The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

テキスト解析に基づく学術論文の定量的評価方法の検討

Consideration of Quantitative Evaluation Method of Scientific Papers based on Text Analysis

清水 勝太[*1]
Shota Shimizu

高間 康史[*1]
Yasufumi Takama

[*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

This paper proposes a method to evaluate scientific papers based on text analysis. Recently, there are many scientific papers on the Web along with the trend of digitization of academic information. Such a situation creates a need for evaluation method of scientific papers as selection criteria. This paper defines Journal similarity and analyzes relationship between Journal similarity and Scimago Journal Rank (SJR). The Journal similarity of a target journal to an authoritative journal is calculated based on distributed representations of texts of papers. When the similarity of a target journal is high, its quality in terms of writing and organization is expected to be high. This paper also proposes to evaluate the consistency of writing of a paper using ROUGE. The experimental results show that journals with high SJR tends to have high Journal similarity with authoritative journals and that Cell Reports is more consistent than Cell in evaluation using ROUGE.

1. はじめに

学術論文は，知識創発にとって重要な情報源の 1 つである．近年，インターネットの普及に伴い，Web での学術論文の出版，古い論文の電子化，学会発表資料の公開などが行われている．電子媒体を通して論文を利用する研究者も増加している．学術図書館研究委員会の調査[SCREAL 2014]において，月に一回以上電子ジャーナルを利用していると答えた研究者は 85%程度に上る．もっとも電子ジャーナルの利用が顕著な薬学分野では 97.2%の研究者が月に一回以上電子ジャーナルを利用している．2007 年時の調査と 2011 年時の調査を比較すると，論文を印刷して読む研究者の割合は減少し，ディスプレイを通して論文を読む研究者の割合が増加している．同様に，American Journal Experts 社の調査[AJE 2016]によれば，2006 年から 2016 年までに学術論文の出版数は倍程度になっている．流通する研究情報が増大することで，人間の処理能力とのギャップが増大し，読むべき論文の選択が研究者にとって以前よりも負荷の大きい作業になっている．その結果，選択基準の 1 つとして論文や学術雑誌の評価指標が利用されている．

学術雑誌の評価指標として，Impact Factor (IF) や Scimago Journal Rank (SJR) [Bote 2012] などが存在する．これらは学術雑誌を掲載論文の被引用数，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り，学術雑誌に対して定量的な評価を与える．しかし，被引用数をもとに算出されることから，出版直後の学術雑誌を評価できないこと，論文単体ではなく，学術雑誌に適用されること，被引用数の多い少数の論文によって学術雑誌の評価が高まり，被引用数の少ない論文も高い評価を得てしまうことなどから，研究の内容評価に用いられることは疑問視されている[Callaway 2016]．また，これらの指標は定期的に刊行される学術雑誌を想定したものであり，学会の予稿には適用されていないものが多い．情報科学の分野では研究サイクルの早さの観点から，学会発表を重視する傾向があるため，これら指標を適用しづらい問題が存在する．

連絡先：首都大学東京大学院
システムデザイン研究科情報科学域
〒191-0065 東京都日野市旭が丘 6-6
E-mail: shimizu-shota@ed.tmu.ac.jp

以上のことから，多量の論文から読むべき論文を選ぶ際，研究内容を考慮した定量的な評価方法が必要と考える．本稿では，学術雑誌間類似度を定義し，既存の評価が高い権威的な論文集合を基準として，類似度による論文の評価を提案する．本稿における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．高い評価を得ている論文集合は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，この様な類似度が高くなると考える．また，要約評価指標 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin 2003,2004] を用いた論文の一貫性に関する評価方法も提案する．

2. 関連研究

2.1 研究評価

本稿では既存の評価指標として，データ利用の観点から SJR を用いる．SJR は，学術雑誌に掲載された論文の被引用数に重み付けを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，PageRank 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ている点で IF と異なる．

この他にも論文あるいは研究を対象とした評価指標が多数提案されている[Moed 2010][West 2010]．しかし，それらのほとんどが論文の内容的な評価を行っておらず，引用関係に基づい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

