情報学分野における教育能力開発を目的とした論文書誌情報の自動分類

安川 美智子（群馬大学）

１．はじめに

大学における専門教育の質を向上するためには、授業担当者が教授する学問分野において、特に教育的な観点から、技術の進歩の傾向を把握しておくことが必要である。授業担当者は、担当科目に関して、十分な知識と研究業績のある研究者であるが、同じ専門分野に属する研究者であっても、大学の研究者と企業の研究者の間で、研究や教育に対する理念や方向性が異なるということがあり得る。具体的には、図１に示すように、学術的な価値の探求と実用的な価値の探求といった２つの異なる方向性があるが、大学で学生が学ぶ専門科目の内容は、学術的な方向と実用的な方向の両方の面から、バランスよく構成されることが望ましい。そこで、日本の情報学分野においては、大学と企業が連携してピアレビューを行い、情報専門教育の質向上のためのカリキュラム標準の策定が行われている[1]。これは世界標準である IEEE/ACM の情報専門教育の標準カリキュラム[2][3]を土台にしたものである。また、近年、情報学分野の専門教育は、分野内のみならず、すべての学問領域の教育全般において必要不可欠なリテラシーとなっており、情報学分野を取り巻く環境は急速に変化している[4]。さらに、最近は技術の変化のために、大学で学んだ専門知識が卒業後６年以内に役立たなくなる場合も多いという指摘がある[5]。

そこで、大学教育の質を向上するためには、研究者は自分の研究の方向性（図１の水平方向や垂直方向）に固執することなく、教育的な観点から学問分野の進歩の方向性（図 1 の点線の矢印）を見極めることが必要である。また、技術の急速な変化に対応するためには、授業担当者がデータ分析の手法とオープンデータを活用し、学問分野の進歩の方向性や傾向(図 1、図 2)を、即時に見極めていくことが重要であると考えられる。

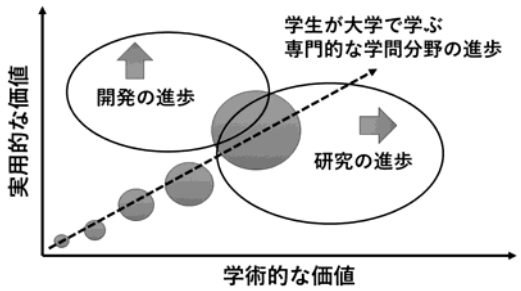

図１　研究、開発、学問分野の進歩の方向性

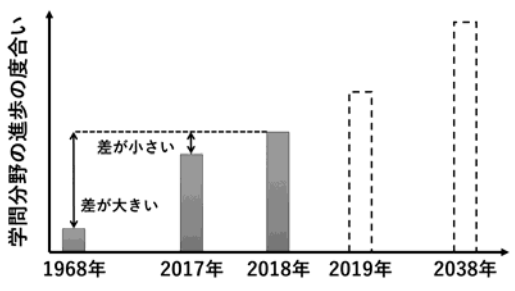

図２　西暦年毎の学問分野の進歩

以上のことを背景として、本研究では、情報学分野における学問分野の進歩の傾向を論文書誌情報のデータベースから分析する手法を提案する。本研究で明らかにする研究上の問い(Research Question)は以下の 2 点である。

[RQ1] 学問分野の進歩の傾向は、機械学習や深層学習の手法により、分析できるか。
[RQ2] 機械的な手法で分析した結果は、人間が目で見てわかる形で確認できるか。

２．実験データと実験方法

本研究では、情報学分野全般の論文書誌情報のオンラインデータベース DBLP をデータ分析の対象として利用した。DBLP に収録されているデータには、研究成果の発表形態のカテゴリが付与されている(表 1)。国際会議(conf)の論文書誌情報は最もデータ数が多く、その次に多いのが研究者のホームページ(homepages)の情報である。計算機科学の論文のプレプリントなど様々な論文が収録されている arXiv(表 1 の corr)は、DBLP ではジャーナル論文のサブカテゴリとなっているが、国際会議やジャーナル論文との重複データが多いことから、本研究では、arXiv 以外のジャーナル論文を分析対象とした。

DBLP には、ジャーナル論文の書誌情報として、論文タイトル、著者名、発表年、ジャーナルタイトル、ページ数、DOI などが収録されているが、本研究では、論文タイトルのみを分析対象とした。また、ジャーナル論文のラベルが付与されているデータの中には、研究論文ではないものも含まれているため、そのようなデータは実験対象から除外することとした。具体的には、文字列長が 30 文字未満の短いタイトルは表 2 に示すように、序文や書評などが多いため、分析対象の論文タイトルは 30 文字以上とした。

