Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 122–127

Procedia Computer Science

www.elsevier.com/locate/procedia

The 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2021)
November 1-4, 2021, Leuven, Belgium

Crawling Parallel Data for Bilingual Corpus Using Hybrid Crawling Architecture

Sai Man Cheok[a,b], Lap Man Hoi[a,b], Su-Kit Tang[a,b,*], Rita Tse[a,b]

[a] School of Applied Sciences, Macao Polytechnic Institute, Macao SAR, China
[b] Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence of Ministry of Education, Macao Polytechnic Institute, Macao SAR, China

Abstract

The quality of translation work mainly depends on the understanding of the words in their domain. If machine translation can accurately translate the words in a domain in different languages, it can even avoid any human communication error. To achieve this, a high-quality bilingual corpus is crucial as they are always the basis of state-of-the-art machine translation system. However, it is complicated to construct the corpus with large amount of parallel data. In this paper, a new crawling architecture, called Hybrid Crawling Architecture (HCA), will be proposed, which efficiently and effectively collects parallel data from the Web for the bilingual corpus. HCA aims at targeted websites, which contains articles in at least two different languages. As it is a mixture of Focused crawling architecture and Parallel crawling architecture, HCA takes advantages over both architectures. In intensive experiments on crawling parallel data of relevance topics, HCA significantly outperforms Focused crawling architecture and Parallel crawling architecture for 30% and 200% respectively, in terms of quantity.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the Conference Program Chairs

Keywords: Parallel Crawler, Focused Crawler, Bilingual Corpus.

* Corresponding author.
E-mail address: sktang@ipm.edu.mo

1877-0509 © 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the Conference Program Chairs
10.1016/j.procs.2021.12.218

1. Introduction

Recently, machine translation technology is commonly used in Natural Language Processing (NLP). It uses computers to effectively translate natural languages with quality and efficiency [1][2]. Natural language translation is domain-based. The meaning of texts in different domains varies. The quality of translation work mainly depends on the understanding of the words in their domain. If machine translation can accurately translate the words in a domain in different languages, it can even avoid any human communication error. To achieve this, a high-quality bilingual corpus is crucial as they are always the basis of state-of-the-art machine translation system [3][4].

In this paper, a new crawling architecture, called Hybrid Crawling Architecture (HCA), will be proposed, which aims at crawling parallel data from targeted websites, which contain articles in at least two different languages. HCA is designed to efficiently collect parallel data of relevant topics from the Web. Based on the URL links and content rules configured, HCA traverses websites for bilingual content. URL links discovered in webpages are distributed to multi-downloader under coordination. If the topic relevance of parallel data is verified successfully by a scheduler, the data will be downloaded.

As it is a combination of Focused crawling architecture and Parallel crawling architecture, HCA takes advantages over both architectures. In intensive experiments on crawling parallel data of relevance topics, HCA outperforms Focused crawling architecture for 30% in terms of quantity. Significantly, it is 200% more than that by Parallel crawling architecture.

In summary, after giving the motivation in this chapter, the background and related works of existing crawling architecture will be introduced in Section 2. As HCA is specially designed for building the bilingual corpus, the design and implementation of its data crawling method will be explained in Section 3. In Section 4, the performance of HCA is evaluated by extensive experiments. Finally, this work is concluded and future work will be given in Section 5.

2. Background and Related Works

The bilingual corpus can be constructed from the parallel data crawled by an efficient web crawler after preprocessing. The data will be used in testing for classification on the future data added to the corpus.

A web crawler is an automated tool that traverses webpages in tree-structural websites for some particular information by following embedded hypertext links in pages. The information is indexed before storing in a large repository for efficient querying.

