Received: 20 January 2021 | Revised: 10 February 2021 | Accepted: 21 February 2021

DOI: 10.1002/int.22413

RESEARCH ARTICLE

WILEY

A novel corpus-based computing method for handling critical word-ranking issues: An example of COVID-19 research articles

Liang-Ching Chen[1,2] | Kuei-Hu Chang[3,4]

[1]Department of Foreign Languages, R.O.C. Military Academy, Kaohsiung, Taiwan

[2]Institute of Education, National Sun Yat-sen University, Kaohsiung, Taiwan

[3]Department of Management Sciences, R.O.C. Military Academy, Kaohsiung, Taiwan

[4]Institute of Innovation and Circular Economy, Asia University, Taichung, Taiwan

Correspondence

Kuei-Hu Chang, Department of Management Sciences, R.O.C. Military Academy, Kaohsiung 830, Taiwan; Institute of Innovation and Circular Economy, Asia University, Taichung 413, Taiwan.
Email: evenken2002@yahoo.com.tw

Funding information

Ministry of Science and Technology, Taiwan, Grant/Award Numbers: MOST 108-2410-H-145-001, MOST 109-2410-H-145-002

Abstract

A corpus is a massive body of structured textual data that are stored and operated electronically. It usually combines with statistics, machine learning algorithms, or artificial intelligence (AI) technologies to explore the semantic relationship between lexical units, and beneficial when applied to language learning, information processing, translation, and so forth. In the face of a novel disease, like, COVID-19, establishing medical-specific corpus will enhance frontline medical personnel's information acquisition efficiency, guiding them on the right approaches to respond to and prevent the novel disease. To effectively retrieve critical messages from the corpus, appropriately handling word-ranking issues is quite crucial. However, traditional frequency-based approaches may cause bias in handling word-ranking issues because they neither optimize the corpus nor integrally take words' frequency dispersion and concentration criteria into consideration. Thus, this paper develops a novel corpus-based approach that combines a corpus software and Hirsch index (H-index) algorithm to handle the aforementioned issues simultaneously, making word-ranking processes more accurate. This paper compiled 100 COVID-19-related research articles as an empirical example of the target corpus. To verify the proposed approach, this study compared the results of two traditional frequency-based approaches and the proposed approach. The results indicate that the proposed approach can refine corpus and simultaneously compute words' frequency dispersion and concentration criteria in handling word-ranking issues.

KEYWORDS
corpus, COVID-19, H-index algorithm, machine learning

1 | INTRODUCTION

Seeking proper algorithms to optimize today's high computational real-world problems is a critical and challenging task that has taken a great deal of efforts in the last decade. For instance, Barshandeh and Haghzadeh[1] proposed a novel hybrid physics-based nature-inspired meta-heuristic algorithm which named as proposed hybrid optimization algorithm (PHOA). They integrated atom search optimization (ASO) and tree-seed algorithm (TSA) to successfully optimize traditional meta-heuristic algorithms, moreover, PHOA was also tested on seven real-life engineering problems and the results of PHOA were superior among traditional algorithms. In addition, Barshandeh et al.[2] proposed a novel hybrid multipopulation algorithm (HMPA) that combined artificial ecosystem-based optimization (AEO) and Harris Hawks optimization (HHO) algorithms, then, adopted Levy-flight strategy, local search mechanism, quasi-oppositional learning, and chaos theory to maximize the efficiency of the HMPA. In their research, HMPA was tested on seven constrained/unconstrained real-life engineering problems, and the calculation results of HMPA were compared with similar advanced algorithms. The results indicated that HMPA was outperformed the other competitor algorithms significantly. To extend the concepts of Barshandeh and Haghzadeh[1] and Barshandeh et al.[2] researches, it is critical to seek optimization algorithms in handling real-life corpus analysis issues, especially during this era of information explosion.

In this modern digital era, corpus building has evolved from manual collection to automatic collection of textual data. To manage its massive textual data, corpus usually combines statistics, machine learning algorithms, or artificial intelligence (AI) techniques; this facilitates the efficiency of data collection, information processing, information retrieval (IR), and so on. Natural languages are one of the most ubiquitous formats of information flow among people. Analyzing, integrating, and reproducing textual data inevitably require importing highly accurate algorithms to process natural languages' semantics and syntax. Corpus-based approaches that embed statistical algorithms, such as frequency calculation and log-likelihood test, are commonly adopted by linguists and data analysts for deciphering linguistic patterns and extracting domain knowledge.[3,4] In addition, in corpus-based approaches, word ranking is an important technique used to define words' importance level and to retrieve critical words from the large textual data; this especially helps discover semantic relationships between lexical units.[5,6]

In the face of novel diseases, it is essential to build specialized medical corpora for integrating, managing, and retrieving massive information related to the diseases; such corpora help further effectively analyze, react, prevent the diseases. For example, COVID-19, a novel disease outbreak in December 2019, has a close genetic form with SARS coronavirus (SARS-CoV), and has caused over 40 million confirmed cases and 1 million deaths by the end of October 2020 (less than a year).[7–12] Leading researchers from various countries are trying to unveil the mystery of the novel disease. As of the end of October 2020, Web of Science (WOS), an internationally renowned academic database, has published more than 35,000 COVID-19-related research articles (RAs); this number keeps rising. No doubt, governments around the world are seeking direct and effective measures to mitigate the pandemic and speed up the cure of the confirmed cases.[13,14] With big textual data about COVID-19 being rapidly distributed, it is critical for humans to rely on machine algorithms to compute important semantic information, thereby, filtering and retrieving critical messages.[15,16] Hence, adopting corpus-based approaches to process and integrate COVID-19-related English-mediated textual data will enhance frontline medical personnel's efficiency of knowledge acquisition and perception.

Since the advent of computer technology, the practicality of corpus-based approaches has received widespread attention and adoption in textual information analysis fields. Frequency criterion is considered as one of the core analytical techniques in corpus-based approaches. However, simply relying on tokens' frequency values to determine their importance may be insufficient; tokens' dispersion and concentration conditions also need to be taken into consideration. For example, in terms of importance, a word occurring 100 times in an RA is not equal to a word occurring 10 times each in 10 RAs because words' dispersion and concentration conditions are different. A potential solution that adopts Hirsch index (H-index) algorithm to integrate and compute the criteria of dispersion and concentration is required to address this issue. H-index algorithm was originally used to quantify the accumulative impacts and relevance of a researcher's scientific research achievements.[17–23] Nevertheless, this algorithm was not only limited to the purposes of evaluating academic achievements but also seen its applications in the fields of risk assessment,[22] medical,[24] and so forth.

Handling critical word-ranking issues using traditional frequency-based approaches may cause distortion and bias because those approaches neither refine the corpus data nor simultaneously compute words' frequency dispersion and concentration criteria, hence, the alleged highly important words with high frequency would be challenged. Thus, this paper proposed a novel corpus-based approach that integrates a corpus software and H-index algorithm as a computation method and evaluation metric that can enhance the accuracy of word ranking, compensate the deficiency of the traditional frequency-based approaches, and further augment the efficacy of corpus-based analysis. To verify the proposed approach, 100 COVID-19-related medical RAs with Science Citation Index (SCI) from WOS were retrieved and compiled as the big textual data and an empirical example which was embedded into the proposed approach. The main reason the researchers adopted this empirical example was that SCI journals represent high-quality academic publications. In addition, understanding the specific linguistic pragmatics of medical RAs will assist frontline healthcare personnel in processing and acquiring important COVID-19 medical messages.

The remainder of this paper is organized as follows: Section 2 describes preliminaries, explains the theoretical framework, and introduces the recent novel disease, COVID-19. Section 3 describes detailed steps of the proposed approach. Section 4 uses COVID-19-related RAs from WOS as the big textual data (i.e., the target corpus) and as an empirical example to verify the proposed approach. Section 5 is the concluding part of this study.

2 | PRELIMINARIES

2.1 | Conventional frequency-based corpus analysis

With the advance of computer technology, corpus development has enabled people to establish algorithms to integrate, manage, and process natural languages from massive textual data, thereby driving the progress of natural language processing (NLP) and AI-related industries. O'Keeffe et al.[25] noted that information on frequency counts of tokens is the basis for understanding core vocabularies that native speakers use frequently and the common combinations of vocabulary usage. Collecting large data (corpora) from native speakers' written texts and discourse transcripts will provide strong evidence for understanding their linguistic patterns. Moreover, ranking words based on their frequency will show the words that are adopted by the majority and the words that are used in day-to-day communications.[26,27] Hence, frequency-based corpus analytical approaches have widely been adopted by linguists, sociologists, text analysts, and so on for extracting strong linguistic evidence for interpreting cultural phenomenon, jargon, genre type, and so on.[28,29] For example, Le and Miller[6] adopted Sketch Engine, a corpus software, to cross-compare four medical corpus sources to extract the most frequently occurring medical morphemes in medical RAs. The resulting data indicated 136 specialized medical morphemes that account for 8.5% of the lexical items in the Medical Web Corpus, and the results offered English as a Foreign Language (EFL) medical students a useful academic resource for enhancing their comprehension of English medical vocabulary. Grabowski[5] used WordSmith Tools 5.0, a corpus software, to present a corpus-driven description of the use and functions of top-50 keywords (i.e., based on keyness values) complemented by a similar description of top-50 lexical bundles (LBs; based on frequency values) in the analysis of specialized corpus which contains patients' prescriptions, outlines of product introduction, clinical trial protocols, and pharmacological RAs. The results provided significant pedagogical value for English for specific purposes (ESP) students and EFL practitioners in the pharmaceutical domain.

