Evaluation of five web search engines in Arabic language

Wissam Tawileh, Thomas Mandl and Joachim Griesbaum
University of Hildesheim
D-31141, Hildesheim, Germany
tawilehw@uni-hildesheim.de

Abstract

To explore how Arab Internet users can find the information in their mother tongue on the web, the five web search engines Araby, Ayna, Google, MSN and Yahoo were tested on an information retrieval evaluation basis with the consideration of the web-specific evaluation requirements. The test used fifty randomly selected queries from the top searches on the Arabic search engine Araby. The relevance of the top ten results and their descriptions retrieved by each search engine for each query were evaluated by independent jurors. Evaluations of results and descriptions were then compared to assess their conformity. The core finding was that Google performed almost all the times better than the other engines. The difference to Yahoo was however not statically significant, and the difference to the third ranked engine MSN was significant to a low degree. The Arabic search engine Araby showed performance on most of the evaluation measures, while Ayna was far behind all other search engines. The other finding was the big differences between search results and their descriptions for all tested engines.

1 Introduction

This work represents an evaluation test performed on multiple web search engines which can deal with Arabic and the specific needs of this language and its speakers as the target users group considered in the test.

The question this test attempts to answer is: which web search engine of the five tested in this study can retrieve the “best” search results for the user judging these results?

The test compares the effectiveness of five different web search engines, two of which are native Arabic engines: Araby and Ayna, and three are international Arabic-enabled engines Google, MSN and Yahoo.

The motivation of the test is the lack of evaluative research of Arabic information retrieval systems, especially on the internet. This is despite the very high growth rates of internet users in the Arab countries, and that most users can not read English which dominates the content on the World Wide Web [Hammo, 2009].

2 Related Work

The evaluation of web search engines has been gaining increased importance and research interest since its early beginnings in 1996. A large number of evaluation experiments has been performed to assess the performance of search engines from different perspectives using varied evaluation measures and test designs.

Most evaluation tests used search queries in English as a dominant language on the web. However, many tests in other languages are also taking place, focusing mainly on the way search engines deal with different languages, their linguistic issues and proper search algorithms aiming to improve the multilingual capabilities of search engines.

Other studies, like this work, focus on the performance evaluation of web search engines from the local users’ point of view. An overview on non-English web information retrieval studies is presented by Lazarinis and others [Lazarinis, 2007; Lazarinis *et al.*, 2009].

Several studies (like [Moukdad and Large, 2001], [Moukdad, 2002; 2004], [Abdelali *et al.*, 2004] and recently [Hammo, 2009]) discussed Arabic information retrieval on the web.

Gordon and Pathak [1999] discussed a collection of web search engines evaluation tests conducted since 1996 against their test methodology and purpose.

Another overview aligned with the recommendations of Tague-Sutcliffe [1992] for general information retrieval evaluation, and Hawking et al. [2001] for web-specific information retrieval evaluation criteria is presented in [Lewandowski, 2008].

The most recent two studies similar to this work in their methodology and design are:

Griesbaum tested three search engines (Google, Lycos and AltaVista) using 50 queries, in German language [Griesbaum, 2004]. Google came in the first place in the overall performance judgment followed by Lycos with no significant difference, and AltaVista came in the last place with a higher difference to Google, but not to Lycos.

The second study has been conducted by [Lewandowski, 2008]. He evaluated five search engines (Google, Yahoo, MSN, Ask.com and Seekport). A set of forty queries, in German language, was created by forty faculty students who were the jurors as well. The study found no significant reason to favor any of the major search engines in terms of performance and concluded that more attention should be paid by search engine companies to the quality of results descriptions.

3 Test Methodology

Tague-Sutcliffe [1992] presented a guide for information retrieval evaluation which helps the experimenters in making the required decisions while planning an evaluation test to ensure the validity of the experiment, the reliability of the results and the efficiency of the test procedures. It assumes to answer ten questions which are discussed here for this study.

3.1 Need for testing

Keeping in mind the limited knowledge of foreign languages, especially English, among Arab internet users, many questions can be stated considering the effectiveness and efficiency of searching the web in Arabic using local Arabic search engines or international Arabic enabled ones.

While these questions can cover the search systems themselves (search algorithms, language handling issues, resources consumption… etc.), another aspect can also be questioned, which is the informativeness and effectiveness of web search engines in the eyes of Arab internet users.

To the best knowledge of the authors, there is no similar published evaluation test before this work.

3.2 Type of test

The test evaluates five online commercial search engines in the Arabic language working on the web as an operational database.

This kind in real use differs from the experimental tests performed in a laboratory environment based on the Cranfield paradigm [Mandl, 2008]. Such tests allow a higher level of control on tests parameters and variables, whereas the test presented here gives an assessment closer to real life and the users' perception of results.

3.3 Variables definition

The independent variables which affect the results of this test are: evaluation criteria, relevance performance measures, queries, information needs and participants (also referred to as users or jurors). These variables are defined in this section to illustrate the general settings of the test.

The dependent variables are the relevance judgments of the jurors which are the indicator of the retrieval performance in this test. These are discussed in the results section.

Evaluation criteria

Like in most information retrieval tests, the main evaluation criterion of search results is relevance. This should measure the ability of tested web search engines to satisfy the users' information needs described in their search queries.

