Micro science and technology fields requiring mathematically trained contributors: Topic modeling using journal paper abstracts

Takashi Ikegawa
The University of Tokyo/Waseda University
Tokyo, Japan
tikegawa@g.ecc.u-tokyo.ac.jp

***Abstract*—A shortage of mathematically trained individuals who can integrate mathematical knowledge with other types of knowledge and skills and thereby contribute to innovation has recently become apparent. Finding clues to help in solving this problem requires a) identification of scientific and technology fields (STFs) that require mathematically trained contributors and b) identification of higher education institutions (HEIs) that provide students an opportunity to develop the mathematical knowledge needed in such STFs. Previous work on this problem has shown that the granularity of STFs is coarse, such as at the information theory level. An in-depth discussion of the educational practices, e.g., project-based learning and industrial internship programs, at mathematical HEIs requires analysis at the micro-STF level. To enable such analysis, a method is presented for discovering topics hidden in a collection of journal paper abstracts. For example, topic modeling using the latent Dirichlet allocation algorithm implemented in Python enabled the discovery of the topics covered studied at the three highest-ranked mathematical HEIs in the information theory field.**

***Index Terms*—Scientific and technology field, higher education institution, topic modeling.**

I. Introduction

Mathematics plays a key role in innovation by providing solutions to many challenging problems in a wide variety of fields [1]–[3]. It has recently become apparent that there is a shortage of mathematically trained individuals who can integrate mathematical knowledge with other types of knowledge and skills and thereby contribute to innovation [3], [4].

A method for solving this problem is herewith presented:

> **Step 1:** Identify scientific and technology fields (STFs) that require innovative thinking from mathematically trained individuals.
> **Step 2:** Identify higher education institutions (HEIs) that provide students an opportunity to develop the mathematical knowledge that is needed in the identified STFs.
> **Step 3:** Identify the educational practices, e.g., curriculum, project-based learning, and industrial internship programs, provided by the identified HEIs that produce mathematically trained graduates who are active in the identified STFs (using surveys including interviews and website searches).
> **Step 4:** Promote widely the excellent educational practices identified in Step 3 to other HEIs.

The author previously presented a method comprising Steps 1 and 2 [5] that can be used to identify the STFs and the HEIs with which mathematically trained contributor are affiliated. It is a bibliometrics method for analyzing a large amount of journal paper metadata. Bibliometrics methods are a quantitative and inexpensive way for discovering the statistical features inherent in a large collection of documents [6].

However, the information granularity of STFs obtained using this method is coarse because journal names are used to identify the STFs. For example, the STFs receiving the greatest contributions from mathematically trained individuals have been shown to include information theory, reliability, fuzzy systems, and neural networks.

To discuss the educational practices of the identified mathematical HEIs in more depth, analysis at the micro-STF level is required. We propose discovering such micro-STFs in a collection of journal paper abstracts.

There are several techniques for revealing the underlying summarized text in a large collection of documents. Topic modeling, which has emerged in natural language processing (NLP), is one of the best techniques for text summarization [7]. Collected documents (or a "corpus" in NLP terms) are mapped to clusters of words, i.e., topics. We assume that each topic is equivalent to a micro-STF and used topic modeling to identify micro-STFs.

II. Related work [5]

In this section, we first briefly explain the previously presented method comprising Steps 1 and 2 [5]. Next, we describe the remaining problem, i.e., the need for analysis at the micro-STF level.

A. Method comprising Steps 1 and 2

Figure 1 illustrates the identification of STFs associated with HEIs by using journal names. There are myriad IEEE journals covering a gamut of topics, such as information theory, fuzzy systems, and vehicular technologies. A journal's name is usually a rough indicator of the topic(s) covered.

