共引用関係における引用論文の文脈情報を考慮した類似論文検索手法

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

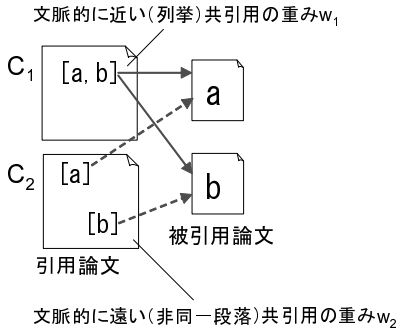
I. はじめに

類似論文検索の代表的な手法に共引用の関係を利用するものがある。この手法は，一つの論文から共に引用された論文間には類似性があるという発想に基づいている。そして，この手法では，一般に算出対象の論文が共引用された回数に基づいて，その類似度を算出する。

しかし，この手法は，共引用関係にある論文ペアの引用論文での扱われ方を考慮しないため，粗く大まかに類似度を算出していると指摘できる。発表者は，引用論文の文脈情報を利用することで，共引用関係を従来よりも精密に扱える尺度を提案してきた [1)2)]。この尺度は，引用論文における引用箇所間の意味的な近さをとらえ，従来は 2 値（共引用関係にある・ない）として扱われてきた共引用関係を多値（強い共引用関係〜弱い共引用関係）として扱うものである。

この尺度を実際に類似論文検索に組み込むことを考えた場合，第 1 図のような状況を対象として，類似度を算出するアルゴリズムを考えなければならない。ここでは論文 a，b が共引用関係にある論文ペアで，このペアの類似度を引用論文 C_1，C_2 からの二つの共引用情報を使って算出することが必要となる。そのためには，(1)「文脈的に近い箇所で共引用された場合の重み w_1」を「文脈的に遠い箇所で共引用された場合の重み w_2」の何倍とするか (2) 論文 a と論文 b の類似度を算出する際に，引用論文 C_1 から共引用された重み w_1 と引用論文 C_2 から共引用された重み w_2 をどのように集計するかが問題となる。

本発表では，(1)「重みの設定」，(2)「類似度の算出」のそれぞれに対して，方法を複数提案し，実際の論文データを用いて検索性能を評価する実験をおこなう。そしてその結果から，適切な類似論文検索手法について考察する。



第 1 図　多値の共引用を用いた類似度の算出

II. 重みの設定方法

発表者がこれまで提案してきた尺度に，論文の構成単位に基づくものがある [1)]。これは，論文を構成する単位のうち「段落」「文」「列挙」に着目し，共引用を「非同一段落」「同一段落」「同一文」「列挙」の 4 種類に分けたものである。発表者のこれまでの研究により，各共引用関係にある論文間の類似性はこの順に強くなることが明らかになっている。

ただし，この構成単位に基づく尺度は，順序尺度であり，検索システムへの組込みを考えた場合，各種類の共引用に，具体的な重みの数値を設定しなければならない。ここでは，以下の三つの重みの設定方法を挙げる。

A. 等間隔に重みを設定する方法

順序尺度である構成単位に基づく尺度を，差が等間隔な尺度とみなして，4 種類の共引用の重みを設定する方法である。この方法で設定をおこなった場合，たとえば，「非同一段落」の重みを 1 とすると，「同一段落」=2，「同一文」=3，「列挙」=4 のようになる。

B. 出現頻度に基づく方法

各種類の共引用の出現頻度に着目し，その逆数に比例した重みを設定する方法である。すなわち，出現頻度が多い種類の共引用ほど重みを弱くし，出現頻度が少ない種類のものほど重みを強くする方法である。

本発表では，発表者のこれまでの実験 [2] で得られている出現頻度の値を利用する。第 1 表の上段が，これまでの実験で得られている各種類の共引用の数で，1055 の引用論文を対象としたものである。出現頻度を用いて重みを設定するにおいて，第 1 表下段のように大きな構成単位の共引用がより小さなものを包含する（同一文共引用は，列挙共引用を包含するなど）ものとして，計算をおこなった。したがって，たとえば「同一文」の重みは，総共引用数（44,198）を同一文共引用数（3,589）で除算することにより求まる。

第 1 表　4 種類の共引用の出現頻度

	列挙	同一文	同一段落	非同一段落
排他	1,870	1,719	6,013	34,596
下位を包含	1,870	3,589	9,602	44,198

C. 類似度指標に基づく方法

各種類の共引用関係にある論文間の類似度を，従来から用いられてきた類似度指標で算出し，それに比例した重みを設定する方法である。たとえば tf∗idf/cosine によって，「列挙」の共引用関係にある論文間の類似度と「非同一段落」の共引用関係にある論文間の類似度をそれぞれ算出し，両者の類似度に基づいた重みを設定することになる。

