symmetry

MDPI

Article

A Keyword-Based Literature Review Data Generating Algorithm—Analyzing a Field from Scientific Publications

Junchao Wang *,†, Guodong Su †, Chengrui Wan, Xiwei Huang and Lingling Sun *

Key Laboratory of RF Circuits and Systems, Ministry of Education, and Zhejiang Provincial Laboratory of Integrated Circuit Design, Hangzhou Dianzi University, Hangzhou 310018, China; guodong@hdu.edu.cn (G.S.); 16042033@hdu.edu.cn (C.W.); huangxiwei@hdu.edu.cn (X.H.)

* Correspondence: junchao@hdu.edu.cn (J.W.); sunll@hdu.edu.cn (L.S.)

† These authors contributed equally to this work.

Received: 5 April 2020; Accepted: 8 May 2020; Published: 1 June 2020

Abstract: A scientific review is a type of article that summarizes the current state of a specific field, which is crucial for promoting the advancement of our science community. Authors need to read hundreds of research articles to prepare the data and insights for a comprehensive review, which is time-consuming and labor-intensive. In this work, we present an algorithm that can automatically extract keywords from the meta-information of each article and generate the basic data for review articles. Two different fields—communication engineering, and lab on a chip technology—were analyzed as examples. We first built an article library by downloading all the articles from the target journal using a python-based crawler. Second, the rapid automatic keyword extraction algorithm was implemented on the title and abstract of each article. Finally, we classified all extracted keywords into class by calculating the Levenshtein distance between each of them. The results demonstrated its capability of not only finding out how communication engineering and lab on a chip were evolved in the past decades but also summarizing the analytical outcomes after data mining of the extracted keywords. Our algorithm is more than a useful tool for researchers during the preparation of a review article, it can also be applied to quantitatively analyze the past, present and help authors predict the future trend of a specific research field.

Keywords: data mining; automation; natural language processing; keyword extraction; scientific review; big data

1. Introduction

With the development of advanced science and emerging technologies, more and more interdisciplinary fields have become the frontiers of scientific research, which are crucial to the welfare of all human beings. For instance, virtual reality (VR) is currently now widely used in surgical training programs for medical students by breaking the boundary between the real world and computer simulations [1–4]. Artificial intelligence (AI), Internet of Things (IoT) and big data are helping radiologists to develop deep neural networks for classification, detection, and segmentation tasks of different diseases that are threatening the health of millions of patients [5,6]. Researchers and engineers from Tissue engineering [7], Genetic engineering [8], Bioinformatics [9], Biological systems engineering [10], Biotechnology [11] and other fields are developing usable, tangible or economically viable Bioengineering products such as bionic eye [12], supramolecular biomaterials [13], among others, for the benefit of all people.

However, the difficulty of these research projects is significantly increased due to the need of researchers and engineers from multiple backgrounds to work closely for achieving the same

pivotal goal. For example, a computer engineer who is developing a deep-learning-based cancer classifier should at least have a basic understanding of how radiologists are detecting cancer using computer tomography (CT) [14,15]. To develop a feasible bionic eye, electronic engineers shall reach tissue engineers for help to make the bionic eye bio-compatible and minimize rejection reaction [12]. These interdisciplinary research projects create an additional barrier for researchers who are not experts in one field but willing to take advantage of the scientific achievement created by others in their own research.

One way to overcome this barrier is to read the latest review article of a specific field to have a preliminary comprehension of the corresponding research progress. A review article (or so-called survey article) is an article that summarizes the current state of understanding on a topic, discusses the recent major advances and discoveries in the corresponding field, and gives ideas of where research might go next to readers [16]. However, there are three issues of the current production of review articles. First, review articles in a certain field are not reported every year. A recently-published review article could provide a comprehensive understanding of the latest research achievements. However, in other cases, the review articles we can find are outdated to a certain extent. Second, preparing the data for a review article is time-consuming and labor-intensive. As shown in Figure 1, authors (usually the experts in a field) would collect the results and data from hundreds of peer-reviewed articles and summarize key points from each of these articles [17]. Thus, even preparing a single data table for a review article could take weeks or months. After that, authors need to find the indications behind these results and data in order to make their own prediction or comments for a certain field. Compared with drafting the manuscript, conducting a thorough literature survey is a much more painful part before submission to the journal. Last but not least, the number of published research articles is increasing tremendously. In 2018, one of the largest research article publisher, Elsevier, has published more than 470,000 articles in 2500 journals [18]. Thus, summarizing the results from hundreds of peer-reviewed articles could only explore a limited fraction of a specific field.

