TSUBAME 共同利用　令和 6 年度　学術利用　成果報告書

利用課題名　大規模学術データ分析基盤の開発
英文：Development of Scholarly Data Analysis Methods

利用課題責任者
桂井　麻里衣

同志社大学理工学部
Faculty of Science and Engineering, Doshisha University
https://mm.doshisha.ac.jp

邦文抄録（300 字程度）
本研究課題は、国内外の学術データ（論文、特許、ウェブページなど）からの効率的な情報抽出を目的とし、データ整形などの前処理からモデル学習、諸タスクでの応用までのパイプラインを構築するものである。本年度はその一環として、テキストやグラフを入力としたニューラルネットワークに基づく特徴量で論文や研究者を表現した。得られた表現を論文や研究者の推薦タスクに応用した結果、テキストに基づく手法（コンテンツベース手法）がグラフベース手法より高い性能を示した。また、日本語論文の抄録文分類モデルを訓練する実験では、英日翻訳技術との組み合わせによるデータ拡張が効果的であることを示した。

英文抄録（100 words 程度）
This research project aims to build a pipeline for efficient information extraction from scholarly data (papers, patents, websites). The pipeline spans from preprocessing and data formatting through model training to information retrieval and recommendation tasks. As part of this framework, neural networks were trained with text and graph inputs to represent different entities—papers and researchers—through output features. When applying the resulting representations to recommendation tasks for papers and researchers, content-based methods demonstrated higher performance than graph-based approaches. Furthermore, in experiments training classification models for Japanese paper abstracts, we demonstrated that data augmentation through the combination with translation technology is effective.

Keywords: scholarly data analysis, science mapping, information retrieval, neural networks, text embeddings

背景と目的

学術論文の数は年々急激に増加しており、人手による情報収集や整理・分析はますます困難となっている。そのため、膨大な学術データの中から所望の情報を効率的に抽出し、分析可能な形にする技術の確立が必要不可欠である。

本研究課題では、学術データの利活用のための主要技術として、本年度はテキストやグラフを入力としたニューラルネットワークに基づき論文や研究者の特徴を表現する手法を構築した。得られた表現を論文や研究者の推薦タスクに応用した結果、テキストに基づく手法（コンテンツベース手法）がグラフベース手法より高い性能を示した。また、グラフのリンク予測の観点での推薦性能と推薦結果の意外性はトレードオフの関係にあることが示唆された。

上記の研究と並行して、学術データのテキストを意味内容に応じて分類するモデルを構築した。これまでに英語のモデルは盛んに研究されてきたものの、それらを日本語に適応させるための学術データ資源は現状不十分である。そのため、効率的なデータセット構築を目的とした情報抽出手法を構築した。実験結果より、翻訳技術に基づくデータ拡張の有効性を示した。

概要

本研究課題は、学術データ（論文、特許、ウェブページなど）からの効率的な情報抽出に向けて、データ整形などの前処理からモデル学習、諸タスクでの応用までのパイプラインを構築するものである。

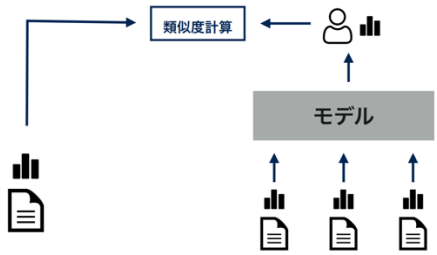

図 1:コンテンツベース手法の概要

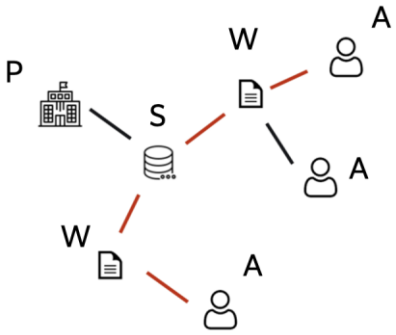

図 2:グラフベース手法とメタパスの概要

【学術データの埋め込み表現】

本年度は特に研究者への情報推薦に着目して研究を実施した。個々の研究者の業績一覧が入手できる場合、そのコンテンツ情報を各人の専門興味のモデル化に用いる手法をコンテンツベース手法とよぶ。一方、共著者情報まで入手できる場合、研究者や論文をノードとしたグラフ構造に基づくモデル化手法を適用できる。近年はこのようなグラフベース手法が注目を集めているが、コンテンツベース手法と同一条件下で性能を比較した例は報告されていない。そこで、学術情報推薦をグラフ内のリンク予測問題とみなし、これら二つのアプローチの性能を比較した。コンテンツベースとしては、図 1 に示すように、論文や研究者をタイトルテキストの埋め込みに基づいて表現する方法を採用した。グラフベース手法については、図 2 に示すように、著者（A）、論文（W）、発表場所（S）、出版社（P）という異なるエンティティ間のメタパスを事前に定義する手法と、メタパスを利用しない手法の 2 種類を用意した。

【文分類モデルの構築】

医学分野の論文抄録を背景、目的、手法、結果、結論のパートへ分類するタスクを実施するために、著者によって手動でラベル付けされている論文を大量収集し、データセットを構築した。そして、英語論文に基づく既存のデータセットを英日翻訳したものを擬似データセットとした。

表 1:著者間のリンク予測性能

手法	HR@1	HR@10	MAP@10
コンテンツベース	0.647	0.966	0.739
グラフベース（メタパス利用なし）	0.067	0.388	0.141
グラフベース（メタパス利用あり）	0.074	0.518	0.181

表 2:抄録文分類の性能評価

手法	Weighted-F1	Accuracy	Macro-F1
比較手法	90.12	90.53	83.49
提案手法	92.19	92.19	88.39

結果および考察

【学術データの埋め込み表現】

著者間のリンク予測実験において、上位 1 件と上位 10 件それぞれの Hit Rate（HR）、上位 10 件の Mean Average Precision を算出した結果を表 1 に示す。いずれの評価指標においても、コンテンツベース手法はグラフベース手法よりも高い性能を示した。グラフベース手法では、利用するメタパスが性能に影響するため、ドメイン知識に基づき慎重にメタパスを設計する必要があると考えられる。

リンク予測結果を精査すると、コンテンツベース手法はグラフベース手法に比べてグラフ内で距離が近いノード間の類似度を大きくする傾向があった。ゆえに、リンク予測を情報推薦に応用する場合、グラフベース手法の方が推薦結果に意外性が高いと考えられる。

【文分類モデルの構築】

構築した日本語論文データセットのみで言語モデルを訓練したものを比較手法、英語論文の翻訳に基づき拡張したデータセットで訓練したものを提案手法として、テ

ストデータでの文分類結果の Weighted-F1、Accuracy、Macro-F1 を算出した。表 2 に示すように、提案手法はいずれの評価指標においても比較手法を上回った。これにより翻訳を介したデータ拡張の有効性が示された。

まとめ、今後の課題

本研究課題では、テキストやグラフを入力としたニューラルネットワークに基づく特徴量で論文や研究者を表現し、論文や研究者の推薦タスクにおける性能を評価した。また、論文抄録の文分類モデル構築実験では、英日翻訳技術との組み合わせによるデータ拡張が効果的であることを示した。

今後の課題として、入力するグラフ構造の複雑さや特徴量の種類が研究者の特徴表現に与える影響を調査する必要がある。また文分類の研究については、広範な分野での実験と他言語への展開を予定している。