ISSN 2186-5647

P-93

論文要旨のテキストマイニング分析による研究トレンドの年次変化の比較に関する一考察

- 大学等研究組織の研究力に関する組織比較研究 -

日大生産工（学部） ○平野 未来　日大生産工　水上 佑治

1 はじめに

文部科学省は2015年、全国86国立大学を研究力をもとに「世界」、「特色」、「地域」に分類、大学に分類に即した運営を課した[1]。本稿は、各大学が分類に即した運営を行うためには、まず、組織の研究活動のトレンドを把握する必要があるとの立場に立ち、その手法を示すことを目的とする。分析では、論文要旨にテキストマイニング分析を施して、組織の研究トレンドの年次変化を示した。分析例では、群馬大学「地域」を中心に考察を行う。

2 研究の方法

論文情報は、Clarivate Analytics社の論文データベースWeb of Science(WoS)を用いて収集した。分析対象は、2004年から2018年までの群馬大学の論文である。論文数の推移を表１に示す。

分析では、Text Mining Studioの単語頻度解析により、3年毎の計15年を対象とし、3文字以上名詞（数詞以外）の上位頻度50単語を抽出した。上位頻度の単語の選定にあたり、2004年から2018年までで、一般的に使用される英単語を除外し、3年毎5回分の解析を行った。

本稿では、Mizukami et.al. (2017)を参考にして、研究の広がりをローレンツ曲線とジニ係数で議論する[2]。一般に、ローレンツとジニ係数は、社会学の分野で富の分配具合の評価で用いられている[2]。本稿では、テキストマイニングで得られた使用頻度の高い語句のリスト（上位50位）をもとにして、このローレンツカーブとジニ係数にて分析を施し研究の広がり具合を評価した。

またことばの繋がりをネットワーク図で図示し、話題の傾向を分析する共起ネットワーク分析を2004~2018年で行った。

3 分析結果

群馬大学の論文の2004年から2018年までの名詞(数詞以外)を対象に、ことばの繋がりを見る共起ネットワーク分析を図１に示す。

また3年毎の計15年を対象とした、単語頻度解析の上位頻度50語から、ローレンツ曲線とジニ係数を図２に示す。それぞれジニ係数は 2004-2006 は 0.221 、 2007-2009 は 0.205 、 2010-2012 は 0.217 、 2013-2015 は 0.228 、 2016-2018は0.222とあまり数値自体に大きい変化は見られなかった。

表1. 群馬大学論文数の推移

年	論文数	年	論文数
2004	578	2012	577
2005	578	2013	609
2006	641	2014	643
2007	578	2015	626
2008	583	2016	644
2009	579	2017	655
2010	566	2018	342
2011	558		

表2. 群馬大学論文単語頻度解析

	2004-2006	2007-2009	2010-2012	2013-2015	2016-2018
1	patient	patient	patient	patient	patient
2	cell	cell	cell	cell	treatment
3	treatment	treatment	treatment	treatment	cell
4	protein	mouse	mouse	mouse	disease
5	presence	protein	protein	disease	mouse
6	tumor	tumor	presence	tumor	tumor
7	mean	condition	tumor	protein	survival
8	mouse	disease	property	presence	cancer
9	property	presence	condition	whereas	outcome
10	order	property	subject	condition	relationship

A Study on the Annual Changes Comparison of Research Trends by Text Mining Analysis of Abstracts of Papers
－ Comparative study on research capabilities of research organizations such as universities －

