引用箇所の間隔に基づいた共引用の検討

江藤　正己†

† 慶應義塾大学大学院文学研究科　〒 108–8345 東京都港区三田 2–15–45
E-mail: †eto@slis.keio.ac.jp

あらまし　類似論文検索に用いられる代表的な類似度指標の一つに，共引用の関係を利用するものがある．ただし，従来の共引用の指標は，引用論文の本文の内容とは無関係に「一つの引用論文で示された共引用関係にある論文同士は全て等しく類似している」と仮定され，本文中の引用のされ方による被引用論文間の関係の違いが考慮されていない．そこで，共引用関係にある 2 論文間の関係を引用箇所に基づいて「非同一段落」「同一段落」「同一文」「列挙」に分け，この順に間隔が短いと定義する．本稿では，引用箇所の間隔が短くなるほど被引用論文間の類似性が強くなるという仮説を設定し，その検証のために各種類の共引用関係にある論文間の類似度を算出・比較する実験をおこなった．実験の結果，間隔の長さに応じて共引用関係にある論文間の類似度が段階的に変化することがわかり，仮説が検証された．
キーワード　類似検索，論文検索，共引用

1. はじめに

1.1 共引用

類似論文検索をおこなうには，論文間の類似度を表す指標が必要となる．これまで提案された指標の中で，代表的な一つに共引用 [1] がある．共引用の類似度指標値は，算出対象の論文のペアが同一の論文から共に引用される回数をもとに求められる．これは，同一の論文によって引用される論文同士には類似性があるという前提によるものである．たとえば図 1 の場合，類似度の算出対象は論文 A と論文 B であり，一般に A と B を共に引用している論文の数によって指標値が求まる．A と B を共に引用している論文の数が多ければ類似度が高くなり，少なければ類似度は低くなる．

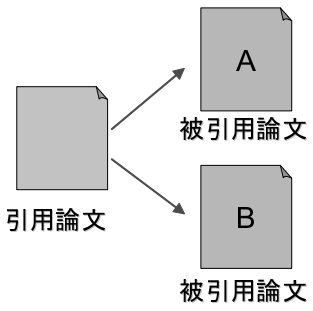


図 1　共引用

共引用は，論文データーベースである *CiteSeer* [2] や *Web of Science* [3] などの実用システムで用いられており，その有用性は高いと言える．また，Web ページ [4] や特許 [5] などの分野にも応用され成果を上げている．

なお，共引用は類似度指標以外にも「ある論文がそれ以前の二つの論文を同時に引用している現象（図 1）」自体も指す言葉である．これらを厳密に使い分けるとすれば「共引用の類似度指標値は，一定数の論文群を調査し，対象のペアを共引用する現象が何回生じたかによって算出される」となる．

1.2 共引用を利用した従来類似度指標の限界

従来の共引用では，引用している論文の文脈を無視し，現象が出現するか否かのみに基づいて類似度が算出される．すなわち，本文における引用のされ方，どのような意味において二つの被引用論文は共引用関係にあるのかは考慮されず,「一つの引用論文で示された共引用関係にある論文間の類似度は全て同じ」と仮定されている．そのため，たとえば,「引用論文の冒頭で引用された論文と末尾で引用された論文間の類似度」は「同一文中で引用された論文間の類似度」と等しい値になる．しかし，被引用論文間の関係は, 引用論文の本文における引用のされ方によって違い，それによって類似性の強弱も異なっているはずである．たとえば，実験の方法を述べている場所で引用されている被引用論文同士は強く類似し，反対に，実験方法を述べている場所で引用された論文と理論的背景を述べている場所で引用された論文はあまり類似していないと推測される．従来の共引用では，そのような関係の違いに関する情報を削ぎ落とした形で類似度が算出されている．

単なる共引用関係を超えて，より精密な類似度の指標を考えようとした場合，引用論文の本文における引用のされ方の情報を利用することは有効であると思われる．論文は，順を追って議論が展開しているため，一般に近くの箇所で引用された論文同士の方が，遠くの箇所で引用された論文同士よりも内容的に類似していると推測できる．引用箇所の近さを用いて，被引用論文間の内容の類似性の強弱をとらえることができれば，それに基づくことで共引用の指標を精密にでき，類似論文検索システムの高度化につながると考えられる．そこで，本稿では，引用論文の本文を解析して被引用論文間の関係の違いを考慮する，

従来の共引用の拡張手法について提案をおこなう．

以下，本稿ではまず 2 章で，引用を扱う際に引用論文の本文の利用を試みている先行研究についてふれる．次に 3 章では，引用論文の本文を用いて共引用を拡張する提案手法について述べる．4 章では提案手法を検証するために，引用箇所の間隔の長さに応じて共引用関係にある論文間の類似度が段階的に変化するかを調べる実験をおこなう．そして，5 章で本稿のまとめをおこない，6 章で今後の課題について言及する．

2. 先行研究

引用論文の本文は，被引用論文に関わる多くの情報をもっていると考えられる．しかしながら，近年まで，大規模論文群を対象にした場合，本文の内容に依拠した形で引用が利用されることはあまりなかった．その原因は，機械可読形式の論文が少なかったことや機械処理で引用を解釈することが難しかったことにあると思われる．しかし，現在は，機械可読形式の論文が増え，本文の内容に依拠した引用を取り扱うことが可能になってきている．そして，引用論文の本文の内容を使って引用から多くの情報を取り出し，検索などに利用する研究もおこなわれはじめている．ここではそのような試みについてふれ，本文を解析して得た情報と引用を結びつけることの意義や有用性を確認する．

