Homework: Search engines exercise

Carlos Andrés Reyes Evangelista - 157068
Ingeniería en Sistemas Computacionales
Universidad de las Américas Puebla

Table of contents

- Exercises
  - 1.3 Open source search engines
    - 1. Apache Lucene Core
    - 2. Sphinx
    - 3. Datapark Search Engine
    - 4. Elastic Search
    - 5. Xapian
  - Web-services search system
    - 1. Udemy
    - 2. KDE Store
    - 3. ArchLinux User Repository
    - 4. Pinterest
    - 5. Google Codelabs

Exercises

1.3 Use the Web to find as many examples as you can of open source search engines, information retrieval systems, or related technology. Give a brief description of each search engine and summarize the similarities and differences between them.
Go to exercise

1.4 List five web services or sites that you use that appear to use search, not including web search engines. Describe the role of search for that service. Also describe whether the search is based on a database or grep style of matching, or if the search is using some type of ranking.
Go to exercise

1.3 Open source search engines

1. Apache Lucene Core

Apache Lucene Core is a cross-platform open source search engine completely based on Java but with implementations in other languages available. It uses pluggable ranking models such as the Vector Space Model and Okapi BM25. With Lucene it is possible to search and sort by fields (title, author, type), it is available under the Apache License so it is free for its use in Open Source and commercial programs. It supports some query types including phrase queries, wildcard queries, proximity queries and range queries.

2. Sphinx

Sphinx is an open source information retrieval software library written in C++, it works for almost any operative system. Its principal purpose is to index contents of a database whose can be retrieved by using SphinxQL, a subset of SQL. Among its principal advantages there is its indexing speed and its native support for certain Database Management System. Its ranking system uses additional factors to BM25.

3. Datapark Search Engine

Datapark Search Engine is a full-featured search engine designed to organize search within a website, group of websites, intranet or local system. It has support for http, https, ftp and other URL schemes. It can index plain and formatted (html, xml) text documents, audio and images types of files natively. It supports query expansion based on editable dictionaries for each language and charset. Its popularity ranks are based on a neural network model (Neo) and vector calculation (Goo).

4. Elastic Search

Elastic Search it is another open source search engine published under Apache License, it is based on Lucene and it is written in Java. It works with an RESTful interface and JSON documents. Among its advantages, its speed and scalability stand out, along with its compatibility with any Java-compatible environment. According to its official website, its speed is obtained thanks to the implementation of «inverted indices with finite state transducers for full-text querying, BKD trees for storing numeric and geo data».

5. Xapian

Xapian it is written in C++, but has bindings to allow its use in some other programming languages. It supports phrase and proximity queries, query expansion, autocorrect misspelled words and has a full range of structured boolean operators. It can index an huge number of file formats. It allows for new documents to be searchable right away. Its ranking system uses Probabilistic, Divergence from Randomness, and Language Modelling families of weighting models.

Web-services search system

1. Udemy

Udemy is an online learning platform, its search function allows the users to seek for courses of their interest by typing keywords (mostly topics) about what they want to learn.

Udemy searches for courses within its database, but it also uses ranking models so that the courses that are more likely relevant to the query are shown first depending on the user language, expected or more recent version, match with tags and topics, among others.

2. KDE Store

KDE Store is a platform to publish, share and download open source stuff for the Linux Desktop Environment KDE Plasma. Its role search allows the user to quickly find any add-on of any category by typing key words corresponding to the category, name of the package or expected features.

It searches for coincidences in the package name, description, category or even files and reviews. It ranks them by giving more weight to packages with the exact name and a best community score overall.

3. ArchLinux User Repository

The AUR it is a repository promoted by users of the Arch community, it contains package's descriptions to build source files in Arch Linux. Its search role is to locate the packages that corresponds or matches the keywords typed.

Since the goal is to reach specific packages, it seeks for matches in the whole repository in a grep style, the name of the package is the more important field to look at, but it also searches for coincidences within the description field. It only uses ranking based on statistical properties.

4. Pinterest

Pinterest is a platform that allows users to create and manage virtual boards where they can store and classify multimedia content with its link to the original website. The use of the search functionality will look over the web to retrieve the *pins* more suitable for the query.

The Pinterest algorithm prioritize pins based on their quality -which is obtained by the popularity and engagement of the pins among other factors-, displaying a series of more accurate images (representing a pin each) with the link of the site they come from available to be clicked and redirected. The more *quality pins* a domain has, the more likely it is to appear in new queries.

5. Google Codelabs

Google Codelabs is a platform that provides small, quick and effective tutorials and guides to get "hands-on coding experience". The role of search of this site is to limit and display results, tutorials and guides that contain any of the keywords or belong to a similar topic.

To retrieve the results of the search, Google Codelabs apply a grep-style of matching with the titles of every course, it is simple to discernate this behavior since the search is performed as soon as the user modifies the search field, so it is visible that a result disappears as soon as a result stop matching the text written in the seach field.