As the structure of website is dynamic and distributed, various crawling architectures have been proposed. Shruti and Parul [5] categorized them into four different types, which are Incremental web crawler, Hidden web crawler, Focused crawlers and Parallel crawlers (Topical crawlers or Topic driven crawlers) [6][7][8][9][10][11][12][13]. Incremental web crawler keeps a local copy of the websites. It continuously and constantly revisits them for any update so as to keep its local collection fresh for higher quality improvement. Hidden web crawler downloads high quality pages hidden from the search engine by using search form analyzer. Focused crawler considers the co-relation of pages as visit priority before downloading them into its repository. Parallel crawler executes multiple processes in parallel to search through the entire web by search engines. It is noted that the proposed HCA is based on Focused crawling architecture and Parallel crawling architecture.

3. Proposed Architecture

Hybrid Crawling architecture (HCA) is a web crawler that is designed to efficiently collect parallel data of relevant topics from the Web. Based on the URL links and content rules configured, HCA traverses websites for articles of bilingual content in the links.

HCA achieves great parallelism that can be applied to a variety of web-oriented applications. A weighting mechanism to process the relevance scores of unvisited web pages to crawl web pages. It saves computing resources and improve download efficiency while extracting specified content from particular websites. It narrows down the amount of crawling effort to be put on the web. The amount of crawling load is reduced.

In the meantime, HCA manages its crawling job using multi-threading, maximizing download performance and minimizing resource consumption. HCA coordinates and distributes URL links to download threads, downloading

html pages from the Web. URL links inside the pages will be followed until all links have been visited. Downloaded pages will then be stored in local repository if its topic passes the relevance verification.

3.1. Design

HCA consists of a number of major components, which are Scheduler, Multithread downloader, Parser and Extractor, and Classifier. Scheduler extracts URL links from html pages for the downloader before storing them. As HCA considers the relevance of topic of articles when crawling, it verifies URL links based on the bilingual criteria on the articles. Once confirmed, URL links are forwarded to Multithread downloader. Downloaded pages will be processed by Parser and Extractor. Based on the training data source, Classifier makes relevance judgments on the page contents, determining if its embedded links are required to be traversed. If it is bilingual articles, the links will be put in Queue for scheduling and storing. Otherwise, it updates the Irrelevance Table. Fig. 1 depicts the architecture of HCA.

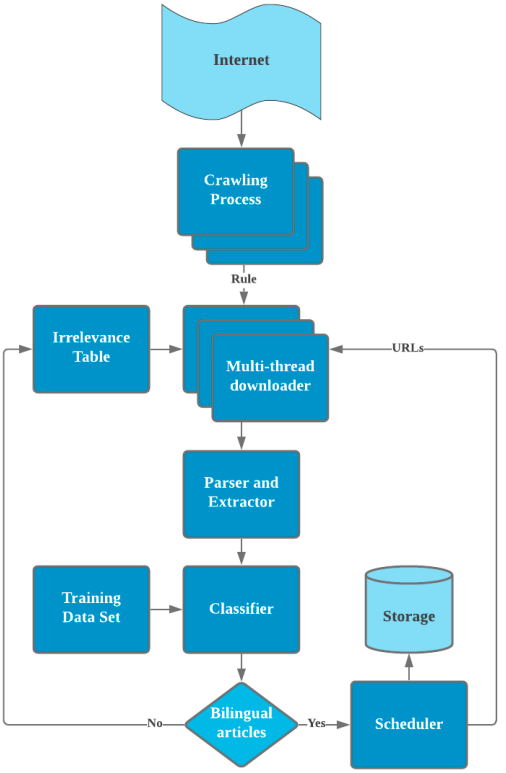

Fig. 1. Architecture of Hybrid Crawling Architecture

In Fig. 1, there are several Crawling processes running in HCA. Each process is responsible for one website (URL link) initially. By extracting URL links in the website home page (e.g., index.html), the links are assigned to Multi-thread downloader as their task. The downloader follows the link and subsequent embedded links in the downloaded pages recursively. The entire website can be traversed.