Relevance assessment is a problematic issue discussed in several studies, and can be influenced by many factors [Schamber, 1994]. A representative assessment can however be done by individual jurors to avoid bias of the researcher in the judgment process.

Search results are evaluated in this test on a binary basis to be "relevant" or "not relevant". Following [Lewandowski, 2008] a search result should satisfy the users' information needs without taking further actions.

Results descriptions are also evaluated on a binary basis as "seems relevant" or "seems not relevant".

To cover the possibility that a search engine may present a result without a description in its results list, the evaluation option "no description available" is also given to jurors.

As this test is specifically designed for Arabic search evaluation and targeting Arab internet users with information needs in their native language, all retrieved documents in languages other than Arabic should be evaluated as "not relevant". Jurors were instructed before they start to judge Non-Arabic results descriptions as "seems not relevant" as well.

Relevance performance measures

Precision is a standard information retrieval evaluation measure used in this test. The other standard evaluation measure, recall, used to evaluate the performance of classical information retrieval systems can not easily be applied to web search engines evaluations as the total number of relevant documents can not be estimated.

As most internet users usually look only at the first one or two results of a query from a search engine, a cut-off value of the first ten results can give reliable evaluation results using the so called top-ten precision.

Precision values will be calculated on both the macro and the micro levels [Womser-Hacker, 1989].

Queries

To be as close as possible to real-life search behavior of Arab web users, a random set of search queries is selected from the most used search queries on the Arabic web search engine (Araby.com).

A collection of fifty queries (a standard for TREC evaluation tests) is a reasonable amount for valid evaluation results. Additional ten queries were reserved for any problems that may occur during the test.

Queries were selected and executed exactly as typed by the original users (as listed in the search engine Araby on 10. March. 2009). No correction or alternative writing methods were suggested.

Information needs

After selecting the random set of search queries, a reconstruction of the information needs behind these queries is necessary to simulate the needs of users originally entered these queries in the search engine and form the relevance judgment criteria. This task is particularly difficult with general short queries.

A group of Arab internet users (mainly students and engineers) were asked to describe their needs of information when searching for given five different queries. All descriptions of each query were then merged to form the relevance judgment criteria. For the search query "Sayings" for example, relevant documents should contain: "Sayings of elders, politicians or celebrities".

Participants

Participants in this test had to be native Arabic speakers and to have average knowledge of internet browsing and usage of web search engines.

A total number of seventy volunteers (53 males and 17 females) filled out the information needs reconstruction forms. To avoid bias in the information needs simulation, users from multiple ages and different education background described information requirements they may associate with the given search queries.

The ideal number of evaluation jurors for fifty queries is equal to fifty, so that each juror can evaluate a single query on all tested search engines. Out of eighty nine invited users (friends and colleagues of the first author), the total number of fifty jurors (42 males and 8 females) from nine Arab countries was achieved.

3.4 Search engines selection

According to the recommendations in [Hawking *et al.*, 2001] for the evaluation of web search engines, the major search engines should be included in the test. As this work tries to explore the suitability of native Arabic web search engines as alternatives to international market leading engines for Arab users, local search engines were tested in addition to the leading international engines.

The most popular five search engines in the Arab countries according to Alexa were selected, two search engines are native Arabic and three are international Arabic-enabled engines. The selected search engines are:

- Araby (www.araby.com)
- Ayna (www.ayna.com)
- Google (www.google.com.sa)
- MSN (www.live.com[1])
- Yahoo! (www.yahoo.com)

3.5 Finding queries

Although there is no published statistics about queries length and complexity for Arabic web searches, the most searched queries on the Arabic search engine Araby showed that Arab users conform to other internet users in using very short and rather unspecific search queries [Jansen *et al.*, 2000]. Search operators (e.g. Boolean operators) were not used in search queries.

3.6 Processing queries

To collect search results from all tested search engines at almost the same time, the queries were processed on a single day one query at a time on all engines with a minimal time interval. This eliminates the possibility of index changes over the tested search engines while processing the single queries which may give one engine an advantage over the others. Results lists were then saved as HTML pages on a local drive.

3.7 Experimental design

The experimental design in this test is based on the repeated-measures design presented in [Tague-Sutcliffe, 1992] and used in [Griesbaum, 2004].

The jurors had to evaluate the top ten search results (from one to ten) for a single query presented by each web search engine without knowing the source of the results to avoid bias caused by users' preferences of a particular search engine they are familiar with.

The second task was to evaluate the search results descriptions of a single different query on the five tested engines. In this part the sources of the results were known to jurors, as they evaluate the descriptions on the locally saved results pages which are identical to the original results pages delivered by the search engines when queries were executed.

3.8 Data collection

The initial design was to collect data from jurors in a laboratory environment on printed evaluation forms. This design faced however difficulties and was replaced with an online survey design as detailed later in the "Pre-Test" section. Data was collected using an online survey service in digital formats which enable different analyses.

3.9 Data analysis

To obtain a binary relevance judgment, not found documents were added to "not relevant" documents in the results evaluation calculations. Results with no descriptions are also considered "seems not relevant".

Using the collected data, the performance of the five tested search engines was evaluated based on the top ten precision. Macro- and micro-precision for the top ten search results were calculated to evaluate the retrieval performance.

Micro-precision values are also calculated for the top ten results descriptions to analyze the conformity of search results and their descriptions by comparing these values and applying measures presented in [Lewandowski, 2008].

