4K2-1

深層学習を用いた論文書誌情報による研究専門分野ラベリング

Research area clustering based on research paper information by deep learning technique

田中 和哉[*1] 荒川 陸[*1] 芝 慎太朗[*2] 森 純一郎[*1] 坂田一郎[*1]
Kazuya Tanaka Riku Arakawa Shintaro Shiba Junichiro Mori Ichiro Sakata

[*1] 東京大学大学院 工学系研究科
Graduate School of Engineering, The University of Tokyo

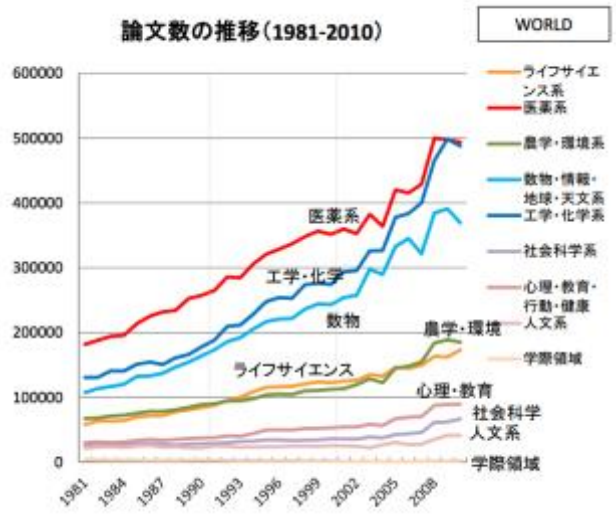
[*2] 東京大学大学院 総合文化研究科
Graduate School of Arts and Sciences, The University of Tokyo

This paper presents a deep-learning based method for research area clustering using abstract information of academic research paper information. We suggested the possibility of the automated distinction of research areas, since the current distinction of research areas is mainly based on journal level, which are defined by database provider. The detailed research area suggested by this research leads the possibility of more correct abstract landscape information for academic information.

1. 背景

科学・技術産業分野に関する情報量、論文(ジャーナル)の出版数など、は図1の通り、年々増加の一途をたどっている。したがって過去の情報量は指数関数的に増えている。

図1 論文数の歴史的推移



(出典: InCites より文部科学省が作成)

その巨大な情報の中から、技術イノベーションには論文などの基礎情報の正確な早期取得が重要である。つまり、技術経営や科学技術政策において、グローバルに流れる大量の電子化された論文情報を如何に、早期に正確に有用な情報を把握、判断するはますます重要であり難易度が上がっている[1]。

一般的に重要度を理解するために取られる手法に論文の被引用数によるその論文や雑誌の重要度を図る手法があり、それによりインパクトファクターや大学ランキング、研究者の評価にも使われている。[2] 被引用数を重要度として使う際に課題としてあげられることの一つに、分野間の差異が大きいことが挙げられる。つまり、違う分野での被引用数を横比較することはナンセンスであり、正確なその論文の分野の把握が望まれる。

しかし、自分の関連する分野を正確に把握することは難し区なってきている。トップジャーナルはもともと学際的である他、学術分野の複雑化、学際分野の増加により投稿ジャーナルがカバーする分野が多岐、変動しがちになっている。しかしながら、現状では、掲載雑誌を基にした論文の研究分野決めが行われており、論文毎での分野選定は専門家が手動で行うほかなく、データベース単位での選定は行われている例は少ない。

以前に当研究の一部著者らが発表した予備的な知見[3]があるが、それは主に直近の日本の論文を基に、単純に学習器を作成したものであるため、今回、深層学習の計算機科学的知見の調整を行い、実際の論文分野特定に使えるものを目指した。

田中和哉：東京大学大学院 工学系研究科
技術経営戦略学専攻

2. 目的

本研究では、深層学習の手法を用いて、論文の概要から論文毎での研究分野選定の手法の探索を行い、実際の論文分野特定に使えるものを目指した。

3. 手法

Clarivate Analytics(旧トムソン・ロイター IP&Science)が提供している自然科学中心の論文検索サイト: Wed of Science Core Collection (大学ライセンス、大学ランキングなどでも使われているデータベース)より論文データを取得した。全論文は投稿ジャーナルを元に分野が割り当てられていている。割り当て方は複数あるが、151 種類のデータベース独自のものを学習データとして用いた。

後処理の関係で取得コーパスの概要(アブストラクト)及び研究分野(マルチラベル)を抽出、gensim により、出現回数 5 回未満、頻度 10%以上の word と幾つかの stop word 除去を行った。その後、NLTK でステミング、Bag of words 等でデータセットを作成した[4]。 本データセットの一部を教師データとして深層学習を行い、精度測定を行った。

4. 実験評価

以前行った研究[3]で用いた下記の可変マルチラベル用の精度測定法と一般的な深層学習で用いられる精度測定の両方を行った。

可変マルチラベル用の精度測定法としては、予測したものが一つでも正解と当たっていたら正解とする。
つまり、予測確率をランキングで表して
1 位のものだけで予測した正解率を正解率①、
1,2 位のものだけで予測した正解率を正解率②

1,2,3 位のものだけで予測した正解率を正解率③とする
e.g.予測確率 B:0.4 A:0.3 C:0.2 D:0.1 正解 A
②と③のみで正例となる。

5. 結果と考察

論文書誌情報、特に概要を用いた研究分野識別器を実装、及び運用可能性を模索することに成功した。

以前の精度測定でも 90%以上の精度を示す[3]ことから、専門的な知見との比較を行ったが、本研究は実際の知見と同様レベルのものとして運用できる示唆を得た。

既存手法が各分野での専門家の識別によるもののため、正解率の包括的かつ正確な比較は行えなかったが、より詳細な分析を行える可能性について、今後の検討課題としたい。

6. 参考文献

[1] S. Iwami, J. Mori, I. Sakata, and Y. Kajikawa, "Detection method of emerging leading papers using time transition," *Scientometrics*, vol. 101, no. 2, pp. 1515–1533, Jul. 2014.

[2] QS: Quacquarelli Symonds, "QS World University Rankings® 2015/16 | Top Universities." [Online]. Available: http://www.topuniversities.com/university-rankings/world-university-rankings/2015. [Accessed: 30-Jan-2016].

[3] K. Tanaka, J. Mori, and I. Sakata, "Research area clustering based on research paper information by machine learning technique," *JSAI (The Japanese Soc. Artif. Intell.*, no. 2E3-3, 2016.

[4] J. Nam, J. Kim, and I. Gurevych, "Large-scale Multi-label Text Classification — Revisiting Neural Networks," *Lect. Notes Comput. Sci. (by Springer)*, vol. 8725, pp. 437–452, 2014.