ArabicWeb16: A New Crawl for Today's Arabic Web

Reem Suwaileh[1], Mucahid Kutlu[1], Nihal Fathima[1], Tamer Elsayed[1], Matthew Lease[2]
[1] Department of Computer Science and Engineering, Qatar University, Qatar
{reem.suwaileh, mucahidkutlu, nihal.fathima, telsayed}@qu.edu.qa
[2] School of Information, University of Texas at Austin, USA
ml@utexas.edu

ABSTRACT

Web crawls provide valuable snapshots of the Web which enable a wide variety of research, be it distributional analysis to characterize Web properties or use of language, content analysis in social science, or Information Retrieval (IR) research to develop and evaluate effective search algorithms. While many English-centric Web crawls exist, existing public Arabic Web crawls are quite limited, limiting research and development. To remedy this, we present *ArabicWeb16*, a new public Web crawl of roughly 150M Arabic Web pages with significant coverage of dialectal Arabic as well as Modern Standard Arabic. For IR researchers, we expect ArabicWeb16 to support various research areas: ad-hoc search, question answering, filtering, cross-dialect search, dialect detection, entity search, blog search, and spam detection.

Keywords

Multi-Dialect; Web Collection; Arabic Retrieval; Ad-hoc Search; Evaluation

1. INTRODUCTION

A central requirement of the Cranfield method for evaluating Information Retrieval (IR) systems is a document collection [7], which is also essential for system development. More specifically, advancing the state-of-the-art in the area of Web search mandates the availability of a large-scale Web collection that is representative of the size and diversity of the Web. A variety of English-centric Web crawls have been previously performed such as VLC2 [11], WT10g [1], Gov2 [4], ClueWeb09 [5], and ClueWeb12[1]. ClueWeb12 intentionally eliminated non-English content.

Approximately 370 million people are estimated to live in the Arab World today, yet the ability of researchers to support this population and their various dialects[2] by advancing the state-of-the-art in Arabic IR (and Arabic Web search in particular) has been greatly restricted by the lack of available Arabic data [9]. Fifteen years ago, the TREC Cross-Language Track created a test collection of Modern Standard Arabic (MSA) news articles from Agence France Press' (AFP) [10]. Great at the time, this test collection has become rather dated and small by today's standards.

The ClueWeb09 Web crawl consists of about 1B Webpages in 10 languages, including 29.2M (2.9%) Arabic pages. This Arabic subset (denoted by *ArClueWeb09*) constitutes the only and largest Arabic Web crawl available for IR research. While this makes it a wonderful resource to researchers, it nevertheless presents several notable limitations.

Most obviously, the contents of this 2009 Web crawl have become somewhat dated in relation to understanding and supporting today's Arabic Web, especially in regard to rapid growth in social media use since then. Secondly, given its coverage of Arabic content was only incidental in crawling, it offers only limited coverage of even the Arabic Web as of 2009. Thirdly, the above two factors conspire to greatly limit its coverage of dialectal Arabic. Finally, our analysis of the ArClueWeb09 subset leads us to estimate that perhaps 14% of its pages are not actually Arabic, with less accurate tools for automatic language detection available in 2009 vs. today. For whatever reason, we know of no analysis having been reported on the content of ArClueWeb09, nor of any IR studies having used it. Taken together, these factors suggest a strong need for an updated crawl of today's Arabic Web that is far more representative of its current state and better supporting research, especially on Arabic IR.

Common Crawl's[3] most recent November 2015 boasts over 1.82 billion URLs, but past analysis suggests that, similar to ClueWeb09, English content dominates the crawl [12]. While Common Crawl could be mined to identify and extract a useful Arabic subset akin to ArClueWeb09, this would address only recency, not coverage.

To address the above concerns, we believe a focused Arabic Web crawl is vital to enable and encourage research on today's Arabic Web. Beyond achieving a massive and modern crawl, we sought to ensure broad coverage of dialectal Arabic. To the best of our knowledge, our new public Web crawl, *ArabicWeb16*, is now the the largest and most representative snapshot of today's Arabic Web. ArabicWeb16 includes about 150M pages crawled over the month of January 2016. To obtain ArabicWeb16, please check the ArabicWeb16 project Web page[4].

