

Comparison the Efficiency of Some Search Engines on Arabic Keywords and Roots

Salah S. Al-Rawi

College of Computers, Al-Anbar University
Ramadi, Iraq
dr_salah_rawi@yahoo.com

Belal Al-Khateeb

College of Computers, Al-Anbar University
Ramadi, Iraq
bird79_79@yahoo.com

Abstract

This paper presents an attempt to show the efficiency of some search engines in dealing with Arabic keywords. This can be achieved by comparing the number of retrieved pages, retrieving time, and stability (in both the number of retrieved pages and the order for each retrieved page) for each one of the selected 20 Arabic keywords (with its roots) that were entered to the selected four search engines at the same time. Google, Yahoo, Al-hoodhood and Ayna were selected as a test bed for the experiment. The obtained results showed that Google was the best search engine among the four selected search engines, the experiments for the results of the stability of the selected search engines took 10 weeks to be obtained.

Keywords: Search engines, Internet, information retrieval, World Wide Web

1. Introduction

Arabic language is being increasingly used on the Internet, despite of significant obstacles. Most Internet users, who for the first time try to read Arabic web sites, have to face difficulties. Most of the difficulties arise from the multiple character sets for representing Arabic and the characteristics of the Arabic script [1].

With the continuing explosive growth of the Internet and the spread of textual information in a multitude of languages other than English on the web, retrieval of documents in these languages is becoming an increasingly significant problem. Rules, theories, algorithms, and retrieval methods designed and developed for English and other morphologically similar languages may or may not apply in different linguistic environments. Nowhere could the problem be sharper than in languages that differ radically from English in morphology and word-formation rules. Words, being the gist of written and spoken information queries, are by far the most fundamental elements of expression, and they form basic components of meaningful information exchanges [2].

Most of available electronic databases were in English, search and retrieval software, indexing methods, and user interfaces were designed specifically for this language. As this is no longer the case, Information Retrieval (IR) systems have been developed for languages other than English, and search engines have increasingly been modified to handle these languages [3][4][5].

Arabic is one language that is likely to present challenges in a traditional IR environment and in popular search engines, because its morphology and words formation rules are radically different from those of English. These rules are based on a root – and – pattern system that has been long thought to be a major factor in hindering IR operations. Finding all possible words that are derived from an Arabic root might not necessarily lead to better IR performance. While researchers on Arabic IR advocated the use of advanced word stemming and root extraction algorithms, the limited scope of their research leave many questions unanswered [6].

This paper explores the handling of Arabic words in English and Arabic search engines and retrieval environment represented by Google, Yahoo, Al-hoodhood, and Ayna, and it presents specific approaches to assessing stemming and root-based retrieval methods to accommodate the peculiarities of Arabic word formation rules within the framework of this environment.

This paper is organized as follows. Section 2 briefly presents the information retrieval. Section 3 describes the search engines. Section 4 describes the implementation that has been done. Experimental designs and their results are discussed in section 5, while section 6 gives the concluding remarks of this work. Finally, section 7 presents some suggestions for future work.

2. Information Retrieval (IR)

Interest in Arabic IR did not materialize until the 1990s. Before that, specialists in Arabic computing focused their efforts on presenting the language in a computer environment and finding solutions for display and coding problems, in the early 1990s. This changed, and research started to appear on the automation of Arabic online library catalogs and on

IR issues [6]. IR involves many strategies each one come with its own features, which can be used to retrieve information effectively, among of these strategies are: *Boolean Search*, *Serial Search*, and *Cluster-Based Retrieval* [7].

Compared to English, redundancy in Arabic was assumed to be higher, because Arabic words are derived from roots according to certain patterns, depending on fixed rules, in addition to suffixes, prefixes and infixes [3]. Also by comparing the results with these from research on English, Arabic was found to have a greater redundancy, and the average word length for Arabic is greater than English, making Arabic potentially more compressible than English [6].

Arabic documents were best indexed by word roots, because root indexing increased recall and bypassed complex problems created by Arabic morphology, a root index term would retrieve all variation of this root and eliminate the need to enter complex search queries [8].

3. Search Engine

Search engine technology has to scale dramatically to keep up with the growth of the web such as the increase number of web pages, documents and web queries posted on the Internet [9] [10].

Evaluation of information retrieval system for the World Wide Web (WWW) environment is a difficult task. The difficulty stems out from the unavailability of standard test data and also the highly subjective nature of the notion of relevancy of WWW pages retrieved with respect to the user's information needs [11].