2.1 引用文章に含まれる語を利用する研究

「引用文章（引用箇所の周辺の文）に含まれる語は，被引用論文の内容を説明する」という発想をもとにいくつかの研究がおこなわれている．Web ページのアンカーテキストを利用する研究 [6] に近い．

この種の研究の初期的なもので，引用文章に含まれる語を単純にキーワードとして被引用論文に付与することで検索性能を向上させる研究が，O'Conner [7] によっておこなわれている．その後，多くの引用文章で用いられる語に強い重みづけをする研究が Bradshaw [8] によって，引用文章の中で被引用論文に関する語と無関係な語を区別する研究が Ritchie [9] らによっておこなわれている．

また，検索以外に引用文章に含まれる語を利用するような研究として，自動的なシソーラス構築を目指すもの [10] や，データマイニングに適用するもの [11] がある．

2.2 引用の役割を自動分類する研究

また，論文群の高度な組織化を目的として，論文間のつながりである引用をいくつかのカテゴリに自動分類する研究もおこなわれている．これらの研究の多くは，それ以前に提案されてきた引用の役割カテゴリ ([12] など) への分類を目的としている．すなわち，ここで挙げる研究は，より一般的には，教師ありの自動分類の研究に相当し，引用文章の特徴とその引用が属するカテゴリを学習して未知の引用に対して自動的に正解カテゴリを付与する研究であるといえる．

この研究に関して，引用文章に含まれる語や句を手がかりとして，引用を分類する試みが，Garzon ら [13], 難波ら [14],Teufel ら [15] によっておこなわれている．また，システムと人間との間でインタラクティブにやりとりしながら分類ルールを作成していく手法が Pham ら [16] によって，有限オートマトンを用いた機械学習による手法が Le ら [17] によって提案されている．

2.3 類似論文検索への応用

2.2 で挙げた難波らは，引用をカテゴリに分類するだけでなく，それを利用して書誌結合を拡張する手法を提案している．書誌結合とは Kessler [18] によって，共引用よりも前に提案された論文の類似度指標である．書誌結合の類似度指標では，図 2 で示すような論文間の関係をもとに，論文 A と論文 B がどれだけ同じ論文を引用しているかによって，引用論文間の類似度が算出される．

難波らの手法では，書誌結合の類似度を算出する際に，それぞれの引用のカテゴリを考慮する．そして，二つの引用論文が同一のカテゴリで被引用論文を引用している（たとえば，論文 A と論文 B が同じ「問題点の指摘」カテゴリで被引用論文を引用している）回数をもとに類似度を算出する．難波らの方法は，引用論文本文の内容に依拠することで書誌結合を精密化するものであり，本研究のアイデアに極めて近いものといえる．

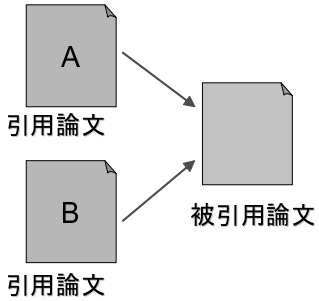


図 2 書誌結合

2.4 共引用文脈分析

本稿がとりあげている共引用の観点から引用と被引用論文の関係をとらえるものとして，共引用文脈分析 [19] と呼ばれるものがある．この研究は，論文集合を組織化するために，その集合を引用している論文の本文を利用する研究である．具体的には，人間が引用論文を読むことにより，そこで引用された被引用論文同士がどのような関係にあるかをとらえ，それをもとに被引用論文集合を組織化する手法である．

このような研究が存在することからも明らかなように，(1) 共引用の関係が一種類のみのものではなく，(2) 本文を用いることで被引用論文間の関係をとらえる情報を引き出すことができることがわかる．しかし，たとえ本文から有用な情報を引き出せるとしても，人間でなければできない共引用文脈分析を，検索が必要となる大規模な論文集合に対して適用することは不可能である．類似論文検索を考えた場合，機械処理を念頭においた共引用文脈分析が必要となる．本研究は，従来の共引用文脈分析を，類似論文検索に適用可能な形でおこなうことを目指すものととらえることもできる．

3. 引用箇所の間隔に基づいた共引用

3.1 本文の内容に依拠した共引用

1.2 でも述べたように，従来の共引用の限界を克服するには，

論文の構成を考慮し，本文の内容に依拠した共引用を考えていくことが望ましい．また，2 章でも述べたように，実際に引用論文の本文の情報を用いることで成果を挙げているような研究も多く存在する．そこで，本稿では本文における被引用間の関係をとらえる新たな共引用の尺度を提案する．そのような共引用の尺度を用いて被引用論文間の類似度を算出することで，共引用に基づく類似度指標を引用論文が示す被引用論文同士の関係をより反映したものへと拡張することができる．