Each paper has author information data, including

- Author name
- Author affiliation
  - Institution name

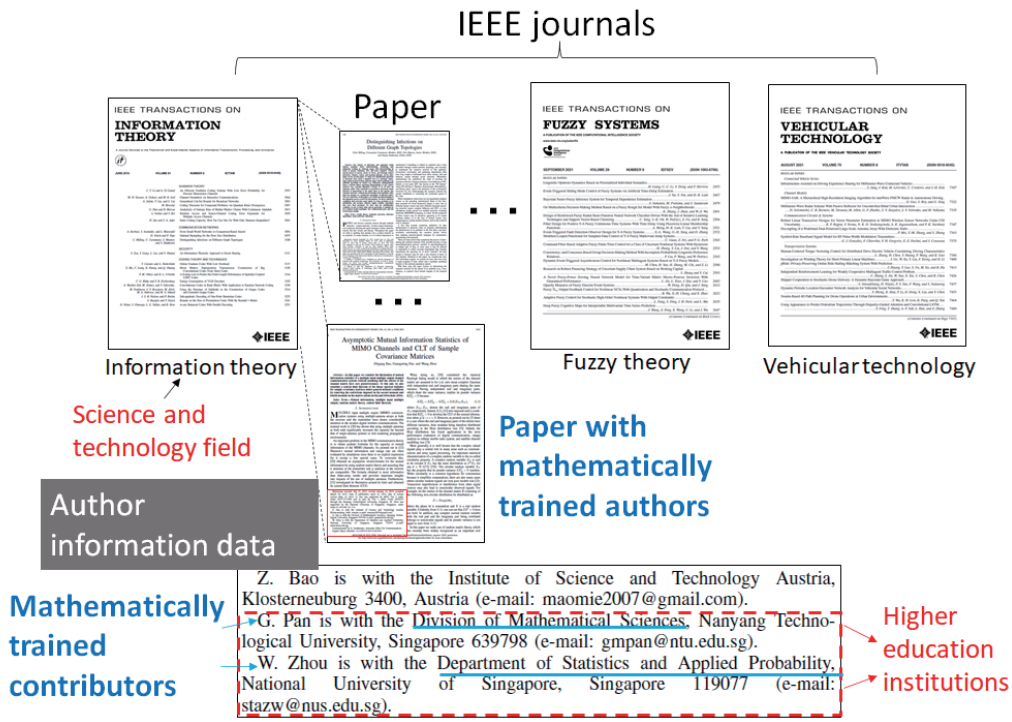

Fig. 1. Identification of scientific and technology fields associated with higher education institution by using journal names.

- Department name
- Sub-department name
- Country name.

In the previous report [5], an author was assumed to be mathematically trained if the author was in the mathematics (sub-)department of an HEI; i.e., the name of the (sub-)department contains at least

- Mathematics or "Math."
- Statistics or "Stat."
- Operations Research or "OR."

Note that journal paper metadata containing author information can be obtained using the Scopus application programming interface (API) [8].

A measure referred to as the "contribution rate of mathematically trained author(s) (CRMT)" was introduced that quantitatively represents the degree of contribution by mathematically trained author(s) to journals' published papers. CRMT is defined as the number of papers with mathematically trained author(s) divided by the number of papers with complete metadata.

Figure 2 shows the CRMT by IEEE journal. The y-axis shows the total number of journal editions from 2011 to 2020, and the x-axis shows the CRMT. The total number of journals is 140. It shows that mathematically trained individuals contributed to STFs, including information theory, reliability, and fuzzy systems, as ascertained from the names of the three highest-ranked journals.

Table I lists the three highest-ranked HEIs with respect to CRMT with which mathematically trained contributors are affiliated for "IEEE Transactions on Information Theory." The identified HEIs are well suited for developing educational programs that produce mathematically trained graduates who are active in fields other than mathematics, i.e., information

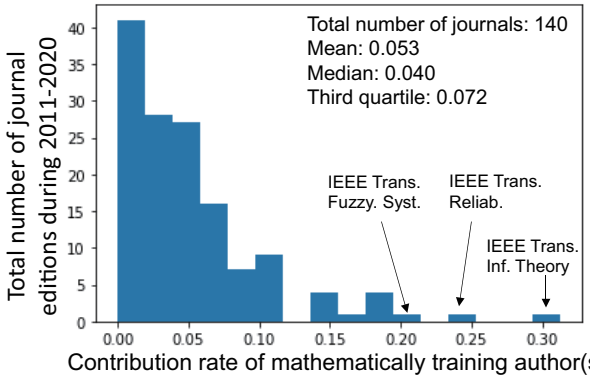

Fig. 2. Histogram of CRMT for IEEE papers published from 2011 to 2020 by journal [5, Fig. 9].