Precision is always reported in formed information retrieval experiments. However, there are variations in the way it is calculated depending on how relevance judgments are made [12].

Bar-Ilan [12] conducted several studies to investigate the search engine stability problems and defined several measures to evaluate search engine functionality over time. Bar-Ilan's measures are based on the technical relevance concept which is the document defined to be technically relevant if it fulfils all the conditions posed by the query [13].

Search engines are updated using a tool commonly reserved to as a spider on robot spiders clean hundreds of thousands of pages a day. Many of them will also follow the links on a page to find information independently. Thus it is possible for a web site to be indexed by a spider even if the web site was not submitted to the search engine [14].

Search engines such as Google [15] [16] is designed to avoid disk seeks whenever possible, and

this has a considerable influence on the design of data structures. In Google the web crawling (downloading of web pages) which is the backbone to the search engine is done by several distributed crawlers. There is a URL (Uniform Resource Locator) server that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the store server which then compresses and stores the web pages into a repository [17].

They many only use a small database from which to create a set of results to the users (Yahoo for example only indexes a very small proportion compare to a billion pages indexes by Google) or they may not be updated particularly quickly (All the web is updated every fortnight or so, while Google is updated monthly). These spider programs may not be very fast, which means that their currency might not be a real reflection of the state of play on the Internet [8] [18].

4. Implementation

In order to maintain a good comparison for Arabic keywords, four search engines were selected for this research, two of them are general search engines (Google [19][20][21][22] and Yahoo [23]) while the others are an Arabic language Search engines (Al-hoodhood [24] and Ayna [25]) that employs stemming and root indexing. The reason for choosing these search engines is that they are widely used as general search engines. The four selected search engines were used to search for a specified word, search for a specified word by its root, and figure out the stability of each search engine in form of number of retrieved pages and the order of each one. Search was designed to compare the performance of Google with Yahoo, Al-hoodhood and Ayna, and evaluate stemming as an alternative to root retrieval.

The experiments were done by using a computer with 1.7 GHz processor, 256 MB RAM, and windows XP operating system.

5. Results and Discussion

This research was being done as a two phase process. The first part is used to determine the speed of loading results, to conduct this phase, twenty different Arabic words were selected, each one with its root, each word will entered as an input in the four selected search engines at the same time and the resulted total number of pages and retrieval time were recorded. Table 1 show the selected (20) words entered all at the same time to the four search engines, and the number of results from each search engine with the relative time spent for searching and retrieving the results.

This process is repeated for the roots of the selected words (as shown in table 2). The purpose of this phase is to maintain a good comparison between the selected search engines in the number of retrieved pages and time point of view, this can be done by summing up the number of the retrieved pages for all the entered search keywords to get the total number of retrieved pages (Total-Pages) and also summing up the time required to retrieve each keyword to get the total time of retrieving (Total-Time) then divide Total-Pages over Total-Time and sort the results in ascending order for the four search engines to know which one of the four selected search engines is faster in retrieving (first one is faster than the second and so forth). This procedure was also applied to table (2) to get knowledge about which search engine is the best from retrieving time point of view.

The second part worked as follows: take five words out of the selected 20 words with its roots, and for each selected word and its root, search for the results in the selected four search engines at the same time, repeat this process for ten weeks and keep the number of retrieved pages for each week. There is no need to keep track of the retrieving time for the selected words at this part as the purpose of this phase is to compare our selected search engines from the stability in retrieving results point of view. So to conduct this phase, we also record the first twenty pages that resulted from each search engines for every week of the ten weeks period. Table (3) shows the stability of each search engines from the point of the number of the retrieved web pages for each word of the selected five words. While tables (4 and 5) and figures (1 and 2) show the stability of each search engines from the point of the order of the retrieved web pages for each word of the selected five words. The results in tables (4 and 5) were calculated by making the twenty pages resulted in the first week as our measure to find how stable the search engine in retrieving the same web pages or not, for example in table 4, Google in the second week retrieve eleven pages from the twenty that were retrieved in the first week, while Yahoo retrieved only four in, and Al-hoodhood retrieved 20, finally Ayna retrieved 13 for the same week. That reflect two points, the first one that Al-hoodhood and Ayna are more stable than Google and Yahoo, while for the second one we clearly see that Google and Yahoo are more flexible in updating their databases (by adding new pages for the same subject).

