Haidar Moukdad
School of Library and Information Studies, Dalhousie University
Halifax, Nova Scotia

Lost In Cyberspace: How Do Search Engines Handle Arabic Queries?

Abstract: The performances of general and Arabic search engines were compared based on their ability to retrieve morphologically related Arabic terms. The findings highlight 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.

1. Introduction

Since the early days of the Web, English has been the lingua franca of Web documents, technology, and search engines and tools. However, the use and spread of other languages are by no means negligible, a fact that contributes to the complexity and importance of investigating information retrieval on the Web. General search engines on the Web are the most popular tools to search for, locate, and retrieve information, and their use has been growing at exponential rates. These engines handle English queries more or less in the same way, but their handling of non-English queries is greatly different from how these queries are handled by non-English search engines—engines that were designed for specific languages.

Most general search engines like AltaVista, AlltheWeb, and Google, allow users to limit their searches to specific languages, and some of them even provide local versions, including fully functional interfaces, to better accommodate the information needs of different regional and linguistic groups. Searchers for non-English documents either use the general search engines or smaller engines specifically designed to handle queries in their respective languages. 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, among which is Arabic, the language covered in this paper.

2. Background

2.1 Related research

The research Users looking for Arabic information on the Web can access search engines and directories that cover Arab countries, or they can use the general search engines to search in Arabic. The question that initiated this research was: how well do the major search engines handle Arabic queries, and to what extent do they accommodate the specific linguistic properties of this language? Researchers who investigated information retrieval in languages other than English on the Web dealt with general search engines and language-specific ones. Describing the development of search engines for Indian languages, Mujoo et al. (2000) discuss the information architecture of these special

engines, and explain the special characteristics of the languages they were developed for, including the use of different scripts and the morphological compositions of words. Moukdad (1999) evaluated AltaVista's handling of the retrieval of Arabic documents by building a small document collection and using a local installation of AltaVista for information retrieval experiments. Adopting a similar approach to non-English information retrieval, Bar-Ilan and Gutman (2003) explored the capabilities of search engines for non-English languages, examining four languages: French, Hebrew, Hungarian and Russian. They concluded that the general search engines largely ignore the special characteristics of non-English languages, and they even sometimes do not handle diacritics used in some European languages. Focusing on the Polish language, Sroka (2000) studied the capabilities of the Polish versions of Infoseek and AltaVista and those of Polish search tools. The only linguistic aspect covered in Sroka's study was the use and handling of diacritics.

2.2 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 and to morphological word variants. Thus it is not surprising that the most common linguistic features provided by search engines are automatic stemming (conflation of morphologically related words) and truncation. While the positive effect of stemming on English information retrieval has yet to be empirically proven (Harman, 1991 and Hull 1996), languages with more complex morphology are more susceptible to the advantages of stemming (Popoviç and Willett, 1992 and Savoy, 1991). As opposed to Arabic and other morphologically complex languages, the English language has morphological rules that can be easily treated in computational and information retrieval environments. English words tend be formed on the basis of a limited and relatively straightforward number of rules, allowing for simple stemming rules in order to retrieve variants of search terms. Conversely, Arabic has a large number of rules that makes retrieving word variants a challenging task and, consequently, stemming and other techniques absolute necessities.

Arabic belongs to the Semitic family of languages, which includes Akkadian, Aramaic, Ethiopic, Hebrew, Phoenician, Syriac and Ugaritic. The Arabic script was derived from the Aramaic via the Nabatean cursive script (Hitti 1963). As is the case with all Semitic languages, the script is written from right to left, and this script has traditionally been represented in converted (Romanised) form in Western academic and computerized environments. As opposed to English and other Western languages, vowels have never become a permanent part of the Arabic writing system, allowing for the occurrence of many homonyms in documents. For example, the written word *Scr* could have any of the following meanings: to feel, poetry, hair or to crack. And the word *clm* could mean flag, science or to know. To handle the Arabic script on the Web, a number of encoding systems has been developed. The most common of these systems are Arabic (Windows), Arabic (ISMO 708), Arabic (DOS), and Arabic (ISO). In addition, Arabic is covered by the Unicode encoding system.