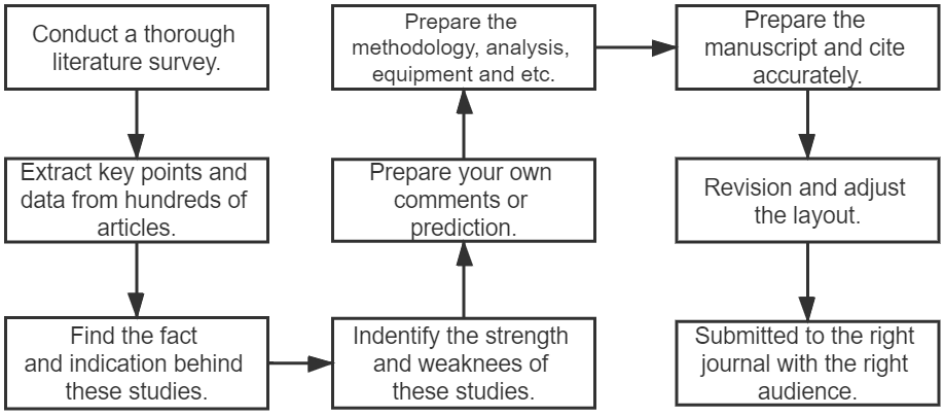

Figure 1. Guidelines for writing a Review Article.

Fortunately, big data analysis has been widely applied in scientific research to find the cause and effect behind these large amounts of numbers [19–22]. Therefore, we are motivated by the question, is it possible that a non-expert researcher could summarize what a field has happened in the past few years and what the major accomplishment of the corresponding field is? More specifically, is it possible for us to generate the data for the review article automatically? In this work, we present an algorithm that generate the data to be further used in a review article by leveraging automatic keyword extraction and similarity calculation. As an example of our algorithm, we summarized the major accomplishments of two different fields, communication engineering and lab on a chip, from two scientific journals, *IEEE transaction on communications* and *Lab on a Chip*, respectively.

Existed Guidelines for Preparing the Literature Review Article

Researchers have developed mature guidelines for writing different kinds of review articles in various fields. Machi et al. provided a classic six-step of how to write a review article [23]. Wee et al. discussed the significance of why and how to write a literature review paper, and focused on the importance of adding value, rather than only providing an overview [17]. Torraco focused on how to organize and write an integrative literature review and discussed the contributions of review articles to the knowledge base of human resource development [24]. Randolph not only summarized the conventional steps for writing a review article but also provided a framework for the self-evaluation [25]. Boote et al. suggested criteria to evaluate the quality of literature reviews for the dissertation of PhD and Master's students [26]. Denney et al. provided the structure, process, and art of writing a review article for both undergraduate and graduate students in the criminal justice field [27]. Levy et al. introduced a framework for conducting and writing an effective literature review in the Information Systems field [28]. Aveyard structured the guidelines of writing a review article by asking ourselves eight questions in the health and social care field [29]. Piper focused on how to write a systematic literature review in the medical field [30].

2. Materials and Methods

The major steps of our proposed method were summarized as follows. (1) Establish the journal library of a target field; (2) Keyword extraction process from title and abstract; (3) Calculate the similarity between different keywords and classify them into proper groups; (4) Analyze the overall development of a target field from post-processed data.

We first used a python-based web crawler to download the published articles from target peer-reviewed journals (Please refer to the supplementary information for details). In this work, we chose two different peer-reviewed journals as examples to illustrate our algorithm. The first journal we chose is *IEEE transaction on communications*, which is one of the top journals in the field of communication engineering [31]. The second journal we chose is *Lab on a chip*, which is one of the top journals in the field of devices and applications at the microscale and nanoscale [32].

During the crawling process, we stored the title, abstract and the published year of each article into a MySQL database (Please refer to the *supplementary information* for details). After that, the rapid automatic keyword extraction (RAKE) [33] algorithm was implemented to extract keywords from both the title and abstract. RAKE is a well-known keyword extraction method that uses a list of stop-words and phrase delimiters to detect the most relevant words or phrases in a piece of text. The pseudocode describing the corresponding process was summarized in Algorithm 1. We set the parameters of RAKE to extract two keywords from the title and three keywords from the abstract. We found that extracted two words from the title is good enough to represent the main novelty and focus of one study. In contrast, only extracting three keywords from the abstract could prevent including unrelated keywords as abstract usually would provide more details of one study than the title. After keyword extraction, each keyword was stored in the MySQL database as well.

Algorithm 1 Extract keywords from target text.

