

Agenda

1. User Interface [Kevin]
2. Information Retrieval Process [Pritom]
3. Indexing & Searching [Petros]
4. Ranking Methods [Kevin]
5. Performance & Example Queries [Oliver]
6. Challenges [all]
7. Future Improvements [Oliver]

Input Data Model

1. Title [String]
2. Date [Date]
3. Views [Integer]
4. Tags [String[]]
5. Related Question [String[]]
6. Question
 1. Question [String]
 2. Code [String]
 3. Votes [Integer]
7. Answer
 1. Answer [String]
 2. Accepted Answer [Boolean]

The screenshot shows a Stack Overflow page for the question "Comparison of full text search engine - Lucene, Sphinx, Postgresql, MySQL?". The question is highlighted with a red box and includes a list of candidates (Lucene, Sphinx, Postgresql, MySQL) and selection criteria (result relevance, speed, ease of use, resource requirements, scalability). The question has 302 votes and was asked 10 years ago. Below the question, there are 8 answers. The top answer, by user 'ajreal', is highlighted with a red box and has 162 votes. It provides detailed information about Sphinx and its advantages over other engines. The right sidebar shows a list of linked and related questions, with the top one being "Full text search in java web application".

Comparison of full text search engine - Lucene, Sphinx, Postgresql, MySQL?

I'm building a Django site and I am looking for a search engine.

A few candidates:

- Lucene/Lucene with Compass/Solr
- Sphinx
- Postgresql built-in full text search
- MySQL built-in full text search

Selection criteria:

- result relevance and ranking
- searching and indexing speed
- ease of use and ease of integration with Django
- resource requirements - site will be hosted on a [VPS](#), so ideally the search engine wouldn't require a lot of RAM and CPU
- scalability
- extra features such as "did you mean?", related searches, etc

Anyone who has had experience with the search engines above, or other engines not in the list -- I would love to hear your opinions.

EDIT: As for indexing needs, as users keep entering data into the site, those data would need to be indexed continuously. It doesn't have to be real time, but ideally new data would show up in index with no more than 15 - 30 minutes delay

mysql postgresql full-text-search lucene sphinx

asked Apr 10 '09 at 10:38

26 MySQL full-text search and transactions are (presently) mutually exclusive. MySQL fulltext indexes require the MyISAM table type, which doesn't support transactions. (As opposed to the InnoDB table type which supports transactions, but not fulltext indexes.) - [Carl G](#) Jan 31 '10 at 20:59

1 PostgreSQL full-text search, [Tsearch](#) does not support phrase search. However, it's on the TODO list [su.msu.ru/~megera/wiki/FTS_Todo](#) - [Gnanam](#) Dec 9 '10 at 14:13

1 Anyone looking at this for Django should checkout the haystack app, [haystacksearch.org](#) - [Keyo](#) Jul 2 '11 at 4:56

4 [slideshare.net/billkarwin/...](#) - [Aoute](#) May 7 '12 at 8:40

23 @CarlG, Just for everybody's reference, MySQL 5.6+ has Full text search support with innodb engine - [DhruvPathak](#) Dec 18 '12 at 13:36

