1A3-3

ニュース記事と特許を利用した科学技術の重要性の評価

Evaluation of the Industrial and Social Impacts of Science and Technology Using Patents and News Articles

飯沼 俊平 Shumpei IINUMA　福田 悟志 Satoshi FUKUDA　難波 英嗣 Hidetsugu NANBA　竹澤 寿幸 Toshiyuki TAKEZAWA

広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

In the fields of scientometrics and citation analysis, several measures for evaluating the industrial relevance or the impact of academic research fields have been proposed. However, they could not evaluate recent industrial relevance or impact of each field, because most of them rely on the citations between research papers and patents. In this paper, we propose a method to evaluate industrial and social impact of research fields. Our method classifies research papers and news articles in terms of various classification systems, such as the International Patent Classification (IPC) or the KAKEN classification index. Then we evaluate the industrial and social impact of each field by comparing the number of documents for each IPC and KAKEN category.

1. はじめに

「学術研究が産業にどれだけ貢献しているか」という観点から，学術研究を評価したり，産業との学術研究の関連性を分析したりする試みが，科学計量学の分野を中心に行われている．学術研究の評価は，優れた研究の発見や，研究の学問的質の維持のために行われ，資金配分を行う研究課題の選定などに用いられてきた．このような評価では，同じ分野の研究者が学問的観点から評価・検証を行うピアレビューが中心であったが，研究開発活動による社会・経済的効果の重要性が増し，実際にどのような効果が生まれたかを評価することが必要とされるようになった．しかし，分野内部で専門家が評価するピアレビューや論文の被引用数にもとづく手法では，研究開発による社会・経済効果を把握することは困難であるため，学術的観点にとらわれず，さまざまな観点からの評価が必要である．

本研究では，「具体的な技術開発目標が実際に達成されたのか」という観点から，特許に着目し，学術論文を特許分類体系に従って自動分類することで，特許分類体系で見た場合の学術論文の発表件数を調査する．また，学術論文自動分類技術を，開発・実用化に関するニュース記事に適用し，特許分類体系に従って分類する．これにより，「どのような分野の技術が社会的な関心を集めているか」という観点から，科学技術の学術界外部への影響度の評価，学術論文と特許の出願傾向との比較を行う．特許分類体系とニュース記事を用いることで，産業界や社会への影響を調査することができると考えられるが，学問分野の分類体系は，資金配分機関により公正なピアレビューが行われるために重要となる．科学研究費助成事業データベース（KAKEN）*1 における「系・分野・分科・細目表」は，科研費の審査のために作成され，適宜改訂が行われており，研究分野の重要性を評価する際の分類体系として有用であると考えられる．本研究では，上述した特許とニュース記事を利用した分析に加え，学術論文を KAKEN の研究分野分類細目に従って自動分類し，KAKEN での採択課題数と学術論文の発表件数の比較を行う．

本論文の構成は以下のとおりである．2 節では，関連研究について述べる．3 節では，学術論文およびニュース記事の自動分類手法について述べる．4 節では，学術論文およびニュース記事を自動分類し，特許の出願数と KAKEN での採択課題との比較を行った実験について述べる．最後に 5 節で本論文をまとめる．

連絡先:
飯沼 俊平，福田 悟志，難波 英嗣，竹澤 寿幸
広島市立大学大学院 情報科学研究科
〒 731-3194 広島市安佐南区大塚東三丁目 4 番 1 号
{iinuma,fukuda,nanba,takezawa}@ls.info.hiroshima-cu.ac.jp

*1 http://kaken.nii.ac.jp

2. 関連研究

本節では，学術研究の産業への影響度の計測，ジャンル横断情報アクセス，学術論文の自動分類に関する関連研究について述べる．

学術研究の産業への影響度の計測

投資や研究資金の分配といった政策決定の際の根拠としての利用という面から，特許や論文などを対象にした技術動向分析技術への関心は近年高まっている．実際に，「学術研究が産業にどれだけ貢献しているのか」という観点から，学術研究を評価したり，産業と学術研究の関連性を分析したりする試みが，科学計量学の分野を中心に行われるようになってきた．その典型的な手法は，Narin らの研究に代表されるように，特許と論文間の引用関係に着目したものである [Narin 94]．Narin らは，アメリカ，イギリス，旧西ドイツ，日本，フランスの 5ヶ国の論文および特許間の引用関係を調査し，論文=科学 (Science)，特許=技術 (Technology) と見なすことで，各国の科学の国内および国外の技術への影響度について分析している．特許と論文間の引用関係を用いた同様の分析は，例えばレーザー医学 [Noyons 94] や宇宙工学 [Schmoch 91] などの特定の分野を対象にしたり，ある特定の地域からの発表論文や出願特許を対象にしたりするなど [Coronado 03]，分析の切り口を変え，様々な側面から行われてきている．

