International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

INFORMATION RETRIEVAL IN ARABIC: AN EVALUATION OF THREE MULTILINGUAL SEARCH ENGINES ON THEIR CAPABILITIES IN DEALING WITH ARABIC SEARCH QUERIES

ANNEGRET M. GROSS

German-Jordanian University, School of Humanities and Languages, BA Translation, Jordan
annegret.gross@ymail.com

ABSTRACT

Most of the Web search engines were developed in the United States and designed for information retrieval (IR) in the English language. With the exponential expansion of the World Wide Web, IR in languages other than English has become increasingly important. Arabic is a language that has been experiencing immense growth on the internet. Many studies have evaluated the IR capabilities of both Arabic and multilingual search engines, but most of the studies are around 15 years old. No recent study has examined present-day search engines with respect to Arabic IR and drawn a parallel to past research. This is where this paper focuses: the first section compares IR search results from 2004 and March 2014 from the three market-dominant search engines Google, Bing and Yahoo Maktoob. Along with an analysis of indexed keywords on the first search results page, this section compares present and past IR performance and investigates how indexing and stemming have changed over this decade. The second section of this paper investigates how the search engines deal with spelling variants of transliterated foreign names, different types of common misspellings, and improper word division, as well as the typographical effect of the Kashida. The section also gives attention to the availability and efficiency of search support tools, such as auto-correction and auto-suggestion.

Keywords: information retrieval, Arabic, search engine, stemming, indexing

1. INTRODUCTION

The rise of the World Wide Web in the 1990's and the ever since soaring amount of web pages brought about the necessity for automated information retrieval (IR) systems. The first web search engines emerged in the mid-1990's. The first search engine to provide a full-text crawler-based web search was developed by Brian Pinkerton at the University of Washington in 1994. WebCrawler was able to index plain text and allowed users to search for any term on the internet. Lycos was launched in 1994 as well, and many other search engines were to follow, including Infoseek in 1994; Excite, AltaVista, and Yahoo in 1995; Inktomi in 1996. Finally, Google was launched in September 1997, which today is the most popular web search engine with more than five billion searches every day (Google Official History, Comscore, 2014).

All these search engines were developed in the United States and as a matter of course, they were initially designed for English-language search queries. However, other languages are very different from English, and it soon turned out that information retrieval in languages other than English was not as effective. Hence, IR in languages other than English became a topic of increasing research interest, in particular from 2000 onwards. Many studies evaluated the effectiveness of search engines in dealing with non-English queries under different perspectives, evaluation methods, and test designs. Bar-Ilan and Gutman (2003), for instance, compared the performance of international and local search engines for Russian, French, Hungarian, and Hebrew. Sroka (2000) did the same for the Polish language. Mujoo, Malviya, Moona, & Prahakar (2000) described the challenges for search engines posed by

ISSN 2304-0777 www.jitbm.com

multiple scripts and encodings for the Indian language. All studies have one thing in common: they came to the conclusion that the international search engines were not sufficiently prepared to handle the characteristics of languages other than English.

Arabic is one of the languages, which are highly challenging for IR systems. However, with an increasing number of Arabic-speaking internet users and growing internet penetration rates, efficient information retrieval in Arabic has become a more pressing issue than ever before. Over the past decade, the internet usage in the greater MENA region has been growing by about 2,500 %, which makes Arabic the fastest growing language on the internet worldwide. From 2000 to 2011, the number of Arabic-speaking internet users rose from 2 million to more than 65 million (Internet World Stats, 2011). For comparison, the second fastest growing language is Russian with 1,826%, followed by Chinese with 1,277%. Far after that come Spanish with 743% and English with 281% (Rotaru, 2011). With about 293 million Arabic speakers from 57 countries and an internet penetration rate of only 40.2% (Internet World Stats, 2012), the growth potential is still not exhausted.

2. PURPOSE OF THIS PAPER

Around 15 years ago, information retrieval in Arabic was a highly active area of research. Arabic search engines were developed and launched into the market. Research studies investigated the information retrieval capabilities of general and Arabic web search engines. However, the Arabic search engines have disappeared from the market, and the major multilingual search engines are available with Arabic user interfaces and, in some cases, specific Arabic web domains. Yet up to March 2014, no study has examined the Arabic versions of the international search engines and compared the status quo with past research results. This is where this paper focuses: the first section takes up specific Arabic search queries as used in a 2004 study. The queries were selected to determine how typical Arabic prefixes and suffixes were dealt with by the major multilingual web search engines. The same queries were now run through the three market-dominating search engines, Google, Bing and Yahoo Maktoob. The aim was to establish a direct comparison between 2004 and 2014 IR performance. The three multilingual search engines were used in their Arabic user interfaces. The test results indicate changes in the IR ratios for each search query, in relation to the past, as well as among the current search engines. An analysis of the indexed keywords on the first page indicates differences in indexing and stemming between the search engines.

The second section of this paper evaluates the search engines under different aspects of Arabic information retrieval, which have not been evaluated yet:

- Different spelling variants of common English names transliterated into Arabic,
- Common spelling errors (typographical errors and improper word division),
- Diacritical marks, including vowel markers,
- Availability and efficiency of search support tools, such as auto-correction and auto-suggestion tools, and
- The Kashida (Arabic typographical effect of elongation).

An analysis of the indexed keywords on the first search result page provides information on indexing and stemming for the second test, too.

3. LITERATURE REVIEW

The majority of research on Arabic IR took place shortly after the emergence of the first search engines around 2000. In 1999, Moukdad investigated AltaVista's performance in dealing with Arabic prefixes. He came to the conclusion that the search did not work as effectively for Arabic queries as for English queries and suggested that "handling prefixes in Arabic words necessitates the development of new information retrieval algorithms for this language" (Moukdad, 1999:219-220). Xu, Fraser and Weischedel (2002) carried out IR experiments on the TREC Arabic corpus, a compilation of 383,872 Arabic documents from Agence France Presse (AFP). They selected their search queries according to orthographic, morphological and semantic particularities of the Arabic language. Their experiments showed that stemming was critical for Arabic IR and improved the retrieval performance by 40%; spelling normalization improved the retrieval performance by 22%. Bushnaq (2003) examined four Arabic (Ayna, Ajeeb, Al-Bahhar, and Arabia) and the English-language search engines (Google, Yahoo, and Alltheweb) by searching for 20 Arabic websites. His tests revealed that the complex morphology and

ISSN 2304-0777 www.jitbm.com

the numerous prefixes of the Arabic language were problematic. Google came first place in terms of search accuracy and quantity of data retrieved. This was followed by Alltheweb and Arabia. Ayna and Arabvista showed poorer results, which the author also contributed to the fact that these two search engines were not spider-based but web directory-based.

Moukdad (2004) compared the performance of three general and three Arabic search engines based on their ability to retrieve morphologically related Arabic words. The general search engines tested were Google, Alltheweb, and AltaVista. The Arabic search engines were Ayna, Albahhar, and Morfix. The author chose a set of eight keywords, which reflected the characteristics of the Arabic language. The test queries included the basic form of the word, as well as prefixed and suffixed variants. Moukdad (2004) measured the IR performance by the search engines' ability to retrieve documents with morphologically related words and their features to avoid neglecting potentially relevant documents. The results indicated that Google retrieved the highest number of documents for exact query searches. This was followed by AltaVista, Alltheweb, and then Albahhar, Ayna, and Morfix. If no morphological search option was available, regular query searches missed a great number of relevant documents. A prefixed or suffixed search term significantly reduced the number of results in these cases.

