¿Qué es solr?

Solr es un motor de búsqueda basado en Apache, escrito en el lenguaje de programación Java y basado en la librería de Java Lucene, que permite integrar motores de búsqueda verticales.

Un motor de búsqueda es por ejemplo Google o Yahoo (son motores de búsqueda generales porque buscan contenido en toda la web).
Un motor de búsqueda vertical es aquel que busca un tipo de contenido en especifico, es decir, que los motores de búsqueda verticales devuelven información en mayor cantidad y los verticales devuelven la información especifica

- **Por poner un ejemplo (NETFLIX)**, si tuviéramos un blog dedicado al cine, Solr nos permitiría encontrar una película determinada dentro del propio blog introduciendo el título de una película o el nombre de alguno de sus actores. Dicha búsqueda la llevaría a cabo recorriendo los documentos, ya sean de texto ya sean bases de datos, que forman parte de la web.
  Así pues, con **Solr** podremos fácilmente crear un **motor de búsqueda** para llevar a cabo búsquedas en webs y bases de datos.

En otras palabras, Solr es una base de datos NoSQL que se utiliza para almacenar datos y consultarlos casi en tiempo real.

Las siglas de **SOLR** son: **S**earching **O**n **L**ucene **R**eplication

En definitiva, nos dice que Solr es sistema que nos da los resultados de una búsqueda basado en la librería Lucene, que es una librería muy utilizada para motores de búsqueda.

Cuales son las características de las busquedas con Solr

Las características principales de las búsquedas con Solr son:

- Filtrado de las búsquedas -> el filtro es una herramienta de búsqueda que permite al usuario restringir su búsqueda. Por ejemplo, en Netflix cuando estamos buscando una película o una serie podemos hacer un filtrado de la búsqueda poniendo el tipo de película/serie que queramos ver: comedia, terror, amor… etc

- Búsqueda por facetas, por el que Solr nos hará sugerencias de filtrado (se conoce como facetado al proceso de la funcionalidad de categorizar los resultados, siendo cada categoría una faceta Esto permite que el usuario pueda descubrir nuevos elementos y afinar los resultados de la búsqueda)

- Clasificación de los resultados de búsqueda -> que consiste en presentar los resultados de búsqueda ordenados por algún criterio (por ejemplo en una tienda de electrónica queremos que los resultados aparezcan en función de la puntuación que les han dado los usuarios)

Destacar también que detecta diferentes idiomas, es capaz de agrupar por términos relacionados de manera automática, permite hacer JOINS (mezcla entre el producto cartesiano y la selección), permite reajustar los términos (es decir, si detecta errores ortográficos en la entrada proporciona una salida con sugerencias con alternativas corregidas, por ejemplo, cuando te equivocas en google buscando) y permite las búsquedas por comodín “?” Para un solo elemento (ejemplo: l?na proporciona de resultado luna, lana, Lena….) y “*” para varios caracteres (ejemplo: prof*, la lista de resultados incluirá todos los términos con esta raíz (profesor, profesorado, profesión)

Cómo funciona Solr

Solr funciona recorriendo los documentos seleccionados e incorporándonos a un índice. Este proceso se llama indexado.

El indexado en Solr sería semejante a crear un índice al final de un libro que incluya las palabras que aparecen en dicho libro y su ubicación, de manera que básicamente llevaríamos un inventario de las palabras que aparecen en el libro y un inventario de las páginas donde aparecen dichas palabras. Este tipo de índice, denominado **índice invertido** , es una forma de estructurar la información que va a ser recuperada por un motor de búsqueda.

En un índice invertido, el buscador crea los índices, o términos de búsqueda, a partir de una serie de documentos, indicando los documentos concretos que los contienen. De esta manera, cuando el usuario teclea un término de búsqueda determinado, el motor de búsqueda creado con Solr le indicará dónde aparece dicho término.
Un índice de Solr acepta datos de muchas fuentes, tales como archivos XML, CSV, archivos Word o PDF.

Así pues, el indexado con Solr consiste en añadir las palabras clave de los documentos que hayamos indicado al índice de Solr.