これについても，発表者のこれまでの実験 [2] で得られている第 2 表の結果を用いる。第 2 表は，第 1 表の上段のデータを対象として，4 種類の共引用関係にある論文間の類似度を，3 種類の類似度指標（「tf∗idf/cosine」，「正規化書誌結合」，「正規化共引用」）と 2 種類の集計方法 (「マイクロな観点」，「マクロな観点」) を用いて算出した結果である。ここでは，第 2 表で 6 通り得られている類似度を平均した値に比例した重みを設定する。

第 2 表　4 種類の共引用の類似度

	列挙	同一文	同一段落	非同一段落
マイクロな観点に基づく集計				
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
マクロな観点に基づく集計				
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

以上，A 節から C 節で述べた三つの方法で得られた，重みの値を第 3 表に示す。なお，「非同一段落」が 1.000 となるよう標準化をおこなっている。

第 3 表　3 方法による 4 種類の共引用の重み

	列挙	同一文	同一段落	非同一段落
等間隔	4.000	3.000	2.000	1.000
出現頻度	23.635	12.315	4.603	1.000
類似度指標	2.101	1.626	1.278	1.000

III. 論文間の類似度の算出方法

従来の手法では，前述した共引用の各種類を考慮しないため，式 [1] のように全種類の共引用回数をそのまま加算して類似度とする。なおここで，P_1 と P_2 は類似度を算出する対象の論文のペアであり，$S(P_1, P_2)$ は P_1 と P_2 との類似度，t_i は共引用の種類 (i=1,...,4 で，t_1=「列挙」，t_2=「同一文」，t_3=「同一段落」，t_4=「非同一段落」) で，$Cocited(P_1, P_2)_{t_i}$ は種類 t_i で論文 P_1 と論文 P_2 が共引用された回数である。

$$S(P_1, P_2) = \sum_{i=1}^{4} Cocited(P_1, P_2)_{t_i} \quad [1]$$

引用論文の文脈情報を考慮して，論文間の類似度の算出する方法について，次の二つを考案

した。

A. 加算法

この方法は，従来手法に沿った方法で，文脈的に近い箇所で共引用される回数が多いものほど類似性が強いという発想による方法である。この方法では，種類毎の回数に重みを乗算した値を加算する。類似度は式 [2] で求まる。ここで，$S_m(P_1, P_2)$ は，重みの設定方法が m(m は，「等間隔」「出現頻度」「類似度指標」のいずれか) の場合の論文 P_1 と論文 P_2 の間の類似度，$W_{m_{t_i}}$ は種類が t_i で重みの設定方法が m の場合の重みである。

$$S_m(P_1, P_2) = \sum_{i=1}^{4} (Cocited(P_1, P_2)_{t_i} * W_{m_{t_i}}) \quad [2]$$

B. 平均法

この方法は，各引用論文からの引用の重みを平均する方法である。すなわち，第 1 図では，w_1 と w_2 の値が平均される。この方法は，算出対象の論文ペアの類似性の強さを，それらが総じてどのような文脈的な近さで共引用されているかによって，求めようとするものである。論文間の類似度は式 [3] で求められる。

$$S_m(P_1, P_2) = \frac{\sum_{i=1}^{4} (Cocited(P_1, P_2)_{t_i} * W_{m_{t_i}})}{\sum_{i=1}^{4} Cocited(P_1, P_2)_{t_i}} \quad [3]$$

IV. 検索性能を評価する実験

提案する 6 手法 (II 章で述べた三つの重みと III 章で述べた二つの算出方法の全組み合わせ) を，従来の手法 (式 [1]) と比較することによって，その性能を評価する実験をおこなった。なお，全ての手法が共引用関係を基にしており，出力される論文自体は同一であるため，手法間での検索性能の差は回答文書の出力順位によってのみ生じる。A 節で実験に用いたデータ，B 節で評価尺度，C 節で実験結果について述べる。

A. 論文データ

検索実験に用いる論文のデータとして，難波らが作成したデータセットを利用した [3]。難波らのデータセットは，1994 年から 1998 年に発表された自然言語処理や計算言語学分野の 330 論文を 10 の排他的なカテゴリに人手で分類したものである。このデータセットを用いた場合，検索質問は 330 論文中の各論文であり，質問論文と同一カテゴリの論文が適合（類似している)，非同一カテゴリの論文がを不適合（類似していない）となる検索実験をおこなうことになる。