3.10 Presenting results

The test motivation, design and methodology were detailed in the previous sections, the test results are analyzed in a dedicated section and the conclusions section gives a summary of the conducted research and future directions for research based on this work.

The complete work is submitted by the first author as a Master thesis at the University of Hildesheim.

4 Performing the test

4.1 Pre-Test

To examine the initial test design, a pre-test was conducted on 03. April 2009 where six participants executed searches for given queries (one by each juror) with the five tested search engines. The search results pages where recorded and the users judged them based on the results descriptions and subsequently, based on the full result documents. The judgments were given on a printed evaluation form.

This design, however, faced the following main problems:

- It was extremely difficult to plan the test timing to suit all users who do not participate as a part of a university course or a job task.
- The pre-test users found the test tasks complicated, confusing and tiring.
- An extra fatigue effect surfaced as a result of the unreliable internet connection on the test location.

All these problems showed that a laboratory test will not be useful or can not be conducted at all at the time and place initially planned. To avoid the disadvantages of the test location, an alternative solution was to involve Arab jurors geographically distributed over multiple countries by performing the test online as shown in the next section.

4.2 Test

All queries were processed on 11. April 2009 in Germany. Results lists and results were saved locally for documentation and prepared on extra web pages for the evaluation process. Jurors only had to visit given links and evaluate the delivered pages digitally on the provided online form.

The responses collection for online surveys was open in the period from 14. April to 12. May 2009. This long period of time can cause variations in the evaluation process due to the highly dynamic nature of the web; it was, however, needed to allow the large number of jurors to find a suitable time slot in their specific location. This effect can be avoided by obligating the participants to work on a certain date, which rises however the questions about users' motivation.

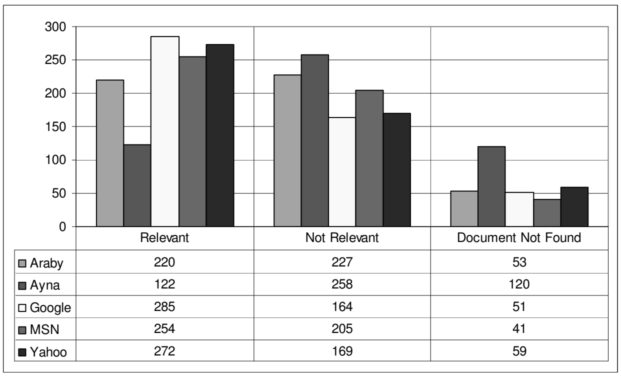
The collected digital data was relatively easy to analyze and process.

[1] Officially replaced on 03.06.09 with a new search service from Microsoft (www.bing.com)

5 Results

5.1 Number of relevant result documents

The first information that can be obtained about the tested search engines from the evaluation data is the number of relevant result documents; this is displayed in (Figure 1).



	Relevant	Not Relevant	Document Not Found
▨ Araby	220	227	53
■ Ayna	122	258	120
□ Google	285	164	51
■ MSN	254	205	41
■ Yahoo	272	169	59

Figure 1: Number of relevant result documents

Google retrieved the largest amount of relevant documents followed by Yahoo then MSN. The native Arabic search engine Araby came in the fourth place with a clear distance to Google. A large gap was found between those four search engines and the other Arabic search engine Ayna, which came in the last place.

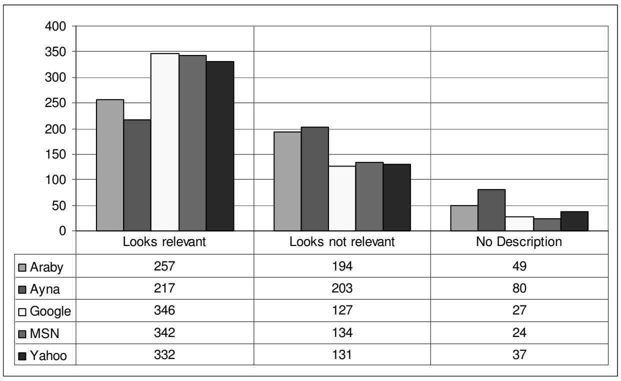
Another finding from the last figure is that both Arabic search engines retrieved a high proportion of absolutely “not relevant” documents (only documents judged as “not relevant” excluding the “not found” documents).

MSN with 8.2% of its results delivered the least search results pointed to lost documents (dead links), followed by Google with 10.2% then Araby which was, with 10.6%, better than Yahoo with 11.8% of dead links in its results lists. Again Ayna came in the last place with more than the double that ratio of all other engines.

These numbers can give an idea on the up to datedness of the search engines indices to a certain extent, but can also be influenced by many factors.

5.2 Number of relevant descriptions

The number of documents among the top ten hits which seemed to be relevant according to their snippet from the five search engines is shown in Figure 2.



