Approach of a Japanese Co-Occurrence Words Collection Method for Construction of Linked Open Data for COVID-19

Yuki Nagai
Department of Information and Computer Engineering, Okayama University of Science (OUS)
Okayama, Japan
t18j057ny@ous.jp

Tetsuya Oda
Department of Information and Computer Engineering, Okayama University of Science (OUS)
Okayama, Japan
oda@ice.ous.ac.jp

Nobuki Saito
Department of Information and Computer Engineering, Okayama University of Science (OUS)
Okayama, Japan
t17j033sn@ous.jp

Aoto Hirata
Engineering Project Course, Okayama University of Science (OUS)
Okayama, Japan
t17p013ha@ous.jp

Masaharu Hirota
Department of Information Science, Okayama University of Science (OUS)
Okayama, Japan
hirota@mis.ous.ac.jp

Kengo Katayama
Department of Information and Computer Engineering, Okayama University of Science (OUS)
Okayama, Japan
katayama@ice.ous.ac.jp

***Abstract*—The Coronavirus Disease (COVID-19) is spreading around the world. There is a possibility that lifestyles will continue to change in the future, and it is essential to take COVID-19 into account in our daily lives as well as in business. However, COVID-19 is causing unprecedented changes in our lives, with no known cure and no established measures to prevent infection. Therefore, it is necessary to consider these response measures based on various information about COVID-19. In this paper, we propose a method for collecting Japanese co-occurrence words with the goal of construct a Linked Open Data (LOD) for COVID-19.**

***Index Terms*—Co-occurrence Words Collection, COVID-19, Linked Open Data.**

I. Introduction

In current year, Coronavirus Disease (COVID-19) is spreading around the world (pandemic). In Japan, the Japanese government has declared a state of emergency and restrictions on going out of the house have been imposed, and people are being asked to refrain from sightseeing, eating out, watching sports games, listening to music at concert halls, etc., which could have an impact on the economy. In addition, major changes in lifestyles are occurring, such as those represented by online work and classes. There is a possibility that lifestyles will continue to change in the future, and it is essential to take COVID-19 into account in our daily lives as well as in business. However, COVID-19 is causing unprecedented changes in our lives, with no known cure, not enough information and no established measures to prevent infection. Therefore, it is necessary to collect various information about COVID-19.

In this paper, we propose a method for collecting co-occurrence words with the goal of construct a Linked Open Data (LOD) [1]–[5] for COVID-19.

II. Proposed System

In this section, we describe the proposed system. In Fig. 1, we show the structure of proposed system. In order to collect the data on COVID-19 in Japan, we proposal a crawling, scraping and indexer systems for collecting the COVID-19 related websites. For the crawling and scraping, we set an upper limit on the number of data collection. First, the crawler chooses a website as a starting point, and reads the HTML and scrapes URL of the website. The extracted strings convert to words using MeCab [6] which a Japanese morphological analysis system. Then, the indexer categorizes the URL, strings and words. The proposed system collects data by repeating these processes.

The proposed system assigns each URL and words the distinguish number for constructing LOD of COVID-19, and the indexer recognizes these numbers. For structure of storing data, we consider Resource Description Framework (RDF) in LOD. The URL and words in the same websites is stored with the same data structure. So, each URL and words linked to each other data makes that searching linked data efficiently. When the indexer stores new data it checks the same stored data for existence already. For example of RDF in LOD, we show the following:

```
@prefix bp: <https://corona.go.jp/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema> .
@prefix geo: <http://www.w3.org/> .
<1> rdfs:label "XXXXXX"@ja ;
bp:title "COVID-19 Information and Resouces" ;
bp:url "http://XXXXX.ac.jp" ;
bp:words "Coronavirus Disease (COVID-19)" ;
```

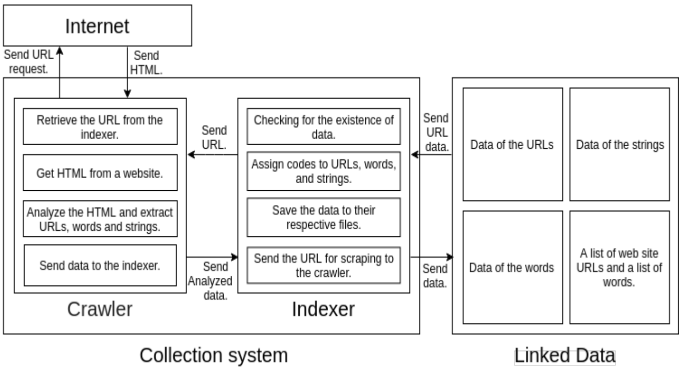
Internet
Send URL request.
Send HTML.
Retrieve the URL from the indexer.
Get HTML from a website.
Analyze the HTML and extract URLs, words and strings.
Send data to the indexer.
Send URL.
Send Analyzed data.
Checking for the existence of data.
Assign codes to URLs, words, and strings.
Save the data to their respective files.
Send the URL for scraping to the crawler.
Send URL data.
Send data.
Data of the URLs
Data of the strings
Data of the words
A list of web site URLs and a list of words.
Crawler
Indexer
Collection system
Linked Data

Fig. 1. The structure of proposed system.

