引用箇所の間隔に基づいた共引用関係にある論文間の類似度

江藤正己（慶應義塾大学大学院文学研究科）eto@slis.keio.ac.jp

Ⅰ．はじめに

A.共引用を用いた検索

代表的な論文検索手法の一つに共引用の関係を利用するものがある。共引用の関係とは，同一の論文によって引用された論文同士の関係を指し，この関係にある論文同士には類似性があるとされている。共引用の関係を利用した検索手法は，ISI や CiteSeer などの実用システムで利用されており，その有用性は高い。

この手法は,「一つの引用論文によって示される共引用の類似度が全て同じ」ということを仮定している。すなわち，論文の本文における引用のされ方とは無関係に，どのような共引用であっても，全て同一の類似度を示すことが前提となっている。

しかし，論文の性質を考えた場合，この仮定には問題がある。論文は，その著者が体系立てて記述したものであり，文脈的に近いものほどまとまって述べられる。そのため，本文の内容と共引用が示す類似度には密接な関係があると考えられる。たとえば,「はじめに」の部分で引用された論文と「方法」の部分で引用された論文間の類似度よりも，双方とも「方法」の部分内で引用された論文間の類似度の方が高いと考えられる。したがって,本文に依拠した形で,それぞれの引用がどのような関係にあるかをとらえた方が，共引用を有効に利用できると思われる。

これまで，本文に依拠した共引用が検索に利用されてこなかったのは，機械処理でそのようなことが難しかったためたためと推測される。しかし，現在は，機械可読形式の論文が増え，処理技術のレベルも向上してきており，本文への依拠が可能になってきている。事実，本文の内容に依拠した形で引用を利用しようとする試みとして，「引用文に含まれる語を検索に利用する研究」[1)]や「引用の役割を自動分類する研究」[2)]などがおこなわれている。

B.目的

筆者は本文の内容に依拠した共引用として，同じような文脈で引用された論文同士であるか否かに着目してきた。そして，特に，図１で示すような，同一箇所で複数の論文を並列に列挙した場合の引用（列挙形式の引用）に着目し，そのような引用によって生じた共引用は強い類似を示すことを明らかにしている[3)]。

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

第１図　列挙形式の引用

この結果をふまえ，本稿では引用箇所の間隔の長さによって，共引用関係にある論文同士がどれだけ同じ文脈で引用されているかを推定し，論文間の類似度を求める手法を考える。そして，そのために，引用箇所の間隔の長さを共引用関係にある論文間の類似度に置き換えることが可能か否かの検証を行う。引用箇所の間隔が長くなればなるほど，それに応じて引用箇所同士が文脈的に異なるようになるため，共引用関係にある論文間の類似度は徐々に低くなると予想される。

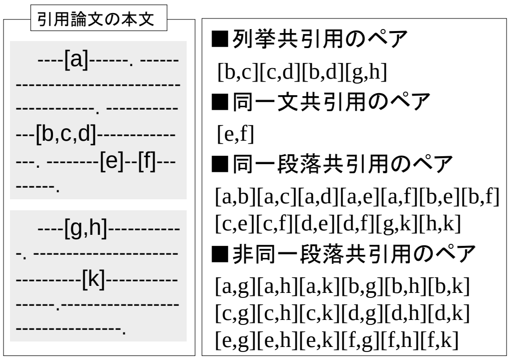
もし，上記したように間隔の長さを類似度に置き換えることができれば，これまで

同一なものとして粗く大まかに利用されてきた共引用が示す類似度を精緻に用いることができる。

Ⅱ．間隔の定義・仮説

引用箇所の間隔のとらえ方として，段落や文などの論文の構造に着目する。段落は一つの文脈的まとまりであるため，同一段落内の引用同士の間隔は，段落をまたぐ引用同士の間隔よりも短いと定義する。これと同様の定義が，同一文における引用や列挙形式の引用においてもできる。また，より小さな構造単位内での引用同士ほど，同じような文脈での引用と考えられるため，引用箇所の間隔が短いと定義する。

そこで，引用箇所の間隔に基づいて，「これまでの共引用」を「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 4 種類に分ける。4 種類は，この順に引用箇所の間隔が長くなる。論文間の類似はペアの形でみるので，第２図左側のような場合は，右側で示すように分類される。なお，同一文共引用としたペアは，同一段落共引用にはならず，列挙共引用としたペアは，同一文共引用，同一段落共引用にはならない。



第２図　共引用のペアの例

引用論文の本文において，それぞれの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，ある特定の共引用関係にある論文ペアに対して，複数の引用箇所の間隔が存在する場合がある。たとえば，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共引用の種類に分類した。

定義をおこなった引用箇所の間隔と論文間の類似度の関係として,「より引用箇所の間隔の長いほど，共引用関係にある論文間の類似度は低い」という仮説が立てられる。これは，列挙よりも同一文，同一文よりも同一段落，同一段落よりも非同一段落の方が，より低い類似度を示す共引用ということである。そこで，4 種類の共引用の類似度を測定，比較し，この仮説を検証する。

Ⅲ．類似度の測定・考察

A.測定対象の収集

検証に必要な引用論文集合，被引用論文集合を CiteSeer が公開しているデータセット CiteSeer Metadata を用いて作成する。このデータセットには論文の書誌事項，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 が含まれている。