	Looks relevant	Looks not relevant	No Description
▨ Araby	257	194	49
■ Ayna	217	203	80
□ Google	346	127	27
■ MSN	342	134	24
■ Yahoo	332	131	37

Figure 2: Number of relevant results description

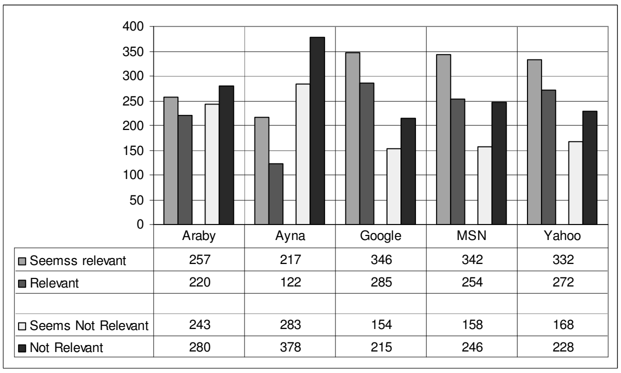
The best judged descriptions were from Google which delivered the highest ratio of relevant descriptions (69.2%). Close after Google was MSN with 68.4% then Yahoo with 66.4% of positively judged descriptions. A clear distance separated these from Araby with 51.4% and Ayna with 43.4%.

The proportion of 40.6% of irrelevant descriptions in the results list of Ayna means that a user may ignore up to this amount of top ten search results because of their descriptions. 38.8% of the descriptions delivered by Araby also gave a negative idea about the results described. With 26.8% for MSN, 26.2% for Yahoo and 25.4% for Google the international search engines gave a lower chance for bypassing results from the first look at their descriptions.

MSN tried to describe the most delivered results out of which 4.8% did not have a description. Google failed similarly to deliver descriptions to 5.4% of presented results and a higher proportion was by Yahoo at 7.4%. Even with 9.8% of results without descriptions, Araby was better than Ayna which delivered 16% of its top ten search results without any description.

5.3 Descriptions-Results conformity

To evaluate how good a search engine can form results descriptions, a comparison between results and results descriptions is needed.



	Araby	Ayna	Google	MSN	Yahoo
▨ Seemss relevant	257	217	346	342	332
■ Relevant	220	122	285	254	272
□ Seems Not Relevant	243	283	154	158	168
■ Not Relevant	280	378	215	246	228

Figure 3: Descriptions-result s conformity

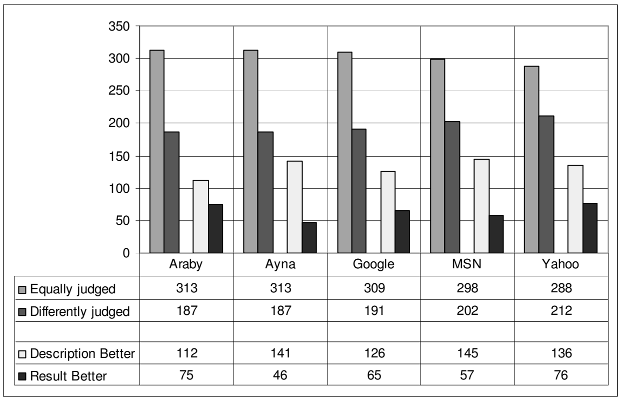
The comparison displayed in (Figure 3) sums not found results to “not relevant” ones and not available descriptions to descriptions “seem not relevant”. It shows that all tested search engines presented more relevant descriptions than real relevant results.

Although the number of relevant descriptions delivered by Araby was lower than the relevant descriptions delivered by the three international search engines, Araby had the lowest difference of 7.8% between the counts of relevant descriptions and relevant results.

For Google, there were 9.7% more relevant snippets than relevant documents, 9.9% for Yahoo and 14.8% for MSN. Ayna exhibited the largest difference of 28% between the numbers of relevant descriptions in comparison with relevant results.

As the relevance judgment of results and results descriptions for each query was done by two different jurors, these results can be influenced as discussed in [Griesbaum, 2004] by formal and contextual variations in the descriptions presentation and by preference factors.

A high number of relevant descriptions does not mean necessarily that they correctly describe the real results and that users could depend on these descriptions to visit relevant results and avoid irrelevant ones. Figure 4 shows the number of documents for which the relevance judgment based on snippet and full document was equal or different. A high judgment consistence of description-result pairs means a well forming of search results descriptions for both “relevant” and “not relevant” results.



	Araby	Ayna	Google	MSN	Yahoo
▨ Equally judged	313	313	309	298	288
■ Differently judged	187	187	191	202	212
□ Description Better	112	141	126	145	136
■ Result Better	75	46	65	57	76

Figure 4: Description-result pairs judgment comparison

The two native Arabic search engines were equally the best in this regard by presenting the highest ratio of consistent description-result pairs. 62.6% of top ten descriptions delivered by each of Araby and Ayna were identically judged as their respective results. The second conformity level was achieved by Google with 61.8% of presented description-result pairs, followed by MSN with 59.6% and lastly Yahoo with 57.6%.

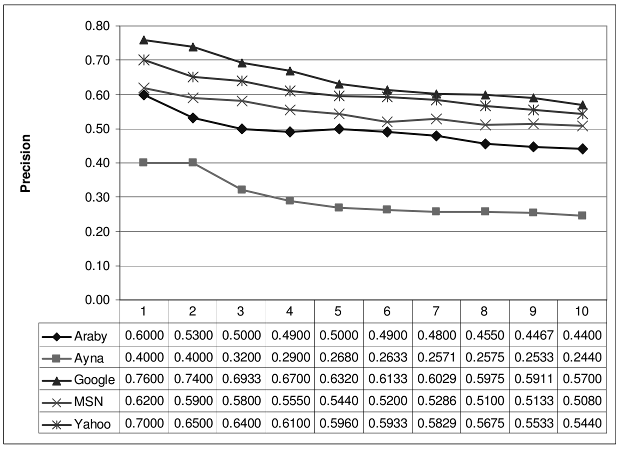
Another output of the last figure is how frequent do the tested search engines tend to present irrelevant results with descriptions that reveal to the user that they can be relevant.