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 (Hegazi and Elsharkawi, 1985), and word formation is a complex procedure that

is entirely based on root-and-pattern system (Hudson, 1986). Using clearly defined patterns, a large number of words can be derived from one root. Like a mathematical formula, using patterns to create different morphological variations from a root is a fairly regular process in which the original letters of the root are constant variables, and changing variables are letters added at the beginning, middle or end of the root. Arabic verbs also have their own pattern system, which assigns different meanings to the original meaning of a verb based on ten forms. Form I (the root) is the simplest form from which nine additional forms can be derived to provide subtle variations in meanings. For example, Form I of *qbl* (to accept) can be changed to: Form III *qabl* (to meet); Form V *tqbl* (to receive); and Form X *astqbl* (to greet).

To make matters more complicated, Arabic nouns and verbs are heavily prefixed. The definite article *al* is always attached to nouns, and many conjunctions and prepositions are also attached as prefixes to nouns and verbs, hindering the retrieval of morphological variants of words (Moukdad, 1999). The prevalence of prefixes in Arabic is a challenging problem for anyone attempting information retrieval on the Web, especially when using general search engines. At the same stripping nouns and verbs of prefixes in Arabic search engines can produce unexpected results, as many words start with one letter or more that can me mistakenly identified as prefixes.

3. Methodology

The model developed for Hebrew by Bar-Ilan and Gutman (2003) was followed in conducting search experiments for this paper. A set of eight Arabic search terms was selected and run in three general search engines (AlltheWeb[1], AltaVista[2], and Google[3]) and three Arabic engines (Al bahhar[4], Ayna[5], and Morfix (the Arabic module)[6]). The searches were conducted in October 2003, using terms that emphasized some of the specific characteristics of Arabic morphology as listed above. The three general search engines allow limiting the search results to Arabic, while the three Arabic search engines allow exclusive search of Arabic documents. AltaVista and AlltheWeb index every word in a Web document and do not perform any automatic conflation of terms. Google uses stemming technology, and it searches for words that are similar to a search term.

Al Bahhar is the successor of ArabVista, and 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, but it based on a powerful indexing system that takes into account the morphological characteristics of the Arabic language. 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 (Morfix, 2003).

Queries using the eight search terms were entered in each of the six search engines and, where appropriate, the linguistic search features/options provided by the engines were used. For example, morphological searching and expanded searching were used in Morfix, and derivational and stripped searching options were used in Al Bahhar. The next section displays the results of the searches and a discussion of how the search engines handled the queries.

4. 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 *jamct* (university). The first term is the exact form of the noun without any prefixes or suffixes, while the remaining four terms are: the noun with definite article attached to it as a prefix; the noun with two prefixes; the noun with one prefix and one suffix; and the noun with three prefixes. The sixth term is the exact form of the noun *byt* (house); the seventh term is *byt* with two prefixes. Finally, the eighth term is a plural noun that starts with two letters that could be mistaken for the definite article as a prefix.

Table 1 shows the results of the eight queries in the three general search engines and the Arabic engine that did not provide advanced linguistic options.

Table 1: Queries in the general search engines and Ayna

Query	AllheWeb	AltaVista	Google	Ayna
jamct (university)	63,684	66,893	132,000	843
aljamct (the university)	52,150	53,012	92,900	694
baljamct (in the university)	5,417	5,659	13,900	274
ljamcty (for my university)	13	13	73	10
wbaljamct (and the university)	25	25	60	0
byt (house)	103,893	103,161	175,000	1288
llbyt (for the house)	3,862	3,913	7,260	555
alwan (colors)	6,572	6,632	11,400	384

For each query, the difference in the number of retrieved documents among engines in Table 1 should be viewed with the understanding that coverages by the different engines are varied. What is important for the purpose of the current research is the number of retrieved documents by each engine for the first five queries, and then for the sixth and seventh query. Entering the exact form of the word *jamct* in AlltheWeb retrieved slightly more than half of the available documents that could be retrieved by using the five forms of the word. Similar results are achieved by AltaVista. Out of the 125,602 possible documents, using the exact form of the word retrieved 66,893. Google retrieved 132,000 documents out of 238,933; and Ayna retrieved 843 out of 1821 (less than 50 percent). The problem was not as severe with the results of *byt* and *llbyt*, and this is due to the fact that the word variant *llbyt* is not as common as *albyt* (the house); had *albyt* been used as a search term, the results of the searches would likely have been similar to those of the variants of the word *jamct.*

The search results listed in Table 2 and Table 3 display the number of documents retrieved by Al Bahhar and Morfix (the two Arabic engines with morphological search options).