Require: Target text, T; The number of extracted keywords, n;
Ensure: The keyword list, $W = [w_0, w_1, ..., w_n]$;
1: r ← Rake(); ▷ Initialize a RAKE object.
2: W ← r.get_ranked_phrases(T); ▷ Extract keywords with customized weights.
3: **return** W;

The next procedure was to determine the similarity between different keywords by calculating Levenshtein distance (LD) [34] and classify them as various classes with a certain meaning. The pseudocode describing the corresponding process was summarized in Algorithm 2. At the beginning of this part of the algorithm, we chose the first keyword as w_i in the functioning ratio

(line 11), and fetch another keyword in the word set as w_k, then calculated the LD between two words. If the LD was larger than the threshold (T_s), then we classified these two words into the same group (an array). If not, we moved to the next keyword and assigned it as w_k and calculate the LD again. The threshold (T_s) here was a critical parameter because if the value is too big, the algorithm might classify two highly related keywords into a different word group. If the value is too small, the algorithm might be over-optimized and put unrelated keywords into the same group. In our algorithm, we tried several different values of T_s, and find $T_s = 0.6$ was a proper value to make our algorithm work. After traversing all the other keywords in the word set, we replaced w_i with another unclassified keyword and repeat the steps above until all the keyword had been classified into a proper group. Until now, we had a series of similar groups that contains the related keywords or phrases. After that, the occurrence of the keywords was calculated by going through a searching process of all the post-classified keywords. Finally, we could use the listed data to draw the developing tendency of a certain discipline in a specific period and find how one field was evolved and advanced.

Algorithm 2 Find related keywords groups by calculating the Levenshtein distance.

Require: The keyword list, $W = [w_0, w_1, ..., w_n]$; Threshold used to define similarity, T_s;
Ensure: Different related keywords groups, $S_0, S_1..., S_j$;

1: $i \leftarrow 0$ ▷ Initialize the index for the main loop.
2: $j \leftarrow 0$ ▷ Initialize the index for different related keyword groups.
3: $W_{used} \leftarrow list()$ ▷ Define a list to store used keywords.
4: **while** $(i \leq n)$ **do**
5: **if** $w_i \notin W_{used}$ **then** ▷ Continue if w_i has not been used.
6: $W_{used}.append(w_i)$ ▷ Add w_i into the used keyword list.
7: $G_j \leftarrow list()$ ▷ Initialize a new list for a related keywords group.
8: $G_j.append(w_i)$ ▷ Append the compared word into G_j.
9: **for** $k = i + 1 \rightarrow n$ **do**
10: **if** $w_k \notin W_{used}$ **then** ▷ Continue if w_k has not been used.
11: $LD \leftarrow Levenshtein(w_i, w_k)$ ▷ Calculate the similarity between w_i and w_k.
12: **if** $LD > T_s$ **then** ▷ Continue if the similarity is larger than T_s.
13: $G_j.append(w_k)$ ▷ Add related keyword into the same group.
14: $W_{used}.append(w_k)$ ▷ Add the keyword into used keyword list.
15: **end if**
16: **end if**
17: **end for**
18: $j \leftarrow j + 1$
19: **end if**
20: $i \leftarrow i + 1$
21: **end while**
22: **return** $S_0, S_1, ..., S_j$;

3. Results

Figure 2 summarized the development of communication technologies from 2001 based on keyword extraction from *IEEE transaction on communications*. Figure 2A showed the three stages of development of communication technologies, Information processing, Higher-level communication technologies and Advanced integration & IoT. Figure 2B depicted the trend of how three stages were evolved.

Table 1 listed the comparison between algorithm-generated keywords and author-selected keywords. *Author-selected keywords* were provided by the authors or editors when articles were published in *IEEE transaction on communications*. We could verify the performance and robustness of our algorithm in the keyword extraction aspect. In general, over 71% algorithm-generated keywords were included in author-selected keywords, which indicated that our algorithm-generated could imitate or replace the keyword-selection process to a certain extent.

Table 1. Comparison between algorithm-generated and author-selected keywords.

