パターン: 引用からの品質

Ver 1.1　2001年3月　1日
Ver 1.2　2001年3月10日

鷲崎 弘宜*　深澤 良彰†

概要　**「引用からの品質」**は，調査対象とする論文の重要性を，本文の詳細や他の文献からの被引用情報を把握せずに，参考文献から判断することで，効率の良い研究調査を行うために用いられる．

1　別名（Also Known As）

Look Bibliography First （まずは参考文献）

2　文脈（Context）

特定の問題領域の研究調査にあたって，論文誌や雑誌に掲載された学術論文を参照することがある．入手した論文について，内容を全て把握する前に，その品質・重要性を把握可能であれば，効率のよい調査が可能となる．現在入手可能な論文の量は増える一方なため，入手した論文全てに目を通し利用することが困難である [1]．さらに，特定の問題領域の情報が整理されたサーベイ論文が，極端に不足している現状もあり，ある論文が重要かどうか取捨選択するための判断指標が強く求められる．

一般に，多くの重要な論文から引用される論文は重要である，と定義される．例えば，研究者の一流の条件として，論文の被引用回数が 100 回以上とされる．この定義は，論文の筆者は引用文献（参考文献）[1]から影響を受けている，という仮定に基づく [2]．論文の引用関係に基づく参照構造と，被引用情報に基づく重要性評価の状況を，図 1に示す．被引用情報を用いる文献の分析手法は，計量書誌学 (bibliometrics) 的調査であり，研究者個人や国・機関の研究業績評価に広く用いられる [3]．分析時に用いる引用索引データベースとして， ISI [4]社の SCI （Science Citation Index）や，NACSIS-IR [5]がある．また，インターネット上の文書の引用情報を自動抽出・索引づけしたデータベースとして， PostScript・ PDF 文書を対象とする ResearchIndex [6]や Cora [7]，および TEX 文書を対象とする PRESRI [1]等が利用できる．

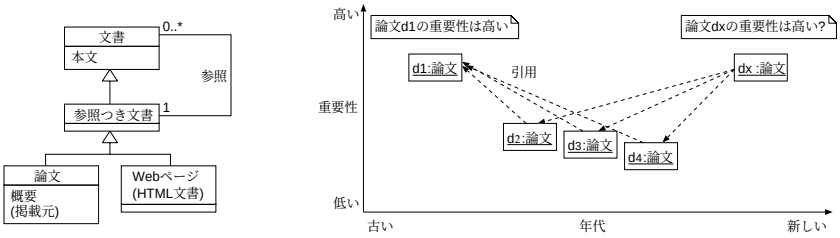

図 1　論文の参照構造（左）　被引用情報による論文の重要性（右）

一方，論文の評価基準として，掲載元に関する情報もまた重要である．論文の掲載元が大会予稿集 (Proceedings) であれば，大会の権威が高い程，その論文が重要と考

*Hironori Washizaki, 早稲田大学大学院理工学研究科

†Yoshiaki Fukazawa, 早稲田大学理工学部

[1]本パターンでは引用タイプ（背景・参考・対比） [1]の区別を考慮しないため，引用文献（References）と参考文献（Bibliography）を同義として扱う．

えられる．学会の論文誌 (Journal) であれば，学会および学会論文誌の権威が高い程重要と考えられる．例えば，ソフトウェア工学の世界では，IEEE Trans on SE や Proc of ICSE に論文が掲載されることが，研究者として一流の条件とされる．

3　問題（Problem）

研究調査にあたり，論文の重要性を適切かつ迅速に，効率よく評価したい．このとき，以下に挙げるそれぞれの**力**（Forces）のバランスをとる必要がある．

主観的判断材料についての制約

- 対象とする論文の本文を読む時間が無い．
- 対象とする問題領域について十分な知識を持たないため，論文の概要記述 (abstract) を一読してもその重要性を判別できない．

客観的判断材料についての制約

- 論文誌および国内外の大会の権威の高さについてある程度の知識を持つ．
- 問題領域に詳しくないため，対象論文には独創的論文（Seminal Paper）としての性質よりも，サーベイ論文（Survey Paper）としての性質が望まれる．
  - 独創性の判断には被引用情報が重要であるのに対し，サーベイ性（網羅性）の判断には自引用情報（参考文献）が重要である．
- 対象論文の被引用情報が得られるとは限らない．
  - 対象論文が発表されたばかりで，引用されるまで時間が経過していない場合がある．また，SCI 等の引用索引データベースは，対象とする領域が限定され，収録対象も著名な論文誌・インターネット上の論文に限られる．
- 対象論文の掲載元情報が得られるとは限らない．
  - インターネット上で論文が多く流通する現状では，対象論文自体の掲載元が不明なことがある．また，学会等で発表されず単に技術記事として公開されているものや，組織内部での文書など，掲載元の情報が存在しないものについても研究調査の対象となることが多い．
- 参考文献の掲載元情報はたいてい得られる．
  - 論文の著者は一般に，第三者が取得可能とするために，掲載元が正確に判明している文献を，その掲載元情報と共に参考文献として記述する．