Table 2: Queries in Al Bahhar

Query	Exact	Derivations	Stripped
jamct (university)	4,635	9,498	8,452
aljamct (the university)	3,332	9,498	8,452
baljamct (in the university)	639	9,498	8,452
ljamcty (for my university)	1	9,498	79
wbaljamct (and the university)	3	84	8,452
byt (house)	4,111	13,133	9,077
llbyt (for the house)	271	15,780	9,090
alwan (colors)	50	3,079	2,115

Table 3: Queries in Morfix

Query	Exact	Morphological	Expanded
jamct (university)	362	592	679
aljamct (the university)	145	592	679
baljamct (in the university)	13	592	679
ljamcty (for my university)	0	592	679
wbaljamct (and the university)	0	592	679
byt (house)	287	2,094	2,118
llbyt (for the house)	14	2,094	2,118
alwan (colors)	17	571	571

Al Bahhar might have been malfunctioning when the searches were performed, and this explains the low numbers for *ljamcty* and *wbaljamct* in "Derivations" and "Stripped" However, the whole picture is what is important. Using the exact word for the first five terms resulted in missing many documents containing morphologically 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*. 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 and were stripped off the retrieved terms.

5. Conclusion

The performance of three general search engines was compared to three engines that were specifically designed to handle the linguistic characteristics of Arabic. A set of queries was entered in the general search engines and in their Arabic counterparts. The query terms were carefully selected to emphasize the specific characteristics of Arabic that differentiate it from English. Criteria for measuring the performance of engines included their ability to retrieve documents containing morphologically related terms, and the features they provide to avoid missing potentially relevant documents. The findings highlight the importance of making users aware of the limitations of general search engines in retrieving Arabic documents, and of the high number of documents that will be lost when only the exact forms of Arabic words are entered as search terms on the

Web. The findings also underscore the need for further research into the feasibility of developing retrieval tools that allow search engines to better handle non-English queries in general and Arabic queries in particular.

ENDNOTES

1. http://www.alltheweb.com
2. http://www.altavista.com
3. http://www.google.com
4. http://www.albahhar.com
5. http://www.ayna.com
6. http://www.morfix.com/arabic/ArabicSearch.asp

REFERENCES

Bar-Ilan, J. and T. Gutman, 2003. How do search engines handle non-English queries? - A case study. The Twelfth International World Wide Web Conference 20-24 May 2003, Budapest, Hungary. [http://www2003.org/cdrom/papers/alternate/P415/BARILAN.HTM]. Accessed January 12, 2004.

Harman, D. 1991. How effective is suffixing? Journal of the American Society for Information Science, 42 (1): 7-15.

Hegazi, M. and A. Elsharkawi. 1985. An approach to a computerized lexical analyzer of natural Arabic. Computer processing of the Arabic Language, Workshop Papers, (Vol. 1). Kuwait: Kuwait Institute for Scientific Research (KISR).

Hitti, P. 1963. History of the Arabs. 8th edition. London: Macmillan.

Hudson, G. (1986). Arabic root and pattern morphology without tiers. Journal of Linguistics. 22, Mar.: 85-122.

Hull, D. 1996. Stemming algorithms: a case study for detailed evaluation. Journal of the American Society for Information Science, 47(1): 70-84.

Morfix. 2003. [http://www.morfix.com/arabic/help.htm]. Accessed December 2, 2003.

Moukdad, H. 1999. An investigation of the necessity of information retrieval algorithms for full-text Arabic databases. Information Science: Where has it Been, Where is it Going? Proceedings of the 27th Annual Conference of the Canadian Association for Information Science, Université de Sherbrooke, June 1999. [Toronto]: CAIS, 207-227.

Mujoo, A., M. Malviya, R. Moona and T. Prahakar. 2000. A search engine for Indian languages. EC-Web 2000, Lecture Notes in Computer Science 1875: 349-358.

Popoviç, M. and P. Willett. 1992. The effectiveness of stemming for natural-language access to Slovene textual data. Journal of the American Society for Information Science, 43: 384-390.

Savoy, J. 1993. Stemming of French words based on grammatical categories. Journal of the American Society for Information Science, 44(1): 1-9.

Sroka, M. 2000. : Web search engines for Polish information retrieval: Questions of search capabilities and retrieval performance. International Information & Library Research, 32: 87-98.