The Normalized Impact Index for Keywords in Scholarly Papers to Detect Subtle Research Topics

Daisuke Ikeda
Kyushu University
Moto-oka 744, Fukuoka,
819-0395, Japan

Yuta Taniguchi
Kyushu University
Moto-oka 744, Fukuoka,
819-0395, Japan

Kazunori Koga
Kyushu University
Moto-oka 744, Fukuoka,
819-0395, Japan
National Institutes of Natural Sciences
Toranomon 4-3-13, Tokyo,
105-0001, Japan

Abstract

Mainly due to the open access movement, the number of scholarly papers we can freely access is drastically increasing. A huge amount of papers is a promising resource for text mining and machine learning. Given a set of papers, for example, we can grasp past or current trends in a research community. Compared to the trend detection, it is more difficult to forecast trends in the near future, since the number of occurrences of some features, which are major cues for automatic detection, such as the word frequency, is quite small before such a trend will emerge. As a first step toward trend forecasting, this paper is devoted to finding subtle trends. To do this, the authors propose an index for keywords, called *normalized impact index*, and visualize keywords and their indices as a heat map. The authors have conducted case studies using some keywords already known as popular, and we found some keywords whose frequencies are not so large but whose indices are large.

1 Introduction

Thanks to the recent open access movement, we can freely access to a huge amount of papers on scholarly repositories, such as institutional repositories maintained by academic institutions. According to IRUS-UK,[1] there exits about 2M items on more than 200 repositories in the UK, as of May 2020. According to NII,[2] there exist more than 2.4M full-text papers on 734 institutional repositories in Japan, as of March 2020. In addition to institutional repositories, we also have disciplinary repositories, such as arXiv.[3]

We can also use a global aggregation servie, which collects papers on repositories. For example, CORE [4] collects papers from more than one thousand data providers in about 150 countries, and provides search APIs, dump files, and search facility for collected papers (Knoth and Zdrahal, 2012). The latest dump file provided by CORE contains 123M metadata items, 85.6M abstracts, and 9.8M full text papers. Some commercial publishers also began to provide APIs for automatic processing.[5]

Basically, items on scholarly repositories are readable PDF files. When research results were published on paper, research papers were final outcomes of the researches. In case of digital media, however, contents of the papers can be an input for automatic processing. We can find many researches which use scholarly papers as input for computer algorithms. For example, some entities, like dataset names, used in papers are automatically extracted (Ikeda and Seguchi, 2017; Ikeda and Taniguchi, 2019), and papers are used to predict research impacts of a new given paper (Baba et al., 2019) and to predict new materials (Tshitoyan et al., 2019).

The final goal of our research is to forecast popular trends in the near future. A typical method for this is to use a clustering algorithm, which is unsupervised learning, and divides target items into groups based on a predefined distance metric. Some approaches use clustering algorithms to divide words in papers into groups, such as the topic model (Griffiths and Steyvers, 2004; Bolelli et al., 2009). Once we introduce a distance metric to data, a target data item is defined as a point in the space defined by the metric, and thus we can compare similarities between any two points. In this sense, this approach uses an absolute distance. There also exit relative approaches, like network structures, in which we know that two items are adjancent. In par-

[1] https://irus.jisc.ac.uk/
[2] https://www.nii.ac.jp/irp/en/archive/statistic/
[3] https://arxiv.org/
[4] https://core.ac.uk/
[5] https://www.elsevier.com/about/policies/text-and-data-mining

ticular, we can naturally construct multiple network structures from papers, like networks of authors, citations, words, and their combinations (Duvvuru et al., 2012; Salatino et al., 2017). However, these researches assume that there are already a number of publications (Salatino et al., 2018). In this sense, these approaches are for topic detection, not for topic forecast.

In this paper, we try to find small topics as a first step toward forecasting future topics. To this end, we propose an index for keywords to measure their impact, assuming a keyword denotes a research topic. We use a relative frequency in the definition of the index to find small topics. As far as the authors know, the frequency of keywords is not directly used to detect topics in research papers, unlike topic or trend detection in general text data. The authors think that this is because a frequency based method requires a list of stop words to remove unnecessary keywords, but it is too costly to construct it for each discipline in case of research papers.

