研究評価指標に関する考察

Consideration on Scientific Indicators

清水　勝太[1]　高間　康史[1]

Shota Shimizu[1], Yasufumi Takama[1]

[1]首都大学東京システムデザイン研究科
[1]Graduate School of System Design, Tokyo Metropolitan University

Abstract: In recent years, many scientific indicators have been proposed for measuring the impact of research, and as the selection criteria of academic journals. Although these indicators give us quantitative evaluation based on the number of citations and downloads of papers, and the number of publications of authors, there is a problem that these indicators are not based on the contents of a paper. In order to support researchers to find appropriate journals for their paper submission and information gathering, a novel indicator that overcomes the above-mentioned problem is needed. This paper compares existing scientific indicators as a preliminary stage for the final objective of research evaluation based on contents.

1　はじめに

近年，インターネットの普及に伴い，学術論文の出版，古い論文の電子化，学会や講演における発表資料の公開など，Web を利用した学術情報の流通が盛んである．American Journal Expert 社の調査[1]によれば，2016 年までの 10 年程度で論文出版数は倍増している．また，最も論文出版数の多い学術雑誌 5 つの内 4 つがオープンアクセスジャーナルである．近年では，学術情報専門の媒体が Social Networking Service (SNS) などを通して，専門家でない人からもよくアクセスされるようにもなっている．このような Web 上での学術情報の流通は今後拡大していくと考えられる．

論文の出版数増大に伴い，様々な問題も発生している．これらの問題は査読と情報収集に関する問題に分類することができる．査読に関する問題の一つとして，投稿論文数増加による査読負担の増大が挙げられる．学術雑誌に投稿される論文が時々刻々と増加しているが，査読を担当する人間はほとんど増えておらず，大量の査読負担が一部の査読者に集中することが懸念されている[2]．情報収集に関する問題としては，流通する学術情報の増大と人間の処理能力のギャップが大きくなることによって，研究者にとって読むべき論文の選定が困難になっていることが挙げられる．

以上の問題に対して，Impact Factor (IF) [3]などの研究評価指標が用いられることがある．例えば，研究者は自身の論文を投稿する学術雑誌をその雑誌の IF で判断することがある．同様に，研究評価指標に基づいて読むべき論文を決めることもある．このように研究評価指標は様々な面での選択基準の一つとして利用されている．

本稿では，増加した論文をその内容に基づき評価することを最終的な目標とし，その前段階として，被引用数に基づく既存研究評価指標と内容に基づく論文評価指標の比較に基づく分析を行い，研究評価指標が満たすべき要件と今後の課題について考察する．

2　研究評価

研究活動の質を管理する方法として，研究評価が長年行われている．査読は最も古く[4]，現在でも行われている研究評価方法の一つである．査読の目的は，学術雑誌の限られた誌面に掲載する論文を選定すること，投稿された研究が科学的に妥当であるかどうかを検討することである．近年では，競争的研究資金の獲得や, 研究活動の説明責任の追及, また，大学や研究機関の評価など，様々な背景から研究評価の必要性が取り上げられている．

現代の研究評価は研究の結果のみならず，そのプロセスやその後の波及効果なども対象とすることがある．しかし，研究成果の主要な発表手段は論文であるので，本稿では，学術雑誌あるいは論文を対象とした研究評価を対象とする．現在主流の被引用数に基づく研究評価指標はいくつか存在するが，そのほとんどが典型的な問題点を抱えている．以下，主な研究評価指標について述べる．

2.1 Impact Factor

Impact Factor (IF) は最も有名な研究評価指標の一つであり，学術雑誌に対して与えられる定量的な指標である．これは対象とする学術雑誌に掲載された論文が直近二年間で，平均してどの程度引用されたかを示す指標で，一年ごとに算出される．例えば，ある学術雑誌のあるy年の IF, I_yは以下のように定められる．

$$I_{\mathrm{y}} = \frac{C_{y-2}^{y} + C_{y-1}^{y}}{P_{y-2} + P_{y-1}}, \tag{2.1}$$

ここで, P_yは該当する学術雑誌がy年に掲載した論文の総数，C_x^yはx年に掲載された論文がy年に引用された回数を示す．

IF は当初，収蔵スペースに限りのある図書館が購読する学術雑誌を決めるための参考として考えられた．そのため，同分野の学術雑誌を比較する際には有用である．一方で，IF は論文単体に与えられる評価ではなく，学術雑誌に与えられる評価であることや，被引用数しか考慮しないため，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]．例えば，2015 年の Nature の IF は 38.1 だが，掲載された論文の 75.5%は 35 回以下の被引用数しか持たない．つまりほとんどの論文が過大評価されている状態である．こういった欠点があるにもかかわらず，IF を研究評価の指標として用いる場面は多数存在している．

