引用論文における引用箇所間の近さをとらえる尺度
Measuring Distance between Cited Papers in a Citing Paper

江藤正己
Masaki ETO

慶應義塾大学大学院文学研究科
Graduate School of Library and Information Science, Keio University
E-mail:eto@slis.keio.ac.jp

従来の共引用では，一つの引用論文における被引用論文間の類似度は全て同じであることを前提としている．しかし，引用論文における引用箇所間の意味的な近さをとらえることで被引用論文間の類似性の強弱を推測できると考えられる．そこで，近さをとらえる尺度として，「物理的距離（文字数）」「（引用箇所の）周辺語間の類似度」「論文構成からみた距離」を設定し，尺度としての適切さを検討する．

Co-citation, which is a typical similarity indicator among papers, has a premise that the degrees of similarity among cited papers are equal in one citing paper. But the similarity strength can be guessed from the distance between the places where cited papers are shown in a citing paper. In this paper, three distance measures(“physical distance”, “co-occurrence of citing words”, and “structural distance”) are introduced for the purpose and these measures are evaluated.

キーワード: 共引用, 論文検索, 類似検索
co-citation, document retrieval, similarity search

1　はじめに

1.1　共引用とは

論文間の類似度を算出する代表的な指標として共引用 [1] がある．この指標では対象となる二つの論文間の類似度は，一般的に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によって求められる．

ただし従来の共引用では，本文における引用のされ方の違いが考慮されない．つまり一つの引用論文においては, どの被引用論文間の類似度も全て同じことが前提になっている．しかし，そのような違いを考慮し被引用論文間の類似性の強弱を推測できれば，共引用の類似度が精密になり，類似論文検索の性能を向上させることができる．

被引用論文間の類似性の強弱を推測するために引用論文本文を解析する方法として，引用箇所間の意味的な近さをとらえることが挙げられる．引用論文の本文において引用箇所間が意味的に近ければ被引用論文間の類似性が強く，引用箇所間が意味的に遠ければ被引用論文間の類似性が弱いと判断できるためである．実際筆者のこれまでの研究により，引用箇所間の意味的な近さを利用することで，被引用論文間の類似性の強弱を推測できる示唆を得ている [2]．

1.2 目的

引用箇所間の意味的な近さをとらえる場合，その尺度としては様々なものが考えられる．尺度が引用箇所間の意味的な近さを適切にとらえていれば，尺度の値と被引用論文間の類似度の間には相関がみられるはずである．そこで，本稿では引用箇所間の意味的な近さをとらえる尺度を複数設定し，相関を調べることによってそれぞれの尺度の適切さを検討する．

2 近さをとらえるための尺度

引用箇所間の意味的な近さをとらえるために，以下の三つの尺度を設定した．

2.1 物理的距離

二つの引用箇所が物理的に短い距離内にあれば，両者は意味的に近いといえよう．そこで，引用箇所間の文字数を距離とする尺度を設定する．その際，論文の長さに対する補正をおこなうため，「文字数」を「論文の総文字数」で除算する正規化処理をおこなった．

2.2 周辺語間の類似度

引用論文中の引用箇所の前後 N 語を周辺語とする．測定対象の二つの引用箇所の周辺語に同じ語が多く含まれていれば，その引用箇所間は意味的に近いと考えられる．そこで，周辺語間の類似度を語の共出現数によって求め，それを距離とする尺度を設定する．具体的には，出現頻度で重み付けをおこない，コサイン係数によって類似度を求める．

2.3 論文構成からみた距離

論文には段落や文などの構造の単位がある．二つの引用箇所が同一の構造単位内にあれば，引用箇所間は意味的に近いと考えられる（たとえば，二つの引用箇所が同一段落内にある）．そして，同一段落よりも，同一文内にある方が意味的に近い（小さな構造単位内にあるほど意味的に近い）とみなすことができる．そこで論文の構造の単位を用いた尺度を設定する．本稿では，構造単位として「段落」「文」「列挙」の三つを用いる．「列挙」とは，[AA99,BB98,CC95] のような一つの引用箇所において複数の論文を並列列挙する形式の引用を指す．「論文構成からみた距離」の尺度は「列挙」「同一文」「同一段落」「非同一段落」の 4 種類から成る．

3 実験

引用論文を解析して 2 節で述べた尺度で引用箇所間の近さを測り，「各尺度の測定値」と「被引用論文間の類似度」との間の相関を調べる．両者の間に相関があれば，尺度として適切と判断できる．

3.1 実験データ

実験には，科学技術分野の主要なデータベースである *CiteSeer* が公開しているデータセット *CiteSeerMetadata* [3] を利用した．このデータセットには，論文の書誌情報，全文を入手するための URL，引用関係の情報が含まれている．

データセットに含まれる論文のうち，タイトルかディスクリプタに「database」が含まれる論文を選び全文のダウンロードを試みた．ダウンロード出来た論文のうち，引用記号が機械処理しやすい（アルファベットと数字から構成される）ものを選び引用論文集合とした．引用論文集合は 1,468 件となった．

3.2 被引用論文間の類似度の算出指標

被引用論文間の類似度は，論文間の類似度に算出によく用いられる「tf∗idf/cosine」「正規化書誌結合」「正規化共引用」の三つの指標を使用した.

