A Corpus Linguistics-based Approach for Estimating Arabic Online Content

Anas Tawileh
Systematics Consulting
anas@systematics.ca

Mansour Al Ghamedi
King Abdulaziz City for Science and Technology
mghamdi@kacst.edu.sa

Abstract

This paper presents the results of a research project for the development of a corpus-based indictor for Arabic online content. The project entailed the construction and analysis of three Arabic language corpora from online and offline sources. These corpora where then utilized in the estimation of the index size of two major search engines. The overlap between these two indices was determined, and an indication of the size of the indexed Arabic web was deduced. The developed indicator suggests that the current size of the indexed Arabic web exceeds 400 million pages.

1 Introduction

In today's interconnected world, the competitive advantages of societies and nations are largely determined by the degree of technology adoption and utilization by these societies. Information and Communication Technologies (ICTs) are considered to be a critical component in the emergence of knowledge societies. The concept of the knowledge society is based on the premises that knowledge can be leveraged to address developmental challenges and improve the quality of life of societies around the world. Mounting evidence supports this argument (Mansell and Wehn, 2000; Stone and Maxwell, 2005), and many countries are launching proactive initiatives to capitalize on knowledge and ICTs for delivering on their development agendas.

It has been suggested that the Arab region is facing particular challenges in its attempts at shifting into knowledge societies. A major challenge that is frequently cited is the severe shortage of relevant local content in the Arabic language. Researchers and practitioners alike argue that without a critical mass of local content, which constitutes the cornerstone of the knowledge society, Arab countries cannot reap the benefits of the global information revolution.

However, despite the importance of understanding the size and quality of Arabic content available online, little research has been done to systematically assess and measure this content in a rigorous manner. This paper presents the results of a research project conducted by King Abdulaziz City for Science and Technology (KACST) and aimed at the development of an indicator for Arabic online content based on computational linguistic corpora. The paper is structured as follows: The next section provides a background for the research and its design, followed by a detailed description of the corpora development process and results. The paper then highlights the findings of the project and provides pointers for further research.

2 Background

The decentralized nature of the Internet's architecture makes it a very dynamic entity that remains constantly in a state of change. New hosts are added to the web on a daily basis, and many others get disconnected for several reasons. This greatly affects the accessibility to the content available on these hosts. However, the general trend has always been one of dramatic growth. The decentralization of the web extends beyond its technical architecture to encompass content production and dissemination. In today's web 2.0, any user connected to the Internet can pro-

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

duce, share and distribute content, using a plethora of tools and platforms for content sharing, such as Flickr, Facebook, YouTube, SlideShare, etc. A study by the International Data Corporation (IDC, 2010) reported that the world's digital output currently stands at 8,000,000 petabytes and may surpass 1.2 zettabytes this year (a zettabyte is 10 to the power of 21).

These characteristics pose significant challenges for attempts to measure the size of content on the web. Appreciating the size of the online content is very important to understand trends in content development and growth, and in supporting the planning and implementation activities of online content initiatives. Several researchers have developed different approaches to estimate the size of the web. The common element of most of these approaches is their reliance on measuring the sizes of online search engines to infer the size of the web at large (Lawrence and Giles, 1998; Gulli and Signorini, 2005; de Kunder, 2006).

Search engines were conceived as tools to facilitate information search and retrieval on the web. The major search engines continuously crawl the Internet to index the content they find in order to make it easily accessible for their users. These indices are then exploited to provide users with the ability to search through these engines and find content relevant to their search criteria.

Given the large, and increasingly growing, number of users on the web, it is difficult to imagine that a single search engine would be capable of indexing all the content available on the web. Therefore, any single search engine would only be able to provide its users with a subset of the information available on the web (Bharat and Broder, 1998). However, due to the dramatic technological developments in the area of information indexing, search and retrieval, the major search engines can serve as an effective tool to connect users searching for information and information relevant to their search criteria. The subset of the Internet that is actively indexed by search engines is typically referred to as "the indexed web" or "the surface web" (He, Patel et al., 2007). Internet usage trends demonstrate that users rely on search engines to find information on the web. Content that is not indexed by search engines is effectively "hidden away" from normal Internet users (Capra Iii and Pérez-Quiñones, 2005). Hence, it can be assumed that a reasonable estimate of the content available on the indexed web will give a good picture of the overall online content available for Internet users.

