A Multi-Threaded Algorithm for Mining the Arabic Web Structure

Mohammad A.R. Abdeen[1], Sami Albouq[2]
[1,2]Islamic University of Madinah, Madinah, Saudi Arabia

Abstract:

The world wide web has become the default destination of acquiring information in just about any knowledge domain. This range of knowledge covers a wide spectrum; from the latest technological advances in medical procedures to most effective drugs and most popular sports cars and personal accessories. The way the web is structured embodies implicit information within it that turns to be useful in information retrieval and in search engine applications. This paper presents an initial attempt at mining the structure of the Arabic portion of the web for the purpose to produce better results of search engines in the Arabic web space. We are presenting a multi-threaded algorithm that provides an initial attempt to reveal the structure of the Arabic web. The algorithm crawls the web and collects interlink information on as many as one million Arabic websites. The algorithm provides an initial analysis of the most "cited" websites and presents a list of the top 20 websites. We have used the website citation count as the measure for ranking a website. It is worth noting that the top ranked websites do not belong to specific categories such as sports, news, lifestyle, or others. Rather the websites on the top of the list represent more of directory websites. Some news websites are included in the top 20 list despite not at the very top of the list. Other categories are represented such as the technology category although represented with only one website at the bottom of the list.

Keywords: Web Structure Mining, Web Graph, Multi-threading, Web Crawling, Arabic Text Processing.

1. Introduction

The 21st century has deservedly gained the title of being the "information age". According to Statista, a leading company in providing market and consumer data, the current size of the world datasphere is about 10 Zetta Byte (Statista GmbH, 2020). The same source predicts that the size of the world datasphere is expected to grow at a size of 50 Zetta Bytes. That is about five folds in five years. This amount of massive data could render useless unless appropriate data mining techniques are used for the purpose of extracting the useful information.

A prime source of this information is the world wide web. According to (Statista GmbH, 2020), the number of websites worldwide are estimated at a staggering number of 1.7 Billion websites. Search engines such as Google, Yahoo, Bing, and many others are the most popular tool for searching the WWW and finding information of interest including food recopies, weather forecasts, status of the current pandemic, to world political news. According to this massive competition of thousands and sometimes hundreds of thousands of websites providing similar service, it became imperative to provide some measure of credibility through the ranking of those competing websites. Many techniques exist that provide such ranking based on the concept of "popularity". In other words, if many websites refer to a specific website, that latter acquire higher credibility than others. One of the known algorithms is the PageRank algorithm (Lawrence Page, et al., 1999).

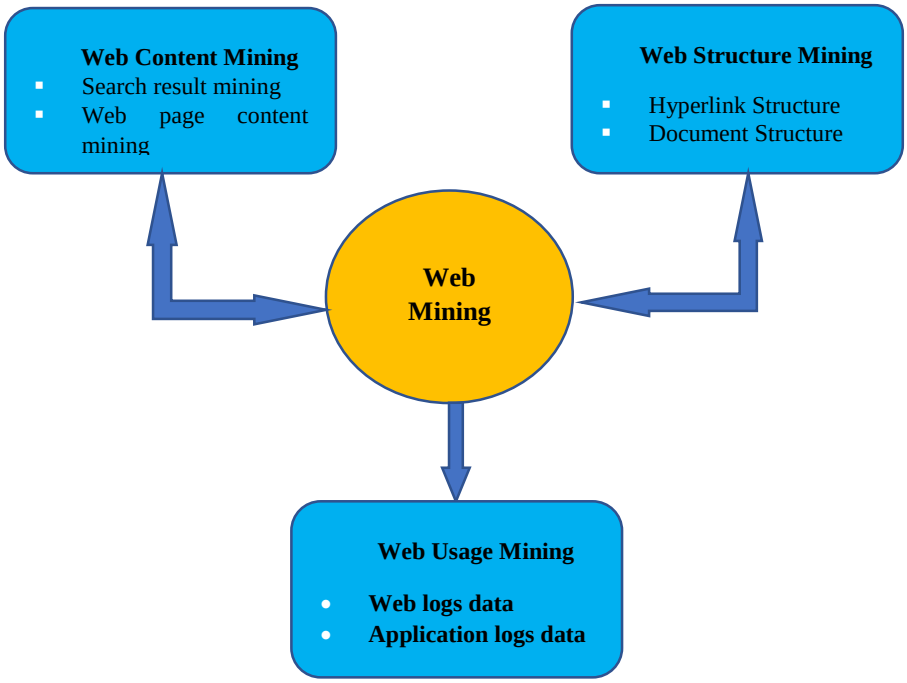
Web mining is area of research that is meant to explore the massive web contents. The web itself is of a graph-like structure. Webpages are considered as nodes in that graph while the links inside each webpage are considered the links of that graph. The area of the web mining deals with three subareas. Those include the web structure mining, web content mining, and web usage mining (Tyagi & Gupta, 2018). This research is meant with the first category, viz. the web structure mining of the Arabic web.

