

NISTEP 定点調査における自由記述のテキストマイニング手法 を用いた基礎的分析

01900990 東京海洋大学 *渡部 大輔 WATANABE Daisuke
文部科学省 楠木 亮介 KUSUNOKI Ryosuke
防災科学技術研究所 松本 拓郎 MATSUMOTO Takuro
01002750 政策研究大学院大学 大山 達雄 OYAMA Tatsuo

1. はじめに

研究を遂行する上で、人材や時間、資金など様々な資源を利用して行われるが、近年では競争的資金や校費などの資源配分の方法が多様化してきている。文部科学省科学技術・学術政策研究所では、産学官の一線級の研究者や有識者への継続的な意識調査を通じて、我が国の科学技術やイノベーション創出の状況変化を把握するために、「科学技術の状況に係る総合的意識調査」(NISTEP 定点調査)を2006年より毎年実施してきた。本研究では、NISTEP 定点調査の自由記述を対象として、文章型データを統計的に分析するテキストマイニングを通じて、研究者の資源配分に関する意見の特徴を計量的に分析することを目的とする。

2. 分析の概要

NISTEP 定点調査では、第2期(2011-15年)と第3期(2016-20年)において、自由記述のデータベースが公開されている[1, 2]。本研究では、それぞれをDB2015, DB202と呼ぶこととする。両社は、ほぼ同じサンプル数であるが、実施年によって自由意見の質問事項が異なることに注意が必要である。

テキストマイニングの分析ツールとしては、KH Coder 3を用いる[3]。独自のタグとして「研究費」と「科研」を設定した。

3. 時点別の比較分析

3.1 抽出語リスト

データの中から語を自動抽出し、多く出現した語を確認する。名詞とサ変名詞を対象とした主要な抽出語の出現数は表1の通りとなっており、DB2015と2020ともに、「資金」「研究費」「予算」が上位に挙がっており、特にDB2020では「資金」「研究費」の順位、出現数ともに大幅に上昇していることが分かる。

3.2 共起ネットワーク

共起とは出現パターンの似通った語であり、共起ネットワークは共起関係を線(edge)で表している。抽出語の出現回数は円の大きさ、距離が近い抽出語同士は同じ色の円で表現され、共起関係の強さは相

関係数とともに線の濃さで表現される。共起関係の尺度として、Jaccard 係数を用いて、サブグラフ検出(modularity)を行った。DB2020の結果を図示すると、図1のようになる。DB2015と2020ともに、中心性高い抽出語として「研究」「大学」が得られた。資源配分に関連した用語として、「資金」は「競争」、「研究費」と「予算」は「配分」が関係の高い用語として得られた。DB2020では「資金」と「研究費」が同じクラスタとしてつながりがより強まったことが分かる。

表1 出現数の多い抽出語

順位	DB2015		DB2020	
	抽出語	出現数	抽出語	出現数
1	研究	29452	研究	29046
2	大学	8280	大学	9366
3	技術	5990	資金	4752
4	科学	5165	人材	3814
5	企業	4434	企業	3445
6	人材	3795	研究費	3091
7	予算	3366	学生	2831
8	分野	3234	社会	2691
9	社会	3090	予算	2670
10	教育	3019	教育	2668
11	資金	2925	状況	2485
23	研究費	2037		
	科研	507	科研	1043

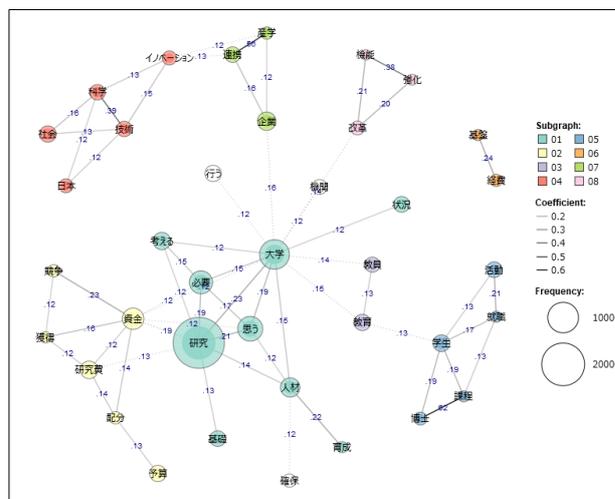


図1 共起ネットワーク (DB2020)

4. 属性別の比較分析 (DB2020)

4.1 出現確率の高い特徴語

これまでの分析で、DB2020の方がより資源配分に関する意見が多く出ていることが分かったことから、本節ではDB2020を対象として分析を行う。特徴的な言葉を抽出するために、類似性測度を Jaccard 係数、集計単位を回答セル数として分析を行う。