自然言語として表現される単語や文といった処理単位を，有限次元の実数値ベクトルとして表現することを分散表現という．通常，分散表現された処理単位は，コーパス中の語彙数より少ない次元数を持ち，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ある．また，分散表現は，分布仮説[Harris 1954]に基づいた「共起している語が似ていれば単語の意味は類似している」という分布類似度の考えに基づいている．

自然言語をベクトル空間上に表現することにより，様々な距離尺度を類似度や重要度のスコアとして利用できる．また，分布類似度の考えから単語や文の意味関係を扱うことができる利点も存在する．さらに，自然言語をベクトルとして扱うことは，単語によって明示的にあらわされるトピックよりも，体裁や言い回しなどの類似度を扱うことに適していると考える．

本稿では，論文をベクトルとして扱うために，分散表現を得るツールとして，学習が高速かつ，可変長の入力を受け付ける fastText [Bojanowski 2016]を用いる．

2.3 要約評価指標

文書間の類似度を測ることで，異なる文書間の差異を計量することができる．自動要約や機械翻訳の分野では，システム出力の評価ために要約結果や翻訳結果と原文の差異を評価する際に文書間類似度を用いる．自動要約システムを評価する指標として ROUGE[Lin 2003,2004]が提案されている．要約の評価は通常，人手で用意された参照要約と要約システムによる要約が一致する度合いで定められる．ROUGE による評価方法を以下の式に示す．

$$R_{ouge}(S,R)=\frac{\sum_{e\in n(S)}C_{match}(e)}{\sum_{e\in n(R)}C(e)} \qquad (1)$$

ここで，Sはシステム出力，Rは参照要約（正解データ），$n(\cdot)$はテキストに含まれる n-gram を示す．$C(e)$は文書中における e の出現頻度，$C_{match}(e)$はシステム要約と参照要約における eの出現頻度の内，小さいものである．

3. 提案手法

論文中の文をベクトルとして扱うことにより，論文間類似度を定義する．さらに論文間類似度を用いて，学術雑誌間類似度を定義する．2 種類の論文集合を用意し，片方を評価基準として用いることにより，論文集合に対する評価方法を提案する．評価基準として，既に高い評価を得ている権威的な論文集合を用いることにより，それらと類似している論文は同様に高い評価を得ると考える．

また，論文において明示的に得られる段落構成に対して要約評価を行うことで論文の一貫性を評価する方法も提案する．

3.1 学術雑誌間類似度

論文中の文に基づく学術雑誌間類似度について述べる．学術雑誌XとYの学術雑誌間類似度$sim_j(X,Y)$を以下に定義する．

$$sim_j(X,Y)=\frac{1}{|X_N||Y_N|}\sum_{p_i\in X_N}\sum_{p_k\in Y_N}sim_p(p_i,p_k) \qquad (2)$$

$$sim_p(p_i,p_k)=\frac{1}{|p_i||p_k|}\sum_{s_m\in p_i}\sum_{s_n\in p_k}sim_s(\boldsymbol{s}_m,\boldsymbol{s}_n) \qquad (3)$$

$$sim_s(\boldsymbol{s}_m,\boldsymbol{s}_n)=\frac{\langle \boldsymbol{s}_m,\boldsymbol{s}_n\rangle}{|\boldsymbol{s}_m||\boldsymbol{s}_n|} \qquad (4)$$

式(2)-(4)において，sim_pは論文間類似度，sim_sは文間類似度を示し，$\boldsymbol{s}_m$，$\boldsymbol{s}_n$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．p_i，p_kは雑誌X，Yに含まれる論文であり，それぞれの雑誌からランダムにN本選択した論文の集合をX_N，Y_Nとする．

また，論文 1 つのテキスト全てを 1 つのベクトルとして扱う場合，雑誌X'，Y'の学術雑誌間類似度$sim_j(X',Y')$を以下の式により定義する．

$$sim_j(X',Y')=\frac{1}{|X'_N||Y'_N|}\sum_{\boldsymbol{p}_i\in X'_N}\sum_{\boldsymbol{p}_k\in Y'_N}sim_s(\boldsymbol{p}_i,\boldsymbol{p}_k) \qquad (5)$$

$$sim_s(\boldsymbol{p}_i,\boldsymbol{p}_k)=\frac{\langle \boldsymbol{p}_i,\boldsymbol{p}_k\rangle}{|\boldsymbol{p}_i||\boldsymbol{p}_k|} \qquad (6)$$

