Scientific Research and Essays Vol. 7(28), pp. 2472-2483, 26 July, 2012
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE11.1708
ISSN 1992-2248 ©2012 Academic Journals

Full Length Research Paper

Estimating the size of Arabic indexed web content

Abdulrahman Alarifi*, Mansour Alghamdi, Mohammad Zarour, Batoul Aloqail, Heelah Alraqibah, Kholood Alsadhan and Lamia Alkwai

Computer Research Institute, King Abdulaziz City for Science and Technology, P. O. Box 6086, Riyadh 11442, Riyadh, Saudi Arabia.

Accepted 7 June, 2012

Various initiatives designed to increase Arabic Web content have been undertaken in recent years, and now search engines are reporting that the Arabic portion of Web content has grown relative to the overall Web content. An accurate estimate of Arabic Web content is crucial for those interested in studying it and enriching it. In this paper, we propose a statistics-based system to estimate the size of Arabic indexed Web content using three popular search engines; Google, Yahoo and Bing. Our system relies on selecting sample words from an Arabic corpus to estimate the size of the Arabic Web content indexed by the search engines and the overlap among them. We have used Arabic Wikipedia as a corpus, as it provides diversified content accessed by a large number of Internet users. Our results show that, as of December 2010, the size of the Arabic indexed Web content was estimated at 2 to 2.1 billion pages.

Key words: World Wide Web, the Web, search engine, index size, Arabic content, Internet, corpus.

INTRODUCTION

Every second, thousands, if not millions, of Web pages are added or removed from the World Wide Web. Bergman estimated that 7.5 million pages per day were added to the Web in 2001 (Bergman, 2001). Google claims to have indexed 60 billion unique uniform resource locators (URLs) in 2005, and 1 trillion in 2008 (Alpert and Hajaj, 2008), and Yahoo claims that its search engine indexed 19.2 billion pages in 2005 (Mayer, 2005). These figures show how much the Web can grow in a very short time. However, we cannot verify the accuracy or reliability of the information involved, since none of the stated companies reveals its highly valued proprietary list of indexed URLs. In fact, what information search engines release to the public and to competitors may reflect their business strategy, which is to gain more users and more Web pages.

The content of the Web can be classified into surface Web content and deep Web content. The deep Web is the part that is not visible to search engines, which is why it is sometimes called the “hidden” Web. The deep Web usually refers to the following (Singh, 2002; Gong et al., 2006; Madhavan et al., 2008):

1) Database content that is stored in tables created by various database management systems, such as Access, Oracle and SQL.
2) Non text files, such as multimedia, images, software, and documents in formats like portable document format (PDF).
3) Unlinked pages, which can only be accessed by typing the URL into a Web browser.
4) Private webs, which consist of content available on sites protected by passwords or other restrictions, such as fees.
5) Dynamically generated content, which are pages accessible through links generated by java script.
6) New Web content, such as blogs, social network pages, etc.

Lately, more and more of what was once considered to be deep Web content is being indexed by search engines and becoming accessible through search queries, such

*Corresponding author. E-mail: aarifi@kacst.edu.sa.

as multimedia files and social network pages.

In contrast, the surface Web can be indexed by search engines. It is sometimes called the “visible” Web, since it is readily accessible to Internet users via search engines. Bergman (2001) studied Web content, and concluded that the deep Web is 400-550 times larger than the surface Web.

In order for search engines to traverse and index the World Wide Web, they use Web crawlers, which are the computer programs at the heart of search engines, as they continuously search the World Wide Web and “crawl” around in it, looking for new pages to add or delete from their indexes. They read and parse Web pages, and then move to hyper links on these pages, looking for more pages to process. The performance of Web crawlers is based on their selection policy, revisit policy, politeness policy, and parallelization policy (Gupta and Johari, 2009; Peisu et al., 2008).

We do not propose to study the size of the deep Web in this paper, but rather to estimate the size of Arabic content on the surface Web. The surface Web is more interesting and informative, since its content be easily searched and accessed. The exact size of the Web cannot be measured, because it is so huge, especially the deep Web portion, and its content is dynamic in nature. The number of indexed pages cannot be precisely determined either, since the indexed data are privately owned by the search engines and details about them are not made public. The fact is, it is only possible to estimate the size of the indexed Web.

We have developed three measurements to estimate the number of Arabic pages indexed by three major search engines; Google, Yahoo and Bing (Hitwise, 2010). We also use these measurements to estimate the overlap among these three search engines. From these figures, we estimate the number of indexed Arabic pages on the Web. In order to do this, we use an Arabic corpus and take into account a range of Arabic accents, as well as the diversity of its elements and contributors. We present our analyses and results, and compare them with different measurements.

Motivation

Recently, national initiatives have been introduced to enrich Arabic Web content, such as King Abdullah’s Initiative for Arabic content (KACST, 2010) and AlWaraq, which is a Web encyclopedia containing numerous books and references (Suwaidi, 2010). One of the main goals in most of these initiatives is to increase the size of Arabic Web content to reflect the importance of Arabic language and the large population of the Arabic world.

The research presented in this paper is motivated by the following:

1) To provide an accurate tool to measure the size of Arabic Web content,
2) To study the growth trends of Arabic Web content.
3) To provide a better understanding of the size of Arabic pages indexed by each of these three search engines: Google, Yahoo and Bing, to help evaluate their relative contribution to enriching Arabic Web content.

In this work, we briefly present the related work that describes the various methods used to estimate the size of the indexed Web. Then, we describe the methodology used to estimate the size of the indexed Arabic pages on the Web and the corpus we have used, and explain the sampling methods that we have adopted in the process and give more information about the Zipf Law distribution method. Then, we present the implementation of our size estimation method, followed by our results and performance evaluations. Finally, we end the paper with our conclusion and directions for future work.

RELATED WORK

Estimating the size of the Web has been a challenge for researchers since 1998. Up to now, several attempts have been made, and a number of results have been reported. Among the many reasons for doing this (Gulli and Signorini, 2005) are the following:

1) It gives a good indication of the current number of Web pages, which is useful information for many researchers.
2) It is essential for the operation of data crawlers and data extractors, since search engines must know when to stop crawling and extracting.
3) It is an important evaluation metric for search engines, showing their coverage of the Web.
4) It is essential information for performing Web compression, ranking, spidering, indexing, and mining.
5) It helps planners and developers who are working on language specific Web content and its market.

There are two ways to achieve this: estimate relative size, and estimate actual size. Relative size estimation is based on search engine performance. The percentage of coverage of each search engine is calculated using various methods and probability models, and, from these percentages, the size of the Web can be inferred. Researchers have also looked at ways to optimize the current approaches to estimation, by reducing the bias caused by search engine sampling, for example. Actual size estimation involves counting the number of Web pages that require searching the whole content of the Web. An efficient and effective Web crawler can provide a better estimation, but the use of such a tool is outside the scope of this research.