This paper is organized as follows: section 2 below gives the required background of the subject and the previous work that discusses the web structure mining in languages other than Arabic. Section 3 gives a detailed description of the algorithm used in this work as well as some implementation highlights. Section 4 presents the results and provides a discussion and reflections on those results. Sections 5 and 6 are the future work and conclusions.

2. Background

Web mining is an area that came about from the area of Data Mining. Data mining is concerned with extracting useful information and discovering knowledge from a massive amount of data whether structured or unstructured. Mining the web has been addressed in many of the available literature. It has three types; mining the usage of the web (Borges & Levene, 1999), mining the contents of the web (Panda, et al, 2016), and mining the structure of the web. **Error! Reference source not found.** below depicts the three types of web mining with their corresponding applications and subtypes. The following sections provides a background on those techniques.

Figure 1: The three types of Web Mining



2.1. Web Content mining

The webpages that exist in on the world Websphere contains various types of information. This includes text, image, audio, or video contents. The area of web content mining is concerned with extracting meaningful information from the variety of content that exist in a given webpage. There has been numerous research work in the area of text mining in the application of document classification, clustering and categorization. These applications have been used for various languages including English (Jindal et. al. , 2015), Chinese (Luo et. al. , 2011), Russian (Sboev et. al., 2016), and Arabic (Abdeen et. al., 2019). These applications are used by various websites such as news website, to categorize various articles in categories such as sports, politics, technology and many others. Image mining is dealt with various techniques of image processing and computer vision. The main steps of the image mining are to preprocess the image and then the feature

extraction process that facilitates the process of knowledge. There exist various applications of image mining such as image retrieval, matching and object and pattern recognition (Nagi, et. al 2018). Web content mining deals with structured, unstructured, or semi-structured data. Various methods and algorithms exist that address the issue of web content mining including the Data Update Propagation (DUP), the Document Object Model (DOM), Association Rules, and other methods. A good survey of those methods exists in (Samuel, et. al, 2019) .

2.2. Web structure mining

Web structure mining is to perform link analysis of the web structure and extract information hidden within this complex yet rich structure. We structure mining starts with collecting information of the available URL set of interest and the sublinks that exist within those URLs. With this information, a graph like structure can be built and graph analysis algorithms can by employed. Several algorithms exist for ranking the websites that exist on the world Websphere. These algorithms include the popular PageRank algorithm used by Google (Lawrence Page, et al., 1999). Other algorithms exist such as the Hyperlink Induced Topic Search (HITS) in which a webpage is classified as either a hub of information (they give many links to other pages) or an authority webpage (they present an authority of information as many other pages link to them) (Kleinberg, 1998). In AHPA algorithm, another function was introduced utilizing two vectors A and H. Vector A

represents the classification of pages from set of webpages while vector H represents classification of the same set of pages as a hub (Stephanides, et.al, 2006).

2.3. Web usage mining

Web usage mining includes mining the data available in the form of server logs that is able to extract user patterns of accessing specific webpages, the previous webpage they came from, the clicking of specific items, and any recurrence of specific user browsing sequence. Massive amount of log data files is generated every day in webservers and mining of this data can be useful in developing marketing strategies, promotions, and possibly a better restructuring of the website for better user interaction (Rita et. al. 2019).