Downloaded pages are investigated by Parser and Extractor based on the content rules assigned. The content rules are guidelines to the extraction of the text content identified and located in the pages. The text content is then verified by Classifier for topic relevance, which assigns scores to URL links found in the content. Based on the Training data set obtained in advance (see Chapter 5), Classifier determines whether the topic relevance of the links meets the assigned topic for the website. If not, the link is then omitted. The URL link is further investigated for bilingual

requirement. If it meets, the link is put in the queue in Scheduler for download and the text content is saved in Storage. Otherwise, the Irrelevance Table is updated with the link for reference by the Multi-thread downloader.

4. Performance Evaluation

The performance of HCA is evaluated by Recall rate, R, and Precision rate, P. R reflects the coverage of focused topics and P measures the correlation of topics [14].

Assume that C is the total number of relevant topics on the Internet, M is the total number of crawled topics, and T is the number of relevant topics in the crawled webpages. R is then expressed as (1).

$$R = T/C \tag{1}$$

where P is expressed as (2).

$$P = T/M \tag{2}$$

Theoretically, C is best fit for evaluating crawling performance. However, C is usually not easily obtained, R is then not common for the evaluation. On the other hand, M is an objective figure that can be summarized from the crawled data. Therefore, P is used as an index to measure the crawling performance.

To illustrate the crawling performance of HCA, an intensive experiment on crawling 9,000 web pages by HCA, Focused crawling architecture and Parallel crawling architecture has been conducted. In the experiment, some websites of relevant topics (sport-related, legislation-related and general news) were identified and assigned. Fig. 2 shows the variation of the precision rate of HCA against Focused crawling architecture and Parallel crawling architecture.

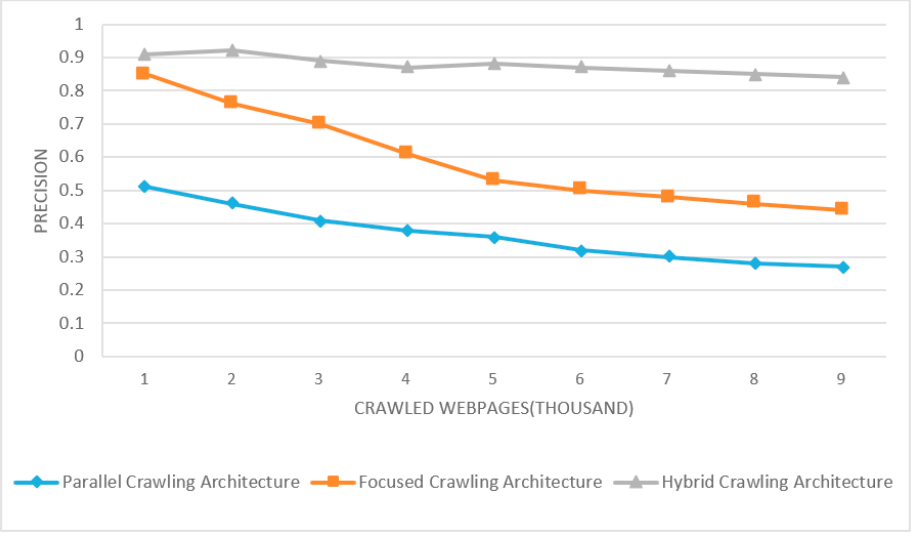

Fig. 2. Precision rate of HCA against existing crawling architectures

As can be seen in Fig. 2, the result showed that HCA maintained the precision rate at a certain level between 0.8 and 0.9 after the crawling 9,000 web pages, while the other two crawling architectures showed a downward trend. If the result is presented in terms of individual relevant topics, the precision rate for the three architectures on the three topics stays at similar level. HCA still outperforms the other two architectures, as shown in Fig. 3.