Traditional corpus-based approach was designed for effectively clarifying, categorizing, and interpreting the patterns of natural languages. Computing word frequency is thus a critical technique that corpus software is capable of (see Equation 1).

Definition 1 (Anthony[30] and Scott[31]). If a_f represents the cumulated value of a token's overall frequency, where f means the sequence of a subcorpus; a means a token's frequency; and a_n means a token's frequency, counted in n subcorpus.

$$\sum_{f=1}^{n} a_f = a_1 + a_2 + a_3 + \cdots + a_n. \tag{1}$$

2.2 | H-index algorithm

H-index algorithm was proposed by Jorge E. Hirsch,[19] a physicist and a professor at the University of California, San Diego in 2005. H-index is an evaluation mechanism that is used to measure a researcher's academic productivity and the citation rate of published articles; the index h is given to represent the number of papers with citation number more than h, it is a useful index to quantify the academic achievements of a researcher. Nowadays, this mechanism has been widely adopted in several academic databases, such as WOS, Google Scholar, Scopus, and even other research fields.[18,20,22] The algorithm computes the interrelationships between publication quantities and numbers of citations, and defines a researcher's academic influence in certain domain. For example, Li et al.[22] adopted H-index algorithm to assess the significance of the urban railroad network structure, which took topology, passenger quantity, and passenger flow correlation of Beijing urban railroad network into consideration to refine rail network structure and decrease operational risks. Gao et al.[17] proposed a weighted H-index (h^w) by constructing an operator H on weighted edges. Moreover, the accumulation of weighted H-index (s^h) in the node's neighborhood defines the spreading influence, then utilized the susceptible–infected–recovered (SIR) model to investigate an epidemic spreading process on 12 real-world networks, and to further define the most influential spreaders. Hanna et al.[24] developed a novel metric for quantifying patient-level utilization of emergency department (ED) imaging. In their research, H-index was adopted to measure a patient's annual ED imaging volume, and the resulting data of patients' H-index values were used as the referential data for mitigating imaging-related costs and improving throughput in the ED. In summary, H-index algorithm integrates multiple considerations to evaluate and to create the values of importance of the research objects, moreover, the definition of Hirsch's H-index algorithm is defined as follows:

Definition 2 (Hirsch[19]). If the value of function f represents citation times of each paper and is ranked in descending sequence (see Equation 2), then find $f(n)$ equal to or larger than n (see Equation 3). The value of H-index has to satisfy this criterion, and can be described as follows:

$$\text{H-index}(f) = \max_n \min(f(1), ..., f(n)), \tag{2}$$

$$f(n) \geq n, \tag{3}$$

where n is the paper numbers, $f(n)$ is the citation times of the paper, and $\max_n \min(f(1), ..., f(n))$ represents citation times of each paper ranked from maximum to minimum.

To understand this algorithm, two examples are given as follows:

Example 1. If a researcher has 10 published articles ($n = 10$) identified as A_1, A_2, A_3, ..., A_{10}, and the citation numbers are randomly given as 9, 5, 50, 20, 6, 8, 6, 4, 1, 0, thus, $f(A_1) = 9$, $f(A_2) = 5$, $f(A_3) = 50$, $f(A_4) = 20$, $f(A_5) = 6$, $f(A_6) = 8$, $f(A_7) = 6$, $f(A_8) = 4$, $f(A_9) = 1$, $f(A_{10}) = 0$. Then, rerank the citation numbers in descending sequence, and they become $f(b_1) = 50$, $f(b_2) = 20$, $f(b_3) = 9$, $f(b_4) = 8$, $f(b_5) = 6$, $f(b_6) = 6$, $f(b_7) = 5$, $f(b_8) = 4$, $f(b_9) = 1$, $f(b_{10}) = 0$. The results indicate that b_6 satisfies the criteria of Equation (2) where $f(b_6) \geq 6$, thus H-index = 6 (see Table 1).

Example 2. The illustrative diagram (see Figure 1) also explains the H-index algorithm; there is a reference line (i.e., it represents that the n paper needs to have at least n citations) on the diagram, the papers' citations have to be over or on the reference line to be included into the value of H-index. $f(b_6)$, in this case, is the sixth paper and is also the last paper on the reference line. Meanwhile, its citation time is six and it satisfies Equation (2), $f(b_6) \geq 6$, thus, the value of H-index is equal to 6.

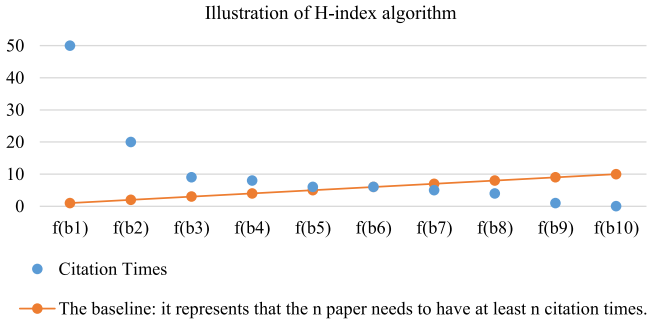


FIGURE 1 Illustration of H-index algorithm [Color figure can be viewed at wileyonlinelibrary.com]

In summary, H-index algorithm presents the estimation of the significance, importance, and wide influence of a researcher's cumulative academic contributions. It has become a standard measurement and a criterion that is unbiased to compare and to evaluate the academic achievements of researchers who are competing in the same research fields.[19]

2.3 | COVID-19

COVID-19, whose original nomenclature was SARS-CoV-2, was renamed by WHO in February 2020. The clusters of first cases of the virus were discovered in Wuhan city, Hubei province, China.[7] Epidemiologists, for now, propose a possibility that the virus which was originally carried by wild animals entered to human-to-human transmission routes because locals in the city have preference for “Yeh-Wei”, meats of wild animals, such as bats, birds, and rodents.[8,10] Upon visiting the possible source location of COVID-19, Huanan market, medical experts found plenty of contaminated carcasses of wild animals stocked and piled for sale. Thus, medical and biological experts speculated that the novel coronavirus may constantly mutate in animal hosts (e.g., bats, pangolins, etc.), then become capable of infecting humans, especially when people process animal carcasses or eat uncooked food ingredients that host the virus.[8] Indeed, many studies have indicated that bats were the initial hosts of COVID-19 because it has over 90% similarity to two SARS-like coronaviruses from bats, bat-SL-CoVZX45 and bat-SL-CoVZX21.[9,12] In terms of etiology, COVID-19 has a genetic form similar to SARS-CoV (i.e., an acute respiratory syndrome coronavirus which broke out in 2002) and MERS-CoV (i.e., middle east respiratory syndrome coronavirus which broke out in 2012),[12,32] but its spike (S) protein has mutated and enabled it to attack the host's immune system, making the host too weak to resist the virus.[33] The comparison of COVID-19 and two prior coronaviruses shows that COVID-19 causes a low fatality rate but has extremely high infectious capability.[34] Yi et al.[12] also pointed out that the majority of the human population lacks the immunity of COVID-19 and is thus susceptible to the novel coronavirus.

Reverse transcriptase polymerase chain reaction (RT-PCR) was initially adopted as the primary criteria for diagnosing COVID-19. However, RT-PCR test method has a high probability of misdiagnosis that may accelerate the pandemic, thus, multiple diagnosing test approaches were integrated with the investigations of travel history survey, disease records, clinical symptoms (see Figure 2), lab tests, and X-ray or computed tomography (CT) for making effective diagnoses.[35] Following the intensification of the COVID-19 pandemic, rapid test

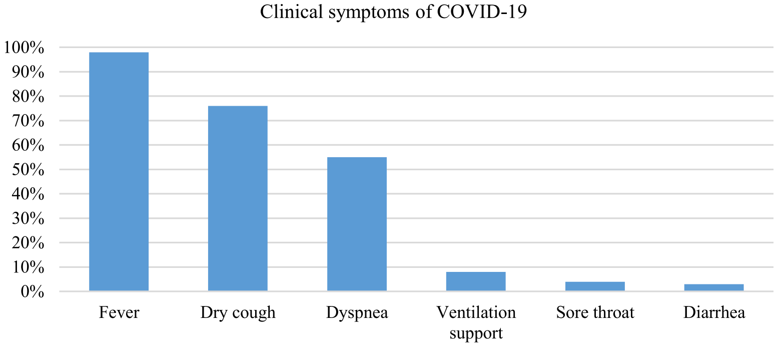


FIGURE 2 Clinical symptoms of COVID-19 [Color figure can be viewed at wileyonlinelibrary.com]

toolkits were invented to rapidly detect RNA, antigen, or antibody of SARS-CoV-2, giving more time to frontline healthcare personnel to respond and cure the confirmed cases. In addition, prior studies pointed out that without protective measures (i.e., surgical masks, respiratory filtrations, etc.), three major transmission routes of inhalation, droplet, and contact routes will cause 57%, 35%, and 8.2% of COVID-19 infection probability.[36] For frontline healthcare personnel, in particular, who treat confirmed cases and have prolonged exposure to the virus emission environment and inhalation of droplets (<10 μm) that contain the virus, their possibility of infection may reach over 80%.[37] Prior research also showed that social distance (1.5–2 m) will not be effective if the virus emission source does not wear any protective equipment because the virus can be spread at least 6 m away via patients' coughing and sneezing.[38,39] Hence, even though the fatality rate of COVID-19 is not extremely high, high infection rates cause difficulties in pandemic response and prevention.