また，IF を部分的に改良したものや，分野ごとの評価値を正規化したものなど，IF をベースラインとした被引用数に基づく評価指標が多数提案されている[6]-[9]. しかし，これらの指標は被引用数に基づくため 2.4 節で後述する問題を抱えている．

2.2 h-index

h-index[10]は論文の著者に対する評価指標である．これは研究者が発表した論文数とその論文がどの程度引用されているかを示す定量的指標であり，定義は「被引用回数がh回以上である論文がh本以上あることを満たす最大の数値h」である．h-index を用いることで，ある著者の論文出版数と被引用数（論文の質とみなされている）を同時に扱うことができる．研究分野や研究慣習の異なりを超えて，著者の持つ研究への量的，質的な貢献度を測ろうというのが h-index による評価の意図である．

h-index の欠点として，10 回引用された論文を 100 本持つ研究者と，100 回引用された論文を 10 本もつ研究者への評価が同じになってしまう点や，研究歴の短い研究者はそもそも論文出版数が少なく，研究歴の長い研究者に対して不利である点などが挙げられる．

2.3 Altmetrics

IF は学術雑誌, h-index は著者に与えられる指標であった．一方，Altmetrics は論文出版数が増えたことや，SNS の利用が増大したことに伴い，論文単位で評価を行うために提唱された[11]．Altmetrics とは，論文や研究成果の影響を，ソーシャルメディアを通じて定量的に測定する手法と，およびこれを用いた研究評価活動のことを指す．これにより，即時的かつ多面的に論文単位での評価が可能となる．Altmetrics の特徴として, 専門家に限らず一般の人に対する研究の影響を測定できる点，被引用数による評価の補完・代替となる可能性，論文発表直後から評価を行える即時性などが挙げられる. Altmetrics 計測の要素として，Altmetrics 計測サービスの Imapct Story は以下の 5 つを挙げている[12].

1. Viewed
   - PDF ファイルなどのダウンロード数
2. Discussed
   - SNS での言及回数
3. Saved
   - Mendeley[1]などでのブックマーク数
4. Cited
   - 論文や Wikipedia での引用数
5. Recommended
   - プレス記事などでの推薦数

これらの項目に基づいて論文の評価を行う．例えば，新規論文が Twitter で言及された回数や，論文 PDF ファイルのダウンロード数で，研究の影響度を測定する．しかし，より新しい論文が有利に測定されやすい点や，SNS での言及回数の測定方法に一貫性がない点など，ソーシャルメディア利用による問題も存在する．

2.4 被引用数に基づく研究評価指標の課題

被引用数に基づく研究評価指標には，以下に挙げる点が問題として付随する. Altmetrics は評価項目として，被引用数も含んでいるので，被引用数に基づく研究評価指標として考える．

1. 研究発表から評価までに時間を要する
2. 分野による研究習慣の異なりに影響を受ける

[1] Mendeley https://www.mendeley.com

3. データベースに依存する
4. 内容に基づいていない
5. 論文の種類による引用のされやすさが異なる
6. 共著者の貢献度を評価できない

これらのうち，1，2 はこれまでに述べた通りである．データベース依存性は被引用数に基づく評価方法の典型的な課題である．例えば，IF は Web of Science に収録された論文にしか適用されず，これに収録されていない論文からの引用も被引用も考慮されていない．

また，どの評価手法でも，論文の内容を考慮しているものはない．論文の内容を考慮せず，被引用数に基づき評価する場合の弊害として，評価が論文の種類に依存することが挙げられる．様々な研究内容について触れるレビュー論文は引用のベースラインとして利便性が高いので引用されやすく，レビュー論文を含む学術雑誌は被引用数に基づく評価が高くなりやすい．

さらに，被引用数に基づく評価では，共著の論文に関して，著者別の貢献度を評価できないことも，問題点として挙げられる．被引用数に基づく評価では，これらの問題が研究評価の妥当性に対して常に存在する．

3 内容に基づく論文評価指標の検討

前節で挙げた被引用数に基づく研究評価指標の問題点のうち，本稿では論文内容を考慮していない点に着目し，これを考慮した評価可能性について検討する．

一般に，情報検索やテキストマイニングなどでは文書の内容としてそのトピックに着目し，トピックの類似性に基づいて文書の検索やクラスタリングを行う．一方，トピックとは異なる観点として，論文の書き方・表現も内容の一種と考えられる．採択率の低いトップジャーナルやトップカンファレンスでは，研究内容の新規性や信頼性だけでなく，論文の書き方についても査読により厳しく評価されている．その結果，十分な研究業績を持つ研究者は論文執筆スキルも高いといえる．従って，トップジャーナルなどに掲載された論文と，書き方に関する類似性の高い論文の質は高い事が期待できる．