In 2006, Moukdad examined the Arabic language processing capabilities of IDRISI and the English search engine AltaVista, using a set of 40 selected Arabic nouns. The experiments showed that nouns entered without a prefix were the greatest problem for AltaVista. With added prefixes, the recall levels increased considerably. IDRISI performed better as it used advanced stemming to strip words of prefixes and suffixes. While most English-language search engines provided stemming for regular English plural forms, Moukdad (2006) noted that irregular plurals of Arabic nouns were another major problem for AltaVista. He further found that the English search engine was not able to handle the Arabic *Kashida*, a typographical effect of elongating the space between two letters, such as in "الحمـــــد." Moukdad (2006) suggested to treat the Kashida as a stop word. Boualem and Abbes (2008) tested Google's Arabic IR efficiency. Their test results indicated that Google was not adapted to handle the particular structure of the Arabic language, especially its complex morphology and the absence of vowels in the Arabic script. The authors stressed the precedence of implementing adequate processing for the lexical, syntactic and semantic singularities of the Arabic language. Al-Rawi and Al-Khateeb (2009) compared the multilingual search engines Google and Yahoo with two Arabic search engines, Al-Hoodhood and Ayna. The authors entered 20 Arabic keywords with their roots into the search engines and evaluated number of retrieved pages, retrieving time, and stability. Google performed best in dealing with Arabic keywords. Then followed Yahoo, Ayna, and Al-Hoodhood. This result indicated that the Arabic search engines were not as well prepared as the general search engines to deal with Arabic search queries.

Tawileh, Mandl and Griesbaum (2010) compared the IR performance of three English-language search engines (Google, MSN, and Yahoo) and two Arabic search engines (Araby and Ayna), using fifty randomly selected queries from the top searches on Araby. Their study indicated that Google performed best in all aspects, followed by MSN and Yahoo. The Arabic search engines were not able to match up to their multilingual counterparts. They showed a significant underperformance in indexing and searching algorithms. Even though there was no remarkable performance difference between Google and Yahoo, the authors considered Google's Arabic interface an advantage over Yahoo, which did not feature an Arabic interface at the time of the test. In addition to the studies mentioned, many more papers have proposed various methodologies and algorithms to master the challenges of Arabic IR, among them Abu-Salem, Al-Omari and Evens (1999), Khoja and Garside (1999), Aljlayl and Frieder (2002), Chen and Gey (2002), Larkey, Ballesteros and Connell (2002, 2005), Al-Shalabi, Kanaan and Al-Serhan (2003), Harmanani, Keirouz and Raheel (2006), Omer and Ma (2009).

4. ARABIC-LANGUAGE PARTICULARITIES POSING DIFFICULTIES TO IR SYSTEMS

With about 293 million native speakers, which is about 4.23% of the world's population, Arabic is the fifth most widely spoken language worldwide after Chinese, Spanish, English, and Hindi (Nationalencyklopedin, 2010). Modern Standard

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

Arabic (MSA) is the official language of 27 Arab states. Arabic is divided into three forms: classical Arabic, Modern Standard Arabic (MSA), and the spoken dialects, which vary according to country and region. Classical Arabic was the prevalent form in pre-Islamic times and is also the language of the Quran. Modern Standard Arabic is derived from classical Arabic and is used for most printed matter, such as newspapers and books. It is also used by Arab speakers from different regions to communicate with each other. The Arabic script consists of 28 letters and is written from right to left. Only consonants and long vowels are written. Short vowels are replaced by vowel signs, so-called diacritics. Arabic has two genders (masculine and feminine), three numbers (singular, plural, and dual), and three grammatical cases (nominative, accusative, and genitive) (Al-Harbi, Almuhareb, Al-Thubaity, Khorsheed, & Al-Rajeh, 2008).

The Arabic language is likely to present a challenge to traditional English search engines for different reasons. Firstly, the redundancy in Arabic is much greater than in English, because Arabic words are derived from roots and formed according to certain patterns. Arabic has about 5 million words, which are derived from approximately 11,300 roots. English, in comparison, has a total of about 1.3 million words with a number of 250,000 distinct words (Al-Maimani, Naamany, & Bakar, 2011). Secondly, the derivation of Arabic words from a usually tri-literal root leads to a high amount of polysemy. The ratio of ambiguity in the Arabic language is found to be larger than in other languages (Abdelali, 2006). An IR search might thus return many documents with ambiguous results, which may not be relevant for the user. Thirdly, the Arabic language is not only rich in polysemes, but also in synonyms. Its synonymy is far higher than that of the English language. Fourthly, the diglossia, resulting from differences between the various dialects and the standard form of Arabic, presents a semantic difficulty when searching for web documents. Fifthly, the greatest challenge for search engines is the complex morphology of the Arabic language, which this paper shall focus on. The reasons for this complexity are identified in the following section.

4.1 Prefix and Suffix Agglutination

Irregular plural forms, so-called broken or internal plurals, are very common in Arabic nouns and adjectives. Broken plurals do not follow the normal plural formation rules. They are made by changing the pattern of letters within the word and often do not resemble their singular counterparts. The regular plural (or external plural) of a masculine noun in nominative case is formed by adding the ending "ون." In accusative and genitive case, it is "ين." Regular female plurals take the suffix "ات" in all three grammatical cases. These suffixes do not occur in irregular plurals. Another peculiarity of the Arabic language is the existence of a special plural form for the number of two, the dual. The dual is formed by adding the suffix "ان" to the singular form in nominative, and "ين" in accusative and genitive case. From a stemming perspective, irregular Arabic plural forms are extremely difficult to handle. In 2008, Sanan, Rammal and Zreik (2008) noted that search engine stemmers were not able to process broken plurals, due the modifications that take place *within* the word and the multitude of possible formation patterns. McCarthy and Prince (1990) suggest more than 70 patterns according to which a broken plural can be made. But even regular plurals are not easy to handle for IR systems. Simply stripping off regular plural or dual suffixes with a stemmer without any verification rules may lead to erroneous results, because in some cases, these suffixes are a part of the word and not a plural marker.

Another morphological difficulty for search engines is the agglutination of certain pronouns and prepositions, as well as the definite article, with the noun they refer to. As a result, a noun can contain up to two prefixes. The fact that the conjunction "و" ("and") is also connected to the subsequent word without a space in between can lead to words with up to three agglutinated morphological particles. The possessive pronouns, on the other hand, are attached to the noun in form of suffixes. In combination with a plural or dual suffix, a noun can contain up to two suffixes. Just as in the case of plural and dual suffixes, a specific letter or combination of letters are not necessarily a pronominal or prepositional affix, but can be a part of the word. Simply stripping it off may result in a different or non-existing word.

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

With all this said, it is obvious that the complexity of affixes merged with nouns and the irregular plurals make stemming highly more challenging in Arabic than in other languages. Table 1 illustrates prefix and suffix agglutination by the example of the Arabic word for "university."

Table 1
Example of the morphological suffix and prefix agglutination in Arabic

Arabic	English	Morphological composition
جامعة	university	جامعة
الجامعة	the university	جامعة + ال
بالجامعة	at the university	جامعة + ال + ب
لجامعتي	for my university	جامعة + ل + ي
وبالجامعة	and at the university	جامعة + ال + ب + و

4.2 Common Orthographic Errors and Spelling Variants

Arabic orthography and spelling may also pose a problem in information retrieval. The diacritic vowel markers to identify short vowels are usually not written, except in those cases where the vowel marker is needed to avoid an ambiguity with another word with the same consonant and/or long vowel spelling. Another point is that people tend to not write certain diacritics. For instance, the Arabic letter Alef with Hamza above or below ("أ" or "إ") or the Alef Madda ("آ") are often written as a plain Alef "ا" without any marker (Abdelali, Cowie & Soliman, 2004). In addition, diacritic points are often omitted on certain letters at the end of a word. It is very common that the Ta Marbuta "ة" is written like the letter "H" ("ه"), and that Yeh ("ي") is written as Alef Maqsura "ى" (Sanan, Rammal & Zreik, 2008). This omission of diacritic signs may render some words ambiguous. Removing the points from the final Ta Marbuta of the Arabic word for "university" ("جامعة") would result in "جامعه," meaning "his mosque."

The frequent inconsistent spelling of foreign proper names transliterated into Arabic constitutes another problem in Arabic IR. Transliterations are often inconsistent due to the many linguistic differences between the languages, especially in phonology, syllabic structure, and lexical stress (Al-Omar, 2013). When Abdelali, Cowie & Soliman (2004) analyzed the online corpora of Agence France Press (AFP), they found the name of the US city "Los Angeles" transliterated in four different ways: " لوس انجيليس" (34 occurrences), "لوس انجلوس" (23), " لوس انجليس" (21), and "لوس انجيلس" (2). The US state name "Carolina" was found transliterated in two different ways: "كارولاينا" (26), "كارولينا" (14).