According to WHO, as of October 31, 2020, there were 45,408,704 confirmed COVID-19 cases and 1,179,363 COVID-19 deaths (see Figure 3). Because targeted therapeutic medicines are still being developed, governments can only presently rely on quarantine policies, and existing indirect medical treatments, thus, making citizens pay attention to personal hygiene, implementing border control measures, encouraging social distance and internet shopping, and so on to decrease close contacts between people and control the COVID-19 pandemic.[40–42]

COVID-19, at the time of this writing, is still a semi-unknown novel disease for medical experts and continues to be explored. To effectively manage the massive medical textual

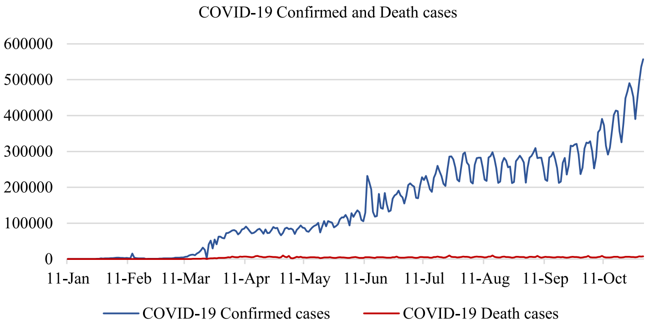


FIGURE 3 COVID-19 confirmed and death cases (data record from January 1 to October 31, 2020) [Color figure can be viewed at wileyonlinelibrary.com]

information about it, it is necessary to create a COVID-19-specialized corpus, integrating appropriate algorithms for information processing and mining.

3 | METHODOLOGY

Traditional corpus-based computing methods for critical word ranking mainly calculate words' frequency values and rank them. Prior studies believed high-frequency words may reflect specific linguistic patterns in certain domains which would benefit EFL speakers in more effective acquisition of domain knowledge when reading English texts.[3,5,6,43,44] Thus, with rapid information flow of COVID-19, establishing COVID-19 specialized corpus for timely acquisition of updated medical knowledge is especially critical for medical care personnel.[7,9,11,14,32] Certainly, as of the end of October 2020, more than 38,000 RAs on COVID-19-related topics had been published in the WOS database; this phenomenon indicated that a large number of research results were produced by leading researchers globally. To effectively integrate and decipher the English-mediated professional textual information and to further improve the efficiency of knowledge acquisition, importing algorithms to compute key natural language semantics is quite critical. Corpus-based and NLP technology hence plays the essential roles at this time for humans to efficiently process the big textual information available.[25,45]

Previous corpus-based studies that focused on calculating words' frequency values may miss important factors and citation rates, which indicate the number of times a word is used by different text creators. For example, a medical-oriented word that occurs 10 times in 10 RAs, respectively (i.e., overall frequency is 100 times), the researchers believe, is more important than a medical-related word that occurs 200 times but only in one RA, this concept is especially critical to healthcare personnel because with time limitations, access to the most critical domain-related words are crucial. Thus, when handling critical word-ranking issues, the following two important conditions must be taken into consideration simultaneously:

(i) *Dispersion*: A word's frequency values that disperse into different subcorpora.
(ii) *Concentration*: A word's frequency values that concentrate on minorities of the subcorpus.

However, taking existing corpus software, such as AntConc 3.5.8,[30] WordSmith Tools 5.0, and so forth, as examples, within its existing algorithms, those are still unable to simultaneously compute these two conditions. Their word-ranking results can only base on frequency value or range value, respectively, hence to make the evaluation of words' importance level exist bias. Therefore, to compensate for the results bias in word-ranking issues of the traditional methods, the researchers propose a novel corpus-based approach that integrates AntConc 3.5.8[30] and H-index algorithm[19] to compute and to evaluate the importance of tokens.

The steps are as follows: in the initial stage of the proposed approach, sample and compile the textual data as the target corpus in a way that suitable for H-index algorithm. Then, adopt Chen et al.'s[46] corpus-based optimizing approach to refine the target corpus. In the middle part of the proposed approach, use AntConc 3.5.8[30] to compute tokens' frequency values and ranges, then, adopt H-index algorithm to integrally compute tokens' dispersion and concentration conditions, and to further obtain their H-index values. Next, rank tokens based on their H-index and frequency values. Postranking results will shed light on the importance of the proposed approach and imply the future possible applications in corpus-based and NLP fields. There are six steps in total in the proposed approach, moreover, detailed descriptions are shown as follows (see Figure 4):

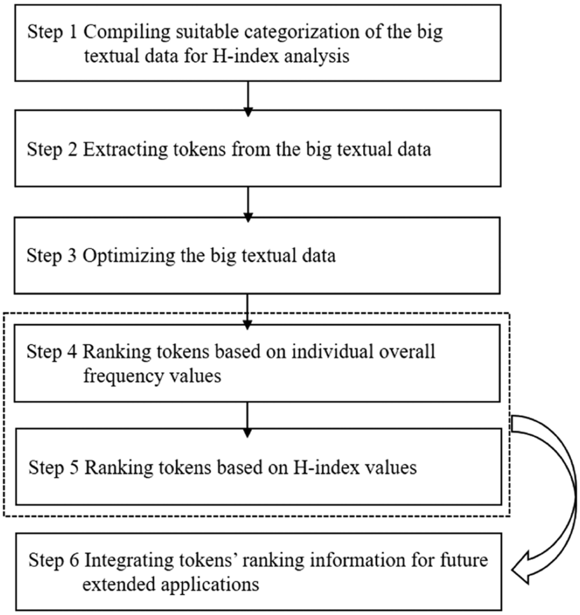

Step 1 Compiling suitable categorization of the big textual data for H-index analysis
Step 2 Extracting tokens from the big textual data
Step 3 Optimizing the big textual data
Step 4 Ranking tokens based on individual overall frequency values
Step 5 Ranking tokens based on H-index values
Step 6 Integrating tokens' ranking information for future extended applications

FIGURE 4 Flowchart of the proposed approach

Step 1. Compiling suitable categorization of the big textual data for H-index analysis.

H-index algorithm is mainly used to explore the citation rate of research papers. In this study, the authors adopt it to explore the usage rate of tokens. In this step, the target corpus (i.e., the big textual data) should be segmented into its basic elements that consider an article as a unit instead of compiling all files into a big file (see Figure 5). Hence, the H-index of tokens will be computed successfully.

Step 2. Extracting tokens from the big textual data.

Using AntConc 3.5.8 as the corpus software to calculate and unveil the composition of the big textual data, the quantitative data will be retrieved and all tokens will be labeled with numbers in this step.

Step 3. Optimizing the big textual data.

Function and meaningless words would decrease the efficiency of corpus-based approaches, hence to retrieve the substantive words which most reflect domain information, a refining process is inevitable. In this step, adopt the function wordlist and machine optimizing process to refine the big textual data,[46] the remaining content words will be processed in subsequent steps.

Step 4. Ranking tokens based on individual overall frequency criteria.

After calculating each token's overall frequency based on Equation (1) by the corpus software, the wordlist in this step will be ranked based on frequency criteria, from highest to lowest frequency sequences.

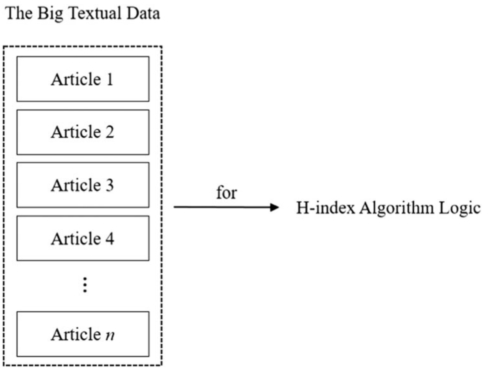


FIGURE 5 Ideal corpus compilation method for H-index algorithm

Step 5. Ranking tokens based on H-index algorithm.

In this step, the researchers adopt the H-index algorithm to compute the significance of tokens. Here, the citation times are considered as the tokens' adoption times (i.e., frequency), thus, the calculation of tokens' H-index is based on a token appearing equal to or more than n times in n RAs. First, based on Equation (2), rank the word frequency of each RA in descending order. Then, based on Equation (3), find a word's H-index value that satisfies the criteria.

Step 6. Integrating tokens' ranking information for future extended applications.

In this step, tokens' H-index and frequency values are integrated and shown on the wordlist, moreover, the sequence of tokens will have to satisfy the following criteria:

1. Ranking tokens based on their H-index values in descending order.
2. If tokens have the same H-index values, then rank their frequency values in descending order.

The proposed approach uses H-index algorithm to compute a token's degree of importance, simultaneously taking the criteria of dispersion and concentration into consideration. In addition, when facing the same H-index values, use tokens' frequency values to define their ranks to avoid hesitation that occurs when defining tokens' degree of importance.

4 | EMPIRICAL STUDY

4.1 | Overview of the compiled big textual data

The big textual data in this paper are 100 RAs that were collected from WOS. This choice was due to WOS that is one of the largest, well-known, and leading databases in the world. Moreover, many academic big textual data analysis researches and NLP researches of scientific fields adopted RAs from WOS as test data.[47–49] Hence, in this study, the researchers chose *Medicine*, *General*, and *Internal*, a category that defined journal citation reports (JCR) for WOS, they then focused on open access (OA) journals ($N = 24$). To process these 24 journals, first, the authors calculated their respective annual publications (data retrieved from 2019.9.1 to 2020.8.31), then, calculated the number

TABLE 1 H-index computing process

Original data		Computing process		
Research paper	Citation times	Research paper	Citation time	H-index result
1	9	3	50	6
2	5	4	20	
3	50	1	9	
4	20	6	8	
5	6	5	6	
6	8	7	6	
7	6	2	5	
8	4	8	4	
9	1	9	1	
10	0	10	0	

of papers that were related to the COVID-19 topic. Finally, they sampled the newest articles from each journal based on ratio and they further compiled the big textual data (see Table 2). The research fields of the sampled journals comprise (1) environmental sciences, (2) public, environmental, and occupational health, (3) infectious diseases, (4) tropical medicine, (5) microbiology, (6) toxicology, (7) healthcare sciences and services, and (8) health policy and services. Furthermore, the collected RAs all had COVID-19 in their titles, and they discussed problems and solutions during the COVID-19 pandemic in line with their research fields. The paper collecting method in this study attempted to reach a balance between domain and genre type as much as possible to make native and EFL healthcare personnel understand the most important and widely used tokens in medical RAs.

