How Do Search Engines Handle Arabic Queries?

How Do Search Engines Handle Arabic Queries?

By:Haidar Moukdad

School of Library and Information Studies,2004

INTRODUCTION

-General search engines on the Web are the most popular tools to search for, locate, and retrieve information.

-These engines handle English queries more or less in the same way.

- But their handling of non-English queries is greatly different.

INTRODUCTION

-Most general search engines like AltaVista, lltheWeb, and Google, allow users to limit their searches to specific languages, and some of them even provide local versions.

-How the general search engines handle non-English queries is an area that has been largely neglected by research on information retrieval on the Web.The neglect is even more apparent in research on non-Western languages like Arabic.

Information Retrieval and the Arabic language

- Information retrieval, as a language-dependent operation, is greatly affected by the language of documents and how a search engine handles the characteristics of this language. Linguistic characteristics that typically have impact on the accuracy and relevancy of Web searches are mainly related to the morphological structures of words.

Information Retrieval and the Arabic language

- Morphologically, Arabic lexical forms (words) are derived from basic building blocks with tri-consonantal roots at their bases. Only about 1200 roots are still in use in modern Arabic.

- word formation is a complex procedure that is entirely based on root-and-pattern system. Using clearly defined patterns, a large number of words can be derived from one root.

Information Retrieval and the Arabic language

- word formation is a complex procedure that is entirely based on root-and-pattern system. Using clearly defined patterns, a large number of words can be derived from one root.

- Arabic nouns and verbs are heavily prefixed

Methodology

- A set of eight Arabic search terms was selected to run in a set search engines.

-Using terms that emphasized some of the specific characteristics of Arabic morphology.

- Three general search engines (AlltheWeb, AltaVista, and Google)and three Arabic engines (Al bahhar, Ayna, and Morfix (the Arabic module)).

Methodology

- Al Bahhar provides options to search for the derivations of a word or for a word stripped of prefixes and suffixes.

- Ayna does not offer information on how its search engine works

- The Arabic module of Morfix allows exact-word searching, morphological searching, and expanded searching. Using morphological searching, all morphological forms of a term(word) would be retrieved. While expanded searching retrieves all the words the share the same root with the search term .

Results and discussion

- The eight queries (search terms) were selected to reflect some of the problematic characteristics of the morphology of the Arabic language that affect information retrieval.
- The first five terms are variants of the noun(جامعة).
- without any prefixes or suffixes(جامعة)
- Noun with definite article attached to it as a prefix (الجامعة)

Results and discussion

- Noun with three prefixes(بالجامعة).

-The noun with one prefix and one suffix (لجامعتي).

-The noun with four prefixes (وبالجامعة).

- The sixth term is the exact form of the noun *byt (بيت)*.

-with two prefixes (للبيت).

Finally, the eighth term is a plural noun that starts with two letters that could be mistaken for the definite article as a prefix (الوان)

Results and discussion

Query	Google	Ayna
جامعة	١٣٢٠٠٠	٨٤٣
الجامعة	٩٢٩٠٠	٦٩٤
بالجامعة	١٣٩٠٠	٢٧٤
لجامعتي	73	١٠
وبالجامعة	60	٠
بيت	١٧٥٠٠٠	١٢٨٨
للبيت	٧٢٦٠	٥٥٥
الوان	١١٤٠٠	٣٨٤

Results and discussion

-Google retrieved 132,000 documents out of 238,933

-(%٥٥).

-Ayna retrieved 843 out of 1821 (less than 50 percent).

Results and discussion

Queries in Al Bahhar

Query	Exact	Derivations
جامعة	4635	9498
الجامعة	3332	9498
بالجامعة	639	9498
لجامعتي	1	9498
وبالجامعة	3	84
بيت	4111	13133
للبيت	271	15780
الوان	50	3079

Results and discussion

Queries in Al Morfix

Query	Exact	Morphological	Expanded
جامعة	362	592	679
الجامعة	145	592	679
بالجامعة	13	592	679
لجامعتي	0	592	679
وبالجامعة	0	592	679
بيت	287	2094	2118
للبيت	14	2094	2118
الوان	17	571	571

Results and discussion

-(Albahhar)Using the exact word for the first five terms resulted in missing many documents containing orphologically related words.

More than 50 percent of the documents were missed by using the exact form of *jamct;*close to 60 percent by using *aljamct, more than 90 percent by using ljamcty; and almost* all documents by using *wbaljamct. Similar results were produced by using the exact*

forms of *byt and llbyt.*

Results and discussion

-In Morfix, it is also clear that using the "Morphological" and "Expanded" search options resulted in significantly higher numbers of retrieved documents.

Finally, unusually high numbers of documents were retrieved by Al Bahhar and Morfix when using the advanced search features with *alwan. Since this noun starts with al, these two letters might have been mistakenly identified by the engines as the definite article* .

Conclusion

- The importance of making users aware of what they miss by using the general engines, underscoring the need to modify these engines to better handle Arabic queries.

- high number of documents that will be lost when only the exact forms of Arabic words are entered as search terms on the Web.

- need for further research into the feasibility of developing retrieval tools that allow search engines to better Arabic queries.

Q &A