To evaluate the proposed index, we use some popular keywords in one discipline, and we check if the proposed indices for them can grasp their popularity. Using this approach, we do not have to consider the issue of stop words. In other words, we try to find some properties among popular topics with the proposed index. For comparison, we also show topic detection by absolute frequency and a standard clustering algorithm.

2 Normalized Impact Index

We assume the range of publication years, $y_1, y_2, \ldots, y_N$, and let $Y = \{y_1, y_2, \ldots, y_N\}$. For $y \in Y$, $D(y)$ denotes the set of papers published in y.

For a word w and a year $y \in Y$, the *normalized impact index*, denoted by $h(w, y)$, is defined as as follows:

$$h(w, y) = \frac{f(w, y)}{|D(y)| \sum_{t=y_1}^{y_N} f(w, t)},$$

where $f(w, y)$ is the number of occurrences (frequencies) of w in $D(y)$.

The proposed index for w and y is a relative frequency, normalized by both the number of publications in y and the total frequency of w among all years. Therefore, we can compare $h(w_1, y_1)$ and $h(w_2, y_2)$.

To understand the meaning of the index, let us assume that $|D(y)| = 1$ tentatively. Then we can treat $h(w, y)$ as a probability since we have $\sum_y h(w, y) = 1$. So, when we depict this index as a bar chart for some w whose height is $h(w, y_i)$, the total area of the bars for w is normalized to 1. Therefore, we can compare any two words w_1 and w_2, in the view point of their trends.

When we consider trends of keywords, it is natural to see temporal changes of the index from some reference year y_1, that is,

$$h(w, y) - h(w, y_1), \tag{1}$$

where $y > y_1$ for $y \in Y - \{y_1\}$. For some $y(\neq y_1)$, if $h(w, y) - h(w, y_1) > 0$ (resp. < 0), then the relative usage of w in y becomes larger (resp. smaller) than that in y_1. This leads to a heat map of the proposed index for keywords.

3 Case Study

In this section, we apply the proposed index to a real dataset to confirm its efficacy. As described in Section 1, a frequency based method suffers from the issue of stop words. To avoid the issue, we check the values of the proposed index for some keywords the authors selected from some specific field. These keywords are already known as popular topics. Therefore, it means that we only check positive examples.

Since the proposed index is defined with relative frequencies, we show the result of topic detection with absolute frequencies for comparison (see Section 3.2). Then, we apply a clustering algorithm to our dataset in Section 3.3, to confirm that a clustering algorithm for keywords can find large topics, not small ones as described in Section 1.

3.1 Dataset

We use a set of abstracts, not the whole papers, from 2000 to 2018, obtained by searching "plasma chemical vapor deposition" at Web of Science. The number of abstracts we obtained is 69,384.

In addition to stop words of English, we also removed tokens starting or ending with special symbols, such as "[" and "+". Then we converted capital letters to lower-case ones.

3.2 Topic Detection by Frequency

As the first case study, we check if a method based on frequency can find a potentially popular topic.

Figure 1 contains four graphs, showing the numbers of papers found by queries at Web of Science. One common line is contained in all graphs

in Figure 1, which is the number of papers found by "plasma chemical vapor deposition". In other

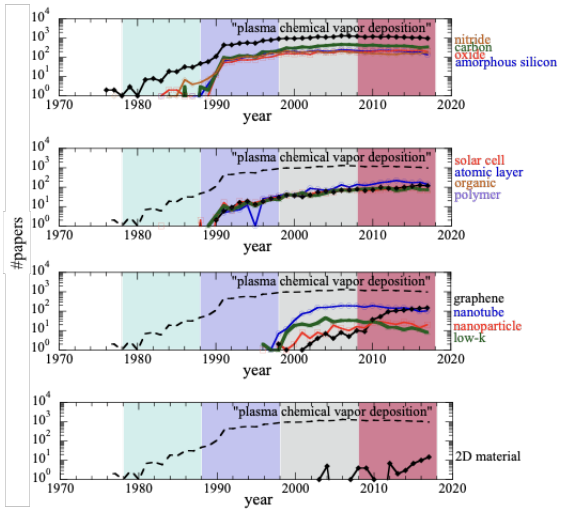

Figure 1: Each graph shows the change of the number of papers found by the corresponding query with "plasma chemical vapor deposition", such as "nitride plasma chemical vapor deposition", as the publication year advances (some data originally from Fig. 6 and 7 in (Iwase et al., 2019)).