[1] lemurproject.org/clueweb12

[2] en.wikipedia.org/wiki/Varieties_of_Arabic

[3] commoncrawl.org

[4] http://qufaculty.qu.edu.qa/telsayed/arabicweb16

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914677

2. CRAWLING PROCESS

Inspired by seed selection of ClueWeb09[5], we constructed our Web crawl using seeds collected from a range of sources (e.g., Wikipedia, Alexa, ArClueWeb09, and Twitter) and selection methods (manual selection and BootCaT [2]) (Section 2.1). We modified Heritrix[6] to prioritize Arabic pages in crawling, and we applied a 3-step language detection pipeline to identify Arabic content (Section 2.2). We also excluded certain types of content during crawling, as well as post-processed data to remove duplicates and further filter out non-Arabic content, among others (Section 2.3).

2.1 Seed Selection

We have collected around 27M seed URLs via several sources and methods:

Wikipedia: Given Wikipedia's high quality, diversity of topics, and authenticity, we downloaded its Arabic pages on October 2nd 2015 and used Wikipedia Extractor[7] to extract approximately 382K URLs of articles.

Manual Selection: We manually harvested around 6.1K popular Arabic websites. These websites are of various categories such as directories, forums, news sources, governmental, academic and question-answering websites. They were obtained from pre-compiled lists on the Web (e.g., Wikipedia) or by issuing Arabic queries against Google.

Alexa: Alexa[8] provides country-specific website rankings based on estimated daily unique visitors per month. We collected the top 500 websites of 12 Arab countries. We obtained roughly 670 seed pages after eliminating duplicates and non-Arabic pages among all countries.

ArClueWeb09: Since ArClueWeb09 contains a relatively large and diverse set of Arabic pages (29M pages), we believed it to be a good addition to our seed list. We performed a cleaning process on this dataset that includes language detection, *Blacklist*[9] URL filtering, and inappropriate-content filtering. From our cleaning process, we found that 2.9M pages were non-Arabic and 95K web pages were either blacklisted or contain inappropriate content. This resulted in 26.1M seed web pages.

Twitter: To support research related to Arabic microblog IR tasks, we collected webpages linked from Twitter. We crawled tweets for one month (Nov. 9th to Dec. 9th 2015) via Twitter's API. We then extracted URLs from Arabic tweets and filtered out URLs of tweets and blacklisted pages. This process resulted in 348K seed pages.

BootCaT [2]: We collected a set of MSA and dialectal queries. The MSA queries contain ArClueWeb09 queries and category names from the DMOZ Open Directory Project[10] and Wikipedia.

In order to collect dialectal queries, we conducted an informal survey among Arab participants. We also used an available list from the *Al-mo3jam* website[11] for each Arab country. In order to avoid biasing data collection toward any one dialect, we randomly selected dialectal words for each country such that the number of selected dialectal words per country is proportional to the estimated size of its internet user population.

We performed an extensive filtering to remove duplicates, non-Arabic queries, queries with more than 5 words, any English word written in Arabic letters, such as ماي سبيس (*MySpace*) and inappropriate terms. The final list of queries contains around 5600 MSA and 1104 dialectal queries. We ran all queries against Google and Bing search engines where we set the language to Arabic and enabled safe mode to eliminate inappropriate content. We retrieved the first 20 results from each search, yielding approximately 24K URLs after eliminating duplicates.

2.2 Language-Focused Crawling

Prioritized Crawling: The relatively small size of ArClueWeb09 (only 3% of ClueWeb09) exemplifies that general Web crawls yield relatively low harvest rate of Arabic pages. Therefore, we modified Heritrix via a method similar to *soft-focusing* [6], decreasing the priority of URLs extracted from non-Arabic pages (as opposed to completely eliminating them). This ensured coverage of Arabic pages that are accessible only through non-Arabic pages. In addition to decreasing priority, our method also increases the cost of pages extracted from non-Arabic pages. If the crawler cannot identify any Arabic pages crawled from a host after a threshold number of attempts, it eliminates that host.