4.2 | Traditional corpus-based computing method for handling critical word-ranking issues

AntConc 3.5.8[30] works like other corpus software; based on Equation (1), it cumulates the sum of words' occurrence times (i.e., frequency values) in the corpus and ranks words. Using the compiled corpus as an example, the traditional method for handling critical word-ranking issues will cause the following problems: (1) function and meaningless words are not eliminated, hence content words are ranked behind and this decreases analytical efficiency, (2) the dispersion condition of frequency is not taken into consideration, (3) the concentration condition of frequency is not taken into consideration. Word-ranking results in Figure 6 indicate that the wordlist is based on words' overall frequency values and ranked in descending orders.

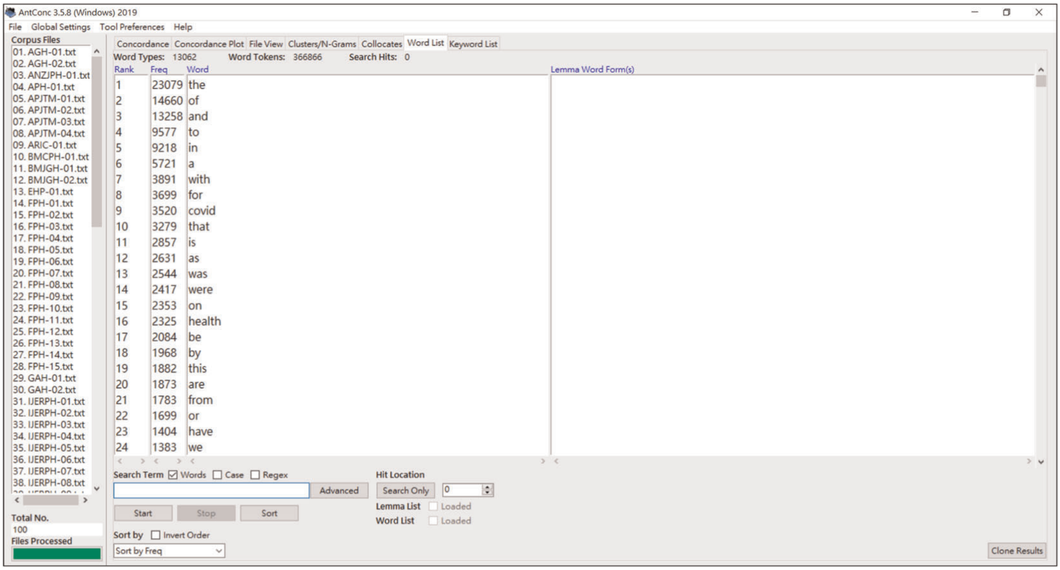
4.3 | The proposed approach

In this section, the compiled big textual data are embedded into the proposed novel corpus-based approach for calculating the actual results of the proposed approach. A detailed description is shown as follows:

TABLE 2 The composition of the big textual data

Topic	Category	Journal	Annual publication	COVID-19-related RAs	Actual collected articles
COVID-19	Medicine, General, and Internal	International Journal of Environmental Research and Public Health	7683	253	41
		Frontiers in Public Health	539	94	15
		Journal of Global Health	228	45	7
		Lancet Global Health	399	43	7
		Lancet Public Health	173	41	7
		Journal of Infection and Public Health	252	27	4
		Asian Pacific Journal of Tropical Medicine	102	22	4
		BMJ Global Health	327	13	2
		Annals of Global Health	97	13	2
		Globalization and Health	108	12	2
		Journal of Nepal Medical Association	172	11	2
		BMC Public Health	1817	8	1
		Journal of Epidemiology	79	5	1
		Antimicrobial Resistance and Infection Control	195	5	1
		Reproductive Health	180	5	1
		Australian and New Zealand Journal of Public Health	114	5	1
		Archives of Public Health	91	4	1
		Environmental Health Perspectives	175	3	1
		Health Expectations	185	2	0
		Conflict and Health	79	2	0
		Tobacco Induced Diseases	65	2	0
		Environmental Health and Preventive Medicine	70	1	0
		Safety and Health at Work	68	1	0
		Gaceta Sanitaria	116	1	0
		Total	13,314	618	100

Abbreviation: RA, research article.



Corpus Files

01. AGH-01.txt
02. AGH-02.txt
03. ANZJPH-01.txt
04. APH-01.txt
05. APJTM-01.txt
06. APJTM-02.txt
07. APJTM-03.txt
08. APJTM-04.txt
09. ARIC-01.txt
10. BMCPH-01.txt
11. BMJGH-01.txt
12. BMJGH-02.txt
13. EHP-01.txt
14. FPH-01.txt
15. FPH-02.txt
16. FPH-03.txt
17. FPH-04.txt
18. FPH-05.txt
19. FPH-06.txt
20. FPH-07.txt
21. FPH-08.txt
22. FPH-09.txt
23. FPH-10.txt
24. FPH-11.txt
25. FPH-12.txt
26. FPH-13.txt
27. FPH-14.txt
28. FPH-15.txt
29. GAH-01.txt
30. GAH-02.txt
31. IJERPH-01.txt
32. IJERPH-02.txt
33. IJERPH-03.txt
34. IJERPH-04.txt
35. IJERPH-05.txt
36. IJERPH-06.txt
37. IJERPH-07.txt
38. IJERPH-08.txt

Total No. 100
Files Processed

Word Types: 13062 Word Tokens: 366866 Search Hits: 0

Rank	Freq	Word
1	23079	the
2	14660	of
3	13258	and
4	9577	to
5	9218	in
6	5721	a
7	3891	with
8	3699	for
9	3520	covid
10	3279	that
11	2857	is
12	2631	as
13	2544	was
14	2417	were
15	2353	on
16	2325	health
17	2084	be
18	1968	by
19	1882	this
20	1873	are
21	1783	from
22	1699	or
23	1404	have
24	1383	we

Search Term ☑ Words ☐ Case ☐ Regex

Hit Location 0

Lemma List ☐ Loaded

Word List ☐ Loaded

Sort by ☐ Invert Order

Sort by Freq

FIGURE 6 Traditional corpus-based computing method used to rank words [Color figure can be viewed at wileyonlinelibrary.com]

Step 1. Compiling suitable categorization of the big textual data for H-index analysis.

To effectively compute the H-index values of each token, the composition of the corpus should consider each article as a unit. To manage the big textual data, first, the researchers gave each journal a codename. For example, *Annals of Global Health* was coded as AGH. The purpose of coding journal names was for rapidly and effectively retrieving sources of tokens, hence, increasing the efficiency of text analysis and mining. Second, the file name of each article paper is given based on a specific rule, for instance, 01. In AGH-01, 01 means the RA's serial number (i.e., from the perspective of the entire big textual data), AGH means journal codename, and –01 represents the RA's serial number in the current journal (see Table 3).

Step 2. Extracting tokens from the big textual data.

Data management of the first step indicated that the principle of coding provides huge convenience when launching AntConc 3.5.8 to process corpus data. The corpus software analyzed all RAs' word types, tokens, and lexical diversity (i.e., types and tokens ratio, TTR; see Table 4). The lexical results of the compiled big textual data indicated that authors from 100 RAs adopted 13,062 word types, and the whole corpus is composed of 366,866 running words. Furthermore, its TTR is approximately equal to 0.0356 (also see Table 4).

Step 3. Optimizing the big textual data.

