Data Analytics and Visualization in Libraries

Rachna Patnaik

Abstract

Libraries play a vital role in supporting research and learning processes. They manage a large store of data about collections to meet the requirements of users. Data in a digital library is growing and data analytics, visualization techniques are trying to make sense of voluminous amount of unstructured data. Space Applications Centre library provides various digital, online services, which generates lot of data in the form of transaction data, metadata and usage data. A large amount of budget is spent in subscription, renewal of e-resources and for accountability; data analytic tools are effective in producing insights from the usage data to understand users and improving library services.

This paper attempts toanalyze library transaction data retrieved from KOHA Integrated library management software and usage data of library website and e-resources at the Library of Space Applications Centre,to make decisions inproviding efficient services by understanding and predicting user needs. The operational data is exported into a warehouse, cleaned and analyzed using Python language and its libraries. Matplotlib a data visualization tool is used to visualize the analysis through scatter plot and bar graph and helped us in the renewal of e-resources.

Keywords: Data, Data Analytics, Data Warehouse, Information, Matplotlib, Predictive Analytics, Python, Visualization

1. Introduction

The role of libraries has changed in the digital environment and to meet the need of users new applications and services have been developed. Each of the elements of a digital library, its collections and services play an important role in determining how the library is used and what impact it will have on users. The never-ending cycle of data generation followed by the increasing need to store, retrieve and analysis of data has been a challenge faced by professionals for decades and it has resulted in the development of powerful tools. Data analytics is one of the powerfultechnique that deserves special attention because of its capability to discover and analyze raw data from the databases (Wikipedia).The users access the services and leave a trail, which websites log as usage data. Data analytics involves many processes that include extracting data and categorizing it in order to derive various patterns, relations, connections and other such valuable insights to arrive at some decision. Data visualization in libraries offers the ability to manage, organize and present data in a graphical or pictorial format collected from a number of sources like transaction logs, e-resources usage and usage of other digital services, which helps in making quick real-time decisions.

2. Data Analytics and Visualization

Data can be something simple, yet voluminous and unorganized. It has no meaning, structure or relationship, so this raw data cannot be used for makingdecisions. Data comes in many forms such as database tables, spreadsheet format, text files, etc. Information is processed data; data becomes information when we add a relationship. Knowledge is the combination of data and information to which experience and expert opinion are added for the benefit of the organization.

Data analytics refers to the set of quantitative and qualitative approaches to derive valuable insights from data. Data mining is a particular data analysis technique that focuses on the extraction of hidden predictive information from large databases rather than purely descriptive purpose. Data mining tools predict future trends and behaviours, allowing proactive, knowledge-driven decisions in an organization.Predictive Analytics is another approach that is beneficial in libraries as it unleashes the data and allows predicting the future. It is the process of collecting, cleaning, transforming and modelling datasets with the help of specialized software fordiscovering useful information, conclusions and supporting decision-making.

Visualization is an art of converting data sets into simple graphs that brings out unseen patterns and connections. Large datasets need an effective way to understand and data visualization helps to understand them quickly. The primary objective of data visualization is to gain insight (hidden truth) into data or information. Some visualization methods are a bar graph, pie chart, histogram, scatter plot etc.

2.1. Benefits of data visualization

1. It simplifies complex quantitative information.
2. It helps analyze and explore big data easily.
3. It helps in identifying areas that need attention or improvement.
4. It identifies the relationship between data.
5. It explores new patterns and reveals hidden patterns in the data.

3. Data warehouse

A data warehouse is a repository of historical data. A data warehouse can perform complex queries and analysis on the information without slowing down the operational systems.A data warehouse is "a copy of transaction data specifically structured for query and analysis" (Kimball, 1996). The data in a data warehouse is static and once entered remains unaltered in any shape or form as it contains a copy of transaction data rather than the actual record generated by the original transaction. Data is entered into the data warehouse from the operational environment and the purpose of the data warehouse is querying, analysis and reporting for the exploration of new relationships, trends and hidden values. It contains raw data from existing databases for supporting queries of management. Thus, the data in a data warehouse contains a cleaned version of the operational data reformatted for analysis.

4. Data Analytics in Libraries

Libraries realized the importance of data analytics to arrive at a realistic decision, based on existing data and information. Data is collected from the operational server as transaction logs, log files etc. and stored in a data warehouse. The librarian writes queries to extract data from the operational data, cleans the data to remove noise and correct inconsistencies and writes the resulting records into either a flat file or a relational database in a structured

format designed specifically for analysis. Once this procedure has been tested, it can be automated to pull data from the operational systems into the data warehouse on a regular basis. While an integrated librarysystem (ILS) is optimized for processing transactions (circulation, purchase, cataloguing, etc), a data warehouse is optimized for analysis. To facilitate complex analysis and visualization, the data in a warehouse is typically modelled multi-dimensionally and data science algorithms can perform analysis on this historical data. Figure 1 shows the architecture of a data warehouse.