Language Detection: To prioritize the URLs in the crawler's queues, we detect the language of each page once downloaded. In addition to pure Arabic pages, we also considered multilingual pages related to Arabic (*e.g.*, a page with few Arabic sentences) as Arabic pages.

To detect the language of a page, we applied a 3-stage pipeline (from most-to-least trusted stages). For a given Web page, our algorithm first checks the HTML code of the page. If tagged as Arabic, we consider it so. Otherwise, we run *LangDetect* [14], also used in ClueWeb12, on the page. If it is still not detected as Arabic, Persian, nor Urdu (languages using Arabic characters), we perform a character analysis. If the page contains Arabic (but not Persian or Urdu) characters, we consider it to be Arabic.

2.3 Crawler Execution Details

Having observed ClueWeb switched crawlers from Nutch[12] in 2009 to Heritrix in 2012, we conducted pilot evaluations with both and ultimately selected Heritrix due to reliability and ease of use and modification. However, since Heritrix does not support distributed crawling (unlike Nutch), it was necessary to add a post-processing step to remove duplicate pages crawled by different Heritrix instances.

Crawling was performed on an 11-node cluster, each having 24 cores (2.5 GHz) and 128GB of RAM. We dedicated 3 nodes for ArClueWeb09, 3 nodes for Wikipedia, and 1 node for Twitter seeds. Rest of seeds were distributed evenly over the remaining 4 nodes. We used default parameter configuration of Heritrix and employed 25 threads on each node.

We began crawling on January 1st, 2016 and stopped on January 30th, 2016. During crawling, we excluded non-textual multimedia, compressed data, pages over 100MB, and Twitter content. In addition to removing duplicates in post-processing, we also filtered out non-target pages (*e.g.*, non-Arabic pages, DNS servers, robot.txt files) and pages

[5] boston.lti.cs.cmu.edu/Data/web08-bst/planning.html

[6] webarchive.jira.com/wiki/display/Heritrix/Heritrix

[7] medialab.di.unipi.it/wiki/Wikipedia_Extractor

[8] www.alexa.com

[9] urlblacklist.com

[10] dmoz.org/World/Arabic

[11] ar.mo3jam.com

[12] nutch.apache.org

that return 3xx, 4xx or 5xx error codes. Crawled pages are stored in compressed WARC files.

3. ArabicWeb16 DATASET

Table 1 presents statistics characterizing ArabicWeb16 and ArClueWeb09 Web crawls. ArabicWeb16 is seen to be 5x larger than ArClueWeb09 with respect to number of pages (150M vs. 30M Arabic pages), and 11x larger with respect to storage requirements of uncompressed files (11TB vs. 1TB). In terms of overlap, 2016 versions of 1.5M pages from ArClueWeb09 can be found in ArabicWeb16. Pages from ArClueWeb09 missing in ArabicWeb16 may arise from several causes: the pages no longer existing, being eliminated due to not being detected as Arabic or other filtering criteria (see Section 2.3), or simply having not yet been crawled.

Table 1: ArabicWeb16 & ArClueWeb09 Statistics

	ArabicWeb16	ArClueWeb09
Data Size	10.8 TB	0.97 TB
Pages	150.9M	29.2M
Domains	768,516	196,776
Dialectal Pages	31.4M	6.2M
Est. AR Pages	97%	86.1%

To evaluate the accuracy of our automatic language detection pipeline, we checked for false positives (pages misdetected as Arabic) by manually inspecting 1000 random pages in ArabicWeb16. We found that 97% of them are indeed Arabic, showing our method's reliability. We also replicated the same method for analyzing ArClueWeb09. Of the 1000 random pages, we found 86.1% were Arabic.

3.1 Diversity of Domains

Table 1 shows that ArabicWeb16 covers nearly 4x more domains than ArClueWeb09 (768K vs. 197K). We further counted the number of pages per domain to construct the histogram in Figure 1. For each given page count shown on the x-axis (in log-scale), the y-axis shows the number of domains (in log-scale) having that many pages. The distribution is roughly similar between ArabicWeb16 and ArClueWeb09, with the greater domain and page count of ArabicWeb16 vs. ArClueWeb09 being somewhat obscured by the histogram bucketing. It is interesting to note ArabicWeb16's far greater prevalence of domains having only a single page, whereas ArClueWeb09 finding more domains with 10-99 pages. We also analyzed the top 100 domains in ArabicWeb16 and, interestingly, found that more than 50% are forums. Overall, the diversity seen in pages per domain suggests ArabicWeb16 will be useful for researchers interested in working with domains of varying depth.