On the basis of Chen et al.'s[46] research, function words, such as *a*, *an*, *the*, *it*, *is*, and so on, would decrease the efficiency of text mining and IR. Indeed, no matter which algorithm is used to calculate the importance of tokens, the irreplaceability of function words in constructing meaningful sentences will cause them to appear in resulting data or even be ranked very high, which directly decreases the accuracy and efficiency of information processing. Thus, the

TABLE 3 Journal codename and data management of RAs

Journal name	Codename	Data management of RAs
Annals of Global Health	AGH	01. AGH-01, 02. AGH-02
Australian and New Zealand Journal of Public Health	ANZJPH	03. ANZJPH-01
Archives of Public Health	APH	04. APH-01
Asian Pacific Journal of Tropical Medicine	APJTM	05. APJTM-01, 06. APJTM-02, 07. APJTM-03, 08. APJTM-04
Antimicrobial Resistance and Infection Control	ARIC	09. ARIC-01
BMC Public Health	BMCPH	10. BMCPH-01
BMJ Global Health	BMJGH	11. BMJGH-01, 12. BMJGH-02
Environmental Health Perspectives	EHP	13. EHP-01
Frontiers in Public Health	FPH	14. FPH-01, 15. FPH-02, 16. FPH-03, 17. FPH-04, 18. FPH-05, 19. FPH-06, 20. FPH-07, 21. FPH-08, 22. FPH-09, 23. FPH-10, 24. FPH-11, 25. FPH-12, 26. FPH-13, 27. FPH-14, 28. FPH-15
Globalization and Health	GAH	29. GAH-01, 30. GAH-02
International Journal of Environmental Research and Public Health	IJERPH	31. IJERPH-01, 32. IJERPH-02, 33. IJERPH-03, 34. IJERPH-04, 35. IJERPH-05, 36. IJERPH-06, 37. IJERPH-07, 38. IJERPH-08, 39. IJERPH-09, 40. IJERPH-10, 41. IJERPH-11, 42. IJERPH-12, 43. IJERPH-13, 44. IJERPH-14, 45. IJERPH-15, 46. IJERPH-16, 47. IJERPH-17, 48. IJERPH-18, 49. IJERPH-19, 50. IJERPH-20, 51. IJERPH-21, 52. IJERPH-22, 53. IJERPH-23, 54. IJERPH-24, 55. IJERPH-25, 56. IJERPH-26, 57. IJERPH-27, 58. IJERPH-28, 59. IJERPH-29, 60. IJERPH-30, 61. IJERPH-31, 62. IJERPH-32, 63. IJERPH-33, 64. IJERPH-34, 65. IJERPH-35, 66. IJERPH-36, 67. IJERPH-37, 68. IJERPH-38, 69. IJERPH-39, 70. IJERPH-40, 71. IJERPH-41
Journal of Global Health	JGH	72. JGH-01, 73. JGH-02, 74. JGH-03, 75. JGH-04, 76. JGH-05, 77. JGH-06, 78. JGH-07
Journal of Infection and Public Health	JIPH	79. JIPH-01, 80. JIPH-02, 81. JIPH-03, 82. JIPH-04
Journal of Nepal Medical Association	JNMA	83. JNMA-01, 84. JNMA-02
Journal of Epidemiology	JOE	85. JOE-01
Lancet Global Health	LGH	86. LGH-01, 87. LGH-02, 88. LGH-03, 89. LGH-04, 90. LGH-05, 91. LGH-06, 92. LGH-07
Lancet Public Health	LPH	93. LPH-01, 94. LPH-02, 95. LPH-03, 96. LPH-04, 97. LPH-05, 98. LPH-06, 99. LPH-07
Reproductive Health	RH	100. RH-01

Abbreviation: RA, research article.

TABLE 4 Lexical data of the compiled big textual data

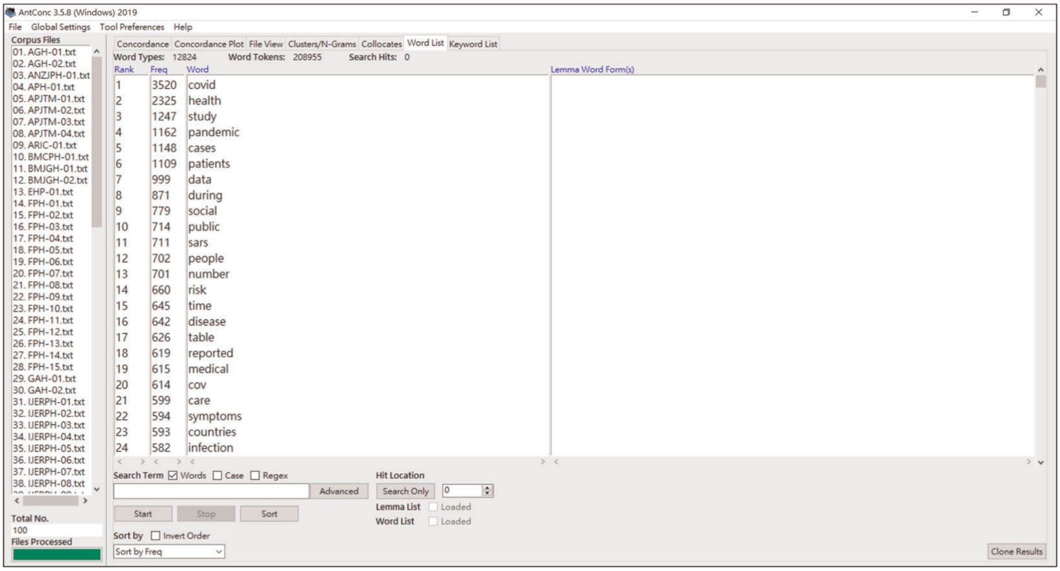
Compiled big textual data				
Data codename	**Numbers of paper**	**Word types**	**Tokens**	**TTR**
AGH	2	1543	7647	0.2018
ANZJPH	1	683	1907	0.3582
APH	1	695	3153	0.2204
APJTM	4	1680	9062	0.1854
ARIC	1	394	989	0.3984
BMCPH	1	731	3108	0.2352
BMJGH	2	2130	10,730	0.1985
EHP	1	868	3333	0.2604
FPH	15	5352	50,993	0.1050
GAH	2	1304	6548	0.1991
IJERPH	41	9124	184,639	0.0494
JGH	7	3263	26,739	0.1220
JIPH	4	1699	9554	0.1778
JNMA	2	973	2763	0.3522
JOE	1	865	3773	0.2293
LGH	7	2905	20,091	0.1446
LPH	7	2411	19,153	0.1259
RH	1	857	2720	0.3151
Whole corpus	100	13,062	366,866	0.0356

Abbreviation: TTR, types and tokens ratio.

researchers adopted Chen et al.'s[46] big textual data refining approach to optimize the compiled big textual data; the refined wordlist on the corpus software shows that meaningful words are ranked to the front (see Figure 7). In addition, the data discrepancy showed that word types of refined data decreased by 238 words (i.e., function words), nevertheless, tokens of refined data decreased 157,911 words, which caused a 43% downsizing in the corpus. Moreover, the lexical diversity was enhanced to 0.0614 (see Table 5). Unexpectedly, when facing highly specialized medical RAs, function words also occupied more than 40% of the corpus. To avoid information distortion, the eliminating procedure for function words is inevitable.

Step 4. Ranking tokens based on individual overall frequency criteria.

After optimizing the compiled big textual data, the authors adopted the refined traditional corpus-based computing method[30] to compute the sum of frequency values of each token (see Figure 7), and to find out each token's frequency values in each RA by the Concordance Plot function of the corpus software. In the Concordance Plot, Concordance Hit represents a token's overall frequency values, and Total Plot (with hits) represents how



Corpus Files
01. AGH-01.txt
02. AGH-02.txt
03. ANZJPH-01.txt
04. APH-01.txt
05. APJTM-01.txt
06. APJTM-02.txt
07. APJTM-03.txt
08. APJTM-04.txt
09. ARIC-01.txt
10. BMCPH-01.txt
11. BMJGH-01.txt
12. BMJGH-02.txt
13. EHP-01.txt
14. FPH-01.txt
15. FPH-02.txt
16. FPH-03.txt
17. FPH-04.txt
18. FPH-05.txt
19. FPH-06.txt
20. FPH-07.txt
21. FPH-08.txt
22. FPH-09.txt
23. FPH-10.txt
24. FPH-11.txt
25. FPH-12.txt
26. FPH-13.txt
27. FPH-14.txt
28. FPH-15.txt
29. GAH-01.txt
30. GAH-02.txt
31. IJERPH-01.txt
32. IJERPH-02.txt
33. IJERPH-03.txt
34. IJERPH-04.txt
35. IJERPH-05.txt
36. IJERPH-06.txt
37. IJERPH-07.txt
38. IJERPH-08.txt

Total No. 100
Files Processed

Word Types: 12824 Word Tokens: 208955 Search Hits: 0

Rank	Freq	Word
1	3520	covid
2	2325	health
3	1247	study
4	1162	pandemic
5	1148	cases
6	1109	patients
7	999	data
8	871	during
9	779	social
10	714	public
11	711	sars
12	702	people
13	701	number
14	660	risk
15	645	time
16	642	disease
17	626	table
18	619	reported
19	615	medical
20	614	cov
21	599	care
22	594	symptoms
23	593	countries
24	582	infection

Search Term ☑ Words □ Case □ Regex

Hit Location 0
Lemma List □ Loaded
Word List □ Loaded

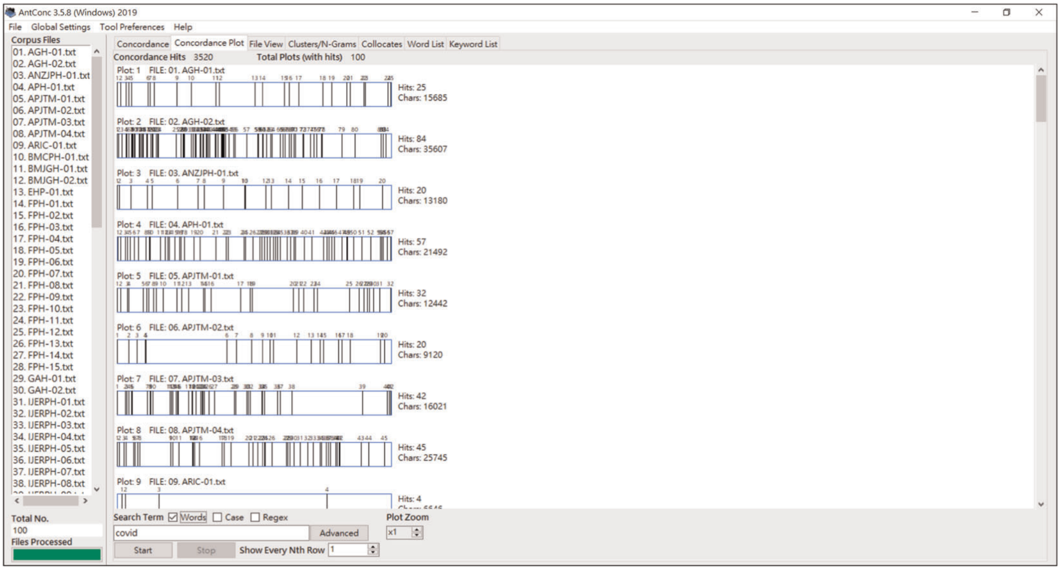
Sort by □ Invert Order
Sort by Freq

FIGURE 7 Refined traditional corpus-based computing method used to rank words [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Data discrepancy between original and refined data

Lexical feature	Original data	Refined data	Data discrepancy
Word types	13,062	12,824	−238 (−1.8%)
Tokens	366,866	208,955	−157,911 (−43%)
TTR	0.0356	0.0614	

Abbreviation: TTR, types and tokens ratio.