Arabic Web content has never been accurately measured. Although, statements from the Economic and

Social Commission for Western Asia (ESCWA) claim that it constitutes only 2% of the size of the Web (ESCWA, 2010), this percentage has never been tested. But, in any case, the amount of Arabic content on the Web is known to be very small. This can be easily seen by searching for a specific word in Arabic. A small number of results will appear, compared to those for the same query in other languages. However, this has been changing over the past few years, because of several initiatives intended to enrich Arabic Web content and the growth of Internet usage in the Arab world. Since then, the size of Arabic Web content has grown significantly.

In a recently published paper, three different corpora were used to estimate Arabic Web content (Tawileh and Alghamdi, 2011), including public Arabic websites, Wikipedia, and contemporary Arabic websites. Google and Yahoo search engines have been used to calculate the number of Web pages in Arabic. Zipf's Law was applied, and the results show an estimated 413 million or more indexed Arabic pages as of April 9, 2011.

Bharat and Broder (1998) introduced a standardized statistical technique for measuring search engine coverage and overlap through random queries, which can be implemented by third-party evaluators using only the public query interface of the search engines. Two procedures are required to implement their idea: sampling, and checking. Sampling selects pages uniformly and at random from the index of a search engine, by firing a random query and selecting a random URL from the first 100 results. Checking determines whether or not a page is indexed by a search engine. To do this, the page is retrieved and analyzed to compute a conjunctive query composed of a small number of the most significant terms on the page. Both sampling and checking use a lexicon of Web words and their frequency of occurrence in documents on the Web, which is built by crawling nearly 300,000 documents. Their experimental results show the estimated size and overlap for HotBot, AltaVista, Excite, and Infoseek -- the largest search engines in use at the time -- as a percentage of their total joint coverage in 1997.

Lawrence and Giles (1998) studied the coverage of six major search engines; AltaVista, Excite, HotBot, Infoseek, Lycos, and Northern Light. They compared the number of documents returned by different search engines and analyzed the queries of employees of the NEC Research Institute. They analyzed the overlap among the search engines to estimate the size of the Web. Their conclusion is that the indexable Web contains 320 million pages.

Aires and Santos (2002) presented their work, measuring the number of pages and Web users in the Portuguese language. To achieve their goal, they used a Portuguese corpus to generate queries for three search engines; AlltheWeb, AltaVista, and Google.

Gulli and Signorini (2005) estimated the size of the public indexable Web and the overlap of several search engines, such as Google, MSN, Yahoo, and Ask, using the approach suggested by Bharat and Border (1998). The DMOZ directory, with more than 4 million pages, was used to collect 2,190,702 terms and their frequency of occurrence in more than 75 languages. All the terms were sorted by frequency of occurrence and divided into blocks of 20 terms. A query term was selected from each block and executed by each search engine. During the checking procedure, search engine interfaces were used to check directly whether or not a URL is indexed. Gulli and Signorini (2005) showed that Google indexes around 68.2% of Web pages indexed by other search engines. Yahoo, MSN, and Ask index 59.1, 49.2 and 43.5%, respectively, and estimated that the size of the indexable Web was 11.5 billion pages.

METHODOLOGY

A corpus is a large collection of structured text, either in a single language (monolingual corpus) or in a number of languages (multilingual corpus), which is used for statistical analysis and fundamental linguistic research (Marcus et al., 1993). In this research, we have used the Arabic Wikipedia as a corpus and calculated the word and document occurrence frequencies of every word in the corpus.

A great deal of effort has been expended by various universities, institutes, and researchers around the world in the creation of large and reliable Arabic corpora. Table 1 shows some of these corpora available in the literature (Al-Sulaiti et al., 2006).

The specifications of the Arabic corpus needed for our research to estimate the size of the Arabic indexed Web content are derived from our research objectives: the corpus is to be extracted from Web content written in Arabic, excluding content written in Arabic script, but in a different language, such as Farsi; the corpus content is to cover different regions of the Arab world; and standard modern Arabic words are to be included. To ensure adequate diversity of words in the Arabic corpus, it is to include words from different fields and disciplines, such as the social sciences, medicine, computer science, engineering, etc.

In this research, we use Arabic Wikipedia as the source for our corpus, since it provides a diversified source of Arabic text and content. This version of Arabic Wikipedia was launched in April, 2002. In February, 2010, it contained more than 120,000 articles, 599,000 pages, 250,000 registered users, and 6,400 media files. As such, it ranked 24th among Wikipedia versions, in terms of the number of articles it contains (Wikimedia, 2010b).

We chose Arabic Wikipedia as the source to build our corpus, for several reasons:

1) It contains the largest number of Arabic articles, and they cover many different fields, which provides a very rich source for building a corpus.
2) It is openly editable, so that individuals from different professions, with different levels of education, and speaking specific dialects, etc., may add to or edit Arabic Wikipedia articles. This feature gives it a broad diversity of words and terms.
3) Its articles are more carefully written and cleaner than the average article found on the Internet.

Table 2 shows our corpus specifications based on the Arabic Wikipedia. It shows the large number of words in our corpus, which

Table 1. Some of the available Arabic corpora.

Corpus	Source	Size
Quranic Arabic	The Holy Quran	77,430 words
QAMUS	Public newspapers on the web	2.5-3 billion words
CLARA (Corpus Linguae Arabicae)	Scanned Arabic books	50 million words
Arabic Newswire Corpus	Articles from the Agence France Presse (AFP) Arabic Newswire	80 million words
A corpus of contemporary Arabic (CCA)	Websites and online magazines	1 million words
Al-Hayat Corpus	Al-Hayat newspaper	18.6 million words
An-Nahar Corpus	An-Nahar Lebanon newspaper	140 million words
Arabic Gigaword Corpus	Arabic news sources (Agence France Presse, Al-Hayat news agency, An-Nahar news agency, Xinhua news agency)	400 million words
DIINAR	Variety of sources	10 million words

Table 2. Specifications of our Arabic Wikipedia-based corpus.

Attribute	Value
Number of articles	135,661 articles
Average size of articles	4.8 KB
Number of words	24,782,158 words
Unique words	251,138 words
Average words/article	183 W/A
Available fields	10

contains 24,782,158 words. Also, it shows that these words are extracted from 135,661 articles in 10 different fields. These numbers reflect the size of the corpus we have used, in comparison to the smaller Arabic corpora available.

Wikimedia provides public “dumps” of Wikipedia's content, which can be used for different purposes, in particular academic research. We used the Arabic Wikipedia dump file, which is available in SQL and XML file format (Wikimedia, 2010a). This dump file was processed to extract words, and then calculate the word and document frequencies for each word. In Figure 1a, we show the frequency of occurrence of the top 20 words in our corpus, while in Figure 1b, we show the top 20 words in terms of document occurrence frequency.