In their study, Lawrence and Giles (Lawrence and Giles, 1998) sent a list of 575 search queries to six search engines, and analyzed the results to determine the overlap between these engines and estimate the size of the indexed web. Gulli and Signorini (Gulli and Signorini, 2005) adopted a similar approach in which they sent 438,141 one term queries to four major search engines. They then analyzed the results to estimate the size of the indexed web based on the relative size estimation of each search engine and the absolute size reported by the search engine.

De Kunder (de Kunder, 2006) argued that these approaches have inherent limitations resulting from their dependence on search engine results and overlap. He proposed a different approach that utilizes linguistic corpora in the estimation of the size of the indexed web. De Kunder implemented his approach in a project that estimated the size of the English and Dutch web.

In this research, we follow a linguistic corpora-based approach similar to de Kunder's, and apply it to the development of an indicator for Arabic online content. A critical dependency for this approach is the availability of relevant Arabic corpora that can be utilized in calculating the indicator. Due to the severe lack of Arabic computer linguistic resources in general, and those specific to online content in particular, we developed two corpora from Wikipedia and the web. The following section describes the adopted methodology and the research process in greater detail.

3 Methodology

To ensure that the estimate of the Arabic online content is reasonable and reflects the actual reality, its development should be based on a sound theoretical foundation. This foundation can be established through the concepts and ideas of computational linguistics. The most relevant of these in Zipf's law (Li, 1992), which states that in a large enough language corpus the frequency of any word is inversely proportional to its rank in the frequency table. To explain this law, the concept of word frequency must be first illustrated. In a language corpus, the frequency in which a specific word occurs in this corpus is

referred to as the word frequency of this word. A frequency table, also referred to as a frequency list, of a particular corpus is a simple table that contains each word along with its corresponding frequency in the corpus.

Based on this concept, if the word frequency of a specific word can be determined in a corpus of a known size, and in another content repository of unknown size, the size of the repository can be deduced by calculating the proportion of the word frequency in each set. Because search engines typically return the number of documents that result from a specific search query, this number can be exploited to estimate the actual size of the search engine's index. This number, however, does not refer to the word frequency in the search results, but rather indicates the total number of documents that contain the search word.

For example, searching for the word "the" on Google returns 23,990,000,000 results ("the" is the word with the highest word frequency in the English language). However, the actual word frequency for "the" will most likely be much higher than this figure, because the word will most probably appear more than once in each web page returned in the search result. To address this issue, a new concept can be introduced, referred to as "document frequency". Document frequency indicates the number of documents in the corpus that include the word being considered, regardless of the number of occurrences of the word in each document. Following this logic, the index size of the search engine Google can be deduced based on the document frequencies in an appropriate corpus. Figure 1 illustrates this approach.

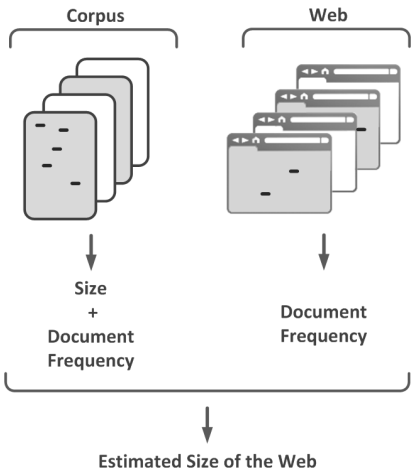

Figure 1. Estimating the Size of the Indexed Web

3.1 Corpora Building

The successful development of the content indicator depends on the availability of high quality Arabic language corpora. The corpora constitute the critical foundation that will provide word frequency statistics for the calculation of the indicator. Therefore, the first step in the development process entailed building a set of Arabic language corpora relevant to the content typical of the web.