75.4% of Ayna’s not matching descriptions gave a better image of the results than they really were, followed by MSN with a close frequency of 71.8% then came Google with 65.9%, closer to Yahoo with 64.2%.

Araby was the search engine that provided the least descriptions which guided the users to results not actually of the same relevance with 59.9% of the total inconsistent descriptions.

5.4 Results mean average precision (micro-precision)

The recall/precision graph plotted in (Figure 5) shows the precision average values for each search engine at the respective rank for the top ten results.



	1	2	3	4	5	6	7	8	9	10
◆ Araby	0.6000	0.5300	0.5000	0.4900	0.5000	0.4900	0.4800	0.4550	0.4467	0.4400
■ Ayna	0.4000	0.4000	0.3200	0.2900	0.2680	0.2633	0.2571	0.2575	0.2533	0.2440
▲ Google	0.7600	0.7400	0.6933	0.6700	0.6320	0.6133	0.6029	0.5975	0.5911	0.5700
× MSN	0.6200	0.5900	0.5800	0.5550	0.5440	0.5200	0.5286	0.5100	0.5133	0.5080
✳ Yahoo	0.7000	0.6500	0.6400	0.6100	0.5960	0.5933	0.5829	0.5675	0.5533	0.5440

Figure 5: Precision graph for the top ten hits

This describes how relevant results were distributed in the top ten results lists.

Google achieved the best values at the top three results ranks, which are usually the most seen by users [Jansen *et al.*, 2000]. 76% of the results delivered by Google at the first ranking place for all queries were judged as relevant, where the average of the sum of relevant results for the top ten results was 57% (rank 10).

The next search engine on the top three ranks was Yahoo with 70% for the first rank and a precision closer to Google at the last rank with 54.4%. Then came MSN which reached 62% at the first rank and 50.8% at the last.

The Arabic search engine Araby reached precision values not far from MSN especially at the first rank with 60% of relevant results at the first place in the results list for the fifty queries. However, the later ranks showed higher differences especially compared to Google. The overall precision at the tenth rank for Araby was at 44%.

Ayna delivered results at the top of results list with lower precision than the results at the last rank of all other engines. The average precision for the first rank results of fifty queries was 40% only. Ayna exhibited a drop in precision after the second position. The precision at the tenth position was merely 24.4%.

The mean average precision values of the top one to ten results for the five tested search engines are shown in (Table 1).

Search Engine	Mean Average Precision
Araby	0.49
Ayna	0.30
Google	0.65
MSN	0.55
Yahoo	0.60

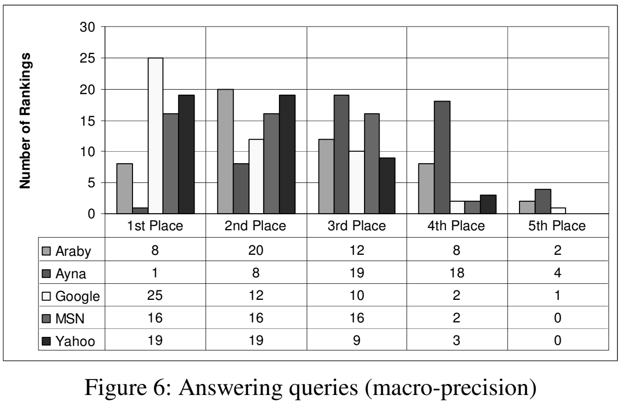
Table 1: Mean average precision

Although Araby performed much better than Ayna, both search engines could not reach an acceptable precision value of 50%, where all other engines stayed above this value even at the last ranking places.

5.5 Answering queries (macro-precision)

To explore which search engine dealt best with every query of the fifty used in the test, the macro-precision is observed. The precision values from all tested search engines for every single query are compared and the engines are ranked accordingly. Search engines with equal precision values for the same query are ranked equally to avoid preferences. The rankings sum comparison should give an overall macro-precision performance view.

These ranking frequencies are displayed in (Figure 6). The numbers show how many times each search engine occupied which ranking place in the comparison.



	1st Place	2nd Place	3rd Place	4th Place	5th Place
▨ Araby	8	20	12	8	2
■ Ayna	1	8	19	18	4
□ Google	25	12	10	2	1
■ MSN	16	16	16	2	0
■ Yahoo	19	19	9	3	0

Figure 6: Answering queries (macro-precision)

The interpretation of the results reveals the over performance of Google with 25 times ranked at the first place. This means that Google achieved the best top ten precision for 50% of all processed queries. Moreover, it maintained its position in the first two ranks for 74% of these queries and had the worst precision for one single query.

Yahoo was equally ranked at the first and second places for 38% of processed queries at each rank, but was never at the last place. MSN came then with distributed precision performance equally at the first three rankings with 16 times at each rank.

Araby was the first search engine for only 8 times and performed at the second and third levels for 64% of the processed queries. It was the worst engine with two queries. Ayna, on the other hand, was mostly at the third and fourth places and could reach the first place for only one query.