The methodology followed in this research is based on that introduced by Kunder (2010), in which he estimates the size of the indexed Web and the size of the indexed Dutch Web using the four search engines: Google, Yahoo, MSN, and Ask. In this methodology, the estimated size of the Web is defined by the following formula:

$$E_p = \frac{DF_s}{DF_c} \times N_c \qquad (1)$$

Where E_p is the estimated number of Web pages, DF_s is the document frequency per search engine, DF_c is the document frequency in the corpus, and N_c is the total number of documents in the corpus.

The idea underlying this formula is that the total number of Web pages can be estimated by knowing the frequency of occurrence of a word in a corpus (document frequency) and the number of returned results for that word when it is queried with a search engine. To be able to apply this formula, we need to build and analyze an Arabic corpus. For each word in the corpus, two measurements are required: word frequency and document frequency. Word frequency is the total number of occurrences of a word in the corpus, while document frequency is the total number of articles in the corpus containing that word.

Next, a collection of words (a sample) is chosen from the corpus. For better results, three sampling methods were used: uniform random word selection, restricted uniform random word selection, and Zipf's Law word selection. In uniform random word selection, a number of words *N* are selected from the corpus where selection is uniformly distributed. Restricted uniform random word selection limits selection to the words with word frequency higher than or equal to 10. In this way, rarely occurring words are not considered in size estimation. The third sampling method is based on the Zipf's Law distribution, where words with higher word frequency are more likely to be selected. Zipf's Law, which is sometimes called the Pareto Law, assumes that the frequency of the event (or word) that occurs most often is nearly twice the frequency of the second most often occurring event (or word). This relation also holds for the frequency of events (or words) occurring after that (Wikipedia, 2012). There are many natural phenomena that follow Zipf's distribution, among them city populations, population incomes, and word frequencies. The Zipf's Law distribution is logarithmic. We have implemented Zipf's Law in word selection using the following

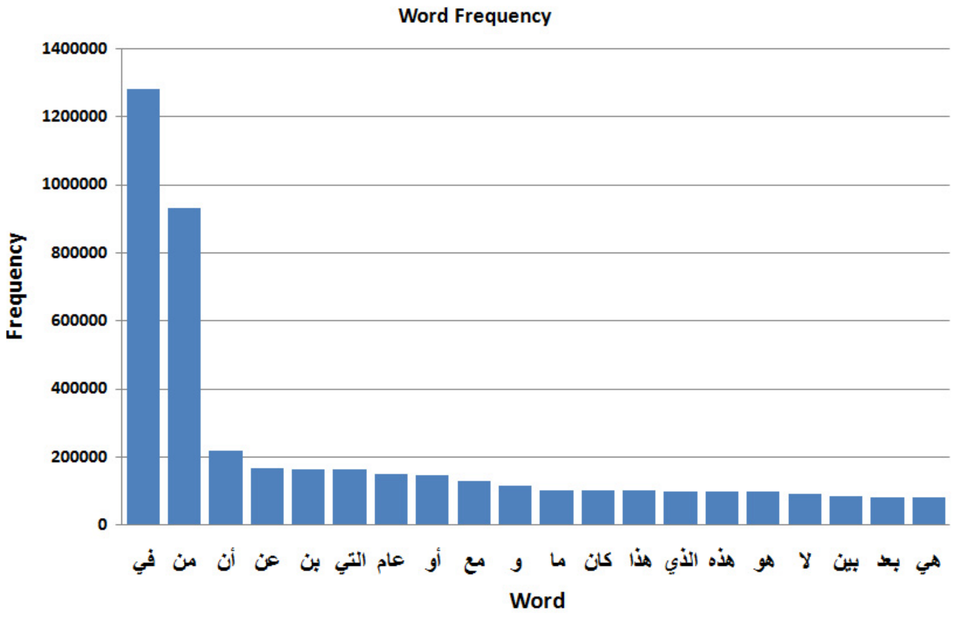


Figure 1a. The top 20 words in terms of frequency.

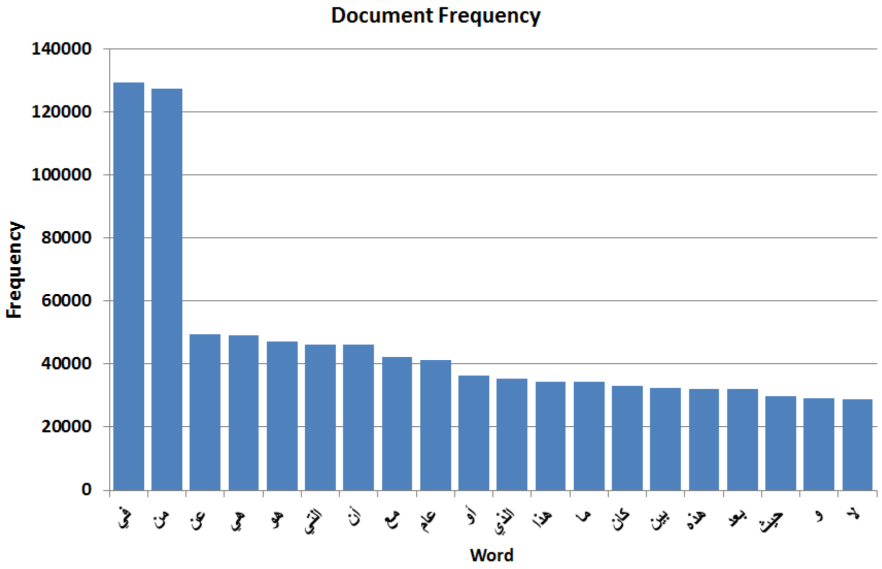


Figure 1b. The top 20 words in terms of document frequency.

logarithmic formula:

$$n^{\log(1.6)} \qquad (2)$$

In order to apply this formula, the list of words in the corpus must be sorted based on their frequency. After that, the positions of selected words are specified based on the sequence of numbers resulting from the logarithmic formula above. Figure 2 shows the selected word sample using the Zipf's Law distribution.

These three sampling methods yield three different word samples, which are used in the next steps of size estimation. Later the processing results of each sample are compared to decide which sample gives better results. Table 3 shows the translation for every word selected by the three sampling methods.