Library data consists of catalog, transaction and usage data. Catalog data means bibliographical data while transaction and usage data is generated by usage of services. Transaction data can be utilized in the analysis of book borrowing record data i.e. analyzing borrower's historical records which deliver great value for library users to read their favourite book.

Over the years the online collection has increased and some critical decisions librarians need to take regarding retaining of the resources, subscribing/ unsubscribing to new resources. Manually studying the usage of library resources is tedious but the dataanalyticsprocess is effective in producing insights from the library usage and visualization techniques are efficient to summarize big data.

Analytical tools can be applied in libraries to imagine the future library from it and will have sufficient information to make right decisions about which kind of books they are going to purchase, what kind of magazines have to be ordered in the coming renewal period of the year or what type of e-journals to be renewed in future. Libraries manage a large database of its documents and a data warehouse can be created with transaction logs (Chen) day, month and year wise to find the day of the week on which maximum check-out was done, or which subject was maximum referred in the last two years. The library may procure a book, which is not referred by patrons so a decision has to be taken to transfer that book to other ISRO/DOS libraries to ensure its effective use.

Libraries share collections across institutional boundaries, but data mining across library collections could open the door to new opportunities for shared collection management and reduce cross-collection redundancies and free up resources to fill gaps in collections.

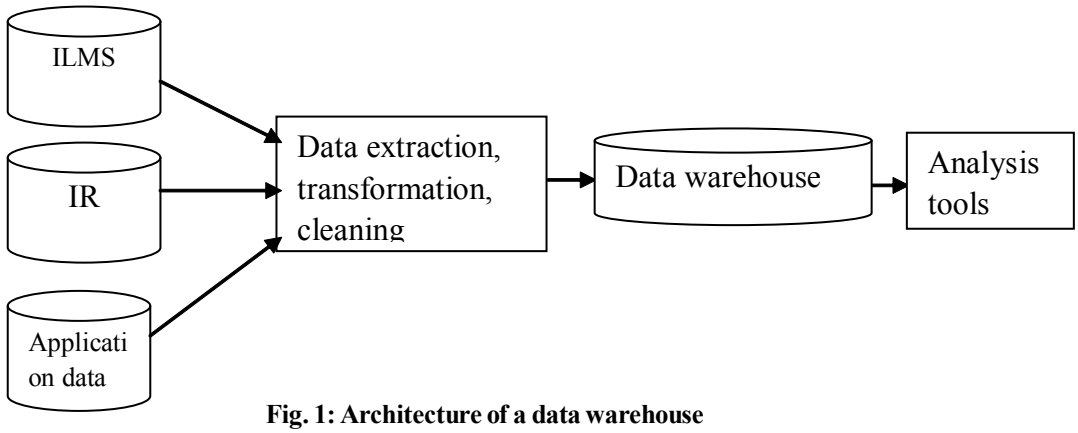

Fig. 1: Architecture of a data warehouse

5. Data Analysis at the Library of Space Applications Centre

Everyday libraries generate transactions that need to be analyzed in many different ways. From the data warehouse, we can get a lot of worthwhile information immediately such as the maximum check-out books on a subject year wise. Data sources, such as transaction logs, circulation data, ILL transactions, reference transactions, and more, can be mined for information that reveals user behaviour and enhances the value of library collections to users. From the circulation transaction log, it is possible to get fast-moving titles and slow-moving titles which is very beneficial to make future decisions regarding acquisition of documents.

Analyzing the data consists of transforming the already summarized data found in the data warehouse into information that can produce useful results. The steps carried out for analysis are:

1. Selecting the data: The first step is to select the data from the data warehouse and transform itinto a particular format.

2. Cleaning and transforming the dataset: The dataset comes with many data quality challenges. The dataset should be pre-processed with various techniques to prepare for analysis and visualization.

3. Analyzing the dataset: For analyzing data we have used Python language, Pandas a Python Data Analysis Library, Numpy a tool for mathematical calculations in Python environment and matplotlib, a data visualization library built on Numpy framework. Python has a number of libraries especially for running the statistical, cleaning and modelling chores. Panda's package facilitates reading of a dataset in a usable format. Matplotlib is a multi-platform data visualization tool that processes high-quality graphics and plots such as histograms, bar charts, pie charts, scatter plots.

The data structure available in Python is a data frame which has rows and columns and delimiter separates columns of a dataset.Datasets of python come in .csv format. Jupyter is a platform for Python development and allows writingprograms in a web browser.