5.6 Number of answered queries

To compare the degree to which each search engine could be helpful for the user, the number of answered queries (queries with at least one retrieved result judged as relevant) is calculated.

(Table 2) shows that Google and Yahoo could answer all processed queries by retrieving at least one relevant result in the top ten results list. The top ten results from MSN for the two queries "Visual illusion" and "Arabic language" included no relevant results.

Search Engine	Answered queries	Not Answered queries
Araby	46	4
Ayna	39	11
Google	50	0
MSN	48	2
Yahoo	50	0

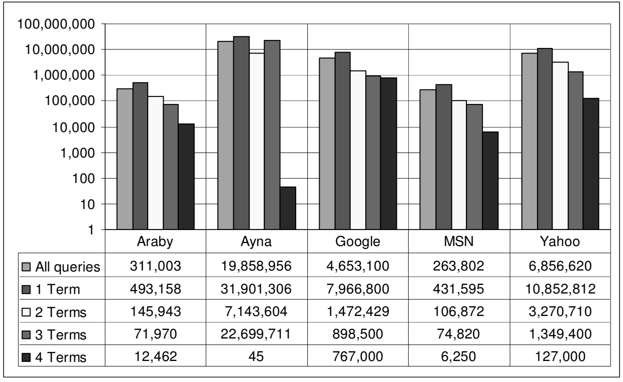
Table 2: Number of answered queries

Araby could not answer the four queries "Olympiad", "Obama", "Visual Illusion" and "Sayings". Although these queries can give an impression that the search engine was of no use for the users who entered these unanswered queries (considering that they only see the top ten results), the findings could be influenced by the subjective judgment and their acceptance can be limited. The results of Ayna seemed, however, clearly disappointing as it performed the worst with 22% of not answered queries.

5.7 Number of retrieved documents

Although a detailed estimation of indices sizes for the tested search engines is not within the scope of this work, a general idea about these indices can be obtained from analyzing the amount of retrieved documents reported by the engines when processing the queries. The average counts of results delivered by each search engine for all queries and classified by query terms count are displayed in (Figure 7).

Ayna reported the largest amount of results for each search query even when it performed the worst as seen in all previous evaluation measures. For the search query "Newspapers" for example, Ayna delivered over 33 Millions of results with a top ten precision value of 0 (i.e. the query was not answered). This may question the indexing method and the retrieval algorithm of this search engine, as presenting over millions of irrelevant results can be a sign of an essential index problem or an improper search algorithm. Yahoo and Google delivered a large amount of results in comparison to Araby and MSN.



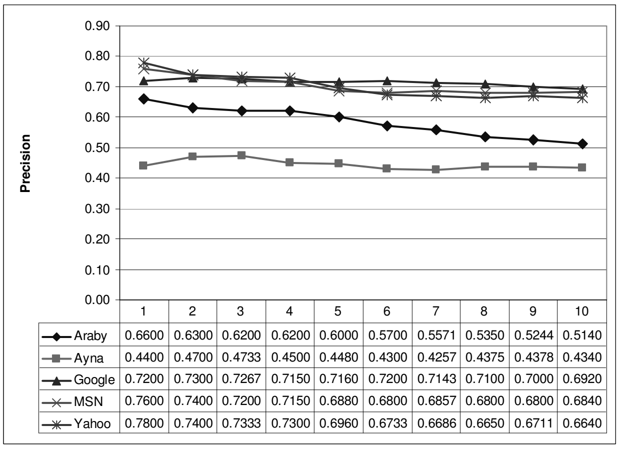
	Araby	Ayna	Google	MSN	Yahoo
▨ All queries	311,003	19,858,956	4,653,100	263,802	6,856,620
■ 1 Term	493,158	31,901,306	7,966,800	431,595	10,852,812
□ 2 Terms	145,943	7,143,604	1,472,429	106,872	3,270,710
■ 3 Terms	71,970	22,699,711	898,500	74,820	1,349,400
■ 4 Terms	12,462	45	767,000	6,250	127,000

Figure 7: Mean number of retrived results

The decrease in average results count for multiple terms queries is clear on all search engines except for Ayna which showed no consistent behavior.

5.8 Descriptions mean average precision

Search results descriptions are evaluated in this work for their importance for users in the decision making to visit a retrieved result. The recall/precision graph of the top ten results descriptions at each result ranking is shown in (Figure 8).



	1	2	3	4	5	6	7	8	9	10
Araby	0.6600	0.6300	0.6200	0.6200	0.6000	0.5700	0.5571	0.5350	0.5244	0.5140
Ayna	0.4400	0.4700	0.4733	0.4500	0.4480	0.4300	0.4257	0.4375	0.4378	0.4340
Google	0.7200	0.7300	0.7267	0.7150	0.7160	0.7200	0.7143	0.7100	0.7000	0.6920
MSN	0.7600	0.7400	0.7200	0.7150	0.6880	0.6800	0.6857	0.6800	0.6800	0.6840
Yahoo	0.7800	0.7400	0.7333	0.7300	0.6960	0.6733	0.6686	0.6650	0.6711	0.6640

Figure 8: Precision graph for the top ten descriptions

All the tested search engines reached better precision values for descriptions than for results themselves (Figure 5) at all cut-off values except for Google at the first two rankings where the average precision for the search results was higher than the average for their descriptions.