この仮説について検証するために，本稿では文献[13]で提案されている，学術雑誌間類似度を用いた予備実験結果について示す．また，単一の雑誌のみを用いた内容に関する評価指標として，同文献[13]で提案されている自動要約指標を用いた手法の予備実験結果も示す．

3.1 学術雑誌間類似度

この手法では，論文中のテキストを文単位でベクトル化し，ベクトル間の類似度をコサイン類似度で定義し，論文間類似度を定義する．さらに，論文間類似度に基づいて学術雑誌間類似度を定義し，学術雑誌の評価を行う．それぞれの定義を以下に示す．

$$S_{\text{Journal}}(X,Y)=\frac{1}{|X_N||Y_N|}\sum_{p_i\in X_N}\sum_{p_j\in Y_N}S_{\text{paper}}(p_i,p_j), \quad (3.1)$$

$$S_{\text{paper}}(p_i,p_j)=\frac{1}{|p_i||p_j|}\sum_{s_k\in p_i}\sum_{s_l\in p_j}S_{\text{sentence}}(\boldsymbol{s}_k,\boldsymbol{s}_l), \quad (3.2)$$

$$S_{\text{sentence}}(\boldsymbol{s}_k,\boldsymbol{s}_l)=\frac{\langle \boldsymbol{s}_k,\boldsymbol{s}_l\rangle}{|\boldsymbol{s}_k||\boldsymbol{s}_l|}, \quad (3.3)$$

ここで，$S_{\text{sentence}}(\boldsymbol{s}_k,\boldsymbol{s}_l)$は文をもとに生成されたベクトル$\boldsymbol{s}_k$，$\boldsymbol{s}_l$のコサイン類似度を示す．$S_{\text{paper}}(p_i,p_j)$は論文$p_i$，$p_j$の論文間類似度を示す．2 件の論文について，文間類似度（式(3.3)）の平均が論文間類似度である．$S_{\text{Journal}}(X,Y)$は学術雑誌X，Yの学術雑誌間類似度を示す．p_i，p_jはそれぞれX_N，Y_Nに含まれる論文を示す．$S_{\text{Journal}}(X,Y)$を算出する際，取得可能な論文数の差，計算量を考慮し，一定数 (N) の論文を各学術雑誌からサンプリングして論文間類似度の算出に用いる．

本稿では，基準となる既存評価指標として，Scimago Journal Rank (SJR) [6] と h-index [10] を用いた．図 3.1，3.2，3.3 に論文間類似度を用いた手法による学術雑誌の評価を示す．

図 3.1 は分子生物学分野の学術雑誌について，トップジャーナルである Cell との学術雑誌間類似度を示したものであり，横軸に SJR の値，縦軸に基準雑誌 (Cell) との類似度を示している．SJR の値が高いほど基準雑誌との類似度が高い傾向が確認できる．