words, this line shows the year-by-year changes of the number of papers containing this query. We call the line for this query the *base line* of this field.

Each of the other lines shows the number of papers found by "plasma chemical vapor deposition" plus the corresponding keyword. For example, the red line in the top graph is obtained by "oxide plasma chemical vapor deposition". These searches are search within the original query, and thus these lines are below the base line. One of the authors chose these additional keywords, based on the heat map in Figure 2 in addition to his expertise. Basically, they are known to be popular topics.

In the four graphs, an upper graph contains keywords whose frequencies are larger. In the top graph of "nitride", "carbon", "oxide", and "amorphous silicon", we see that these keywords are large topics in this field and the shapes of graphs are similar to the base line. Compared to the top graph, the second one contains smaller topics, but they have emerged in early 90s, and increased its publications steadily.

Compared to the two top graphs, keywords for the other two graphs are relatively new topics, and thus the numbers of papers containing these topics are much smaller. In particular, the number of the papers about "2D material", meaning 2 dimensional materials, is quite small. In spite of its small frequency, this topic has potential to be big in this field because "2D material" is a more conceptual word than "graphene", which is a 2D material, and the Nobel Prize was awarded to researchers studied graphene in 2010.

Therefore, methods based on the frequency of a keyword can not find such a trend at very early stages.

3.3 Topic Detection by Clustering

Next, we consider a clustering algorithm as a method to find research topics.

For a clustering algorithm, we used Non-negative Matrix Factorization (NMF), which decomposes a given matrix $\boldsymbol{V}$ into two matrices $\boldsymbol{W}\boldsymbol{H}$, where all emelements in those matrices are required to be non-negative (Lee and Seung, 1999).

Using the set of abstracts, we can construct a term-document matrix $\boldsymbol{V}$, where w_{ij} is the frequency for the ith term in the jth document, that is the jth document d_j has $w_{1j}, w_{2j}, \ldots$ as its elements.

Let D and V be the number of documents and one of vocabularies, respectively. Then, the size of $\boldsymbol{V}$ is $D \times V$. When we apply NMF to $\boldsymbol{V}$, we have to specify a parameter K, which defines the sizes of two matrices: $D \times K$ and $K \times V$ for $\boldsymbol{W}$ and $\boldsymbol{H}$.

We can see $\boldsymbol{W}$ as a weight matrix and $\boldsymbol{H}$ as a base matrix, and an original document is expressed as a weighted linear combination of base elements. In this expression, we can see that a base matrix consits of K base vectors.

Table 1 shows the top 10 keywords with largest weights for each base vector, where we set $K = 10$. There exist K topics, each of which has 10 keywords with the top 10 largest weights in the topic.

From this table, we can find many major topics in this field. For example, the first cluster contains "chemical vapor deposition", and the second and 10th ones "carbon nanotubes" and "thin film", respectively, both of which are major materials used in this field. However, we can not find minor topics from this decomposition.

3.4 Topic Detection by the Proposed Index and Heat Map

In this section, we detect topics using the normalized impact index and its visualization.

