2E3-3

論文書誌情報を用いた機械学習による研究分野クラスタリング

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

田中　和哉[*1]　　森　純一郎[*1]　　坂田一郎[*1]
Kazuya Tanaka　　Junichiro Mori　　Ichiro Sakata

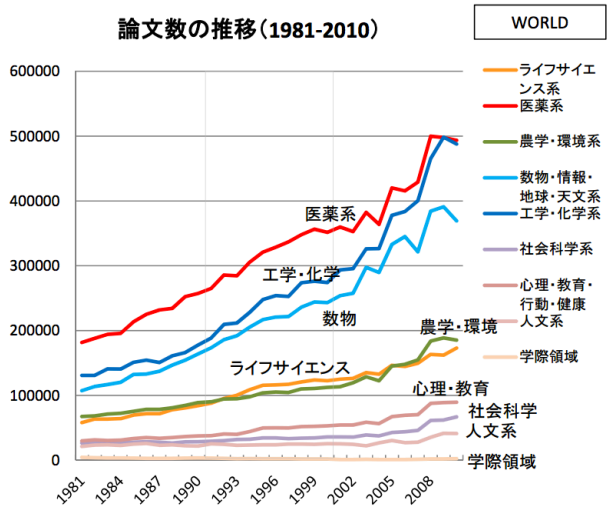
[*1] 東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

This paper presents a method for research area clustering based on abstract information of academic research paper information. The current distinction of research areas is mainly based on journal level, and this study approached distinction of research areas on paper level by machine learning and deep learning technique. We suggested the possibility of the automated distinction of research areas.

1. 背景

科学・技術産業分野に関する情報量は年々増加の一途をたどっており、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: 米国トムソンロイター社 InCites より文部科学省が作成)

その増加している情報の中から、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取り込み、早期に正確に有用な情報を把握する成長・成功への鍵となる[1]。

正確な把握のために一般的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ない。

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，坂田・森研究室

2. 目的

本研究では、機械学習、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った。

3. 手法と実験

米国トムソンロイター社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。データ数の関係から、2015 年の論文のうち日本の大学・研究所が執筆したもの(89, 875 件)を取得した。全論文は投稿ジャーナルを元に分野が 151 種類割り当てられている。（この 151 種類は参考に本論文の最後に記す。）

後処理の関係で取得コーパスのうち最新の 46,467 件の概要（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 でデータセットを作成した[3]。

データセットのうち 85%を教師データとして、3 層の深層学習にて 151 分類の学習器を生成した。

各パラメータとしては
[21771,1000,sigmoid],[1000,151,softmax]　Cost function は Cross entropy、Momentum は 0.9、Batch は 50 個ずつで 10epoch 学習させた。

4. 実験

今回データセットが可変のマルチラベルであることから下記のような正解率を定義した。

前提: 予測したものが一つでも当たってたら正解とする。
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。

今回のデータセットでは、下記のような結果を得た。
正解率①0.64 正解率②0.82 正解率③0.91
正解率③までを含めると論文ごとの分野識別はほぼ可能となった（cf. SVM では正解率③が 0.40）

5. 考察

現行の論文分野選定が投稿ジャーナル一括でつけたものが主なため正確な比較はできないが、専門分野より詳細な分野ラベル付けも一部可能となったと思われる。
(長文のため概要ではなくタイトルを示す)
例 1
タイトル
The challenge presented by progestins in ecotoxicological research
元データベースの研究分野ラベル
Environmental Sciences & Ecology
Engineering
学習器の予測した研究分野ラベル
Environmental Sciences & Ecology
Pharmacology & Pharmacy
Toxicology
タイトルから示されるように、学習器の予測した研究分野ラベルの方は正解した上に、詳細な研究分野分析を行えている例を見出すことができた。