A sample of URLs is then created using the search engines involved in the experiments Google, Yahoo and Bing, by querying each search engine with each word in the sample. The first 10 returned URLs from each query are added to the URL sample. This yields three URL samples, where the size of each sample is:

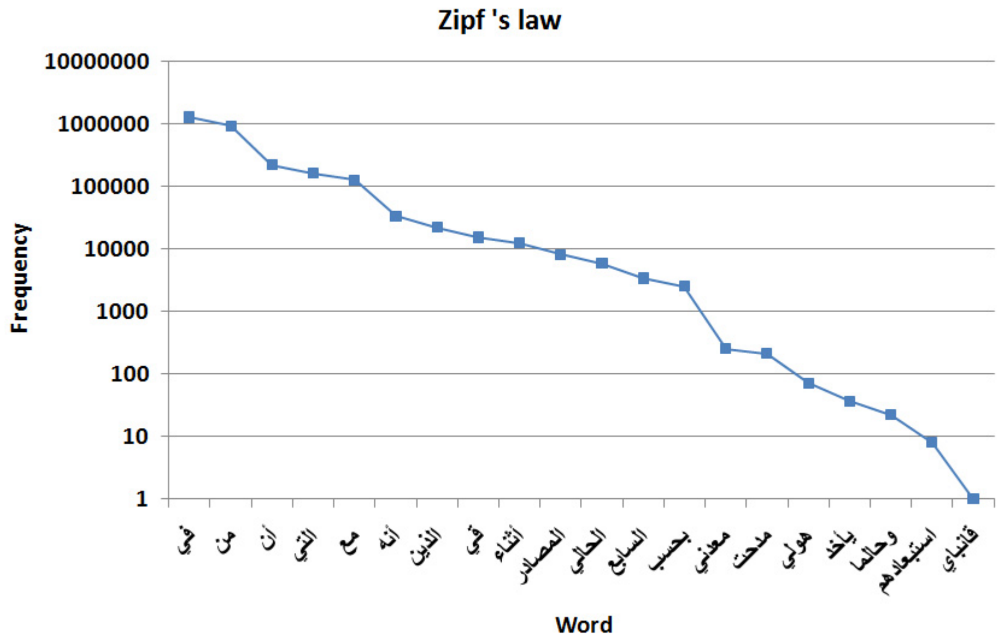


Figure 2. Word sample using Zipf's Law.

Table 3. Word translation.

Arabic word	Translation	Arabic word	Translation	Arabic word	Translation
في	In	هذا	This (Male)	أثناء	During
من	Of/From	الذي	Which (Male)	المصادر	Sources
أن	That	هذه	This (Female)	الحالي	Present
عن	About	هو	He	السابع	The seventh
بن	Son	لا	No	بحسب	According to
التي	Which (Female)	بين	Between	معدني	Metallic
عام	Year	بعد	After	مدحت	A male name
أو	Or	هي	She	هولي	Holly
مع	With	حيث	Where	يأخذ	Takes
و	And	أنه	That He	وحالما	Once the
ما	What	الذين	Whose	استبعادهم	Excluded them
كان	Was	قي	In	قانباي	A male name

$$\begin{aligned} \left|\text{URL Sample}\right| &= \text{number of search engines} \times \\ &\quad \text{number of URLs} \times \\ &\quad \left|\text{words sample}\right| \\ &= 3 \times 10 \times N \end{aligned} \tag{3}$$

The coverage of each search engine is calculated by querying each search engine with every URL in the URL sample. However, for each search engine, we exclude the URLs that were originally generated by that search engine. There will definitely be an overlap among the search engines, in terms of coverage. Consequently, the cross overlap must be calculated to obtain the percentage of unique URLs indexed by each search engine, which represents the number of URLs found in one search engine, but not in the others. The process of calculating the cross overlap is performed as follows: each URL returned by search engine X is checked, whether or not it is indexed by search engine Y, by sending that URL as a query to engine Y. Then, the first result returned by engine Y is compared with the query URL. If it is an exact match, then this URL is also indexed by Y. The coverage of each search engine is now represented by the percentage of URLs from the URL samples that are indexed by each search engine.

After that, a list of words is selected from the word corpus using the Zipf's Law distribution, and Equation 1 is applied for each word. In order to determine the weighted estimated size for each search

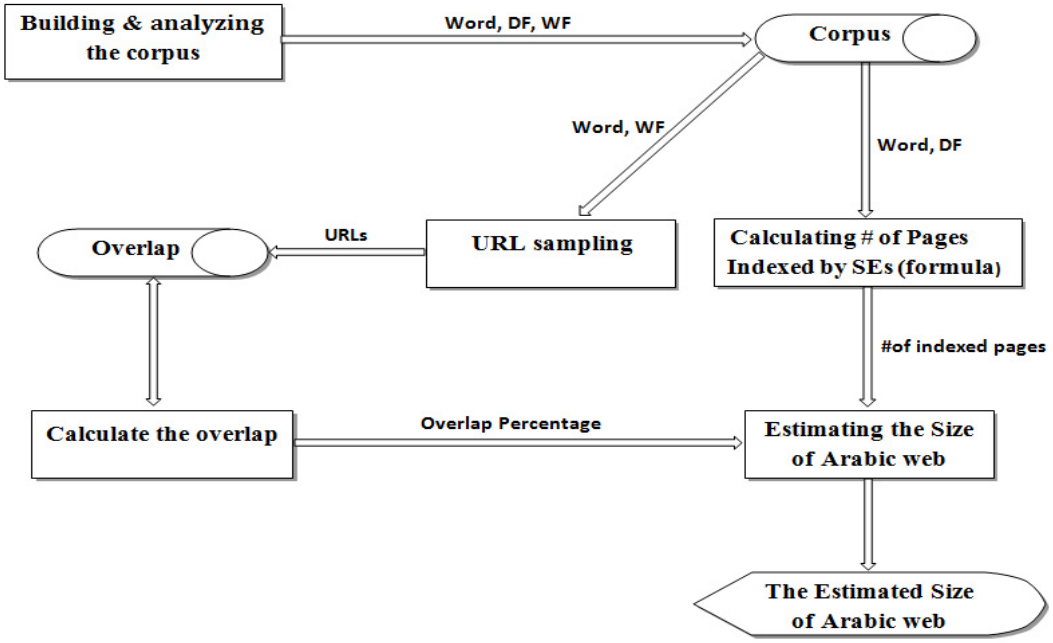

Building & analyzing the corpus
Word, DF, WF
Corpus
Word, WF
Word, DF
Overlap
URLs
URL sampling
Calculating # of Pages Indexed by SEs (formula)
#of indexed pages
Calculate the overlap
Overlap Percentage
Estimating the Size of Arabic web
The Estimated Size of Arabic web

Figure 3. Flow Chart of the processes involved in estimating the size of Arabic content on the Web.

engine, the sum of all weighted estimated sizes related to that search engine is divided by the number of words selected. The result of this step gives the total number of pages indexed by each search engine.

The final step is to estimate the number of indexed Arabic pages on the Web, rather than the number found by each search engine. This is achieved by adding the weighted estimated sizes produced by all the search engines, and then deducting the overlap of the coverage of the search engines. Figure 3 shows the sequence of steps involved in this approach.