Keywords (Occurrence)	Included in Author Keywords	Not Included in Author Keywords
Channel (1026)	channel estimation error, imperfect channel state information, wireless channels, channel uncertainty, fading channels, regularized channel inversion	mimo channel estimation, cdma channel estimates, multichannel random access, ideal channel reciprocity, channel shortening detection
Network (914)	ad hoc networks, elastic optical networks, optical networks, random networks, cognitive radio networks, heterogeneous cellular networks, clustered wireless sensor networks	multiuser multiple antenna relaying networks, antenna relay networks, multistage switching networks, device mobile network, energy harvesting networks
Communication (702)	distributed communication, molecular communication, narrowband powerline communications, visible light communications, machine type communications, millimeter wave communication systems	green wireless communications, field passive rfid communication, digital communication systems, digital processing enabled massive mimo communications
System (436)	ofdm systems, uwb systems, multiuser mimo systems, cellular systems, multicarrier systems, communication systems, interference coupled wireless systems	mimo spatial multiplexing systems, joint time frequency spreading systems, time coded mimo systems, cdma cellular system
Codes (392)	convolutional codes, lattice codes, nonlinear codes, linear block codes, rotationally invariant codes, linear block precoding, rateless codes, ldpc codes	correcting codes, restricted codes, matched spreading codes, adaptive turbo codes,multilayer codes
Allocation (384)	power allocation, resource allocation, optimal power allocation, green energy allocation, spectrum allocation, antenna allocation, bandwidth allocation	constrained power allocation, femtocell allocation, efficient uplink resource allocation, dynamic backhaul resource allocation, decentralized energy allocation, structured spectrum allocation
Wireless (313)	heterogeneous wireless networks, clustered wireless sensor networks, wireless power transfer, wireless energy harvesting, selfish wireless networks	multicarrier optical wireless communication channel, green wireless communications, vehicle wireless communication channel, underwater wireless optical communication links
Coding and decoding (297)	channel coding, source coding, turbo decoding, iterative decoding, soft decoding, multiple description coding	stack decoding, siso decoding, metric multiuser decoding, spectrum domain decoding, suboptimal siso decoding
Estimation (282)	channel estimation, frequency estimation, frequency offset estimation	noncoherent sequence estimation, likelihood sequence estimation, coarse frequency offset estimation
Detection (259)	multiuser detection, detection, noncoherent detection, iterative detection, differential detection, signal activity detection	likelihood detection, carrier detection, synchronized detection, continuous error detection, adaptive iterative detection, multiple symbol differential detection
Analysis (240)	performance analysis, interference analysis, error rate analysis, convergence analysis, delay analysis, capacity analysis, queueing analysis	trapping set analysis, ber performance, asymptotic error rate analysis, markov analysis, collision analysis, uplink capacity analysis
Cooperative (239)	cooperative routing, cooperative transmission, cooperative diversity, cooperative networks, cooperative spectrum sensing, cooperative beamforming	performance cooperative demodulation, forward cooperative diversity, cooperative design, hop cooperative diversity network, scale cooperative broadcast network
Transmission (207)	optimal transmission range, cooperative transmission, supervised transmission, transmission capacity, rateless transmission, transmission strategy	fde block transmission, ofdma transmission system, carrier block transmission, data packet mobile downlink transmission, transmission subspace tracking
Modulation (183)	coded modulation, turbo decoding, differential modulation, continuous phase modulation, adaptive modulation	nonorthogonal multipulse modulation, multiuser demodulation, multilevel coded modulation, feedback differential demodulation, adaptive modulation

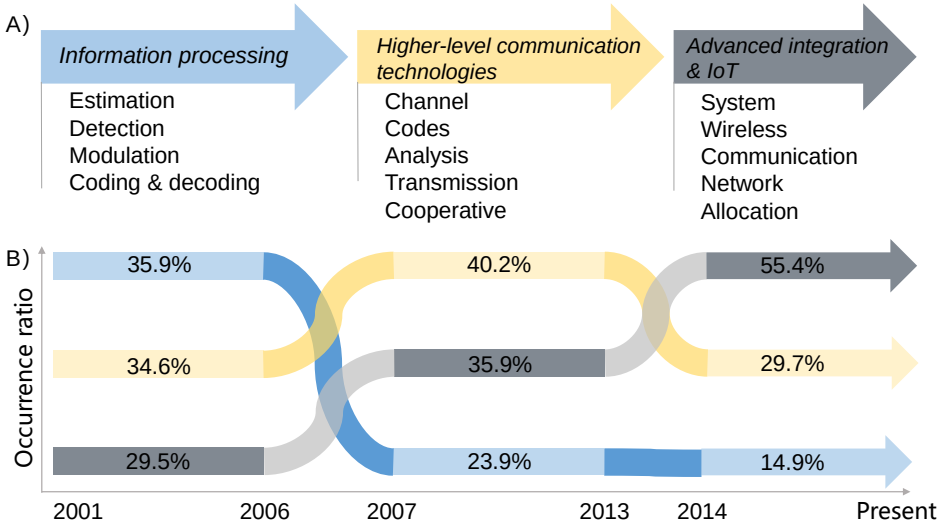

Figure 2. Development of communication technologies from 2001 based on keyword extraction from *IEEE transaction on communications*. (**A**) The three stages of development trend. Keywords of Information processing, Higher-level communication technologies and Advanced integration and Internet of Things (IoT) were showed under blue, yellow and gray arrows, respectively. (**B**) The occurrence ratio of extracted keywords in the field of Information processing, Higher-level communication technologies and Advanced integration and IoT along with time. The x-axis showed the timeline of communication technology development, while the y-axis showed the rank of three stages based on their corresponding keywords occurrence ratio.

Figure 3 summarized the development of Lab on a Chip technologies from 2003 based on keyword extraction from *Lab on a Chip*. Figure 3A showed the three stages of development of LoC technologies, which are Early days of miniaturization, Growth of LoC technologies and Rise of LoC applications. Figure 3B depicted the trend of how three stages were evolved.

