Vol.2017-DBS-165 No.26
Vol.2017-IFAT-128 No.26
2017/9/20

オープンアクセスジャーナルの評価指標に関する予備的検討

清水勝太 [1,a)]　高間康史 [1,b)]

概要：近年，多数のオープンアクセスジャーナルが誕生し，Web 上でアクセス可能となっている．それらの中には，学際的研究や萌芽的研究など新しい研究トピックの論文投稿先，情報収集先として有望なものもあれば，十分な査読がなされていない可能性があるものなどもあり，玉石混交の状況にある．Impact factor や Scimago Journal Rank などの評価指標が手掛かりとして有効であるが，従来指標の多くが被引用数などに基づくものであり，歴史の浅いものが多いオープンアクセスジャーナルの評価に適していないと考える．本発表では，論文の内容に基づく評価指標の確立を目的として，論文間の内容的類似度と，既存評価指標との関係について分析した結果について報告する．

キーワード：オープンアクセスジャーナル，学術論文，研究評価

1. はじめに

近年，インターネットの普及に伴い，オンラインでの学術論文の出版，古い論文の電子化など，研究者が研究発表，情報収集の場として web を使用する機会が増大している．研究者は自身の研究成果が高く評価されること，研究費の獲得に有利にはたらくことなどから研究発表の場に採択率の低い学術会議や学術雑誌を選ぶことがある．また，採択率の低い学術会議や学術雑誌自体も採択率を低くすることで権威を保とうとする傾向がある．しかし，こうした状況は研究発表や情報収集の機会損失につながる．同様に，萌芽的，学際的な研究がトップカンファレンスやトップジャーナルに認められにくいという傾向も存在している[1].

オープンアクセスジャーナルとはオンライン上で無料閲覧可能な学術雑誌のことである．雑誌の運営費用を読者から回収するのではなく，論文を掲載する著者から回収する方式のものが多い．さらに，発行媒体や発行頻度に制限を持たないため大量の論文を掲載する雑誌が増加し，年間に 3 万本の論文を掲載するようなものまで出現している[*1]．オープンアクセスジャーナルは，研究発表，情報収集の機会損失を改善する手立ての一つであると言える．しかし，大量の論文を掲載，出版するため，内容に問題のある偽の論文を受理する雑誌が多数存在する[2]など，査読の信頼性に問題があることも指摘されている．

学術雑誌に掲載される研究の評価は査読や研究評価指標に基づく．論文の査読は研究者同士で行われるピアレビューが一般的であるが，これが一部の査読者に集中する傾向にあることが指摘されており[3]，学術雑誌，オープンアクセスジャーナルの増加によってその持続可能性が危惧されている．また，既存の研究評価指標として Impact Factor(IF)[*2]や Scimago Journal Rnak(SJR)[4]などが存在する．これらは学術雑誌を，その被引用数，引用元の論文が掲載されている雑誌などから評価する指標である．これらは学術雑誌に対して定量的な評価指標を与えるが，引用数を基に算出されることや，論文単体ではなく学術雑誌に適用されることなどから，研究の内容評価に用いられることは疑問視されている[5].

本稿では分子生物学分野で最高峰と認められている学術雑誌 Cell に掲載された論文と，オープンアクセスジャーナルに掲載された論文をそれぞれ fastText[6]を用いて分散表現し，内容的類似度とその雑誌の SJR の値との関係を分析した結果について述べる．ここで，本稿における内容的類似度とは，論文が扱うトピックの類似度ではなく，論文の構成や体裁，言い回し等に関するものを指す．レベルの高い学術雑誌であれば，著者の論文執筆スキルや経験が高いことが想定されるため，この様な内容的類似度が高くなると考える．この仮定を検証することを本稿の目的とする．

本稿では以下の構成をとる．2 節では関連研究について述べる．3 節では実験方法について述べる．4 節で実験結果について述べ，5 節にまとめと今後の計画を述べる．

2. 関連研究

2.1 研究評価

学術雑誌の評価指標に関する研究[5]では，IF の問題点について指摘している．少数の論文が大量に引用されることによって，その雑誌の引用による評価が上昇する結果，同じ雑誌に掲載される被引用回数の少ない論文も高評価を得るということが問題視されている．