１．引用論文集合の作成

データセット内でタイトルかディスクリプタに語「database」を含む論文を選び，全文をダウンロードした。プログラムで処理する都合上，その中で，引用記号に

① 大括弧に囲まれているもの
② 括弧内が数字とアルファベットの組み合わせであるもの

該当例…[CACS94] [Bon97b]

非該当例…1),(1),[1],(CACS94) ,[Bon]

を用いているものを引用論文集合とした。その数は，1,468 件となった。（ただし，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，実際に類似度の測定に用いることができたのは，1,055 件である)。

２．被引用論文集合の作成

データセットに含まれる引用関係の情報を用いて，被引用論文集合を作成した。その数は，4,592 件である。なお，詳しくは後述するが，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)を用いる都合上，被引用論文の全文をダウンロードした。

B.類似度を測定する方法

１．測定に用いる指標

類似度を測定する指標としては，実際に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。さまざまな指標で測定をおこない，検証の信頼性を高める。

「従来の共引用」を用いることは，4 種類の共引用ペアの類似度を「従来の共引用」の考え方に基づいて算出することを意味する。また，「書誌結合」と「従来の共引用」については，正規化処理をおこなった指標も用いる。

それぞれの類似度は以下のように算出した。ただし，求める類似度を S,類似度の測定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。

「tf*idf/cosine」

$$S = \frac{\vec{P_1}\cdot\vec{P_2}}{\left|\vec{P_1}\right|\times\left|\vec{P_2}\right|}$$

ここで，ベクトル$\vec{P}$ 中の各要素（語）に対する重みは以下のとおり。

Pに含まれる各語の重み=

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

ただし，全文をダウンロードできた全ての論文（15,713 件）を総文書とする。

「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」

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

「正規化共引用」

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

２．類似度の算出方法

まず，引用論文毎に 4 種類の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，４種類の共引用についての類似度をすべての引用論文で平均する。したがって，各共引用の類似度$\tilde{S}_t$は次式で求まる。ただし，n は総引用論文である。

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

第１表　測定に用いた共引用の数

	これまでの共引用		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	引用論文	共引用ペア
tf*idf/cosine	927	18740	772	14430	500	2670	264	813	317	827
正規化書誌結合	972	33662	810	26254	564	4620	333	1298	411	1490
正規化共引用	1055	44198	892	34596	650	6013	392	1719	465	1870
書誌結合	1055	44198	892	34596	650	6013	392	1719	465	1870
従来の共引用	1055	44198	892	34596	650	6013	392	1719	465	1870

C.測定結果・考察

前節で示した指標・方法を用いて，類似度を測定した。測定に用いた共引用の指標毎の数は第 1 表で示すとおりである。なお，本稿で提案した共引用の利点を明らかにするための比較対象として,「これまでの共引用」についても類似度の測定をおこなった。測定の結果を第 2 表に示す。

第 2 表　類似度の測定結果

	これまでの共引用	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	0.195	0.177	0.204	0.251	0.332
正規化書誌結合	0.109	0.094	0.129	0.163	0.232
正規化共引用	0.163	0.142	0.178	0.223	0.273
書誌結合	0.420	0.374	0.572	0.799	1.057
従来の共引用	6.856	5.676	7.820	12.449	17.028

この結果からわかるように，どの指標で測定した場合においても,「列挙」「同一文」「同一段落」「非同一段落」の順に，共引用関係にある論文間の類似度が低下する傾向がみられた。このことから，仮説は検証されたといえる。

また，測定結果から，４種類の共引用のそれぞれの間の類似度の異なりは小さくないこともわかる。これは，すなわち，それぞれの種類の共引用を使い分けることの有効性を示唆するものである。

そして，このことから「これまでの共引用」は，構造に基づいた間隔の観点からみた場合に差が現れるような共引用の関係を，全てひとくくりにしていたこともわかる。今回の結果は「従来の共引用」が粗くおおまかなものであったことを数値的に示したものといえる。

以上のことから，共引用を構造からみた引用箇所の間隔に基づいて利用することの意味は大きいことがわかった。間隔に基づいて共引用をより精緻にとらえることで，共引用の利用価値を高めることができると考えられる。

Ⅳ．まとめ・今後の課題

本稿では，論文の構造に基づいて，引用箇所の間隔を定義した。そして，その定義に基づき,「これまでの共引用」を「列挙」「同一文」「同一段落」「非同一段落」の四つの共引用に分けた。その後，それぞれの共引用関係にある論文間の類似度は，引用箇所の間隔に長さに応じて変化することを実験によって確認した。このことから，引用箇所の間隔を，共引用関係にある論文間の類似度に置き換えられることがわかった。

この結果をふまえると，引用箇所の間隔を考えることで，重みを持った共引用を論文検索に利用することができると予想される。そのための課題として，間隔に基づいた共引用のそれぞれに対して,どの程度の重みを設定していくかが挙げられる。この課題を考えるにあたり，複数の引用論文における評価（たとえば，ある引用論文では列挙共引用であり，別の引用論文では非同一段落であった共引用のペアの総合的な類似度評価）を検討していく必要があると思われる。

最終的には，重み付き共引用に基づいて検索を行うシステムを構築して,「従来の共引用」との検索性能を比較し，その有用性を検証する予定である。

引用文献

1)Shannon Bradshaw. Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes. Proceeding of the 7th European Conference on Research and Advanced Technology for Digital Libraries , p499-510(2003).

2)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,No.11,p.2640-2649(2001).

3)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,p.9-12(2005).