IMPLEMENTATION

The Arwiki XML file, arwiki-latest-pages-articles.xml.bz2, which was last updated on October 11, 2010, at 10:25:38, is 146.2 megabytes in size and was downloaded from Wikimedia to create the required corpus. This XML file was then parsed into a text file using the Apache Xerces Java open source XML parser. We decomposed the file into smaller parts (33 text files) for ease of handling, using Text File Splitter 1.5.0. For each of the 33 text files, a PHP function was applied, which calculates the frequency of every word in a file. This resulted in a two-column array, one containing all the words in the file, and the other recording the word frequencies. The resulting 33 tables were merged into a single table in the SQL database. Each word has a single record in the table, where the first field holds the word itself and the second field records the frequency for that word as it appears in the entire Arwiki File.

We also calculated the document frequencies for each word, and added them to our table. For example, for a word with a word frequency of 1, its document frequency must be 1 as well. This enabled us to avoid the long process of eliminating a large number of rare and misspelled words. For every word with a word frequency higher than 1, the number of articles containing that word is counted. To achieve this, all the articles are searched for each word, and, as soon as the word has been found, the document frequency is increased, the rest of the article is ignored, and the search moves to the next article.

As a result of the previous steps, we have a table of words with their word and document frequencies. We then created a new table in the database for calculating the coverage of each search engine, and the overlap among the search engines. A number of words were selected from the corpus, using the three different sample selection algorithms.

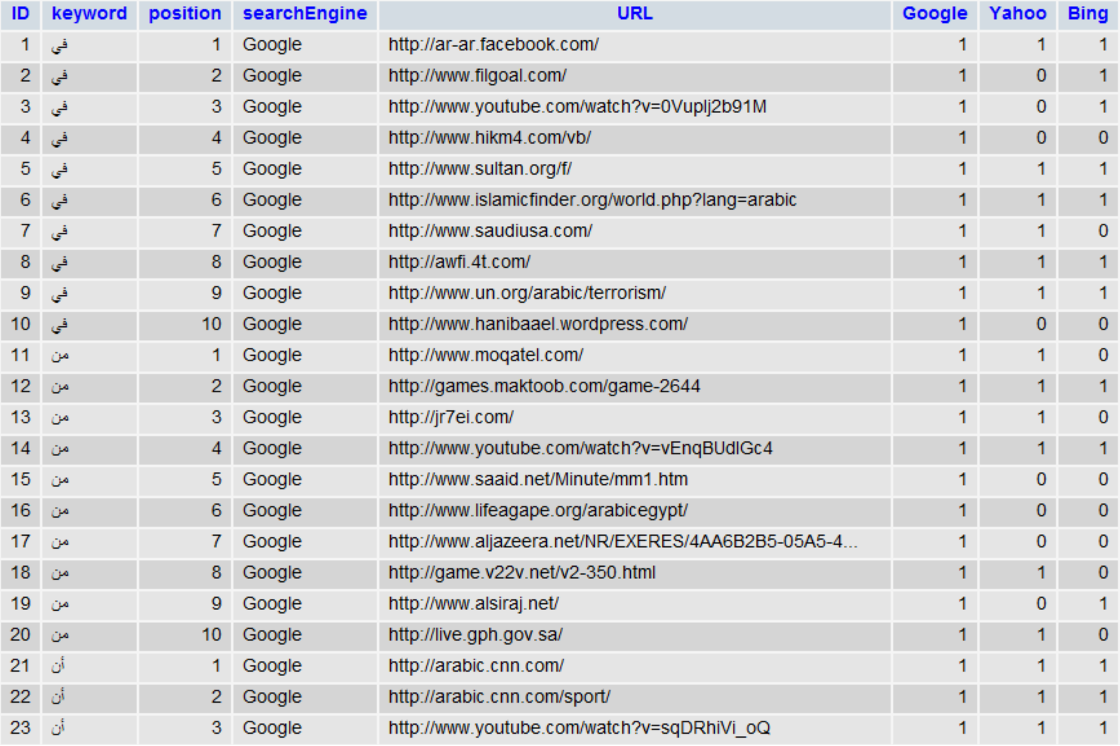
Each word in the sample was sent to each of the search engines as a query, using the application programming interface (API) they provide. These APIs make it possible to automate this process and simplify the accessing and querying of search engines. The table was populated with the resulting URLs. Then, we calculated the coverage for each search engine, which is the percentage of URLs indexed by each search engine from all the URLs sampled. Figure 4 shows the resulting table. We also calculated the cross overlaps, which is the percentage of URLs found by one search engine, but not by the others.

Then, to estimate the number of the Arabic pages indexed by each search engine, a list of words was selected using the Zipf Law distribution, and, for each selected word, formula 1 was applied. This process was repeated every day during the experiment, and these figures were stored in a table in the database. Finally, we added the number of Web pages indexed by a specific search engine to the number of pages uniquely indexed by the other two search engines (cross overlap percentage). The result is the estimated size of Arabic indexed Web content.

To display the final results, a website was created, which is shown in Figure 5. The portal shows graphs and figures giving the estimated size of Arabic indexed content on the Web. The information on the website is updated automatically by the underlying PHP code.