To increase the accuracy and relevance of the indicator, more than one corpus will be used in its estimation. This will diversify the content base for the linguistic statistics and increase the corpora representation of the actual web. Three corpora were built as part of this project, using materials from the following sources:

- The Arabic Wikipedia
- The Open Directory Project
- The Arabic Contemporary Corpus

The choice of these corpora intends to diversify the fingerprints used in the calculation of the index. For example, the corpus of Wikipedia articles will have more words than the Open Directory Project because the articles on Wikipedia are mostly self-contained, while on the web pages indexed by the ODP, the same piece of content may spread over several pages. By incorporating these differences in the calculation method, more robust estimates can be made.

The corpora building process also includes the analysis and identification of the word and document frequencies in each corpus. A web-based application was developed to handle this task which will generate a table of words and their word and document frequencies in a database format. This frequency table forms the basis for the calculation of the Arabic content indicator.

Particular care was taken in the construction of the corpora from web content in order to ensure the quality of the collected materials. For example, while some sites might be targeting the Arab region, or be actually based in an Arab country, their content may not be necessarily presented in Arabic. The corpus building process will disqua-

lify any website not written in Arabic and exclude it from the corpus. Another consideration is the accessibility of the websites. Because search engines do sometimes retain links to inaccessible websites in their indices, these websites must be excluded from the corpus. This was achieved by incorporating a timeout limit on the web page retrieval in the application. Additionally, web pages that contain a very small number of words (less than 10 words) were not considered. These pages are mostly navigation panels or website introductions, and including them in the corpus will skew its linguistic statistics.

3.1.1 The Web Corpus

This corpus was intended to capture content from the public Arabic web. There are very little similar resources already available, and therefore a decision was taken to construct this corpus by an extensive crawling process on the Arabic web. To start the process, an initial list of Arabic language websites was required. The Open Directory Project (ODP) is the "largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors" (ODP, 2011). All the data available in the ODP is available under an open license that permits reuse and adaptation.

The complete directory of the ODP was downloaded and analyzed to identify the Arabic language links. These links were then extracted from the downloaded XML file and inserted into a MySQL database table. Upon completion, the "Directory" table contained information about **7,595** Arabic web sites, and was ready for use in the crawling process. The crawling process then followed the links in each of these sites and extracted their content, removed any markup, converted the content's encoding into Unicode and stored it in a database.

After the completion of the crawling process, all pages that contain less than 10 words were removed from the database. Pages that contain such a small number of words are usually redirects to other pages or of the type similar to "Under Construction", and would provide little value for the corpus.

By the end of the process, **75,560** pages were added to the corpus, with a total size of 530.1 MB. This constitutes the actual size of the corpus.

In order to utilize the web corpus built by the crawling process, word frequency statistics should be calculated for use in the estimation of the indicator. This process started by extracting the word occurrences in the page content in the corpus, and creating a table for these words. This exercise resulted in **659,756** unique words being added to the word list table. Word and document frequency for each of these words were then calculated by specifically designed scripts.

3.1.2 The Wikipedia Corpus

The second corpus was extracted from the Arabic Wikipedia (http://ar.wikipedia.org). To build this corpus, a complete dump of the database of the Arabic Wikipedia was downloaded from the official Wikipedia download site (http://download.wikimedia.org/arwiki/latest/) on the 24th of July 2010.

All articles that contained less than 10 words were deleted from the database, resulting in a total corpus size of **95,140** articles and 213.3 MB.

The same process described in the previous section was followed to generate the word list, document frequency and the word frequency tables for the Wikipedia corpus, using the same scripts. The extracted word list contained **760,690** distinct words, and the document and word frequencies were calculated for each.

3.1.3 Corpus of Contemporary Arabic

The third corpus was the Corpus of Contemporary Arabic (CCA), developed by Dr Latifa Al-Sulaiti at the University of Leeds (Al-Sulaiti, 2010). The content of this corpus was obtained from Arabic web sites, and although its size is rather small, it is useful to include so that the coverage of the proposed indicator is diversified even further.

The corpus was downloaded on the 25th of July 2010 from the author's website: http://www.comp.leeds.ac.uk/eric/latifa/research.htm. The contents of the corpus, totaling **377** articles, were imported into a database table, and the same process for extracting the word list was

followed. The word list for the CCA contained **82,878** words. The document and word frequencies for each of these words were calculated and added to the appropriate database tables.

