ON THE DESIRABLE FEATURES OF THE OPEN SOURCE ARABIC SEARCH ENGINE

حول الميزات المطلوبة لمحرك البحث العربي مفتوح المصدر

By

Adnan Yahya

Birzeit Univesrity

Palestine

1

OUTLINE

- **Introduction and Background.**

- **Why an Arabic Search Engine?**

- **Top Search Engine Ranking Factors.**

- **Basic Search Features in Search Engines.**

- **Arabic support Elements.**

- **Domain Specific Bias**

- **Conclusion.**

2

INTRODUCTION AND BACKGROUND

- Lots changed since we last meeting

- IR and Web search have been on the rise also in academic curriculum

- Even commercial search engines are doing much better by the day

- We need to respond: how much and how fast?

MOTIVATION AND BACKGROUND (CONT)

- Better to have clear goals, strategic and tactical: The main goals of the product the features it should have/will develop to have in the future

- Need a review mechanism, even after the product is functional for continuous improvement

- Sustainability issues need addressing

- Can sustainability be done thorough sustained academic collaboration starting from us and expanding to include many more in the Arab area and beyond?

- Define success parameters: short and longer term and maybe a monitoring scheme

4

WHY AN ARABIC SEARCH ENGINE?

- Search engines come in all shapes and forms.
- 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, large query logs, vast computational and storage power and the latest research results.

WHY AN ARABIC SEARCH ENGINE?

○ An Arabic OS search engines is likely to be research oriented and benefit from any flexibility that allows additional features by users.

○ The intended audience may define the functionality the engine has 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 and needed resources.

○ The latter may be rectified if such systems are deployed, say for commercial usage.

6

WHY AN ARABIC SEARCH ENGINE?

- So, 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 maybe 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.

WHY AN ARABIC SEARCH ENGINE?

- It must also allow for adaptability to user needs and environment.

- In a sense it is desirable that the search engine be an element of the research infrastructure for Arabic Information Retrieval.

- Must have an intuitive and uncluttered, easy to use interface and must have helpful options to broaden or tighten the search through query reformulation/relevance feedback or based on user profiling or session history, if available.

TOP SEARCH ENGINE RANKING FACTORS
TOP RANKING FACTORS BASED ON 150 MARKETERS VIEWS

- **Domain-Level, Link Authority Features:** Based on link/citation metrics such as quantity of links, trust, domain-level PageRank, etc. 8.22/10

- **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

- **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

- **Page-Level, Keyword-Agnostic Features:** Content length, readability, Open Graph markup, uniqueness, load speed, structured data markup, HTTPS, etc. 8.57/10

TOP SEARCH ENGINE RANKING FACTORS
TOP RANKING FACTORS BASED ON 150 MARKETERS VIEWS

- **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

- **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

- **Domain-Level Keyword Usage:** Exact-match keyword domains, partial-keyword matches, etc. 4.97/10

- **Domain-Level, Keyword-Agnostic Features:** Domain name length, TLD extension, SSL certificate, etc. 4.09/10

10

TOP SEARCH ENGINE RANKING FACTORS
TOP RANKING FACTORS BASED ON 150 MARKETERS VIEWS

- **Page-Level Social Metrics:** Quantity/quality of tweeted links, Facebook shares, Google +1s, etc. to the page 3.98/10
- **Mobile optimization** is a major trend in current search engine features.
- **Keeping user data** for future help: as in Google or not as in DuckDuckGo?
- 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.

TOP SEARCH ENGINE RANKING FACTORS
TOP RANKING FACTORS BASED ON 150 MARKETERS VIEWS

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

- 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.

Basic Search Features in Search Engines

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

13

ARABIC SUPPORT ELEMENTS

- 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.
- Address issues like named entity recognition (NER) and disambiguation resulting from absence of diacritics.
- Transliteration of foreign names as well as Foreign writing of Arabic names.

14

ARABIC SUPPORT ELEMENTS

- 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.

- 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

- If the search engine is be directed at Academic Research environment, take that into account in the implementation/search/ranking.

- 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.

DOMAIN SPECIFIC BIAS

- Access to Related Languages resources such as 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 make any sets we use available to researchers and allow them to add resources and search for them on one of the sites accessible to the search engine (as opposed to LDC style licensing).

- 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 team goes.

- Having the general guidelines from the outset may help define the work terms and ways the effort will develop.

- Having the product accepted by the research community is essential: so may need to add features and materials.

Thank you

Questions?