ここで，$\boldsymbol{p}_i$，$\boldsymbol{p}_k$は各論文の全テキストによるベクトルとする．

本稿では，評価基準となる論文集合として，権威的な学術雑誌または，高 h-index の著者による論文集合を用いる．2 種類の論文集合から選択する論文数Nは 4 節に後述するように実験ごとに変更した．

3.2 要約評価指標による論文評価

学術論文には abstract，あるいは summary として全体の要約を示す部分が存在する．本文とみなせるパラグラフを参照要約，全体の要約を示す部分をシステム要約とみなし，要約評価指標を学術論文に適用することで，本文と要約部分の一貫性を評価する．学術論文における一貫性は読み手の理解を助長するため，評価指標として有効であると考える．

本稿では，要約評価指標として ROUGE を用いて論文の一貫性評価を行う．論文Pの本文をR_P，要約部分をS_PとしてPの要約評価スコア$S_S(P)$を以下の式で定義する．

$$S_S(P)=R_{ouge}(S_P,R_P) \qquad (7)$$

本稿では，ROUGE を適用する際，単語 n-gram をユニグラムとして評価を行う．

表1. 使用した学術雑誌

学術雑誌	分野	SJR 値
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
Journal of Machine Learning Research	計算機科学	1.627
Cell and Bioscience	分子生物学	1.621
Journal of Artificial Intelligence Research	計算機科学	1.016
Cells	分子生物学	0.889
i-Perception	計算機科学	0.725
Computational Linguistics	計算機科学	0.546
Open Biochemistry Journal	分子生物学	0.394
Journal of Computer Science	計算機科学	0.255
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

表2. 使用した高 h-index の著者

著者	分野	h-index
A	計算機科学	145
B	計算機科学	137
C	計算機科学	106
D	計算機科学	99
E	計算機科学	77

4. 評価実験

4.1 使用データ

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抽出し，不要部分の削除など，前処理を行った論文のテキストデータをベクトル化する．本稿では権威的な学術雑誌として，SJR 値の高い，Cell，Nature，Journal of Finance を用いた．同様に権威的な論文集合として，h-index の高い著者による論文を使用した．表 1 に使用した学術雑誌を，表 2 に使用した高 h-index の著者を示す．

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直近 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．

4.2 前処理

PDF 形式の論文ファイルから抽出したテキストデータに対して，可能な限り，雑誌名や発行年月日などの不要な情報を排除した．同様に論文中に現れるストップワードと制御文字を削除し，ピリオドで終わる文字列を一文とみなしてベクトル化する際の単位として用いた．論文間類似度を計算するための，fastText の学習パラメータについては，学習モデルを skip-gram，文字 n-gram は 3-6 のデフォルト値を用いた．ベクトルの次元数，学習コーパスについては実験ごとに変更して評価を行った．

4.3 論文間類似度と SJR に基づく評価

4.3.1 異分野間雑誌の評価

Cell と異なる分野の学術雑誌 4 誌との学術雑誌間類似度と SJR の関係を分析する．fastText の学習パラメータとして次元数 300，テキストコーパスを Wikipedia，論文 1 つの全テキストを 1 つのベクトルとした．雑誌からのサンプル数Nは 16 とした．結果を図 1 に示す．

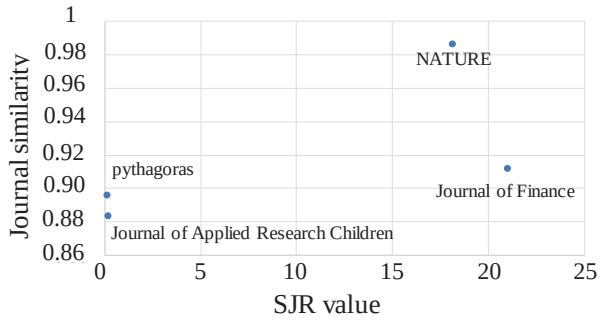

図1.異分野雑誌の学術雑誌間類似度

図 1 から SJR 値が高い雑誌ほど Cell との類似度が高い傾向にあることがわかる．しかし，Nature と Journal of Finance に関して傾向が逆であることがわかる．これは，Cell と Nature の分野が部分的に重なるため，Cell と Nature に共起する単語が Cell と Journal of Finance に共起する単語より多く，Cell と Nature の学術雑誌間類似度が高く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評価を行う場合は分野を考慮した補正が必要であることが示唆される．さらに，論文の全テキストを単一のベクトルとして扱うため，多数の単語情報が圧縮され，共通の単語によるトピックの影響を排除しきれなかったと考える．

