fill a service gap by enforcing standards and best practices in the big-dataera because they could create trustworthy data repositories for researchers.

3. IS LIBRARY DATA A BIG DATA?

"Big data" is one of the most popular terms these days. The hospitals, manufacturers, colleges, banks, retailers and governments are all collecting those so called "big data". Libraries are also doing it. Of course, the ultimate goal for doing this is to use these data to provide new useful services or to improve efficiency.
The concept of "Big data" was first coined by Laney in 2001 in his research note. Laney described the characteristics of big data as which cannot be processed by traditional data managementtools .

Three Vs were first used to characterize the big data. With further study on big data, the "Three V's" have been expanded to "Five V's": volume, velocity, variety, veracity (integrity of data), value (usefulness of data) and complexity (degree of interconnection among data structures) by some researchers. However, the most important are still the first three. If we only consider the static collection in libraries, it might be hard for us to relate it to big data. In addition, the database management systems should be enough to store and to process library data, therefore, based on the definition of big data, there is no need for big data technology such as distributed systems toanalyze the data in library . In this section, we try to analyze the properties of data sets in library and to understand how close they are related to big data.

A.Volume:

According to Wikipedia, ‘Big data’ refers to data sets whose size is beyond the ability of traditional software tools for capturing, managing, and processing the data. However, the actual size is a moving target, which could range from a few dozen terabytes to many petabytes of data. The size of big data varies depends on the discipline. Some recently developed big data applications include healthcare, transportation, and entertainment, all of which involve enormous collections of data. It seems to us that each library has limited collections. For example, the National Geological Library of China has only 710,000 collections which are much smaller than those in other fields. On the other hand, library collects a lot of “small research data”, which are created by individual researchers. Those tens of thousands of small-data producers in aggregate may well produce as much data (or more, measured in bytes) as the big data.

Moreover, library collections have a close tie to the linked data which forms larger web of big data. British library studied the linked data of library collections and tried to model the people, events, places which are related to holdings in the library. The US Library of Congress has done a similar work. On the other hand, data schema could be created from library collections. For example, the relationships from co-authors, citations, geo-location, dates, named entities, subject classification, institution affiliations, and publishers could be easily extracted based on the books or journals.

Those relationships could then be connected to other works, people, patents, events, etc. Creating, processing and making available this graph is big data.
In general, the data stored in library certainly can be classified as large since it has hundred years of collections on one hand, contains tens of small research data as well and the data captured during users using the library service.

B. Velocity:

The velocity characteristics of big data could also be found in the data from library. Library maintains multiple copies of files on servers and on tape, in geographically distributed locations. Therefore, there are movements of files between and within organizations. There are more and more researches going on and the research data come in and join the dataset dynamically. On the other hand, the library data need to be processed fast so that researchers could use it with value and ordinary users could receive the search results they need right away.

C. Variety:

In general, libraries contain different types of data: books, journals, reports, notes, maps, films, pictures, audios etc. Some are unstructured. Unstructured data consists of language-based data (e.g., notes, twitter messages, books) and non -language-based data (e.g., pictures, slides, audios, videos). Even for digital research data, they have every imaginable shape and form, from scans of historical negative photographs to digital microscope images of unicellular organisms taken hundreds at a time at varying depths of field. On other hand, as a matter of course libraries collect a host of usage and transactional data created by users as they interact with their systems and services. They are awash with this type of data –and are waking up the potential value that can be extracted from what at the moment is largely, unstructured data. Therefore, the characteristics of variety the big data obtains could also be found in the library data. Besides those mentioned characteristics, the library data also have other properties.

D. Data Less Organized:

It appears to us that the data such as books, journals in library are well organized since users could use categories to look for what they need. However, the situation is different for those research data stored in libraries.

The research data in libraries seem to be disorganized, less described, and in formats poorly suited to long - term reuse . Researchers are used to their own process to produce these unorganized data. Those data are often managed by the project. Once projects complete with publication of articles or reports, research data are often locked into digital closets being unorganized.

E.Non-Standard Data and Data Format:

Research data often lack of standard and format. They depend on the disciplines and individual libraries. Although a few disciplines might have created data standards, due to a strong centralized data repository, such as political and social research, in most disciplines, there often do not exist data standards, particularly for those researches which are individualized: i.e. each researcher defines the parameters which are important to the project. The data format is another issue. Researchers use their own format for the data they collect. Even for the same researcher, different data formats might be used for different projects, which pose difficulty to integrate those data.

F. Summary:

In conclusion, library data could be treated as big data without doubt due its property of large volume; high velocity and obvious variety. In addition, library data are often less organized, lacking of standards and unique formats.