5. MULTILINGUAL AND ARABIC SEARCH ENGINES

5.1 Former Arabic Search Engines

Around 2000, a range of Arabic search engines were launched. These search engines were designed to handle the typicalities of the Arabic language, which the general search engines missed out on. These Arabic search engines, many of them mentioned in Section 3, have disappeared from the market. Ayna was the first Arabic search engine to be launched in 1997 by Ayna Corporation. In 2009, Ayna was announced to be the most visited search engine for the Arabic-speaking Middle Eastern and North African market. The company stated that the search engine employed "innovative technology to handle the complexities of the languages of the Middle East" (Basis Technology, 2009, para. 3), as well as special search algorithms, morphologic and orthographic analysis, automatic language identification, entity extraction, name indexing, and name translation (Basis Technology, 2009). Ayna was formerly available at ayna.com. ArabVista (formerly available at arabvista.com) was launched in 2000 by Emirati company Emirates Internet & Multimedia. It was later on replaced by Al-Bahhar (formerly available at albahhar.com). Al-Bahhar provided options to search for derivations of a word as well as prefix and suffix

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

truncation (Bushnaq, 2003; Moukdad, 2004). Al-Hoodhood was launched in its Beta version in June 2005 by ATA Software Ltd., a London-based company specializing in Arabic business software. Al-Hoodhood was a morphological search engine with an Arabic user interface only (ATA Software Ltd., nd). Even though the website is still online at www.alhoodhood.com, the web search is no longer functional.

Unlike the search engines mentioned above, which could only search through a limited database of sites, Araby.com was the first Arabic crawler-based search engine. It was able to crawl through Arabic web pages and index results. Araby was launched in 2006 by Jordanian computer services company Maktoob Inc. and was available araby.com. Maktoob Inc. was sold to Yahoo in 2009. Araby was said to be superior to the other Arabic search engines, due to its advanced algorithms tailored to the Arabic language. Araby was capable of detecting different grammatical forms of Arabic words, recognize stop words, prefixes, and other typical features of the Arabic language. It was also able to recognize spelling mistakes and to provide auto-suggestions (Al Bawaba, 2006).

5.2 Current Search Engine Market

The current search engine market is clearly dominated by Google Search, which is the most-used internet search engine worldwide (eBizMBA, 2014). On the US market alone, Google owns the greatest share, by far, with 67%, as shown in Figure 1. Microsoft's search engine, Bing, and Yahoo follow behind with 17.9% and 11.3%, respectively. The last two places are occupied by Ask with 2.7% and AOL with 1.2%.

Figure 1. US search engine market share in percent as of July 2013.

Google's dominance in the search engine business has made its name a common verb in everyday language: "to google" is used synonymously for "searching the web." Google's success is mainly based on its search algorithms. Around 2000, Google brought up an innovative algorithm called "PageRank," enabling it to provide more relevant and precise search results than any other engine (Brin & Page, 1998). Google is known for investing huge sums into the constant upgrading and improvement of their search algorithms to remain the most sophisticated search tool on the market (WeSearch, 2014). This chimes indeed with the majority of research findings mentioned in Section 3., in which Google was found to consistently produce above-average IR performance. Another reason for Google's popularity may be its availability in a multitude of country-specific web domains and languages. Users can access Google at about 200 international top level domains (TLDs) and numerous multi-language interfaces (Sottimano, 2012). Google is available at a total number of 15 TLDs for the Arabic-speaking countries, including an Arabic version at their Israeli domain. The only three Arab countries without a separate domain yet are Algeria, Syria, and Yemen at the time of this paper.

Yahoo is currently available in localized interfaces for 65 countries (Yahoo!, 2014). As mentioned above, Yahoo acquired Jordanian computer company Maktoob Inc. along with their search engine Araby in 2009. Yahoo's Arabic website, which targets the whole of the Arabic-speaking countries, is called Yahoo Maktoob and can be accessed at either www.maktoob.com or at http://maktoob.yahoo.com.

Bing (www.bing.com) can be displayed in 35 languages and its search function can retrieve results in 40 languages, including Arabic. Bing offers a settings button that allows the user to select the language they want to have the search engine displayed in. In 2009, Microsoft and Yahoo concluded a deal to have Yahoo Search powered by Bing, with Microsoft providing web, video, and image listings to Yahoo (Yahoo! Help, 2013).

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

6. PRACTICAL SEARCH ENGINE TESTING: RESEARCH QUESTIONS AND TEST SETTINGS

The search engines tested in this paper are the three multilingual search engines that are currently prevalent in the market: Google Search (www.google.jo), Yahoo Maktoob (www.maktoob.com), and Bing (www.bing.com). Google's Jordanian interface with the language set to Arabic was used, as the test was carried out in Jordan. Yahoo Maktoob was used with its English language interface, because the Arabic interface does not indicate the number of results retrieved. Bing was used in its Arabic language setting. For the practical testing, different search queries were selected according to specific criteria, which shed light on the research objectives described above. These search queries were entered into the search engines during the same period of time, in March 2014. As a first step, the number of retrieved documents for each search query was noted in a table. Second, the indexed keywords of the first results page of each search engine were counted and analyzed to obtain more detailed information about the search engines' stemming and indexing methods. All tested search engines highlight indexed keywords in bold. The default results page of each search engine shows 10 links to presumably relevant pages, including title of the page, URL, and one to two lines of preview text. Keywords appearing in any type of advertisement were excluded from the analysis.

6.1 First Test: 2004 and 2014 Comparison

The first test aimed to assess changes in IR performance of multilingual search engines over the ten-year period from 2004 to 2014. The test also aimed to identify differences in IR performance, as well as in indexing and stemming, between the three major present-day multilingual search engines. An analysis of the indexed keywords on the first results page of each current search engine provided more detailed information about the indexed keywords.

6.1.1 First Test Settings

To compare current and past search engine IR performance in Arabic, this paper takes the test results of Moukdad's 2004 study as a baseline. Moukdad (2004) examined the information retrieval capabilities of one Arabic search engine, Ayna, and three multilingual search engines: AlltheWeb, AltaVista, and Google. To do so, he used a set of selected Arabic affixed and non-affixed search queries. AltaVista and AlltheWeb were both acquired by Yahoo in February 2003 and March 2004, respectively. Both web links (www.altavista.com and www.alltheweb.com) now refer to Yahoo's internet search site. Since Yahoo acquired both engines along with their search technology, the results retrieved with the current Yahoo Maktoob version were compared with the results obtained by AltaVista. As the retrieval performance of AlltheWeb and AltaVista was quite identical in the 2004 study, this study focused on AltaVista only.

For the present experiment, the same set of Arabic search queries used by Moukdad (2004) was run through the multilingual search engines Google, Yahoo Maktoob and Bing to obtain up-to-date search results. These search results were compared to Moukdad's results from ten years ago in Table 2. Since the number of documents available on the World Wide Web grows exponentially by the day, it is not possible to compare the gross quantity of the retrieved query data from 2004 and 2014. For this reason, the number of retrieved documents was converted into a percentage value. In accordance with Moukdad (2004, p. 4), it was assumed that the sum of all retrieved documents corresponds to the total amount of retrievable documents for a specific set of search queries. The percentage value of the retrieved results per query in relation to the total number of results allowed conclusions to be drawn about changes in information retrieval performance (IR perf. in %) for each search query and for each search engine. An analysis of the indexed keywords on the first search results page of each of the three current multilingual search engines in Table 3 provides more insight into differences in indexing and stemming for each search engine.

