# テキストマイニングと k-means 法を用いた
# 現在の AI のアンケート調査分析の可視化に関する一考察

| | | | |
|---|---|---|---|
| | 電気通信大学 | *舛井　海斗 | MASUI Kaito |
| | 群馬大学 | 松野　省吾 | MATSUNO Shogo |
| | 電気通信大学 | 伊集院　大将 | IJUIN Hiromasa |
| 05001342 | 広島大学 | 長沢　敬祐 | NAGASAWA Keisuke |
| 01704740 | 電気通信大学 | 山田　哲男 | YAMADA Tetsuo |

## 1. はじめに

近年、AI (Artificial Intelligence、人工知能) は商品やサービスをはじめとして広く社会に組み込まれるようになった[1]。AI とは、人間が現時点で得意としていることを、いかにしてコンピュータに行わせるかの分野[2]であり、人々は急速に普及し始めた AI に期待と恐れを抱いている[1]。高橋ら[3]は、そうした恐れを解決するために、人々の AI に対する危惧を尊重した上で、不安を含めて理解を深めることが重要としている。したがって、人々の言葉からテキストマイニングを用いることで、現在の AI を理解する必要がある。

テキストマイニング[4]とは、コンピュータを用いて大量のテキストデータから定量的な情報を取り出す分析手法の総称である。先行研究として、舛井ら[5]は学生と研究者を対象とした現在の AI に関するアンケート調査とテキスト分析を用いて、人々が考える現在の AI のメリットとリスクや、Z 世代が抱く現在の AI について分析を行った。しかし、形態素解析と単語頻度解析が中心であったため、全体的なアンケートの傾向を示したが、傾向をグループ化して示すには至らなかった。

クラスター分析[6]とは、近いデータを持つデータは同一区分に、離れた値を持つデータは別の区分になるように区分けを定める方法である。クラスター分析をテキストマイニングに適用することで、大量のデータを定量的かつ客観的に分類することが可能になる。

本研究では、人々が抱く現在の AI に関するアンケート調査について、大量のテキストデータから情報を取り出すテキストマイニングと、k-means 法によるクラスター分析を行うことで、人々が現在の AI をどのように捉えているかの可視化に関する一考察を行う。

## 2. 研究方法

### 2.1 クラスター分析の手順

図 1 に、クラスター分析の手順を示す。Step1 として、舛井ら[5]と同様に、米国 Purdue 大学の Prof. Nof らによって行われた自動化に関するアンケート[7]にもとづき、学生へ現在の AI のアンケート調査を実施した。本研究ではアンケート設問のうち、「あなたは現在の AI をどのように定義しますか」という設問に対する、対象学生 199 人の全回答のテキストデータを利用する。

Step2 では、日本語形態素解析エンジンである MeCab[8]と MeCab 用の辞書である mecab-ipadic-NEologd[9]を用いて形態素解析を行い、アンケート回答から名詞の情報を抽出する。さらに、word2vec と 2021 年 9 月 1 日の日本語 Wikipedia のダンプデータ[10]を用いて 200 次元の単語分散表現を獲得し、アンケート回答から得られた名詞に適用する。

Step3 で、各名詞に紐づいた単語分散表現から、k-means 法を用いてクラスター分析を行い、各クラスター内の単語からグループ名称を与える。Step4 では、抽出した名詞の単語分散表現に対して、高次元のデータを視覚化するための次元削減アルゴリズムである t-SNE (T-distributed Stochastic Neighbor Embedding) と、Step3 で得られたクラスター分析の結果を適用し、2 次元に可視化する。

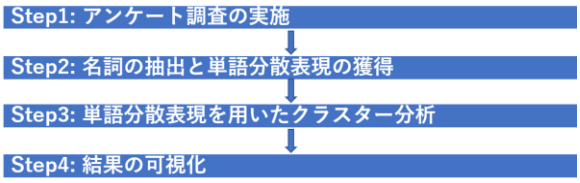


図 1. テキストマイニングと k-means 法を用いた現在の AI のアンケート調査分析の可視化の手順

### 2.2 クラスター分析

クラスター分析[6]とは、N 個のデータについて、K 組の変数値 (K 次元のデータ) が求められている時に、近いデータを持つデータは同一区分に、離れた値を持つデータは別の区分になるように区分けを定める方法である。その中で、区分数を特定して、その区分数になる区切り方のうち、最も良い場合を見出そうとする非階層的手法の 1 つに k-means 法がある。

k-means 法では、まず無作為に K 個のデータを代表値に選び、各データがどの代表値に近いかでグループ化を行う。次に、各クラスターの重心を新たに代表値とし、各データの再分類を行う。この処理を、再分類が起こらなくなるまで行い、その時点のクラスターを最も良い区分とする。

本研究では、k-means 法の設定として、クラスター数 K=5、トライアル数を 300 とし、アンケート回答から得られた単語の分散表現のクラスター分析を行う。また、今回得られたアンケート回答に出現する単語の種類は 394 種類であり、そのうち全名詞の 259 種類を分析対象とする。

## 3. 結果

図 2 は、アンケート回答内に出現する Cluster 2, 3, 4, 5 の名詞の単語分散表現をクラスター分析した結果に応じて、名詞の分散表現を 2 次元に圧縮することで可視化した分散表現図である。