To compare the overall precision performance of the results descriptions, the mean average precision for the top one to ten results descriptions from the five search engines is displayed in (Table 3).

Search Engine	Mean Average Precision
Araby	0.49
Ayna	0.30
Google	0.65
MSN	0.55
Yahoo	0.60

Table 3: Mean average precision for the top one to ten results descriptions

5.9 Descriptions-Results comparison

To explore the search engines performance differences in retrieving search results and presenting these results, the evaluations of results and descriptions are compared using measures introduced by Lewandowski [2008].

Mean distance deviance

The distance deviance DRdist$_n$ shows how precision of search results vary from the precision of results description, where n is the number of results or descriptions observed. The following table shows the mean values:

Search engine	DRdist$_{10}$
Araby	0.09
Ayna	0.15
Google	0.08
MSN	0.16
Yahoo	0.10

Table 4: Mean distance deviance of top ten descriptions and results

MSN and Ayna showed the largest difference between descriptions and results precision then Yahoo and Araby. Google had the lowest average precision difference.

We compare the individual description-result pairs on the basis of absolute evaluation values for results and descriptions displayed in (Table 5).

Description	Result	Araby	Ayna	Google	MSN	Yahoo
Relevant	Relevant	146	77	221	198	197
(a)						
Relevant	Not relevant	113	124	127	146	137
(b)						
Not relevant	Relevant	76	123	66	58	77
(c)						
Not relevant	Not relevant	165	158	86	98	89
(d)						
Total number of documents (e)		500	500	500	500	500

Table 5: Individual evaluation counts for description-result pairs

Dividing the pair counts (a, b, c, d) by the total number of documents (e), the precision-result comparison measures can be calculated as shown in (Table 6).

Comparison measure	Araby	Ayna	Google	MSN	Yahoo
Description-result precision (a/e)	0.29	0.15	0.44	0.40	0.39
Description-result conformance (a+d)/e	0.62	0.47	0.61	0.60	0.57
Description fallout (c/e)	0.15	0.25	0.13	0.12	0.15
Description deception (b/e)	0.33	0.32	0.17	0.20	0.18

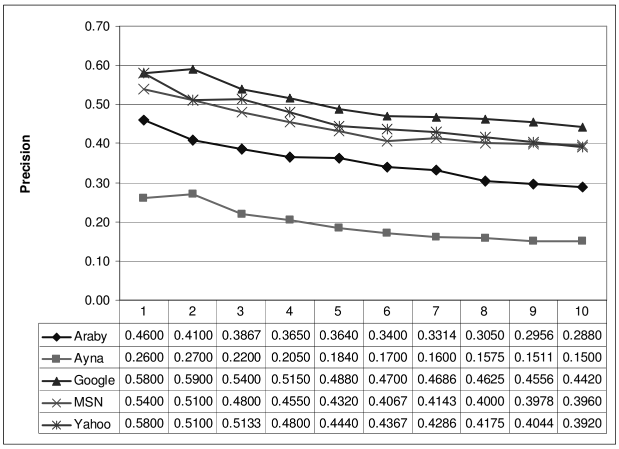
Table 6: Description-result comparison measures

The best case is when the search engine delivers relevant documents with descriptions that make them appear relevant to the user.

Description-result precision

Google had the highest description-result precision (super precision) followed by MSN then Yahoo. Araby followed with a clear gap, where Ayna was far behind all other engines.

The recall/precision graph for relevant results described with relevant descriptions is plotted in (Figure 9).



	1	2	3	4	5	6	7	8	9	10
Araby	0.4600	0.4100	0.3867	0.3650	0.3640	0.3400	0.3314	0.3050	0.2956	0.2880
Ayna	0.2600	0.2700	0.2200	0.2050	0.1840	0.1700	0.1600	0.1575	0.1511	0.1500
Google	0.5800	0.5900	0.5400	0.5150	0.4880	0.4700	0.4686	0.4625	0.4556	0.4420
MSN	0.5400	0.5100	0.4800	0.4550	0.4320	0.4067	0.4143	0.4000	0.3978	0.3960
Yahoo	0.5800	0.5100	0.5133	0.4800	0.4440	0.4367	0.4286	0.4175	0.4044	0.3920

Figure 9: Precision graph for the top ten relevant results and descriptions

Yahoo performed the best for example at the first rank where Google leads with clear difference at the second rank. MSN kept close to the other two engines on all ranks. Araby precision was clearly lower than the three international search engines, and Ayna, which performed at the second rank better than the first rank, was far behind all other engines.

Description-result conformance

From (Table 6) one can see that Araby performed the best by giving the highest amount of "right" described results followed by Google and MSN with small differences, and then came Yahoo followed by Ayna at the last place describing less than a half of delivered results correctly.

Description fallout

Considering that a user may not visit a search result if its description seemed to be irrelevant, the description fallout measures the chance of missing relevant results because of their descriptions.

Most of the tested search engines performed very closely in this regard except for Ayna which described 25% of its relevant search results with seemingly irrelevant descriptions.

Description deception

A high value of description deception can show that the search engine does not provide proper descriptions for irrelevant retrieved results and may cause a frustrating impression to the user who feels misled by the search engine.