Table 2

Search results for affixed and non-affixed Arabic search queries by Google and AltaVista as of 2004 compared with the same search query results by Google, Yahoo Maktoob, and Bing as of March 2014, in number of retrieved documents (No. of results) and information retrieval performance in percent (IR Perf. in %)

		Results as of 2004				Results as of 2014					
		Google		AltaVista		Google		Yahoo Maktoob		Bing	
#	Search query	No. of results	IR Perf. in %	No. of results	IR Perf. in %	No. of results	IR Perf. in %	No. of results	IR Perf. in %	No. of results	IR Perf. in %
	Query Set 1										
1	جامعة (university)	132,000	55.2%	66,893	53.3%	293,000,000	39%	15,800,000	79.9%	24,200,000	81.0%
2	الجامعة (the university)	92,900	38.9%	53,012	42.2%	86,600,000	11.5%	3,350,000	16.9%	5,060,000	16.9%
3	بالجامعة (at the university)	13,900	5.8%	5,659	4.5%	185,000,000	24.6%	595,000	3.0%	594,000	2.0%
4	لجامعتي (for my university)	73	0.03%	13	0.01%	2,530,000	0.3%	14,800	0.07%	14,200 (هل تقصد بجامعة؟)	0.05%
5	وبالجامعة (and at the university)	60	0.02%	25	0.02%	184,000,000	24.5%	4,720	0.02%	4,750	0.02%
	TOTAL	**238,933**	100%	**125,602**	100%	**751,130,000**	100%	**19,764,520**	100%	**29,872,950**	100%
	Query Set 2										
6	بيت (house)	175,000	96.0%	103,161	96.3%	126,000,000	28.6%	5,220,000	83.8%	10,900,000	91.1%
7	للبيت (for the house)	7,260	4.0%	3,913	3.7%	315,000,000	71.4%	1,010,000	16.2%	1,070,000	8.9%
	TOTAL	**182,260**	100%	**107,074**	100%	**441,000,000**	100%	**6,230,000**	100%	**11,970,000**	100%

Note. Results as of 2004 adopted from "Lost in cyberspace: how do search engines handle Arabic queries?," by H. Moukdad, 2004, *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science, Winnipeg, 2004*. Copyright 2004 by CAIS 2004.

ISSN 2304-0777

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

www.jitbm.com

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

Table 3
First page indexed keywords by Google, Yahoo Maktoob, and Bing as of March 2014

			Number and type of first page indexed keywords as of March 2014			
#	Search query		Indexed keywords	Google	Yahoo Maktoob	Bing
1	جامعة (university)		Exact query (جامعة):	14	6	14
			Basic form with article (الجامعة):	2	0	0
			Basic form with normalized Ta Marbuta (جامعه):	1	8	1
			Plural with article (الجامعات):	2	0	0
2	الجامعة (the university)		Exact query (الجامعة):	9	6	13
			Basic form without article (جامعة):	12	0	0
			Basic with normalized Ta Marbuta (جامعه):	1	0	0
			Plural with article (الجامعات):	1	0	0
3	بالجامعة (at the university)		Exact query (بالجامعة):	0	3	11
			Exact query with normalized Ta Marbuta (بالجامعه):	0	4	0
			Basic form with article (الجامعة):	14	0	0
			Basic form without article (جامعة):	7	0	0
			Basic form with normalized Ta Marbuta (جامعه):	1	0	0
4	لجامعتي (for my university)		Exact query (لجامعتي):	7	13	12
			Query without prefix (جامعتي):	6	0	0
			Query without prefix and normalized Alef Maqsura (جامعتى):	1	0	0
			Exact query with spelling mistake of Alef Madda (لجامعتى):	0	0	1
5	وبالجامعة (and at the university)		Exact query (وبالجامعة):	0	7	5
			Exact query with normalized Ta Marbuta (وبالجامعه):	0	5	5
			Basic form without article (جامعة):	9	0	0
			Basic form with article (الجامعة):	7	0	0
7	للبيت (for the house)		Exact query (للبيت):	9	16	22
			Basic form with article (البيت):	16	0	0

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

6.1.2 First Test Results

The number of retrieved documents shown in Table 3 indicates that Google has the greatest quantitative IR performance, by far, for all search queries in both 2004 and 2014. Bing followed in second place, and Yahoo Maktoob in third place. Over the 10-year period from 2004 to 2014, Google was able to significantly extend its reach. While Google retrieved approximately twice as many documents as AltaVista in 2004, it yielded 38 times more results than AltaVista successor Yahoo Maktoob for Query Set 1 in 2014, and 70 times more results for Query Set 2. These quantitative IR results correspond to previous research findings discussed at the beginning of this paper, in which Google was consistently found to be the search tool with the widest reach (cf. Bushnaq, 2003; Moukdad, 2004; Al-Rawi & Al-Khateeb, 2009).

The most distinctive test outcomes in IR performance over the ten-year period of time for specific search queries will now be discussed. Starting with prefixed Search Query 3 (“بالجامعة”), Google retrieved 5.8% of potentially retrievable documents in 2004 as opposed to 24.5% in 2014. An even larger difference was observed in double-prefixed Search Query 5 (“وبالجامعة”), in which Google’s performance increased from 0.02% to 24.5%. Compared to its predecessor AltaVista, Yahoo Maktoob produced similar performance for these two search queries: 4.5% in 2004 and 3% in 2014 for Search Query 3 and an unchanged 0.02% for Search Query 5. Bing’s IR performance did not differ considerably from that of Yahoo Maktoob in these two examples. A decrease in Google’s IR performance, on the other hand, could be noted for unaffixed Search Query 1 (“جامعة”), as well as for the same word with a definite article in Search Query 2 (“الجامعة”), where the retrieval ratios decreased from 55.2% in 2004 to 39% in 2014 and from 38.9% to 11.5%, respectively.

Query Set 2 confirms that Google’s retrieval performance changed significantly over the last decade for prefixed search queries. While the search engine yielded the majority of results for non-affixed Search Query 6 (“بيت”) in 2004 with a staggering 96%, the ratios have reversed in 2014. The majority of results were returned for the prefixed variant in Search Query 7 (“للبيت”) with 71.4%. In contrast to Google, Yahoo Maktoob showed unchanged high levels of IR performances of around 80% when searching for basic forms as in Search Query 1 and in Search Query 6. Bing’s IR performance in these two cases did once again not differ much from those of Yahoo Maktoob.

In case of both prefixed *and* suffixed Search Query 4 (“لجامعتي”), all search engines generated the fewest results with no more than 0.3% and no remarkable difference compared to the past. It is noteworthy that Bing made an auto-suggestion for Search Query 4, proposing to search for “**بجامعة**” instead of “لجامعتي.”

On the whole, Google in particular showed significant changes in its IR performance over the ten-year period. Its retrieval ratio increased in all cases in which a prefixed Arabic noun was used as the search query (Search Queries 3, 5, and 7). The search engine generally yielded fewer results than a decade ago when the basic word (Search Queries 1 and 6) and the basic word with a definite article (Search Query 2) were entered. Yahoo Maktoob's changes in IR performance compared to its predecessor AltaVista were not as significant as those of Google, by far. The most remarkable changes, as previously noted, occurred in the performance increase for Search Query 1 and the performance decrease for Search Query 2. Bing’s performance was relatively similar to that of Yahoo Maktoob, which may be due to the fact that Yahoo Search is powered by Bing. However, Bing showed a slightly poorer performance compared to Yahoo Maktoob in case of prefixed Search Queries 3 and 7.

While the retrieval ratios of Google and AltaVista were quite similar in 2004, Google now retrieved significantly more data for prefixed search queries than in the past. Google's IR performance for prefixed search queries was also well above those of the other two current search engines tested: Google performed 35% better on average as compared to Yahoo Maktoob and Bing for Search Queries 3, 5, and 7. These findings suggest that Google has implemented major changes in its indexing methods for the Arabic language, which enabled it to increase its retrieval performance, in particular for prefixed Arabic search queries.