6. Conclusion

From analyzing tables 1 and 2, by summing up the results of each search engine and dividing it by the

sum of the retrieving time, one can conclude that Google is the best search engine in dealing with Arabic keywords. Yahoo is the second, while Ayna comes third and Al-hoodhood is the last one. The results show that Google is faster and can retrieve a large number of results comparing with others, and that reflects that even there are many search engines that are special in dealing with Arabic keywords, but these search engines still have a limited ability comparing with the general purpose ones (Google and Yahoo).

From analyzing table 3, we can see that Google is the best search engine in dynamic update of web pages with stability in dealing with Arabic keywords. Yahoo is the second, while Ayna comes third and Al-hoodhood is the last one (no update occurs in al-hoodhood during the search time). The results shows that Google has the ability of rapid dynamic update to its database in a short time comparing with others, while one can easily notice that Al-hoodhood are the slower one in the that update.

From analyzing tables 4 and 5, we can conclude that Google is the best search engine in retrieving the same results from week to week with dynamic update of web pages in dealing with Arabic keywords. Yahoo is the second, while Ayna comes third and Al-hoodhood is the last one (no update occurs in al-hoodhood during the search time).

7. Future Work

A web search engine is a very rich environment for research ideas. These issues will be looked at in an attempt to define a way to search the web in a more meaningful manner. The present and future issues in developing a web search are:

- 1- Designing smart algorithms to decide what old web pages should be re-crawled and what new ones should be crawled.
- 2- Developing a metasearch engine that improves the efficiency of web searches by downloading and analyzing each document and then displaying results that show the query terms in context. This helps users more readily determine if the document is relevant without having to download each page.
- 3- For solving Arabic language problems, we must be able to handle Unicode, which is just one out of several possible encoding sets.
- 4- Another important consideration is how the system handles simultaneous search and database updates/indexing in real time. Most current web search systems use some very limited "parallel processing"

techniques and replication technology to handle performance scalability issues.

- 5- Supporting query refining.
- 6- Add more search engines together with using additional samples in the experiments.

8. References

- [1] Sanan, M., Rammal, M., Zreik, K., Internet Arabic Search Engines Studies, *3rd International Conference on Information and Communication Technologies: from Theory to Applications*. ICTTA, Damascus, 2008, pp. 1-8
- [2] Maier, Robert J., Glossary of Information-Retrieval Terminology, *IEEE Transactions on Engineering Writing and Speech*, 1970, pp. 34-36.
- [3] Moukdad, H., Stemming and root-based approaches to the retrieval of Arabic documents on the Web, *Webology*, 2006, Article 22.
- [4] Douglas Comer, *Computer Networks and Internets with Internet applications*, prentice hall international, INC., USA, 2008.
- [5] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, UK, 2008.
- [6] Saba Abdul Khaliq Al-Khadady, Internet and Arabic Search Engines M.Sc. Thesis, Iraq, 2002.
- [7] Van Rijsbergen C. J., *Information Retrieval*, Butterworth, London, 1979.
- [8] Khalid Shaker Jassim, Comparison of Efficiency of some search Engines on the Internet, M.Sc Thesis, Iraq, 2005.
- [9] Lewis Mackenzie, *Communication and Networks*, McGraw-Hill, USA, 1998.
- [10] Stott D. And Moran D., *Information and Communication*, Springer, London, 2000.
- [11] Massimo Marchiori, *The Quest For correct information on the web: Hyper search Engines*, Department of Pure Application Mathematics University of Padova, Italy, 2000.
- [12] Bar-Ilan J., Evaluating the stability of the search tools Hotbot and Snap: a case study, *Online Information Review*, Emerald, Bradford, ROYAUME-UNI, INIST-CNRS, Cote INIST, 2000, pp. 439-450.
- [13] Mike Thelwall, The Responsiveness of Search Engine Indexes, *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics* 2001.
- [14] Mark Levene, *An Introduction to Search Engines and Web Navigation*, Pearson Education, UK, 2006.
- [15] Danny Sullivan, Search Engine Features For Webmasters [online] available from <<http://searchenginewatch.com/showPage.html?page=2167891>> [5 Dec 2002].
- [16] Danny Sullivan, How Search Engines Work [online] available from <<http://searchenginewatch.com/showPage.html?page=2168031>> [14 Mar 2007].
- [17] Sengey Brin and Lawrence page, *The Anatomy of large-scale Hypertextual web search Engine*, Computer Science Department, Stanford University, 1994.
- [18] Multi-search Engines - a comparison [online] available from <<http://www.philb.com/msengine.htm>> [2003].
- [19] Google [online] available from <<http://en.wikipedia.org/wiki/Google>>.
- [20] All About Google [online] available from <http://www.google.com/about.html>.
- [21] Google Help Central [online] available from <http://www.google.com.au/help>.
- [22] Danny Sullivan, Major Search Engines and Directories [online] available from <<http://searchenginewatch.com/showPage.html?page=2156221>> [28 Mar 2007].
- [23] Linda Barlow, A Helpful Guide to Search Engines [online] available from <<http://www.monash.com/spidap3.html>> [5 Nov 2004].
- [24] <http://www.alhoodhood.com/about.html>.
- [25] <http://www.aynacorp.com/About/6.html>.