本研究で用いる SJR[4]では，学術雑誌に掲載された論文の被引用数に重みを考慮した評価指標を採用している．この重みの基準は権威のある学術雑誌，学術会議に対し，Page Rank に類似するアルゴリズムを適用して求めている．また，SJR は分野の異なる雑誌間の評価に対して補正をかけ，算

1 首都大学東京大学院システムデザイン研究科
Graduate School of System Design Tokyo Metropolitan University

a shimizu-shota@ed.tmu.ac.jp
b ytakama@tmu.ac.jp

*1 PLOS 社が運営する雑誌の 2014 年の総掲載数は 3 万本を超えた
https://www.plos.org/
*2 Web of Science Core Collection 収録雑誌に対する一年間に平均どの程度引用されたのかを測る指標

出された数値によってすべての雑誌を一様に評価する．

SJR は，ある雑誌に掲載された論文が，権威のある雑誌に掲載された論文からどの程度の引用を行っているかを考慮し，学術雑誌の評価に使用する指標である点で IF と大きく異なる．

この他にも論文あるいは研究のための評価指標が多数提案されている[7][8]．しかし，これら研究評価指標のほとんどが論文の内容的な評価を行っておらず，その引用に関連する項目によって評価を定めている．また，これらは論文単体に対して評価が与えられず，対象の論文がどのような学術雑誌に掲載されているか，という点からしか論文の評価を考えることができないデメリットが存在する．

2.2 分散表現

分散表現とはテキストコーパスを基に単語を，その意味を表すベクトルとして表現することである．これは，同じ文脈に出現する単語が意味を共有するという分布仮説に基づいている[9]．分散表現によって得られたベクトルは通常，高次元の実数ベクトルであり，同一の単語や文でも学習データによって異なるベクトルとして表現される．

近年では，ニューラルネットワークを用いた分散表現の学習の研究が盛んである[10][11] [12]．word2vec[11][12]では与えられた単語の周辺予測に基づく skip-gram，周辺単語から中心単語を予測する CBOW(Continuous Bag of words)といった，それまでの研究よりも比較的単純な学習モデルが提案され，大規模なデータに対して分散表現を獲得可能となった．また，その後の研究により単語に限らず，一定の長さを持つフレーズや文の分散表現が得られるようになった[13]．

また，分散表現を用いて文書要約の研究が行われている．野口らは，単語の分散表現を基に，文や文書の分散表現を求め，類似度によってスコアを付けて，要約する手法を提案している[14]．田口らは，文書間の距離を最小化するように分散表現の学習手法を考慮し，要約によってその評価を行っている[15]．ニューラルネットワークベースの分散表現を用いる，ことによって，意味的に元の文書に近い要約文を得ることができるため有用であることが示されている．同様に，単語や文といった粒度の異なるものを一様にベクトル空間上に表現できるため，様々な距離の尺度を類似度や重要度のスコアとして利用できる．文書に分布するとされる意味を一定次元のベクトルに埋め込むことで，文書間の関係性を考えることができるのが分散表現の利点の一つである．

3. 実験方法

web 上から収集した pdf 形式の論文ファイルからテキストを抜き出し，不要部分の削除など，前処理を行った論文ファイルをベクトル化する．ベクトル化された論文を用いて各学術雑誌間の類似度を計算する．
権威的な学術雑誌とオープンアクセスジャーナルとの類似度と，既存の評価指標による評価との関係性を分析する．

本稿では学術雑誌間の類似度を計算するために，学習が高速かつ，単語，文を分散表現可能なライブラリの fastText[6]を用いる．

表1. 使用した学術雑誌

学術雑誌	分野	SJR
Cell	分子生物学	27.696
Journal of Finance	経済学	20.973
NATURE	学際的分野	18.134
Cell Reports	分子生物学	8.507
Cell and Bioscience	分子生物学	1.621
Cells	分子生物学	0.889
Open Biochemistry Journal	分子生物学	0.394
Journal of Applied Research of Children	心理学	0.173
BUSSINESS THEORY and PRACTICE	経済学	0.165
Journal of Chemical and Pharmaceutical Sciences	分子生物学	0.124
pythagoras	数学	0.101