3.2 Selection of Search Engines

To further improve the reliability of the Arabic content indicator, and increase its representation of the indexed web, two search engines were utilized in the calculations rather than one. The selection of these search engines depends on two factors: their popularity among Arab Internet users, and the size of their search index.

Based on these criteria, the selected search engines for use in the calculation of the Arabic content indicator are Google and Yahoo!. According to ComScore, these two search engines command 65% and 20% (respectively) (comScore, 2009) of the global search engine market. Microsoft's recently launched search engine, Bing, was also considered, but the numbers of the results it returned indicate a very small index size for Arabic content, and hence it was excluded from the selection. The situation of the Arabic search engine Ayna was very similar to that of Bing. The small index size of these two search engines makes their impact on the estimated size of the indexed Arabic web negligible, and hence they were not considered.

It is crucial when more than one search engines are used to estimate the size of the indexed web that the overlap between these search engines is identified and accounted for. Naturally, the two search engines would have crawled and indexed the same pages. For example, the homepage of Wikipedia – www.wikipedia.org – can be found in the indices of both Google and Yahoo!. At the same time, the different search engines will have indexed different websites due to the differences in their crawling and information retrieval mechanisms. This is evident in the different number of results returned by the two search engines for the same search term.

To determine the overlap between the two search engines, a sample of URLs was generated by applying a logarithmic selection to the three corpora to extract 200 words from each corpus. The logarithmic selection was chosen to provide the widest coverage of the corpora according to Zipf's law. Each of these words was then sent to both search engines, and the first 10 result URLs were collected. These results were then compared to determine the overlap between the two search engines.

3.3 Application Development

The Arabic online content indicator will be calculated and made available on the web on a daily basis. A dedicated application was developed to perform these calculations and present the results in a web-friendly format, including a textual and graphical representation.

The application uses a logarithmic Zipf word selection process to obtain a representative selection of words from the three corpora that were built. The logarithmic selection ensures that the selected words are representative of the actual distribution of the corpora, as it ensures that the selection covers words of widely varying document frequencies. To compile this list, the application starts with a normal series (1,2,3,4,..etc) and for each number in the list, it will calculate the anti-log(1.6). This will give the following distribution:

1, 2, 3, 4, 7, 10, 17, 27, 43, 69, 110, 176,.. etc

The application then selects the word that corresponds to each of these locations in the word list ranked by document frequency. For each selected word, the application will record the word, its word frequency and its document frequency. This Zipf-generated list will form the basis for the calculation of the Arabic online content indicator.

When the application is invoked, it will fetch the words from the Zipf-generated list, and send each word to both search engines. The application will then record the number of search results returned for each word. The size of the indexed Arabic web is then calculated, with the proper adjustments for the search engines' overlap and other considerations, such as dead links.

To ensure that only Arabic content is returned as a result of the search query, a language filter is applied to the search engine restricting the search language to Arabic. This is very helpful to eliminate sites that may include an Arabic word in their meta-data for example, but their actual content is presented in another language.

The estimate calculated by the application is then stored in the application's database, along with a timestamp to document the date and time of the

measurement. This information is utilized to present the indicator in a textual format, as well as a graphical form. Figure 2 shows the graphical representation of the Arabic Content Indicator.