図 3.2，3.3 は h-index を基準指標とした学術雑誌間類似度を示している．基準となる著者の論文と対象となる学術雑誌の分野は計算機科学分野とした．

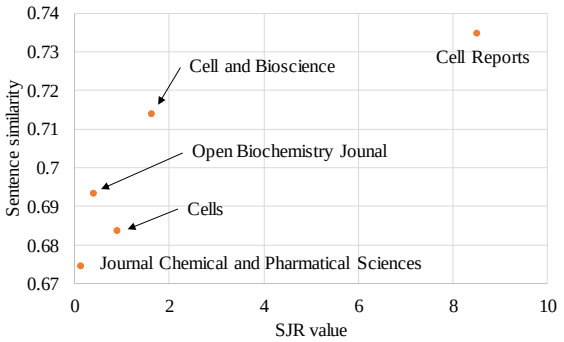

図3.1 分子生物学分野雑誌の学術雑誌間類似度

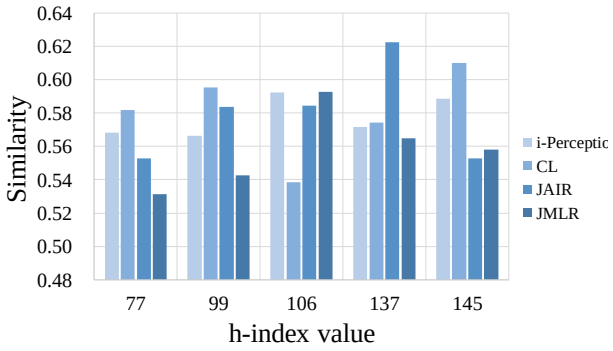

図3.2 著者別の学術雑誌間類似度

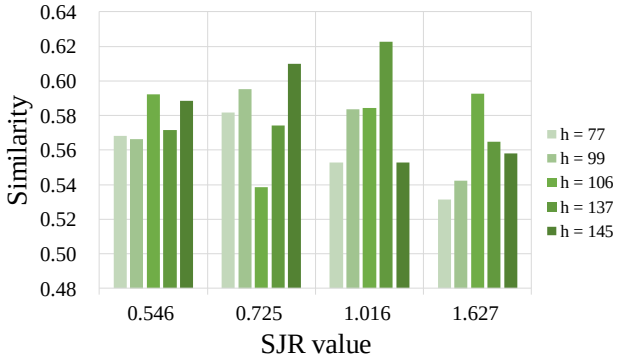

図3.3 雑誌別の学術雑誌間類似度

凡例に雑誌名と著者の h-index 値を示している. 両図から h-index と SJR, 学術雑誌間類似度の間に一貫した傾向がないことがわかる. ここから, h-index を基準にしても既存評価指標と相関のある評価は行えないと考えられる.

3. 2 要約評価指標に基づく評価

自動要約システムの評価は, システムによる要約（評価対象）と人手の要約（正解データ）との一致度合いに基づいて行われる. Abstract, Summaryといった論文の要約セクションを正解データ, 研究内容を示す本論セクションを評価対象としてROUGE (Recall-Oriented Understudy for Gisting Evaluation) [14][15]を適用することで, 要約セクションと本論セクションの一貫性が評価可能と考える. ROUGEによる正解データRと評価対象Sに関する評価値$S_{\mathrm{ROUGE}}(S,R)$は式(3.4)で定義される.

$$S_{\mathrm{ROUGE}}(S,R)=\frac{\sum_{e\in n_{\mathrm{gram}}(S)} C_{\mathrm{match}}(e)}{\sum_{e\in n_{\mathrm{gram}}(R)} C(e)}, \qquad (3.4)$$

ここで, $n_{\mathrm{gram}}(\cdot)$はテキストに含まれる単語n-gramを示す. $C(e)$は文書中におけるeの出現頻度, $C_{\mathrm{match}}(e)$は正解データRと評価対象Sに含まれるeの共起回数を示す. 式(3.4)は Recall と同じ定義であるが, 同様に Precision を算出することもでき, そこから F-value も求めることができる.

表3.1 Cell の要約評価

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

表3.2 Cell Reports の要約評価

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

表 3.1, 3.2 に Cell, Cell Reports の要約評価を示す. Summary を正解データ, 研究内容を示す Introduction, Results を評価対象としている. 両表からどちらの雑誌でも, Results の方が Recall の値が高いことがわかる. これは, Introduction には研究背景として, 当該論文以外の研究内容が相対的に多く含まれるのに対し, Results に示される研究成果の主要な部分は Summary にも含まれるためと考える.

4 研究評価指標の比較と考察

3 節の結果を踏まえて, 表 4.1 に研究評価指標の比較を示す. 2.4 節で述べた各問題点の解消を要件とみなし, それぞれに対する各評価指標（手法）の相対的, 定性的な性能比較を○, △, ×で示す.

評価までの時間に対しては, h-index が最も性能が悪く, 論文が引用されるまで評価が与えられない. IF は 2 年ごとに評価値が与えられ, Altmetrics は発表された瞬間からダウンロード数などが与えられ,

表4.1 研究評価指標の比較
“-” は評価なしを示す.

	Impact Factor	h-index	Altmetrics	S_{Journal} / ROUGE
評価までの時間	△	×	○	○
異分野間対応	×	-	-	△
DB非依存性	×	×	○	△
著者の評価	×	△	×	×
内容に基づく評価	×	×	×	○
論文単位の評価	×	-	○	△

3 節の手法 (S_{Journal} / ROUGE) も評価対象が発表されてすぐに評価が可能である．

分野の異なりに対して IF は前述のとおり，影響を受けるので最も評価が悪い．h-index，Altmetrics は分野間の異なりを対象としないので評価なしとした．3 節の手法は分野の異なりに部分的に対応している [13].