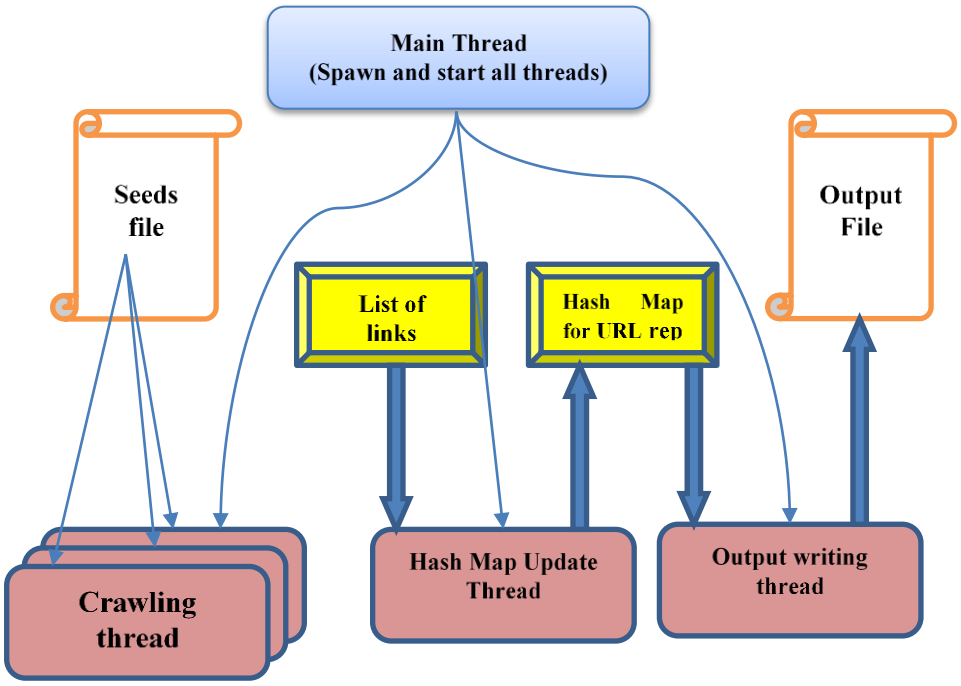
3. The Arabic Web Structure Mining

In this section we discuss the methodology and algorithm used for mining the Arabic web content. By Arabic web we mean those pages whose primary language is Arabic.

3.1 The Algorithm

The main components of the algorithm presented in this work are shown in Figure 2 below.

Figure 2: A schematic showing the components of the algorithm



The figure shows the following components:

1. The main thread: This function of this thread is to mainly create necessary global data structures and input and output files handles. It also creates thread groups for crawling the web and accessing the global data structures (L1 and M1).
2. The web crawling thread group: this thread pool is created with the size equals the initial seed URLs (Uniform Resource Locator). The purpose is to have each seed handled by one thread to improve throughput and to employ the multicore hardware we are targeting. Each thread in this thread pool fetches the URL from the seed file, checks if its content is Arabic and if so, fetches the sublinks included in this website to serve as future seeds.
3. The list L1 management thread.
4. The map M1 management thread.

Error! Reference source not found. below shows the algorithm developed in this work to mine the structure of the Arabic Web.

Figure 3: The Multi-Threaded Arabic Web Structure Mining Algorithm

Procedure: **RepCalThr** synchronized run( )

1. While(true)
2. Read url from List **L1**
   3. If url in Map **M1**
      a. Increment url repetition value
      b. Remove url from List **L1**
   4. else
      a. add url to Map **M1**
      b. initialize url repetition to **1**
   5. End if
6. End while

Procedure: **fileIOThr** synchronized run( )

1. While(true)
   a. Thread sleep(60 seconds)
   b. Foreach element in Map M1
   c. If repetition < 100
      Store url and rep in file_l
   d. else if rep >= 100 < 1000
      store url and rep in file_m
   e. else store url and rep in file_h
2. End while

Procedure: Main

1. Create output files
2. Initialize input/output file handles
3. Initialize global data structures
   urlList **L1** and Map **M1**
4. Initialize Thread pools
   CrawlThr, LinkT, MapThr, OutThr
5. Read Input Seed file
6. Assign a thread for every seed URL
7. Run all threads

Procedure: **CrawlThr** synchronized run( )

1. While(true)
   a. Get next URL in L1
   b. open a connection to the webpage(url)
   c. foreach sublink in webpage
      Check if Arabic content, if so continue
      Get all sublinks
      Store subink in L1
   d. End foreach
2. End while

3.2 Initial seed URLs

The initial seed URLs are obtained by doing an initial manual search for Arabic websites that publish various classes of Arabic websites (similar to a directory pages). The used search term in this case is:

دليل المواقع العربية

Which is translated as "A guide to Arabic websites".

The top 10 list of websites as per the performed Google search of those "directory" websites is given in **Error! Reference source not found.** below:

Table 1: The seed websites used in the algorithm as per Google search results

URL	Rank
https://www.raddadi.com/	1
http://dalil.info/	2
https://eyoon.com/	3
https://www.hvips.com	4
https://www.202020.net/	5
http://bawaba.khayma.com	6
http://www.sultan.org/a/	7
http://dir.m5zn.com/	8
https://www.dm0z.com/	9
https://www.alrwabt.com/	10

It is to be noted that the list of URLs and the map M1 data structures are shared memory ones and they need to protected by locking mechanisms (such as semaphores or mutex). We have used in this implementation the Java synchronization mechanism that allows to design methods that access shared variables without the need of the developer to manage the locking mechanisms. This approach is simpler and guarantees mutual exclusion.