Corpus Files
01. AGH-01.txt
02. AGH-02.txt
03. ANZJPH-01.txt
04. APH-01.txt
05. APJTM-01.txt
06. APJTM-02.txt
07. APJTM-03.txt
08. APJTM-04.txt
09. ARIC-01.txt
10. BMCPH-01.txt
11. BMJGH-01.txt
12. BMJGH-02.txt
13. EHP-01.txt
14. FPH-01.txt
15. FPH-02.txt
16. FPH-03.txt
17. FPH-04.txt
18. FPH-05.txt
19. FPH-06.txt
20. FPH-07.txt
21. FPH-08.txt
22. FPH-09.txt
23. FPH-10.txt
24. FPH-11.txt
25. FPH-12.txt
26. FPH-13.txt
27. FPH-14.txt
28. FPH-15.txt
29. GAH-01.txt
30. GAH-02.txt
31. IJERPH-01.txt
32. IJERPH-02.txt
33. IJERPH-03.txt
34. IJERPH-04.txt
35. IJERPH-05.txt
36. IJERPH-06.txt
37. IJERPH-07.txt
38. IJERPH-08.txt

Total No. 100
Files Processed

Concordance Hits 3520 Total Plots (with hits) 100

Plot: 1 FILE: 01. AGH-01.txt Hits: 25 Chars: 15685
Plot: 2 FILE: 02. AGH-02.txt Hits: 84 Chars: 35607
Plot: 3 FILE: 03. ANZJPH-01.txt Hits: 20 Chars: 13180
Plot: 4 FILE: 04. APH-01.txt Hits: 57 Chars: 21492
Plot: 5 FILE: 05. APJTM-01.txt Hits: 32 Chars: 12442
Plot: 6 FILE: 06. APJTM-02.txt Hits: 20 Chars: 9120
Plot: 7 FILE: 07. APJTM-03.txt Hits: 42 Chars: 16021
Plot: 8 FILE: 08. APJTM-04.txt Hits: 45 Chars: 25745
Plot: 9 FILE: 09. ARIC-01.txt Hits: 4

Search Term ☑ Words □ Case □ Regex
covid
Show Every Nth Row 1
Plot Zoom x1

FIGURE 8 Interface of Concordance Plot: *COVID* as an example [Color figure can be viewed at wileyonlinelibrary.com]

many RAs adopted a token. Take *COVID* as an example, its Concordance Hit is 3520 (i.e., overall frequency values) and Total Plot (with hits) is 100 which means *COVID* was adopted by 100 RA authors (see Figure 8). Hence, in this step, the authors obtained three important factors which include overall frequency values, frequency values in each RA, and how many RAs adopted a token. These factors are critical and will be calculated by the H-index algorithm in the following step.

Step 5. Ranking tokens based on H-index algorithm.

In this step, the researchers used the wordlist to compute tokens ($N = 420$) that had frequency values over 100. Take *mortality* as an example, the authors recorded frequency values of *mortality* of each RA as original data, and sorted each frequency from highest to lowest, then it was found that $f(9) \geq 9$; that satisfied the criteria of Equation (3), thus, the value of H-index was given as 9 (see Table 6). This computing approach is used to calculate a token's overall adopting rates and evaluate its importance level more accurately. Then, they recorded tokens' H-index values in Excel software for a ranking process.

It was found that after using the H-index values to rank tokens, the sequences of the wordlist had been changed significantly because H-index calculated authors' adoption rate in each RA and reinterpreted the importance of tokens. However, tokens' H-index values often produced the same value. If the same H-index values are encountered, the authors would sort tokens by their frequency values again. That is, this paper considers H-index and frequency values simultaneously to make the important calculation of tokens more accurate.

Step 6. Integrating tokens' ranking information for future extended applications.

The wordlist of Step 5 showed the combinations of token's H-index and frequency values. The tokens' ranking issue handled by the proposed approach redefine their importance level, hence, these data provide the important referential indicators for future applications, such as IR, NLP, big data analysis, machine learning, deep learning, and so on. By this study, the authors propose a novel corpus-based approach that integrates a corpus software and H-index algorithm to calculate which tokens are important in medical RAs. The resulting data will improve native and EFL medical researchers' learning and processing efficiency of medical RAs.

4.4 | Comparison and discussion

When competitor methods (i.e., the traditional frequency-based approach[30] and the refined traditional frequency-based approach[46]) in handling word-ranking issues only based on words' frequency values or range values, respectively, to determine their sequences, namely, traditional methods do not integrally take a word's dispersion and concentration criteria into account. This deficiency will cause critical word-ranking results exist bias, in addition, the importance levels of high-frequency critical words will be challenged. Hence, to improve the accuracy and efficiency of the big textual data analysis, this section uses the collected COVID-19-related RAs from WOS as the empirical example (i.e., test data) to discuss the difference between the traditional frequency-based approach,[30] the refined traditional frequency-based

TABLE 6 An example of a token's H-index computing process

	Original data		Computing process		
Token	Articles	Frequency	Articles	Frequency	H-index result
Mortality	1	2	6	90	9
	2	1	38	36	
	3	1	20	31	
	4	3	25	30	
	5	2	37	21	
	6	90	30	13	
	7	1	33	12	
	8	3	22	10	
	9	4	15	9	
	10	5	18	6	
	11	1	23	6	
	12	2	10	5	
	13	1	39	5	
	14	2	9	4	
	15	9	31	4	
	16	2	4	3	
	17	1	8	3	
	18	6	26	3	
	19	1	1	2	
	20	31	5	2	
	21	1	12	2	
	22	10	14	2	
	23	6	16	2	
	24	1	2	1	
	25	30	3	1	
	26	3	7	1	
	27	1	11	1	
	28	1	13	1	
	29	1	17	1	
	30	13	19	1	
	31	4	21	1	
	32	1	24	1	
	33	12	27	1	

(Continues)

Token	Original data		Computing process		H-index result
	Articles	Frequency	Articles	Frequency	
	34	1	28	1	
	35	1	29	1	
	36	1	32	1	
	37	21	34	1	
	38	36	35	1	
	39	5	36	1	

approach[46] and the proposed approach (see Table 7). In addition, the top 50 words ranked by the three approaches are also presented to show the discrepancies between them (see Table 8). First, refined corpus data are compared to show which approaches are able to make content words ranked higher. Second, frequency dispersion criteria are compared to show that the proposed approach can compute frequency dispersion criteria, thus, making word-ranking results more accurate. Lastly, calculating frequency concentration criteria is compared to show that the proposed approach can compute frequency concentration criteria, thereby, compensating the blind side of truly defining high-frequency words' importance level.

1. *Refining corpus data*

   According to Table 8, raw data contain many functions and meaningless tokens, such as *the*, *of*, *and*, *to*, *in*, and so forth. The traditional frequency-based approach[30] calculated all tokens' frequency values, it was unable to identify which tokens contain more substantial meanings for humans. To enable the corpus-based approaches to rank critical words with substantial meanings, the refined traditional frequency-based approach[46] and the proposed approach have eliminated function and meaningless words. Hence, based on Table 8, refined data show content words that have general or domain-oriented purposes. It makes corpus analytical results more meaningful and enhances its efficiency in retrieving critical words.

2. *Calculating frequency dispersion criteria*

The authors adopted the proposed approach to compute the top 420 tokens whose frequency values reached more than 100, respectively, from the wordlist of the refined data.

