Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing

Nilufar Abdurakhmonova
Uzbek linguistics department
National University of Uzbekistan
Tashkent, Uzbekistan
n.abduraxmonova@nuu.uz

Ismailov Alisher
Innovative Educational department
Andijan machine building institute
Andijan, Uzbekistan
alisherismailov1991@gmail.com

Guli Toirova
Uzbek linguistics department
Bukhara State University
Bukhara Uzbekistan
tugulijon@mail.ru

***Abstract*— over the past decade, the amount of information on the internet has increased. A large amount of unstructured data, referred to as big data on the web, has been created. Finding and extracting data on the internet is called information retrieval. In the search for information, there are web crawler tools, which are a program that scans information on the internet and downloads web documents automatically. Search robot applications can be used in various fields, such as news, finance, medicine, etc. In this article, we will discuss the basic principle and characteristics of search engines as an example to build parallel corpora, as well as the classification of modern popular crawlers, strategies and current applications of crawlers. Finally, we will end this article with a discussion of future directions for research on crawlers.**

Keywords*— WWW, Web Crawler, Crawling techniques, Information Retrieval, Search engine

I. Introduction

There are a number technologies developed in various interdisciplinary fields Computational linguistics, language technology and NLP. There is no doubt that language is more important for all spheres to get ultimate data. In globalization epoch linguistic resources are increasing drastically in digital format by the internet.

At present Computational linguistics is more significant area to develop Uzbek in the different scope of community. This fear was shared by a number of scholars' works. To tackle linguistic problems in this field we need various scientific approaches and computer methods. However, Uzbek is considered lack of resource language in the world, today more investigations draw upon more attention to develop studies in Computational linguistics.

We with our group created Uzbek corpus and it is available now in this site http://uzbekcorpus.uz/ . There are 1.5 million words in different chronological literal styles, official, scientific and publicly styles written texts. Corpus included Latin and Cyrillic texts due to applied both graphemes are available to use even official and education process. It also comprises parallel corpora (English-Uzbek, Turkish-Uzbek, Russian-Uzbek and Korean-Uzbek), learner corpus, authorship corpus. In spite of being many resources in the corpus (terminology, dictionaries, thesaurus etc.) it is necessary to apply computer tools as possible to fast our procedure and to obtain sufficient results in NLP. Our article is focused on general outlook of applying tool parallel corpora using crawler tools for creating machine translation database.

We introduce first of all insight overlooking related works in Section 2 then described special peculiarities web crawlers in Section 3. To clarify distinguishes between crawlers given brief information types of the in Section 4. Finally, we give conclusion to summarize all outlines in our article in Section 5.

II. RELATED WORKS

A number of scientists consider that one of the merits of contemporary computer technologies to obtain any kind knowledge is information retrieval system. Today it can be said that all kinds data are available to look up and get information by tools. Recently they are used as with name as *web crawler, web spider or web robot.* The internet as a source for the database plays significant role in all sphere to catch information as possible format (pdf, doc, html etc.). However, this data cannot be used locally for access because all data in cloud databases. It is as a software via using hyperlinks for collecting data, in this case webpages are used as source in order to compile data. This tool searches key words vertically which connecting by indexes. Today regarding to accuracy and fastness of these search systems have become more fad in natural language processing [1]. There are a number of scientists who subscribe to the view that there is the search engines can automatically navigate through information on the internet, and the search engine is inseparable from the crawler. The most important role of a web crawler is to crawl the internet's big data, find effective information, and store the required information in a local database. Web crawlers mainly contain downloaders, extractors, schedulers, and crawl queues. The structure of the web crawlers are a) Scheduler will provide download URL; b) The downloader then fetches the information from the internet and sends the information to the extractor.; c) extractor strategy according to instruction from extracting information to getting information and next level in URL; d) then the next level URL into the waiting queue, the waiting queue to proceed to send the URL of the heavy filtering and sorting operation to the list, after waiting for the scheduler calls [2-3].

A number programming languages works with web crawler frameworks. For site example JAVA for Apache Nutch, webmagic, Heritriz3, WebCollector, craw1er4j, Spiderman, SeimiCrawler, jsoup-Gecco, and htmlunit and Python for Scrapy, pyspider, Newspaper, and Crawley. Moreover, cola, Portia, python selenium, QueryList, phpspider, and PHPCrawl for PHP.
We would like note that Yanfei Qi draw a comparison between aforementioned GitHub data of web crawler frameworks on conducting research to build online English corpus by web crawler [4].