TABLE I
THREE HIGHEST-RANKED HEIS WITH RESPECT TO CRMT WITH WHICH MATHEMATICALLY TRAINED CONTRIBUTORS ARE AFFILIATED FOR "IEEE TRANSACTIONS ON INFORMATION THEORY."

Rank	HEI name		Country
	Institution name	Department name	
1	Nanyang Technological University	School of Physical and Mathematical Sciences	Singapore
2	National University of Singapore	Department of Mathematics	Singapore
3	Stanford University	Department of Statistics	United States

theory.

Remark 1 The STFs associated with HEIs are useful for identifying the excellent educational practices, which is Step 3 (described in Section I). This information contributes to the formulation of hypotheses when Step 3 is performed. For example, if the HEIs with which mathematically trained individuals are affiliated are found to produce a large number of papers in the field of information theory, we can hypothesize that these HEIs are providing relevant educational programs, such as courses and project-based learning related to information theory.

B. Analysis at micro-STF level

As explained above, identification of the educational practices of the identified mathematical HEIs in more depth requires analysis at the micro-STF level.

Remark 2 Making the information granularity of STFs finer would increase the precision of the hypotheses formulated in Step 3. For example, if the granularity was reduced from the broad level of "information theory" to the finer level of cryptography would enable the formulation of the hypothesis that educational programs specific to cryptography are being provided. This in turn would make the identification of excellent educational programs more effective and efficient.

The method presented here enables the discovery of micro-STFs.

III. Topic modeling using journal paper abstracts

Our proposed topic modeling method is aimed at discovering topics, i.e., micro-STFs, on the basis of groups of words. We focus on journal paper abstracts as documents because they are easily obtained via the Scopus Abstract Retrieval API [9].

Figure 3 depicts topic modeling using journal paper abstracts. It produces three-fold output: a set of topics, the frequency of words for each topic, and the distribution of topics in each abstract. For example, the topic model in the figure identified the

- words "coupling problem," "Rényi resolvability," and "maximal guessing coupling" for the topic "coupling," and
- the words "write-once memory (WOM) code," "wiretap pattern," and "quantum rate distortion" for the topic "WOM."

IV. Results and Discussion

We focused on papers with mathematically trained author(s) published between 2011 and 2020 in "IEEE Transactions on Information Theory." To discover the topics, we used the Scikit-learn package, which is an easy-to-use and proficient Python library implementing the latent Dirichlet allocation (LDA) algorithm [11]. We assumed that three topics (see Appendix for the topic modeling procedure used).

Tables II (a), (b), and (c) list the topics discovered and the words identified in "IEEE Transactions on Information Theory" for the three highest-ranked HEIs with respect to CRMT:

- School of Physical and Mathematical Sciences, Nanyang Technological University,
- Department of Mathematics, National University of Singapore, and
- Department of Statistics, Stanford University.

The STFs, i.e., topics, discovered using the proposed method are more granular than those using the previous method comprising Steps 1 and 2 [5]. Although the authors of the aforementioned HEIs publish papers in the same STF, their topics differ (see Tables II (a), (b), and (c)). The journal's STF might therefore not be an adequate descriptor of their

TABLE II
Topics discovered and words identified for "IEEE Transactions on Information Theory" by HEI.