しかし，ある文献が他の文献から引用されるまでには一定の期間が必要であるため，こうした引用に基づく手法では，最新の影響度を測ることができる保証が必ずしもない．本研究では，文献間の引用ではなく，文書分類技術を使うことにより，この問題の解決を試みる．

ジャンル横断情報アクセス

ジャンル横断検索や，文書分類に関して様々な研究が行われてきた．NTCIR-3 で実施された技術動向調査タスクでは，与えられた新聞記事と関連する特許を検索する，という課題が設定された [Iwayama 02]．このタスクでは，各ジャンルの文書で使われる用語の違いを考慮することが 1 つのポイントとなる．たとえば,「社長」という単語は新聞記事中に比較的高頻度で出現するが，特許中では出現頻度が低い．このため，一般的に使われる逆文書頻度（IDF）を単語の重み付けに用いた場合，同じ単語でも新聞記事と特許では重要度が大きく異なる．ジャンル間の用語の違いは，論文と特許にも存在する．例えば,「磁気記録媒体」のように特許中では一般的に使われるが，論文中では全く用いられない．難波らは，特許，論文間の引用関係や，特許から自動作成したシソーラスなどを用いて，論文用語を特許用語に自動変換する手法を提案している [難波 09]．たとえば，論文用語「フロッピーディスク」は特許用語「磁気記録媒体」に自動変換され，この技術を用いることで論文用語で関連特許の検索が可能となる．

NTCIR-7 および NTCIR-8 特許マイニングタスクは，ある分野の特許と論文から技術要素と効果の対を抽出し，技術動向マップを自動的に作成することを目標としている [Nanba 10]．特許マイニングタスクでは，次の 2 つのサブタスクを設定している．

- 学術論文分類サブタスク: 論文抄録に，国際特許分類（IPC）コードを自動付与し，ある分野に関する論文と特許の網羅的収集を可能にする．
- 技術動向マップサブタスク: 技術要素とその効果を表す表現を，論文と特許から自動抽出し，技術動向マップとしてまとめる．

学術論文分類サブタスクでは，論文に付与する国際特許分類（IPC）コード数が 30,855 件と非常に多く，また，訓練用データが 350 万〜450 万件と膨大であるため，多くのグループが k-Nearest Neighbor （k-NN）法を用いた．本研究では，1 つめのサブタスクで用いられたジャンル横断文書分類技術を用いて，特許分類体系から見た学術論文とニュース記事の傾向を分析する．

学術論文の自動分類

Fukuda らは，科学研究費助成事業データベース（KAKEN）の研究分野が付与された採択課題データを訓練用データとして用い，学術論文の自動分類を行っている [Fukuda 13]．特に，技術要素とその効果を表す表現を重要と考え，k-NN 法を用いることで，高い分類精度を得ている．本研究では，福田らの自動分類技術を用いて KAKEN での採択課題と学術論文の傾向を分析する．

3. 学術論文およびニュース記事の自動分類

本研究では，次の 2 つを明らかにすることを目標としている．

- 国際特許分類から見た,「特許の出願傾向」,「学術論文」,「ニュース記事」の違い．
- KAKEN の研究分野分類から見た，採択課題数と学術論文の発表件数の違い．

1 つめの分析を行うために，k-NN 法に基づく分類器を用いて，学術論文とニュース記事に国際特許分類（IPC）コードを付与する．IPC は，国際的に統一されて用いられている分類体系であり，特許文献の技術内容によって上から順に「セクション」,「クラス」,「サブクラス」,「メイングループ」,「サブグループ」の 5 階層から構成分類されている．本研究では「サブクラス」までの分類コード（643 カテゴリ）の付与を目的とする．

