研究データ論文の抄録を用いた被引用数推定方式 ：Scientific Data 掲載の抄録を例に

甲斐尚人[1] 義久智樹[1] 新原俊樹[2] 矢野英人[1] 田主英之[1]

概要：オープンサイエンス時代の到来により，研究データの公開，利活用に向けた取り組みが盛んに行われている．公開もしくは共有された研究データの分野融合型研究への利活用を考慮した，異分野研究の研究者にも伝わりやすい抄録の記述方法の開発も今後期待される．これまで学術論文の論文誌のように抄録，本文，参照文献という形式をとる，研究データに特化した論文誌が登場してきている．本研究では研究データに特化した論文誌である「Scientific Data」の抄録に着目し，抄録を構成する英文の品詞の出現数，単語数，キーワード数と研究データ論文の被引用数を重回帰分析することで，各品詞等が研究データの利活用に及ぼす影響を考察した．また，それらの結果をもとに，説明変数を名詞，動詞，その他品詞，単語数，キーワード数に設定し，目的変数を被引用数として機械学習を行い，被引用数を予測する分類器を開発した．これにより，今後の研究データ利活用に向けた，研究データ公開の際の抄録記述の留意点についての議論に繋がることを期待する．

キーワード：研究データ論文，抄録，被引用数，サイエンティフィックデータ

Citation Estimation Method Using Abstracts of Research Data Articles : Using Abstracts of Scientific Data Articles as An Example

NAOTO KAI[†1] TOMOKI YOSHIHISA[†1] TOSHIKI SHIMBARU[†2]
HIDETO YANO[†1] HIDEYUKI TANUSHI[†1]

Abstract: With the trend of open science, efforts have been made to open and utilize research data. Considering the use of published or shared research data for interdisciplinary research, it is expected to develop a method of writing abstracts that can be easily understood by researchers in different research fields. Journals specializing in research data that have a format of abstract, text, and references like academic journals have emerged. In this study, we focus on the abstract of "Scientific Data", a journal specialized in research data, and examine the influence of each part of speech on the utilization of research data through multiple regression analysis of the number of occurrences of the part of speech, the number of words and the number of keywords in the abstract, and the number of citations to the research data article. Based on these results, we set the explanatory variables as the number of occurrences of nouns, verbs, the other parts of speech, the number of words, and the number of keywords in the abstract, and developed a classifier to estimate the number of citations by machine learning with the number of citations as the objective variable. We hope that this will lead to a discussion of the issues that need to be considered when writing abstracts for publication of research data for future use of research data.

Keywords: Research data article, Abstract, Number of citations, Scientific Data

1. はじめに

研究データは日々増大し，世界的なオープンサイエンスの潮流から保管や流通を含む研究データ管理が大きな問題となっている．また，日本においてもその重要性が認識され，様々な取り組みが始まっている．研究データ管理には，研究データの公開までに留まらず，公開された研究データの利活用や研究データを生み出す研究者の評価にまで繋げるエコシステムのサイクルが求められている．

研究データの利活用は，研究効率化や異分野融合といった研究の促進に繋がることが期待されている一方で，研究データはそれぞれの細かな研究領域に深く結びついたものであり，その研究領域を専門としない研究者が，研究データの概要や価値などを理解するのは容易ではない．そのため，研究データの利活用を推し進めるためには，研究概要の理解を促すメタデータ(特にデータ記述)が重要である．データ記述は多様な研究データを説明するために重要なメタデータであるが，どのような記述方法が研究データの利活用に有効であるかといった議論は十分に行われていない．

そのような中で，近年，資金配分機関の要請や研究データの流通を目的としたデータリポジトリやデータジャーナルが世界的に普及しつつある．代表的なものに，世界に先駆けて整備されたデータリポジトリの一つである「Edinburgh DataShare」，分野に制限されないデータジャーナルである「Scientific Data」や「Data in Brief」，ある研究領域に特化したデータジャーナルの「Earth System Science Data」,「Chemical Data Collections」などがある．[1] [2] [3] [4]