1.2 でも述べたように論文は体系立てて記述がおこなわれる．そのため，意味的に近い内容のものはまとまって述べられると考えられる．そして,「同じ意味のまとまり内で引用された被引用論文同士」と「異なる意味のまとまりで引用された被引用論文同士」では，前者の方が強く類似していると想定することができる．すなわち，1.2 で挙げた「同一文中で引用された論文間の類似性」と「冒頭で引用された論文と末尾で引用された論文間の類似性」を比較した場合，前者が後者よりも強いことが予想される．

本稿では，この意味のまとまりをとらえるために，段落や文といった論文の構造に着目する．論文では，同じ意味のまとまりが，一つの構造を形成して表現されると考えられるためである．同一の構造内で引用されているか否かを調べることで，同じ意味のまとまりで引用されているか否かをとらえることができる．

以上のことから，引用箇所の構造を調べることによって，被引用論文同士の類似性の強弱を推測できるのではないかと考えられる．たとえば，図 3 のような引用をおこなっている場合，論文 b と論文 c は同じ構造内で引用されているため強く類似し，論文 a と論文 d は異なる構造で引用されているため，弱く類似していると判断される．もしこの推測が正しければ，引用論文の本文を構造の視点から解析することで，被引用論文間の類似性の強弱を判別できることになる．

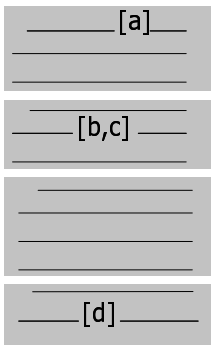


図 3　引用論文本文の例

そこで本稿では，論文の構造をとらえた共引用の尺度が被引用論文間の類似性の強弱を適切に反映できるものであるかについて検証をおこなう．なお，詳しくは 3.3 で述べるが，本稿ではどの構造単位内で共引用されているかを「引用箇所の間隔の長さ」で表す．そして，同一文内で共引用された場合と同一段落内で共引用された場合では，前者の間隔は後者の間隔よりも短いと表現する．

引用箇所の間隔の長さを用いて，検証事項を理論仮説として定義すれば,「引用箇所の間隔が短くなればなるほど共引用関係にある論文間の類似性は強くなり，間隔が長ければ長いほど共引用関係にある論文間の類似性は弱くなる」となる．この理論仮説が検証されれば，引用箇所の間隔を用いて共引用関係にある論文間の類似性の強弱を推測できるため，それを用いることで共引用の関係を利用した類似度指標をより精密なものにすることができる．

3.2　論文の構造

3.1 でも述べたように，本稿で提案する手法は，段落や文といった論文の構造に着目するものである．ここでは，この論文の構造に関して詳細に述べる．

論文における一般的な構造の単位として,「章」「節」「段落」「文（一文)」の四つを挙げることができる．ただし，このうち「章」「節」は，論文によって，著者によって，まとまりの意味合いが大きく異なることが多い．この二つは構造の単位として不安定であるため，引用箇所の間隔をとらえるのに利用するのに不適当と判断した．

さらに，引用の観点からとらえた構造の単位であり,「文」よりもさらに小さな構造単位として「列挙」を用いる.「列挙」とは，同一箇所で複数の論文を並列に列挙した引用 (図 4，[20] より) をとらえた構造単位である．筆者のこれまでの研究により，列挙形式の引用で引用された論文間の類似度は，その他の形式で引用された論文間の類似度よりも高い値になることが明らかになっている [21].

> A very few recent papers address techniques that adapt to dynamic environment[Zell90,Pang93,Brow92, Brow93,Meht93b].

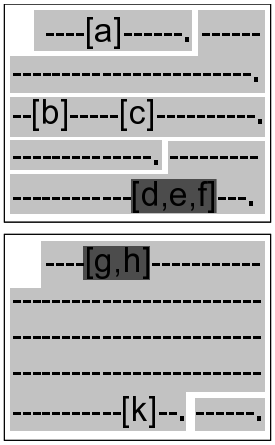
図 4　列挙形式の引用の例

以上のことから，引用箇所の間隔を求めるのに利用する構造単位として,「段落」「文」「列挙」の三つを用いることとした．

3.3　論文の構造を用いてとらえた引用箇所の間隔

「段落」「文」「列挙」の三つの構造単位を用いることで，非同一段落共引用，同一段落共引用，同一文共引用，列挙共引用の 4 種類の共引用から成る尺度を設定することできる．たとえば，図 5 左のような本文があった場合，それぞれの種類の共引用は図 5 右で示したものとなる．構造の単位が小さいほどそこで述べられる意味のまとまりが強いと考えらため，4 種類の共引用はここで示した順に引用箇所の間隔が短くなると考える．なお，構造が包含関係を持つため，小さな構造単位の共引用は大きな構造単位の共引用に含まれる (たとえば，列挙共引用は，同一段落共引用に含まれる) ともいえるが，一つの共引用は最も小さな構造単位の種類の共引用にのみ属すると考える．

この引用箇所の間隔は構造単位を基準とするため，必ずしも表層的な距離の縮尺とは一致しない．たとえば，図 5 において，



■非同一段落共引用
例［a, g］

■同一段落共引用
例［a, f］

■同一文共引用
例［b, c］

■列挙共引用
例［d, e］

図 5　間隔に基づいた共引用の例

a-f と f-g の間の表層的な距離は，前者の方が後者よりも長い．しかし，a と f は，段落という同じ構造内に含まれるが，f と g は異なる構造に含まれる．引用箇所の間隔の観点では，同じ構造内にあるもの同士の間隔の方を短いととらえるため，a-f の方が f-g よりも間隔は短いと考える．