Google's indexing and stemming system must differ significantly from those deployed by Yahoo Maktoob and Bing to achieve this significantly better performance for prefixed Arabic search queries. The analysis of the first page indexed keywords in Table 3 explains the differences in IR performance from Table 2 clearly: As for Search Query 3 and 5 (“بالجامعة” and “وبالجامعة”), Google did

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

not index any single exact term the way it was typed in. Instead, most of the indexed keywords are composed of stemmed, basic forms with and without an article ("جامعة"/"الجامعة"). The same holds true for Search Query 7 ("للبيت"), for which about half of the indexed keywords came again from stemmed variants with an article. Yahoo Maktoob and Bing, on the other hand, indexed almost exclusively the exact search terms. Stemming here can only be noted to the point that the Ta Marbuta ("ة") at the end was treated as a stop word, leading to an almost equal number of results for the variant with Ta Marbuta and the one with the substituted letter "H" ("ه") in Search Query 5. Ta Marbuta normalization can be noted for Google, too. However, only a maximum of one normalized keyword was indexed per search query (see Search Query 1, 2, and 3).

Furthermore, the analysis of the first page indexed keywords in Table 3 revealed that Google, by far, deploys the broadest stemming methods among the three multilingual search engines. In almost all the search queries examined, Google Search was able to return the widest range of multiple stemmed variants of the search term. This gap becomes particularly obvious when looking at the indexed keywords in Search Queries 2, 4, 6, 7, and 8. In these cases, Yahoo Maktoob and Bing displayed the results of the exact search query only, and thus missed out on many other relevant documents containing different morphological variants of the search terms.

6.2 SECOND TEST: TRANSLITERATION VARIANTS OF FOREIGN PROPER NAMES, COMMON MISSPELLINGS, AND THE KASHIDA

Settings for the second test were chosen to assess how the three major present-day multilingual search engines Google, Yahoo Maktoob and Bing handle search queries of proper names that were transliterated into Arabic with different spelling variants, search queries that include common spelling mistakes, as well as search queries that are typed in with a Kashida. In particular, the second test aimed at answering the following research questions:

- How do the search engines handle different spelling variants of proper names, of which some variants are more common than others?
- How do the search engines handle search queries with common errors of misspellings and improper word division as compared to their correctly spelled counterparts?
- Do the search engines offer search support tools, such as auto-correction / auto-suggestion, for spelling variants and common spelling errors?
- How do the search engines handle the typographical effect of the Kashida?

6.2.1 Second Test Settings

To answer these questions, five query sets were selected. The query sets were composed of common everyday words in the Arabic language that either represent different spelling variants of a foreign proper name, and a correctly spelled word versus a frequent misspelling (typographical error relating to the transposition of letters and accidental stroke of a wrong keyboard key, frequent substitution of a letter, and improper word division).

Query Set 1 aimed to determine how foreign proper names transliterated into Arabic as different spelling variants are treated by the IR systems. As mentioned in Section 4.2, the Arabic writing system is characterized by the absence of short vowels; only long vowels are written. This feature inevitably leads to several possible transcriptions when names are transferred to Arabic from languages whose vowel system is highly divergent from the phonetic pattern of the Arabic language. Query Set 1 features different transcription variants of the US-American city name of "Los Angeles." Search Queries 1 to 4 differ in the second part of the name "Angeles," which is transliterated with different sequences of long vowels and consonants. Search Queries 5 to 6 are both spelling variants without any long inner-word vowel. The difference between Search Queries 5 and 6 is that in Search Query 6, the diacritical sign Damma was used above the "L" in "Los" to mark the short vowel "o" ("لُس").

Query Sets 2 to 5 represent typical spelling errors of common Arabic nouns. Query Set 2 consists of the correctly spelled Arabic word for "appointment" in Search Query 7, as well as a common misspelling (Shaalan, Allam, & Gomah, 2003, p. 243) in Search Query 8, in which the third

ISSN 2304-0777 www.jitbm.com

and the fourth consonant have been confounded. Query Set 3 is the Arabic term for "almsgiving" in the Islamic religion. Search Query 9 is spelled correctly, and Search Query 10 includes a common typographical error, in which the initial letter "ص" is written with a dot on top as "ض." This typographical error occurs frequently, because the two letters are located adjacent to each other on the Arabic keyboard (Shaalan, Allam, & Gomah, 2003, p. 242). Query Set 4 was chosen to investigate how the search engines handle the substitution of the Arabic letter "Ya" ("ي") with Alef Maqsura ("ى") at the end of a word. As a result of this error, the Arabic word for "lawyer" ("المحامي," Search Query 11) is often misspelled as "المحامى" (Search Query 12) (Abdel-Rasool, 2013, p. 653). Query Set 5 is another typical Arabic-Islamic expression, which can be translated as "God willing." Search Query 13 is correctly spelled as "إن شاء الله" with three individual words, the first word starting with Alef Kasra. Search Query 14 ("انشاء الله") is a common misspelling (Al-Fedaghi & Amin, 1992, p. 180). The first error lays in the improper division of the words: the first and the second word are written as one, resulting in a two-word term. The second spelling error is the missing Hamza below the initial Alef, an equally common misspelling, as mentioned in Section 4.2. This query selection is thus adequate to test how common, incorrectly divided words are treated in the IR process, and to focus on the indexing of diacriticized versus non-diacriticized characters. Finally, Query Set 6 is the Arabic word for "the praise," a term which is very often used in the phrase "praise Allah" ("الحمد لله"). For this test, the word was used with a Kashida before the last letter in order to determine how this typographical elongation is handled by the present-day search engines.

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777
www.jitbm.com

Table 4

Search results for search queries of different Arabic spelling variants, common misspellings, and the Kashida, from Google, Yahoo Maktoob, and Bing as of March 2014. With an analysis of the first page indexed keywords

#	Search query	No. of results by search engine			Auto-suggestions by search engine			First page indexed keywords by search engine					
		Google	Yahoo Maktoob	Bing	Google	Yahoo Maktoob	Bing	Google		Yahoo Maktoob		Bing	
	Query Set 1: Los Angeles												
1	لوس انجيليس	1,030,000	12,100	12,100	Showing results for لوس انجلوس Search instead for لوس انجيليس	---	---	لوس انجيليس لوس انجلوس لوس أنجلس لوس انجليس	0 17 5 5	لوس انجيليس لوس	12 5	لوس انجيليس لوس	12 3
(1a)	Searched instead for لوس انجيليس	85,900,000			Did you mean: لوس انجلوس								
2	لوس انجلوس	995,000	258,000	208,000	---	---	---	لوس انجلوس لوس أنجلس لوس انجليس	17 6 5	لوس انجلوس لوس [انجليس /انجلس] /انجلس/جت]	14 8	لوس انجلوس لوس [انجليس /انجلس]	16 4
3	لوس انجليس	4,360,000,000	149,000	157,000	Did you mean: لوس انجلوس	---	---	لوس انجليس لوس انجلوس لوس انجلس	15 7 2	لوس انجليس لوس [انجلوس/ انجلس]	11 3	لوس انجليس لوس [انجلوس/ انجلس]	14 2
4	لوس انجيلس	1,030,000	3,010	3,010	Showing results for لوس انجلوس Search instead for لوس انجيلس	---	---	لوس انجيلس لوس انجلوس لوس انجلس	5 17 6	لوس انجيلس لوس [انجلوس/انجليس / انجلس / انغليس]	6 10	لوس انجيلس لوس [انجلوس/ انجلس/انجليس]	6 12
(4a)	Searched instead for لوس انجيلس	2,520,000,000			Did you mean: لوس انجلوس								
5	لس انجلس	352,000,000	821,000	622,000	---	---	---	لس انجلس لوس انجلس لوس انجلوس لوس انجليس	0 16 4 2	لس انجلس	23	لس انجلس	29
6	لس انجلس (with diacritical mark Damma)	1,210,000,000	821,000	622,000	---	---	---	لس انجلس لوس انجلوس لوس انجلس لوس انجليس	0 12 7 2	لس انجلس	26	لس انجلس	29