No.	The top 10 keywords with largest weights in a topic
1	deposition, chemical, vapor, rate, process, high, gas, using, PECVD, pressure
2	carbon, growth, nanotubes, CNTs, field, emission, electron, catalyst, grown, chemical
3	silicon, layer, solar, amorphous, cells, layers, chemical, cell, nitride, high
4	films, deposited, thin, properties, spectroscopy, optical, amorphous, content, using, x-ray
5	surface, surfaces, roughness, layer, chemical, contact, treatment, energy, morphology, atomic
6	plasma, power, gas, density, treatment, enhanced, using, pressure, hydrogen, discharge
7	C, degrees, temperature, annealing, growth, substrate, temperatures, si, low, rights
8	diamond, growth, microwave, substrate, CVD, high, nucleation, quality, substrates, grown
9	coatings, coating, properties, DLC, chemical, using, deposited, wear, elsevier, reserved
10	film, thin, thickness, substrate, deposited, stress, structure, dielectric, nm, ratio

Table 1: The top 10 keywords with largest weights in a topic found by NMF.

Figure 2 shows a heat map, defined by (1), for keywords in our dataset. One column corresponds

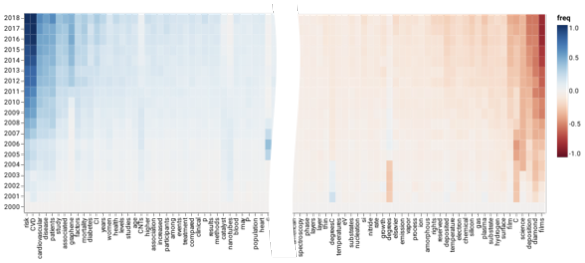

Figure 2: The heat map shows values of (1) for each keyword extracted from our dataset, where one column corresponds to one keyword, and a cell in the column indicates the value of (1).

to one keyword, and each row to one year. We only show the left and right parts of the heat map because the original figure is too wide since there are many keywords.

Each cell shows the difference between the normalized impact index of that year and the reference year, 2000, for some word. That is, it shows the value of (1), where blue (resp. red) cells are positive (resp. negative) values, meaning the relative frequency of the corresponding year for the word is larger (resp. smaller) than that of the reference year.

Figure 3 shows temporal changes of the proposed indices for some selected keywords, some of which appear in Figure 1 and the other ones are chosen from the heat map.

"graphene", "2D", "nanotube", "low-k" (low dielectric constant), "h-BN" (hexagonal boron nitride), and "GaAs" are names of materials, and "interconnect" and "fuel" are the keywords of the plasma chemical vapor deposition (CVD for short) applications, where "interconnect" refers as inter-

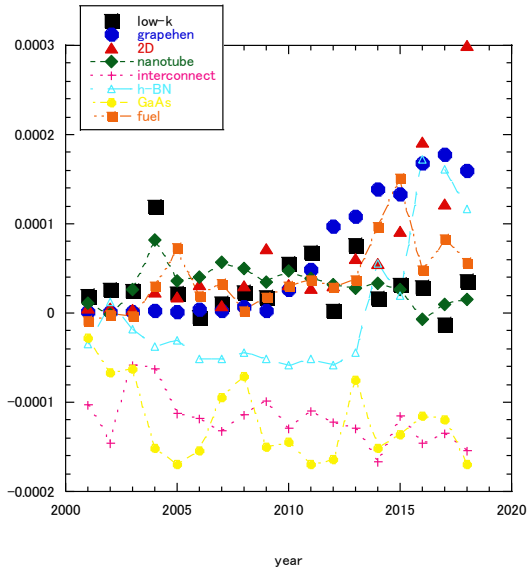

Figure 3: The graph shows the temporal change of the proposed index for some keywords, such as "low-k".

connect in semiconductor devices and "fuel" as fuel cells.

For interconnect, the proposed index was negative and decreased from 2000. Plasma CVD as interconnect process technology has been losing interest. The proposed index for fuel increases continuously and there was temporary booming in 2000 and 2015.

Both "nonotube" and "low-k" appeared in the third graph of Figure 1. From this graph, we can see sharp rises of their frequencies. However, from the proposed index for these keywords, we can not say these topics are actively examined in papers.

As shown in Figure 1, "2D" has its small frequency although it has potential to be a big trend because unique characteristics of 2D materials have been found then the research of 2D materials seems

to become active as the trigger of the graphene Nobel Prize. On the other hand, the proposed index for "2D" rises sharply in Figure 3.