TABLE 7 A comparison of corpus-based approaches

Methods	Refining corpus data	Calculating frequency dispersion criteria	Calculating frequency concentration criteria
The traditional frequency-based approach[30]	No	No	No
The refined traditional frequency-based approach[46]	Yes	No	No
The proposed approach	Yes	Yes	Yes

TABLE 8 The top 50 tokens of the compared three approaches (partial data)

Raw data			Refined data						
The traditional frequency-based approach[30]			The refined traditional frequency-based approach[44]			The proposed approach			
Rank	Frequency	Token	Rank	Frequency	Token	Rank	H-index	Frequency	Token
1	23,079	the	1	3520	COVID	1	39	3520	COVID
2	14,660	of	2	2325	health	2	28	2325	health
3	13,258	and	3	1247	study	3	21	1247	study
4	9577	to	4	1162	pandemic	4	20	1162	pandemic
5	9218	in	5	1148	cases	5	18	1109	patients
6	5721	a	6	1109	patients	6	17	1148	cases
7	3891	with	7	999	data	7	17	871	during
8	3699	for	8	871	during	8	16	999	data
9	3520	COVID	9	779	social	9	15	702	people
10	3279	that	10	714	public	10	14	779	social
11	2857	is	11	711	SARS	11	14	701	number
12	2631	as	12	702	people	12	14	660	risk
13	2544	was	13	701	number	13	14	645	time
14	2417	were	14	660	risk	14	14	642	disease
15	2353	on	15	645	time	15	14	599	care
16	2325	health	16	642	disease	16	13	714	public
17	2084	be	17	626	table	17	13	711	SARS
18	1968	by	18	619	reported	18	13	619	reported
19	1882	this	19	615	medical	19	13	614	CoV

(Continues)

Raw data			Refined data						
The traditional frequency-based approach[30]			The refined traditional frequency-based approach[44]			The proposed approach			
Rank	Frequency	Token	Rank	Frequency	Token	Rank	H-index	Frequency	Token
20	1873	are	20	614	CoV	20	13	594	symptoms
21	1783	from	21	599	care	21	13	593	countries
22	1699	or	22	594	symptoms	22	13	541	one
23	1404	have	23	593	countries	23	13	491	transmission
24	1383	we	24	582	infection	24	12	582	infection
25	1275	not	25	570	population	25	12	570	population
26	1247	study	26	561	participants	26	12	561	participants
27	1213	at	27	546	first	27	12	533	high
28	1177	their	28	541	one	28	12	499	analysis
29	1162	pandemic	29	533	high	29	12	439	clinical
30	1148	cases	30	531	control	30	11	626	table
31	1112	it	31	527	used	31	11	615	medical
32	1111	an	32	526	results	32	11	546	first
33	1109	patients	33	506	based	33	11	526	results
34	999	data	34	499	analysis	34	11	506	based
35	957	more	35	498	case	35	11	498	case
36	921	which	36	491	transmission	36	11	471	information
37	871	during	37	471	information	37	11	470	research
38	861	can	38	470	research	38	11	466	related
39	834	has	39	466	related	39	11	465	higher

Raw data			Refined data						
The traditional frequency-based approach[30]			The refined traditional frequency-based approach[44]			The proposed approach			
Rank	Frequency	Token	Rank	Frequency	Token	Rank	H-index	Frequency	Token
40	814	also	40	465	higher	40	11	456	virus
41	805	these	41	456	virus	41	11	404	age
42	792	p	42	452	studies	42	11	404	associated
43	788	may	43	451	use	43	11	404	confirmed
44	779	social	44	446	two	44	11	355	factors
45	762	been	45	442	coronavirus	45	11	340	model
46	752	they	46	439	clinical	46	10	531	control
47	746	had	47	432	outbreak	47	10	527	used
48	737	who	48	431	measures	48	10	451	use
49	731	all	49	420	CHINA	49	10	446	two
50	731	other	50	406	new	50	10	442	coronavirus

According to Table 8, there were significant differences in token ranking between the traditional corpus-based computing approaches[30,46] and the proposed approach. The traditional corpus-based computing approaches[30,46] only calculated a token's total frequency values to define its rank and importance; however, the frequency dispersion criteria were not taken into consideration; that is, a token with high frequency may not be widely adopted or used by the RA authors, or may be concentrated in very few RAs or even possibly occur in only one RA. Nevertheless, the proposed approach not only used H-index to compute the dispersion and concentration criteria of frequency simultaneously, but also used frequency values to distinguish tokens that had the same H-index values. Therefore, after taking all criteria into considerations, the proposed approach is more rigorous and accurate. Interestingly, tokens, such as *COVID*, *health*, *study*, *pandemic*, *reported*, *infection*, *population*, *participants*, and *case*, still remain in their original ranks when compared with the refined traditional frequency-based approach and the proposed approach; that is, after being calculated using the two approaches, their frequency and H-index values were both extremely high, hence those tokens' importance was unquestionable.

The calculation results of the proposed approach redefine the importance of tokens ($N = 420$) that were compared with the traditional corpus-based computing approaches.[30,46] In other words, the authors found only 11 tokens (2.6%) that remained at original ranks and only nine tokens (2.1%) among them in the top 50 wordlists (see Table 8), 15 tokens (3.5%) that moved forward more than 100 ranks, respectively, 196 tokens (46.6%) that moved forward from 1 to 99 ranks, respectively, 14 tokens (3.3%) that moved backward more than 100 ranks, respectively, and 184 tokens (43.8%) that moved backward from 1 to 99 ranks, respectively. In other words, the proposed approach successfully re-evaluates the importance of tokens and makes more than 97% changes by adopting H-index algorithm which simultaneously took the dispersion and concentration criteria of frequency into consideration (see Table 9).

The nine tokens (2.1%) in the top 50 wordlists indicate that these tokens were extremely critical and they had unquestionable importance rather than the fault of the proposed approach as they showed no differences when compared with the traditional corpus-based computing approaches.[30,46] Those tokens are important because they were adopted by many RA authors and occurred with very high frequency in the compiled big textual data. Moreover, the proposed approach made tokens' sequence moves forward and backward which implicated the traditional corpus-based computing approaches[30,46] caused the distortion when handling token ranking issues. For example, *efforts* were ranked at 349 based on its calculation results of the traditional corpus-based computing approaches[30,46] (frequency = 113), but after being computed by the proposed approach (H-index = 7; frequency = 113), its rank moved forward at 179, that is, it moved forward by 170 sequences. In other words, the importance of *efforts* was promoted, yet, originally, its actual importance level was

TABLE 9 Changes of token ranks ($N = 420$)

Data discrepancy	Token numbers	Proportion
Tokens stay at the original ranks	11	0.0262
Tokens move forward more than 100 ranks	15	0.0357
Tokens move forward from 1 to 99 ranks	196	0.4667
Tokens move backward more than 100 ranks	14	0.0333
Tokens move backward from 1 to 99 ranks	184	0.4381
Tokens' H-index value equal to 1	2	0.0048

underestimated by the traditional corpus-based computing approaches.[30,46] Another instance, *news*, was ranked at 125 based on its calculation results in the traditional corpus-based computing approaches[30,46] (frequency = 229). However, its most occurrence times were concentrated in few RAs, *news* occurred 180 times (78%) only in an RA that was coded as 63. IJERPH-33 in the compiled data. After computing by the proposed approach (H-index = 3; frequency = 229), its rank moved backward to 410, that is, it moved backward by 285 sequences. The data discrepancy indicates that its actual importance level was overestimated by the traditional corpus-based computing approaches.[30,46] The distorted results were caused by the traditional corpus-based computing approaches[30,46] because those did not take tokens' frequency dispersion criteria into consideration, whilst defined tokens' importance level was only based on their total frequency. On the contrary, the proposed approach took tokens' frequency dispersion criteria into consideration, hence, they produce more accurate evaluation results, and define tokens' importance level more precisely.

3. *Calculating frequency concentration criteria*

The proposed approach can also handle tokens' frequency concentration criteria. For example, as discovered, *hyponatremia* was ranked at 231 based on its calculation results in the traditional corpus-based computing approaches[30,46] (frequency = 153), and *tobacco* was ranked at 391 based on its calculation results in the traditional corpus-based computing approaches[30,46] (frequency = 104). Nevertheless, after computing by the proposed approach, both words' H-index values were equal to 1 (see Table 9); hence, their post rank moved backward at 419 and 420, respectively (i.e., they became the last important two words among 420 tokens), they moved backward by 188 and 29 sequences, respectively. Even if *hyponatremia* and *tobacco* had more than 100 occurrence times in the compiled big textual data, they were adopted by only one RA each. In other words, their importance was almost negligible because there is extremely low probability that people will encounter those two words in future COVID-19-related RAs. Therefore, the traditional corpus-based computing approaches[30,46] again overestimated the tokens' importance level.

To conclude this section, tokens' importance level computation has affected the analysis and development of big data management and processing, search engines, and other relative AI industries. If the frequency value is the only criteria for ranking tokens' importance level, the assessment of their importance will be inaccurate and distorted. Hence, we proposed the novel corpus-based approach in this paper, which integrates a corpus software and H-index algorithm to take tokens' frequency dispersion and concentration criteria into consideration simultaneously, thus, accurately and comprehensively handling the token ranking issue.

5 | CONCLUSION

Traditional corpus-based computing methods still present some analytical doubts during corpus processing, for example, refining corpus data, computing frequency dispersion criteria, and computing frequency concentration criteria. Those may cause a decrease in corpus data processing efficiency, and more seriously, the evaluation of tokens' importance level may be biased as frequency value is the only indicator used for handling word-ranking issues in traditional corpus-based computing methods. Thus, to compensate the blind side of the traditional methods, this paper proposed a novel corpus-based approach that integrates a corpus software and H-index algorithm to refine corpus data, to calculate tokens' frequency dispersion and concentration criteria, and further to handle word-ranking issues.

The significant contributions of the proposed approach are listed as: (1) the proposed approach is able to refine corpus data via machine processing to eliminate function and meaningless words, (2) the proposed approach is able to compute tokens' frequency dispersion criteria; moreover, when facing tokens with the same H-index values, tokens' frequency values are the second criteria used to rank, hence, it makes word-ranking process more accurate and to avoid hesitance situations occurring in the ranking process, (3) the proposed approach is able to compute tokens' frequency concentration criteria, such as in cases where a token has high-frequency values but is over-concentrated in certain RAs; hence, H-index = 1 indicates that H-index algorithm precisely evaluates a token's importance level, whilst, frequency values overestimate a token's importance level and cause ranking results distortion. Furthermore, in relation to textual analysis in COVID-19-related RAs, the proposed approach also helps native and EFL frontline healthcare personnel to integrate and retrieve professional medical knowledge, and to further enhance their information processing efficiency.