1 大阪大学
Osaka University
2 西南学院大学
Seinan Gakuin University

[5] [6]データジャーナルは通常の論文誌と同様に，文字数やキーワードの設定など一定のルールに基づき，研究者に抄録の記載を求めるが，データリポジトリにおいてはその記載方法は研究者に委ねられている場合が多い．先にも述べたとおり，研究データ論文の抄録は研究概要の理解を促進し研究データ論文の被引用数に影響する重要な要素であり，これまでの研究においても着目してきた．[7] [8]

2. 本研究の目的

このような現状を踏まえ，研究データの抄録から研究データ論文の被引用数の増加を予測することが可能であると仮定し，機械学習を活用した被引用数の推定を試みた．本研究では，具体的には「Scientific Data」の抄録に着目した．抄録を構成する英文の品詞の出現数，単語数，キーワード数から研究データ論文の被引用数を推測する分類器の開発を目的として，まず重回帰分析によって研究データの被引用数に及ぼす要素を考察した．また，それらの結果をもとに，説明変数を名詞，動詞，その他品詞，単語数，キーワード数に設定し，目的変数を被引用数として機械学習による被引用数予測を行った．

本研究成果を足掛かりとして，被引用数の増加に寄与する抄録の記載方法，つまり研究者が求める情報の抽出に繋がることを期待している．

3. 関連研究

本研究に対して，以下のような関連研究が存在する．

3.1 抄録の重要性に関する研究

Jinfang Niu[9]はデータ記述（メタデータの一種であり研究データジャーナルの抄録に該当）が十分でないとデータ再利用の促進を妨げるとした．データ記述に不足が生じる原因は，データを作成する者とデータを利活用しようとする者との利害関係が，データ作成者の情報提供の意欲を減衰させることを挙げた．また，コミュニケーションの減少によって，必要な情報に対して優先順位を決める必要が出てくる．そのため，データ作成者の暗黙な情報は優先度が下がってしまうことも再利用の促進を妨げるとした．これらの Jinfang による暗黙な情報と研究データの被引用数の関係性の示唆は，抄録を構成する品詞などの情報から被引用数を推測する本研究に繋がるものであると考える．

3.2 論文要約に与える影響に関する研究

本研究で着目した抄録は，研究データを詳細に説明する本文を簡潔にまとめたものであり，つまり，研究データ論文の要約と捉えることもできる．Kam-Fai Wong ら[10]は，文章の要約作成を目的に，文の表層的特徴と内容的特徴に着目した．そして，それらを組み合わせた要約性能が，組み合わせる前の単独の要約性能より改善したことを示した．表層的特徴として着目した点は，具体的には，文書における文の位置，文中の単語数などである．内容的な特徴については出現頻度の高い語などに着目している． 結果として，表層的特徴に着目した要約性能は適合率（precision）が 48%程度で，内容的な特徴に着目した要約性能は適合率（precision）が 40%程度となり，表層的特徴に着目した方が要約性能としては良い結果となった．また，二つの特徴を合わせた要約性能は適合率（precision）が 57%程度となり性能が改善されることが確認された．本研究は，抄録が被引用数に与える影響を分析することから，Wong らの研究方法に近い．しかし，Wong らは品詞そのものに着目しているわけではないため，本研究の分析手法とは異なる．

3.3 抄録の分析による傾向把握に関する研究

松本ら[11]は，論文の抄録をデータとし，緩和ケアにおける看護研究の動向について，雑誌掲載年と用語の共起ネットワークで関連性を調査するなどテキストマイニングによる分析で明らかにした．また，小山ら[12]は年次学術集会の抄録を計量テキスト分析し，リハビリテーション診療の特徴を調査した．このように論文の抄録に含まれる頻出語など論文の傾向等を把握する研究は数多く行われているが，それら論文の抄録がが論文閲覧者に与える影響に焦点を当てているわけではない．

