Application of Big Data Technology to Library data:A review

A.Kaladhar
Research Scholar
Dept. of Library and Information Science
JNTUK
Kakinada (A.P)
Email:librarian@svecw.edu.in and

B.R. Doraswamy Naick
Dept. of Library and Information Science
Associate Professor
UCE- JNTUK
Kakinada (A.P)

K. Somasekhara Rao
Dept. of Library and Information Science
Andhra University
Vishakapatnam (A.P)

Abstract - *As per changing trends the primary duty of libraries is to preservation of data or information in the digital form to act as juncture to its users, research institutions, universities and to the government. The data in large quantity which is in the form of raw data to be prepared in to required formats as per its user's requirement. At this situation the librarians or badly need to know how to convert, evaluate and brought before to the users in the ultimate form of information or knowledge. Mere creation of knowledge is not enough in the digital era that should be useful to maximum extent, perceptible and reachable. The new concept and challenge to the librarians is powerful analytics of "Big Data". Big data is a Information revelation tool which presents data in a different way and facilitate information mining to its users what they really intended to have. In this manuscript we have discussed various forms of datasets and their characteristics, consulted IT (Information Technology) experts on relevancy of Big data technology to the library operations. We identified so many issues related to it and explained.*

Key Words: Big data, Information, Data mining, Information technology, Library Data

I. **Introduction**

The library collection generally consists of books, e-books, online journals, research papers, conference proceedings, seminar papers, institutional repositories etc., both in the electronic and physical formats which are specially meant for research scholars, and other users to fulfill their information needs. The ever increasing nature of data in to information and to knowledge necessitated Big data analytics. For example if we take Polavaram Project being constructed in the state of Andhra Pradesh, a project engineer is having 2,523 records of physical and electronic books and journals, recordings, maps, and field trip notes, however, most of these 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. However as library professionals though we have many efficient and effective techniques in managing such type of data there is no much research on using metadata to organize digital assets by using tools like big data and cloud computing technology. Big data is gaining momentum in the information and communication era. General opinion is 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. The purpose of this paper is to review the researches which have been done in the usage of big data technology in library in order to provide a basis for further work in future.

II. What is Big Data ?

Big Data is a technique to capture, store, distribute, manage and analyze datasets that traditional data management methods are unable to handle. The term Big data first coined by Laney in his research note. He characterized by three Vs: Volume, Velocity, and Variety.

The first V, refers to the volume of data. In general the size of the data sets of big data is massive compared to regular data. There is no fixed definition for the size, i.e. how big of data will be treated as big data. Hence, the size may 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 petabyte sized data sets. The second V velocity refers to 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. There have been a lot of researches on big data in general and its applications during last a few years. The important significance of research on big data has been well recognized: The big data technology enable us to acquire deeper, more valuable insights from the data and make more timely decisions. The hardware and software for storing and analyzing big data is cheaper and available 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. So much of research work is going on to explore the issues and possibility of big data in library, ProQuest tried to understand the behavior 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-data era because they could create trustworthy data repository to researchers. It is worth to point out that a particular conference on big data at library was held in Georgetown University in 2013 and a lot of issues have been discussed during the conference.

III. Do you consider the Library Data as 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. 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 to analyze 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. The library data also have other properties like..

a. Improper organization of Data

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.

b. **No Standard Data Formats and Data**

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.

IV. Problems identified with Library Big Data

The data exists in the library no doubt is a big data but it is different from other data fields like hospitals and business. Research on Big data particularly in libraries is relatively new. Hence definitely there will be problems in processing of data, transformation, analysis and presentation. The technology used in library big data might be different from that in other areas. One example is that should we create full-test indexes for millions/billions of files to support full discovery in library? In order to apply Big data techniques in library 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 both library data and big data as listed below.

A. **Lacking of Data Analysts**

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. 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, 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. On the other hand, a large set of data often contains some dirty or false data. Therefore, correctly removing those data needs some work. Due to heterogeneous types and formats of research data, integrating them become a very tough job. Many types of research data are considerably less usable when they are in their raw state than after they have had filters or algorithms or other processing performed on them. Those work need budget to build tools and provide other supports as well.

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 concept.

V. What is the use of Library Big data

Businesses are analyzing big data looking for improved ways of selling products and services, what can it do for library? From the DBMS's view, it would make sense for library to find a tool to store the data, another one for indexing it, and yet another to run queries against it. In addition to those traditional features, more functions could be added. For example, with library big data, we might be able to advise local business owners seeking market information, help students run statistics for a project, help researchers to manage large datasets effectively. In the library two aspects of data mining could be achieved: one is using data stored in the library and another is using the data collected during the process when users use the library service. Some of those are listed as below.

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. Based on the data, the decision could be more useful. For example, 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 user satisfaction by providing better service, and efficient usage of library resources.

B. **New Data Format**

Resource Sharing is one of the important goals of library. For example, OCLC (Online Computer Library Center) 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. 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**

From a single work, like a research paper, or a book, the relationships from co-authors, citations, geo-location, dates, named entities, subject classification, institution affiliations, publishers and historical circulation information could be easily extracted.

D. **User Behavoir 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.

VI. Conclusion

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. 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.

Big data in library might have less challenge to study, but more challenge to engage with it due to budget and technical issues. The research data are increasing very fast, and more and more researchers wish to use collections as a whole, mining and organizing the information in novel ways. 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.

REFERENCES

- Affelt, A., 2015, The accidental data scientist: big data applications and opportunities for librarians and information professionals, Medford, New Jersey. 9781573875110.

- Bell, S., 2013, Promise and Problems of Big Data, *Library Journal*. March 13.

- Heidorn, P. Bryan, 2008, Shedding light on the dark data in the long tail of science. *Library Trends* 57:2, pp. 280-299.

- Huffine, R., 2015, The Next Frontier: Federal Librarians and Data, Information Today; Jan/Feb2015, Vol. 32 Issue 1, p.1.

- ProQuest, 2014, "Big data": A strategy for improving library discovery. *ProQuest Blogs,*doi:http://www.proquest.com/blog/2013/241902501.html

- Schwartz, M., 2013, What Governmental Big Data May Mean For Libraries." *Library Journal*. May 30.

- Shield, M., 2004, Information literacy, statistical literacy and data literacy. *IASSIST Quarterly.* 28(2), 6-11.

- Teens, M. and Goldner, M., 2013, Libraries' Role in Curating and Exposing Big Data, Future Internet, 5, 429-438.

- Zikopoulos, P., Eaton, C., DeRoos, D., Deutsch, T. & Lapis, G., 2012, *Understanding Big Data*. The McGraw Hill.