同様に，c-d 間と h-k 間では，表層的な距離は h-k 間の方が長い．しかし，構造の観点からみた場合，h と k は同一「文」内で出現し，c と d は「文」よりも大きな同一「段落」内で出現する．小さな構造で出現する場合の方が間隔は短いととらえるため，間隔の長さは h-k 間の方が短いと考える．

なお，引用論文の本文において，被引用論文が出現する回数は１回とは限らない．そのため，ある特定の論文ペアに対して，複数の引用箇所の間隔が存在する場合がある．場合によっては，特定のペアが，ある段落では「列挙共引用」であり，別の段落では「同一文共引用」であることもある．一つの引用論文において，特定のペアに複数の間隔が存在する場合は，その中で最も間隔の短い種類の共引用にのみ分類することとする．

3.4 作業仮説

3.3 で定義した引用箇所の間隔に基づいた共引用を用いると，3.1 で述べた理論仮説「引用箇所の間隔が短くなればなるほど共引用関係にある論文間の類似性は強くなり，間隔が長ければ長いほど共引用関係にある論文間の類似性は弱くなる」は作業仮説「共引用関係にある論文間の類似度は，非同一段落共引用・同一段落共引用・同一文共引用・列挙共引用の順に高くなる」に具体化される．ここで類似度について次のように考える．

仮説を検証するためには，各種類の共引用関係にある論文間の類似性の強さを求めて，それらを比較する必要がある．そこで，本稿では各種類の共引用関係にある論文間の類似性の強弱を，従来から用いられており一定の評価がなされている tf∗idf/cosine などの論文の類似度指標を用いて算出する（指標の詳細については，4.4 で述べる）．もし類似度の値が間隔の長さに応じて段階的に変化すれば，仮説が検証されたことになる．

そこで仮説を検証するために，引用論文の本文を解析し，それぞれの共引用を引用箇所の間隔に基づいて分類し，各種類の共引用関係にある論文間の類似度を算出・比較する実験をおこなった．その詳細について，次章で述べる．

4. 実　　験

4.1 実験で用いたデータ

実験には，*CiteSeer* が公開しているデータセット *CiteSeer Metadata* [22] を利用した．このデータセットには，約 57 万件分の論文の書誌事項，論文の引用関係に関する情報，論文全文を入手するための URL などが含まれている．

4.2 引用論文集合の作成

データセット内の論文の中から，タイトルかディスクリプタに語「database」を含む論文を選び，その全文のダウンロードを試みた．ダウンロードできた 13,551 件の中から，本文の解析をプログラムで処理し易い，引用記号が

- 大括弧に囲まれている
- 括弧内が数字とアルファベットの組み合わせ

該当例　… [CACS94] [Bon97b]
非該当例　… 1),(1),[1],(CACS94) ,[Bon]

であるものを条件として抽出をおこない，引用論文集合とした．その数は，1,468 件となった．

4.3 被引用論文集合の作成

データセットが持つ引用関係の情報をもとに，引用論文集合が引用している論文を収集した．これを被引用論文集合とする．被引用論文集合の数は，4,592 件となった．なお，引用論文集合と被引用論文集合には重なりがある．また，詳しくは 4.4 で述べるが，共引用関係にある論文間の類似度を算出する方法の一つとして，語の共出現数 (tf∗idf/cosine) を用いるため，被引用論文についても全文のダウンロードを試みた．

4.4 類似度の算出に用いる指標

類似度を算出する指標として，多くの先行研究や実用システムで利用されており，一定の評価がなされていると判断される「tf∗idf/cosine」「書誌結合」「従来の共引用」を用いる．「tf∗idf/cosine」は語の共出現頻度 (図 6) を用いて，「書誌結合」は同じ論文を引用している数 (図 7) によって類似度を算出する指標である．「従来の共引用」は，図 8 のように *CiteSeer Metadata* の約 57 万件の論文から，何回共引用されているかを求める（ここで，「従来の共引用」を用いて類似度の算出をおこなうことは，提案尺度で 4 種類に分類した被引用論文ペアの類似度を，従来の共引用指標で算出することを意味する）．

なお，「書誌結合」と「従来の共引用」については，コサイン係数で正規化処理をおこなった指標も用いる．この正規化処理は，「書誌結合」では引用数を「従来の共引用」では被引用数を補正するためのもので，先行研究でも用いられているものである [23]．以上示した五つの類似度指標で算出をおこなうことで，多様な類似性の観点から検証し，実験の信頼性を高める．

各指標の類似度の具体的な算出は次のようにおこなった．ここでは，求める類似度を S，類似度の算出の対象となる論文を P_1，P_2 とし，P が引用している論文集合を $citing(P)$，その数を $count(citing(P))$，P を引用している論文集合を $cited(P)$，その数を $count(cited(P))$ とする．