4. 方法

4.1 対象データの選定

本研究の分析では，文献検索ツールである Scopus を利用した．データセットを詳細に発表することを目的にしたオープンアクセスジャーナル「Scientific Data」に掲載されている研究データ論文に対して，数なくとも 1 回は引用されている 721 件の研究データ論文を対象にした．「Scientific Data」は，Nature Publishing Group によって 2014 年 5 月に創刊された，科学のすべての領域を対象にした研究データジャーナルである．出版元の Nature Publishing Group は，研究データの広範囲な利活用を促進し，データを公開した研究者にクレジットを与えることを目指す，としている．論文のタイプとして Data Descriptor という新しい形式を採用しており，本研究では研究データ論文として記載している．

本研究では，具体的に創刊の 2014 年 5 月から 2018 年までに収載された研究データ論文を対象とした．研究データ論文は一般的な学術論文と同様に，冒頭に抄録があり，その他研究データを詳細に説明する本文，参照文献で構成されている．特に抄録は，学術論文などの内容を簡潔に示し

ており，本文を読み進めるか判断する重要な文章である．これら 721 件の研究データ論文の抄録を，樋口[13]が開発した計量テキスト分析またはテキストマイニングのためのフリーソフトウェアである KHcoder を用いて，形態素解析を行った．形態素解析エンジンは「Stanford POS tagger」[14]を使用した．形態素解析による品詞の出現頻度によって，例えば，形容詞や副詞が多用されていた場合，説明文の曖昧さを増加させることが考えられるが，一方でそれらの品詞を多用することで説明文の限定性を緩和し，より多くの研究者の目に留まる可能性もある．また，一般的な文章で最も多く含まれる名詞も重要な単語であり，読み手の情報量に影響を与える可能性が高いと考えられる．

まず，分析の対象とする研究データ論文の抄録を 2018 年までとした理由は以下である．図 1～図 4 で示すとおり，研究データ論文の発行から十分に時間が経過している 2014 年と 2015 年の被引用数は，研究データ論文が発行されてから 5 年経過する際にその伸びが鈍化するか否かの分かれ目となっていることが推測できる．特に、10 件を超えるか否かがその分岐点であることが推測される．したがって，被引用数が 10 件以上ある研究データ論文 460 件と正例として，10 件より少ない研究データ論文 261 件を負例とすることとした．次の表 1 は，本研究で対象となる研究データ論文の各年度毎の正例，負例の数を示したものである．

表 1　各発行年毎の研究データ論文の被引用数

発行年	正例	負例	全論文
2014年	34	14	48
2015年	50	23	73
2016年	75	43	118
2017年	135	59	194
2018年	166	122	288
計	460	261	721