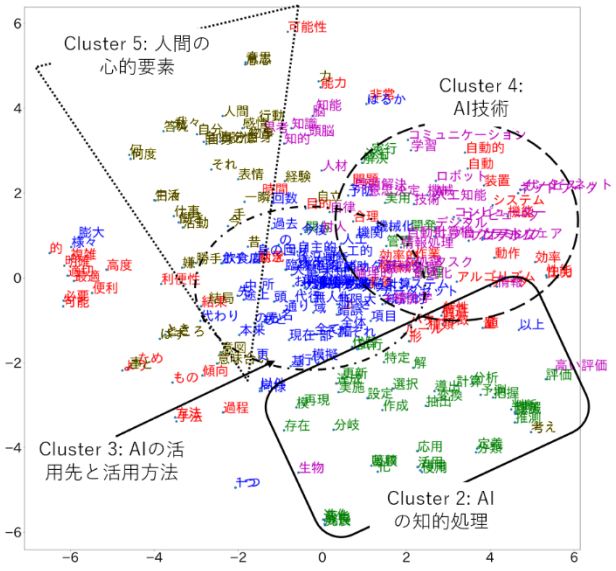


図 2. Cluster 2, 3, 4, 5 の名詞の単語分散表現のクラスタリング結果を 2 次元に圧縮することで可視化した図

図 2 より、分類された 5 つのクラスターごとに、単語がまとまっていることが確認できた。また、各クラスターの分布より、横軸は右側に Cluster 2 の AI の知的処理や、Cluster 4 の AI 技術に関するクラスターがあった。中央には、人間と AI が交わるように、Cluster 3 の AI の活用先と活用方法のクラスターが存在し、左側に人間の心的要素の Cluster 5 が存在した。このことから、横軸は機械的か人間的かを示していると考えた。

一方で、縦軸を考えると、最も高い位置に Cluster 5 の人間の心的要素のクラスターが存在し、続いて Cluster 4 の AI 技術、Cluster 3 の AI の活用先と活用方法、Cluster 2 の AI の知的処理と並んだ。このことから、縦軸は人間の関与する割合を示していると考えた。例えば、同じ AI に関連するクラスターでも、AI の技術を生み出すのは人間であり、活用には人間と AI の両方が必要で、知的処理は AI が独立して行えるものであるためだと考えられる。

さらに、人工知能や AI といった単語が分類された Cluster 4 の AI 技術のクラスターと、Cluster 3 の AI の活用先と活用方法のクラスターは距離が近く、AI とそれらのサービスが密接に関わっていることがわかった。一方で、Cluster 2 の AI の知的処理と Cluster 5 の人間の心的要素のクラスターには最も距離があり、意味合いに隔たりがあることがわかった。

## 4. まとめと今後の課題

本研究では、人々が抱く現在の AI に関するアンケート調査について、テキストマイニングおよびクラスター分析を行うことで、人々が現在の AI をどのように捉えているかの可視化に関する一考察を行った。

今後の課題として、他の設問に対する回答の分析や、名詞以外の品詞の分析、所属分野を分けた分析が挙げられる。

**謝辞** アンケートにご協力いただいた方々に、感謝の意を表する。本研究の一部は、日本学術振興会科研費基盤研究(A)JP18H03824 の助成を受けたものである。

## 参考文献

[1] 独立行政法人情報処理推進機構 AI 白書編集委員会. AI 白書 2020. 株式会社角川アスキー総合研究所. 2020.

[2] W. Ertel. Introduction to Artificial Intelligence. Springer, Cham. 2017.

[3] 高橋大志, 津本周作, 堤富士雄, 松尾豊, 栗原聡, 北川源四郎, 椹木哲夫, 林勲. 横幹連合・日本人工知能学会共催 パネル討論「人工知能と横幹知」. 横幹第 14 巻第 2 号. pp.100-112 (2021)

[4] 牛澤賢二. やってみようテキストマイニング[増訂版]. 朝倉書店. 2021.

[5] 山田哲男, 舛井海斗, 松野省吾, 長沢敬祐, 伊集院大将, 石垣綾, 稲葉通将, 井上全人, 于亜婷, 岡本一志, 北田皓嗣, 周蕾, 杉正夫, 滝聖子, 中嶋良介, 仲田知弘, 大戸一藤田恵理, 山田周歩, Z 世代が抱く現在の AI に関するアンケートテキスト分析の研究と課題, 第 12 回横幹連合コンファレンス, A-3-5, 12 月, オンライン (2021)

[6] 上田尚一. クラスター分析. 朝倉書店. 2017

[7] S. Y. Nof. Springer Handbook of Automation. Springer-Verlag Berlin Heidelberg. pp.14-47. 2009.

[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. https://taku910.github.io/mecab/ (アクセス日：2021 年 12 月 23 日)

[9] mecab-ipadic-NEologd 公式ページ. https://github.com/neologd/mecab-ipadic-neologd. (アクセス日：2021 年 10 月 25 日)

[10] 日本語 Wikipedia ダンプデータ. https://dumps.wikimedia.org/jawiki/latest/ (アクセス日：2021 年 9 月 17 日)