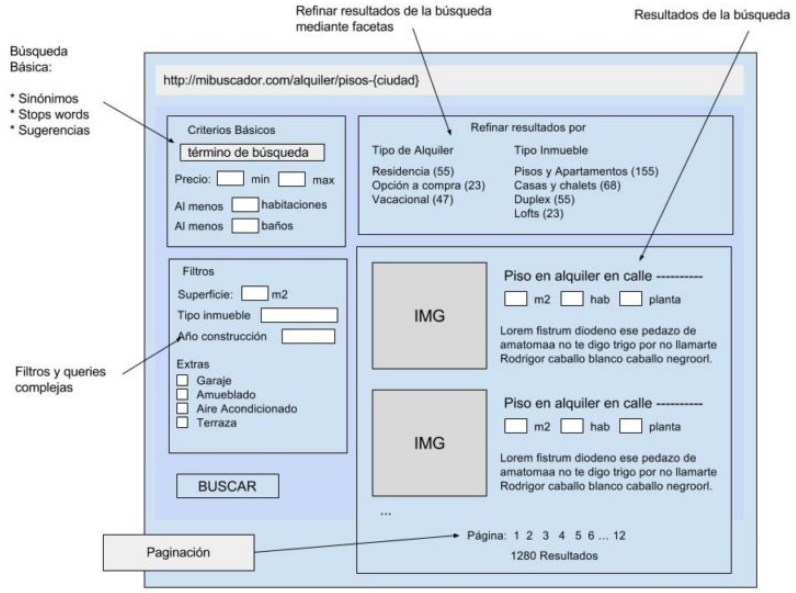
Solr en lugar de buscar en el texto mismo, realiza la búsqueda de la palabra clave buscada en el índice, y a continuación nos indica en qué documentos se encuentra dicha palabra clave
Este tipo de indice se llama indice invertido porque la estructura de los datos se basa en las palabras clave en lugar de basarse en la página.

- Entonces el funcionamiento es el siguiente:

  Se realiza la **indexación** y antes de añadir los datos al indice pasan por un **analizador** de campos donde se filtran y tokenizan para facilitar el proceso de búsquedas. Así se descomponen los campos en tokens que se pueden buscar.
  En la fase de **mapeo**, Solr asigna las consultas del usuario a los documentos almacenados en la base de datos para encontrar los resultados adecuados.
  Por último, se **clasifican** los resultados según la puntuación de relevancia, de forma que los más relevantes aparecen en la parte superior.

EJEMPLO DE FUNCIONAMIENTO

Empezaremos con una aplicación de búsqueda “típica” de pisos en alquiler, e iremos viendo el funcionamiento de SOLR pasito a pasito:



En nuestro buscador, el usuario ingresa un término de búsqueda y esto produce unos resultados que SOLR puede agrupar en categorías y filtrar por otros criterios de búsqueda. A esta funcionalidad de categorizar los resultados se le conoce como “facetado”, siendo cada categoría una faceta. Esto permite que el usuario pueda descubrir nuevos elementos y afinar los resultados de la búsqueda. Para el ejemplo de la imágen, los resultados están facetados por **Tipo de Alquiler** y **Tipo de Inmueble**.

Algunas paginas que usan Solr: Netflix, La NASA…. Etc

Funciones de Solr

Reajuste de términos también para grupos de palabras: el sistema detecta errores ortográficos en la entrada y proporciona resultados para una alternativa corregida.

Joins (uniones): una mezcla entre el producto cartesiano (en la búsqueda se consideran varios términos en cualquier orden) y la selección (solo se muestran los términos que cumplen una determinada condición), es decir, una sintaxis compleja de variables booleanas.

Agrupación de términos relacionados temáticamente.

Clasificación en facetas: el sistema clasifica cada elemento de información individual según varias dimensiones. Por ejemplo, a un texto lo asocia con palabras clave como el nombre del autor, el idioma y la longitud del texto, junto a los temas de los que trata el texto, así como a una clasificación cronológica. La búsqueda en facetas permite al usuario utilizar varios filtros y obtener así una lista de resultados personalizada.

Búsqueda con “comodín”: ¿un carácter representa a un elemento indefinido o a varios de estos elementos en una cadena? Para un carácter se utiliza “?” y para varios “*”. Por ejemplo, si se introduce el fragmento de una palabra más el comodín para varios caracteres, como prof*, la lista de resultados incluirá todos los términos con esta raíz (profesor, profesorado, profesión), de modo que el usuario obtiene resultados para estos temas. La relevancia resulta de la delimitación del tema de su biblioteca o de otras delimitaciones de búsqueda. Por ejemplo, si los usuarios buscan "l?na", obtendrán resultados como luna, lana o lena, pero no, en cambio, palabras como liana o leona, ya que “?” solo sustituye a una letra.

Reconocimiento de texto en muchos formatos, desde Microsoft Word hasta PDF y contenido enriquecido indexado, pasando por editores de texto.

Detección de diferentes idiomas.

Bibliografía

https://solrtutorial.es/que-es-solr.html

https://aprenderbigdata.com/solr/

https://www.ionos.es/digitalguide/servidores/configuracion/apache-solr/

https://solrtutorial.es/indexado-solr.html