Topic	Words identified
Coupling	Coupling problem, Quantum rate distortion, Rényi resolvability, Guessing coupling, Maximal guessing, Maximal guessing coupling, Preservation region, UMP code, Distance measure
Rényi conversion	Rényi conversion rate, Conversion rate, Rényi conversion, Classical bit, Generalized class, Generalized class hash, Harvested energy, Differentially uniform permutation, Uniform permutation
Write-once memory (WOM)	WOM code, Wiretap pattern, Distortion measure, Chain rule, Two-write WOM code, Two-write WOM, Noisy channel, Coding region, Second-order coding region

(a) School of Physical and Mathematical Sciences, Nanyang Technological University

Topic	Words identified
Optimal nonlinearity	Optimal nonlinearity, DoF region, Receive antenna, One-bit measurement, Face randomly, Face randomly projected, Projected polytope, Randomly projected, Hard threshold
Signal corruption	Signal corruption, Signal recovery, Mean variance, Upper triangular, Transmission capacity, Outage probability, Potential path, Recovery structured, Recovery structured corruption
Non-coherent capacity	Non-coherent capacity, Estimation error, Optimal spreading, Harvested energy, Domain interaction, Deletion probability, Deletion channel, Input distribution, Unconstrained case

(b) Department of Mathematics, National University of Singapore

Topic	Words identified
LCD	LCD code, Self-orthogonal code, Family binary LCD, Infinite family binary, Code self-orthogonal, Code self-orthogonal code, Binary LCD code, Binary LCD, LCD code self-orthogonal
Power permutation	Power permutation, FHSS FH set, FHSS FH, Binary GCP, ZCZ width, Optimal OB-ZCP, Low differential, Low differential uniformity, Differential spectrum
PAPR	Low PAPR, SNC low PAPR, SNC low, CSSS SNC, CSSS SNC low, Complementary set, Bent idempotent, Spectral null, Anti-self-dual bent function

(c) Department of Statistics, Stanford University

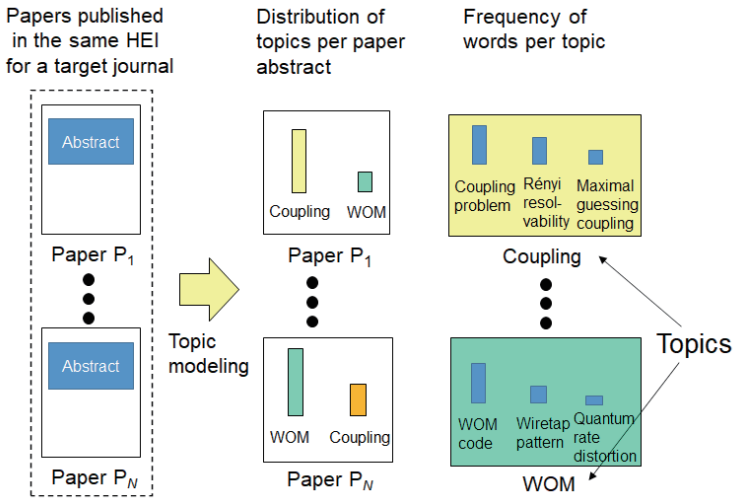

Fig. 3. Topic modeling using journal paper abstracts.

work[1]. Suitable STFs with respect to information granularity are required in order to identify the excellent educational practices that produce mathematically trained contributors.

V. Conclusion

We have presented a method for discovering the micro fields of science and technology (STF) to which mathematically trained individuals contribute, given by the higher educational institutions (HEIs) with which they are affiliated. The identification of STFs at a more granular level (i.e., micro-STFs) enables more effective and efficient identification of the excellent educational practices at the associated HEIs because it enables the formulation of better hypotheses.

In the proposed method, micro-STFs are treated as topics. We demonstrated that topic modeling using LDA enables the discovery of the topics covered at the three highest-ranked mathematical HEIs in the information theory field. However, the topics discovered might not be useful because of much micro-STF.