This paper exists a major limitation that is waiting for future researches to overcome, for example, without the assistant of existing software, H-index computing process still relies on human processing, once the data are too bounteous, it will cause a great burden on data analysts. Hence, in terms of future perspective, this paper suggests that future corpus-based and NLP research can import H-index algorithm to corpus program (i.e., software) for processing big textual data. It will enhance accuracy and efficiency in handling word-ranking issues, and aid accurate retrieval of critical words from the big textual data.

ACKNOWLEDGMENTS

The authors would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this study under Contract Nos. MOST 108-2410-H-145-001 and MOST 109-2410-H-145-002.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

ORCID

Liang-Ching Chen http://orcid.org/0000-0002-7896-1990
Kuei-Hu Chang http://orcid.org/0000-0002-9630-7386

REFERENCES

1. Barshandeh S, Haghzadeh M. A new hybrid chaotic atom search optimization based on tree-seed algorithm and Levy flight for solving optimization problems. *Eng Comput*. 2020. https://doi.org/10.1007/s00366-020-00994-0
2. Barshandeh S, Piri F, Sangani SR. HMPA: an innovative hybrid multi-population algorithm based on artificial ecosystem-based and Harris Hawks optimization algorithms for engineering problems. *Eng Comput*. 2020. https://doi.org/10.1007/s00366-020-01120-w
3. Chang CF, Kuo CH. A corpus-based approach to online materials development for writing research articles. *Engl Specif Purp*. 2011;30(3):222-234.
4. Gholami J, Zeinolabedini M. Peer-to-peer prescriptions in medical sciences: Iranian field specialists' attitudes toward convenience editing. *Engl Specif Purp*. 2017;45:86-97.
5. Grabowski L. Keywords and lexical bundles within English pharmaceutical discourse: a corpus-driven description. *Engl Specif Purp*. 2015;38:23-33.
6. Le CNN, Miller J. A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes. *Engl Specif Purp*. 2020;58:102-121.

7. Gorbalenya AE, Baker SC, Baric RS, et al. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol*. 2020;5(4):536-544.
8. Guan W, Ni Z, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708-1720.
9. Lu RJ, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565-574.
10. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. *Lancet*. 2020;395(10223):470-473.
11. Wolfel R, Corman VM, Guggemos W, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature*. 2020;581(7809):465-469.
12. Yi Y, Lagniton PNP, Ye S, Li EQ, Xu RH. COVID-19: what has been learned and to be learned about the novel coronavirus disease. *Int J Biol Sci*. 2020;16(10):1753-1766.
13. Forrest JI, Rayner CR, Park JJH, Mills EJ. Early treatment of COVID-19 disease: a missed opportunity. *Infect Dis Ther*. 2020;9(4):715-720.
14. Huff C. Covid-19: Americans afraid to seek treatment because of the steep cost of their high deductible insurance plans. *BMJ—Br Med J*. 2020;371:m3860.
15. Cheng X, Cao Q, Liao SS. An overview of literature on COVID-19, MERS and SARS: using text mining and latent Dirichlet allocation. *J Inf Sci*. 2020;2020:0165551520954674.
16. Glowacki EM, Wilcox GB, Glowacki JB. Identifying #addiction concerns on Twitter during the COVID-19 pandemic: a text mining analysis. *Subst Abus*. 2021;42(1):39-46. https://doi.org/10.1080/08897077.2020.1822489
17. Gao L, Yu SB, Li MH, Shen ZS, Gao ZY. Weighted h-index for identifying influential spreaders. *Symmetry—Basel*. 2019;11(10):1263.
18. Hauer MP, Hofmann XCR, Krafft TD, Zweig KA. Quantitative analysis of automatic performance evaluation systems based on the h-index. *Scientometrics*. 2020;123(2):735-751.
19. Hirsch JE. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA*. 2005;102(46):16569-16572.
20. Hu GY, Wang L, Ni R, Liu WS. Which h-index? An exploration within the Web of Science. *Scientometrics*. 2020;123(3):1225-1233.
21. Kung SC, Chien TW, Yeh YT, Lin JCJ, Chou W. Using the bootstrapping method to verify whether hospital physicians have different h-indexes regarding individual research achievement a bibliometric analysis. *Medicine (Baltimore)*. 2020;99(33):e21552.
22. Li XL, Zhang P, Zhu GY. Measuring method of node importance of urban rail network based on h index. *Appl Sci—Basel*. 2019;9(23):5189.
23. Pluskiewicz W, Drozdzowska B, Adamczyk P, Noga K. Scientific quality index: a composite size-independent metric compared with h-index for 480 medical researchers. *Scientometrics*. 2019;119(2):1009-1016.
24. Hanna TN, Duszak R, Chahine A, Zygmont ME, Herr KD, Horny M. The introduction and development of the H-index for imaging utilizers: a novel metric for quantifying utilization of emergency department imaging. *Acad Emerg Med*. 2019;26(10):1125-1134.
25. O'Keeffe A, McCarthy M, Carter R. *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press; 2007.
26. Motschenbacher H. Corpus linguistic onomastics: a plea for a corpus-based investigation of names. *Names*. 2020;68(2):88-103.
27. Seracini FL. Phraseology in multilingual EU legislation: a corpus-based study of translated multi-word terms. *Perspect—Stud Transl*. https://doi.org/10.1080/0907676X.2020.1800058
28. Gholaminejad R, Sarab MRA. Academic vocabulary and collocations used in language teaching and applied linguistics textbooks a corpus-based approach. *Terminology*. 2020;26(1):82-107.
29. Zhang XM, Kotze H, Fang J. Explicitation in children's literature translated from English to Chinese: a corpus-based study of personal pronouns. *Perspect—Stud Transl*. 2020;28(5):717-736.
30. Anthony L. *AntConc* (*Version 3.5.8*). Corpus Software. 2019. https://www.laurenceanthony.net/software/antconc/
31. Scott M. PC analysis of key words—and key key words. *System*. 1997;25:233-245.

32. Paules CI, Marston HD, Fauci AS. Coronavirus infections—more than just the common cold. *JAMA—J Am Med Assoc*. 2020;323(8):707-708.
33. Wan YS, Shang J, Graham R, Baric RS, Li F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J Virol*. 2020;94(7):e00127-20.
34. Zhu N, Zhang DY, Wang WL, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727-733.
35. Ye Z, Zhang Y, Wang Y, Huang ZX, Song B. Chest CT manifestations of new coronavirus disease 2019 (COVID-19): a pictorial review. *Eur Radiol*. 2020;30(8):4381-4389.
36. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, et al. A machine learning model to identify early stage symptoms of SARS-CoV-2 infected patients. *Expert Syst Appl*. 2020;160:113661113661.
37. Jones RM. Relative contributions of transmission routes for COVID-19 among healthcare personnel providing patient care. *J Occup Environ Hyg*. 2020;17(9):408-415.
38. Morawska L, Cao JJ. Airborne transmission of SARS-CoV-2: the world should face the reality. *Environ Int*. 2020;139:105730.
39. Setti L, Passarini F, De Gennaro G, et al. Airborne transmission route of COVID-19: why 2 meters/6 feet of inter-personal distance could not be enough. *Int J Environ Res Public Health*. 2020;17(8):2932.
40. Czeisler ME, Garcia-Williams AG, Molinari NA, et al. Demographic characteristics, experiences, and beliefs associated with hand hygiene among adults during the COVID-19 pandemic—United States, June 24–30, 2020. *MMWR—Morb Mortal Wkly Rep*. 2020;69(41):1485-1491.
41. Lee CY, Wang PS, Huang YD, Lin YC, Hsu YN, Chen SC. Evacuation of quarantine-qualified nationals from Wuhan for COVID-19 outbreak—Taiwan experience. *J Microbiol Immunol Infect*. 2020;53(3):392-393.
42. Peak CM, Kahn R, Grad YH, et al. Individual quarantine versus active monitoring of contacts for the mitigation of COVID-19: a modelling study. *Lancet Infect Dis*. 2020;20(9):1025-1033.
43. Dang TNY. High-frequency words in academic spoken English: corpora and learners. *ELT J*. 2020;74(2): 146-155.
44. Kempen G, Harbusch K. Mutual attraction between high-frequency verbs and clause types with finite verbs in early positions: corpus evidence from spoken English, Dutch, and German. *Lang Cogn Neurosci*. 2019; 34(9):1140-1151.
45. Jelodar H, Wang YL, Orji R, Huang SC. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE J Biomed Health Inform*. 2020;24(10):2733-2742.
46. Chen LC, Chang KH, Chung HY. A novel statistic-based corpus machine processing approach to refine a big textual data: an ESP case of COVID-19 news reports. *Appl Sci—Basel*. 2020;10(16):5505.
47. Lin HX, Wang XT, Huang ML, et al. Research hotspots and trends of bone defects based on Web of Science: a bibliometric analysis. *J Orthop Surg Res*. 2020;15(1):463.
48. Carmona-Serrano N, Lopez-Belmonte J, Cuesta-Gomez JL, Moreno-Guerrero AJ. Documentary analysis of the scientific literature on autism and technology in Web of Science. *Brain Sci*. 2020;10(12):985.
49. Li ZQ, Poon H, Chen W, Fan JT. A comparative analysis of textile schools by journal publications listed in Web of Science (TM). *J Text Inst*. https://doi.org/10.1080/00405000.2020.1824434

How to cite this article: Chen L-C, Chang K-H. A novel corpus-based computing method for handling critical word-ranking issues: An example of COVID-19 research articles. *Int J Intell Syst*. 2021;36:3190–3216. https://doi.org/10.1002/int.22413