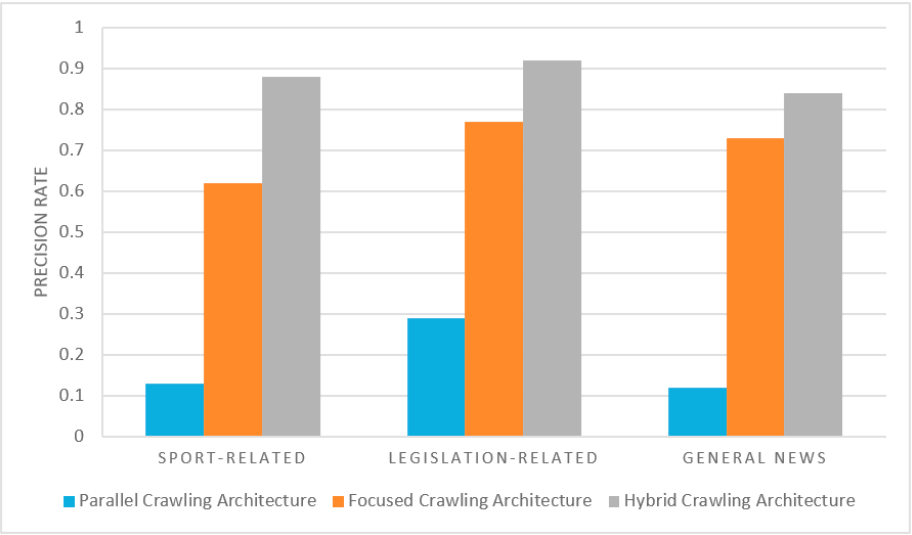

Fig. 3. Precision rate of crawling on relevant topics

It is remarkable that the crawling speed of HCA is significantly improved by the number of crawling threads in the experiment. The number of relevant web pages collected by HCA is about 30% more than that by Focused crawling architecture. It is even 200% more than that by Parallel crawling architecture. Fig. 4 shows the crawling speed of HCA against existing crawling architectures on relevant pages.

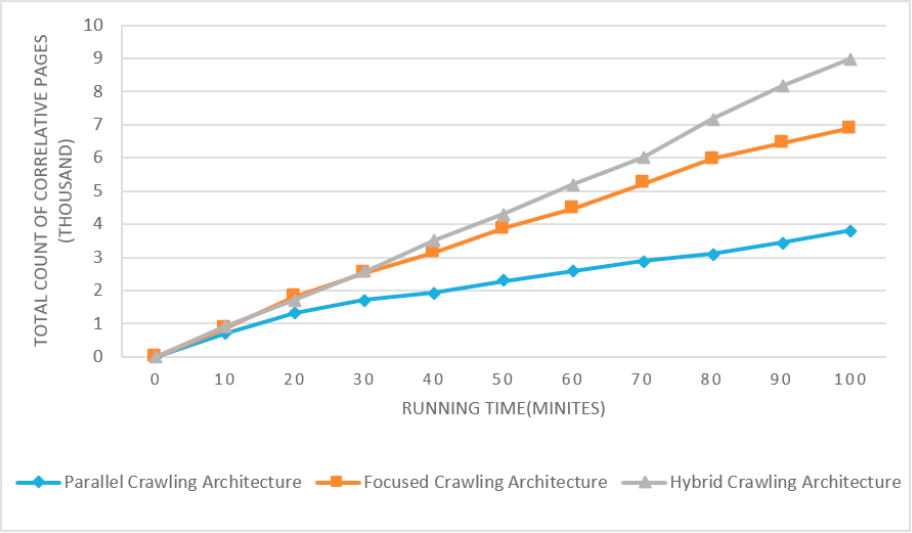

Fig. 4. Crawling speed of HCA against existing crawling architectures on relevant pages

With this significant performance in crawling, parallel data of particular topics from the Web can be collected efficiently and effectively using HCA.

5. Conclusion

In this paper, we present the Hybrid Crawling Architecture (HCA) that collects parallel data with relevant topics from the web for the construction of bilingual corpus for training machines learning systems. Evaluation result shows that HCA outperforms Focused crawling architecture and Parallel crawling architecture for over 30% and 200% respectively, in terms of quantity.