4.ISSUES WITH BIG DATA IN LIBRARY:

It is clear now that library contains big data which is valuable. However, the big data is different from the data in other fields such as hospitals, business, as mentioned above. Big data research in library is relatively new. Therefore, there might exist some issues or difficulty in the process of data transformation, curation, analysis, and presentation. At less, the technology used in library big data might be different from that in other areas such as storage, software and personnel. One example is that should we create full-test indexes for millions/billions of files to support full discovery in library? Can library staff develop the expertise to provide guidance to researchers in using analytical tools? In order to apply data science efforts in library collections, there are some works which should be done, such as, but not limited to:

□_ Central data repositories, where data are stored, maintained, and cataloged;
□_ Data standards, to which collected data should follow;
□_ Data communities, which collect, maintain, and curate data;
□_ Analysis tool

There are some issues which are common to library big data research as listed below.

A._ Lacking of Data Scientists:

According to studies, USA might not be able to fill half of the positions of data scientists and data managers by 2018 . The situation in library might be same. The key issue is that data analysts need not only the skills of statistics and computer science, but also skills of domain knowledge and collaboration ability. Therefore, the challenges faced by librarians are the ability to manage the information of big data. It seems that short-course training might not be sufficient.

B._ Ability of Adopting Big Data:

Big data comes in various fields. However, a lot of companies are not ready for it. According to the study, more than half of organizations could not handle the big data currently due to lack of personnel and platform . Research of library big data is even much slower than that in other disciplines . The key reason is that the digital libraries tend to be self-contained organizational units and they try to stay back from new technology.

C. Budget Issues:

Although more and more people understand the great benefit of using big data analysis, the IT investment such as analytics servers, high-performance computing servers are needed. Majority of US government organizations have not had plan for investment in big data mainly due to budgetary issue . It seems that most of library administrations have not yet placed big data on the table because of shrinking budgets as well. Research data managed by projects are paid less attention due to the challenge of human resources. Moreover, a lot of research data which were produced ten year ago is still analogue, such as the laboratory notebook for biology research or geology work. Digitizing these resources is not a simple task, which need a lot of time and personnel resource.

D. Technical Challenges:

Big data involves techniques such as capturing, storing, processing and presenting data. Data in the library have different types and might be in various statues. Some data might be waiting for digitalization. For geological data, data capturing often face challenge. For example, digitalizing field trip notes and geological maps is still an issue. On the other hand, a large set of data often contains some dirty or false data. Therefore, correctly removing those data needs some work.

E.Privacy:

Big data is mining the data and discovering knowledge. There should be a privacy issue. On the other hand, new risks of system intrusions might arise due to the accessibility of a large amount of data. Data security issues have not been well considered for library big data research.

F. Big Data Not For All Organizations:

It is clear that the organizations that plan to use big data need to have a relative large investment in IT infrastructure and personnel. Therefore, small library without enough budget support might need to share the resource with other organizations. On the other hand, big data is relative new and traditional analytic approach still dominates majority of organizations. With regards to the individual research data, small library might not have enough resource to support direct interaction with research faculty. Therefore, it might be hard to integrate all the data from all researchers in the organization.

G. Summary:

A lot of libraries begin to introduce big data technology. It is necessary to realize that there are still some issues such as budgetary issue, technique issue etc. Other common issues faced by most big data researches include: the misunderstanding of time within the data, the scale of big data, the inability of big data analysis to deal with non-linear dynamics, the question of causality and the challenges of inter-disciplinary. On the other hand, faculty in all disciplines are increasingly creating and/or incorporating big data into their research and institutions or libraries are creating repositories and other tools to manage it. There are many challenge to effectively manage and curate this research data which are similar and different to managing document archives such as books or journals.

5.WHAT CAN WE DO WITH LIBRARY BIG DATA:

Big data is a hot topic during these days. While businesses are analyzing big data looking for improved ways of selling products and services, governments are analyzing telecommunication and financial data to track money launderers and terrorists . Hospitals use it to prevent illness. The financial industry uses it to detect credit card fraud. Airlines use it to fill seats. What can it do for library?

A. Data-Driven for Decision Making:

Data -driven approach, which takes the data as the basis, to make decision or recommendation, is a common method used in many areas. For example, it is used in the database design or software design. It is now the key approach for library big data. Based on the data, the decision could be more useful. For example, based on the loan transactions customers borrow or search, the library could use collaborative data mining techniques and text analytics to optimize the collections (books or journals) to generate better search results and to make recommendation for the books. At the end, this approach would improve the customer satisfaction by providing better service, and efficient usage of library resources.

B. New Data Format:

Sharing data and make data accessible is one of the important goals of library. However, a lot of data need to be re-done, particularly those data collected in old time. Digitalizing them is the first step, by scanning or microfilming. Other efforts critical to increased data access and reusability involve building tools and infrastructure. For example, the USGS is deploying several innovative applications which could help geological scientists better manage their data. One example is Science Base, a data-management platform allowing the uploading and cataloging data. USGS is developing Science Base which is also used for release of its official data to public.