The index for "h-BN" has negative values until 2012, which seems to have lost the interest of researchers, but after that it increases rapidly. In fact, "h-BN" has been studied as a 2D semiconductor material recently. In this sense, "h-BN" can be seen as a 2D material family, and so it is convincing the sharp rise for "h-BN".

4 Conclusion

In this paper, we have introduced an index to find keywords, which express small topics, using relative frequencies. As visualization, the difference of the proposed index from the reference year, 2000 in this paper, is depicted as a heat map. Therefore, we can easily find subtle topics even if their absolute frequencies are not so large. We have conducted case studies using the proposed index, and confirmed that some keywords, which are already known as popular, show sharp rises of the proposed index.

As described in Section 3, we have only checked popular keywords. So it is an important future work to check all keywords whose values of the proposed index.

Even if we find some keywords with high values of the proposed index, you might want to check their absolute frequencies. Therefore, it is also important to develop a visualization tool which enables to check both the absolute frequency and the proposed index. Similarly, it is an important future work for the tool to introduce a grouping facility, which groups a different keywords in a hierarchical way, and then we can grasp transitions of topics with flexible granularity with the tool. To do so, we can use some vocabulary system, like one in (Salatino et al., 2019), or word embeddings to measure the distances between two keywords.

Acknowledgments

In this paper, the authors used data from Web of Science, a product of Clarivate.

References

Takahiro Baba, Kensuke Baba, and Daisuke Ikeda. 2019. Citation Count Prediction using Abstracts. *Journal of Web Engineering*, 18(1–3):207–228.

Levent Bolelli, Şeyda Ertekin, and C. Lee Giles. 2009. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation. In *Advances in Information Retrieval (ECIR 2009)*, Lecture Notes in Artificial Intelligence 5478, pages 776–780.

Arjun Duvvuru, Sagar Kamarthi, and Sivarit Sultornsanee. 2012. Undercovering research trends: Network analysis of keywords in scholarly articles. In *Proceedings of Ninth International Conference on Computer Science and Software Engineering*, pages 265–270.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Daisuke Ikeda and Daisuke Seguchi. 2017. Automatically Extracting Keywords from Documents for Rich Indexes of Searchable Data Repositories. In *Proceedings of the 12th International Conference of Open Repositories*.

Daisuke Ikeda and Yuta Taniguchi. 2019. Toward Automatic Identification of Dataset Names in Scholarly Articles. In *Developments in Open Science and Research Data Management: 8th International Conference on Data Science and Institutional Research*.

Taku Iwase, Yoshito Kamaji, Song Yun Kang, Kazunori Koga, Nobuyuki Kuboi, Moritaka Nakamura, Nobuyuki Negishi, Tomohiro Nozaki, Shota Nunomura, Daisuke Ogawa, Mitsuhiro Omura, Tetsuji Shimizu, Kazunori Shinoda, Yasushi Sonoda, Haruka Suzuki, Kazuo Takahashi, Takayoshi Tsutsumi, Kenichi Yoshikawa, Tatsuo Ishijima, and Kenji Ishikawa. 2019. Progress and perspectives in dry processes for emerging multidisciplinary applications: how can we improve our use of dry processes? *Japanese Journal of Applied Physics*, 58(SE).

Petr Knoth and Zdenek Zdrahal. 2012. CORE: Three Access Levels to Underpin Open Access. *D-Lib Magazine*, 18(11/12).

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Angelo Salatino, Francesco Osborne, Thiviyan Thanapalasingam, and Enrico Motta. 2019. The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In *Proceedings of the 23rd International Conference on Theory and Practice of Digital Libraries*, Lecture Notes in Computer Science 11799.

Angelo A. Salatino, Francesco Osborne, and Enrico Motta. 2017. How are topics born? understanding the research dynamics preceding the emergence of new areas. *PeerJ Computer Science*, 3(e119).

Angelo A. Salatino, Francesco Osborne, and Enrico Motta. 2018. AUGUR: Forecasting the Emergence of New Research Topics. In *Proceedings of the*

18th ACM/IEEE on Joint Conference on Digital Libraries, pages 303–312.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature*.