3.1 使用データ

本稿では表 1 に示す学術雑誌を用いた．使用した学術雑誌は全て英語で書かれたもので，SJR データベースによる分類によれば，分野は分子生物学，経済学，心理学，数学，学際的分野である．Cell, NATURE, Journal of Finance 以外の学術雑誌は全てオープンアクセスジャーナルである．また，各雑誌の SJR の値は 2015 年の値を使用した．2015 年の SJR 値の算出に用いられるのは 2012 年から 2014 年の 3 年間に掲載された論文であるので，その期間に掲載された取得可能な論文を分析対象とした．

3.2 前処理

学術雑誌間の類似度を求める際，同一分野の雑誌間と異なる分野の雑誌間で異なるアプローチをとった．同一分野（分子生物学）に対しては，論文構成を分析し，論文自身の研究内容と関連する，イントロダクション，実験，結果，結論のみを分析対象とした．また，論文中に表れるストップワード，制御文字を削除した後，ピリオドで終わる文を一文として文単位で分散表現を求めた．

異なる分野の論文では論文構成が異なるため，機械的な研究内容部分の抽出が難しいため，ストップワード，制御

Vol.2017-DBS-165 No.26
Vol.2017-IFAT-128 No.26
2017/9/20

文字を削除する処理のみ行った．これにあわせて，出現単語の違いが考えられるため，ベクトル化する文を長くするように，論文単位で分散表現を求めた．

論文をベクトル化する際の fastText のパラメータに関して，学習の際の文字 n-gram は 3-6 を用いた．ベクトルの次元数は，同一分野間は 100，異なる分野間は長文をベクトル化するため 300 とした．学習モデルは skip-gram とした．

学習コーパスの違いによる類似度の違いを評価するため，後述する自己類似度について，学習コーパスとして収集した Cell の論文ファイルと英語の wikipedia のデータを用いて分析を行った．同一分野間の論文に対しては Cell の論文ファイルを学習コーパスとし，異分野間の論文に対しては英語の wikipedia のデータを学習コーパスとして分析を行った．

3.3 学術雑誌間類似度

学術雑誌XとYの類似度$sim_j(X,Y)$を以下で定義する[16]．

$$sim_j(X,Y) = \frac{1}{|X_N||Y_N|}\sum_{p_i \in X_N}\sum_{p_k \in Y_N} sim_p(p_i,p_k) \qquad (1)$$

$$sim_p(p_i,p_k) = \frac{1}{|p_i||p_k|}\sum_{s_m \in p_i}\sum_{s_n \in p_k} sim_s(\boldsymbol{s_m},\boldsymbol{s_n}) \qquad (2)$$

$$sim_s(\boldsymbol{s_m},\boldsymbol{s_n}) = \frac{\langle \boldsymbol{s_m},\boldsymbol{s_n}\rangle}{\|\boldsymbol{s_m}\|\|\boldsymbol{s_n}\|} \qquad (3)$$

ただし，sim_pは論文間の類似度，sim_sは文間の類似度を示し，$\boldsymbol{s_m}$，$\boldsymbol{s_n}$ はそれぞれ論文p_i，p_kに含まれる文のベクトルとする．論文p_i，p_kは雑誌X，Yの論文であり，それぞれランダムに選択した集合をX_N,Y_Nとする．

二種類の学術雑誌から論文を一本ずつ選択し，選択した論文の文同士の平均類似度が論文間の類似度となる．二種類の学術雑誌から論文を選択する際，取得可能な論文数の差，類似度の計算量，計算時間を考慮し，学術雑誌X，Yから選択する論文数をそれぞれ$X_N = Y_N = 16$とした．