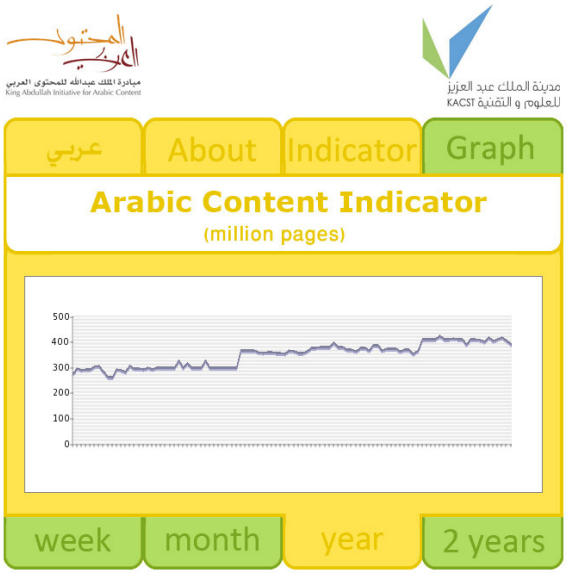

Figure 2. Arabic Content Indicator

As of the 9th of April 2011, the estimated size of the indexed Arabic web reported by the indicator exceeds 413 million pages. The application also recorded an increase of around 40% in the size of the indexed Arabic web over the six months period leading to this date. The fluctuations in the indicator's value are a natural result of the dynamic nature of the web. These happen due to changes in connectivity, hosts going offline or becoming inaccessible.

4 Conclusions and Future Work

This paper presented the design, development and implementation of a linguistic corpora-based approach for the estimation of the size of online Arabic online content. One of the major contributions of this work is the construction and analysis of three Arabic language corpora required for the development of the indicator. These corpora can be utilized by Arabic computer linguistics researchers in their studies, and constitute a significant foundation that can be further developed and built upon.

Another contribution is the development and implementation of the indicator's application that leverages the results of corpora analysis to estimate and report the size of the indexed Arabic web. The Arabic online content indicator is an important element in gauging the size of the Arabic content on the web, and in observing and monitoring its changes over time. It is also of pivotal importance to initiatives that aim at improving Arabic content, as it enables the assessment and evaluation of the impact of these initiatives and projects, as well as informs their design and future planning.

The data that will be collected by the indicator's application is expected to have great value to researchers and practitioners working on Arabic content, and will likely lead to future research and development in this domain. Future work may include the analysis of the trends in the growth of Arabic online content and identifying correlation or causalities into factors that may affect these trends. Another interesting area might be investigating the quality of the Arabic online content to determine redundancy in such content and the topics they cover. While evaluating quality is a challenging endeavor, research into this area that leverages emerging trends of crowdsourced evaluation is very interesting and could shed light into the relevance of online content rather than focus only on quantity. The indicator can also be adapted to include geo-location data that maps the geographic source of the Arabic content and explore the comparative intensity of content production in different parts of the Arab world. The concepts and tools developed as part of this research can also be utilized in measuring the size of online content in other languages.

References

Al-Sulaiti, L. 2010. "Corpus of Contemporary Arabic." Retrieved 25 July, 2010, from http://www.comp.leeds.ac.uk/eric/latifa/research.htm.

Bharat, K. and A. Broder. 1998. "A technique for measuring the relative size and overlap of public web search engines." Computer Networks and ISDN Systems **30**(1-7): 379-388.

Capra Iii, R. G. and M. A. Pérez-Quiñones. 2005. "Using web search engines to find and refind information." Computer **38**(10): 36-42.

comScore. 2009. "Global Search Market." Retrieved 1 February, 2011, from http://www.comscore.com/Press_Events/Press_Releases/2009/8/Global_Search_Market_Draws_More_than_100_Billion_Searches_per_Month.

de Kunder, M. 2006. "Geschatte grootte van het geïndexeerde World Wide Web."

Gulli, A. and A. Signorini. 2005. The indexable web is more than 11.5 billion pages, ACM.

He, B., M. Patel, et al. 2007. "Accessing the deep web." Communications of the ACM **50**(5): 94-101.

IDC. 2010. "The digital universe decade - are you ready?" Retrieved 1 February 2011, from http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm.

Lawrence, S. and C. L. Giles. 1998. "Searching the world wide web." Science **280**(5360): 98.

Li, W. 1992. "Random texts exhibit Zipf's-law-like word frequency distribution." IEEE Transactions on Information Theory **38**(6): 1842-1845.

Mansell, R. and U. Wehn. 2000. Knowledge societies: Information technology for sustainable development, United Nations Pubns.

ODP. 2011. "The Open Directory Project." Retrieved 1 February, 2011, from http://www.dmoz.org/docs/en/about.html.

Stone, D. and S. Maxwell. 2005. Global knowledge networks and international development: bridges across boundaries, Routledge.