III. ATRIBUTES OF WEB CRAWLER

According to [5] web crawler has a number advantages in many aspects. It should be noted that its advantages make ease the work of machine and human interface. The following points support this standpoint [6]:

1) Synchronization of the tool within machines in distributed environment;

2) It has specific point that it does not disaffect it speediness of data according to capability to add more machine or networks;

3) Tool's effectiveness is availabilty to download files having accessed a site;

4) User has chance to except or include according to demand. So, it is available to high recourse how much as possible;

5) Tool uses frequency of web pages providing update resources according to URL changing information what it is loaded recently;

6) Extensible crawlers made specific composition for modification protocol of data and its forma as.

IV. CLASSIFICATION OF WEB CRAWLERS

Universal crawler applies a lot of web pages from different source in the internet. Free download access web pages are important to search and save information in one disk. Google PageRank is tool as universal crawler. But it is time consuming and labor consuming work [7]. Another one is focus or thematic crawler which is directed only exact web pages to expand information fast [8]. Additionally, incremental web crawler is more focused to approach searching files based on the latest results which existing one as data texts. It retrieves expired versions of web pages and replaced new files. One of the big advantages of this type of tool is that it improves crawler productivity and minimalize physical memory usage for data. Moreover, each time before crawling, it analyzes before existing the same data. Distributed or nodes crawler run on each computer. It organizes work system between each computer. Distributed web crawlers have three types of modes. First is *master-slave mode* which is hosted in one machine and it controls set of computers. So, head machine as "master" is responsible for managing URLs, joining together machines by distributed tasks. Other computers as "slaves" do tasks and pass report results to master machine and none of the machines communicate each other. Second one is *offline mode.* There is no host machine, each machine communicates each other by either one directed transmission like ring structure. Second type is each machine communicate each other. Third type is hybrid mode that a combination abovementioned two modes. In mixed mode, the host is responsible for assigning tasks to other machines, but other slave machines can also communicate with each other and also have task assignment functions. Another one type is *parallel distributed crawler*. In order to reduce time and large search results it is communicated with local and broadly networks of global system. *IoT (internet of things) Web Crawler* as a tool comprises two types of public IoT and ad hoc IoT.

The Shodan search engine [10] collects data from the following ports: HTTP, FTP.SSH. Telnet. LUTS. LUTS. A SIP. RTSP. Shodan's search engine crawlers collect metadata about devices, while Google's crawlers collect data for websites. Web crawler functions have procedure as following three phases: Webpage acquisition and analysis, how to apply data storage and web search strategy. A common function of a web crawler is to mimic a browser to make an HTTP request. The crawler then contacts the web server via an HTTP request. After receiving a response from the server, the crawler analyzes and saves the web page. Web page scraping is the process of denoising a web page. All types of data are stored on the Internet. Web page denoising means extracting text content from web pages. The next phase is data storage. It is saved either local or data according to the seize of information.

V. USAGE WEB CRAWLER TO BUILD CORPORA OF THE TEXTS

It is worth bearing in mind that national domains (webpages) are source as document of files to build corpora for natural language processing by web crawlers. Specialty of web crawler for corpus building it has opportunity to compile texts exact key words according terminology. Today there are different types of Corpus builder to use texts as material for natural language processing. One of like tools is BootCaT created by Baroni and Bernardini (2004) was experienced by building corpora for English and Italian languages. The metadata is used for determining the language of the webpages. Tamura et el. (2005) also used TextCat tool identify the language in UTF-8 in Japan language.

One is of the tools is Heritrix which this web crawling tool collects national libraries as one database.

In [9] focused to build parallel corpora by using hybrid crawling architecture. This tool contains scheduler, multithread downloader, Parser and Extractor, and Classifier. According to this study, compared three domains (sport, legislation, and general news) of available in the sites. The statistical indicator showed high rate hybrid crawling architecture than parallel crawling and focused crawling.

The bilingual corpora for Uzbek language is crucial task today to increase quality of machine translation by hybrid parallel crawler. In this case to use parallel translated texts which have been already done by human is more applicable in different sites especially in Russian and English. Efficiently using crawlers in this purpose we need translation dictionaries and each query is indexed by based on parallel texts which segmented before.

CONCLUSION

To date, scientists have conducted many studies [12, 13, 14, 15, 16] on the web crawler; however, there is still a need for crawler performance research. One of the main problems with web crawlers is fixed search approaches. If a web crawler needs to organize the crawling of a web page among different websites on the Internet, fixed search engines cannot efficiently crawl the data. Web crawl performance improvements are yet to be explored. Another barrier to web crawling is some limitations for a more detailed crawl topic, so improving the choice of topic keywords from a semantic point of view has become a trending research topic for crawl technology in the future.