所属機関別に分析した結果は表 2 のように、大学等は「資金」「研究費」、公的研究機関は「資金」「予算」が上位となった一方、民間企業等とその他では資源配分に関しては記述少ない。

職位別に分析した結果は表 3 のように、主任研究員・准教授は「資金」「研究費」、研究員・助教は「資金」が上位となった一方、社長・役員・学長等、部課長・教授では資源配分に関しては記述少ない。

4.2 対応分析

2つの成分を抽出した上で、差異が顕著な語(上位50語)を対象として2次元のプロットを行うことで、抽出語とともに属性の類似性を分析する。所属機関別に分析した結果は図2のようになる。これより、大学等が中央に近い位置していることから、全ての属性に関係が見られる一方、他の3つの属性の間の類似性はとても低いことが分かる。

資源配分に関連した用語に着目すると、「資金」は中央に近い位置していることから、全ての属性に関係することが分かる。大学等は「研究費」の他に、「科研」「経費」「時間」という抽出語と関係が見られる。公的研究機関は「予算」の他に、「組織」「機関」「事務」という抽出語と関係が見られる。

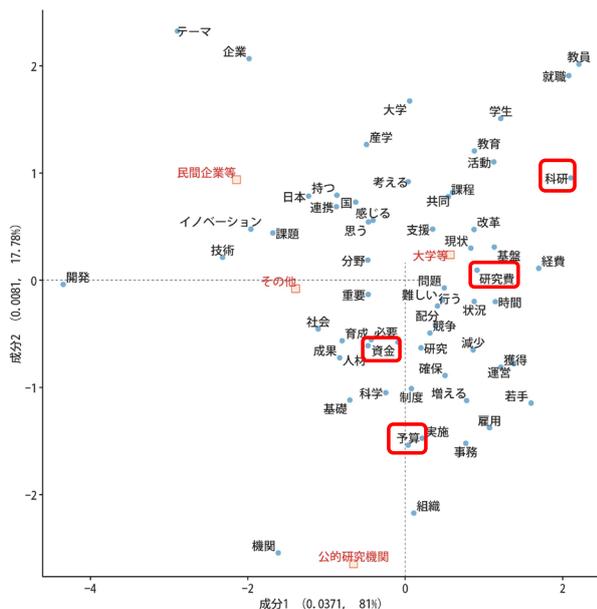


図2 対応分析 (所属機関別, DB2020)

表2 出現確率の高い抽出語 (所属機関別, DB2020)

順位	大学等	公的研究機関	民間企業等	その他
1	研究	研究	企業	政策
2	大学	必要	大学	繋がる
3	資金	機関	思う	施策
4	考える	資金	必要	兆し
5	研究費	人材	技術	含む
6	学生	予算	開発	動き
7	教育	社会	人材	見える
8	状況	基礎	感じる	提案
9	行う	成果	連携	理解
10	教員	技術	イノベーション	経営

表3 出現確率の高い抽出語 (職位別, DB2020)

順位	社長・役員、学長等	部課長、教授	主任研究員、准教授	研究員、助教
1	大学	社会	研究	研究
2	必要	成果	思う	思う
3	企業	基礎	資金	感じる
4	人材	連携	感じる	大学
5	連携	科学	学生	資金
6	考える	評価	研究費	状況
7	技術	イノベーション	時間	学生
8	社会	問題	考える	若手
9	基礎	重要	活動	多い
10	イノベーション	競争	状況	行う

5. おわりに

本研究では、NISTEP 定点調査の自由記述を対象として、抽出語リスト、共起ネットワーク、特徴語、対応分析などテキストマイニングを通じて、研究者の資源配分に関する意見の特徴を計量的に分析した。その結果、研究者の資源配分に関する用語の出現確率が高い属性として、所属機関別では大学等と公的研究機関、職位別では主任研究員・准教授と研究員・助教ということが明らかになった。今後、質問項目との対比を含めた分析を行うとともに、抽出語を基にした自由記述に関する内容の分析を行う必要がある。

謝辞

本研究はSciREX 共進化実現プログラム(第IIフェーズ)「研究業績の評価に基づく資源配分効果の動態的特性分析のための基礎研究」の一環として実施された。

参考文献

- [1] 文部科学省科学技術・学術政策研究所「科学技術の状況に係る総合的意識調査(NISTEP 定点調査2015)」, NISTEP REPORT, No.166, 2016.
- [2] 文部科学省科学技術・学術政策研究所「科学技術の状況に係る総合的意識調査(NISTEP 定点調査2020) データ集」, NISTEP REPORT, No.190, 2021.
- [3] 樋口耕一: 社会調査のための計量テキスト分析—内容分析の継承と発展を目指して【第2版】、ナカニシヤ出版、2020.