Open-Source Search Engines

S. M. Vahidipour

University of Kashan

- Elasticsearch is a highly scalable open-source full-text search and analytics engine based on Lucene.
- It is developed in Java and provides a distributed, multi-tenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.
- Using Elasticsearch you can store all kinds of documents, search, and analyze big volumes of data quickly and in near real time.

- **Programming language:** Java, but have client available in .NET (C#), PHP, Python, Apache Groovy and many other languages.
- **License:** Apache License 2.0
- **Ranking of search results:** Relevance Scoring

- **Indexing style:** Elasticsearch is distributed, which means that indices can be divided into shards and each shard can have zero or more replicas.
- Re-balancing and routing are done automatically.
- Related data is often stored in the same index, which consists of one or more primary shards, and zero or more replica shards.
- Once an index has been created, the number of primary shards cannot be changed.

- Complete Elasticsearch Masterclass with Logstash and Kibana
- Complete Guide to Elasticsearch
- ElasticSearch, LogStash, Kibana ELK #1 - Learn ElasticSearch

- It is an open source enterprise search platform programmed in Java to provide full-text search, real-time indexing, hit highlighting, dynamic clustering, faceted search, database integration, and rich document (e.g., Word, PDF) handling.
- Since it is based on Lucene, Solr extensively uses the Lucene's Java search library at its core for full-text indexing and search, and has REST-like HTTP/XML and JSON APIs support, thus making it usable from most popular programming languages.
- It is designed for scalability and fault tolerance. Solr's external configuration allows it to be tailored to many types of applications without programming in Java, and it's plugin architecture helps support more advanced customization.

- **Programming language:** Java
- **License:** Apache License 2.0
- **Ranking of search results:** It is based on td.if scoring model as used in Lucene, which involves various factors like Term Frequency, Inverse Document Frequency, Coordination Factor and Field length.

- Getting Started with Enterprise Search Using Apache Solr (Rated 4 / 5 by 271 students)
- Learn Apache Solr with Big Data and Cloud Computing (Rated 3/5 by 119 students)

- Lucene is one of the more established open source search engines out there with a text search engine library that is written purely in Java. Its indispensable software can be used for any application that requires full-text search.
- Lucene can be used across platforms, has a configurable storage engine (Codecs) and has many powerful query types such as proximity queries and phrase queries.
- At the moment, its open source project is available for free download and Twitter is actually using Lucene for real time search.
- **Programming Language:** Originally Java but ported to other languages such as: Delphi, Perl, C#, C++, Python, Ruby, and PHP.
- **License:** Apache Software Foundation.
- **Ranking of search results:** Versatile (follows popular choices).
- **Indexing style:** multiple-index searching with merged results.
- Even in 2017, Lucene is still preferred by some employers.
- Click here to see the **demand for lucene** in your location.

Sphinx

- Sphinx is an open source full text search server that is programmed with relevant search quality and integration simplicity.
- Sphinx allows flexible testing whereby its indexing features include full support for SBCS and UTF-8 encodings, stopword removal and optional hit position removal (hitless indexing); morphology and synonym processing through word forms dictionaries and stemmers; exceptions and blended characters; and many more.
- Sphinx has an easy application integration that is derived from 3 different APIs.
- It has a native library for many programming languages, a pluggable storage engine for MySQL and an application query that uses MySQL client library and syntax.
- Websites such as Craigslist, Living Social, MetaCafe and Groupon has adopted Sphinx for its searches.
- **Programming Language:** C++
- **License:** GPLv2 and commercial
- **Ranking of search results:** Versatile
- **Indexing style:** SQL database indexing and Non-SQL storage indexing.
- To know who is hiring for Sphinix near you, click here.

Xapian

- Xapian, termed as an open source probabilistic information retrieval library, provides a full text search engine library for programmers.
- It possesses a wide range of structured Boolean search operators which are allocated based on probabilistic weights. There are also Boolean filters to restrict a probabilistic search.
- Xapian's search engine also has the dexterous ability to support the search's word synonyms explicitly and as an automatic form of query expansion.
- Currently, Xapian is used as a search engine for the Library of the University of Cologne and Die Zeit (A popular German newspaper)

- **Programming Language:** C++
- **License:** GNU General Public License
- **Ranking of search results:** Flexible (important words become more probable than unimportant words)
- **Indexing Style:** Filing system

- Indri is an open source search engine that prides itself through its state-of-the-art text search and a rich structured query language for text collections of up to 50 million documents (single machine) or 500 million documents (distributed search).
- Indri is multi platform and is applicable in Linux, Solaris, Windows and Mac OSX.
- Indri is can support UTF-8 encoded text and is able to parse PDF, HTML, XML and TREC documents. It also recognizes text annotations.
- One of Indri's significant involvements is being the search engine component of the Lemur toolkit. The Lemur toolkit came from the partnership between the Center for Intelligent Information Retrieval and the Language Technologies Institute at Carnegie Mellon University. The partnership between the 2 institutions developed the Lemur Toolkit, an open-source (BSD license) software framework for building language modeling and information retrieval software.
- **Programming Language:** Java, PHP, or C++
- **License:** BSD style license
- **Ranking of search results:** Versatile (Explicit term weighting and Robust query language)
- **Indexing style:** Flexible indexing with tokenization

Zettair

- Written and designed by the Search Engine Group at RMIT University, Zettair is a compact and fast text search engine which allows you to index and search HTML (or TREC) collections. It also formatted for simplicity as well as speed and flexibility, and one of its fundamental features is the ability to handle large amounts of text.
- Other features that Zettair has are its Boolean, ranked and phrase querying, Modular C API and it's easy to use command-line interface. Not to mention the search engine is applicable for many platforms including Solaris and Linux.

- **Programming language:** C
- **License:** BSD style License
- **Ranking of search results:** simple and straightforward
- **Indexing style:** Single Executable (when an index doesn't exist, Zettair will create one for you based on the parameters you provide)