3.2 Diversity of Dialects

With regard to coverage of dialectal Arabic, we estimate the distribution of dialectal content vs. MSA in our crawl by training a multi-class *Naïve Bayes* classifier using the `Scikit-learn` [13] library. We distinguish MSA from 4 common geographical dialects: Egyptian, Gulf, Levantine and Maghrebi. Using several dialectal Arabic datasets [8, 3], we sampled a balanced dataset for MSA and each dialect, roughly 7k tweets and comments per category. Since our training dataset is not covering all Arabic dialects, we introduced a 6^{th} category, named *others*, and classified pages as *others* if their classification confidence score is < 0.5.

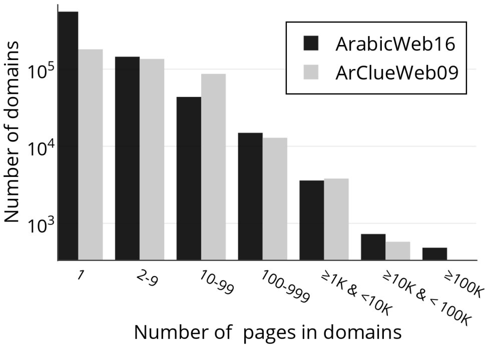

Figure 1: Histogram of pages per domain (log-scale)

We classified the first 150 Arabic words from each page. Table 2 shows the dialectal distribution in both datasets. In ArClueWeb09, we applied our post-processing step (see Section 2.3) and ran the classifier on that filtered dataset.

Table 2: Distribution of MSA and Dialects

Dialect	ArabicWeb16	ArClueWeb09
MSA	119M (79%)	19.9M (76%)
Egyptian	9M (6%)	1.5M (6%)
Gulf	7.6M (5%)	2.7M (10%)
Levantine	7M (5%)	1.5M (6%)
Maghrebi	5M (3%)	0.4M (1.6%)
Others	2.8M (2%)	0.1M (< 1%)

ArabicWeb16 has 5 times more pages with dialects than ArClueWeb09, that is, the ratio of dialectal content is proportional to the ratio of dataset sizes. However, considering the number of internet users of each country[13], the distribution of dialects in ArabicWeb16 is a better representative of dialects. For instance, while Egypt has the highest number of internet users among Arab countries, pages with Egyptian dialect is also more than others in ArabicWeb16, which is not the case in ArClueWeb09.

3.3 Diversity of Page Content

To estimate the proportion of different Web page types in ArabicWeb16, we asked CrowdFlower.com *contributors* to classify pages into the following category schema we defined:

Informational: Web pages whose main purpose is to provide information (*e.g.*, Wikipedia). Information can vary from scientific articles to event schedules.

Discussion & Opinion: Web pages with discussions, opinions, interviews, etc., often on social platforms.

News and Media: Web pages that provide different topics of news and articles from around the world.

Online Services: Web pages for online applications, or platforms for payment and shopping. These Web pages may list services or products to buy or use, user-guides, etc.

Organizational: Institutional Web pages describing owners' interests, activities, news or services, etc.

Entertainment: Web pages with a main purpose to provide entertainment to users (*e.g.*, games, movies).

Other: Web pages not fitting any of the above types.

We randomly sampled 1000 pages from ArabicWeb16 and ArClueWeb09. Broken links and non-Arabic web pages were

[13] internetlivestats.com/internet-users/

identified and excluded from the task in both datasets. We limited the job to Arabic-speakers and moderately-rated contributors who passed test questions with ≥ 80% accuracy.

We requested 3 judgments per page. The overall *Fleiss's Kappa* inter-annotator agreement values for ArabicWeb16 and ArClueWeb09 are 0.57 (moderate agreement) and 0.37 (fair agreement), respectively.