4　解決（Solution）

対象論文の参考文献を見て，掲載元の権威が高いと判断可能な文献を多く引用している程，対象論文の重要性が大きいと判断する．すなわち，重要な論文を多く引用する論文は重要である．これは，著者の問題領域に関する造詣の深さを，参考文献は反映するという仮定に基づく．各掲載元の権威の高さの判断は，一般的な常識に基づくが，**インパクトファクター**（後述，8.2 節）を用いることもできる．なお，参考文献の数は誌面のスペースの大きさから制約を受けるため，全体における割合から判断することが妥当である．

例として，参考文献が図 2・図 3になっている論文 d_A・d_B を考える．論文 d_A が，国内の研究会・全国大会レベルの大会予稿集を主に参考文献としているのに対し，論文 d_B は，海外論文誌や著名な大会予稿集を参考文献としている．そこで，d_A の著者は国内の手近な文献のみを参考としているのに対し，d_B の著者は国内外の文献を深

くサーベイしていると判断し，d_B は d_A に対してより重要性が高いと判断する．

参考文献
[1] 鷲崎弘宜, “ある適当な研究の報告”, 情報処理学会研究会報告, SE1XX-YY, 1999
[2] 深澤良彰, “何らかの研究報告”, 電子情報通信学会第 ZZ 回総合大会, 2000
[3] .....

図 **2**　論文 d_A の参考文献

参考文献
[1] H.Washizaki, “A Suitable Research”, IEEE Trans on SE, Vol.X, No.Y, 1999
[2] Y.Fukazawa, “A Certain Research Report”, Proc of OOPSLA’ZZ, 2000
[3] .....

図 **3**　論文 d_B の参考文献

5　注意（Caution）

論文の体裁を良く見せるために，本文の内容と無関係に，文脈を無視して著名な論文を引用する場合が考えられ，この場合は，概要記述や本文を注意深く読む必要がある．この問題を，**「スパム参考文献」** [2]として提案し，**「引用からの品質」**の安易な適応では，うまく対処できない場合があることに注意する．類似する問題として，論文誌の**インパクトファクター**値を意図的に上げるために，論文の編集委員会が投稿者に対して同論文誌からの引用を奨励する行為が指摘される [9].

6　適用例（Example）

ある学会論文誌と研究会報告から論文を 10 篇ずつ無作為に抽出し，各論文の参考文献より重要性を評価し，査読プロセスに基づく重要性の高さに合致するか調べた．まず，論文の掲載元の権威の高さに応じた掲載元重要度 $W(d)$ を用意し，ソフトウェア工学関連の論文誌に関する各値を，経験則より表 1に設定した．ただし，専門的な知識を必要とする状況を考え，論文以外の一般の書籍が参考文献として挙げられている場合は重要性算出に加えていない．次に，論文 d_i の重要度 $E(d_i)$ および論文群 D の平均重要度 $E_{avg}(D)$ の定義を以下に行った．

d : 論文，D : 論文群，$\mid D \mid$: D中の論文数

$d_i \rightarrow d_k$: d_kを引用するd_i，N_{d_i} : d_iの総参考文献数

$$E(d_i) ::= \frac{1}{N_{d_i}} \sum_{d_k \mid d_i \rightarrow d_k} W(d_k) \qquad E_{avg}(D) ::= \frac{1}{\mid D \mid} \sum_{d_i \in D} E(d_i)$$

表 **1**　掲載元種別と掲載元重要度（抜粋）

論文 d の掲載元種別	査読の有無	掲載元重要度 $W(d)$
IEEE/ACM Trans. on SE/SEM	有	1.0
Proc. of ICSE, TOOLS, ECOOP	有	0.8
電子情報通信学会論文誌	有	0.5
国内シンポジウム・ワークショップ	有 / 無	0.3
情報処理学会研究会報告	無	0.1

学会論文誌と研究会報告の各 10 編ずつの論文群について，総参考文献数と平均重要度を計測した結果を表 2に示す．評価の結果，平均重要度の値は学会論文誌が研究

[2] word spam に由来する．宣伝などを目的として，Web ページに検索頻度の高いキーワードを意図的に羅列することを word spam または spamdexing と呼ぶ [8].