例 2
タイトル
Comparison of the specificity, stability, and PCR efficiency of six rice endogenous sequences for detection analyses of genetically modified rice
元データベースの研究分野ラベル
Food Science & Technology
学習器の予測した研究分野ラベル
Agriculture
Plant Sciences
Biotechnology & Applied Microbiology
今回は不正解ではあるが、学習器の予測した研究分野ラベルはむしろ、より詳細な分析を行えた可能性もあり、今後の検討課題としたい。

6. まとめ

研究結果より，論文書誌情報、特に概要を用いた研究分野識別器を実装することに成功した。既存手法が各分野での専門家の識別によるもののため、正解率の比較は行えなかったが、学習器の予測した研究分野ラベルはより詳細な分析を行えていたり、むしろ、より詳細な分析を行えた可能性もあり、今後の検討課題としたい。

7. 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] 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.

8. 参考 – 151 分野

1 Agriculture（農学）
2 Allergy（アレルギー）
3 Anatomy & Morphology（解剖学、形態学）
4 Anesthesiology（麻酔学）
5 Anthropology（人類学）
6 Behavioral Sciences（行動科学）
7 Biochemistry & Molecular Biology（生化学、分子生物学）
8 Biodiversity & Conservation（生物多様性保全）
9 Biophysics（生物物理学）
10 Biotechnology & Applied Microbiology（バイオテクノロジ、応用微生物学）
11 Cardiovascular System & Cardiology（循環器系、心臓学）
12 Cell Biology（細胞生物学）
13 Critical Care Medicine（集中治療医学）
14 Dentistry, Oral Surgery & Medicine（歯科学、口腔外科、口腔内科）
15 Dermatology（皮膚病学）
16 Developmental Biology（発生生物学）
17 Emergency Medicine（救急医学）
18 Endocrinology & Metabolism（内分泌学、新陳代謝）
19 Entomology（昆虫学）
20 Environmental Sciences & Ecology（環境科学、生態学）
21 Evolutionary Biology（進化生物学）
22 Fisheries（水産業）
23 Food Science & Technology（食品科学、食品技術）
24 Forestry（林学）
25 Gastroenterology & Hepatology（消化器病学、肝臓学）
26 General & Internal Medicine（一般医療、内科学）
27 Genetics & Heredity（遺伝学、遺伝）
28 Geriatrics & Gerontology（老年医学、老年学）
29 Health Care Sciences & Services（ヘルスケア科学、サービス）
30 Hematology（血液学）
31 Immunology（免疫学）
32 Infectious Diseases（感染症）
33 Integrative & Complementary Medicine（統合医療、代替医療）
34 Legal Medicine（法医学）
35 Life Sciences & Biomedicine（生命科学、生体臨床医学） - その他のトピック
36 Marine & Freshwater Biology（海洋生物学、淡水生物学）
37 Mathematical & Computational Biology（数理生物学、計算生物学）
38 Medical Ethics（医学倫理）
39 Medical Informatics（医療情報学）
40 Medical Laboratory Technology（臨床検査室技術）
41 Microbiology（微生物学）
42 Mycology（菌類学）
43 Neurosciences & Neurology（神経科学、神経学）
44 Nursing（看護）
45 Nutrition & Dietetics（栄養、栄養学）
46 Obstetrics & Gynecology（産科学、婦人科学）