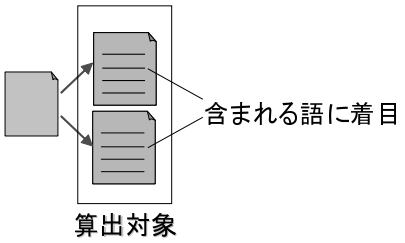


図 6 「tf∗idf/cosine」による評価

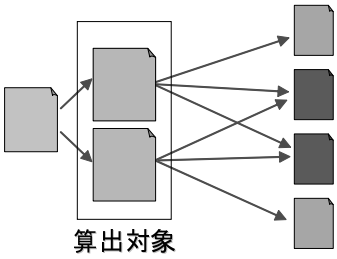


図 7 「書誌結合」による評価

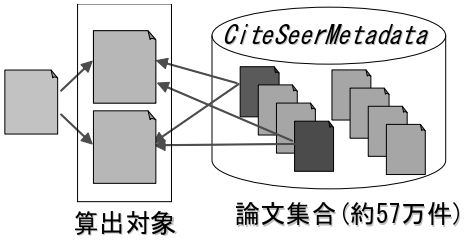


図 8 「従来の共引用」による評価

「tf∗idf/cosine」

$$S = \frac{\overrightarrow{\mathrm{P}_1} \cdot \overrightarrow{\mathrm{P}_2}}{|\overrightarrow{\mathrm{P}_1}| \cdot |\overrightarrow{\mathrm{P}_2}|}$$

ここで，ベクトル $\overrightarrow{\mathrm{P}}$ 中の各要素（語）に対する重み W は以下のように求める．

$$W = log(\frac{\text{その語の出現回数}}{P \text{ の延べ語数}} + 1) * (log(\frac{\text{総文書数}}{\text{出現文書数}}) + 1)$$

ただし，全文をダウンロードできた全ての論文 (引用論文集合＋被引用論文集合) (15,713 件) を総文書とする．また，各語を POSTagger ソフト *MontyLingua* [24] を用いて原型に変換してから，計算をおこなった．

「書誌結合」

$$S = count(citing(P_1) \cap citing(P_2))$$

「正規化書誌結合」

$$S = \frac{count(citing(P_1) \cap citing(P_2))}{\sqrt{count(citing(P_1)) \times count(citing(P_2))}}$$

「従来の共引用」

$$S = count(cited(P_1) \cap citied(P_2))$$

「正規化共引用」

$$S = \frac{count(cited(P_1) \cap citied(P_2))}{\sqrt{count(cited(P_1)) \times count(cited(P_2))}}$$

4.5 各種類の共引用の類似度集計方法

実験は，引用論文の本文から得られた引用箇所の間隔の情報が類似性の強弱を判別できるか調べるためにおこなうものである．そのため，以下のように引用論文を単位とした方法で各種類の共引用に分類される論文間の類似度を集計した．この集計方法は，全 5 指標において共通である．

まず，引用論文毎に算出対象の種類に該当する共引用のペアの類似度を平均する．この結果を $S_t(P)$ とする．t は,「列挙共引用」「同一文共引用」「同一段落共引用」「非同一段落共引用」のいずれかである．次に，当該種類の共引用が存在した引用論文のすべての類似度を平均する．したがって，各共引用の類似度の集計値 $\widetilde{S}$ は次式で求まる．ただし，n は当該種類の共引用を含む引用論文の総数である．

$$\widetilde{S} = \frac{\sum_{k=1}^{n} S_t(P_k)}{n}$$

t を列挙共引用，算出指標を tf∗idf/cosine とした場合の類似度集計過程は図 9 のようになる．

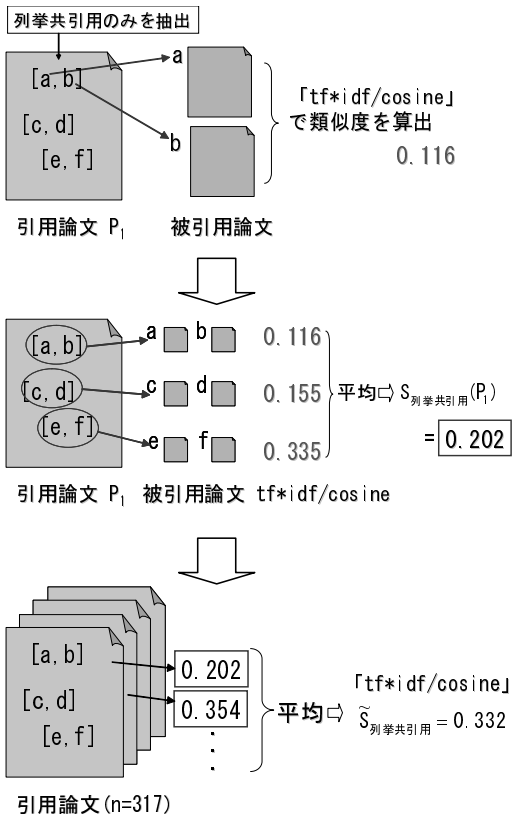


図 9 類似度の算出・集計の例

4.6 各種類の共引用の比較結果

4.4 で述べた指標，及び 4.5 で述べた方法を用いて類似度の算出・集計をおこなった．算出・集計に用いた共引用関係にある論文のペア数及び引用論文数は表 1 で示すとおりである．なお，引用論文集合中にページがページ番号の昇順に並んでいない論文などがあったため，利用することができたのは引用論文集合中の 1,055 件であった．