4.3.2 同分野間雑誌の評価

Cell と同分野の学術雑誌との学術雑誌間類似度と SJR の関係を分析する．fastText の学習パラメータとして次元数 100，テキストコーパスを Cell，論文中の一文を 1 つのベクトルとした．雑誌からのサンプル数Nは 16 とした．結果を図 2 に示す．

図 2 において SJR 値が高い雑誌ほど，Cell との学術雑誌間類似度が高い傾向にあることがわかる．SJR 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．ここから評価の与えられていない学術雑誌の論文に対して，同分野の権威的な学術雑誌を基準とした評価を行えると考える．

4.4 論文間類似度と h-index に基づく評価

権威的な論文集合として，表 2 に示した h-index の高い著者による論文を使用し，SJR との関係に関する分析を行った．fastText の学習パラメータは次元数 300，テキストコーパスを Wikipedia，論文中の一文を 1 つのベクトルとした．論文集合からのサンプル数Nは 5 とした．図 3 に著者別の学術雑誌間類似度，図 4 に雑誌別の学術雑誌間類似度を示す．

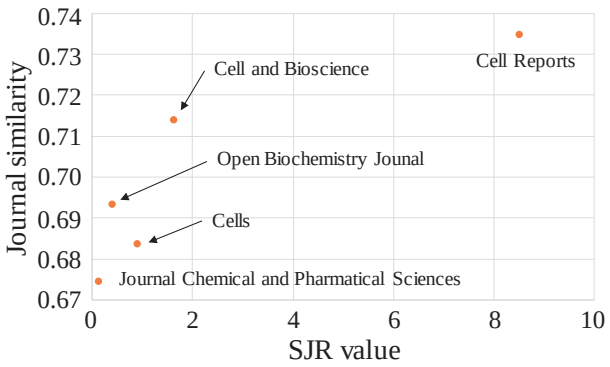

図2.同分野雑誌の学術雑誌間類似度

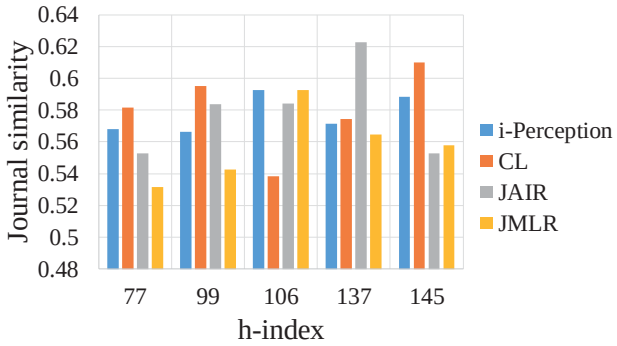

図3.著者別の学術雑誌間類似度

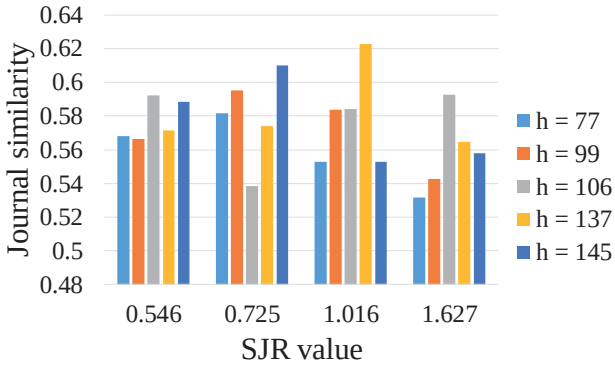

図4.雑誌別の学術雑誌間類似度

表3. ROUGE による Cell の要約評価値

	Precision	Recall	F-value
Introduction	0.106	0.506	0.175
Results	0.016	0.710	0.031

表4. ROUGE による Cell Reports の要約評価値

	Precision	Recall	F-value
Introduction	0.113	0.515	0.185
Results	0.018	0.713	0.035