The data is extracted from Library Management Software of most circulated itemsfor the period 1/1/ 2017 to 15/07/2019 and is analyzed using Python and its libraries.Figure 2 shows a snapshot of the Excel file retrieved from KOHA software The scatter plot graph is generated using Matplotlib, whichhelped to make the decision of purchasing multiple copies of the title. Figure 3 shows excerpt of data analysis using python code.

	A	B
	Title	Total
	Microwave Remote Sensing: Active and Passive	368
	Fundamentals of Remote Sensing	316
	IEEE Trans. on Geoscience & Remote Sensing	273
	Fundamentals of Remote Sensing	244
	International Journal of Remote Sensing	222
	IEEE Trans. on Microwave Theory & Techniques	196
	Remote Sensing of Environment	177
	Young Scientist	158
	Young Scientist	128
	Microwave Filters	117
	Physics	113
	Premchand; sampurna kahaniyan.	111
	IEEE Trans. on Antennas & Propagation	105
	Global positioning system: theory and applications	89
	World atlas of coral reefs	88
	Satellite Meteorology:	87
	Design of Geosynchronous Spacecraft	85
	Remote Sensing: Optics and Optical Systems	80
	Satellite Communications Systems	78
	PE & RS	77

Fig. 2: Excel file of most circulated items from KOHA software

	20/	Title	Total
0	1	Microwave Remote Sensing: Active and Passive	368
1	2	Fundamentals of Remote Sensing	316
2	3	IEEE Trans. on Geoscience & Remote Sensing	273
3	4	Fundamentals of Remote Sensing	244
4	5	International Journal of Remote Sensing	222
5	6	IEEE Trans. on Microwave Theory & Techniques	196
6	7	Remote Sensing of Environment	177
7	8	Young Scientist	158
8	9	Young Scientist	128
9	10	Microwave Filters	117
10	11	Physics	113
11	12	Premchand; sampurna kahaniyan.	111
12	13	IEEE Trans. on Antennas & Propagation	105
13	14	Global positioning system: theory and applicat...	89
14	15	World atlas of coral reefs	88
15	16	Satellite Meteorology:	87
16	17	Design of Geosynchronous Spacecraft	85
17	18	Remote Sensing: Optics and Optical Systems	80
18	19	Satellite Communications Systems	78
19	20	PE & RS	77

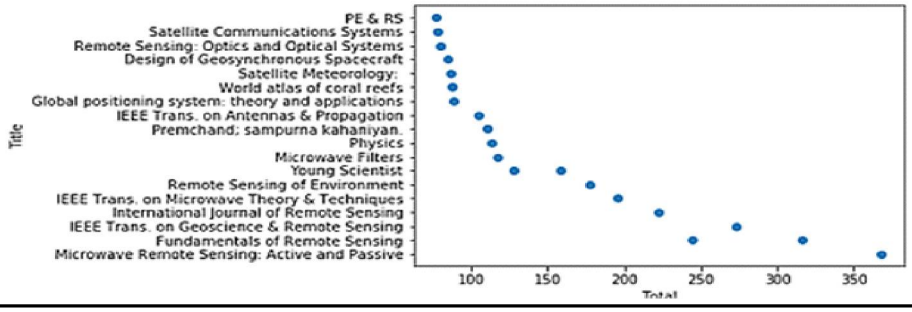

Fig. 3: Data analysis of most circulated items

Librarians arerequired to provide details to higher management with issues likewhich category of users are accessing their services; which collection of resources they are accessing; and how the digital library service can be improved to better serve their users. Data analytics tools can be applied to e-resources usage which is beneficial for predicting future library service. The e-resource data is collected from RemoteXs software and analyzed for predicting renewals of e-resources.SAC Library is subscribing to various e-resources, which are accessible on-campus and off-campus. The usage data of off-campus access is analyzed and it evident that for July 2019 that maximum usage is of IEEE and SPIEe-resources.Figure 4 depicts the usage of e-resources for the month of July 2019 which is retrieved using RemoteXs software.