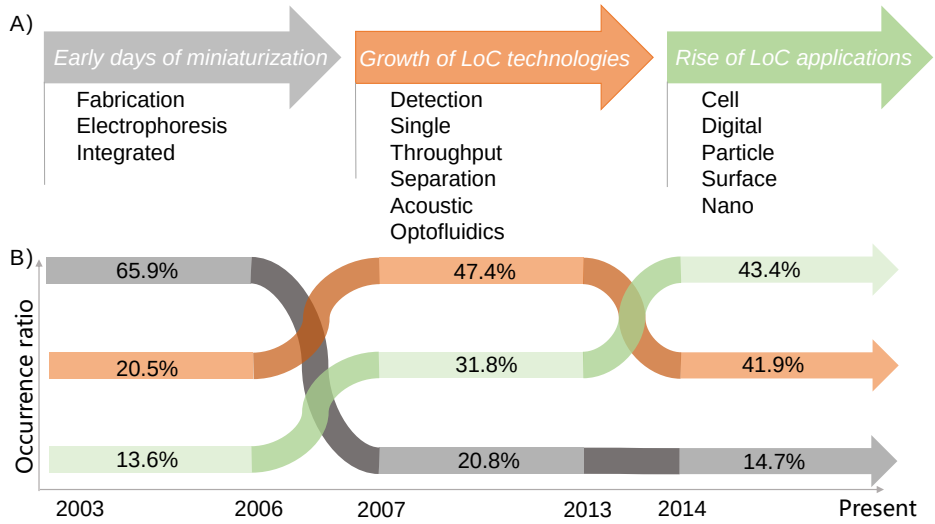

Figure 3. Development of Lab on a Chip (LoC) technologies from 2003 based on keyword extraction from *Lab on a Chip*; (**A**) The three stages of development trend. Keywords of Early days of miniaturization, Growth of LoC technologies and Rise of LoC applications were showed under gray, red and green arrows, respectively. (**B**) The occurrence ratio of extracted keywords in the field of Early days of miniaturization, Growth of LoC technologies and Rise of LoC applications along with time. The x-axis showed the timeline of LoC development, while the y-axis showed the rank of three stages based their corresponding keywords occurrence ratio.

Further details of the top extracted keywords and their occurrence from *Lab on a Chip* were showed in Table 2. We have listed those keywords and their specific sub-keywords to demonstrate how those keywords can be used to provide information for a scientific review. For example, the keyword *detection* may refer to seven different sub-keywords, electrical detection, cell detection, biological detection, sensitive detection, amperometric detection, fluorescence spectrum detection, fluorimetric lead detection, which were directly extracted by our algorithm.

Table 2. Extracted keywords and sub-keywords from *Lab on a Chip*.

Keywords (Occurrence)	Sub-keywords
Detection (68)	electrical detection, cell detection, biological detection, sensitive detection, amperometric detection, fluorescence spectrum detection, fluorimetric lead detection
Cell (66)	cell lysis cell analysis, cell genetic analysis, whole cell array, rare cell isolation , chip cell culture, cell type interactions
Single (39)	single particles, single cell, single cell immobilisation, single dna molecules, single stem cells, single cell isolation
Particle (33)	wavelength particle trapping, inertial particle focusing, particle manipulation, particle inertial migration, efficient particle separation, particle tracking
Throughput (30)	high throughput screening devices, high throughput drug testing, high throughput separation, throughput tracking throughput rheology, throughput diffusion
Separation (22)	multidimensional separations, multiway separation, electrokinetic separation electrophoretic separation, chemical separations, 2D separations
Optofluidic (21)	optofluidic lasers, optofluidic sers chip, droplet optofluidic imaging, optofluidic imaging system, optofluidic waveguide, optofluidic chip, optofluidic ultrahigh
Nano (17)	nanoparticle assay, nanoparticles, nanoparticle separation
Surface (15)	surface acoustic waves, standing surface acoustic wave microfluidics, surface acoustic wave driven fluid motion, surface acoustic wave enabled pipette
Digital (13)	single digital microfluidic reactor chip, digital microfluidic chip, digital diffraction detection, digital biology, thermal digital microfluidic device, magnetic digital microfluidics
Fabrication (10)	nanofabrication, design fabrication, fabrication platform, situ fabrication, rapid fabrication
Integrated (10)	chip integrated system, integrated microsystem, integrated disposable dye clad leaky waveguide sensor, integrated optical leaky waveguide sensor, integrated microfluidic processor, integrated microfluidic device, integrated microfluidic uv absorbance detector
Acoustic (10)	acoustic radiation force, acoustic streaming, acoustic trapping, driven acoustic streaming
Electrophoresis (9)	capillary electrophoresis, microchip electrophoresis, microscale capillary electrophoresis, two-dimensional capillary gel electrophoresis, dielectrophoresis

4. Discussion

The first field we investigated was communication engineering and we chose *IEEE transaction on communications* to test our algorithm. All the articles published in *IEEE transaction on communications* had listed a few keywords which were provided by the authors or editors. Thus, we could verify the keyword extraction performance by comparing algorithm-extracted keywords with author provided keyword. Totally 11,199 articles from 1984 to 2019 were crawled and processed. And we specifically analyzed the keywords in the period from 2001 to 2019.