ISSN 2304-0777

	Query Set 2: Appointment												
7	اجتماع (correct spelling)	145,000,000	2,090,000	2,210,000	---	---	---	اجتماع الاجتماع اجتماعا الاجتماعات	12 7 4 1	اجتماع	12	اجتماع	17
8	اجمتاع (spelling error)	145,000,000	1,470	1,460	Showing results for اجتماع	---	---	اجمتاع اجتماع الاجتماع اجتماعا الاجتماعات	0 9 6 4 1	اجمتاع	15	اجمتاع	13
	Query Set 3: Almsgiving												
9	صدقة (correct spelling)	15,000,000	912,000	958,000	---	---	---	صدقة الصدقة صدقه للصدقة	9 10 3 4	صدقة صدقه	8 10	صدقة صدقه	7 13
10	ضدقة (spelling error)	15,000,000	540	538	Showing results for صدقة	---	---	ضدقة صدقة الصدقة صدقه للصدقة	0 8 9 4 4	ضدقة ضدقه	2 7	ضدقة ضدقه	2 7
	Query Set 4: The lawyer												
11	المحامي (correct spelling)	50,400,000	523,000	520,000	---	---	---	المحامي محامي المحامين المحامى محام المحام والمحامي	18 7 6 2 2 1 1	المحامي المحامى	12 1	المحامي	8
12	المحامى (spelling error)	28,300,000	523,000	519,000	---	---	---	المحامى محامى المحامي محامي والمحامي	5 1 16 1 1	المحامى المحامي	1 12	المحامى المحامي	0 13

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

	Query Set 5: God willing									
13	إن شاء الله (correct spelling)	65,400,000	3,110,000	3,100,000	---	---	---	---	---	
14	انشاء الله (spelling error)	25,500,000	4,130,000	4,110,000	---	---	---	---	---	
	Query Set 6: The praise									
15	الحمـــد	146,000,000	3,140,000	3,010,000	---	---	---	---	---	

	إن شاء الله ان شاء الله إن Insha Allah إن (Arabic)	9 9 1		إن شاء الله إنشاء الله ان شاء الله	16 11 8		انشاء الله إنشاء الله إن شاء الله ان شاء الله	15 4 5 1		الحمـــد الحمد حمد
	إن شاء الله ان شاء الله إن يكتب / Insha Allah إن (Arabic) الله (رسول الله	6 8 2 1		انشاء الله إنشاء الله إن شاء الله ان شاء الله الله (إن شاء الله) انشاء (انشاء الله)	15 5 0 0 3 1		انشاء الله إنشاء الله إن شاء الله ان شاء الله	0 11 0		الحمـــد الحمد حمد
	إن شاء الله إنشاء الله ان شاء الله	16 11 8		انشاء الله إنشاء الله إن شاء الله ان شاء الله الله انظر إلى رحمة الله بك / ومنه قول الله تعالى / يقول الله في الحديث إنشاء إذا أشابهين إنشاء / معنى الفعل إنشاء	6 10 11 8 3 2		الحمـــد الحمد حمد	0 31 2	

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

6.2.1 Second Test Results

The second test results confirmed that Google is the search engine with the highest number of retrieved documents per query. Yahoo Maktoob and Bing showed very similar performance, sometimes the one and sometimes the other engine yielded slightly more search results. For the different Arabic spelling variants of the US city name "Los Angeles" in Query Set 1, Google retrieved the most results for Search Query 3, with more than 4 billion documents found. This was followed by Search Query 6 with more than 1 billion documents found. Yahoo Maktoob and Bing achieved the highest number of results for Search Queries 5 and 6. All three search engines retrieved the second highest number of documents with Search Query 6. Google returned the lowest number of results in Search Query 2. Yahoo Maktoob and Bing yielded their lowest numbers in Search Query 4 and Search Query 1. Yahoo Maktoob and Bing were only able to produce 0.3% of the number of documents that Google retrieved with Search Query 4, and 1.2% of what Google yielded with Search Query 1. In Search Query 3, the quantitative gap between the search engines became even more prominent. Whereas Google produced more than 4 billion results, Bing generated only 0.004% (157,000) of the search results as compared to Google. Yahoo Maktoob produced only 0.003% (149,000) of Google's search results.

For the spelling variants in Search Queries 1 and 4, Google provided auto-corrections. It automatically displayed the search results for what it obviously considers the most adequate spelling variant of the city name. The corresponding notification to the user, which appeared above the search results, stated: "Showing results for لوس انجلوس" ("عرض النتائج عن لوس انجلوس"). Below this notification, the user finds an option to search for the term they initially entered. In Search Query 4, for example, this option reads: "Search instead for لوس انجيلس" ("البحث بدلاً من ذلك عن لوس انجيلس"). However, a verification run in Search Queries 1a and 4a, in which the initially entered spelling was used by Google, revealed that the auto-corrected options did not produce a higher number of results. On the contrary, both verification tests yielded far more results than the suggested auto-corrections. Search Query 1a produced 83 times more results than its auto-corrected counterpart in Search Query 1, and Search Query 4a returned even 2,446 times more potentially relevant documents than did Search Query 4.

For Search Query 3, Google notified the user that they might have made a typographical error by giving an auto-suggestion: "Did you mean: لوس انجلوس" ("هل تقصد : لوس انجلوس"). However, this auto-suggestion appeared for the spelling variant that actually yielded the highest quantity of results within Query Set 1. Equally perplexing is the fact that Google generated the lowest number of results in Search Query 2, where the auto-corrected and auto-suggested spelling variant of the city name was actually typed in as search term.

The analysis of the first page of indexed keywords, shown in Table 4, provides more detailed information about the quantitative results discussed above. For the transcriptions in Query Set 1, Yahoo Maktoob and Bing mostly indexed the exact search terms as typed in. This became particularly obvious in Search Queries 5 and 6. In Search Queries 1 to 4, these two search engines partly indexed the word "Los" individually with different words following it (listed in square brackets). The words that followed "Los" were often different spelling variants of "Angeles," which actually led to a broader range of relevant results. In Search Query 2, however, this method resulted in Yahoo Maktoob also indexing a site called "Loose Chat," spelled "لوس جت" in Arabic, an obviously irrelevant search result. Google indexed the widest range of different spelling variants in all cases. Interestingly, Google did not index the exact search term at all for Search Queries 1, 5, and 6, but exclusively indexed other spelling variants.

The results from Search Queries 5 and 6 were highly interesting due to the dramatic difference in the numbers of documents Google retrieved. The non-diacriticized and non-vowelized transcription variant in Search Query 5 returned far fewer results (352,000,000 hits) than the identically spelled term *with* the diacritical mark Damma above the initial "L" ("لُ") in Search Query 6 (1,210,000,000 hits). This means that Google displayed 3.4 times as many results for diacriticized Search Query 6 than it did for non-diacriticized Search Query 5. Yahoo Maktoob and Bing retrieved the same quantity of documents for both search queries. This quantitative gap is not explicable by the first page indexed keywords. As shown in the analysis of the keywords in Table 4, Google did not index the exact search term at all, with or without diacritical sign. In Search Query 5, the most highly indexed spelling variant was " لوس

International Journal of Information Technology and Business Management
29th Aug. 2015. Vol.40 No.1

ISSN 2304-0777 www.jitbm.com

انجلس" (16 hits). The first word "Los" was transcribed with a long vowel, and the second word "Angeles" was non-vowelized. Both words were non-diacriticized. Equally inclusive are the indexed keywords for Search Query 6, in which Google's earlier auto-suggested spelling variant "لوس انجلوس" produced the highest number of hits with 12 indexed keywords. These findings indicate that Google seems to use different indexing methods for Arabic words with and without diacritical marks, but this is not reflected in the indexed keywords.

Query Sets 2 and 3 illustrate how common spelling errors are handled by the search engines. The results make it obvious that Google was the only search tool to have an auto-correction system for common misspellings (cf. Search Query 8 and 9). This automatic spelling correction enabled Google to always produce the same number of results for the incorrectly spelled query as it did for the correctly spelled one: 145 million hits each for Search Queries 7 and 8, and 15 million hits each for Search Queries 9 and 10. Yahoo Maktoob and Bing do not offer any spelling correction. The lack of this feature led to drastically fewer results for the misspelled words in Search Queries 8 and 10, as compared to the correct spellings. Yahoo Maktoob and Bing yielded only 0.07% of the Google results for Search Query 8 and 0.06% for Search Query 10.