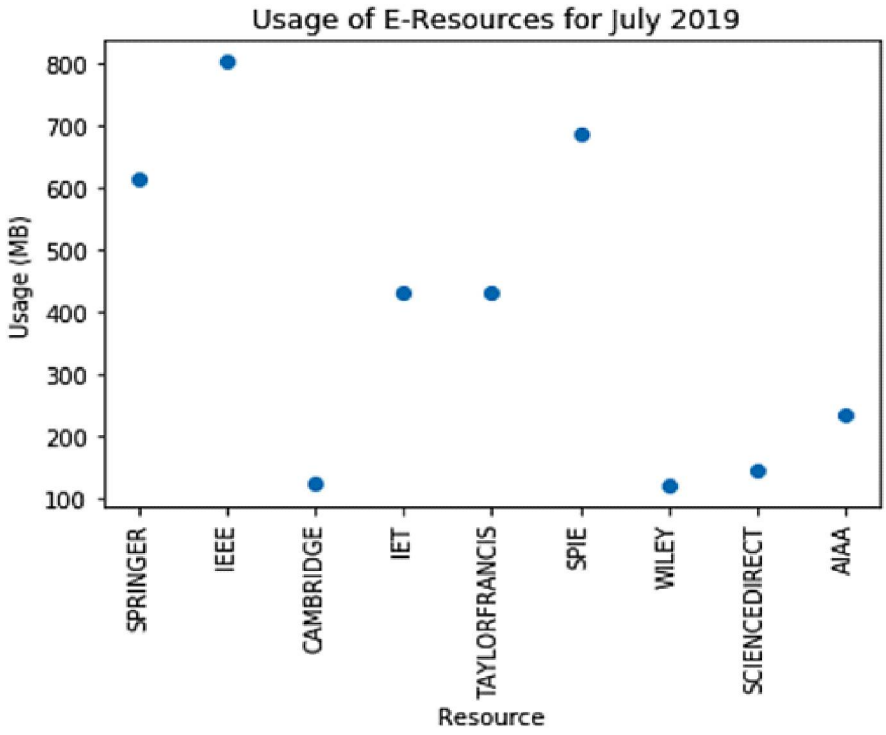

Fig. 4: Data Visualization of E-Resources Usage

Similarly, library website usage data from January to July 2019 is collected and analyzed and a bar graph is plotted for management reporting as shown in figure 5 here.

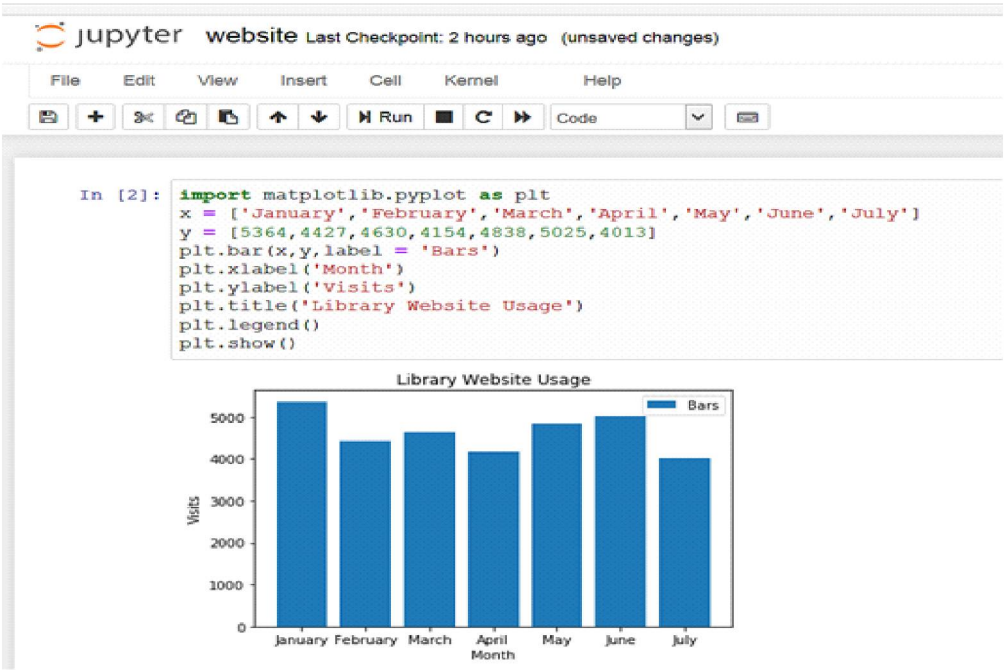

Fig. 5: Data Visualization of Website Usage

6. Conclusion

Data analytics and visualizationare in fact too young to define in any permanent way, but how successfully we implement the concept, will have a great impact and influence among librarians of future generation conceptualize their mission in the digital world. In future, print resources will remain an important component of a library'scollection and an increasing number of resources will be available inonline/digital form only. Since a lot of data will be in digital form data analytics tools can be applied for libraries. Before committing to these technologies on a large scale, libraries need to determine how datamining fits with existing resources and organizational goals. Thus, datamining tools are most beneficial to librariesthat are interested in purchasing digital resources rather than physical materials.

References

1. Anon. Data warehouse from Wikipedia, the free encyclopedia.
2. Kimball, Ralph (1996). The data warehouse toolkit.
3. Chen, Hsin-Liang. Library assessment and data analysis in a big era of data: Practices and Policies.
4. Palmer Joy (2014). Developing a shared analytics service for academic libraries. Insights 27(2), July 2014.

About Author

Rachna Patnaik, Sci/Eng "SG" & Head, Library & Documentation Division, Space Applications Centre, Ahmedabad 380054
Email: rachna@sac.isro.gov.in