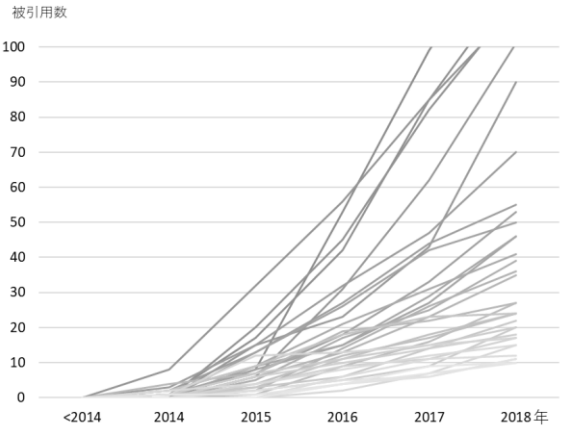

図 1　2014 年発行研究データ論文の被引用数（10 件以上）

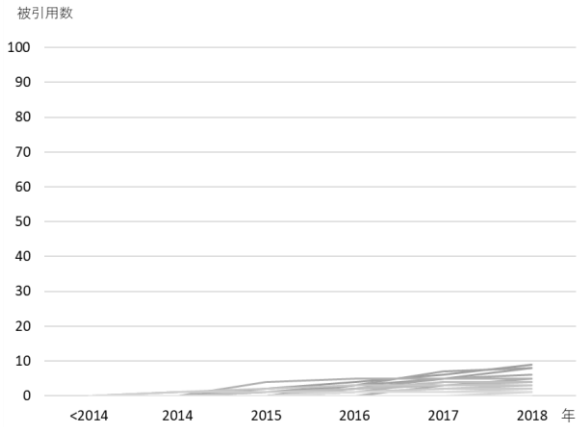

図 2　2014 年発行研究データ論文の被引用数（10 件未満）

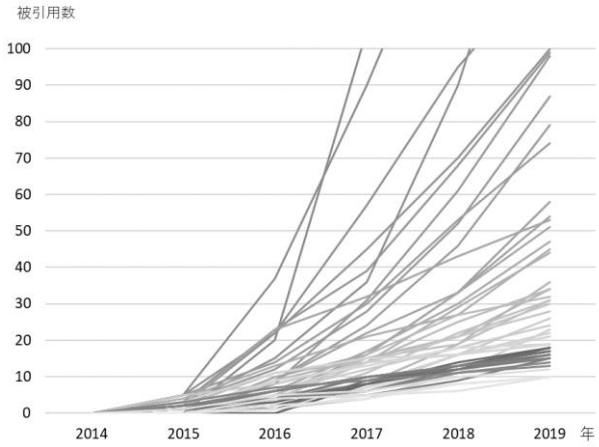

図 3　2015 年発行研究データ論文の被引用数（10 件以上）

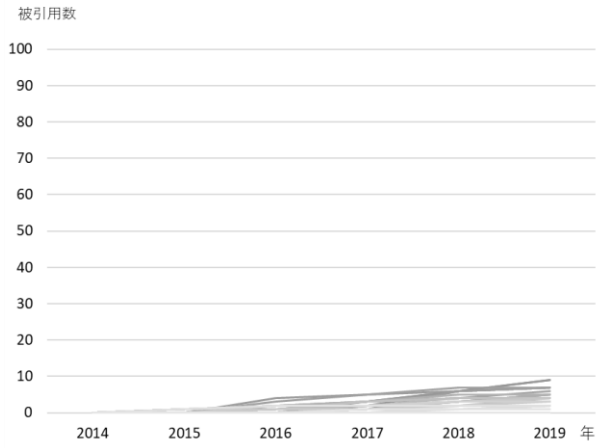

図 4　2015 年における被引用数の累計（10 件未満）

4.2 分析方法

まず，形態素解析によって，各抄録を構成する名詞，固有名詞，動詞，形容詞，副詞の出現回数を明らかにした．接続詞や前置詞などの出現回数も確認したが，名詞，固有

名詞，動詞，形容詞，副詞に比べて，出現回数が少ないため本研究で使用する品詞から除いた．

機械学習による被引用数の推測を行うために，R version 4.2.2（R Core Team 2022）を使用して重回帰分析（有意水準 5%）を行った．形態素解析によって分類した品詞に加え，抄録の単語数，キーワード数を追加し，それらを説明変数として，被引用数へ大きな影響を与える要素を抽出する．機械学習による識別実験の実行環境は，Web ブラウザ上で Python を記述, 実行できる Google Colab を使用した．表 2 のとおり，721 件のデータに対して，学習データとテストデータを 7 ： 3 の割合で分割した．

表 2　学習データとテストデータ

	学習データ	テストデータ
正例	317	143
負例	187	74

また，機械学習の分類モデルをサポートベクターマシン（SVM, LinearSVM），K 近傍法（KNN），パーセプトロン（Perceptron），多層パーセプトロン（MLP），ナイーブベイズ（NB），XGboost（xgb），ロジスティック回帰（LR），決定木（tree），ランダムフォレスト（Random）として識別性能を求めた．