会報告を上回った．学会論文誌は，査読プロセスに基づき一定以上の品質を持つ論文を採択するため，得られた平均重要度は現実の重要性を数値的に反映しているといえる．なお，学会論文誌の平均重要度の値は，表 1での論文誌の掲載元重要度の値に近く，経験則に基づく重みづけはある程度妥当である事も分かった．この平均重要度の値を，表 1における掲載元重要度の値に再帰的にフィードバックすることで，より精度の高い重要性の判断が可能となると思われるが，**「引用からの品質」**はあくまで迅速な判断を目的とするためここでは行わない．

表 **2**　総参考文献数と平均重要度

論文群 D の抽出元	総参考文献数	平均重要度 $E_{avg}(D)$
学会論文誌	98	0.448
研究会報告	45	0.353

7　結果（Consequences）

「引用からの品質」は以下のような利点を持っている．

- 扱いの容易さ
  論文の引用情報が網羅されたデータベースを必要とせず，学会・論文誌等の権威の高さについて知っている者であれば，容易に論文の重要性を評価できる．
- ある程度の妥当性
  10 編ずつの計 20 編の論文を検査したところ，我々が知るところの各論文誌の重要性と，ある程度一致していることが分かった．

「引用からの品質」の欠点として以下のような点があげられる．

- 権威の高さ判断の難しさ
  各参考文献の掲載元の権威の高さの設定は，**インパクトファクター**値が得られないとき評価者に委ねられるため，よく知らない国際会議などを，どの程度権威が高いか判断することが難しい場合がある．
- 一般書籍の扱い
  一般の書籍であっても，調査対象として十分に有用であるものも多く，何らかの方法で加味する必要がある．しかしながら，参考文献として挙げられ，かつ，未知の書籍についてその権威の高さを判断することは困難である．

8　関連パターン（Related Pattern）

8.1　Evaluate Papers Fast

「Evaluate Papers Fast」[10]は，大会のプログラム委員として投稿論文を査読する際に，題名 → 概要記述 → 参考文献 → 関連研究 → 導入 → 結論の順に読み進めることで，本文に時間を割く前に，すばやく新規性・重要性を判断可能であることを示唆する．**「Evaluate Papers Fast」**では，利用者が問題領域への知識を持つことを暗黙の前提とするのに対して，**「引用からの品質」**は，知識を持たない状況を考慮するため，概要記述よりも前に参考文献を読むことが妥当な解決となる．

8.2　インパクトファクター

インパクトファクター（Impact Factor**）**[9]は，学術雑誌が学術界に対しどれだけ影響を与えているかを，過去 2 年間の被引用情報を用いて算出したもので， ISI 社か

らJCR（Journal Citation Report）として一覧がCD-ROM形式で発行される．短期決戦型のテーマが有利なことや，レビュー論文（サーベイ論文）の比率が高い程有利であるなど，種々の欠陥が指摘されるものの，業績評価等における簡便な手法として，論文自身の掲載元のインパクトファクター値が広く用いられている．**「引用からの品質」**の解決において，参考文献の掲載元の権威の高さを判断する際に，掲載元のインパクトファクター値が得られるならば，その値を比較判断に用いることが可能である．例えば，日本版インパクトファクター値が，情報処理学会論文誌が0.288，電子情報通信学会論文誌D-I/IIが0.278という調査報告[12]があり，この報告に基づくなら，情報処理学会論文誌は電子情報通信学会論文誌よりも権威が高いと判断できる．インパクトファクターは以下に定義される．

$IF_a(x)$：学術雑誌aの西暦x年のインパクトファクター値
$R_a(x,y)$：学術雑誌aでx年に発表された論文がy年に引用された数
$C_a(x)$：学術雑誌aでx年に発表された論文数

$$IF_a(x) ::= \frac{R_a(x,x-1)+R_a(x,x-2)}{C_a(x-1)+C_a(x-2)}$$

8.3 Webページのランキング手法

WWW検索エンジンでは，検索語との類似性に加え，Webページ間のハイパーリンク構造を考慮して，検索結果を順序付ける手法が用いられる[8]．Webページは文書の一種であり，他への参照情報をリンクとして持つ点で論文と類似するが，論文の引用が固定的であるのに対し，Webページのリンクは動的である点について異なる．

- **PageRank**
  PageRank[13]は，検索エンジンGoogleが用いるWebページの重要度評価手法であり，多くの重要なWebページからリンクされるWebページは重要である，という考えを評価方針とする．Webページ間のリンク情報が網羅されたリポジトリの存在を前提とし，また，計算の規模が大きく時間がかかる．Webページd_iのPageRank $PR(d_i)$は以下に定義される．

  d：Webページ，q：ユーザがランダムにページを選択する確率
  $d_k \to d_i$：d_iをリンクするd_k，N_{d_k}：d_k内の総リンク数

  $$PR(d_i) ::= q + (1-q)\sum_{d_k | d_k \to d_i} \frac{PR(d_k)}{N_{d_k}}$$

- **HITS他**
  HITS（Hypertext Induced Topic Search）[14]では，Webページの信頼性評価にあたり，外部からの被リンクの多さに加えて，多くリンクされるWebページに対する自リンクの多さも考慮している．これは，信頼性の高いページへのリンクの多いページは，ユーザにとって道標（hub）となる，という考えに基づく．すなわち，重要なWebページに多くリンクするWebページは重要である多くリンクされるページをオーソリティ（authority），オーソリティに多くリンクするページをハブ（hub）と呼ぶ．オーソリティ度の大きさがWebページの最終的な信頼性評価となる．アルゴリズムの定義を以下に示す．