3.3 集計方法

集計は次の二通りの方法でおこなった. (集計方法 1) 測定した「引用箇所間の近さ」とその「被引用論文間の類似度」との間の相関を求める．(集計方法 2) 最初に「引用箇所間の近さ」と「被引用論文間の類似度」のそれぞれを引用論文毎に平均し，その平均値同士の相関を求める.

3.4 算出・集計結果

算出・集計に用いたデータ数は，表 1 となった.（被引用論文の全文を入手できないなどの理由により）類似度の算出ができなかったものや二つ以上の引用箇所を引用論文から特定できなかったものは，対象から除外している.

表 1　算出・集計に用いたデータ数

	tf∗idf/cosine	正規化書誌結合	正規化共引用
集計方法 1	35,064	33,626	44,198
集計方法 2	1,014	973	1,055

3.4.1 物理的距離

「物理的距離」と「被引用論文間の類似度」の相関係数を求めた．相関係数が -1 に近ければ，引用箇所間の近さを適切にとらえていると判断される．結果を表 2 に示す.

表 2　物理的距離と被引用論文間の類似度との相関

	tf∗idf/cosine	正規化書誌結合	正規化共引用
集計方法 1	-0.075	-0.107	-0.067
集計方法 2	-0.124	-0.127	-0.105

3.4.2 周辺語間の類似度

引用箇所の周辺語の語数 N として，15・25・50・100・150・200 の 6 種類を設定して，各 N において「周辺語間の類似度」を求めた.「周辺語間の類似度」と「被引用論文間の類似度」の相関係数を求めた結果が表 3 である．この尺度では，相関係数が 1 に近いほど適切と判断される.

3.4.3 論文構成からみた距離

「論文構成からみた距離」と「被引用論文間の類似度」の相関をみた結果が表 4 である．この尺度は順序尺度であるため，それぞれの種類毎の類似度の平均値を示す.

4 考察

4.1 各尺度の適切さに関する検討

実験した尺度のうち，文字数からみた「物理的距離」に関してはほとんど相関がないといえる．したがって，この尺度は引用箇所間の意味的な近さを適切にとらえていないと判断される．次に，「周辺語間の類似度」では相関がみられた．このため，「周辺語間の類似度」は，引用箇所間の意味的な近さをとらえていると判断できる．三つ目の尺度である「論文構成からみた距離」では，構造単位が小さいものほど類似度が高くなっていることがわかり，引用箇所間の近さをとらえていることが分かった.

4.2 共引用ペアの分布による尺度の比較

相関がみられた「周辺語間の類似度」と「論文構成からみた距離」に関して，尺度の値毎の共引用ペアの分布状況を調べて比較した．なお，「周辺語間の類似度」では，「tf∗idf/cosine」の集計方法 1 で最も相関係数の高かった $N = 100$ の場合の分布状況を調べた．その結果が表 5, 表 6 である.

表 3　周辺語間の類似度と被引用論文間の類似度との相関

		15 語	25 語	50 語	100 語	150 語	200 語
集計方法 1	tf∗idf/cosine	0.259	0.290	0.321	0.352	0.347	0.337
	正規化書誌結合	0.228	0.258	0.267	0.278	0.290	0.292
	正規化共引用	0.214	0.256	0.278	0.307	0.298	0.295
集計方法 2	tf∗idf/cosine	0.352	0.384	0.382	0.364	0.381	0.382
	正規化書誌結合	0.187	0.217	0.238	0.232	0.204	0.206
	正規化共引用	0.209	0.241	0.245	0.230	0.258	0.264

表 4　「論文構成からみた距離」の尺度値毎の被引用論文間の類似度の平均値

		列挙	同一文	同一段落	非同一段落
集計方法 1	tf∗idf/cosine	0.292	0.234	0.185	0.153
	正規化書誌結合	0.200	0.156	0.109	0.075
	正規化共引用	0.248	0.204	0.155	0.119
集計方法 2	tf∗idf/cosine	0.349	0.257	0.214	0.178
	正規化書誌結合	0.234	0.163	0.129	0.093
	正規化共引用	0.274	0.222	0.179	0.142

表 5　「周辺語間の類似度」の尺度値における共引用ペアの分布

周辺語間の類似度	∼0.1	∼0.2	∼0.3	∼0.4	∼0.5	∼0.6	∼0.7	∼0.8	∼0.9	∼1.0
共引用ペアの数	8,854	14,622	9,850	4,624	1,991	832	474	447	1,079	1,425

表 6　「論文構成からみた距離」の尺度値における共引用ペアの分布

種類	非同一段落	同一段落	同一文	列挙
共引用ペアの数	34,596	6,013	1,719	1,870

5　まとめ

実験結果から,「周辺語間の類似度」と「論文構成からみた距離」が尺度として適切さを持つことが分かった．また，両者を共引用ペアの分布状況からみた場合，前者の方が共引用のペアを詳細に分類していることがわかった.

ただし，それぞれの尺度が類似論文検索の性能向上に寄与することができるか，あるいはどちらがより適切な尺度であるのか判断は，実際の検索システム上でさらに検証する必要がある．なお，必ずしも一方の尺度のみに限定する必要はなく，両者を組み合わせることも有用であろう.

参考文献

[1] Small,H. 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, no. 4, 1973, p. 265-269.

[2] 江藤正己. 引用箇所の間隔に基づいた共引用の検討. 電子情報通信学会第 18 回データ工学ワークショップ/第 5 回日本データベース学会年次大会 (DEWS2007), 広島,2007, L1-1.

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