5. 結果

5.1 被引用数に影響を与える要素の推定

重回帰分析（有意水準 5%）を行った結果を表 3 に示す．名詞（t 値=3.182, p 値=0.002）の出現数，単語数（t 値=4.418, p 値=0.000）が多い研究データ論文ほど被引用数が増加し，また動詞（t 値=-2.035, p 値=0.043），キーワード（t 値=-2.646, p 値=0.008）の出現数が少ない研究データ論文ほど被引用数が増加する結果となった．

表 3　R による重回帰分析結果

説明変数	推定値	標準誤差	t 値	P値
名詞	0.0057	0.0018	3.1820	0.0015
固有名詞	0.0035	0.0026	1.3350	0.1822
形容詞	0.0049	0.0030	1.6600	0.0974
副詞	0.0041	0.0069	0.5870	0.5573
動詞	-0.0064	0.0031	-2.0350	0.0422
単語数	0.0026	0.0006	4.4180	0.0000
キーワード数	-0.0056	0.0021	-2.6460	0.0083

5.2 機械学習の各モデルによる識別性能

前項の重回帰分析の結果から，被引用数に与える影響が大きい要素が明らかになった．そのため，名詞、動詞、その他品詞，さらに単語数及びキーワード数を説明変数として，目的変数である被引用数を機械学習によって識別する実験を行った．

表 4 は，各モデルの最適なパラメータによって得られた適合率（precision），再現率（recall），正解率（accuracy）を示す．適合率と再現率の調和平均の F 値（F-measure）も評価指標として使用した．サポートベクターマシン（SVM, LinearSVM），多層パーセプトロン（MLP），XGboost（xgb），ロジスティック回帰（LR），ランダムフォレスト（Random）の分類モデルにおいて，80%を超える F 値で被引用数を識別可能であることがわかった．その中でも，多層パーセプトロン（MLP）モデルで被引用数を推定した場合, F 値 0.853 と正解率 0.788 と最も高く，機械的識別が十分高い F 値と正解率で分類が可能であることがわかった．

表 4　各モデルによる被引用数の推定結果

モデル（パラメータ）	パラメータ値	適合率（precision）	再現率（recall）	F値（F-measure）	正解率（accuracy）
SVM（C）	0.1	0.696	1.000	0.821	0.696
LinearSVM（C）	0.05	0.711	0.927	0.805	0.687
KNN（n_neighbors）	3	0.747	0.821	0.782	0.682
Perceptron	—	0.692	0.536	0.604	0.512
MLP（hidden_layer_sizes）	(5, 2)	0.822	0.887	0.853	0.788
NB	—	0.724	0.887	0.797	0.687
xgb（n_estimators）	10	0.738	0.954	0.832	0.733
LR（C）	20	0.714	0.927	0.807	0.691
tree	—	0.695	0.861	0.769	0.641
Ramdom	—	0.716	0.967	0.823	0.710

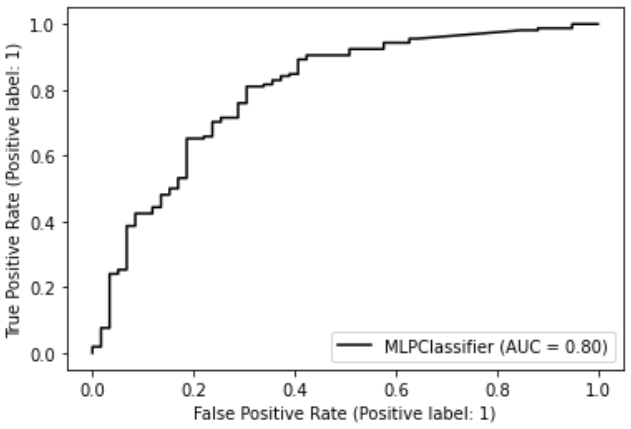

図 5　分類モデル MLP による識別性能

6. おわりに

本研究では，研究データの公開，利活用を促進する目的で創刊されたオープンアクセスジャーナル「Scientific Data」に掲載されている研究データ論文の抄録に着目し，抄録の品詞，単語数，キーワードによる，研究データ論文の被引用数の推定を行った．