分類器は特許検索システムを利用しており，検索システムは，内容語（名詞，動詞，形容詞）を索引語とし，類似性尺度として BM25 を採用している．入力された学術論文またはニュース記事に対し，特許検索システムが出力した上位 k 件の文書 $\{d_1, d_2, \ldots, d_k\}$ から，式 1 に基いて IPC コードを順位付けし，スコアが最も高い IPC コードを自動付与する [Xiao 08]．

$$score_{Listweak}(c) = \sum_{i=1}^{k} occur(c, d_i) sim(q, d_i) r^i \qquad (1)$$

式 1 において，$occur(c, d_i)$ は文書 d_i に IPC コード c が付与されている場合に 1，そうでなければ 0 となる．$sim(q, d_i)$ は，入力された文書（論文またはニュース記事）と特許検索システムが出力した文書 d_i との類似度を表す．r^i は，順位の低い文書に対するペナルティ係数であり，$r = 0.95$ を用いる．

2 つ目の分析を行う際は，要素技術と効果に関する記述を重要と考えて単語の重み付けを行い，k-NN 法を用いた Fukuda らの自動分類手法（[Fukuda 13]）を用いて，KAKEN の研究分野分類細目に従って学術論文に科研費コードを付与する．

4. 実験

特許出願件数との比較のため，学術論文分類器を構築し，学術論文とニュース記事への IPC コードの付与，集計を行った．また，分野ごとの採択課題数と論文の発表件数との比較を行うため，学術論文に科研費コードを付与し，集計を行った．

4.1 分類器の構築と評価

学術論文に IPC コードの付与を行うための分類器を構築した．k-NN 法を適用する際に利用した文書データを表 1 に示す．日本国公開特許公報および米国特許データは全文データに加え，IPC コードが付与されており，それぞれ，日本語論文，英語論文の分類器に利用した．

構築した分類器は NTCIR-8 で使用されたデータを用いて評価した．日本語論文用分類器を日本語サブタスクで使用された論文データに適用した結果，精度 0.815 が得られた（k=300）．同様に，英語論文用分類器を英語サブタスクで使用された論文データに適用した結果，精度 0.656 が得られた（k=300）．

表 1: 分類器に用いた文書データ

データ名	年	文書数
日本国公開特許公報	1993-2012	6,910,194
米国特許	1993-2012	2,895,149

4.2 IPC コードと科研費コードの付与

分析に用いたデータを表 2 に示す．学術論文データとして，JST 科学技術文献データ（2003 年〜2012 年），約 700 万件を用いた．文献データには，タイトル，著者名，出典情報が記載されている．ニュース記事は，読売新聞記事データ（1993 年〜2012 年）のうち，見出しに「開発」または「実用化」を含む記事 8,674 件を用いた．新聞記事に加え，TechCrunch[*2] の

*2 http://techcrunch.com/

記事（2005 年 6 月〜2013 年 12 月）120,596 件を分類対象とした．TechCrunch は米国のニュースサイトであり，主に IT 系のニュースを配信している．比較対象として，日本国特許公開公報（1993 年〜2012 年）での出願件数を用いた．

上述した JST 科学技術文献データと読売新聞記事データに，日本語用論文分類器を用いて IPC コードの付与を行った．同様に，英語論文用分類器を TechCrunch の記事に適用し，IPC コードの付与を行った．また，Fukuda らの k-NN 法ベースの分類器を利用して，科学技術文献データに科研費コードを付与した．Fukuda らの分類器では，論文の表題を対象とした研究分野分類で精度 0.827 を得ている．

表 2: 分析に用いたデータ

対象データ	付与コード	付与件数
JST 科学技術文献データ（書誌情報）	国際特許分類	約 85 万
	科研費コード	6,533,269
読売新聞（開発・実用化記事）	国際特許分類	8,674
TechCrunch（IT 系英文ニュース）	国際特許分類	120,596

4.3 結果

国際特許分類分野別の科学技術文献データ数（上位 10 件）を表 3 に示す．特に，G06F（電気的デジタルデータ処理），A61K（医薬品），C12N（微生物，酵素）に分類された文献数が多いことから，国際特許分類で見た場合，これら分野での研究が活発に行われていると考えられる．

表 3: 国際特許分類分野別の科学技術文献データの数（上位 10 件）

IPC	説明	件数
G06F	電気的デジタルデータ処理	94,943
A61K	医薬品	65,059
C12N	微生物，酵素	62,034
G01N	材料の調査または分析	49,391
H01L	半導体	42,070
H04N	画像通信・テレビジョン	18,847
G02F	光の制御	17,428
A61B	診断，手術，個人認識（医学）	15,845
C01B	非金属元素およびその化合物	15,613
A01G	園芸	12,728

次に，国際特許分類分野別の日本国特許の出願件数，科学技術文献データ数およびニュース記事（読売）の件数上位 10 件の分野を表 4 に示す．ニュース記事では，A23L（食品，食料品）や A61K（医薬品）が上位に存在し，A01G（園芸）などの特許や科学技術文献データでは上位 10 件に含まれなかった分野が見られる．国際特許分類では，セクション A は「生活必需品」と定められており，日用品の実用化，開発に関する研究が社会的関心を集めやすいことが分かる．特許と科学技術文献データを比べると，分野の割合にばらつきが見られることから，研究の活発さ（文献数）は必ずしも特許出願には結びつかないと考えられる．G06F（デジタルデータ）や A61K（医薬品），H04N（画像通信・テレビ）に関しては，すべてのジャンルで上位 10 中に含まれている．これらの分野は，研究が特許に結びつきやすく，社会的に見ても重要であるといえる．TechCrunch の記事を分類した結果では，約 9 割の記事が G06F（デジタルデータ）または G06Q（データ処理）に分類された．

図 1 に KAKEN の採択課題と科学技術文献データの分野の割合を示す．KAKEN では，研究分野が,「分野」,「分科」,「細目」の 3 階層で分類されており，Fukuda らの分類器により第 3 階層の「細目」レベルまで分類されたが，カテゴリ数が非常に多いため，第 1 階層のカテゴリ分類結果を示している．文献データは「工学」の割合が最も高いが，採択課題では「医歯薬学」の割合が最も多く，資金配分機関がこの分野を特に重要と考えていることが分かる．国際特許分類体系でみた場合でも，「医薬品」よりも「デジタルデータ」に関する特許出願数，科学技術文献データ数が多かった．医学分野は，他の分野に比べると多くの分野に細分化されるため，分野ごとの論文件数でみると数が少なくなることが原因の一つであると考えられる．ただし，どの分野をどれだけ細分化するかはカテゴリ設計者の観点によるため，一つの文献に対し複数のカテゴリを付与し，分析を行う必要がある．

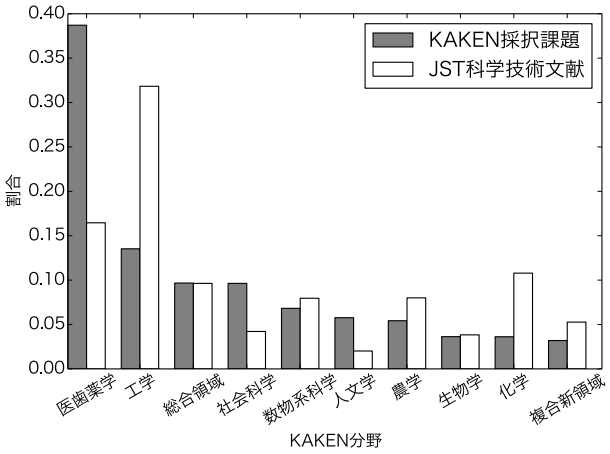

図 1: KAKEN 採択課題と科学技術文献データの比較

5. おわりに

本研究では，文書を他ジャンルの分類体系で分類することで，従来とは異なる観点からの比較が可能となることを示した．学術論文とニュース記事に IPC コードを付与することで，特許の出願件数との比較を行った結果，特許の出願や論文とくらべ，ニュース記事では食品や医薬品などの生活との結びつきが強い分野での実用化，開発に関する記事が多く，それらの分野に対する社会的関心が高いことなどが分かった．また，学術論文に科研費コードを付与し，KAKEN の採択課題数と科学技術文献データ数を比較した．文献の多さと，採択課題数には相関が見られなかった．