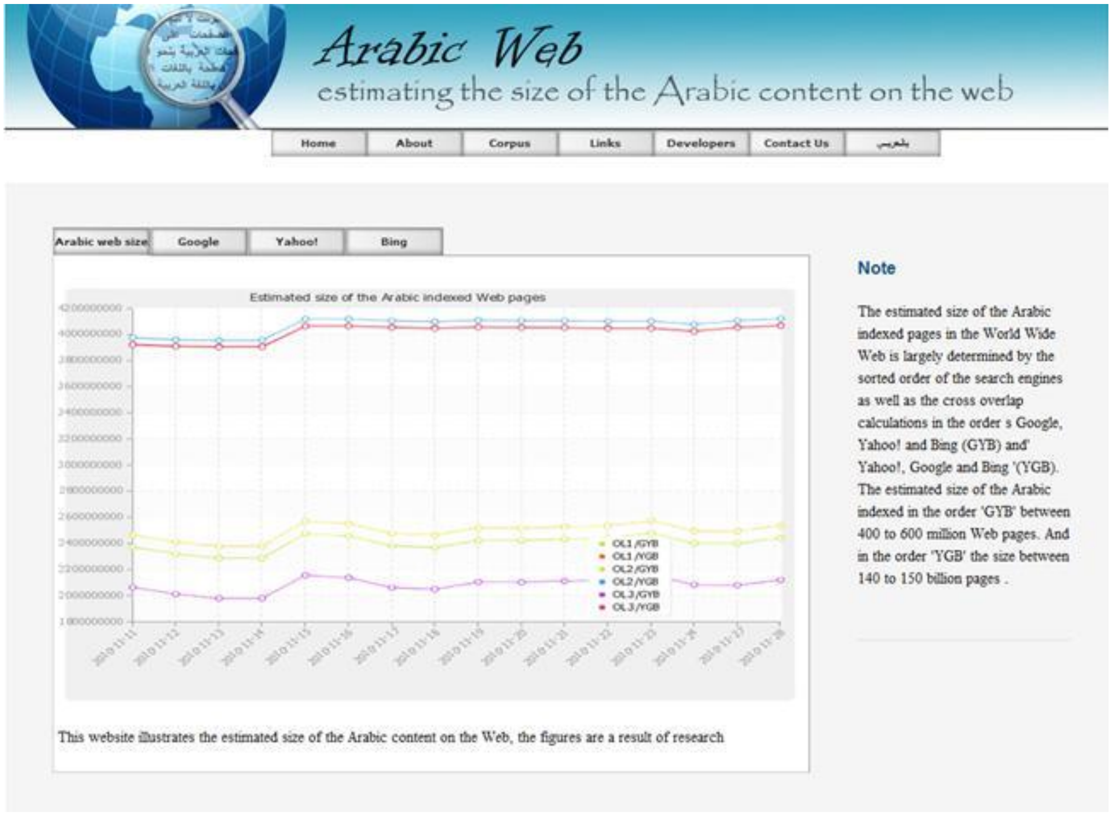
RESULTS AND PERFORMANCE EVALUATION

The experiment resulted in an estimation of the size of



ID	keyword	position	searchEngine	URL	Google	Yahoo	Bing
1	في	1	Google	http://ar-ar.facebook.com/	1	1	1
2	في	2	Google	http://www.filgoal.com/	1	0	1
3	في	3	Google	http://www.youtube.com/watch?v=0Vuplj2b91M	1	0	1
4	في	4	Google	http://www.hikm4.com/vb/	1	0	0
5	في	5	Google	http://www.sultan.org/f/	1	1	1
6	في	6	Google	http://www.islamicfinder.org/world.php?lang=arabic	1	1	1
7	في	7	Google	http://www.saudiusa.com/	1	1	0
8	في	8	Google	http://awfi.4t.com/	1	1	1
9	في	9	Google	http://www.un.org/arabic/terrorism/	1	1	1
10	في	10	Google	http://www.hanibaael.wordpress.com/	1	0	0
11	من	1	Google	http://www.moqatel.com/	1	1	0
12	من	2	Google	http://games.maktoob.com/game-2644	1	1	1
13	من	3	Google	http://jr7ei.com/	1	1	0
14	من	4	Google	http://www.youtube.com/watch?v=vEnqBUdlGc4	1	1	1
15	من	5	Google	http://www.saaid.net/Minute/mm1.htm	1	0	0
16	من	6	Google	http://www.lifeagape.org/arabicegypt/	1	0	0
17	من	7	Google	http://www.aljazeera.net/NR/EXERES/4AA6B2B5-05A5-4...	1	0	0
18	من	8	Google	http://game.v22v.net/v2-350.html	1	1	0
19	من	9	Google	http://www.alsiraj.net/	1	0	1
20	من	10	Google	http://live.gph.gov.sa/	1	1	0
21	أن	1	Google	http://arabic.cnn.com/	1	1	1
22	أن	2	Google	http://arabic.cnn.com/sport/	1	1	1
23	أن	3	Google	http://www.youtube.com/watch?v=sqDRhiVi_oQ	1	1	1

Figure 4. Overlap database.



Note

The estimated size of the Arabic indexed pages in the World Wide Web is largely determined by the sorted order of the search engines as well as the cross overlap calculations in the order s Google, Yahoo! and Bing (GYB) and' Yahoo!, Google and Bing '(YGB). The estimated size of the Arabic indexed in the order 'GYB' between 400 to 600 million Web pages. And in the order 'YGB' the size between 140 to 150 billion pages .

This website illustrates the estimated size of the Arabic content on the Web, the figures are a result of research

Figure 5. A snapshot of our website that shows the estimated size of Arabic content on the Web.

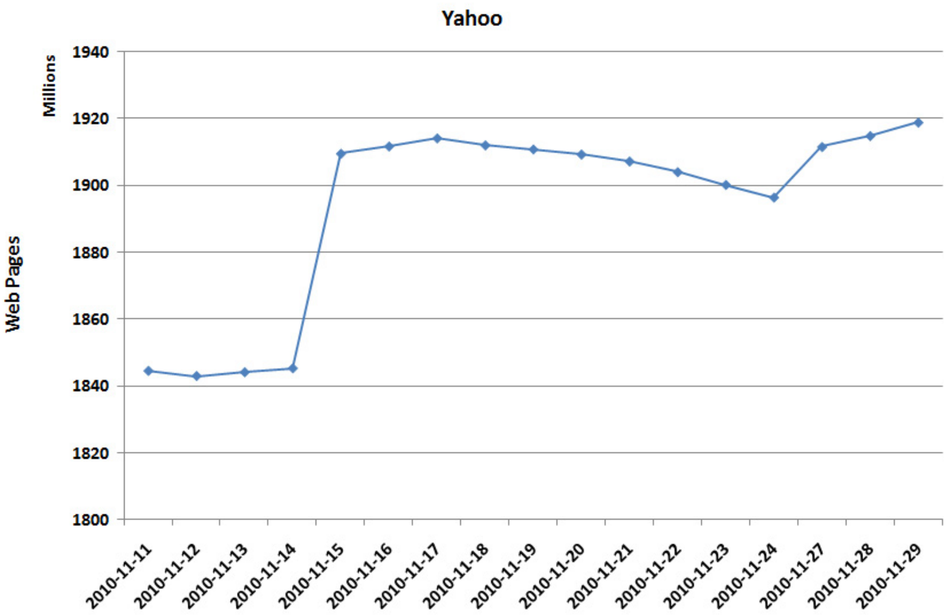


Figure 6. Yahoo's estimated Arabic index size.