表 1　類似度算出・集計に用いたデータ数

	非同一段落共引用		同一段落共引用		同一文共引用		列挙共引用	
	引用論文数	共引用ペア数	引用論文数	共引用ペア数	引用論文数	共引用ペア数	引用論文数	共引用ペア数
$tf*idf/cosine$	772	14430	500	2670	264	813	317	827
正規化書誌結合	810	26254	564	4620	333	1298	411	1490
正規化共引用	892	34596	650	6013	392	1719	465	1870
書誌結合	892	34596	650	6013	392	1719	465	1870
従来の共引用	892	34596	650	6013	392	1719	465	1870

表 2　類似度集計結果

	非同一段落共引用	同一段落共引用	同一文共引用	列挙共引用
$tf*idf/cosine$	0.177	0.204	0.251	0.332
正規化書誌結合	0.094	0.129	0.163	0.232
正規化共引用	0.142	0.178	0.223	0.273
書誌結合	0.374	0.572	0.799	1.057
従来の共引用	5.676	7.820	12.449	17.028

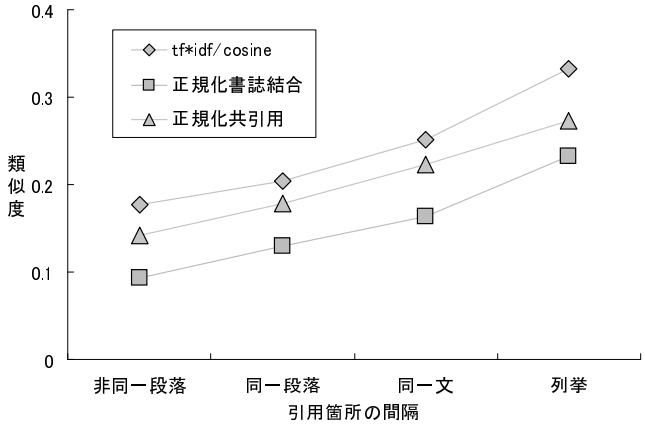


図 10　各種類の共引用の類似度比較

なお,「tf*idf/cosine」ではペアの一方の被引用論文の全文が入手できなったもの,「正規化書誌結合」ではペアの一方の被引用論文が引用数 0 であったものについては，類似度の算出をすることが不可能なため，対象から除外した．このため，この二つの指標は他の指標よりも算出対象数が少なくなっている．

算出の結果が表 2，図 10 である．図 10 は，類似度が 0 〜 1 の範囲内で正規化される,「tf*idf/cosine」「正規化書誌結合」「正規化共引用」の三つをグラフ化したものである．この表 2，図 10 からも明らかなように，算出をおこなった「tf*idf/cosine」「正規化書誌結合」「正規化共引用」「書誌結合」「従来の共引用」のどの指標においても,「非同一段落共引用」「同一段落共引用」「同一文共引用」「列挙共引用」の順に類似度が高くなることが分かった．なお，全指標毎に全ての共引用の種類間で，平均値の差の検定をおこなったが，有意水準 1% で帰無仮説は全て棄却された．

表 3　これまでの共引用の類似度算出・集計に用いたデータ数

	引用論文数	共引用ペア数
$tf*idf/cosine$	927	18740
正規化書誌結合	972	33662
正規化共引用	1055	44198
書誌結合	1055	44198
従来の共引用	1055	44198

4.7　これまでの共引用の尺度との比較

本稿で提案した尺度は，これまでの共引用の尺度では一種類しかなかったものを引用箇所の間隔に基づいて 4 種類に分けるものである．そこで，これまでの共引用の尺度と引用箇所の間隔に基づいた共引用の尺度を比較する実験をおこなった．比較は，前述の実験と同様 1,055 件の引用論文を使い，4.4 の指標を用い，4.5 の方法でおこなった．これまでの共引用は 4 種類の共引用を全て同一とみなすので，その算出となる共引用ペア数は表 1 における 4 種類の共引用ペア数の合計したものになる（表 3）.

これまでの共引用と引用箇所の間隔に基づいた共引用を指標別に比較したものが図 11 である．図 11 で示されるように，これまでの共引用では算出された類似度が高い共引用ペアと低い共引用ペアをひとくくりにまとめているが，引用箇所の間隔に基づいた共引用ではそれらを区別していることを確認できる．

4.8　考　　察

4.6 の実験結果より，算出をおこなったどの指標においても，引用箇所の間隔の長いものほど共引用関係にある論文間の類似性は弱く，間隔が短いものほど類似性が強いことが分かった．このことにより，仮説は検証されたといえる．したがって，本稿で提案した引用箇所の間隔の尺度によって，共引用関係にある論文間の類似性の強弱を推測可能なことが明らかになった．

また，各指標によって算出された論文間の類似度の差は，0 〜 1 の範囲で類似度が正規化される三つの指標に着目した場合でも，極端に小さいものでないといえる．これは引用論文の本文を解析して，被引用論文間の関係の違いをみる提案尺度の有効性を示唆するものといえる．

そして，4.7 の実験結果より，これまでの共引用の尺度がとらえていない類似性の強弱差を提案尺度がとらえていることを確認できた．4.7 の実験は，これまでの共引用が引用箇所によ