According to the metrics of the presence of keywords, we can see a three-stage distribution of keywords in the field of communications in Figure 2. The three stages were *Information processing* (blue arrow), *Higher-level communication technologies* (yellow arrow) and *Advanced integration & IoT* (gray arrow). And we summarized all the 13 keywords listed in Figure 2A as well as the corresponding sub-keywords which were highly consistent with author-provided keywords and algorithm-provided keywords in Table 1. The x-axis of Figure 2B showed the timeline of communication technology development, while the y-axis showed the rank of three stages based on their corresponding keywords occurrence ratio. The figure demonstrated clearly that Information processing was the most popular

research topic at the first stage of communication technology, but with the further development of basic knowledge around 2006, the rise of higher-level communication technologies became more attractive from 2007 to 2013 and researches on channels, codes and so on has become the key focus. During the analyzed 12 years of development, Advanced integration & IoT had steady growth and gradually showed its importance. With the growing need of higher-level technologies and the application of internet of things, the stage of Advanced integration & IoT has gathered increased research focus and became the most popular research topic in the field of communication engineering staring 2014.

In the first stage (2001 to 2006), the most popular topics were channel coding, decoding, signal modulation, channel estimation, and signal detection, which were fundamental research topics of communications. In the second stage (2007 to 2013), keywords like codes, cooperative, channel and analysis showed up the most times. When referring to the meaning or details behind these keywords, we could find that different types of codes were investigated and the corresponding standards were established in this period. For instance, low-density parity-check codes, time-spatial codes, and some polar codes were popular research topics. The keyword *cooperative* included multi-user cooperative communications, cooperative relay system, cooperative beamforming technology, and cooperative beamforming wireless system. Besides, there was a lot of research about signal channels. Keywords related to the channel included fading channel, non-linear channel, Multi-input Multi-output (MIMO) channel and mixed channel. In addition, keyword *analysis* was important in all three stages and especially in the second stage. Capacity, error, delay and timing-jitter analysis made up the content of the analysis of the keyword. Compared with the previous period, researchers moved forward to do higher-level research and tried to analyze the performance of previously developed techniques in communications. The third stage was from 2012 to the present. In this stage, the advanced integration of different kinds of network and internet of things became the focus in the community. The number of research topics related to MIMO has increased significantly, including millimeter network, wireless cellular networks, relay networks, and sensor networks, as well as other networks. All we can see from the three-stage period was the development of communications. At first, researchers were focusing on theoretical studies. After that, they were moving to solve the signal transmission part of communications. In the third period, researchers took the system and different kinds of networks as a whole into consideration and IoT became the most attractive application and research topics. The trend analyzed by our algorithm matched the development pattern of communications. And the example provided by us can be seen as the development of a new generation of communication technology.

To evaluate the performance of our method, it was necessary to compare the similarities and differences between algorithm-generated keywords and author selected keywords. From Table 1, we can easily find out that some algorithm-generated keywords are exactly the keywords that the author selected, and others that are not included in the author-selected keywords show more explicit details. These results demonstrated the ability of the method sorting out keywords for author-selected keyword replacement.

The second field we investigated is microfluidics/Lab on a Chip (LoC) and we chose *Lab on a Chip* as our source journal. *Lab on a Chip* did not ask authors to list any keywords so it was a good example to test our algorithm in an emerging field.

Based on the extracted data from *Lab on a Chip*, we concluded the development of LoC field into three stages as well in Figure 3. The three stages were *Early days of miniaturization* (gray arrow), *Growth of LoC technologies* (red arrow) and *Rise of LoC applications* (green arrow). And we summarized all the 14 keywords listed in Figure 3A as well as the corresponding sub-keywords which were highly consistent with author-provided keywords and algorithm-provided keywords in Table 2. The x-axis of Figure 3B showed the timeline of LoC technology development, while the y-axis showed the rank of three stages based on their corresponding keywords occurrence ratio. From 2003 to 2006 (the first stage), researchers were focusing on the fabrication or integration of microfluidic devices. The top keywords in this stage were *fabrication*, *integrated* and *electrophoresis*. Thus, we used *Early days of Miniaturization* to describe the first stage. The next stage started from 2007 and continued until 2013.