4. Results

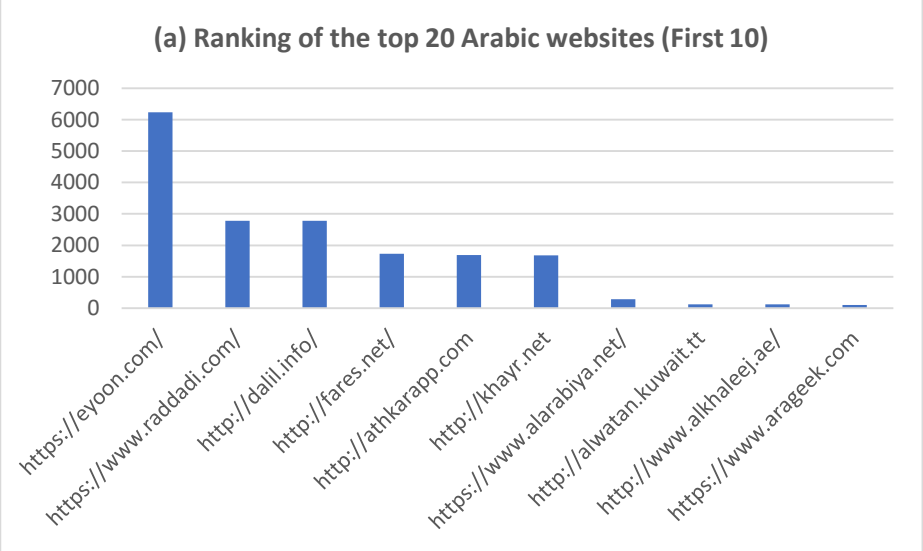
The algorithm presented in this work is implemented with Java programming language and run on a multi-core HP ProLiant server with dual processor and 24 cores. The algorithm utilized the Java

List and Hash Map data structures. It is to be noted that these data structures are shared memory ones and access to them needs to be synchronized in order to avoid corrupting their contents as they present a critical section.

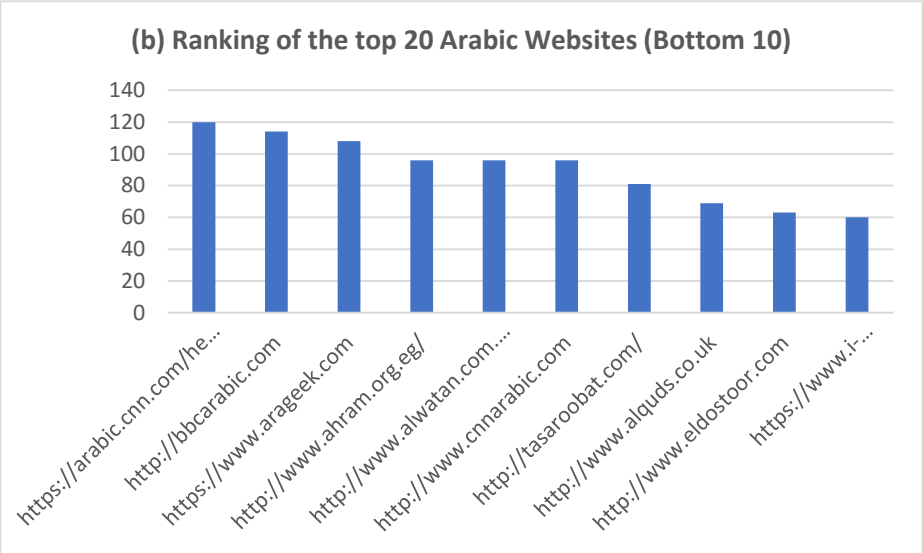
We have collected over one million links as a result of crawling the internet using this algorithm. We classified the output result according to the citation of each URL. The Hash Map stores the URL as its key and the frequency (of occurrence in other websites) as the corresponding value of the key. The map M1 is stored in a file and is sorted in a descending order. Figure 4 below shows the top 20 most cited web sites. It is to be noted that the top ranked websites are those who do not have a specific category e.g. news and sports websites, but rather generic and directory-like websites (eyoon.com and raddadi.com). Some other websites of religious nature such as athkarapp.com and khayer.net, made it to the top of the list of most frequently referred websites. News websites come in the lower half of the list such as eldostoor.com and alwatan.com.sa. An interesting result of the experiments is that Arabic version of foreign news website such as cnnarabic.com came ahead of several credible and established Arabic newspapers such as ahram.org.eg.