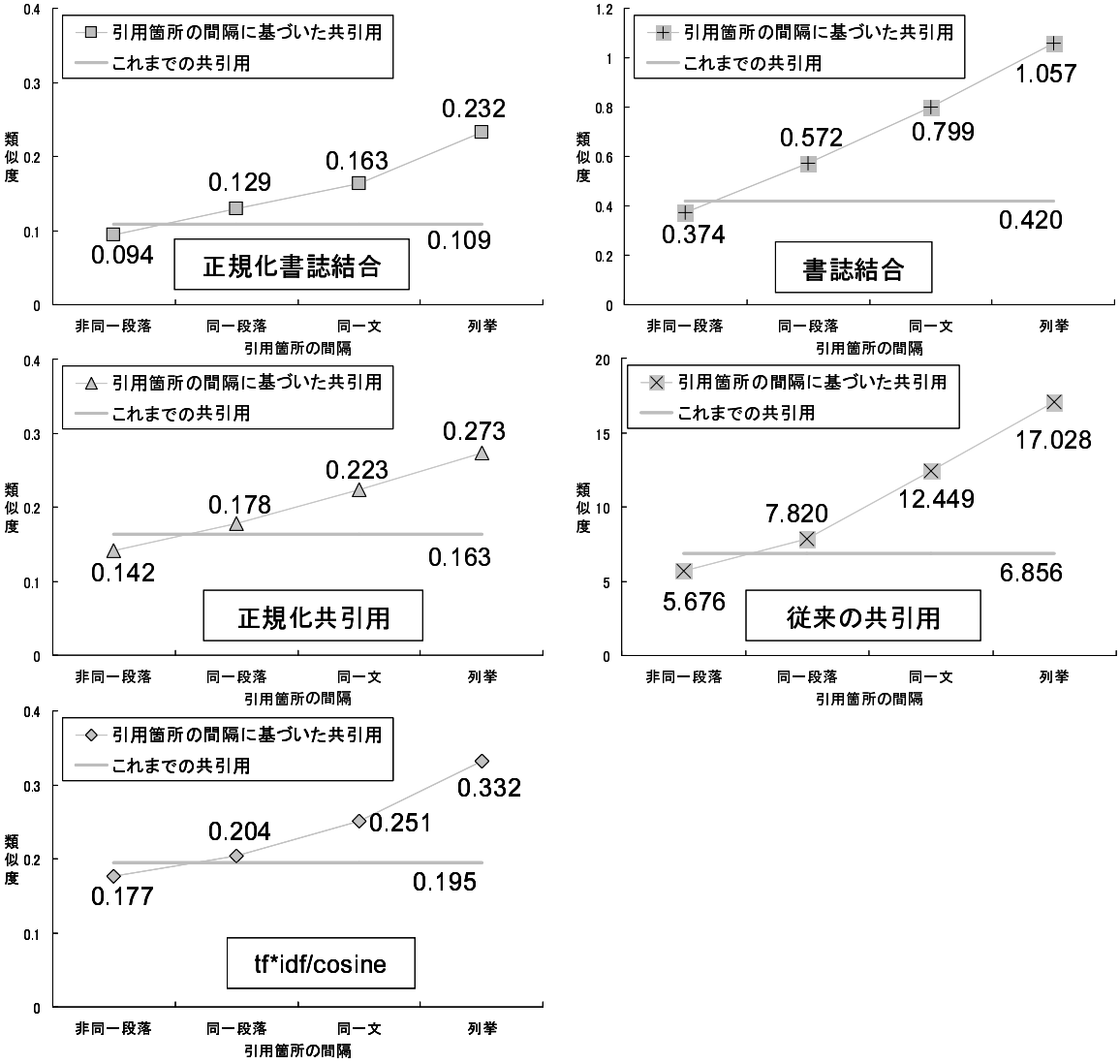


図 11　これまでの共引用の尺度との比較

る被引用論文間の関係の違いを考慮せず，全ての共引用関係をひとくくりにした粗い手法であることを数値的・視覚的に示したものである．

5. ま と め

本稿では以下のことをおこなった．(1) 引用論文の本文の内容に依拠した共引用として，論文の構造からみた引用箇所の間隔を利用することを提案した．(2) これまでの共引用を引用箇所の間隔の長さによって 4 種類に分けた．(3) 実験をおこない，引用箇所の間隔の長さに応じて共引用関係にある論文間の類似度が段階的に変化するか否かを検証した．また，これまでの共引用と引用箇所の間隔に基づく共引用との比較もおこなった．(4) 間隔に応じて類似度が段階的に変化することが分かり，引用箇所の間隔で共引用の類似性の強弱を推測できることを明らかにした．また，これまでの共引用ではとらえられない類似性の強弱差を提案手法がとらえられること確認した．

6. 今後の課題

本研究は，引用箇所の間隔に基づいた共引用を用いた類似論文検索システムの構築を目指すものである．システムでは，引用箇所の間隔に基づいた共引用の尺度を使って，論文間の類似度指標値を算出することになる．その算出のためには，次の二つの課題を解決していかなければならない．

一つ目は，重みの設定方法である．今回の結果から，引用箇所の間隔が短ければ強い重みを設定し，引用箇所の間隔が長ければ弱い重みを設定することが有効であることが分かった．ただし，実際の検索システムにおいては，重みに実際の数値を与えなければならない．そのため，各種類の共引用にどのように重み設定するかについて検討する必要がある．