47 Oncology（腫瘍学）
48 Ophthalmology（眼科学）
49 Orthopedics（整形外科学）
50 Otorhinolaryngology（耳鼻咽喉科学）
51 Paleontology（古生物学）
52 Parasitology（寄生生物学）
53 Pathology（病理学）
54 Pediatrics（小児科学）
55 Pharmacology & Pharmacy（薬理学、薬学）
56 Physiology（生理学）
57 Plant Sciences（植物学）
58 Psychiatry（精神医学）
59 Public, Environmental & Occupational Health（公衆衛生学、環境衛生学、労働衛生学）
60 Radiology, Nuclear Medicine & Medical Imaging（放射線学、核医学、医用画像）
61 Rehabilitation（リハビリテーション）
62 Reproductive Biology（生殖生物学）
63 Research & Experimental Medicine（研究、実験医学）
64 Respiratory System（呼吸器系）
65 Rheumatology（リウマチ学）
66 Sport Sciences（スポーツ科学）
67 Substance Abuse（物質乱用）
68 Surgery（外科学）
69 Toxicology（毒物学）
70 Transplantation（移植）
71 Tropical Medicine（熱帯医学）
72 Urology & Nephrology（泌尿器学、腎臓学）
73 Veterinary Sciences（獣医学）
74 Virology（ウイルス学）
75 Zoology（動物学）
76 Astronomy & Astrophysics（天文学、宇宙物理学）
77 Chemistry（化学）
78 Crystallography（結晶学）
79 Electrochemistry（電気化学）
80 Geochemistry & Geophysics（地球化学、地球物理学）
81 Geology（地質学）
82 Mathematics（数学）
83 Meteorology & Atmospheric Sciences（気象学、大気科学）
84 Mineralogy（鉱物学）
85 Mining & Mineral Processing（採鉱、選鉱）
86 Oceanography（海洋学）
87 Optics（光学）
88 Physical Geography（自然地理学）
89 Physics（物理学）
90 Polymer Science（高分子科学）
91 Thermodynamics（熱力学）
92 Water Resources（水資源）
93 Acoustics（音響学）
94 Automation & Control Systems（オートメーション、制御システム）
95 Computer Science（コンピューターサイエンス）
96 Construction & Building Technology（土木技術、建築技術）
97 Energy & Fuels（エネルギー、燃料）
98 Engineering（工学）
99 Imaging Science & Photographic Technology（イメージングサイエンス、写真技術）
100 Information Science & Library Science（情報科学、図書館学）
101 Instruments & Instrumentation（機器、計装）
102 Materials Science（物質科学）
103 Mechanics（力学）
104 Metallurgy & Metallurgical Engineering（冶金、冶金工学）
105 Microscopy（顕微鏡検査）
106 Nuclear Science & Technology（核科学、核技術）
107 Operations Research & Management Science（オペレーションズリサーチ、経営科学）
108 Remote Sensing（リモートセンシング）
109 Robotics（ロボット工学）
110 Science & Technology（科学、技術） - その他のトピック
111 Spectroscopy（分光学）
112 Telecommunications（電気通信）
113 Transportation（交通運輸）
114 Architecture（建築）
115 Art（芸術）
116 Arts & Humanities（芸術、人文） - その他のトピック
117 Asian Studies（アジア研究）
118 Classics（古典）
119 Dance（ダンス）
120 Film, Radio & Television（映画、ラジオ、テレビ）
121 History（史学）
122 History & Philosophy of Science（科学史、科学哲学）
123 Literature（文学）
124 Music（音楽）
125 Philosophy（哲学）
126 Religion（宗教）
127 Theater（演劇）
128 Archaeology（考古学）
129 Area Studies（地域研究）
130 Biomedical Social Sciences（生医学社会科学）
131 Business & Economics（ビジネス、経済学）
132 Communication（通信）
133 Criminology & Penology（犯罪学、刑罰学）
134 Cultural Studies（文化研究）
135 Demography（人口統計学）
136 Education & Educational Research（教育学、教育研究）
137 Ethnic Studies（民族研究）
138 Family Studies（家族研究）
139 Geography（地理学）
140 Government & Law（政府、法学）
141 International Relations（国際関係）
142 Linguistics（言語学）
143 Mathematical Methods In Social Sciences（社会科学の数学的手法）
144 Psychology（心理学）
145 Public Administration（行政学）
146 Social Issues（社会問題）
147 Social Sciences（社会科学） - その他のトピック
148 Social Work（社会事業）
149 Sociology（社会学）
150 Urban Studies（都市研究）
151 Women's Studies（女性学）

(研究分野 – Web of Science ヘルプ より
https://images.webofknowledge.com/WOKRS59B4_2/help/ja/WOS/hp_research_areas_easca.html)