International Journal of Librarianship and Administration.
ISSN 2231-1300 Volume 16, Number 1 (2025), pp. 01-08

https://dx.doi.org/10.37622/IJLA/16.1.2025. 1-8

Application of API supported open source full-text integrated search tools in digital transformation of Libraries

Gautam Mukhopadhyay

Librarian, ChandrapurCollege, Purba Bardhaman,West Bengal, India.

Abstract

This paper presents a discussion on the implementation of the open source web information retrieval techniques commonly known as search engines like Apache Solr, Elasticsearch in library digitization. The internet has turned into a global transformation tool. Networks, online web interactions and activities on social media platform opened a new way to access for the users' community because of the accessibility of internet. It revolutionizes in the world of information systems and services. Searching through the big datasets has become a big issue now. It is imperative to apply effective search engines for quick retrieval of information from the large amount of web resources. Times of process, storage and retrieval of large-scale data have become longer over time. Quite a number of search engines have been introduced to retrieve relevant data accurately and quickly. Filtration process gradually becomes a complex performance while manipulating with huge amount of datasets. There are several full-text open source search engines e.g. Apache Lucene, Apache Solr, Elasticsearch,YaCy etc. to process large-scale datasets. Nature and form of resources have been changed with the change of mode of access to them. Generation of resources has been rapidly increased with the advent of Information and Communication Technology (ICT).

Keyword(s): Integrated-search tools, web-based server, search-engines, transformation

Abbreviation(s): ICT (Information and Communication Technology),ILS (Integrated Library System), LSP (Library Services Platform)

1 Introduction

Nature and form of resources have been changed with the change of mode of access to them. Generation of resources has been rapidly increased with the advent of Information and Communication Technology (ICT). The internet has turned into a global transformation tool. Networks, online web interactions and activities on social media platform opened a new way to access for the users' community because of the accessibility of internet. It revolutionizes in the world of information systems and services. Searching through the big datasets has become a big issue now. It is imperative to apply effective search engines for quick retrieval of information from the large amount of web resources. Times of process, storage and retrieval of large- scale data has become longer over time. Quite a number of search engines have been introduced to retrieve relevant data accurately and quickly. Filtration process gradually becomes a complex performance while manipulating with huge amount of datasets. There are several full-text open source search engines e.g. Apache Lucene, Apache Solr, Elasticsearch, YaCy etc. to process large-scale datasets.

This paper presents a discussion on the implementation of the open source web information retrieval techniques commonly known as search engines like Apache Solr, Elasticsearch in library digitization.

2 Objectives

The main objectives of this paper is to

i. To give a brief idea about the various relevant concepts relating to the open source search engines;
ii. To define search engine indexing;
iii. To discuss the needs of the search engines in library digitization;
iv. To highlight the traditional ILS vs sophisticated LSP and
v. To present a brief discussion on the application of the open source information retrieval techniques i.e. search engines.

3 Open source full-text integrated search tool (Search Engine)

Generally, a search engine consists of two essential components. Firstly, the algorithm which

performs a search and returns results and the second one is the database that holds all of those potential results. Open source simply denotes a product which is licensed to permit modifications and redistribution of its source code. Now, open source search engines allow one to change the code of the retrieval and ranking technique. The algorithm programmes (whose source code are open) are such a technique that can be applied for a variety of information retrieval are called Open Source Search Engines or Web Search Engines. Apache Lucene, Solr, Elasticsearch, Xapian, Nutch, Terrier and others are examples of Open Source Search Engines. When we are talking about open source search technology, the widely known open source search engines are Solr and Elasticsearch. Although both of them use Lucence as their indexing and search process. Search engine is a web-based tool that enables particular enquirers to pass a query to a database for their bit of information they need from the internet or World Wide Web. The search engine generates an index on the basis of rankings of the web pages and sites as per their popularity and number of web browsers. In fact, a search engine is a software system that relates and arranges the queries in a systematic order for a bit of information specified in web query term. Information seekers feed their particular search terms on the web search engine and it returns the huge amount of information related to the query or keyword.

Apache Lucene

Apache Lucene is java based high performance search engine with the concept of "a document contains fields of text" at its core logical architecture. It was developed by the Apache Software Foundation. Apache Lucene is an open source search engine software library where 'Apache' is a HTTP server, a widely used open source web server software suite and the term 'Lucene' denotes a full-text java-based search engine and API indexing of documents. Lucene enables the search functionality to a website or application easier by adding content to a full-text index. Lucene stores data by spreading its index across several on-disk files those are organized into logical 'segments' that represents subsets of documents across the corpus.