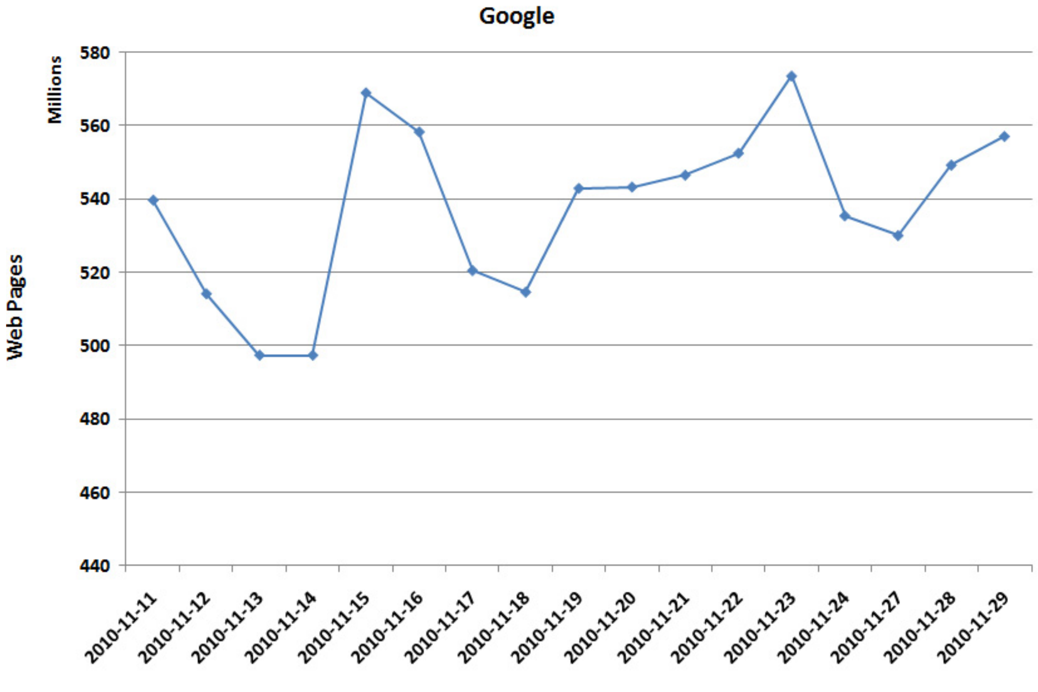


Figure 7. Google's estimated Arabic index size.

Arabic content indexed by the three search engines involved: Google, Yahoo and Bing, at 557,112,990, 1,918,909,680 and 254,924,429 pages, respectively. Figures 6, 7 and 8 show the index size of these search engines over twenty days, starting on November 11 and ending on November 29, 2010.

Figure 6 shows that, the number of Arabic Web pages indexed by Yahoo was around 1.8 billion pages at the beginning of the experiment. In the first three days of the experiment, there was a large increase in this number, to

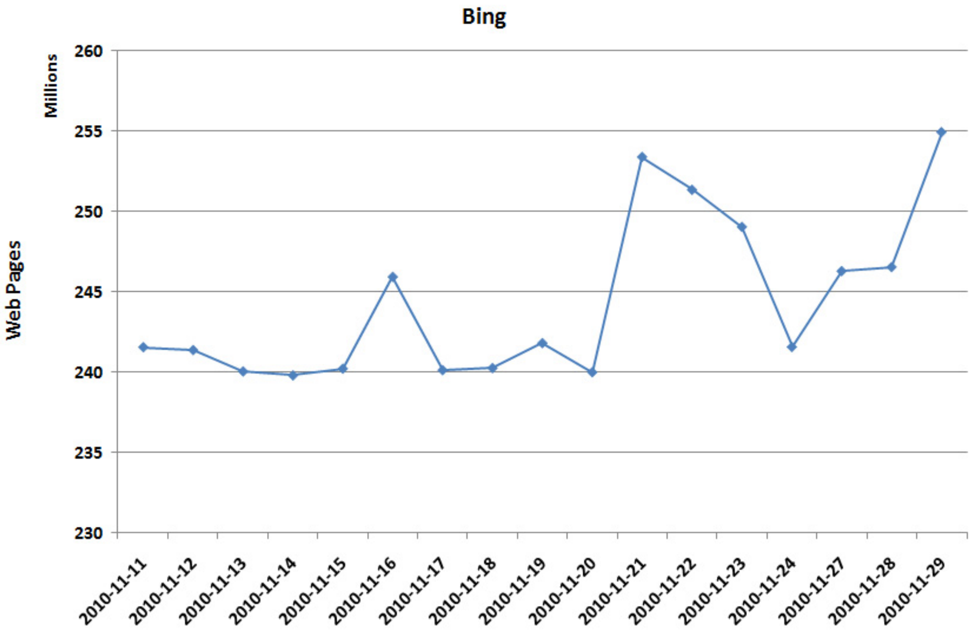


Figure 8. Bing's estimated Arabic index size.

1.9 billion pages on November 14, 2010. The content remained stable during the remainder of the experiment.

The number of Arabic pages indexed by Google averaged 550 million pages, although we noted major changes in the results during the experiment, as shown in Figure 7.

Bing indexed a smaller number of Arabic pages – a maximum of 255 million pages at its peak, as shown in Figure 8.

There are a number of possible reasons for the fluctuations we noted in the search engines results over time, such as the following (Mettrop, 2001; Vronis, 2005):

1) Search engines use different indexes according to Wouter Mettrop, although the indexes of a search engine are periodically synchronized, they are usually not 100% identical,
2) Different indexes are used in different parts of the world,
3) Retrieval algorithms are performance-dependent. The more queries the search engine receives, the more chance there is of fluctuations,
4) Duplicate pages are removed.

In addition, we have noted an increase in the number of Arabic Web pages indexed by all three search engines from November 14 to November 17, 2010. This may be because of an increase in Arabic content due to special events taking place in Arabic regions.

To estimate the size of the Arabic content indexed on the Web, two measurements were performed with the following orderings: Google, Yahoo and Bing (GYB) and Yahoo, Google and Bing (YGB). Each of these orderings was tested on the overlap percentages related to the three URL samples. This resulted in six individual experiments being performed, as shown in Figure 9. The estimated size of the Arabic indexed Web content in the GYB ordering is between 1.2 and 1.3 billion pages, and in the YGB ordering, it is between 2 and 2.1 billion pages.

The reason we obtained different total estimations from these measurements is that the overlap was subtracted in different sequences in the two calculations. The fact is that the total estimated size of the Arabic indexed Web content is underestimated because the overlap is overestimated. It appears that starting the calculations with Yahoo's figure and then adding the number of unique Web pages in Google and Bing yields higher numbers than starting with Google’s figure and adding the number of unique Web pages in Yahoo and Bing. This is because, according to our previous calculations, Yahoo seems to index a larger number of Arabic Web pages. After performing the experiment on the three overlap samples, we found that sample 2 (restricted uniform random word selection) gives the largest number of pages. This can be justified by the fact that the other two sampling techniques (uniform random word selection and Zipf's Law word selection) may be affected by extreme word frequency values.