二つ目は，論文間の類似度の算出方法である．実際の類似論文検索システムでは，たとえば図 12 の論文 a と論文 b のように，複数の論文から引用された被引用論文同士の類似度を算出することになる．図 12 は，a と b は一方の引用論文では列挙共引用されており，もう一方の引用論文では非同一段落共引用さ

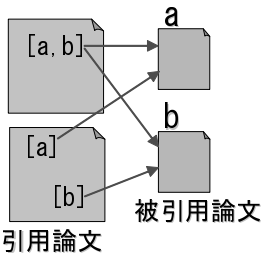


図 12　複数の引用論文による評価

れている状態を示している．このような場合，論文 a と論文 b の類似度をどのように計算するのかについて考える必要がある．

これらの課題に対処したうえで，最終的には「従来の共引用」と「引用箇所の間隔に基づいた共引用」の間で類似論文の検索性能比較実験をおこなう予定である．なお，この際には，ユーザーによる評価もおこない定性的な評価をおこなうことも検討している．

謝　　辞

本研究において，慶應義塾大学文学部の原田隆史助教授，岸田和明教授，田村俊作教授，細野公男名誉教授にご指導を賜りました．また，同大学理工学部の遠山元道専任講師から貴重なご意見を頂きました．ここに深く感謝の意を表します．

文　　献

[1] H. Small, "Co-citation in the scientific literature:a new measure of the relationship between two documents," Journal of the American Society for Information Science, vol. 24, pp. 265-269, July Aug. 1973.

[2] CiteSeer,http://citeseer.ist.psu.edu/

[3] Web of Science, http://www.thomsonscientific.jp/products/wos/index.shtml

[4] M. Beigbeder, T. Lafouge and C. Prime-Claverie, "Transposition of the cocitation method with a view to classifying web pages," Journal of the American Society for Information Science and Technology, vol. 55, pp. 1282-1289, Dec. 2004.

[5] K. Lai and S. Wu, "Using the patent co-citation approach to establish a new patent classification system," Information Processing and Management, vol. 41, pp. 313-330, Mar. 2005.

[6] O. A. McBryan, "GENVL and WWWW: Tools for taming the web,",Proceedings of the First International World Wide Web Conference,pp79-90, May 1994.

[7] J. O'Connor, "Citing statements: computer recognition and use to improve retrieval," Information Processing and Management, vol. 18, pp. 125-131, 1982.

[8] S. Bradshaw, "Reference directed indexing: Redeeming relevance for subject search in citation indexes." in ECDL, 2003, pp. 499-510.

[9] A. Ritchie, S. Teufel and S. Robertson, "How to find better index terms through citations," in Proceedings of the Workshop on how can Computational Linguistics Improve Information Retrieval? 2006, pp. 25-32.

[10] J. W. Schneider, "Concept symbols revisited: Naming clusters by parsing and filtering of noun phrases from citation contexts of concept symbols," Scientometrics, vol. 68, pp. 573-593, Dec. 2006.

[11] N. I. Preslav, S. S. Ariel and H. A. Marti, "Citances: Citation sentences for semantic analysis of bioscience text," in Proceedings of the SIGIR '04 Workshop on Search and Discovery in Bioinformatics, 2004.

[12] I. Spiegel-Rösing, "Science Studies: Bibliometric and Content Analysis," Social Studies of Science, vol. 7, pp. 97-113, 1977.

[13] M. Garzone and R. E. Mercer, "Towards an automated citation classifier," in AI '00: Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, 2000, pp. 337-346.

[14] 難波英嗣, 神門典子, 奥村学, "論文間の参照情報を考慮した関連論文の組織化," 情報処理学会論文誌, vol. 42, pp. 2640-2649, 2001.

[15] S. Teufel, A. Siddharthan and D. Tidhar, "Automatic classification of citation function," in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006, pp. 103-110.

[16] S. B. Pham and A. G. Hoffmann, "A new approach for scientific citation classification using cue phrases." in Australian Conference on Artificial Intelligence, 2003, pp. 759-771.

[17] M. Le, T. B. Ho and Y. Nakamori, "Detecting citation types using finite-state machines." in PAKDD, 2006, pp. 265-274.

[18] M. M. Kessler, "Bibliographic coupling between scientific papers," American Documentation, vol. 14, pp. 10-25, 1963.

[19] H. Small, "Co-citation context analysis and the structure of paradigms, " Journal of Documentation. vol. 36, no. 3, 1980, p. 183-196.

[20] V. Soloviev, "An incremental memory allocation method for mixed workloads," Information Systems, vol. 21, pp. 369-386, 1996.

[21] 江藤正己, "列挙形式で引用された論文間の類似特性,"日本図書館情報学会, 三田図書館・情報学会合同研究大会発表要綱, pp. 9-12,2005.

[22] CiteSeer.PSU OAI,http://citeseer.ist.psu.edu/oai.html

[23] T. Couto, M. Cristo, M. A. Goncalves, P. Calado, N. Ziviani, E. Moura and B. Ribeiro-Neto, "A comparative study of citations and links in document classification," in JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, 2006, pp. 75-84.

[24] H. Liu, "MontyLingua," ver. 2.1, 2004.,http://web.media.mit.edu/~hugo/montylingua/