Table 1: Loading speed of the selected search engines on Arabic Keywords

	Google		Yahoo		Al-hoodhood		Ayna	
	Results	Time (sec.)	Results	Time (sec.)	Results	Time (sec.)	Results	Time (sec.)
يبدأ	2,630,000	0.55	1,480,000	0.22	41,904	2.000	3,182,550	0.991
قتلت	531,000	0.37	338,000	0.14	13,007	1.000	641	0.7369
مظلوم	810,000	0.35	291,000	0.17	1,879	1.000	599,270	0.5383
علمنا	625,000	0.38	342,000	0.11	8,058	1.000	715	0.2934
منطق	1,620,000	0.18	621,000	0.13	10,543	1.000	1,308,300	0.4291
الموتى	378,000	0.17	300,000	0.17	8,715	1.000	5,366,480	0.4687
شهادة	2,260,000	0.19	1,060,000	0.17	29,384	1.000	5,488,980	0.2081
الصاعقة	109,000	0.21	59,400	0.10	11,480	1.000	326	0.4435
بكاء	320,000	0.02	213,000	0.16	3,728	1.000	1,058,400	0.3454
حكما	227,000	0.08	348,000	0.12	4,298	1.000	301	0.2497
مخطوف	20,000	0.36	8,190	0.45	158	1.000	182	0.0489
بيوت	835,000	0.90	456,000	0.12	12,210	1.000	15,224,790	0.396
كتابة	8,400,000	0.35	7,290,000	0.15	281,677	1.000	26,972,050	1.154
نجوم	3,240,000	0.55	1,850,000	0.10	19,879	1.000	11,157,300	0.2697
تنزيل	1,490,000	0.53	725,000	0.08	7,249	1.000	3,369,240	0.2071
صورة	11,700,000	0.07	4,770,000	0.12	124,130	2.000	68,771,010	0.2313
اخبار	23,000,000	0.64	18,600,000	0.11	50,881	1.000	69,492,290	0.308
زمان	11,600,000	0.49	4,530,000	0.10	22,911	1.000	7,636,160	0.2994
ايام	3,600,000	0.57	5,240,000	0.11	28,853	1.000	8,698,480	0.2855
سلام	9,220,000	0.34	3,460,000	0.12	69,798	1.000	25,194,820	0.1831

Table 2: Loading speed of the selected search engines on Arabic roots

	Google		Yahoo		Al-hoodhood		Ayna	
	Results	Time (sec.)	Results	Time (sec.)	Results	Time (sec.)	Results	Time (sec.)
بدا	3,490,000	0.45	2,720,000	0.11	78,448	3.000	4,975,460	0.2354
قتل	4,080,000	0.44	2,920,000	0.23	87,941	4.000	5,375,300	0.2518
ظلم	1,650,000	0.09	730,000	0.16	11,548	1.000	2,130,520	0.227
علم	9,520,000	0.14	4,310,000	0.27	210,369	9.000	45,192,700	0.8705
نطق	624,000	0.26	266,000	0.22	5,301	1.000	676,690	0.5099
مات	2,790,000	0.11	1,240,000	0.20	43,565	2.000	2,681,770	0.1814
شهد	1,530,000	0.43	947,000	0.11	26,478	2.000	2,329,950	0.4935
صعق	28,800	0.12	23,800	0.38	156	1.000	86	0.2234
بكي	331,000	0.14	240,000	0.15	3,661	1.000	415	0.708
حكم	5,250,000	0.24	3,440,000	0.27	149,919	7.000	12,096,630	1.7774
خطف	403,000	0.18	280,000	0.13	6,412	1.000	886,410	0.2275
بيت	6,800,000	0.04	3,850,000	0.27	89,188	1.000	15,803,970	0.6742
كتب	10,400,000	0.04	5,222,000	0.24	304,339	1.000	33,520,410	0.504
نجم	1,730,000	0.26	1,090,000	0.18	18,752	1.000	10,378,200	0.3234
نزل	1,180,000	0.28	755,000	0.21	26,647	1.000	941,290	0.2456
صور	17,000,000	0.19	7,330,000	0.09	140,149	2.000	74,913,160	0.1938
خير	20,200,000	0.51	8,630,000	0.15	39,221	1.000	70,907,410	0.4054
زمن	2,940,000	0.25	1,620,000	0.22	42,182	2.000	186	0.0449
يوم	19,500,000	0.37	12,900,000	0.03	291,731	15.000	91,490,840	0.5056
سلم	1,520,000	0.34	982,000	0.20	44,441	5.000	5,410,090	0.3271