Google performed the best in this regard by providing the least amount (17%) of irrelevant results associated with descriptions that let them look relevant. The next search engines were Yahoo then MSN with close values (18% and 20% respectively). Clearly, both Arabic search engines performed worse with 32% of results with misleading descriptions for Ayna and 33% for Araby.

6 Conclusions

As overall result, it can be concluded that Google was the best search engine in all measures except for the number of not found documents, number of results with no descriptions, descriptions fallout and the conformance of results and descriptions.

MSN and Yahoo exchanged the second and third ranking places with regards to most evaluation measures except for the number answered queries where they performed equally and for the number of not found document and the conformance of results and descriptions where Yahoo fell back to the fourth rank.

Moreover, MSN performed best in terms of not found documents count, count of results with no descriptions and description fallout.

Although Araby was mostly on the penultimate place, it showed now significant precision difference to MSN, and delivered the best conformance of results and their descriptions, performed better than Yahoo in terms of not found documents count and was equal to it in description fallout. The only last place given to Araby was in description deception.

The underperformance of Ayna was a remarkable trouble sign. The search engine with the large promotion campaign seemed to suffer from very serious problems in both its indexing and searching algorithms and it obviously would need substantial improvement.

This test found that there is mostly no significant reason to prefer Google to Yahoo in terms of search performance in Arabic language. One should however keep in mind that Yahoo does not offer an Arabic interface (by the time of the test) which can affect its acceptance. Arab users may also still consider MSN as a potential alternative search engine especially when interested in particular performance aspects.

The more important finding of the test is that both tested native Arabic search engines could not proof their ability to compete as a local alternative to international search services. Even when Araby had some good results, a wide space for improvement still exists.

The results of this state of the art work can be considered for further evaluations and research of Arabic search engines, particularly with the absence of similar published studies for this language.

References

[Abdelali *et al.*, 2004] Ahmed Abdelali, Jim Cowie, and Hamdy S. Soliman. Arabic Information Retrieval Perspectives. In: *Proceedings of JEP-TALN 2004 Arabic Language Processing*, Pages 19-22. April. 2004.

[Gordon and Pathak, 1999] Michael Gordon and Praveen Pathak. Finding Information on the World Wide Web: the Retrieval Effectiveness of Search Engines. *Information Processing & Management*, 35:141-180.

[Griesbaum 2004] Joachim Griesbaum. Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. In: *Information Research*, 9(4). <http://informationr.net/ir/9-4/paper189.html> (Accessed 10.08.09)

[Hammo 2009] Bassam H. Hammo: Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. In: *Information Retrieval*, 12(3):300-323, June 2009, Springer, Netherlands.

[Hawking *et al.*, 2001] David Hawking, Nick Craswell, Peter Bailey and Kathleen Griffiths. Measuring Search Engine Quality. In: *Information Retrieval*, 4(1):33-95 Springer, Netherlands.

[Jansen *et al.*, 2000] Bernard J. Jansen, Amanda Spink and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the Web. In: *Information Processing & Management*, 36(2):207-227.

[Lazarinis, 2007] Fotis Lazarinis. Web retrieval systems and the Greek language: do they have an understanding? In: *Journal of Information Science*, 33(5):622-636, Sage Publications Inc.

[Lazarinis, 2009] Fotis Lazarinis, Jesus Vilares, John Tait and Efthimis Efthimiadis. Current research issues and trends in non-English Web searching. In *Information Retrieval*, 12(3):230-250, June. 2009.

[Lewandowski, 2008] Dirk Lewandowski. The Retrieval Effectiveness of Web Search Engines: Considering Results Descriptions. In: *Journal of Documentation*, 64(6):915-937.

[Lewandowski and Höchstötter, 2008] Dirk Lewandowski, and Nadine Höchstötter. Web Searching: A Quality Measurement Perspective. In: A. Spink and M. Zimmer (Editors.): *Web Searching: Multidisciplinary Perspectives*. Pages: 309-340, Springer, Berlin.

[Mandl, 2008] Thomas Mandl. Recent Developments in the Evaluation of Information Retrieval Systems: Moving Toward Diversity and Practical Applications. In *Informatica - An International Journal of Computing and Informatics*, 32:27-38.

[Moukdad, 2002] Haidar Moukdad: Language-based retrieval of Web documents: An analysis of the Arabic-recognition capabilities of two major search engines. In: *Proceedings of the 65th Annual Meeting of the American Society for Information Science and Technology*, 18-21. November. 2002, Philadelphia, PA. Medford: Information Today, Poster p. 551.

[Moukdad, 2004] Haidar Moukdad. Lost in Cyberspace: How do search engines handle Arabic queries? In: Access to Information: Technologies, Skills, and Socio-Political Context. Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science, Winnipeg, 3-5. June. 2004.

[Moukdad and Large, 2001] Haidar Moukdad, and Andrew Large. Information retrieval from full-text Arabic databases: Can search engines designed for English do the job? Libri 51(2):63-74.

[Schamber, 1994] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*. 29:3-48.

[Tague-Sutclife, 1992] Jean Tague-Sutclife. The Pragamatics of Information Retrieval Experimentation, Revisited. In: *Information Processing & Management*, 28(4): 467-490, Elsevier.

[Womser-Hacker, 1989] Christa Womser-Hacker. Der PADOK Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information-Retrieval-Systemen. Hildesheim, Georg Olms.