DBLP のデータベースは日々、更新されている。2019 年のデータは出版済みのものが随時、収録されているが、年度の途中はデータの収録状況が不完全なため、西暦年が 2018 年以前のものを分析の対象とした。

ジャーナル論文のタイトルには、研究の重要な特徴を表す単語が用いられている場合が多いことから、ある年と別の年の間で、学問分野の進歩（図 2）の度合いが大きければ、ジャーナル論文のタイトルには顕著な異なりがあり、自動分類の精度が高くなるはずである。そこで、機械学習および深層学習の手法を用いて、論文タイトルを「論文の新旧（最先端の論文と古典的論文)」の 2 つのカテゴリに自動分類する実験を行った。

表 1　DBLP の発表形態と発表数

発表形態	データ数
conf	2,485,385
homepages	2,345,632
journals(corr 以外)	2,098,927
journals/corr	226,632
others	150,050

表 2　文字列長が 30 文字未満の論文タイトル

論文タイトル	データ数
Editorial	4,981
Preface	2,671
Foreword	1,030
Introduction	885
Guest Editorial	647
Book Reviews	523

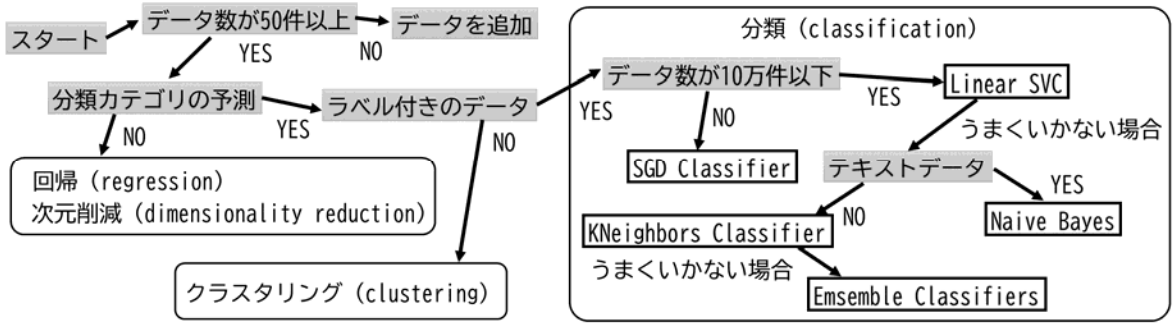

図 3　　scikit-learn の分類タスク用のアルゴリズム早見図(参考文献[6]の図より作成)

具体的には、Python のオープンソース機械学習ライブラリ scikit-learn[6]の文書分類タスク用の 5 つのアルゴリズムの実装(SDG Classifier, KNeighbor Classifier, Ensemble Classifier(Random Forest), Naive Bayes, Linear SVC(SVM Classification))を実験に利用した（図 3)。また、最近の深層学習の手法である Transformer[7]と BERT[8]のアルゴリズムを実装したプログラム[9]を利用した実験も行った。

文書分類の実験において、「新旧」の 2 つのカテゴリとしては、「新」カテゴリを 2018 年とした。また、「旧」カテゴリは、1 年前(2017 年)、5 年前(2013 年)、約 20 年前(1998 年～2003 年)、約 25 年前(1993 年～1998 年)、約 30 年前(1988 年～1998 年)、約 40 年前(1978 年～1988 年)、40 年以上前(1936 年～1978 年)として、実験を行った。実験の条件を統一するため、すべての分類実験において、新旧のカテゴリの文書数は、ランダムに選んだ 25000 件（学習用およびテスト用をそれぞれ 50%）とした。

機械学習の手法である Random Forest は、特徴量ごとの重要度を算出できる。そこで、各分類実験において、重要度の高い上位 30 個の特徴単語を抽出し、論文タイトルの中のどの単語が分類の根拠として用いられているかを調べた。

また、深層学習の手法である Transformer では、注意機構(Attention mechanism)と呼ばれる特徴量を関連づける仕組みがあり、特徴量の重みを可視化できる。この仕組みを利用して、分類カテゴリの予測が誤りであった場合と正しかった場合に、どの単語が重要とみなされたかを調べた。