Future work includes

1) developing appropriate methods for identifying excellent educational practices by using the discovered STFs associated with HEIs that have been producing mathematically trained contributors,
2) determining the optimal degree of information granularity for STFs, and
3) improving and/or developing algorithms for discovering STFs with the optimal degree of information granularity as well as determining the optimal tuning of the (hyper-)parameters, including the number of topics.

[1]For example, the scope of "IEEE Transactions on Information Theory" includes coding theory, data compression, signal processing, pattern recognition, cryptography, and quantum information theory [12].

VI. Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19K02871.

Appendix

This appendix provides a brief description of the procedure used for topic modeling.

Step 1: Collect documents, i.e., journal paper abstracts, by extracting the JSON-object "abstract" from IEEE metadata acquired using the Scopus Abstract Retrieval API.

Step 2: Perform the following preprocessing:

1) Remove punctuation in collected documents,
2) Lowercase text in collected documents,
3) Tokenize collected documents; i.e., split documents into sentences and split sentences into words, and
4) Remove stop words, which are defined as commonly used words that do not provide much meaningful context on their own.

Step 3: Create n-grams from the tokenized documents that contain n consecutive words. For example, 1-grams include "information"; 2-grams concatenate two 1-grams such as "coupling problem"; and 3-grams comprise three 1-grams such as "compresses sensing model." We created 2- and 3-grams from the tokenized documents.

Step 4: Convert a collection of n-grams into a matrix of term counts, namely a dictionary for NLP, which is the main input to the LDA topic model.

Step 5: Build the LDA model using the dictionary provided by the number of topics and hyperparameters including α, which controls the prior distribution over the topic weight of each document.

Step 6: Visualize the results using the LDA model.

REFERENCES

[1] Society for Industrial and Applied Mathematics, "Mathematics in Industry Report," ccessed May 1, 2021. [Online]. Available: http://www.siam.org/Portals/0/Publications/Reports/Mathematics_Industry_2012.pdf

[2] National Research Council, "The Mathematical Sciences in 2025," The National Academies Press, 2013.

[3] P. Bond, "The Era of Mathematics: An Independent Review of Knowledge Exchange in the Mathematical Sciences," Accessed May 1, 2021. [Online]. Available: https://webarchive.nationalarchives.gov.uk/ukgwa/20220208115701/http://epsrc.ukri.org/newsevents/pubs/era-of-maths/

[4] The Conference Board, "US Labor Shortages: Challenges and Solutions," Accessed May 1, 2021. [Online]. Available: https://conference-board.org/topics/labor-shortages/us-labor-shortages-report-2020

[5] T. Ikegawa, "Science and technology fields and higher education institutions with mathematically trained contributors: Metadata analysis of IEEE papers," in *Proceedings of 2021 IEEE Frontiers in Education Conference (FIE)*, pp. 1–9, 2021.

[6] O. Ellegaard and J.A. Wallin, "The bibliometric analysis of scholarly production: How great is the impact?," *Scientometrics*, vol. 105, pp. 1809–1831, 2015.

[7] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation: A survey," *ACM Comput. Surv.*, vol. 54, no. 7, article 145, pp 1–35, 2022.

[8] Elsevier Developer Portal, "API Interface Specification," Accessed May 1, 2021. [Online]. Available: https://dev.elsevier.com/api_docs.html

[9] Elsevier Developer Portal, "Abstract Retrieval API," Accessed April 11, 2022. [Online]. Available: https://dev.elsevier.com/documentation/AbstractRetrievalAPI.wadl

[10] M. E. Rose and J. R. Kitchin, "Pybliometrics: Scriptable bibliometrics using a Python interface to Scopus," *SoftwareX*, vol. 10, p. 100263, 2019.

[11] "Scikit-learn: Machine learning in Python," Accessed April 14, 2022. [Online]. Available: https://scikit-learn.org/stable/

[12] IEEE Transactions on Information Theory, "Aims and scope," Accessed April 11, 2022. [Online]. Available: https://www.itsoc.org/it-trans