Mirai Hirano, Yuji MIZUKAMI

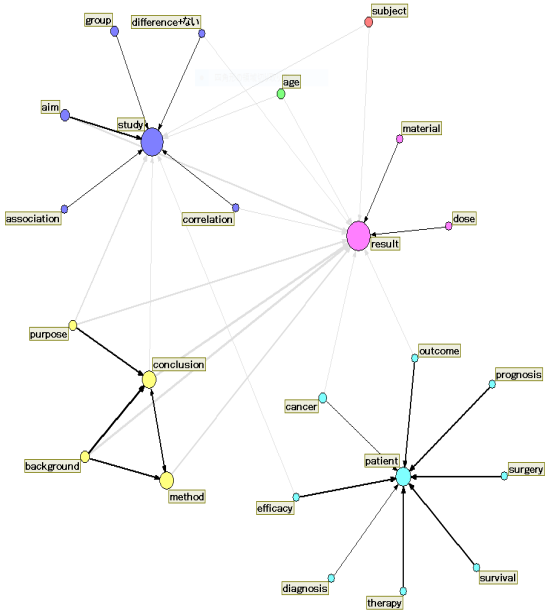

図１．共起ネットワーク

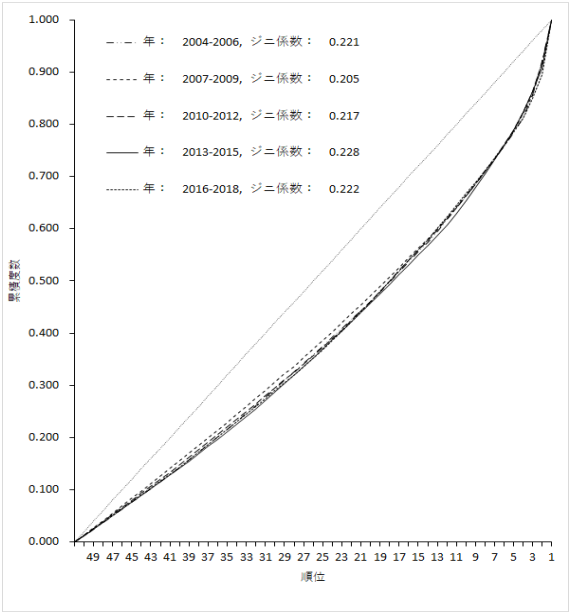

図２．　ジニ係数とローレンツ曲線

4　考察

本稿の分析の結果、３つの傾向を示すことができた。まず、図１の群馬大学の論文から共起ネットワーク分析を行ったところ、"patient"などが強い繋がりを示しており、"cancer", "therapy"など医学関係の用語が傾向として表れることを示した。また表２から５つの単語が3年毎の計15年で全てで上位10単語に入る結果となった。またその５つの単語も医学関係の用語であった。

次に図２から、ジニ係数で議論すると、2007年から2012年にかけて研究の広がりの減少、または、一部の研究への集中を示唆する結果を得た。その後、2013年から2015年にかけて研究の広がりが増加、2016年以降は、2006年以前の水準になっている。

そしてローレンツカーブで議論すると、2007年から2009年が他のデータに比べ異なる傾向を示している。具体的には、語句の使用頻度が上位5位までに集中している傾向がある。"patient", "cell", "protein", "treatment", "mouse"という語句を含む研究が他の年に比べて活発化していたと考えられる。

以上のことより、群馬大学では医学の分野に長けた論文が多いことがわかる。またその各年で流行している話題名などにも強く影響されると考えられる。

謝辞

本研究は統計数理研究所共同研究プログラム(30-共研-4212 研究者の異分野融合度と多様度の客観的な評価指標研究の深化)の助成を受け、統計数理研究の資産を活用したものである。

参考文献

[1]「国立大学法人の現状・取組・課題」、一般社団法人 国立大学協会, 2015

[2]Yuji Mizukami, Yosuke Mizutani, Keisuke Honda, Shigenori Suzuki, Junji Nakano, An International Research Comparative Study of the Degree of Cooperation between disciplines within mathematics and mathematical sciences, Springer, Behaviormetrika, Vol.44, issue 2, pp.385-403, 2017

[3]　アマルティア・セン（2000）『不平等の経済学』（鈴村興太郎・須賀晃一訳）東洋経済新報社