３．実験結果

機械学習および深層学習による文書分類の精度(Accuracy)を表 3 に示す。MLkn、MLsdg、MLrf、MLnb、MLsvm は、それぞれ、 KNeighbor Classifier, SDG Classifier , Ensemble Classifier(Random Forest), Naive Bayes, Linear SVC(SVM Classification)に対応する。また、DLtr と DLtrb は、「(Bert を適用しない)Transformer」と「Bert を適用した Transformer」に対応する。また、図 4 と図 5 の棒グラフは、機械学習による分類と深層学習による分類の精度を示している。実験の結果から、最先端の論文と古典的論文は、年数の差が大きいほど、分類の精度が高いことが確認できた。

表 3　機械学習および深層学習による文書分類の精度

#years	MLkn	MLsdg	MLrf	MLnb	MLsvm	DLtr	DLtrb
1	0.51	0.53	0.52	0.53	0.52	0.53	0.54
5	0.57	0.61	0.58	0.60	0.60	0.63	0.63
20	0.71	0.75	0.73	0.74	0.75	0.75	0.74
25	0.76	0.79	0.77	0.79	0.79	0.77	0.76
30	0.77	0.81	0.78	0.80	0.81	0.80	0.81
40	0.83	0.86	0.84	0.85	0.86	0.85	0.87
40+	0.85	0.90	0.88	0.89	0.90	0.91	0.92

表 4 は、ランダムフォレストによる文書分類の特徴語である。All は、すべての分類において、重要な特徴語に含まれていた単語である。太字・下線付きの単語は、当該分類においてのみ、重要な特徴語とみなされた単語である。たとえば、iot（IoT; Internet of Things）は、2018 年の論文を 2013 年の論文と区別する上で、重要な特徴語であるとみなされている。具体的に調べてみると、2018 年と 2013 年の論文タイトルで IoT を含むものは、それぞれ 1117 件と 19 件であった。これに対して、略語ではない Internet of Things が、2018 年と 2013 年で、それぞれ 928 件と 94 件であった。2013 年当時は、IoT という略語が研究者コミュニティにおいて、まだ定着していなかったと考えられる。

図 6 は、「最先端の論文（2018 年）」と「古典的な論文（40 年以上前）」の論文タイトルを、深層学習の手法により分類した場合の重要単語の可視化である。上の 2 つ（予測が失敗した例）は、特徴的な単語が含まれておらず、末尾のピリオドに強く反応している。一方、下の 2 つ（予測が成功した例）は、最先端の論文に出現する adaptive や wireless が重要とみなされていることから、特徴的な単語への注意が有効に機能したと考えられる。

機械学習や深層学習の手法の分類結果は、単語レベルでの検証が可能ではあるが、特徴語として提示された単語は網羅的に抽出された専門用語ではない。各カテゴリの文書群をもとに、技術の進歩の傾向を詳しく調べるためには、専門用語（キーワード）自動抽出のツール[10]を使って、特徴語分析を行うことが必要であると考えられる。

４. おわりに

本発表では、情報学分野の授業担当者が、教育の質を向上させるための取り組みとして、何を行うべきかを検討し、教育的な観点から学問分野の進歩の傾向を分析する手法を提案した。また、情報学分野の書誌情報データベース DBLP と、機械学習および深層学習の手法を用いた実験を行い、提案法の実行可能性を検証した。

今後の課題としては、「最先端の論文」と「古典的な論文」の論文書誌情報から専門用語（キーワード）を抽出し、重要度の高い文献等（ハイレベルのジャーナルに掲載されている論文、教科書や参考書）に対する情報アクセスを効果的、かつ、効率的に行える手法を検討していくことを考えている。