REFERENCES

[1] Pan, X. Y., et al. "Survey on research of themed crawling technique." Application Research of Computers 37.5 (2019).

[2] Rungsawang, Arnon, and Niran Angkawattanawit. "Learnable topic-specific web crawler." Journal of Network and Computer Applications 28.2 (2005): 97-114.

[3] Kozanidis, Lefteris. "An ontology-based focused crawler." International Conference on Application of Natural Language to Information Systems. Springer, Berlin, Heidelberg, 2008.

[4] Yanfei, Q. “Construction of Online English Corpus Based on Web Crawler Technology” Hindawi Wireless Communications and Mobile Computing Volume 2022, Article ID 7589727, 8 pages https://doi.org/10.1155/2022/7589727.

[5] Pavai, G., and T. V. Geetha. "Improving the freshness of the search engines by a probabilistic approach based incremental crawler." Information Systems Frontiers 19.5 (2017): 1013-1028.

[6] Deka, Ganesh Chandra. "NoSQL Web Crawler Application." Advances in Computers. Vol. 109. Elsevier, 2018. 77-100.

[7] Yu, Linxuan, et al. "Summary of web crawler technology research." Journal of Physics: Conference Series. Vol. 1449. No. 1. IOP Publishing, 2020.

[8] Sun, L. W., G. H. He, and L. F. Wu. "Research on web crawler technology." Computer knowledge and technology 6.15 (2010): 4112-4115.

[9] Sai, M., Lap, Man Hoi, Su-Kit Tang, Rita Tse “Crawling Parallel Data for Bilingual Corpus Using Hybrid Crawling Architecture” Procedia Computer Science, Volume 198, 2022, 122-127.

[10] Boldi, Paolo, et al. "Ubicrawler: A scalable fully distributed web crawler." Software: Practice and Experience 34.8 (2004): 711-726.

[11] Cho, Junghoo, Hector Garcia-Molina, and Lawrence Page. "Efficient crawling through URL ordering." Computer networks and ISDN systems 30.1-7 (1998): 161-172.

[12] Ismailov, A., Jalil, M. A., Abdullah, Z., & Abd Rahim, N. H. (2016, August). A comparative study of stemming algorithms for use with the Uzbek language. In 2016 3rd International conference on computer and information sciences (ICCOINS) (pp. 7-12). IEEE.

[13] Jalil, M. M., Ismailov, A., Abd Rahim, N. H., & Abdullah, Z. (2017). The Development of the Uzbek Stemming Algorithm. Advanced Science Letters, 23(5), 4171-4174.

[14] D. Mengliev, V. Barakhnin, and N. Abdurakhmonova, “Development of Intellectual Web System for Morph Analyzing of Uzbek Words,” Applied Sciences, vol. 11, no. 19, p. 9117, Sep. 2021, doi: 10.3390/app11199117.

[15] N. Abdurakhmonova, U. Tuliyev and A. Gatiatullin, "Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz," 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670043.

[16] N. Abdurakhmonova, U. Tuliyev and A. Gatiatullin, "Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz," 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670043.

[17] Agostini, Alessandro, Timur Ravilevich Usmanov, Ulugbek Khamdamov, Nilufar Abdurakhmonova and Mukhammadsaid Mamasaidov. “UZWORDNET: A Lexical-Semantic Database for the Uzbek Language.” GWC (2021).

[18] D. Sulevmanov, A. Gatiatullin, N. Prokopyev and N. Abdurakhmonova, "Turkic Morpheme Web Portal as a Platform for Turkology Research," 2020 International Conference on Information Science and Communications Technologies (ICISCT), 2020, pp. 1-5, doi: 10.1109/ICISCT50599.2020.9351500.

[19] Khusainov, Aidar, Dzhavdet Suleymanov, Rinat Gilmullin, Alina Minsafina, Lenara Kubedinova and Nilufar Abdurakhmonova. “First Results of the TurkLang-7 Project: Creating Russian-Turkic Parallel Corpora and MT Systems.” CMCL (2020).

[20] Abdurakhmonova N, Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. Literature, Education. 2018(3):68.

[21] Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.

[22] Abdurakhmonova N. Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation. Journal of Social Sciences and Humanities Research.2017;5(03):89-100.