On the other hand, reformatting library data so that it can work with other online resources that users might intend to connect with . For example, OCLC (Online Computer Library Centre) has been working to produce a Google- like “knowledge card” based on the reformatted library data and the card can be linked to from the outside . Library data could become linked data in order to achieve interoperability on the Web. Without reformatting the data, it might be hard to effectively achieve such purpose.

Another example is that British Library studied the linked data of the library collections and modelled the people, events, places which are related to holdings in the library.

C. Data Standardisation and Data Modelling:

Metadata is the key characteristics for the data in database. However, there is no metadata or standard for research data. Building data with metadata could certainly foster sharing and remixing of library data (research data). It seems that using Entity-Relationship Diagram to model the key concepts in the library holdings is another useful approach. First we identify the entities and then assign the relationships between each two entities. Each entity has properties which should be also identified. This is a process of database modelling. This is also a process that we transition from just having lots of data to what is known as big data where data are connected. Understanding the relationships among the entities could help us to link all different types of data in the library holdings. It then could allow the library users to find items of interest among the holdings in ways which were not previously possible.

D._ Library Data Visualization:

Library data could be selected and visualized by tools such as Tableau dashboard, to present to users as user’s need. On the other hand, librarian in the university library could use data visualization to compare sections of the library collection, expenditures in those areas, with the number of majors in them. The possible unbalance in the collections or budgeting in those areas might be able to determine and then provide planning advice.

E.User Behaviour Study:

As mentioned previously, the information of library collections could be mined through big data technology . On the other hand, it is possible to record and track library user’s activity and to store that data in large-scale data storage, and then conduct data analysis. The result could then be used to potentially improve the overall user experience, and user satisfactory of library service.

F. Summary:

Big data technology could be used in library. Main process involves collection selection, organization, description and modelling, storage, presentation or visualization. Of course data analysis is also important. Data reformatting, data standardization and data modelling need a lot of work, and of course, the return from the work would be enormous. In addition, the amount of storage and processing has grown the complexity of the library data and the challenges of working with it have also accelerated. Another issue is that this work could only be done by "data scientists", not traditional librarians, although librarians have always been great at information management and organization. The reason is that they need to understand all of the following: the Internet, databases, analytics, visualization, and data curation.

6. PAST:

Library data has been analyzed through data mining and prediction's have been obtained .In the scenario of missing of a book the data mining is used and we have predicted whether the lost book can be found or it'll be missing forever. The tool is used in this scenario is Rapid miner.

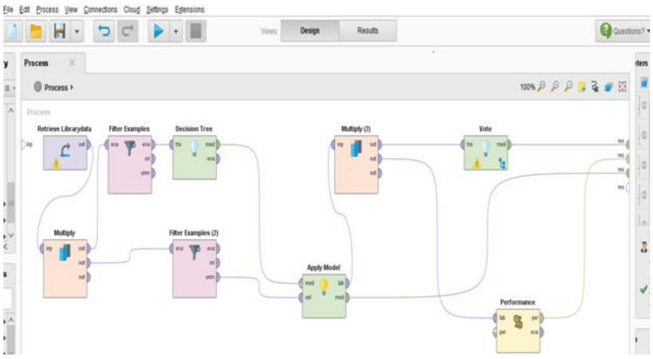


Fig 6.1 how it is done

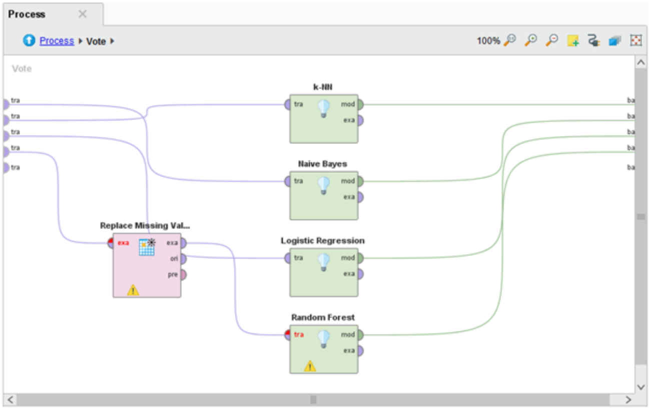
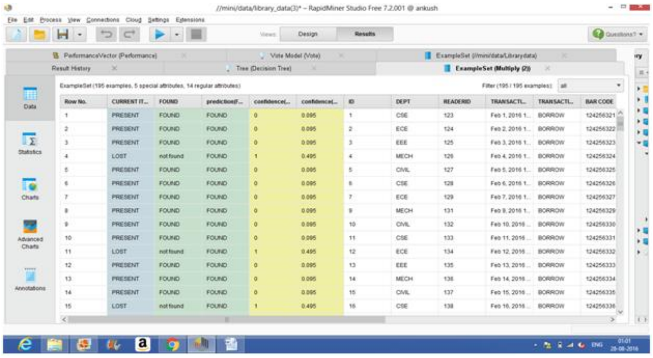