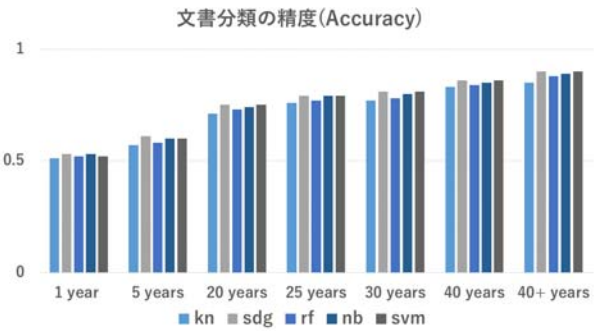

図 4　機械学習による文書分類の精度

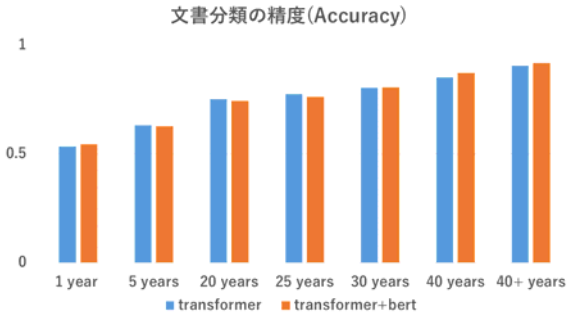

図 5　深層学習による文書分類の精度

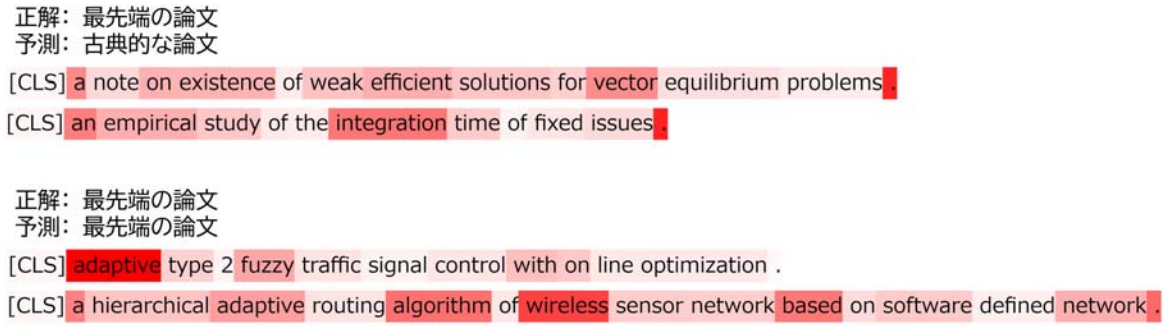

図 6　BERT を適用した文書分類タスクの Self-Attention の可視化

表 4　ランダムフォレストによる文書分類の特徴語

#years	特徴語
1	application, automatic, cloud, control, detection, hybrid, neural, object, optimization, parallel, power, scheme
5	application, automatic, control, **deep**, detection, efficient, fuzzy, hybrid, image, **iot**, scheme, **time**
20	application, cloud, control, detection, efficient, energy, fuzzy, hybrid, improved, note, optimization, scheme
25	application, automatic, cloud, control, detection, hybrid, neural, object, optimization, parallel, power, scheme
30	cloud, computer, control, detection, efficient, energy, improved, neural, note, optimization, parallel, power
40	cloud, computer, **design**, fuzzy, hybrid, image, improved, **program**, scheme, **smart**, study, **wireless**
40+	cloud, computer, **corresp**, efficient, fuzzy, hybrid, image, **language**, **mobile**, optimization, scheme, study
All	adaptive, algorithm, analysis, approach, based, data, framework, learning, method, model, multi, network, new, novel, problem, sensor, system, using

【謝辞】
本研究は JSPS 科研費(JP18K11986)、および、統計数理研究所共同研究プログラム(2019-ISMCRP-2045)の助成を受けたものです。

【参考文献】
[1] “カリキュラム標準”, 情報処理学会, 情報処理教育委員会, https://www.ipsj.or.jp/annai/committee/education/j07/ed_curriculum.html

[2] “Computing Curriculum Efforts”, IEEE/CS, https://www.computer.org/volunteering/boards-and-committees/professional-educational-activities/curricula

[3] “Advancing Education”, ACM, https://www.acm.org/education

[4] "Informatics for All: The Strategy", CECE, https://portalparts.acm.org/hippo/cecereport.pdf

[5] “Post-College Path No Longer So Clear”, VOA Education, https://learningenglish.voanews.com/a/post-college-path-no-longer-so-clear/5091058.html

[6] “Choosing the right estimator”, scikit-learn developers, https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

[7] “Attention Is All You Need”, Ashish Vaswani, et al., https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[8] “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Jacob Devlin., https://arxiv.org/abs/1810.04805

[9] “つくりながら学ぶ! PyTorch による発展ディープラーニング”, 小川雄太郎, マイナビ出版, 2019. https://github.com/YutaroOgawa/pytorch_advanced

[10] 専門用語（キーワード）自動抽出システム 言選 Web, Hiroshi Nakagawa, et al., http://gensen.dl.itc.u-tokyo.ac.jp/index.html