Solr

Solr stands for Searching On Lucene with Replication. Solr is a HTTP based search application. This is a web-based user interactive search tool and a generally accepted very fast open source search platform built on Apache Lucence. Solr is a process of creating a searchable index that list every terms and documents that consist those terms. It works with Java environment settings. Solr helps to store, index and search full-text data quickly. It has additional features like advanced full-text search capabilities, analytics of the indexed data

and REST-like API search server. Solr is based on several standard interfaces e.g. CSV, XML, HTTP etc. Solr is also capable to adapt users' needs all while simplifying configuration. One of the advantages is that the Solr search server offers real-time indexing to make sure one see his/her content when one wants to see it.

Elasticsearch

Elasticsearch is open source search platform based on REST-like web interface that provides a distributed full-text search engine. It is written in Java language. It offers near real-time scalable search and supports multitenancy. It is largely used for full- text search, business analytics, log analytics, security intelligence. Elasticsearch runs as a standalone system by users of various professions. High availability and massive distribution are the two major features of Elasticsearch.

4 Apache Lucene vs. Solr vs. Elasticsearch

Apache Lucene is a high-performance, multi-faceted, stable, full-featured text java library search engine and building non-relational database (NoSQL) solution. It provides algorithms for creating indexes and performing search and ranking of documents based on query. It offers a new querying model and support document faceting as well. On the other hand, Solr and Elasticsearch are built over Lucene and fully distributed fault-tolerant document-oriented search engines. Both of them are web-servers designed to perform searches with index documents, high-performance and provide other features like faceting. Both are supported by the API end points for indexing and searching queries. Solr can provide enterprise search by connecting to various data sources whereas Elasticsearch provides host of other use cases like machine learning, log analytics etc. For Elasticsearch, setup is easier and it is required to do different REST calls. It has a better documentation and client libraries in different languages.

5 Search Engine Indexing

The index is a central database where the search engine organizes and saves the information stored by the crawlers. Indexing is a process that includes linking the keywords and specific trafficators of a web page to its identifier on the web. The indexing is done in a database which facilitates quick information retrieval process.

The search engine first selects the terms and the indexes a webpage by analyzing the occurrence of the words after filtering out irrelevant words appearing in the content of the visitor's web page. The crawlers analyzed the data based on its quality. If the data is precise and relevant, it finds a place in the index. Another process of indexing is algorithmic process

which considers the relationships of words in the indexed web page. Search engines also provide advanced search options which include filtering and sorting to refine the search results. The ranking of the web pages to be shown in the search results is from most relevant to the least relevant.

6 Application Programming Interface (API)

API is a set of routines, protocols, and tools for building software applications. A software intermediary that makes it possible for software components to interact with one another, leading to the ability to share data over a network. The main importance of API is that this allows the capabilities of one computer programme to be used by another. It means that the APIs are the ways by which different systems, programmes and platforms are able to communicate and carry out wide variety of tasks, for example, better integration, ease of integration, improved services and mechanizing tasks etc. APIs allow information seekers to use and reuse information in new techniques and on a larger scale. In the perspective of digital libraries, the interfaces are enabling attractive new projects and reducing complexities in routine tasks.

7 Need of search engines in digital transformation of Libraries

A search engine is an application that searches for, and retrieves, data based on some criteria, especially one that searches the web for documents having specified terms. It acts as a search tool consisting of programmes that help information users find the information they are searching for using keywords on the web platform. There are different forms and sizes of digital resources in the modern ICT era. As they have been growing in an exponential rate, it has become increasingly important and difficult to obtain the particular data or image. Academic digital libraries gradually realize the utility of full-text search approaches to find the specific information from the large-scale treasure house of data. Existing systems need to be tested on the basis of performance. Institutional digital repository search engines confront big challenges in maintaining continual services. Just a few days ago, relational databases were the main solution for the data processing and retrieval. But they were not programmed for large-scale big data querying. Open source search engines are solution to the new structural design and functionalities including storage, extraction, indexing and filtration of huge amount of data in digital libraries. A search engine needs to be designed that properly stores, indexes and searches data, so that the end user can rapidly access the information. Search engine indexing is a process in which data is collected, parsed and stored in order to support fast and accurate information retrieval [Gonzalez, 2012]. Intuitive and interactive user interface can retrieve information efficiently. For precise and relevant result set of

particular query, it is essential to introduce a ranking system which will be able to sort the specified query by relevance. Different search engines have been trying to find out quite a number of solutions of a set of arisen issues relating to the effective and efficient search results. But it is difficult to select a best solution among the existing ones.