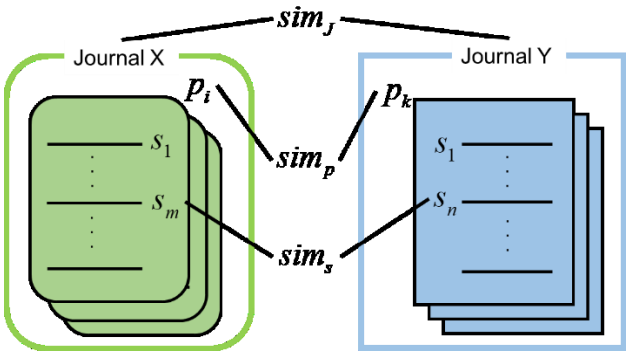

図1．学術雑誌間類似度

異なる分野間での学術雑誌間類似度を求める際，論文を一つのベクトルとするため，学術雑誌X'とY'の類似度sim_j'の定義は式(4)とする．

$$sim_j(X',Y') = \frac{1}{|\boldsymbol{p_i}||\boldsymbol{p_k}|}\sum_{\boldsymbol{p_i} \in X}\sum_{\boldsymbol{p_k} \in Y} sim_s(\boldsymbol{p_i},\boldsymbol{p_k}) \qquad (4)$$

ここで，$\boldsymbol{p_i}$，$\boldsymbol{p_k}$は共に単一のベクトルであるため，類似度計算には式(3)を用いる．

4. 実験結果

4.1 学習データによる類似度の違い

同一の単語や文でも学習コーパスによって異なるベクトルとして表現されることが，学術雑誌間類似度に与える影響を考慮する必要がある．したがって，学術雑誌間の類似度を求める前に，学習データによる類似度の違いを考察する．Cell の論文をベクトル化する際に，学習データを Cell にしたものと，英語の wikipedia にしたものを用意し，それぞれ自己類似度を計算した．

自己類似度は，コサイン類似度の分布の違いを確認するため，各文間の類似度（式(3)）を求め，その分布として定義する．論文間類似度（式(2)）の分布ではないことに注意されたい．Cell に関して，学習データに Cell を用いた場合と，英語の wikipedia を用いた場合を，それぞれ図 2，図 3 に示す．

それぞれの分布の分散は，学習データが Cell の場合，2.23×10^{-3}，wikipedia の場合は8.13×10^{-4}であり，Cell を学習データに用いた方が広がりのある分布になることがわかった．これはベクトル化するテキストのドメインと学習データのドメインの重なりが，分散表現に幅を持たせ，その幅が類似度に表れているためだと考えられる．

上記から，後述の，同一分野の学術雑誌間類似度を考える際は，分解能の高い類似度を得るために学習データを Cell とした．逆に，異なる分野では出現単語など，ドメイ

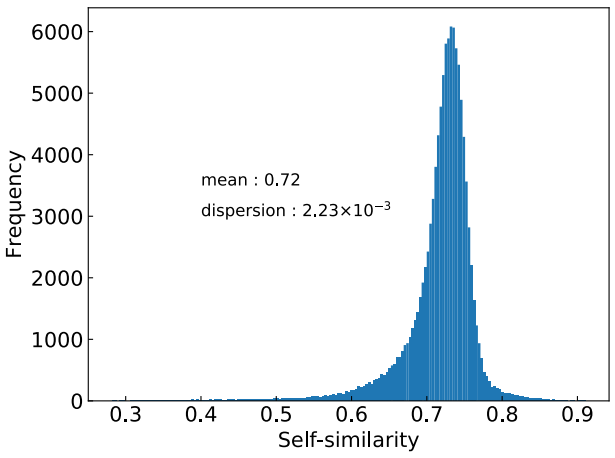

図2．学習データに Cell を用いた Cell の自己類似度

Vol.2017-DBS-165 No.26
Vol.2017-IFAT-128 No.26
2017/9/20

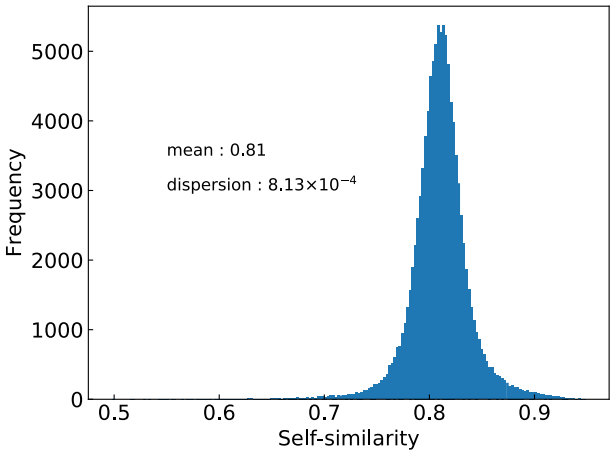

図3. 学習データに wikipedia を用いた Cell の自己類似度

ンの違いを踏まえ, 学習データを wikipedia として分析を行った.