ただし，難波らのデータセットには，330 論文を引用している論文の情報は含まれていない。そのため，*CiteSeer Metadata*[4] (*CiteSeer* の論文データセット）を用いて，難波らのデータセット中の論文を複数引用している論文の情報を得ることを試み，234 の引用論文の情報を得た。この 234 の論文は，難波らのデータセット中の 135 論文を共引用している。234 の論文の本文を解析した結果，135 の論文に対するのべ 552 回の共引用（「列挙」24 回，「同一文」21 回，「同一段落」54 回，「非同一段落」453 回）を特定した。

135 論文のそれぞれが質問論文となるが「質問論文と共引用関係にある論文に同一カテゴリがなく，適合論文がない」場合や，「質問論文と共引用関係にある論文が 1 つしかなく出力順位が生じない」場合は，性能の比較評価に沿わないと考え除外した。結果，質問論文は 88 となった。

B. 評価尺度

評価尺度としては，MRR(Mean Reciprocal Rank) を用いた。この尺度は，質問論文に対する順位付き回答論文のリストのうち，適合論文の順位の逆数に基づいて算出される。なお，回答リスト中に複数の適合論文が含まれていた場合は，各順位の逆数を合計し，総適合論文数で除算することにより平均を求めた。また，同一の順位がある場合は，それらの平均順位を当

該回答論文群の順位とした（たとえば，3 位の回答論文が二つあった場合は，両者は 3.5 位となる）。本発表における MRR は式 [4] のように定義され，質問論文毎にこの値が求まる。ここで，R_k は適合論文の順位の値 (k=1,...,N, N は適合論文数) である。

$$MRR = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{R_k} \qquad [4]$$

C. 実験結果

88 質問論文を用いて，各手法で質問論文と同一カテゴリの論文を検索（順位付きの出力）し，MRR で評価をおこなった。そして，従来手法（式 [1]）の MRR と提案する 6 つの手法のそれぞれの MRR とを，質問毎に比較しその優劣を判定した。その結果が第 4 表である。第 4 表において，優は提案手法が従来手法の MRR を上回っていた質問数，劣はその逆，試行数は両者の合計（88 の質問のうち MRR が同値になったものを除外した数）である。また，従来手法と提案手法の優劣数の差に対する符号検定を試みた結果として，その p 値を示す。

第 4 表　従来の共引用の手法との比較

重み	算出方法	優	劣	試行数	p 値
等間隔	加算	28	19	47	0.243
出現頻度	加算	27	17	44	0.174
類似度指標	加算	30	13	43	0.014
等間隔	平均	29	28	57	1.000
出現頻度	平均	29	28	57	1.000
類似度指標	平均	29	28	57	1.000

実験の結果，提案 6 手法の全てで従来手法を上回る結果を得た。このうち特に，重みを「類似度指標」，算出方法に「加算法」を用いた場合の結果が最も良く，かつ有意水準 5% で従来手法と差があることがわかった。

V. 考察

実験の結果より，引用論文の文脈情報を考慮することで，従来の共引用の手法よりも良い検索結果を得られることが明らかになった。そして，「類似度指標」に基づいて重みを設定する方法と，「加算法」で論文間の類似度を算出する方法を組み合わせることで最も良い結果を示すことが分かった。

なお，平均法が従来手法と比較して，あまり差が生じなかった理由の一つとしては，非同一段落の共引用の割合が高く，同順位になる（類似度が 1.000 となる）ものが多かったためと考えられる。同順位の場合は，共引用回数に基づいて順位に差をつけるなどの手法の修正が必要と予想される。

ただし，加算法の場合でも，算出対象の論文が全て「非同一段落」の共引用されていた場合は，結果的に提案手法と従来手法と間に差が生じない。したがって，「非同一段落」共引用をさらに細かくとらえる等の方法を検討しなければならないと思われる。これについては，発表者が別の文脈情報のとらえ方として提案している [2]，引用箇所周辺の語を利用することなどが考えられる。

また，従来の共引用を用いた検索手法には，共引用関係にある論文それぞれの被引用数で正規化した手法が存在する。被引用数の情報を提案手法に組み込み，従来の正規化による手法との比較をすることも今後の課題である。

引用文献

1) 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, 広島, 2007-02-28/03-02, L1-1.

2) 江藤正己. 引用論文における引用箇所間の近さをとらえる尺度. 情報知識学会第 15 回年次大会, 国文学研究資料館, 2007-05-25/26, 情報知識学会誌, 2007, vol.17, No.2, p.65-68.

3) 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. 2001, vol. 42, no. 11, p. 2640-2649.

4) CiteSeer.PSU OAI, `http://citeseer.ist.psu.edu/oai.html`