Table 3 presents the number of pages receiving ≥ 2 agreeing labels for each page type. The number of annotated pages in ArClueWeb09 is 5x smaller (as its size). This is mainly because of higher percentage of broken links and non-Arabic pages. We notice that ArabicWeb16 has more pages annotated as Discussion & Opinion. This can be due to the increasing popularity of social media platforms (*e.g.*, tumblr) in recent years. Overall, ArabicWeb16 appears to provide better diversity in terms of page content.

Table 3: Distribution of Different Web page Types

Web page Type	ArabicWeb16	ArClueWeb09
Informational	113 (12.80%)	23 (12.92%)
Discuss. & Opinion	295 (33.41%)	37 (20.79%)
News and Media	93 (10.53%)	9 (5.06%)
Online Services	36 (4.08%)	7 (3.93%)
Organizational	8 (0.91%)	2 (1.12%)
Entertainment	28 (3.17%)	4 (2.25%)
Other	310 (35.11%)	96 (53.93%)

4. ENABLING NEW RESEARCH

A primary goal in constructing ArabicWeb16 is to enable further research in Arabic IR by providing a sound dataset that supports various IR tasks. We envision ArabicWeb16 can be employed for research related to (at least) the following areas: ad-hoc web search, question answering, filtering, cross-dialect search, dialect detection, spam detection, entity-oriented search, and blog track tasks.

To further elaborate, ArabicWeb16 contains many forums, including question-answering sites given as seeds, as well as many informational pages such as Wikipedia, which can usefully support question answering research. In addition, the large dialectal content clearly supports cross-dialect search. While we filtered blacklisted pages in selection of seeds, we intentionally did not filter out spams. Leaving spam present in ArabicWeb16 makes it useful for (Arabic) spam detection research. Finally, ArabicWeb16 contains approximately 19M blog pages (as determined by checking domain names), clearly providing significant content for blog search research.

5. CONCLUSION AND FUTURE WORK

The ever-increasing scale of the Web, shifting patterns of user search and information-sharing behaviors, and emergency of new types of content (e.g., blogs and tweets) creates an ever-growing need to continuously adapt and refine IR methods. Progress in Arabic IR has been impaired in recent years vs. other languages due to lack of suitable data supporting research. Creating a vast Arabic IR collection will therefore create new opportunities and pave way for more further advancements and enhancements in Arabic IR.

Our ongoing work includes construction of search topics and collection of corresponding relevance judgments in order to provide researchers not just with documents to search, but a complete test collection for IR experimentation.

Acknowledgments

This work was made possible by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. We thank the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for computing resources enabling this research.

6. REFERENCES

[1] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing & Management*, 39(6):853–871, 2003.

[2] M. Baroni and S. Bernardini. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the Language Resources and Evaluation Conf. (LREC)*, 2004.

[3] H. Bouamor, N. Habash, and K. Oflazer. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1240–1245, 2014.

[4] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information retrieval*, 10(6):491–508, 2007.

[5] J. Callan, M. Hoy, C. Yoo, and L. Zhao. The ClueWeb09 Dataset, 2009. Presentation Nov. 19, 2009 at NIST TREC. Slides online at boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf.

[6] S. Chakrabarti, M. Van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11):1623–1640, 1999.

[7] C. W. Cleverdon. The evaluation of systems used in information retrieval. In *Proceedings of the international conference on scientific information*, volume 1, pages 687–698. National Academy of Sciences, 1959.

[8] R. Cotterell and C. Callison-Burch. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 241–245, 2014.

[9] K. Darwish and W. Magdy. Arabic Information Retrieval. *Foundations and Trends in Information Retrieval*, 7(4):239–342, 2014.

[10] F. C. Gey and D. W. Oard. The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries. In *Proc. of the Tenth Text REtrieval Conference (TREC 10)*, 2001.

[11] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the TREC-8 Web Track. In *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*, 1999.

[12] V. Kolias, I. Anagnostopoulos, and E. Kayafas. Exploratory Analysis of a Terabyte Scale Web Corpus. *arXiv preprint arXiv:1409.5443*, 2014.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] N. Shuyo. Language detection library for java, 2010. http://code.google.com/p/language-detection/.