

インダストリー4.0：

トピックモデル分析によるビックデータ研究に関する一考察

日大生産工（学部） ○伊藤 慎太郎 日大生産工 水上 祐治

1. はじめに

ビックデータとは、従来のデータベース管理システムなど記録や保管、解析が困難であった巨大データ群のことを指す。近年のコンピュータやソフトウェア技術の進歩により活用が進み、それに伴いビックデータ研究の論文数が増加している[1]。本稿はビックデータ研究の動向を、テキストマイニングのトピックモデル分析を中心にした分析を行い考察するものである。

2. 先行研究と本稿

中村ら[2]は、テキストマイニング分析を用いてX線天文衛星SUZAKUの年代別の研究トレンドを抽出、その年次的推移を示す手法を示した。また、Mizukamiら[3]は、インダストリー4.0に着目して、その技術要素あるビックデータ、IoT、AIの研究が活発化していることを指摘している。

本稿では、ビックデータ研究を題材として、テキストマイニング分析の諸分析を施し研究動向を調査する。分析手法は、コレスポンディング分析に加えて、トピックモデル分析を併せて用いる。また、トピックモデル分析と階層的クラスタ分析の結果を比較し、分析精度を検証する。

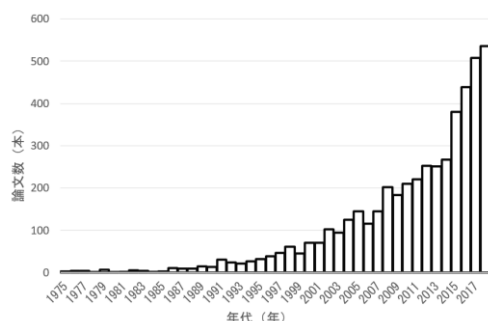


図 1：ビックデータ研究の論文数の推移

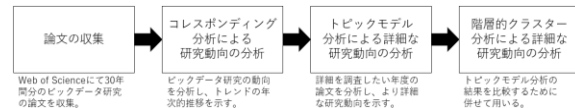


図 2：分析チャート

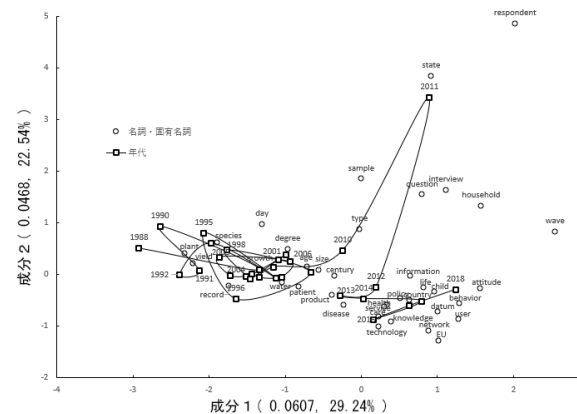


図 3：コレスポンディング分析

3. 研究方法

分析データは、オンラインデータベース Web of Science に掲載されている、1988 年から 2018 年までの 30 年間に発表されたビックデータ研究の論文とする。

分析では、まずテキストマイニングソフトである KH Coder 3 を用い、収集した論文の要旨をコレスポンディング分析し、ビックデータ研究のトレンドの年次的推移を示す。

次に、詳細を調査したい年度について統計ソフト R を用い、その年度に発表された論文の要旨をトピックモデル分析する。同様に、トピックモデル分析と比較するために、KH Coder 3 を用いて階層的クラスタ分析を行う。