Query Set 4 confirmed that all three multilingual search engines are able to tackle the substitution issue of the Arabic letter "Ya" at the end of a word, where it is often written without diacritics. All search engines indexed correctly spelled forms with a diacriticized "Ya" when the search term was entered with the substituted Alef Maqsura. The analysis of the first page indexed keywords indicates that Google and Yahoo Maktoob mainly indexed the correctly spelled variant both in Search Queries 11 and 12. Bing was the only search engine to not index any incorrectly spelled form at all. Yahoo Maktoob and Bing yielded an identical number of results for the correctly spelled term and the incorrectly spelled variant. Google, on the other hand, returned only slightly more than half as many results for incorrectly spelled Search Query 12 as it did for the correct term in Search Query 11. The main reason for this quantitative discrepancy lies in the higher number of morphological variants indexed in Search Query 11, for which seven different morphological forms were indexed with a total of 37 indexed keywords on the first page. In Query 12, only five morphological variants were indexed resulting in a total of 24 indexed keywords on the first page. In Search Query 11, the dual form "المحامين" was indexed 6 times. The search term without an article, "محام," generated 2 hits, and one indexed keyword was a common spelling error (Abdel-Rasool, 2013, p. 653), "المحام" with a missing "ي" at the end of the word. In addition, the indefinite term "محامي" yielded 7 hits in Search Query 11, but only one hit in Search Query 12.

Next, the handling of improper word division was tested with Query Set 5. Google found more results for the correct form in Search Query 13 - approximately 2.5 times as many as for erroneous Search Query 14. Surprisingly, Yahoo Maktoob and Bing displayed slightly more hits for the improperly divided search term. For the three-word variant in Search Query 13, Google also indexed the two-word variant with a quite considerable quantity of 11 times, compared to 24 indexed keywords for its three-word counterpart. Yahoo Maktoob and Bing mostly indexed the exact search term as in previous test queries. Due to the indexing of individual words of the expression, all search engines also indexed irrelevant word combinations, such as "إن يكتب" ("that he writes") or "رسول الله" ("Allah's prophet") (cf. Search Query 13, Yahoo Maktoob). No search engine provided an auto-correction for the improperly divided words. As far as the diacritization of the letter Alef is concerned, the test shows that all search engines ignored the diacritical mark Hamza below the initial Alef. Both diacriticized and non-diacriticized characters are among the indexed keywords.

Test results for Query Set 6, which investigated the handling of the Kashida, show that none of the search engines foundered on the occurrence of this typographical effect. All indexed keywords were written without a Kashida, i.e., the present-day search engines have implemented appropriate functions to systematically ignore the Kashida and treat it as a stop word. This constitutes an improvement of Arabic IR capabilities since Moukdad's 2006 research, when the Kashida posed a problem (Moukdad, 2006).

CONCLUSION

This study examined different aspects of Arabic information retrieval in the three market-dominant,

ISSN 2304-0777 www.jitbm.com

multilingual search engines: Google, Yahoo Maktoob, and Bing. The test results show that Google consistently retrieves the highest number of documents per search query. Google is well ahead of its major market competitors Yahoo Maktoob and Bing, with Yahoo Maktoob performing slightly better than Bing. The comparison of 2004 and 2014 test results for specific Arabic search queries show that Google maintained its significantly wider reach than other search engines. Google was also able to greatly extend its reach by generating up to 70 times more results in 2014 for the same search query. Google now yields a higher percentage of hits in particular for prefixed search terms. In case of prefixed *and* suffixed queries, both Google and Yahoo Maktoob continue to perform poorly with no noticeable difference from ten years ago. However, Yahoo Maktoob now outperforms Google percentage-wise when a non-affixed, basic noun is entered into the search field. Google's quantitative IR performance in these cases has decreased since 2004.

Google is the only engine to provide search support tools, such as auto-correction and auto-suggestion, for common spelling errors and different spelling variants. As far as common typographical errors are concerned, Google's auto-correction is very efficient and enabled it to provide an equally high number of results for correctly and incorrectly spelled search terms. Yahoo Maktoob and Bing do not provide any search support, which leads to a drastically lower number of results in the case of a spelling error. On the other hand, a number of peculiarities were observed in Google's search support for different spelling variants of proper foreign names transliterated into Arabic. The auto-correction did not trigger more relevant results than the entered search term. In addition, diacritcized and non-diacriticized variants led to different quantitative results. Currently, none of the search engines offer correction for common errors of improper word division. All search engines employ character normalization for letters which are often substituted and written without diacritics, such as Alef and Ya. The Kashida is now treated as a stop word by all search engines tested. This no longer presents an IR problem as it did in the past.

The analyses of the first page indexed keywords indicate that Google uses broader stemming than Yahoo Maktoob and Bing. This enables the search engine to index different morphological variants of a search term, resulting in a higher number of indexed keywords per search query. Yahoo Maktoob and Bing mainly index the exact search term. This causes them to miss many other relevant documents containing different morphological variants of the entered query. However, when searches for common multi-part expressions and names were performed, all search engines also indexed individual words only, which led to some irrelevant results.

Overall, Google is the best-prepared multilingual search engine on the market to handle the challenges of Arabic information retrieval. Its outstanding quantitative retrieval performance, as well as its broad stemming and indexing, place Google well ahead of its main competitors, Yahoo Maktoob and Bing, in the number of documents retrieved per search query. With the availability of search support tools, such as auto-suggestion and auto-correction, Google also offers the highest degree of user-friendliness.

References

1. Abdel-Rasool, A. G. (2013). Computational technologies of Arabic language and its contribution in e-learning – study on grammar and spelling proofing tools. *US-China Education Review A, September 2013, 3*(9), 649-660.

2. Abdelali, A. (2006). *Improving Arabic information retrieval using local variations in Modern Standard Arabic*. Socorro, NM: New Mexico Institute of Mining and Technology.

3. Abdelali, A., Cowie J. & Soliman, H. S. (2004). Arabic information retrieval perspectives. *Proceedings of the 11th Conference on Natural Language Processing, JEP-TALN*, 119-132. Retrieved from http://aune.lpl-aix.fr/jep-taln04/proceed/actes/arabe2004/TAAA13.pdf

4. Abu-Salem H., Al-Omari M. & Evens, M. (1999). Stemming methodologies over individual query words for an Arabic IR system. *Journal of the American Society for Information Science, 50*, 524 – 529.

5. Al Bawaba (2006). Maktoob beats competition and strengthens commitment: Araby is first Arabic search engine. Retrieved from http://www.thefreelibrary.com/Maktoob+beats+competition+%26+strengthens+commitment%3A Araby+is+first...-a0183539445

ISSN 2304-0777 www.jitbm.com

6. ALDayel, A. & Ykhlef, M. (2013). Arabic users' attitudes toward Web searching using paraphrasing mechanisms. *International Research Journal of Computer Science and Information Systems (IRJCSIS), 2*(2), 34-39.

7. Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M.S. & Al-Rajeh, A. (2008). Automatic Arabic text classification. *Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data,* 77-83. *Retrieved from http://eprints.soton.ac.uk/272254/1/Arabic-Classification.pdf*

8. Aljlayl, M. & Frieder, O. (2002). On Arabic search: improving the retrieval effectiveness via a light stemming approach. *Proceedings of the 11th International Conference on Information and Knowledge Management*, 340-347. Retrieved from http://wenku.baidu.com/view/c595ad250722192e4536f680.html

9. Al-Fedaghi, S. & Amin, A. (1992). Automatic correction of spelling errors in Arabic. *Journal of Kuwait University (Sciences), 19*(2), 175-194. Retrieved from http://pubcouncil.kuniv.edu.kw/kjs/files/10Apr2013104941Automatic%20correction%20of%20spelling%20errors%20in%20Arabic.pdf

10. Al-Maimani, M.R., Naamany, A.A., & Bakar, A.Z.A. (2011). Arabic Information Retrieval: Techniques, tools and challenges. IEEE GCC Conference and Exhibition (GCC), 541-544. doi:10.1109/IEEEGCC.2011.5752576

11. Al-Maskari, A., Sanderson, M. & Clough, P. (2007). Arabic users' satisfaction with the online information as obtained from Google. *Proceedings of the 6th International Conference on Conceptions of Library and Information Science (CoLIS).* Retrieved from http://www.seg.rmit.edu.au/mark/publications/my_papers/CoLIS07.pdf

12. Al-Omar, N.A.M. (2013). Transliterating proper nouns: facts and implications. *Linguistics, Culture & Education, 2*(1), 119-132.