After solving the fabrication and integration issues, the research focus was moved to several LoC technologies, such as *(cell/particle) detection*, *(cell/particle) separation*, *acoustic (microfluidics)*, *optofluidics*. In this stage, the researcher was focusing on developing or inventing LoC-based technologies. Thus, we used *Growth of LoC technologies* to characterize this stage. Since 2014, researchers were more focusing on find specific applications for LoC technologies. From the extracted keywords in this stage, we could see the biological application keywords (*cell, particle*), chemical analysis keywords (*surface, nano*) and so on.

5. Conclusions

In conclusion, we propose an algorithm to automatically extract keywords from the meta information of peer-reviewed journals and use these keywords as the basic data for scientific review articles. Our algorithm consists of two parts. We use the RAKE algorithm to achieve keyword extractions and use Levenshtein distance to classify different related keywords into the same group. We first applied our algorithm in the field of communication engineering. From the keyword data, we divided the communication engineering into three stages from 2001 to present and summaries the extracted keywords with their occurrence in a table that could be further used in a review article. We also compared the algorithm-generated keywords with author-selected keywords, which showed the robustness of our method. What's more, we tested our algorithm on an emerging field, Lab on a Chip. Based on the extracted keywords, we found a clear path of how LoC technology was developed in the past decades. Researchers from LoC were solving the fabrication or integration issues first, then moving to develop new LoC-based technology, and finally trying to apply LoC technology to solve real-world problems.

5.1. Limitations

The proposed algorithm has certain constraints and limitations. The extracted keywords are usually terminology. We still need human experts to illustrate the further significance behind those raw keywords. For example, everyone knows 5G technology is developed for the next-generation mobile Internet connection, yet few people have a basic understanding of *Modulation*, *Coding*, or *decoding* (Figure 2A). Besides, the extracted keywords from one single journal could be biased to predict the overall development of one specific field. Different research journals in the same field could have distinct scopes and aims, which would tend to publish the articles more suitable to their scopes and aims. This issue could be amended by the implementation of our algorithm on multiple journals in the same field and analyze the extracted keywords together.

5.2. Future Directions

With the tremendous increment of published research articles, our algorithm could be one solution to reduce the labor-intensive literature survey process. We are confident that our algorithm could help provide not only the data but also the point of view that researchers might need a lot of time to find out during the process of understanding a new field or writing review articles. Moreover, nothing is stopping us to apply our proposed method only in the scientific field. It will be interesting to extend our method to more general fields and find the indications behind it. For example, it would be easy to apply our method to the print media (e.g., Newspaper) to find what kinds of news would be more attractive to people in different generations and connect the findings with economic development.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-8994/12/6/903/s1.

