On the Desirable Features of the Open Source Arabic Search Engine.

حول الميزات المطلوبة لمحرك البحث العربي مفتوح المصدر

By

Adnan H. Yahya

Birzeit University, Palestine

April 2017

Why an Arabic Search Engine:

Search engines come in all shapes and forms and the commercial have major companies with major resources dedicated to their improvement on all evaluation points: relevance of returned results, recall levels and accuracy as well as cross lingual support and even user profiling. They also make use of the valuable resources at their disposal: large document collections/corpora, large query logs, vast computational and storage power and the latest research results. Open source search engines are likely to be research oriented and may benefit greatly from any added flexibility that allows additional features by users. The intended audience may define the functionality they have and the ease with which additional features can be added. One element that may be lacking is the extensive collections of user profiles and user queries. They may not also be able to scale due to the associated costs. But the latter may be rectified if such systems are deployed, say for commercial usage.

So it should be clear that an open source Arabic search engine needs to maintain maximal flexibility in terms of future development and usability by researchers also to test novel approaches to Arabic and Cross Lingual search . It must be stable for future updates, even if major, and robust in terms of increased volume of indexed documents and query streams. It must also allow for adaptability to user needs and environment based on documents already indexed and queries posed. In a sense it is desirable that the search engine be an element of the research infrastructure for Arabic Information Retrieval.

From the user's perspective, the search engine should offer Relevant results to the user needs (results you are actually interested in, not necessarily closest to the query text), must have an intuitive and uncluttered, easy to use interface and must have helpful options to broaden or tighten the search direct through query reformulation/relevance feedback or based on user profiling or session history if available.

As far as Arabic is concerned so much can be incorporated and desirable:

We may befit from accounting for factors viewed by many as important for preferring one engine over the other. For example the following is the list of **Top Search Engine Ranking Factors (2015) based om 150 marketers views on** which features of websites and webpages are associated with higher search rankings

https://moz.com/search-ranking-factors/survey

1. **Domain-Level, Link Authority Features:** Based on link/citation metrics such as quantity of links, trust, domain-level PageRank, etc. 8.22/10
2. **Page-Level Link Metrics:** PageRank, Trust metrics, quantity of linking root domains, links, anchor text distribution, quality/spamminess of linking sources, etc. 8.19/10

3. **Page-Level Keyword & Content-Based Metrics:** Content relevance scoring, on-page optimization of keyword usage, topic-modeling algorithm scores on content, content quantity/quality/relevance, etc. 7.87/10
4. **Page-Level, Keyword-Agnostic Features:** Content length, readability, Open Graph markup, uniqueness, load speed, structured data markup, HTTPS, etc. 8.57/10
5. **User Usage & Traffic/Query:** Data SERP engagement metrics, clickstream data, Visitor traffic/usage signals, quantity/diversity/CTR of queries, both on the domain and page level 6.55/10
6. **Domain-Level Brand Metrics:** Offline usage of brand/domain name, mentions of brand/domain in news/media/press, toolbar/browser data of usage about the site, entity association, etc. 5.88/10
7. **Domain-Level Keyword Usage:** Exact-match keyword domains, partial-keyword matches, etc. 4.97/10
8. **Domain-Level, Keyword-Agnostic Features:** Domain name length, TLD extension, SSL certificate, etc. 4.09/10
9. **Page-Level Social Metrics:** Quantity/quality of tweeted links, Facebook shares, Google +1s, etc. to the page 3.98/10

Additionally, Mobile optimization is a major trend in current search engine features.

https://searchenginewatch.com/2016/02/25/say-goodbye-to-google-14-alternative-search-engines/

One Should also decide if to follow Google or should not (like DuckDuckGo) by keeping user data for future help.

https://searchenginewatch.com/2016/02/25/say-goodbye-to-google-14-alternative-search-engines/

And the way the engine uses click statistics to evaluate the results in the longer term instead of static evaluations. That is using Log analytics for personalized results.

Context based and semantic search

Use deep learning and information extracted from other sources like Wikipedia to improve the search: a la IBM Watson.

https://www.searchtechnologies.com/blog/top-enterprise-search-trends-2016

Search or Metasearch: One could base the effort on a new search engine or utilize the results taken from other proven search engines to answer user queries; Each has its pros and cons, and a major drawback is if the search engines, especially commercial ones are willing to cooperate.

http://www.thewindowsclub.com/meta-search-engine-list

One would need also to consider the basic search features and benefit from the currently common ones as detailed in the table below:

Basic Search Features in Search Engines;

http://www.thewindowsclub.com/meta-search-engine-list

Search Engine	Boolean	Default	Proximity	Truncation	Fields	Limits	Stop	Sorting
Google	-, OR	and	Phrase	No Auto stem word in phrase	intitle, inurl, link, site, more	Language, filetype, date, domain	Varies	Relevance, site
Bing	AND, OR, NOT, ( ), -, +	and	Phrase	No Auto stem	intitle, inurl, link, site, more	Language, filetype, date, domain	No	Relevance, site
Blekko	–	and	Phrase	No	site	date, slashtags	No	Relevance, date
Procog	–	and	Phrase	No			No	Relevance
Gigablast	AND, OR, AND NOT, ( ), +, –	and	Phrase	No	title, site, ip, more	Domain, type	Varies, + searches	Relevance
Exalead	AND, OR, NOT, ( ),-	and	Phrase, NEAR	Yes and stems	intitle, inurl, link, site	Language, file type, date, domain	Varies, + searches	Relevance

http://www.searchengineshowdown.com/features/byfeature.shtml

Added: Family Filters:

In as far as Arabic support is concerned one may want to emphasize the following:

1- Ability to support Cross Lingual Retrieval, may be one of pair of languages at a time. One may want to work with cross lingual similarity say through ESA based on Wikipedia.
2- One may need to address issues like named entity recognition (NER) and disambiguation resulting from absence of diacritics. Transliteration of foreign names may be an important issue as well as Foreign writing of Arabic names.
3- If one is to derive info from links then there needs to an account of the fact that the links language is usually Latin-based and thus there may be some need for translation.

4- Normalization issue in terms of Arabic confusion letters (Hamza, Alef, Ta marbouta/Ha and so on), absence of diacritical marks in the standard writing, the prevailing mixing of dialects and local slang for entities that may vary between Arab countries.

Domain Specific Bias:

Given that the search engine may be limited to the Academic/Research environment, one may want to take that into account in the implementation/search/ranking. This may be useful for disambiguation of entities and queries even without user profiling. Queries like “admissions requirement” شروط القبول will be interpreted as university related rather than more general as the phrase may imply.

Access to Related Languages resources: one of the big issue seemingly haunting researchers is that of freely accessible datasets: both how to reach them and how to contribute. Since the project may involve working with such sets it may be a good idea to try to make any sets we use avialable to researchers and maybe to be able to add resources and search for them on one of the sites accessible to the search engine (as opposed to LDC style leicensing). This may help researchers unify their datasets and improve the quality and encourage competition.

Conclusion: It may be good to specify the properties of the engine early in the game and to try to tune them as the tean goes. Having the general guidelines from the outset may help define the work terms and ways the effort will develop.

References:

https://searchenginewatch.com/2016/02/25/say-goodbye-to-google-14-alternative-search-engines/

http://www.thewindowsclub.com/meta-search-engine-list

http://www.thewindowsclub.com/meta-search-engine-list