4.2 異なる分野間での学術雑誌間類似度

表 1 に示した学術雑誌のうち，Cell と異なる分野の 4 つの雑誌の学術雑誌間類似度を図 4 に示す.

図 4 から SJR 値が高い雑誌ほど Cell との類似度が高い傾向があるが，NATURE と Journal of Finance に関しては傾向が逆である．これは分野が部分的に重なっているので，NATURE と Cell に共起する単語は Journal of Finance と Cell に共起する単語より多く，このような結果になったと考えられる．このことから，異分野の学術雑誌に対して学術雑誌間類似度から SJR 値を推定する場合は，分野の類似度を考慮した補正が必要であることが示唆される.

また, どの雑誌の学術雑誌間類似度の値も 1 に近いのは，論文一本という多量の情報を 300 次元のベクトルに圧縮するため，次元数に対して文章情報が多く，ベクトル空間上の狭く, 近い部分に情報が圧縮されたためだと考えられる.

類似度をブロードな分布として考えるためには，ベクトル空間に埋め込む単語，文の長さとベクトルの次元数のバランスを考える必要がある．同様に，学習コーパスが次元

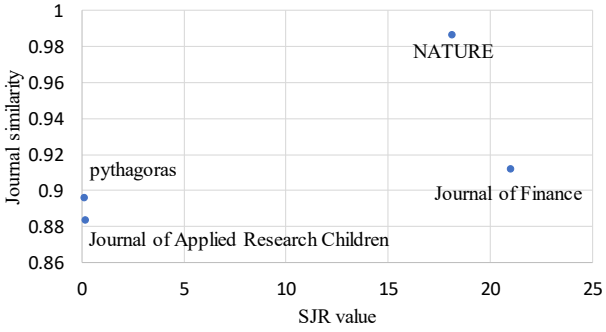

図4. 異分野の学術雑誌間類似度

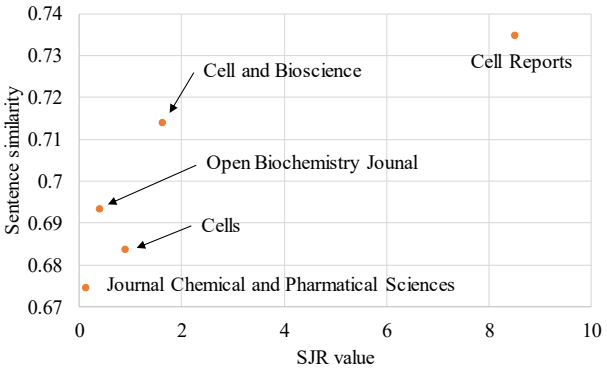

図5. 同一分野の学術雑誌間類似度

数に対して小さければ，余剰な次元を持つ埋め込みが行われてしまうため，次元数と学習コーパスの大きさのバランスも分散表現を得るために考慮が必要である.

4.3 同一分野での学術雑誌間類似度

Cell を基準に，表 1 に示した分子生物学分野の学術雑誌間類似度と SJR 値の関係を，図 5 に示す.

図 5 から，既存の評価指標によって高い評価を得ている学術雑誌を基準に考えると，同一分野の学術雑誌の評価は内容的な類似度とも対応があるといえる．このことから，既存の指標による評価がなされていない学術雑誌でも，同一分野の既存指標による評価が高い学術雑誌と内容的に類似していれば，既存指標による高い評価が得られる可能性がある.

4.4 パラグラフ別の学術雑誌間類似度

学術雑誌間のパラグラフ別の類似度について述べる．ここで使用した学術雑誌は，同じ出版社から発行されているため，見かけの構成が同じである Cell と Cell Reports とした．比較対象とするパラグラフは論文自身の研究内容を示す 3 つにした．また，文をベクトル化した場合と段落をベクトル化した場合の比較を行った．それぞれ表 2，表 3 に示す.