Industry 4.0:
A study on the Big Data research by topic model analysis

Shintaro ITO, Yuji MIZUKAMI

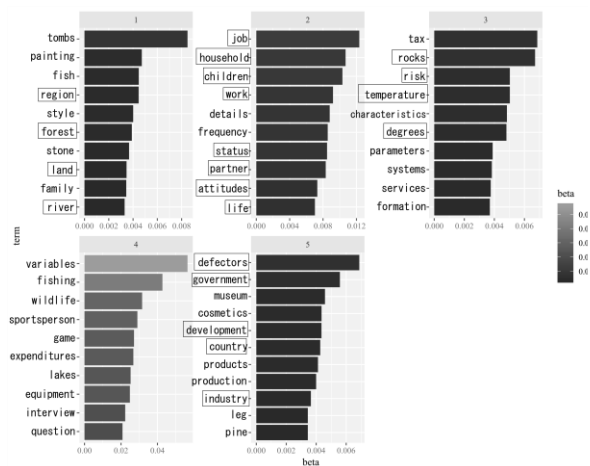


図 4：トピックモデル分析

4. 分析結果

4.1. コレスポンディング分析

初期のビックデータ研究のトレンドはplant, yieldなどが示す産業関連であった。2000年前後になるとspecies, water, degreesなどが示す環境関連がトレンドに加わった。2010年台前半ではpatientやdiseaseなどが示す医療関連がみられるようになり、2010年後半になるとtechnology, networkなどが示す技術関連や、life, child, behaviorなどが示す生活様式関連がトレンドとなった。また、時代の推移によって扱われる分野が多岐にわたっている。加えて、2011年のトレンドのみ異様に突出した。

4.2. トピックモデル分析

コレスポンディング分析にてトレンドが突出していた2011年の論文について、トピックモデル分析をした。分析の結果、regionやforestといった地球の自然に関して、jobやhouseholdといった生活様式や人生設計に関して、riskやtemperatureといった地球温暖化に関して、defectorsやgovernmentといった政治政策に関しての計4トピックが明確に抽出できた。一方で、図4中のトピック4に関しては、トピックの詳細が不明瞭であった。そこで、このトピックに関わりの強い論文を調査した結果、要旨が同一の関連論文が5本連続で投稿されており、これが強く結果に影響を及ぼしていた。

4.3. 階層的クラスター分析

トピックモデル分析と異なり、頻出度合が基準を満たした単語が抽出され各階層に分類さ

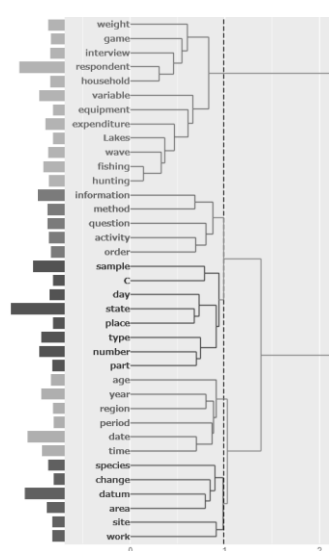


図 5：階層的クラスター分析

れる。そのため、各階層の名詞の繋がりやトピックの詳細を示すには向かない結果となった。

5. 考察と結果

本稿は、ビックデータ研究に関して、その動向を時系列で示して考察するものである。分析の結果、1988年からの30年間に於いて、初期のplant, yield等

が示す産業から、technology, network等が示す技術関連、そして、life, child, behavior等が示す生活様式関連のトレンドに継続的に移行していることが示せた。また、2011年には、他の年には見られない特異な特徴が示されたので、詳しく分析を行い、region, forest等の地球環境、job, household等の生活様式、risk, temperature等の地球温暖化、そして、defectors, government等の政治政策の計4トピックが明確に抽出できた。この年近辺で、それらトピックに関する研究が盛んに行われていることが示された。

本稿では、中村らのコレスポンデンス分析の結果を用いて情報収集の精度を高めた上で、トピックモデル分析とクラスタリング分析を用いて詳細な分析を行っている。この手法を用いることで、前処理の負荷軽減効果を示唆することができた。

参考文献

- 1) 日経コンピュータ、「すべてわかるビックデータ大全 2017」,日経 BP、2016
- 2) 中村匡佑, 水上祐治, 大島昭子, 本多啓介, 中野純司: “論文要旨のテキストマイニング分析による探査衛星の測定データを基にした研究動向の把握”
- 3) Yuji Mizukami, Keisuke Honda and Junji Nakano, “Study on Research Trends on the Internet of Things Using Network Analysis”, IJAMS, Vol.10, No.1, pp.27-35, 2018
- 4) 小林雄一郎: 「Rによるやさしいテキストマイニング」, オーム社(2017)