図 3 から h-index と学術雑誌間類似度の間に一貫した傾向は確認できない．同様に図 4 から SJR 値と学術雑誌間類似度についても一貫した傾向を確認できない．つまり SJR 値や h-index の大小は，直接的に学術雑誌間類似度の値に影響しない．h-index は論文をもとに定まるが，論文は単一の著者によって書かれるわけではなく，同一著者による論文集合は学術雑誌のように扱えないと考える．よって高 h-index の著者による論文集合は評価基準として適していないと考える．

4.5 要約指標による評価

Cell と Cell Reports を用いて，要約指標による論文の評価について検討する．Cell，Cell Reports の論文における Introduction，Results の記載内容を要約評価におけるシステムの出力，Summary を参照要約（正解データ）とみなして，各論文に対して ROUGE を用いて要約評価を行った．表 3 に Cell，表 4 に Cell Reports の要約評価を示す．

表 3，4 からどの比較においても適合率より再現率の方が高い．これは Summary と Introduction，Results で同じ単語が多く使われている．また，Introduction よりも Results の方が，再現率が高く，Results と Summary で同じ単語がより多く使われていると考えられる．さらに，それぞれの評価値に着目すると，Cell よりも Cell Reports の方が，全ての評価値が高く，要約評価の面からみて，一貫性のある学術雑誌だといえる．

5. おわりに

本稿では，分散表現を用いたテキスト解析による定量的な学術論文の評価方法を提案した．また，既存の評価指標 SJR と提案手法の関係を分析することで，その使用可能性を検証した．

評価実験から，Cell を基準として同分野の学術雑誌に対して提案手法を適用することで，SJR が高ければ，学術雑誌間類似度も高い傾向を示す結果が得られた．異分野の学術雑誌に対する実験では，分野の違いによるトピックの影響を考慮しなければならないことがわかった．

高 h-index の著者による論文を基準とした実験では，SJR に対しても h-index に対しても一貫した傾向が確認できず，論文の評基準として適していないことがわかった．これは論文が複数の著者によって執筆されることなどが原因であると考えられる．

要約指標による一貫性の評価では，ROUGE を適用することで，論文の一貫性を評価できることを示した．論文の相対的な比較において要約指標による評価を利用できる可能性があると考える．

本稿で検討した評価手法は，他の尺度による評価を得ている論文を基準として相対的に評価するアプローチであり，近年普及しつつある多数のオープンアクセスジャーナルの評価方法として，使用可能性があると考える．

今後の課題として，他の指標や評価方法と比較して評価の妥当性の検討，引用関係を考慮した評価方法の検証などが挙げられる．

参考文献

[SCREAL 2014] 学術図書館研究委員会（SCREAL），学術情報の取得動向と電子ジャーナルの利用度に関する調査 2011，SCREAL 調査報告書，2014.

[AJE 2016] American Journal Experts, “AJE Scholarly Publishing Reports:2016,” American Journal Experts, 2016.

[Bote 2012] V. P. Guerrero-Bote and F. Moya-Anegón, “A further step forward in measuring journals’ scientific prestige: The SJR2 indicator,” Journal of Informetrics, vol. 6, no. 4, pp. 674–688, 2012.

[Callaway 2016] E. Callaway, “Publishing elite turns against impact factor,” Nature, vol. 535, no. 14, pp. 210–211, 2016.

[Moed 2010] H. F. Moed, “Measuring contextual citation impact of scientific journals,” Journal of Informetrics, vol. 4, no. 3, pp. 265–277, 2010.

[West 2010] J. D. West, T. C. Bergstrom, and C. T. Bergstrom, “The eigenfactor metrics™: A network approach to assessing scholarly journals,” College and Research Libraries, vol. 71, no. 3, pp. 236–244, 2010.

[Harris 1954] Z. S. Harris, “Distributional Structure,” WORD, no. 2-3, pp. 146–162, 1954.

[Bojanowski 2016] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[Lin 2003] C. Y. Lin and E. Hovy, “Automatic evaluation of summaries using N-gram co-occurrence statistics,” NAACL ’03, pp. 71–78. 2003.

[Lin 2004] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” WAS 2004, pp. 25–26, 2004.