Fig 6.2 methods used


//mini/data/library_data(3)* – RapidMiner Studio Free 7.2.001 @ ankush
File Edit Process View Connections Cloud Settings Extensions
Design
Results
Questions?
PerformanceVector (Performance)
Vote Model (Vote)
ExampleSet (//mini/data/Librarydata)
Result History
Tree (Decision Tree)
ExampleSet (Multiply (2))
Data
Statistics
Charts
Advanced Charts
Annotations

ExampleSet (195 examples, 5 special attributes, 14 regular attributes)

Filter (195 / 195 examples): all

Row No.	CURRENT IT...	FOUND	prediction(F...	confidence(...	confidence(...	ID	DEPT	READERID	TRANSACTI...	TRANSACTI...	BAR CODE
1	PRESENT	FOUND	FOUND	0	0.095	1	CSE	123	Feb 1, 2016 1...	BORROW	124256321
2	PRESENT	FOUND	FOUND	0	0.095	2	ECE	124	Feb 2, 2016 1...	BORROW	124256322
3	PRESENT	FOUND	FOUND	0	0.095	3	EEE	125	Feb 3, 2016 1...	BORROW	124256323
4	LOST	not found	FOUND	1	0.495	4	MECH	126	Feb 4, 2016 1...	BORROW	124256324
5	PRESENT	FOUND	FOUND	0	0.095	5	CIVIL	127	Feb 5, 2016 1...	BORROW	124256325
6	PRESENT	FOUND	FOUND	0	0.095	6	CSE	128	Feb 6, 2016 1...	BORROW	124256326
7	PRESENT	FOUND	FOUND	0	0.095	7	ECE	129	Feb 7, 2016 1...	BORROW	124256327
8	PRESENT	FOUND	FOUND	0	0.095	9	MECH	131	Feb 9, 2016 1...	BORROW	124256329
9	PRESENT	FOUND	FOUND	0	0.095	10	CIVIL	132	Feb 10, 2016 ...	BORROW	124256330
10	PRESENT	FOUND	FOUND	0	0.095	11	CSE	133	Feb 11, 2016 ...	BORROW	124256331
11	LOST	not found	FOUND	1	0.495	12	ECE	134	Feb 12, 2016 ...	BORROW	124256332
12	PRESENT	FOUND	FOUND	0	0.095	13	EEE	135	Feb 13, 2016 ...	BORROW	124256333
13	PRESENT	FOUND	FOUND	0	0.095	14	MECH	136	Feb 14, 2016 ...	BORROW	124256334
14	PRESENT	FOUND	FOUND	0	0.095	15	CIVIL	137	Feb 15, 2016 ...	BORROW	124256335
15	LOST	not found	FOUND	1	0.495	16	CSE	138	Feb 16, 2016 ...	BORROW	124256336

Fig 6.3 Outcome

7. ADVANTAGES:

1.User satisfactory services offered by libraries.

2.More suggestions could be given to user regarding books and articles .

3.Less burden on library staff.

8. DISADVANTAGES:

1)More storage and processing of data.

2)Could only be handled by Data scientists unlike traditional librarians.

3)More ETL processes.

9. CONCLUSION AND FUTURE SCOPE:

The ability to collect and analyze massive amounts of data will be a competitive advantage across all industries, including library. The big data currently might be suitable only for those organizations with large set of data and funding.

The traditional DBMS or data analytics might be still a dominant approach. In future, we will survey the actual platforms or technologies used in library big data.

10. REFERENCES:

1. Affelt, A., 2015, The accidental data scientist: big data applications and opportunities for librarians and information professionals, Medford, New Jersey. 9781573875110.
2. Armour, F., 2012, Introduction to big data, presentation at the symposium Big Data and Business Analytics: Defining a Framework, Centre for IT and Global Economy, Kogod School of Business, American University, Washington, DC.
3. Bell, S., 2013, Promise and Problems of Big Data, LibraryJournal. March 13.
4. Eliot, S.; Rose, J., 2009, A Companion to the History of theBook; John Wiley and Sons: Hoboken, NJ, USA, p. 90.
5. Gartner, Inc., 2012, Press Release. Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015. http://www.gartner.com/newsroom/id/2207915.
6. Heidorn, P. Bryan, 2008, Shedding light on the dark data in the long tail of science. Library Trends 57:2, pp. 280-299.