8 Traditional ILS vs Sophisticated LSP

Integrated Library System (ILS) manages mainly print materials. ILS includes several features like i) Acquisition (ordering, acquiring, invoicing physical/printed resources etc.); ii) Cataloging (Processing i.e. classifying and indexing physical materials); iii) Circulation (lending printed documents to users and receiving them back);iv) Presenting reports via SQL; v) Serials control (tracking journals); vi) Advanced Search (keyword as phrase) and OPAC (Public interface for users).

Library Services Platform (LSP) denotes a type of library resource management system with a set of new sophisticated features that differ from the genre of integrated library system. A variety of descriptions is used for this set of new products such as web-scale management solutions, uniform management systems, or merely services platform. Marshall Breeding coined the term Library Services Platform (LSP) in 2011 to describe a new set of library software that are making a different approach to library resource management in comparison with integrated library systems. He described the term "The library services platform in general will replace multiple incumbent products, including the integrated library system, any formal or informal products or processes to managing electronic resources, and knowledge bases of e- content resources" (Breeding, 2015). Some of the new set of library projects have already been launched by the vendors or providers, for example, Alma by Ex Libris (January, 2011),World Share Management Services by OCLC (July, 2011), Sierra by Innovative Interfaces (2012), Open Skies by VTLS (2012), Open Library Environment (OLE) by Kuali in 2013 (presently, FOLIO),and Intota by Serials Solutions (2013).

The fundamental difference between the older ILS offerings and the library services platform is that the ILS products were predominantly designed around the management of print resources. In the age of digital era, the ILS products are unable to be reconfigured well enough to smoothly and efficiently operate the integration of

all the workflows having different features, for both print as well as digital. Notably, the traditional ILS does not have the feature of new cloud computing technologies and architectures whereas LSP serves as a digital public interface. Different technological approaches we may have seen in the browser-based web-scale services platform.

9 Conclusion

Implementing and proper integration of searching tools into the ILS software is challenging. There are many reasons why libraries might not want to implement the newly emerging search offerings. The reasons may be the limited capacity, self- hosting not available, data privacy etc. Open-source search engines are not so easy to use because they offer data privacy, freedom and are self-hosted. Apache Solr can search through the full-text of your documents, including the title, body, and metadata. It is distributed search and one of the features is real-time indexing.. Moreover, Solr can be made use of in a distributed cluster to scale to meet the needs of big libraries. It also comes with an Apache 2.0 license. Meilisearch is designed to be simple to implement and integrate as it is customizable and powerful. Elasticsearch is known for its scalability, speed and ease of use. Quite a number of academic institutions and other organizations including e-commerce sites and government agencies use Elasticsearch. However, some open source search engines offer numerous features that make it a suitable choice for a variety of applications, including accuracy, speed, robust, customization, and ease of use and so on. For example, Swirl is open-source software that uses AI to simultaneously search multiple content and data sources. Open Search is an open-source, community-driven search which offers a highly scalable system for fast access and interact with huge volumes of data supported by an integrated visualization tool, Its OpenSearch Dashboards feature makes it possible to explore data by the users at ease.

References

[1] Sweller, John (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science*, 12(2); 257–85.

[2] Merriënboer, Jeroen J. G. van and Paul, Ayres (2005). Research on cognitive load theory and its design implications for e-learning. *Educational Technology Research and Development*, 53(3);5–13.

[3] Schnotz, Wolfgang and Kürschner, Christian (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, 19(4); 469–508.

[4] Hollender, Nina et al. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior*, 26(6); 1278–88.

[5] Condit Fagan, Jody (2010). Usability studies of faceted browsing: a literature review. *Information Technology & Libraries*, 29(2); 58–66.

[6] Yang, Sharon Q and Hofmann, Melissa A (2011). Next generation or current generation? a study of the OPACs of 260 academic libraries in the USA and Canada. *Library Hi Tech*, 29(2); 266–300.

[7] Gonzalez,ZGet al. (2012). International BusinessMachineCorporation, Search Engine Indexing. U.S. Patent Application; 13/713, 765.

[8] Breeding, Marshall (2015). Library Services Platform: a maturing genre of products. *Library Technology Reports*, 51(4); 40.

[9] Osborne, Hollie M and Cox, Andrew (2015). An investigation into the perceptions of academic librarians andstudents towardsnext-generationOPACs andtheirfeatures. *Program:Electronic LibraryandInformationSystems,* 51(4); 2163.

[10] Comeaux, David J (2017). Web design trends in academic libraries: a longitudinal study. *Journal of Web Librarianship,* 11; 1–15.