13. Al-Rawi, S. S. & Al-Khateeb, B. (2009). Comparison the efficiency of some search engines on Arabic keywords and roots. *Proceeding of the 2nd Internation Conference on Applications of Digital Information and Web Technologies (ICADIWT '09),* 163-168. doi:10.1109/ICADIWT.2009.5273982

14. Al-Shalabi R., Kanaan G., & Al-Serhan, H. (2003). New approach for extracting Arabic roots. Proceedings of the International Arab Conference on Information Technology (ACIT'2003), 42-59.

15. Bar-Ilan, J. & Gutman, T. (2003). How do search engines handle non-English queries? – A case study. *Proceedings of the International World Wide Web Conference (WWW)*, 78-87. Retrieved from http://www2003.org/cdrom/papers/alternate/P415/415.pdf

16. Arabterm, United Nations Multilingual Terminology Database of the Arabic Translation Service (2014). Retrieved from http://unterm.un.org/DGAACS/arabterm.nsf/WebView/A4D1BB323BC0478285256D4E005884BF?OpenDocument and http://unterm.un.org/DGAACS/arabterm.nsf/WebView/CD7986F447AEE92485256D4E00589FF5?OpenDocument

17. ATA Software Ltd. (nd). 20 years of Arabic software development. Retrieved from http://www.atasoft.com/documents/33.html

18. Basis Technology. (2009). Ayna.com, the most visited Arabic search engine, selects Basis Technology for multi-language text analysis [Press release]. Retrieved from http://www.basistech.com/about-us/news/press-releases/2009-12-02-ayna-search-engine/

19. Boualem, M. & Abbes, R. (2008). Information retrieval in Arabic language. Retrieved from http://www.malek-boualem.info/wp-content/uploads/sites/2/2013/03/Malek-BOUALEM-Ramzi-ABBES-v2.pdf

20. Brin, S. & Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine. Proceedings of the 7th International World Wide Web Conference (WWW),* 107-117. doi:10.1016/S0169-7552(98)00110-X

21. Bushnaq, A. (2003). Evaluation arabischer Webseiten: Informationsangebote im Bereich Medien und Kultur. *Ibn Rushd Forum for Freedom of Thought, 4*. Retrieved from http://www.ibn-rushd.org/forum/suchmas.html

ISSN 2304-0777 www.jitbm.com

22. Chen A. & Gey, F. (2002). Building an Arabic stemmer for information retrieval. *Proceedings of the 11th Text Retrieval Conference (TREC 2002), NIST*, 631-639.

23. eBizMBA. (2014). Top 15 Most Popular Search Engines | March 2014. Retrieved from http://www.ebizmba.com/articles/search-engines

24. Google Official History, Comscore (2014). Google Annual Search Statistics. Retrieved from http://www.statisticbrain.com/google-searches/

25. Harmanani, H.M., Keirouz, W.T. & Raheel, S. (2006). A rule-based extensible stemmer for information retrieval with application to Arabic. *The International Arab Journal of Information Technology, 3(3)*, 265-272.

26. Internet World Stats. (2011). Internet world users by language. Retrieved from http://www.internetworldstats.com/stats7.htm

27. Internet World Stats. (2012). Internet penetration in the Middle East. Retrieved from http://www.internetworldstats.com/stats5.htm

28. Khoja, S. & Garside, R. (1999). *Stemming Arabic text*. Lancaster, UK: Computing Department, Lancaster University.

29. Larkey, L. S., Ballesteros, L. & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*, 275-282.

30. Larkey, L. S., Ballesteros, L. & Connell, M. E. (2005). Light stemming for Arabic information retrieval. Arabic Computational Morphology: Knowledge-based and Empirical Methods. Retrieved from https://www.academia.edu/5217995/Stemming_techniques_of_Arabic_Language_Comparative_Study_from_the_Information_Retrieval_Perspective

31. McCarthy, J. & Prince, A. (1990). Foot and word in prosodic morphology: the Arabic broken plural. *Natural Language and Linguistic Theory, 8*(2), 209-283.

32. Moukdad, H. (1999). An investigation of the necessity of information retrieval algorithms for full-text Arabic databases. *Proceedings of the 27th annual conference of the Canadian Association for Information Science (CAIS 1999)*, 207-227.

33. Moukdad, H. (2004). Lost in cyberspace: how do search engines handle Arabic queries? *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science (CAIS 2004)*, 1-7.

34. Moukdad, H. (2006). Stemming and root-based approaches to the retrieval of Arabic documents on the Web. *Webology*, ***3***(1). Retrieved from http://www.webology.org/2006/v3n1/a22.html

35. Mujoo, A., Malviya, M. K., Moona, R. & Prahakar, T. (2000). A search engine for Indian language. *Proceedings of the 11th International Conference on Database and Expert Systems Applications (EC-WEB 2000)*, 349-358. Retrieved from http://link.springer.com/chapter/10.1007%2F3-540-44463-7_30

36. Nationalencyklopedin. (2010). Världens 100 största språk 2010. Retrieved from http://www.ne.se/spr%C3%A5k/v%C3%A4rldens-100-st%C3%B6rsta-spr%C3%A5k-2010

37. Omar, M. A. H. & Ma, S. (2009). Stemming algorithm to classify Arabic documents. Journal of Communication and Computer 7(9) (2010/09) Symposium on Progress in Information & Communication Technology, 111-115. Retrieved from http://spict.utar.edu.my/SPICT-09CD/contents/pdf/SPICT09_B-1_1.pdf

38. Rotaru, A. (2011, June 21). The foreign language internet is good for business [Blog post]. Retrieved from http://www.scottmclay.co.uk/foreign-language-internet-good-business/

39. Sakhr Software. (2009). IDRISI search engine. Retrieved from http://www.sakhr.com/images/datasheets/idrisi.pdf

40. Sanan, M., Rammal, M. & Zreik, K. (2008). Internet Arabic search engine studies. *Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, 1-8.

41. Shallan, K., Allam, A. & Gomah, A. (2003). Towards automatic spell checking for Arabic. *Proceedings of the 4th Conference on Language*

ISSN 2304-0777 www.jitbm.com

Engineering, Egyptian Society of Language Engineering (ELSE), 240-247.

42. Singh, N. & Pereira, A. (2005). *The culturally customized Web site: customizing Web sites for the global marketplace*. Oxford, UK/Burlington, MA: Elsevier Butterworth-Heinemann.

43. Sottimano, D. (2012, August 2). Google ccTLDs and associated languages & codes reference sheet [Blog post]. Retrieved from https://www.distilled.net/blog/uncategorized/google-cctlds-and-associated-languages-codes-reference-sheet/

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

45. Tawileh, W., Mandl, T. & Griesbaum, J. (2010). Evaluation of five web search engines in Arabic language. *Proceedings of Workshop Information Retrieval* (*LWA 2010)*. 221-228.

46. WeSearch. (2014). Why is Google so successful? Retrieved from http://www.wesearchsg.org/why-is-google-so-successful.php

47. Xu, J., Fraser, A. & Weischedel, R. (2002). Empirical studies in strategies for Arabic retrieval. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, 275-282.

48. Yahoo!. (2014). Yahoo international. Retrieved from http://everything.yahoo.com/world/

49. Yahoo! Help. (2013). The Yahoo and Microsoft search alliance. Retrieved from https://help.yahoo.com/kb/search/SLN2251.html?impressions=true

50. List of Tables

List of Figures