Table 3: The stability of the four search engines on the selected 5 words in terms of retrieved pages

	Google	Yahoo	Al-hoodh	Ayna
Week1	403,000	280,000	6,412	886,410
Week2	445,000	329,000	6,412	1,046,640
Week3	493,000	238,000	6,412	677,670
Week4	482,000	252,000	6,412	1,335,250
Week5	667,000	338,000	6,412	1,335,250
Week6	510,000	240,000	6,412	1,335,250
Week7	402,000	220,000	6,412	1,214,710
Week8	475,000	304,000	6,412	1,089,760
Week9	462,000	335,000	6,412	861
week10	423,000	300,000	6,412	862

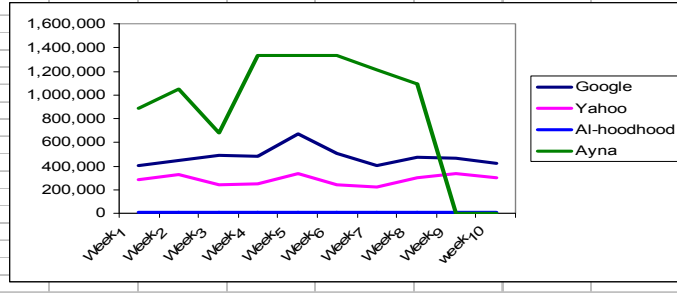


Table 4: The stability of the four search engines on the selected 5 keywords in terms of the order of retrieved pages

	Google	Yahoo	Al-hoodhood	Ayna
Week1	20	20	20	20
Week2	11	4	20	13
Week3	11	7	20	11
Week4	0	15	20	5
Week5	0	16	20	14
Week6	14	13	20	20
Week7	11	6	20	18
Week8	0	2	20	6
Week9	8	14	20	20
Week10	15	18	20	20

Table 5: The stability of the four search engines on the selected 5 roots in terms of the order of retrieved pages

	Google	Yahoo	Al-hoodhood	Ayna
Week1	20	20	20	20
Week2	11	6	20	7
Week3	8	11	20	8
Week4	0	15	20	5
Week5	0	11	20	8
Week6	10	9	20	20
Week7	10	11	20	20
Week8	0	5	20	17
Week9	15	19	20	4
Week10	18	19	20	16

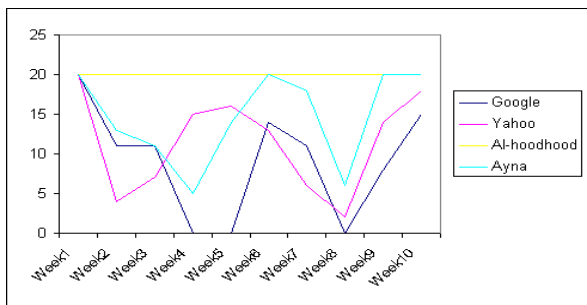


Figure 1: The stability of the four search engines on the selected 5 keywords in terms of the order of the retrieved pages

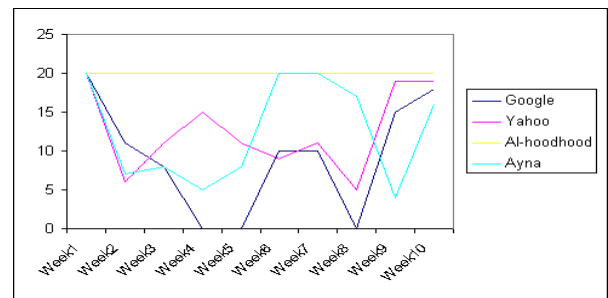


Figure 2: The stability of the four search engines on the selected 5 roots in terms of the order of the retrieved pages