In the future, the work will be extended to the construction of bilingual corpus. The improvement of the efficiency and accuracy of machine translation systems is highly expected using the bilingual corpus.

Acknowledgement

This work was supported in part by the Macao Polytechnic Institute - Edge Sensing and Computing: Enabling Human- centric (Sustainable) Smart Cities (RP/ESCA-01/2020).

References

[1] Anandika, A., Mishra, S. P.: A Study on Machine Learning Approaches for Named Entity Recognition. International Conference on Applied Machine Learning (2019). doi:10.1109/icaml48257.2019.00037.

[2] Nahar, S., Huda, M. N., Nur-E-Arefin, M., Rahman, M. M.: Evaluation of machine translation approaches to translate English to Bengali. In: 20th International Conference of Computer and Information Technology (2017). doi:10.1109/iccitechn.2017.8281851.

[3] Tse, R., Mirri, S., Tang, S.-K, Pau, G, Salomoni, P.: Building an Italian-Chinese Parallel Corpus for Machine Translation from the Web. In: 6th EAI International Conference on Smart Objects and Technologies for Social Good (GOODTECHS), pp. 265-268 (2020) doi: 10.1145/3411170.3411258.

[4] K. I. Chan, N. S. Chan, S. -K. Tang and R. Tse, "Applying Gamification in Portuguese Learning," 2021 9th International Conference on Information and Education Technology (ICIET), 2021, pp. 178-185, doi: 10.1109/ICIET51873.2021.9419612.

[5] Sharma, S., Gupta, P.: The anatomy of web crawlers. In: International Conference on Computing, Communication & Automation. (2015). doi:10.1109/ccaa.2015.7148493.

[6] Deshmukh, S., & Vishwakarma, K. (2021). A Survey on Crawlers used in developing Search Engine. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iciccs51141.2021.9432368.

[7] Shi, Z., Shi, M., & Lin, W. (2016). The Implementation of Crawling News Page Based on Incremental Web Crawler. 2016 4th Intl Conf on Applied Computing and Information Technology/3rd Intl Conf on Computational Science/Intelligence and Applied Informatics/1st Intl Conf on Big Data, Cloud Computing, Data Science & Engineering (ACIT-CSII-BCD). doi:10.1109/acit-csii-bcd.2016.073.

[8] Sundarde, S., & Rathod, P. R. (2016). Smart crawler for hidden web interfaces. 2016 Online International Conference on Green Engineering and Technologies (IC-GET). doi:10.1109/get.2016.7916710.

[9] Yan, W., & Pan, L. (2018). Designing focused crawler based on improved genetic algorithm. 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI). doi:10.1109/icaci.2018.8377476.

[10] Langhi, J. G., & Jadhav, S. (2018). Parallel Crawling for Detection and Removal of DUST Using DUSTER. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). doi:10.1109/iccubea.2018.8697837.

[11] Iliou, C., Kalpakis, G., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2016). Hybrid Focused Crawling for Homemade Explosives Discovery on Surface and Dark Web. 2016 11th International Conference on Availability, Reliability and Security (ARES). doi:10.1109/ares.2016.66.

[12] Kumar, M., Bhatia, R., & Rattan, D. (2017). A survey of Web crawlers for information retrieval. WIREs Data Mining and Knowledge Discovery, 7(6). doi:10.1002/widm.1218.

[13] Hernandez, J., Marin-Castro, H. M., & Morales-Sandoval, M. (2020). A Semantic Focused Web Crawler Based on a Knowledge Representation Schema. Applied Sciences, 10(11), 3837. doi:10.3390/app10113837.

[14] B., Carterette.: Precision and Recall. In: Liu L., Özsu M.T. (eds) Encyclopedia of Database Systems. Springer, New York, NY (2018). https://doi.org/10.1007/978-1-4614-8265-9_5050.