まず，抄録をそれぞれ形態素解析し，品詞の出現数，単語数，キーワード数から，被引用数に大きく影響を与える説明変数を特定するために，研究データ論文の被引用数を目的変数とする重回帰分析を実施した．重回帰分析によって, 研究データ論文の被引用数に影響を与える説明変数が, 名詞, 動詞, 単語数, キーワード数であることがわかった．名詞と単語数は数が多いほど被引用数を増加させ，動詞とキーワード数は数が少ないほど被引用数を増加させることがわかった．ここから，一般的な文章と同様に，情報を直接的に伝達する品詞である名詞が多いほど被引用数の増加に貢献することが推測された．しかし，キーワード数は反対の傾向を取ることから，厳選されたキーワードを記載することが最も被引用数の増加に寄与する可能性があることがわかった．今後の課題として，説明変数とした名詞の単語とキーワードで使用されている単語について，対応関係を分析することが求められる．

重回帰分析による結果をもとに，名詞，動詞，それ以外の品詞はその他品詞としてまとめ, 説明変数を名詞, 動詞, その他品詞，単語数，キーワード数と設定し，機械学習による被引用数の推定実験を行った．その結果，多くの分類モデルで 80%を超える F 値（F-measure）で分類可能であることを明らかにした．その中でも多層パーセプトロン（MLP）を分類モデルとした場合, F 値（F-measure）0.853, 正解率（accuracy）0.788 と最も高く，機械的識別が十分高い F 値と正解率で分類が可能であることがわかった．

参考文献

[1] Data Share, https://datashare.ed.ac.uk/, 最終アクセス日 2022.12.1.
[2] ヒリナスキエヴィッチ, イアン, 新谷洋子. Scientific Data データの再利用を促進するオープンアクセス・オープンデータジャーナル, 情報管理, 2014, 57-9, p. 629-640.
[3] Scientific Data, https://www.nature.com/sdata/, 最終アクセス日 2022.12.1.
[4] Data in Brief, https://journals.elsevier.com/data-in-brief, 最終アクセス日 2022.12.1.
[5] Earth System Science Data, https://www.earth-system-science-data.net, 最終アクセス日 2022.12.1.
[6] Chemical Data Collections, https://www.sciencedirect.com/journal/chemical-data-collections, 最終アクセス日 2022.12.1.
[7] 甲斐尚人, 研究データ利活用促進のための暗黙知継承に関する着眼点応用への検討, 情報処理学会研究報告．インターネットと運用技術（IOT）, 2022, 56, 28, p.1-6.
[8] Kai Naoto, Shimbaru Toshiki. Characteristic Analysis of Data Description in Highly Cited Research Data. IIAI Letters on Institutional Research, 2022, Vol. 1, p. 10-18.
[9] Jinfang Niu. Overcoming inadequate documentation. Proceedings of the American Society for Information Science and Technology, 2010, vol.46, issue1, p. 1-14.
[10] Kam-Fai Wong, Mingli Wu, Wenjie Li. Extractive Summarization Using Supervised and Semi-supervised Learning. Proceedings of the 22nd International Conference on Computational Linguistics, 2008, p. 985–992.
[11] 松本啓子, 櫻井大輔, 井上玲子．緩和ケアに関する看護研究の動向 ：テキストマイニングを用いた抄録の内容分析．東海大学健康科学部紀要, 2018, 23, p.107-112.
[12] 小山哲男, 道免和久．中枢神経系疾患のリハビリテーション診療の特徴．リハビリテーション医学, 2021, 58, 3, p.317-325.
[13] 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.
[14] Stanford POS tagger, https://nlp.stanford.edu/software/tagger.shtml, 最終アクセス日 2022.12.1.
[15] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 最終アクセス日 2022.12.1.