TABLE I
EXPERIMENTAL PARAMETERS

Parameters	value
Depth of search	3
Waiting time[$second$]	10
Minimum length of string	1
Maximum length of strings	10000

III. CASE STUDY

In this section, we present a case study of the proposed system. For experiment, the proposed system crawl and scrape websites. The total number of websites that were crawled and scraped is 654. Through scraping, we can collect the Japanese strings of 473747 [$lines$] and the number of words are 542999 [$words$]. We show the experimental results and the visualization results of proposed system in Fig. 2 and Fig. 3.

Fig. 2 shows experimental results for the number of appearance vs. words. We show that the 10 most frequently occurring words in the collected data in Fig. 2. The most common Japanese word in collected data means “infection”.

The co-occurrence network in Fig. 3 is that larger circles indicate a higher frequency of words and smaller circles indicate a lower frequency of words. For analysis of the co-occurrence network, we remove noise words in the set of co-occurrence word. The minimum number of co-occurrence words is set to 250 [$times$] and if the co-occurrence words that appeared more than 250 [$times$] is analysis. For experimental results, the number of co-occurrence words is 181 and the number of extracted words was 150.

IV. CONCLUSION

In this paper, we proposed a Japanese co-occurrence words collection method with the goal of construction of LOD for COVID-19. For experimental results, the proposed method can collect the Japanese co-occurrence words. In the future, we will collect the more data to improve the reliability of co-occurrence words, and construct the vocabulary foundation for COVID-19.

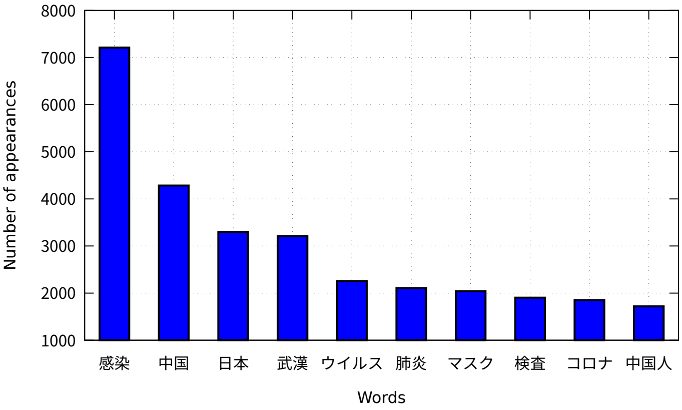

Fig. 2. Experimental results.

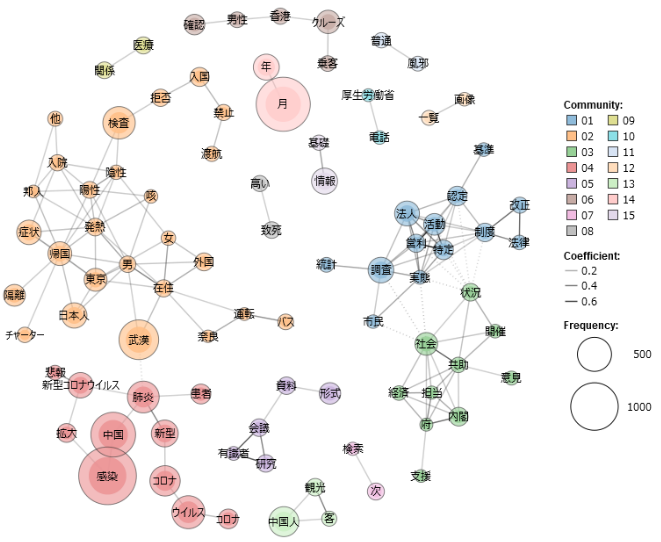

Fig. 3. Visualiuzation results of co-occurence network.

ACKNOWLEDGEMENT

This work was supported by Grant for Promotion of Okayama University of Science (OUS) Research Project (OUS-RP-20-3).

REFERENCES

[1] C. Gutierrez, et. al., “Introducing Time into RDF”, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 2, pp. 207-218, 2007.
[2] J. Fernández, et. al., “Compact Representation of Large RDF Data Sets for Publishing and Exchange”, Proc. of The 9-th International Semantic Web Conference (ISWC-2010), pp 193-208, 2010.
[3] I. Ermilov, et. al., “Linked Open Data Statistics: Collection and Exploitation”, The 4-th International Conference on Knowledge Engineering and the Semantic Web (KESW-2013), pp. 242-249, 2013.
[4] P. Ristoski, et. al., “Mining the Web of Linked Data with RapidMiner”, Journal of Web Semantics, Vol. 35, No. 3, pp. 142-151, 2015.
[5] M. Kamdar, et. al., “Enabling Web-scale Data Integration in Biomedicine Through Linked Open Data”, npj Digital Medicine, Vol. 2, No. 90, pp. 1-14, 2019.
[6] T. Maki, et. al., “Resource Propagation Algorithm Considering Predicates to Complement Knowledge Bases in Linked Data”, International Journal of Space-Based and Situated Computing, Vol. 8, No. 2, pp. 115-121, 2018.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE