２Ａ０５

科学技術イノベーション政策立案のためのデータプラットフォーム
－テキストマイニングによる科学技術分野の同定－

○原田 裕明[*1], 小柴 等[*2], 池内 健太[*3], 原 泰史[*4], 黄 俊揚[*4], 黒田 昌裕[*1,4]
(*1 科学技術振興機構, *2 科学技術・学術政策研究所,
*3 経済産業研究所, *4 政策研究大学院大学)

1. 概要

科学技術イノベーション政策における「政策のための科学」(SciREX)推進事業の一環として、エビデンス・データを基に政策オプションを合理的に導出する研究を進めており[1][2]、特にその基盤となるエビデンス・データベースとして、科学技術用語をキーにして、その分野の競争的研究資金、その成果である論文・特許などの成果を一気通貫に検索できるシステムを構想した。そのシステムについて先に予備調査をおこなったところ、時代とともに変化する科学技術用語に対して、あらかじめ設定された分類コードだけでデータセット間を対応づけることが難しいことが明らかになった[3]。

今回、その解決策として、科学技術用語と科研費の研究課題の間のように関係があいまいで流動的な部分に対して、テキストマイニング技術を応用して対応付けを学習し、同定する手法を用いることにした。この同定で対応付けをおこなった部分と、データセット間ですでに明確に対応付けが整備されている部分（研究課題とその成果等）と統合することにより、科学技術用語からその分野の競争的資金、その成果を通して検索、集計することが原理的には可能となる。

本稿ではこの同定手法を組み込んだ試作システムの概要とその検索性能について報告する。

2. データプラットフォームの全体像

図１に今回試作したデータプラットフォーム・システムの概要を示す。複数のデータセットを使用しており、それぞれの提供元も掲示した。

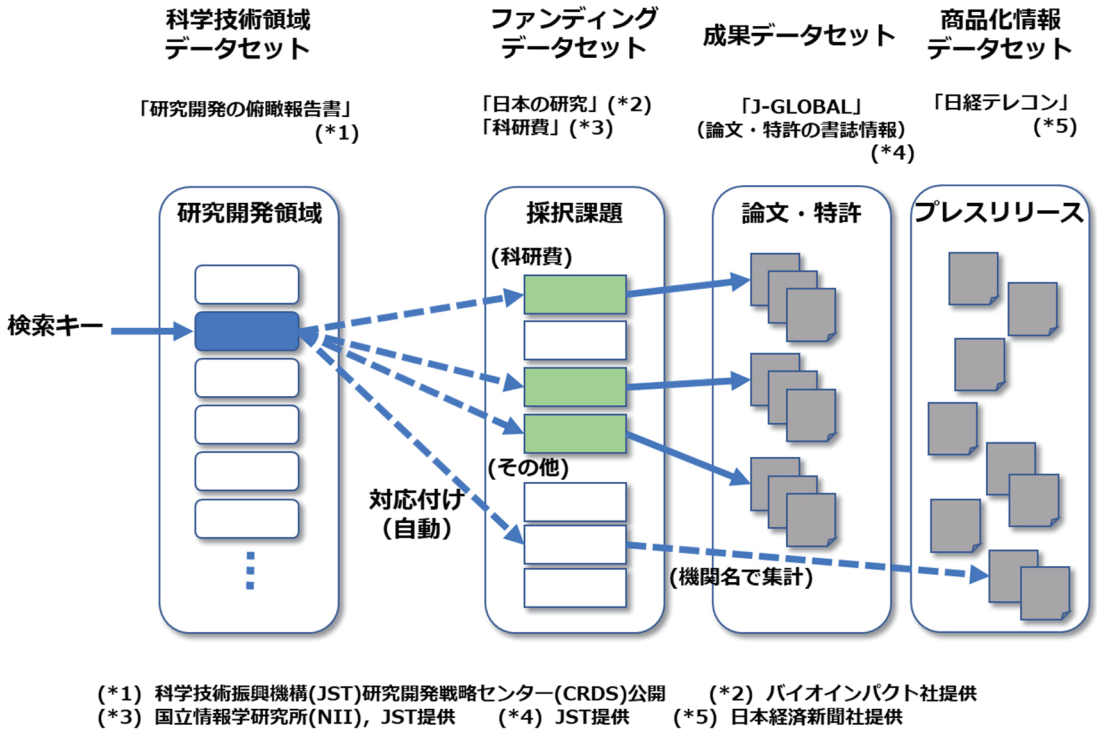


図 1. データプラットフォームの概要

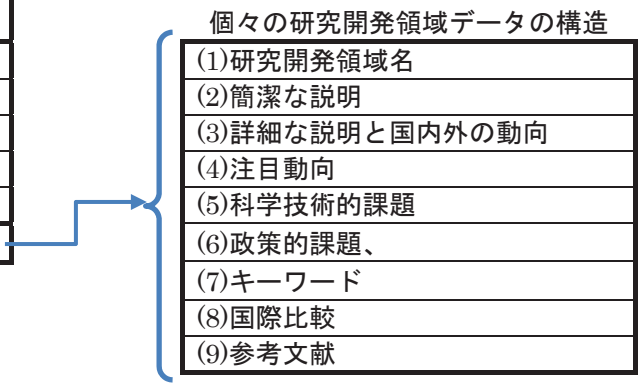
[*1] E-mail: hiroaki.harada@jst.go.jp

科学技術用語には用語を特徴付ける説明文章が付属している必要があるため、JST-CRDS が発行している「研究開発の俯瞰報告書（2015 年版）」[1]に記載されている研究開発領域名を科学技術用語として利用した。俯瞰報告書は表 1 のように「環境エネルギー」、「ライフサイエンス・臨床医学」、「ナノテクノロジー・材料」、「情報科学技術」、「システム科学技術」の 5 つの分野に分けられ、合計 355 件の研究開発領域を解説している。個々の研究開発領域は、(1)研究開発領域名～(9)参考文献、という 9 項目に構造化されている。

なお科学技術用語の事典としては今回の俯瞰報告書に限らず、幅広い科学技術分野をカバーしている、個々の用語の説明が詳細で構造化されている、などの条件を満たすデータセットであれば同様に使用できる。

表 1.　俯瞰報告書の構成

分野	研究開発領域数
環境エネルギー	92
ライフサイエンス・臨床医学	77
ナノテクノロジー・材料	41
情報科学技術	91
システム科学技術	54
合計	355



他方、ファンディング・データセットとして参照した”日本の研究.com”[2]には現在、国内で公開されているほとんどの競争的研究資金公募情報が網羅されている。個々の研究課題は、表 2 のように(1)研究課題名、(2)研究課題情報、(3)研究概要に構造化されている。表 2 の(2)研究課題情報は対象とするすべての競争的研究資金事業に共通の形式である。なお競争的研究資金のうち、科学研究費（科研費）[3]については NII、JST 提供のデータセットを使った。研究成果である論文、特許の一覧表からリンク可能なものは書誌情報データセット J-GLOBAL[4]へリンクしている。

表 2.　ファンディング・データにおける研究課題の構造

項目	サブ項目	内容
(1)研究課題名		
(2)研究課題情報	研究課題番号	事業ごとに付与された管理番号
	研究期間	○年度～○年度
	競争的研究資金の事業名	○○省○○事業○○プログラム等
	研究者名（代表者名）、所属機関名	
	研究費	各年度額および全期間の総額
	関連キーワード	事業ごとに付与されたもの
	成果論文の一覧表	J-GLOBAL へのリンク（可能なもの）
	成果特許の一覧表	J-GLOBAL へのリンク（可能なもの）
(3)研究概要	目的、方法、成果等	事業ごとに様式は異なる

3.　同定手法

テキストマイニングは検索のキーとなる科学技術用語と、競争的資金に採択された研究課題の間の類似性を抽出する。

これらの対象に対して、次の手順で処理をおこなう。

---

1 https://www.jst.go.jp/crds/report/report02/index.html　2017 年版も発行されている。

2 “日本の研究.com”（株式会社バイオインパクト）　https://research-er.jp/

3 科学研究費助成事業（日本学術振興会）　https://www.jsps.go.jp/j-grantsinaid/index.html

4 科学技術総合リンクセンター（JST）　http://jglobal.jst.go.jp/

① 専門用語辞書の作成….”日本の研究.com”キーワード辞書等を利用
② 専門用語辞書をもとに科学技術用語の形態素解析[5]と単語出現回数ベクトルの作成…. 特に 1)研究開発領域名と 7)キーワードに出現した単語には大きな重みを付けた上で、log スケールに変換
③ 研究課題の形態素解析と単語出現回数ベクトルの作成…. 特に 1)研究課題名に出現した単語には大きな重みを付けた上で、log スケールに変換
④ 両者間のコサイン類似度計算….類似度の高い組合せを同定結果として選出する

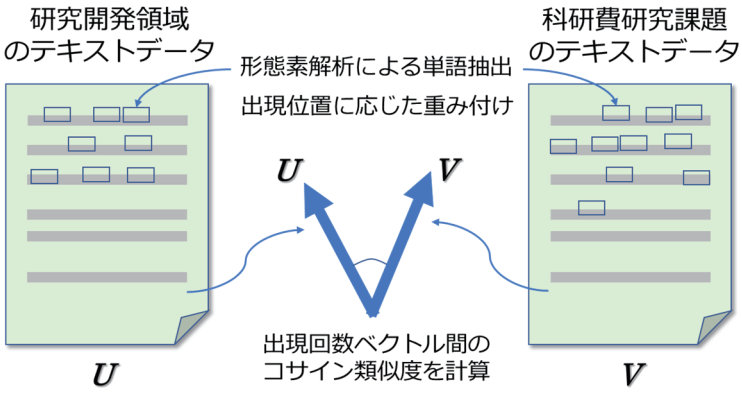


図 2. テキストの同定処理の概要

以上の同定方法を用いて、任意の「研究開発領域名」（科学技術用語）から「研究課題」（ファンディング・データ）へ自動対応付けをおこなうことにより、特定の領域に関連する研究課題、研究者、研究機関、研究費、リンクしている成果論文・特許の書誌情報もまとめて収集できる。さらに図 1 のように「プレスリリース」情報も加えて多面的な検索ができるように構成した[6]。

データプラットフォームの主な機能、ユーザーインタフェース等については[4]に詳しい。

4. エキスパートによる判定

図 3. エスキパート判定システムの画面例

今回提案した自動対応付け方法が、各研究分野の専門家の判断とどの程度の一致があるかを確認するために、エキスパートによるサンプル実験をおこなった。

「研究開発の俯瞰報告書」の「環境エネルギー」、「ライフサイエンス・臨床医学」、「ナノテクノロジー・材料」、「情報科学技術」の 4 分野を編纂したエキスパート計 5 名[7]により、4 分野から 2 領域ずつ計 8 領域を選択し、各領域に同定された選出課題各 100 件の適合度を判定してもらった（計 800 件）。図 3 には領域「バイオマス」に同定された研究課題がランダム順に表示された例を示す。

表示された研究課題の概要文を 1 件ずつ参照しながら、”その領域に属する典型的な研究課題”とみなされるものを”○”, ”近いもの”を”△”（以上を真陽性）、”まったく違うもの”を”×” 、”判断できない”を”？”（以上を偽陽性）とマークしてもらうという方法を採った。この実験により、検索の適合率（precision）が推察できる。（現状では全数確認していないため、再現率（recall）は不明である。）

[5] オープンソース MeCab を使用。http://taku910.github.io/mecab/
[6] プレスリリースについては、データセットの著作権を配慮して、現在は研究機関名ごとに集計したプレスリリース件数のみ表示している。
[7] いずれも JST 研究開発戦略センターに所属。

表 3 に判定結果を示す。800 件に対する平均値として真陽性（＝適合率）は約 5 割となった。しかし領域によってバラツキがあり、たとえば「リサイクル技術」、「免疫」は適合率 7 割を越えている一方で、「新型炉」は 3 割に達していない。このような適合率の全体的な向上とバラツキの減少が自動対応付けにとっての当面の目標である。

適合率の低い領域についてその原因を調べると、たとえば「新型炉」と「高分子医薬品」にはともに単語「核」が出現することが多い、等の傾向が見られた。ここから、辞書の強化やシソーラス[8]を活用する、等の対策が効果を持つことが期待できる。

表 3.　エスキパート判定の結果　（数値はいずれも%）

分野	判定対象とした研究開発領域名	適合 ○	やや適合 △	不適 ×	不明 ？
情報科学技術	CPS/IoT アーキテクチャー	25	35	36	4
	CPS/IoT セキュリティー	4	38	57	1
ライフサイエンス・臨床医学	免疫	62	14	21	3
	高分子医薬品（抗体医薬）	11	25	62	2
ナノテクノロジー・材料	超低消費電力ナノエレクトロニクス	54	8	38	0
	データ駆動型材料設計（マテリアルズ・インフォマティクス	16	16	68	0
環境エネルギー	新型炉（核融合を含む）の研究・開発	11	10	79	0
	リサイクル技術（都市鉱山含む）	57	28	15	0
平均		30	22	47	1
判定（適合率）		真陽性（52）		偽陽性	

5.　今後の課題

以上のデータプラットフォーム構築について、今後対応すべき課題をまとめる。

(1) 同定方式の性能改善：　テキストマイニングによる自動対応づけは膨大な研究課題の検索を大幅に省力化できる効果がある。反面、4 で述べたエキスパート判定も踏まえ、検索性能（適合率、再現率）を広範に測った上で、上記の辞書強化など偽陽性減少の対策を立てる必要がある。本来の政策オプション導出の目的のために許容できる適合率、再現率の程度を見きわめる必要もある。

(2) 政策オプション導出に向けたシステム統合：　現状のデータプラットフォームの機能はファンディングからその研究成果までの検索支援にとどまっている。最終目的である政策オプション導出に至るためには、さらに(a)科学技術イノベーションが社会・経済に与える影響を考察するシナリオプラニング、(b)そのシナリオプラニングに寄与する予測技術[5]、(c)シナリオに沿った経済的影響のシミュレーション、などを構築して組み合わせていく必要がある。

【謝辞】競争的研究資金に関する情報をご提供いただいた株式会社バイオインパクト殿、科研費および J-GLOBAL に関する情報をご提供いただいた JST 情報企画部殿に感謝いたします。

【参考】

[1] JST 研究開発戦略センター「変動の時代に対応する科学技術イノベーション政策のためのエビデンスの整備と活用に向けて」、CRDS-FY2015-RR-01、2015.3

[2] JST 研究開発戦略センター「科学技術イノベーション政策の科学における政策オプションの作成～ICT 分野の政策オプション作成プロセス～」、CRDS-FY2015-RR-07、2016.3

[3] 原田裕明、他「科学技術イノベーション政策立案のためのデータプラットフォームー投資と成果のデータ対応調査」、研究・イノベーション学会第 31 回年次学術大会 2G16、東京、2016.11.5

[4] 原泰史、他「特許・論文・ファンドデータを活用した研究活動可視化システムの開発および運用－SciREX 政策形成インテリジェント支援システム」、産学連携学会第 15 回大会 0616E1515-4、宇都宮、2017.6.15

[5] 小柴等、他「政策形成に資する科学技術予測支援システムの開発と現状」、産学連携学会第 15 回大会 0616E1515-2、宇都宮、2017.6.15

[8] 意味の上下関係、同義/類義関係などによって体系づけた類語辞書。