Inter-subdiscipline Analysis Based on Mathematical Statements

Rui Wang, Xiaoling Zhou
Center for Applied Mathematics
Tianjin University, Tianjin, China
rwang_ruiwang@tju.edu.cn
zhouxiaoling0727@gmail.com

Jian Wu
Department of Computer Science
Old Dominion University
Norfolk,VA
j1wu@odu.edu

Ou Wu*
Center for Applied Mathematics
Tianjin University, Tianjin, China
wuou@tju.edu.cn

ABSTRACT

A mathematical paper contains various mathematical statements, including definitions, theorems, lemmas, and so on. The mining of mathematical literature currently focuses on formulas and disregards statements. The present study investigates the (automatic) subdiscipline classification for mathematical statements. The classification results are applied into inter-subdiscipline analysis, including proportion and dependency analyses. First, a statement learning data is directly compiled from mathematical textbooks with a little human labeling to train an effective subdiscipline classifier. Second, a relatively large corpus, namely, analysis data, is compiled from mathematical journals. The classification results on the analysis data are subsequently used to quantify the inter-subdisciplinary relationships and conduct proportion analysis. Lastly, the dependency of different subdisciplines is analyzed and dependency chains among subdisciplines can be obtained.

KEYWORDS

mathematical statements; dependency chain; deep learning

ACM Reference Format:
Rui Wang, Xiaoling Zhou, Jian Wu, and Ou Wu. 2020. Inter-subdiscipline Analysis Based on Mathematical Statements. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398574

1 INTRODUCTION

Mathematical literatures contains a substantial number of mathematical statements that include definitions, theorems, lemmas, etc. [5]. The development of professional mathematical digital libraries, which can index and search formulas, has recently aroused great interest [1]. Little work on mathematical statements are available. Nevertheless, the mining of statements should further be explored because it can be used to quantify the connections among different subdisciplines.

In this study, we investigate how to (automatically) determine the subdiscipline category of a mathematical statement based on deep learning, which is regarded as subdiscipline classification. The main contributions of our work are summarized as follows. 1) A relatively new literature mining problem, namely, subdiscipline classification for mathematical statements, is investigated. To our knowledge, this work is the first to focus on the mining of mathematical statements. 2) The process of applying the statement classification results to analyze the inter-subdisciplinary relationships is investigated. 3) Two data corpora are compiled. One is used to train the classifier, the other is for analysis.

2 DATA AND METHOD

2.1 Datasets

Two corpora consisting of mathematical statements are compiled, namely, learning data and analysis data. The learning data is gathered to train a classifier and the analysis data is utilized to explore the inter-subdisciplinary relationships.

`Learning data.` The Classification and Code of Disciplines[1], indicates 25 mathematical subdisciplines in total. Ten representative subdisciplines are considered, namely, algebra, geometry, topology, number theory, function, algebraic geometry, combinatorics, probability, functional analysis, and ordinary differential analysis. After the refinement, 49 categories are obtained. A total of 260 textbooks in PDF format are downloaded from the Internet[2]. Mathematical statements and their associated categorical labels can be directly extracted from selected textbooks. Finally, 37,720 statements are obtained.

`Analysis data.` A total of 5,844 papers of the 10 subdisciplines are downloaded from 12 specialized journals. A total of 121,949 mathematical statements are extracted from these papers.

2.2 Methods

There are three technical components including statement extraction, subdiscipline classifier training, and inter-subdiscipline dependency quantification.

Statement extraction is mainly based on keywords and identifiers. The content between the beginning keyword and the ending identifier is extracted as a mathematical statement.

In subdiscipline classifier training, three deep learning models are compared, including Long Short-Term Memory

*Corresponding author.

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/06.
https://doi.org/10.1145/3383583.3398574

[1] http://www.bdp.cas.cn/pjyjl/kjpj/201608/P020160830307329828774.pdf
[2] https://github.com/wasdfghjklr/Supplementary

Table 1: The classification macro F1 scores.

Model	10-category set	49-category set
CNN	0.89	0.70
Transformer	0.66	0.43
LSTM	0.91	0.81
Random assignment	0.09 (1/10)	0.021 (1/49)

Networks (LSTM), Convolutional Neural Network (CNN), and Transformer[2].