表2. 文をベクトル化した場合のパラグラフ別類似度

Cell	Cell Reports	類似度
Introduction	**Introduction**	**0.772**
Introduction	Results	0.738
Introduction	Summary	0.767
Results	**Introduction**	**0.746**
Results	Results	0.733
Results	Summary	0.729
Summary	Introduction	0.761
Summary	Results	0.729
Summary	**Summary**	**0.768**

Vol.2017-DBS-165 No.26
Vol.2017-IFAT-128 No.26
2017/9/20

表3. 段落をベクトル化した場合のパラグラフ別類似度

Cell	Cell Reports	類似度
Introduction	**Introduction**	**0.954**
Introduction	Results	0.942
Introduction	Summary	0.933
Results	Introduction	0.940
Results	**Results**	**0.964**
Results	Summary	0.933
Summary	Introduction	0.926
Summary	Results	0.927
Summary	**Summary**	**0.930**

表 2，表 3 においてパラグラフごとに最も類似度が高い組み合わせは太字で表す．表 2 から，文をベクトル化した場合のパラグラフ別の類似度は Cell の Results に対する組み合わせ以外は同一のパラグラフの類似度が最も高くなった．これに対して，表 3 ではすべての組み合わせで同一のパラグラフが最も高い類似度を示している．これは，類似度を考える際，比較対象の粒度を考慮する必要があることを示している．パラグラフに関する類似度を考えるにはパラグラフをベクトル化する方が，もっともらしい結果が得られる．

また，表 3 において各パラグラフで最も類似度の差がないパラグラフは Summary であるが，これは Summary が他のパラグラフの内容を反映しているためだと考えられる．

4.5 自己類似度に関する考察

自己類似度は平均値が 1 に近く，分散が小さければ，学術雑誌に掲載される論文が，類似度の観点から，まとまりをもっていると考えられる．表 4 に分子生物学分野の学術雑誌の SJR 値と自己類似度の分散を示す．自己類似度の分散は最も小さい Cell Reports の値で正規化している．表 4 から SJR 値の小さい学術雑誌は自己類似度の分散が大きい傾向にあることがわかる．

表 4 において，正規化分散の値が他の雑誌と比べて特に

表4. 自己類似度の正規化分散

雑誌名	SJR	正規化分散
Cell	27.696	1.17
Cell Reports	8.507	1
Cell and Bioscience	1.621	2.06
Cells	0.889	1.72
Open Biochemistry Journal	0.394	6.07
Journal of Chemical and Pharmaceutical Sciences	0.124	1.68

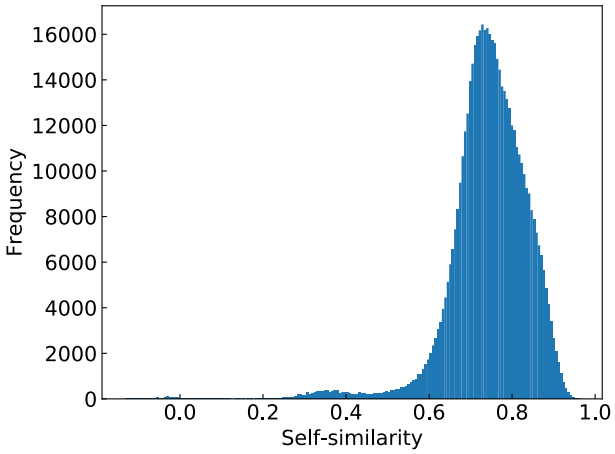

図6. Open Biochemistry Journal の自己類似度

大きい Open Biochemistry Journal について考察するために，自己類似度を図 6 に示す．図 6 から Open Biochemistry Journal では自己類似度の分布に多峰性が確認できる．これより，低類似度側の峰の存在が分散を大きくしている一要因だと考えられる．

以上から SJR 値の高い学術雑誌は自己類似度の分散が小さくなる傾向があると考えられる．この性質を利用することで，その傾向を学術雑誌の評価指針の一つとして利用できる可能性があると考える．