  オーソリティ度 $a(d_i) ::= \sum_{d_k | d_k \to d_i} h(d_k)$，$\sum_d a(d)^2 = 1$
  ハブ度 $h(d_i) ::= \sum_{d_j | d_i \to d_j} a(d_j)$，$\sum_d h(d)^2 = 1$

また，文献 [15]における**参照重要度**は，**HITS** と同様に，Web ページの重要性評価にあたり，Web ページの外部への自リンクの多さを考慮することで，検索精度の向上を試みている．Web ページの参照重要度 $R(d)$ の定義を以下に示す．

$$S_i : d_i\text{の索引語と検索質問の類似度}$$
$$w_{ij} : d_i\text{から}d_j\text{へのリンクの重み}，\ \alpha : S_i\text{の重み}$$
$$R(d_i) ::= \alpha S_i + \sum_{d_j | d_i \to d_j} w_{ij} R(d_j) + \sum_{d_k | d_k \to d_i} w_{ki} R(d_k)$$

以上より，Web ページの重要性の評価にあたり，被リンク情報だけでなく，自リンク情報も考慮した手法の有用性が，認識されつつあると云える．Web ページは文書の一種であることから，この結果は，**「引用からの品質」**が，参考文献のみを考慮した解決をとることの妥当性を示していると言える．

8.4 高価な事実

「明日出会うかもしれないもの」[11]における**「高価な事実」**は，関心事の直接的な把握が経済的に困難なとき，他の経済的に観測容易な事実から目的とする関心事を推測することで，直接的な把握を回避可能なことを示唆する．**「高価な事実」**を論文の重要性判断に当てはめると，論文の被引用情報の把握が経済的に困難なとき，自引用情報（参考文献）の把握は容易なため，**「引用からの品質」**の解決に合致する[3]．

謝辞 井上健氏，平鍋健児氏にシェファーディングをしていただいたことを深く感謝致します．また，JapanPLoP ライティングワークショップの参加者の方から数多くのご意見を頂いたことに感謝致します．

参考文献

[1] 難波英嗣, 奥村学, “論文間の参照情報を考慮したサーベイ論文作成支援システムの開発”, 自然言語処理, Vol.6, No.5, 1999
[2] M.MacRoberts, B.MacRoberts, “Brief Communication: Citation Content Analysis of a Botany Journal,” JASIS, Vol.48, No.3, 1997
[3] 根岸正光, “研究評価とビブリオメトリックス”, 情報の科学と技術, Vol.49, No.11, 1999
[4] ISI（Institute for Scientific Information）, http://www.isinet.com/
[5] 国立情報学研究所, NACSIS-IR, http://www.nii.ac.jp/ir/
[6] NEC Research Institute, ResearchIndex, http://www.researchindex.com/
[7] JUSTSYSTEM, Cora, http://cora.whizbang.com/
[8] 原田昌紀, “サーチエンジンにおける検索結果のランキング”, bit, Vol.32, No.8, 2000
[9] 山崎茂明, “インパクトファクターをめぐる議論: 正しい理解と研究への生かし方”, 情報管理, Vol.41, No.3, 1998
[10] J.Palsberg, “Evaluate Papers Fast,” http://c2.com/cgi/wiki?EvaluatePapersFast
[11] 友野晶夫, “明日出会うかもしれないもの”, http://www.kame-net.com/jplop/
[12] 根岸正光, 西澤正己, 礒谷峰夫, 蓑毛堅一郎, 浅野正一郎, “電子情報通信学会の評価 –学会誌の引用数分析を中心に–”, 電子情報通信学会誌, Vol.82, No.5, 1999
[13] 馬場肇, “Google の秘密 –PageRank 徹底解説–”, http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html
[14] J.Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” ACM-SIAM Sympo. on Discrete Algorithms, 1998
[15] 大野潮満, 黄瀬浩一, 松本啓之亮, “参照重要度に基づく WWW 検索”, 情報処理学会研究会報告自然言語処理, NPL135-1, 2000

[3] この関連性は誤っているかも知れない．**「引用からの品質」**が目的とする把握したい関心事は，直接的には，論文の重要性であって被引用情報ではない．