Inter-subdiscipline dependency is based on sub-subdiscipline dependency. Let p and q represent two subdisciplines, and q_i be the *ith* sub-subdiscipline of q. The sub-subdiscipline dependency of p on q_i is calculated as follows defined by us:

$$q_{p \leftarrow q_i} = \frac{M_{q_i}}{M_p}, i \in \{1, 2, ..., N_q\} \quad (1)$$

where M_p is the number of statements that belong to p. M_{q_i} is the number of statements that are classified to q_i from p. N_q is the number of sub-subdisciplines in q. The inter-subdiscipline dependency ($D_{p \leftarrow q}$) of p on q can be quantified by the dependency of p on all the sub-subdisciplines of q. $D_{p \leftarrow q}$ is calculated with the following formula:

$$D_{p \leftarrow q} = \sum_{i=0}^{N_q} d_{p \leftarrow q_i}, p, q \in \{0, 1, ..., 9\}, p \neq q \quad (2)$$

3 EXPERIMENTS

3.1 Subdiscipline Classification

In LSTM[2], the Glove[3] [3] is used to initialize the word embedding. The dimension of each word vector is set to 256 and the output dimension of fully connected layer is set to 500. The dropout is set to 0.5, the batch size is set to 64, and the optimizer used is Adam Optimizer with a learning rate of 0.001. The max sentence length of LSTM is set to 200. The Glove word vector is also used in CNN and Transformer. Table 1 presents the average results of 10 runs on the two sets. LSTM performs better than others on both sets.

3.2 Proportion Analysis

In this section, we take the algebraic geometry as an example to conduct the proportion analysis, because the algebraic geometry occupies a central place in modern mathematics [4]. The trained (49-category) classifier is utilized to classify all the statements contained in algebraic geometry papers in the analysis data. Figure 1 illustrates the subdiscipline proportion tree of the algebraic geometry. The top four subdisciplines related to the algebraic geometry are number theory, algebra, geometry, and topology.

3.3 Dependency Analysis

In this section, we calculate the dependency between 10 subdisciplines and derive the dependency chains. Figure 2 shows the graph of mutual dependencies of the 10 subdisciplines. The graph illustrates that most subdisciplines depend on the algebra and the geometry because they are two basic

[3]https://nlp.stanford.edu/projects/glove/

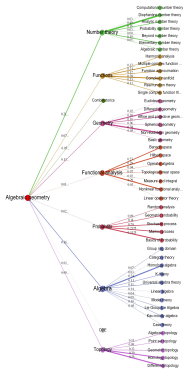

Figure 1: The subdiscipline proportion tree of the algebraic geometry. The values on the edges indicate the weights representing the strength of subdiscipline proportions.

subdisciplines. Dependency chains can be also obtained. Figure 2 reveals a typical chain from the analysis data among subdisciplines, that is, functional analysis (***1***) → function (***2***) → number theory (***3***) → algebra (***4***) → algebraic geometry (***5***).

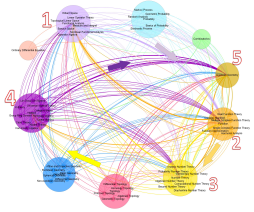

Figure 2: The dependency graph of the 10 subdisciplines.

4 CONCLUSION

This study investigates the (automatic) subdiscipline classification for a mathematical statement and performs inter-subdiscipline proportion and dependency analyses. The analysis results can help scholars understand mathematical researches more systematically. Our future work will focus on increasing the learning data and the analysis data by adding more subdisciplines.

5 ACKNOWLEDGMENTS

This work is supported by the Frontier science and technology innovation project (2019QY2404),NSFC (61673377), and Tianjin Nature Science Fund (19JCZDJC31300).

REFERENCES

[1] B. Mansouri et al. 2019. Characterizing searches for mathematical concepts. In *Proc. JCDL*. 57–66.
[2] H. Sak et al. 2014. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. In *arXiv:1402.1128*.
[3] J. Pennington et al. 2014. Glove: Global vectors for word representation. In *Proc. EMNLP*. 1532–1543.
[4] L. LI. 2011. The Application of Advanced Mathematics in Different Disciplines. *Sichuan University of Arts and Science Journal* 2 (2011).
[5] P. Swinnerton-Dyer. 2005. The justification of mathematical statements. *Proc. Philos. Trans. R. Soc. A* 363, 1835 (2005), 2437–2447.