Author Contributions: J.W. and G.S. contributed equally to this work. J.W. contributed for the idea , data mining and manuscript; G.S. contributed for the idea and manuscript; C.W. contributed for the data collection and data mining; X.H. contributed for paper review and manuscript; L.S. contributed for paper review. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by National Natural Science Foundation of China No.61827806, Zhejiang Provincial Natural Science Foundation of China under Grant No.LQ20F040003 and the Fundamental Research Funds for the Provincial Universities of Zhejiang No.GK199900299012-008.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gallagher, A.G.; Ritter, E.M.; Champion, H.; Higgins, G.; Fried, M.P.; Moses, G.; Smith, C.D.; Satava, R.M. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Ann. Surg.* **2005**, *241*, 364. [CrossRef] [PubMed]
2. Ziegler, R.; Mueller, W.; Fischer, G.; Göbel, M. A virtual reality medical training system. In *Computer Vision, Virtual Reality and Robotics in Medicine, Proceedings of the First International Conference, CVRMed'95, Nice, France, 3–6 April 1995*; Springer: Berlin, Germany, 1995; pp. 282–286.
3. Hamza-Lup, F.G.; Rolland, J.P.; Hughes, C. A distributed augmented reality system for medical training and simulation. *arXiv* **2018**, arXiv:1811.12815.
4. Izard, S.G.; Juanes, J.A.; Peñalvo, F.J.G.; Estella, J.M.G.; Ledesma, M.J.S.; Ruisoto, P. Virtual reality as an educational and training tool for medicine. *J. Med. Syst.* **2018**, *42*, 50. [CrossRef] [PubMed]
5. Park, S.H.; Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **2018**, *286*, 800–809. [CrossRef]
6. Chartrand, G.; Cheng, P.M.; Vorontsov, E.; Drozdzal, M.; Turcotte, S.; Pal, C.J.; Kadoury, S.; Tang, A. Deep learning: a primer for radiologists. *Radiographics* **2017**, *37*, 2113–2131. [CrossRef] [PubMed]
7. Ward, B.; Brown, S.; Krebsbach, P. Bioengineering strategies for regeneration of craniofacial bone: A review of emerging technologies. *Oral Dis.* **2010**, *16*, 709–716. [CrossRef]
8. Daher, M.; Rezvani, K. Next generation natural killer cells for cancer immunotherapy: the promise of genetic engineering. *Curr. Opin. Immunol.* **2018**, *51*, 146–153. [CrossRef]
9. Vamathevan, J.; Apweiler, R.; Birney, E. Biomolecular Data Resources: Bioinformatics Infrastructure for Biomedical Data Science. *Annu. Rev. Biomed. Data Sci.* **2019**, *2*, 199–222. [CrossRef]
10. Gosak, M.; Markovič, R.; Dolenšek, J.; Rupnik, M.S.; Marhl, M.; Stožer, A.; Perc, M. Network science of biological systems at different scales: A review. *Phys. Life Rev.* **2018**, *24*, 118–135. [CrossRef]
11. Donohoue, P.D.; Barrangou, R.; May, A.P. Advances in industrial biotechnology using CRISPR-Cas systems. *Trends Biotechnol.* **2018**, *36*, 134–146. [CrossRef]
12. Blanckaert, J.; Glorieux, C.; Puers, R. Bionic Eye Lens. U.S. Patent App. 10/123,869, 13 November 2018.
13. Webber, M.J.; Appel, E.A.; Meijer, E.; Langer, R. Supramolecular biomaterials. *Nat. Mater.* **2016**, *15*, 13. [CrossRef] [PubMed]
14. Lakhani, P.; Sundaram, B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **2017**, *284*, 574–582. [CrossRef] [PubMed]
15. Ma, K.; Sun, L.; Wang, Y.; Wang, J. Classification of blood cancer images using a convolutional neural networks ensemble. Eleventh International Conference on Digital Image Processing (ICDIP 2019). *Int. Soc. Opt. Photonics* **2019**, *11179*, 1117903.
16. Gülpınar, Ö.; Güçlü, A.G. How to write a review article? *Turkish J. Urol.* **2013**, *39*, 44. [CrossRef] [PubMed]
17. Wee, B.V.; Banister, D. How to write a literature review paper? *Transp. Rev.* **2016**, *36*, 278–288. [CrossRef]
18. RELX. 2018 RELX Group Annual Report. 2019. Available online: https://www.relx.com/~/media/Files/R/RELX-Group/documents/reports/annual-reports/2018-annual-report.pdf (accessed on 23 May 2019).
19. Ioannidis, J.P.A.; Baas, J.; Klavans, R.; Boyack, K.W. A standardized citation metrics author database annotated for scientific field. *PLoS Biol.* **2019**, *17*, e3000384. doi:10.1371/journal.pbio.3000384. [CrossRef]
20. Hirsch, J.E. Does the h index have predictive power? *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19193–19198. [CrossRef]
21. Bornmann, L.; Daniel, H.D. Does the h-index for ranking of scientists really work? *Scientometrics* **2005**, *65*, 391–392. [CrossRef]
22. Anderson, C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Available online: https://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed on 3 March 2020).
23. Machi, L.A.; McEvoy, B.T. *The Literature Review: Six Steps to Success*; Corwin Press: Thousand Oaks, CA, USA, 2016.

24. Torraco, R.J. Writing integrative literature reviews: Guidelines and examples. *Hum. Resour. Dev. Rev.* **2005**, *4*, 356–367. [CrossRef]
25. Randolph, J. A guide to writing the dissertation literature review. *Pract. Assess. Res. Eval.* **2009**, *14*, 13.
26. Boote, D.N.; Beile, P. Scholars before researchers: On the centrality of the dissertation literature review in research preparation. *Educ. Res.* **2005**, *34*, 3–15. [CrossRef]
27. Denney, A.S.; Tewksbury, R. How to write a literature review. *J. Crim. Justice Educ.* **2013**, *24*, 218–234. [CrossRef]
28. Levy, Y.; Ellis, T.J. A systems approach to conduct an effective literature review in support of information systems research. *Inf. Sci.* **2006**, *9* 181–212. [CrossRef]
29. Aveyard, H. *Doing A Literature Review in Health and Social Care: A practical Guide*; McGraw-Hill Education: London, UK, 2014.
30. Piper, R.J. How to Write a Systematic Literature Review: A Guide for Medical Students. 2013, pp. 1–8. Available online: http://sites.cardiff.ac.uk/curesmed/files/2014/10/NSAMR-Systematic-Review.pdf (accessed on 1 March 2020).
31. IEEE Transactions on Communications. Available online: https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=26 (accessed on 29 April 2020).
32. Lab on a Chip. Available online: https://www.rsc.org/journals-books-databases/about-journals/lab-on-a-chip/ (accessed on 29 April 2020)
33. Rose, S.; Engel, D.; Cramer, N.; Cowley, W. Automatic keyword extraction from individual documents. *Text Min. Appl. Theory* **2010**, *1*, 1–20.
34. Levenshtein, V.I. Binary codes capable of correcting spurious insertions and deletions of ones. *Probl. Inf. Transm.* **1965**, *1*, 707–710.