It is also interesting to see one category addressing technology despite being in the bottom of the list (i-electrician.com).

Figure 1: Graphs (a) and (b) showing the top 20 most referred Arabic Websites according to their frequency



(a)



(b)

5. Conclusions and Future Work

In this work we have presented a multi-threaded algorithm to mine the structure of the Arabic web. The algorithm is implemented using Java programming language and its data structures. The algorithm was tested on an HP ProLiant Server with dual-processes and 24 cores. We have collected information on over one million Arabic websites by crawling the web using the developed multi-threaded algorithm. The results presented herein showed the top 20 websites according to the number-of-referrals (aka citation) measure. The results showed that the top-ranking websites do not belong to specific website categories such as sport, news, culture, or travel. They rather belong to a directory like websites with presented articles/stories of no specific category (e.g. raddadi.com). Some of the top ranked websites belong to the category of religion which reflects one nature of the Arabic region. News websites came in the bottom half of the list with Arabic release of foreign news websites (such as CNN or BBC) preceding some of the established and traditional Arabic news websites. The "technology" category is represented by one website at the bottom of the list.

It is to be noted that this study is preliminary despite giving fair representation of the facts on the ground. As a future work, we will extend the experimentation to browse as much as 10 Million Arabic websites and provide the top ranking of those browsed. This will give a broader and more precise representation of the structure of the Arabic web.

Acknowledgment

This research is funded by the Islamic University of Madinah, Tamayoz initiative grant 23/40.

References

[1] Statista GmbH, (July 2020). Annual size of real time data in the global datasphere, [Online]. Available: https://www.statista.com/statistics/949144/worldwide-global-datasphere-real-time-data-annual-size/

[2] Lawrence Page, et al. (1999) The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*,

[3] Tyagi N., Gupta S.K., (2018) "Web structure mining algorithms: A survey" in Big Data Analytics, *Springer*, pp. 305-317.

[4] Borges J. and Levene M. (1999) "Data mining of user navigation patterns". In: *Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling*, pp. 31–36.

[5] Panda B., et al. (2016) "A comparative study on serial and parallel web content mining." *International Journal of Advanced Networking and Applications* 7.5: 2882.

[6] Jindal R., Malhotra R., and Jain A., (2015) "Techniques for text classification: Literature review and current trends." *Webology*, vol. 12, no. 2.

[7] Luo X., Ohyama W., Wakabayashi T., and Kimura F., (2011) "A study on automatic Chinese text classification," in 2011 *International Conference on Document Analysis and Recognition. IEEE*, pp. 920–924.

[8] Sboev, A., Litvinova, T., Gudovskikh, D., Rybka, R., & Moloshnikov, I. (2016). Machine learning models of text categorization by author gender using topic-independent features. Procedia Computer Science, 101, 135-142.

[9] Abdeen, M. A. R., et. al, (2019) "A closer look at Arabic Text Classification" *International Journal of Advanced Computer Science and Applications* (IJACSA), 10(11).

[10] Nagi R. S., et. al. (2018) "A Review and Comparative Analysis on Image Mining Techniques." *Image Segmentation*: 51.

[11] Samuel, M. O. et. Al (2019) "A Systematic Review of Current Trends in Web Content Mining." *Journal of Physics: Conference Series*. Vol. 1299. No. 1. IOP Publishing.

[12] Roy, R., & Giduturi, A. (2019). Survey on Pre-Processing Web Log Files in Web Usage Mining (No. 1927). EasyChair.

[13] Sardhara R. and lakhataria K. L., (2018) "Web Structure Mining: A Novel Approach to Reduce Mutual Reinforcement," 2018 3rd *International Conference and Workshops on Recent Advances and Innovations in Engineering* (ICRAIE), Jaipur, India, pp. 1-6.

[14] Kleinberg, J. (1998) "Authoritative sources in a hyperlinked environment", *In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 668-677.

[15] Stephanides, G., Cosulschi, M., Gabroveanu, M., & Constantinescu, N. (2006, April). AHPA-calculating hub and authority for information retrieval. In 22nd International Conference on Data Engineering Workshops (ICDEW'06) (pp. 36-36). IEEE.