In Table 4, we used sample 2 to calculate the overlap among the search engines, and then assumed that Google had the greatest coverage, at 557,112,990 Web pages. After that, the unique URLs in Yahoo and Bing

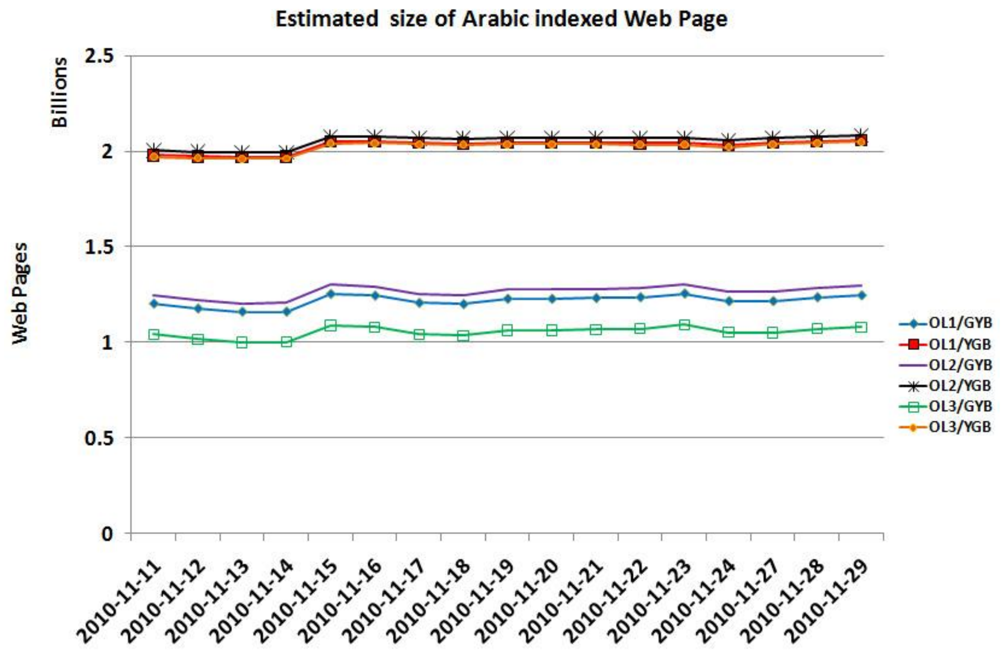


Figure 9. Estimated size of Arabic content on the Web.

Table 4. Estimated Arabic Web OL2/(GYB) as of 29-11-2010.

Search engine	Coverage percentage	Number of URLs	Results
Google	100	557,112,990	557,112,990
Yahoo	36.65	1,918,909,680	703,280,398
Bing	13.30	254,924,429	33,904,949
Total URLs			1,294,298,337

Table 5. Estimated Arabic Web OL2/(YGB) as of 29-11-2010.

Search engine	Coverage percentage	Number of URLs	Results
Yahoo	100	1,918,909,680	1,918,909,680
Google	23.48	557,112,990	130,810,130
Bing	13.30	254,924,429	33,904,949
Total URLs			2083660312

were added as a percentage of their total number of Arabic indexed Web pages. This calculation estimates the size of the Arabic indexed content on the Web at 1,294,298,337 Web pages.

In Table 5, we used the same word sample (sample 2) with the aim of comparing the results when we apply a different order: YGB. This calculation estimates the size of the Arabic indexed content on the Web at 2,083,660,312 Web pages.

Although, our system uses three main search engines and different sampling techniques, the system has two essential limitations. First, it estimates the size of the Arabic content on the surface Web, rather than that of the deep Web. Because of the dynamic nature and huge size of the deep Web, it is very hard to estimate its size, and doing so is beyond the scope of this paper. Second, the system relies on the search results provided by the search engines, which are not stable, and its fluctuations affect the accuracy of our system.

CONCLUSION AND FUTURE WORK

The estimated size of the Arabic content on the Web is approximately 2,083,660,312 pages. This shows that Arabic content constitutes a representative portion of the size of the Web, considering the number of Arabic speakers in the world. Trends also show that this portion is growing very rapidly. As for the search engines involved in the experiment, Yahoo appears to have the largest index of Arabic content. However, the accuracy of the results obtained depends on the accuracy of the search engines queries and their APIs. Applying the methodology on different corpora and using different sampling algorithms may be considered in the future, to study the effect of each on the results we have obtained.

REFERENCES

Aires R, Santos D (2002). Measuring the Web in Portuguese. Proc. EuroWeb Conf. pp. 198–199.

Al-Sulaiti, Latifa, Atwell, Steven E (2006). The design of a corpus of contemporary Arabic. Int. J. Corpus Linguist. 11(2):135–171.

Alpert J, Hajaj N (2008). We knew the Web was big. Available at: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Bergman MK (2001). The deep Web: Surfacing hidden value. J. Electronic Publishing, 7(1).

Bharat K, Broder A (1998). A technique for measuring the relative size and overlap of public Web search engines. Comp. Netw. ISDN Syst. 30(1-7):379–388.

ESCWA (2010). Report of the final meeting of the project on promotion of the digital Arabic content industry through incubation. Available at: http://www.escwa.un.org/

Gong Z, Zhang J, Liu Q (2006). Hidden Web database exploration. In Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications. pp. 838–843.

Gulli A, Signorini A (2005). The indexable Web is more than 11.5 billion pages. Proceedings of the 14th International Conference on the World Wide Web. ACM, New York. pp. 902–903.

Gupta P, Johari K (2009). Implementation of Web crawler. Proceedings of the Second International onference on Emerging Trends in Engineering and Technology, 16-18 December, 2009. pp. 838–843.

Hitwise (2010). Hitwise US -- leading search engines. Available at: http://www.hitwise.com/datacenter

Kunder MD (2010). The size of the World Wide Web. Available at: www.WorldWideWebSize.com

Lawrence S, Giles CL (1998). Searching the World Wide Web. Sci. 280(5360):98–100.

Madhavan J, Ko D, Kot L, Ganapathy V, Rasmussen A, Halevy A (2008). Google's deep Web crawl. In Proceeding of VLDB Endow. 1:1241–1252.

Marcus MP, Marcinkiewicz MA, Santorini B (1993). Building a large annotated corpus of English: the Penn Treebank. Comput. Linguist. 19:313–330.

Mayer T (2005). Our blog is growing up, and so has our index. Available at: http://www.ysearchblog.com/2005/08/08/our-blog-is-growing-up-and-so-has-our-index/

Mettrop W (2001). Internet search engines fluctuations in document accessibility. J. Doc. 57(5):623–651.

Peisu X, Ke T, Qinzhen H (2008). A framework of deep Web crawler. The 27th Chinese Control Conference. pp. 582–586.

KACST (2010). King Abdullah Initiative for Arabic Content. Available at: http://www.econtent.org.sa

Singh M (2002). Deep Web structure. IEEE Internet Comput. 6(5):4–5.

Tawileh A, Alghamedi M (2011). A Corpus-based Linguistics Approach for Estimating Arabic Online Content. In Proceeding of the Conference on Human Language Technology for Development.

Suwaidi MA (2010). Alwaraq Encyclopedia. Available at: http://www.alwaraq.net

Vronis J (2005). Web: Google adjusts its counts. Available at: http://blog.veronis.fr/2005/03/web-google-adjusts-its-counts.html

Wikimedia (2010a). Arabic Wikipedia dumps. Available at:http://download.wikimedia.org/arwiki/

Wikimedia (2010b). List of Wikipedias. Available at: http://meta.wikimedia.org/wiki/List_of_Wikipedias

Wikipedia (2012). Zipf's Law. Available at: http://en.wikipedia.org/wiki/Zipf's\textunderscorelaw