IF，h-index はその算出がデータベースに依存するので，データベース非依存性の評価が最も悪く，Altmetrics では論文が多様なメディアにより流通されてから評価が行われるので非依存性が高いと判断される．一方，3 節の手法は，データベースには依存しないが，評価対象のジャーナルに収録された論文を取得・分析する必要があるため中程度の評価とした．

著者の評価に対しては，全ての評価指標で十分ではないといえる．例として，複数の共著者に対して，貢献度を差別化できない点が挙げられる．h-index は著者個人に評価を与えるので中程度の評価としたが，各論文への貢献度を判断できない点は他の指標と同様である．

内容に基づく評価に対しては，既存評価指標は全て十分ではないといえる．3 節の手法は現時点で十分な性能を持つとは言えないが，相対的に高い評価としている．

論文単位の評価に対しては，IF は論文単体を対象としていないので低い評価となる．h-index は論文出版数に基づき算出されるものの，論文に対して評価を与えないので，評価なしとした．3 節の手法は，ROUGE を用いる方法が論文単体に対応しているといえる．

以上より，IF，あるいはそれに類する被引用数に基づく研究評価指標を研究や研究者の評価として用いることの妥当性は低く，内容に基づく評価指標の導入が必要と考える．

5 おわりに

本稿では，被引用数に基づく評価指標と内容に基づく評価指標による学術雑誌・論文の評価について比較考察を行った．

考察の結果，被引用数に基づく評価指標は共通の課題を抱えていることを明らかにした．これら課題の解消を要件として，各研究評価指標を比較した結果，代表的な研究評価指標として頻繁に用いられる IF は，ほとんどの要件に対して評価が悪いことを示した．h-index も同様に，ほとんどの要件に対して，不十分であるか，評価なしであった．Altmetrics はそれらに比べ，対応している評価要件が多い，論文の内容に基づいて評価が行えない．一方，3 節の手法は内容に基づいて評価が行える利点があり，他の要件に対しても，相対的に高い評価となることを示した．

本稿で示した比較結果は相対的かつ定性的なものであり，3 節で示した手法により信頼性の高い評価が可能かどうかは検証されていない．定量的な評価実験などを通じ，評価指標としての確立を目指すことが今後の課題である．

参考文献

[1] American Journal Expert.: AJE Scholarly Publishing Reports:2016, American Journal Expert, (2016)
[2] M. Kovanis, R. Porcher, P. Ravaud, and L. Trinquart.: The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise, PLoS ONE, Vol. 11, No. 11, (2016)
[3] E. Garfield.: The History and Meaning of the Journal Impact Factor. Journal of the American Medical Association, Vol. 295, No. 1, pp. 90-93. (2006)
[4] R. Spier.: The history of the peer-review process, Trends in biotechnology, Vol. 20, No. 8, pp. 357-358, (2002)
[5] E. Callaway.: Publishing elite turns against impact factor, Nature, Vol. 535, No. 14, pp. 210-211, (2016)
[6] 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)
[7] R. M. Alguliyev and R. M. Aliguliyev.: Modified Impact Factors, Journal of Scientmetric Research, Vol. 5, No. 3, pp. 197-208, (2017)
[8] J. D. West, T. C. Bergstrom, and C. T. Bergstrom.: The Eigenfactor Metrics™: A Network Approach to Assessing Scholary Journals, College & Research Libraries, Vol. 71, No. 3, pp. 236-244, (2010)
[9] H. F. Moed.: Measuring contextual citation impact of scientific journals, Journal of Informetrics Vol. 4, No. 3, pp. 265-277, (2010)
[10] J. E. Hirsch.: An index to quantify an individual's scientific research output, Proceedings of the National Academy of Sciences of the United States of America, Vol. 102, No. 46, (2005)
[11] J. Priem, D. Taraborelli, P. Groth, and C. Neylon.: altmetrics: a manifesto, http://altmetrics.org/manifesto/, (2010), 最終アクセス 2018 年 11 月 8 日.
[12] Impact Story.: A new framework for altmetrics, http://blog.impactstory.org/31524247207/, (2012), 最終アクセス 2018 年 11 月 8 日.
[13] Shota Shimizu and Yasufumi Takama.: Preliminary Investigation on Quantitative Evaluation Method of Scientific Papers based on Text Analysis, In Proceedings of ACM MEDES'18, (2018)
[14] C. Y. Lin and E. Hovy.: Automatic Evaluation o summaries using N-gram co-occurrence statistics, In Proceedings of NAACL'03, Vol. 1, pp.71-78, (2003)
[15] C. Y. Lin.: ROUGE: A package for automatic evaluation of summaries, In Proceedings of WAS2004, pp. 74-81, (2004)