

# 論文の内容における特徴量を用いた 高引用論文の早期発見に向けて

川島大樹(株式会社NTTデータアイ) 来間一郎(株式会社日立製作所) 新田佳菜(株式会社東芝)

## 高引用論文発見の問題点

価値が高い(被引用数が高い)論文の早期発見  
①公開論文数増加により、自力で価値の  
高い論文を見つける難易度が高まった  
②従来の被引用数予測では「過去1~2年の  
被引用数実績」を特徴量として扱っていた

## 手法・ツールの適用による解決

①被引用数予測に必要な情報を自動抽出する  
②論文の内容から得られる情報(特徴量)だけで  
将来の被引用数を予測できるようになる

## 既存特徴量・新規特徴量

既存特徴量(被引用数以外)

表層的な情報がほとんどで、  
内容の良し悪しが反映されない

- 著者数
- 所属機関数
- 参考文献数
- 参考文献の新しさ
- ページ数
- キーワード数
- タイトル中のコロン区切り有無
- タイトルの文字数
- アブストラクト中の文字数
- 本文中の図表の数 etc...

提案する新規特徴量

論文の内容理解に踏み込んだ特徴量として、  
「背景・課題の理解しやすさ」など、 $4 \times 4 = 16$ 通りの特徴量を定義

### 論文内セクション

背景・課題

提案手法

評価手法・データ

結果・考察・結論

### 良い論文の指標

理解しやすさ

簡潔明瞭に整理

情報の十分性

妥当性評価や再現に十分な情報

新規性

既存研究にない着眼点や手法

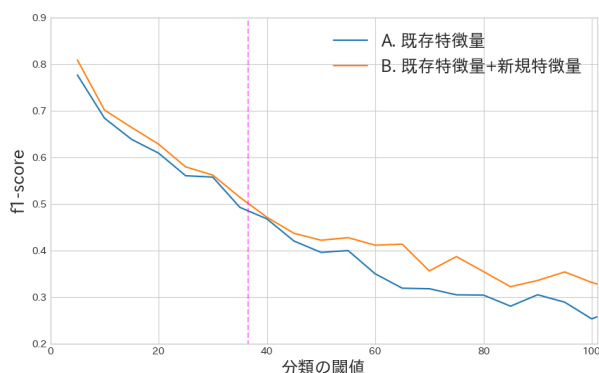
重要性

重要な社会課題の解決に寄与

→ 特徴量の定義と論文テキストを入力として、GPT-4を用いて定量化

## 予測精度の評価結果

A. 既存特徴量	B. 既存特徴量 + 新規特徴量
0.4921917	<b>0.5137554</b>



Random Forestを用いた分類モデルで高引用論文を予測した結果、**新規特徴量追加により精度が改善**  
→新規特徴量のスコアについても概ね妥当であった

## 考察・課題

- **新規特徴量追加により、予測精度は僅かに改善**
  - 新規特徴量は予測に有用と考えられる
- **予測精度としてはやや不十分な結果**
  - 学習データや特徴量の追加によって更なる予測精度の改善が見込める
  - 予測タスク自体の難度が高かった可能性も
- **今後の課題**
  - (1) 別の分析手法を用いる
  - (2) 別の特徴量を用いる
  - (3) 専門家によるLLM算出スコアの妥当性確認
  - (4) モデル予測精度不良の原因の深掘り