8 Answers

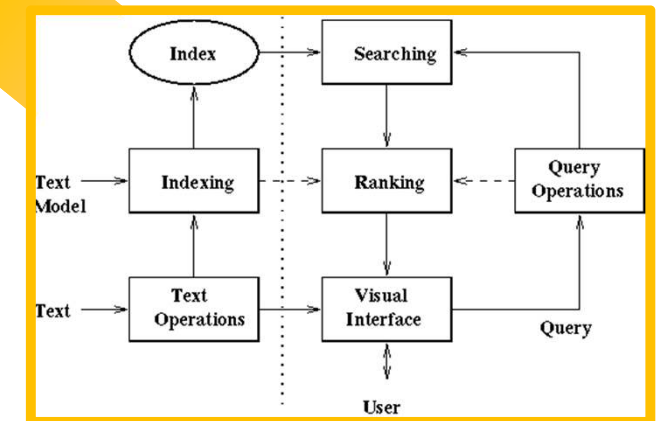
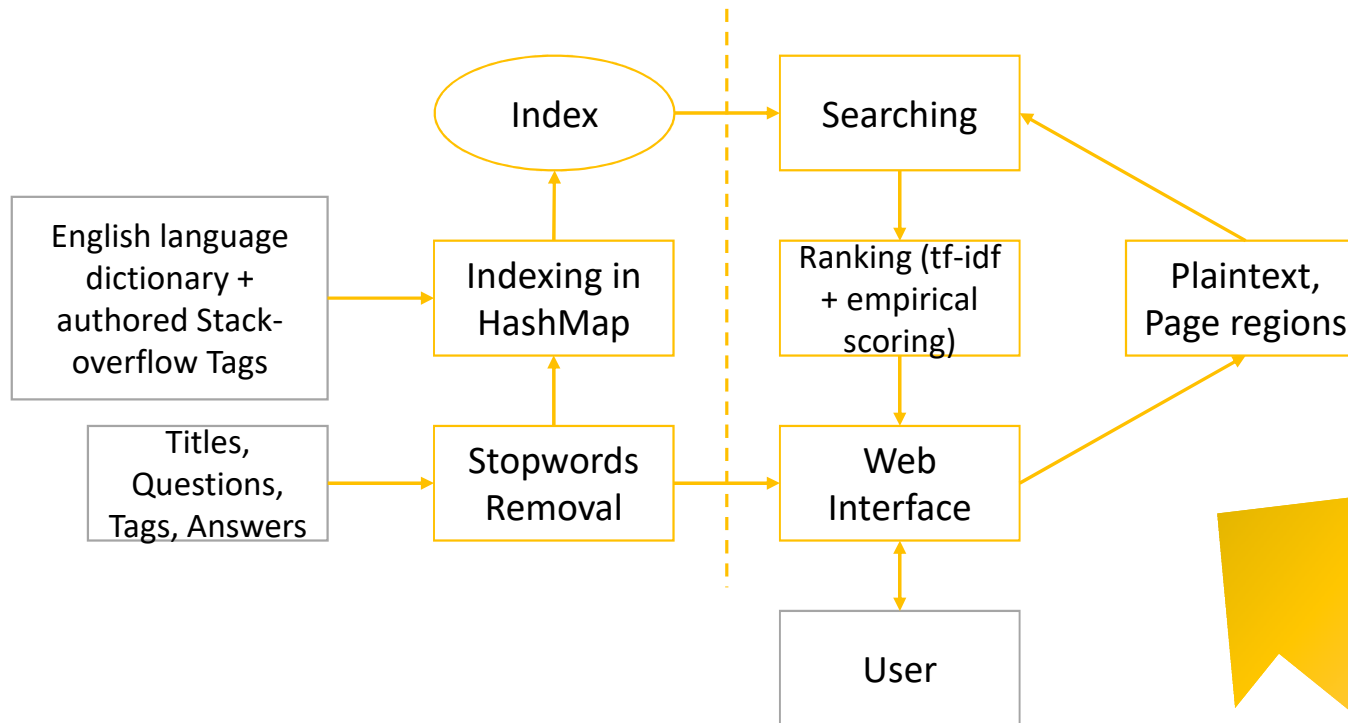
Good to see someone's chimed in about Lucene - because I've no idea about that.

Sphinx, on the other hand, I know quite well, so let's see if I can be of some help.

- Result relevance ranking is the default. You can set up your own sorting should you wish, and give specific fields higher weightings.
- Indexing speed is super-fast, because it talks directly to the database. Any slowness will come from complex SQL queries and un-indexed foreign keys and other such problems. I've never noticed any slowness in searching either.
- I'm a Rails guy, so I've no idea how easy it is to implement with Django. There is a Python API that comes with the Sphinx source though.
- The search service daemon (searchd) is pretty low on memory usage - and you can set limits on [how much memory](#) the indexer process uses too.
- Scalability is where my knowledge is more sketchy - but it's easy enough to copy index files to multiple machines and run several searchd daemons. The general impression I get from others

Hot Network Questions

- How can I support myself financially as a 17 year old with a loan?
- Why are prions in animal diets not destroyed by the digestive system?



Modern Information Retrieval © Addison-Wesley-Longman Publishing co.

1999 Ricardo Baeza-Yates, Berthier Ribeiro-Neto

Example User Query Methodology

1. Think of developer's information need and corresponding queries
2. Use Google Search for only Stackoverflow results (one day for calibration)
3. Filter for date range 06/05/2017 to 06/05/2019
4. Cross-validate results with Stackoverflow built-in ElasticSearch results
5. Return ranked top 5 questions

User query text	nginx reverse-proxy
Regional Search	-
Result 1	https://stackoverflow.com/questions/54151/nginx-reverse-proxy
Result 2	https://stackoverflow.com/questions/54038
Result 3	https://stackoverflow.com/questions/54123/for-nextcloud-not-work
Comments	

User query text	java ide
Regional Search	-
Result 1	https://stackoverflow.com/questions/51803/docker-nginx-php7-fpm-and-xdebug