5. まとめと今後の計画

学術論文を分散表現することにより，既存の評価指標と内容的類似度の関係性の分析を行った．分散表現する際の学習データによる違いを調査した結果では，対象学術雑誌のドメインと重なるコーパスを用いれば，類似度の分散が大きくなることを示した．

異なる分野の学術雑誌間の比較では，SJR 値が高いほど，SJR が高い雑誌（Cell）との類似度が高くなる傾向が確認できたが，比較対象論文の分野の異なりの程度を考慮する必要があることも示された．同一分野の学術雑誌間の比較でも，同様の傾向が確認できたが，部分的に対応が異なる学術雑誌が存在した．

また，学術雑誌間のパラグラフごとの類似度を比較した結果からは，論文構成とベクトル化を行う文章の粒度を考慮する必要があることがわかった．

さらに，SJR 値の高いものは自己類似度の分散が小さくなる傾向があることを示した．これより，分野の異なりや，分散表現のための学習コーパスを考慮すれば，自己類似度の分散を学術雑誌の評価指針の一つとして利用できる可能性が示された．

今後の計画として，分子生物学分野以外の分野での学術雑誌の比較，既存の評価指標をベースにしない評価方法と

して，テキスト解析による論文評価に適用可能な属性の検討を行う予定である．

参考文献

[1] Paul Wouters : “Journal ranking biased against interdisciplinary research,” https://citationculture.wordpress.com/2011/11/15/journal-ranking-biased-against-interdisciplinary-research/ 最終アクセス 2017 年 8 月 16 日
[2] John Bohannon : “Who`s Afraid of Peer Review?,” Science, Vol.342, Issue6154, pp. 60-65, 2013
[3] Michail Kovanis, Raphaël Porcher, Philippe Ravaud, Ludovic Trinquart : “The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise,” PLOS ONE, Vol. 11, e0166387, 2016.
[4] Vicente P. Guerrero-Bote and Félix Moya-Anegón : “A further step forward in measuring journals’ scientific prestige: The SJR2 indicator,” Journal of Informetrics, Vol. 6, Issue 4, pp674-688, 2012.
[5] Ewen Callaway : “Beat it, impact factor! Publishing elite turns against controversial metric,” NATURE, Vol.535, pp.210-211, 2016.
[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov : “Enriching Word Vectors with Subword Information,” arXiv preprint arXiv:1607.04606, 2016.
[7] Ludo Waltman, Nees Jan van Eck, Thed N. van Leeuwen, Martijn S. Visser : “Some modifications to the SNIP journal impact indicator,” Journal of Informetrics, Vol. 7, Issue 2, pp.272-285, 2013.
[8] J.D.West, Theodore C. Bergstrom, Carl T. Bergstrom : “The eigenfactor metrics™: A network approach to assessing scholarly journals,” College and Research Libraries, Vol. 71, Issue 3, pp.236-244, 2010.
[9] Zellg S. Harris : “Distributional Structure,” WORD, Vol. 10, Issue 2-3, pp.146-162, 1954.
[10] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin : “A Neural Probabilistic Language Model,” Journal of Machine Learning Research, Vol. 3, pp.1137-1155, 2003.
[11] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean : “Efficient Estimation of Word Representations in Vector Space,” arXiv:1301.3781, 2013.
[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean : “Distributed Representations of Words and Phrases and their Compositionality,” arXiv:1310.4546, 2013.
[13] Quoc V. Le and Tomas Mikolov : “Distributed Representations of Sentences and Documents,” arXiv:1405.4053, 2014.
[14] 野口　正樹，谷塚　太一，小林　隼人 :“分散表現を用いたヤフー知恵袋の要約，” 言語処理学会第 21 回年次大会，pp.1084-1087, 2015.
[15] 田口　雄哉，重藤　優太郎，新保　仁，松本　裕治 :“抽出型文書要約における分散表現の学習，” 言語処理学会第 23 回年次大会，pp.497-500, 2017.
[16] 太田　貴久，増山　繁 :“模倣レポート判定に用いる文書間類似度の考案，” 言語処理学会第 10 回年次大会，pp.729-732, 2004.