研究分野に対する社会的関心を調査するために，ニュース記事を用いたが，TechCrunch（IT 系ニュース）に適用した場合，当然のことながら，そのほとんどが G06F（デジタルデータ）または G06Q（データ処理）に分類された．さらに，関連分野内で注目を集めている技術を把握するためには，グループレベルの IPC コードを付与する必要がある．ただし，ニュース記事では専門性が高くなるに連れて，技術そのものよりもその技術がどう使われるか，といった点に着目される可能性があるため，請求項や詳細説明との類似度のみを用いた手法では分類コードの付与が難しいと考えられる．この点が改善できれ

表 4: 特許の出願件数と，学術論文，開発・実用化に関するニュース記事の比較（上位 10 件）

日本国特許		科学技術文献データ		ニュース記事（読売）	
IPC	説明	IPC	説明	IPC	説明
H01L	半導体	G06F	デジタルデータ	G06F	デジタルデータ
G06F	デジタルデータ	A61K	医薬品	G06Q	データ処理
H04N	画像通信・テレビ	C12N	微生物，酵素	A23L	食品，食料品
G03G	電子写真	G01N	材料の調査・分析	A61K	医薬品
G11B	情報記憶	H01L	半導体	H04N	画像通信・テレビ
G02B	光学装置	H04N	画像通信・テレビ	C12N	微生物，酵素
B41J	タイプライタ	G02F	光の制御	G01N	材料の調査・分析
A61K	医薬品	A61B	診断，手術（医学）	H04M	電話通信
G01N	材料の調査・分析	C01B	非金属元素	A01G	園芸
H01M	電池	H01M	電池	G09B	教育用器具

ば，IPC コードの付与だけでなく，ニュース記事を該当特許そのものに結びつけることにつながり，さらに詳細な分析が可能となるだろう．

謝辞

本研究成果は，2014 データサイエンス・アドベンチャー杯（主催：SAS Institute Japan 株式会社， 独立行政法人科学技術振興機構）で得られたものです．

参考文献

[Coronado 03] Coronado, D. and Acosta, M.: The Effects of Regional Scientific Opportunities in Science-Technology Flows: Evidence from Scientific Literature Cited in Firms' Patent Data, in *ERSA conference papers ersa03p321*, European Regional Science Association (2003)

[Fukuda 13] Fukuda, S., Nanba, H., Takezawa, T., and Akiko, A.: Classification of Research Papers Focusing on Elemental Technologies and Their Effects, in *Proceedings of the 6th Language & Technology Conference*, LTC'13, pp. 366–370 (2013)

[Iwayama 02] Iwayama, M., Fujii, A., Kando, N., and Takano, A.: Overview of Patent Retrieval Task at NTCIR-3, in *Working Notes of the 3rd NTCIR Workshop Meeting, Part III: Patent Retrieval Task*, pp. 1–10 (2002)

[Nanba 10] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-8 Workshop, in *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp. 293–302 (2010)

[Narin 94] Narin, F., Olivastro, D., and Stevens, K. A.: Bibliometrics / Theory, Practice and Problems, *Evaluation Review*, Vol. 18, pp. 65–76 (1994)

[Noyons 94] Noyons, E. C. M., Raan, van A. F. J., Grupp, H., and Schnoch, U.: Exploring the Science and Technology Interface: Inventor-author Relations in Laser Medicine Research, *Research Policy*, Vol. 23, pp. 443–457 (1994)

[Schmoch 91] Schmoch, U., Kirsch, N., Lay, W., Plescher, E., and Jung, K. O.: Analysis of Technical Spin-off Effects of Space-related R&D by Means of Patent Indicators, *Acta Astronautica*, Vol. 24, pp. 353–362 (1991)

[Xiao 08] Xiao, T., Cao, F., Li, T., Song, G., Zhou, K., Zhu, J., and Wang, H.: KNN and Re-ranking Models for English Patent Mining at NTCIR-7, in *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 333–340 (2008)

[難波 09] 難波 英嗣, 釜屋 英昭, 竹澤 寿幸, 奥村 学, 新森 昭宏, 谷川 英和：論文用語の特許用語への自動変換, 情報処理学会論文誌. データベース, Vol. 2, No. 1, pp. 81–92 (2009)