文脈情報に基づく共引用関係を利用した文書検索手法の特徴

江藤 正己 †

† 亜細亜大学（非常勤）

共引用関係を利用した文書検索手法は，共引用関係を2値的に扱うため，文書得点の算出が粗く，手法の発展性に限界がある．また，語に基づく検索では得られない適合文書を検索できる可能性が示唆されているものの，その検証はあまりなされていない．そこで，本稿では，文脈情報に基づき共引用関係を精密に扱う検索手法を提案し，テストコレクションによる検索実験から，提案手法の特徴の検証とその評価をおこなった．その結果，提案手法の特徴として，(1) 語に基づく検索では得られない適合文書を検索できること (2) 従来手法より検索結果の順位を適切に出力できること (3) 共引用回数が少ない適合文書を検索結果の上位に順位付られることが明らかになった．

Document Retrieval Method Using Context Based Co-citation Relationship

Masaki ETO †

† Asia University(part-time)

In this paper the author proposes a sophisticated document retrieval method using context based co-citation relationship. To evaluate the effectiveness of the proposed method, two experiments were conducted. The first experiment is to compare documents retrieved by word with the ones retrieved by co-citation relationship. The second one is to compare the proposed method with the traditional method by using binary co-citation relationship based on two typical metrics (MAP and nDCG), and by analyzing documents ranked top 10. The experiments showed that the proposed method will (1) retrieve relevant documents which cannot be retrieved by using word, (2) rank retrieved relevant documents more adequately, and (3) rank relevant documents highly, even though the number of the frequencies of their co-citation is few.

1 共引用関係を利用した検索

1.1 検索手法の概要

文書検索の代表的な手法の一つに共引用関係を利用するものがある[1]．共引用関係とは，同一の文書から引用された二つの文書間の関係のことを指す．たとえば，図1のような引用関係があった場合，文書Bと文書Cが共引用関係にある．この手法では，「共引用関係にある文書同士には何らかの関連がある」という発想により，文書Bを検索キーに文書Cを検索する（あるいはその逆）ことがおこなわれる．

共引用関係を利用した検索手法では，一般的に二つの文書が共引用された回数に基づいて，文書得点（検索キーとなる文書と検索対象の文書との類似度）を算出する．すなわち，図1において，文書Bを検索キーとした場合の文書Cの文書得点は，

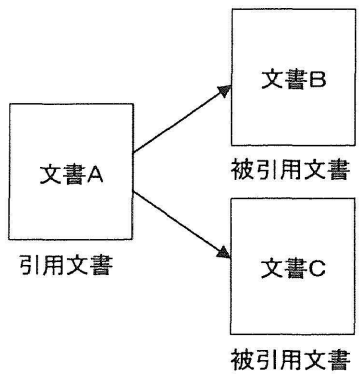

図1 共引用関係

文書Bと文書Cをともに引用している文書が多ければ高く，少なければ低くなる．

共引用関係を利用した検索が有効であることは既往研究において報告されており[2][3]，アルゴリ

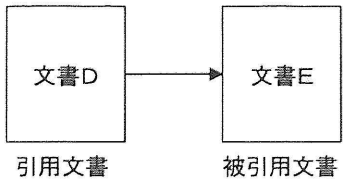

図 2　直接引用関係

ズムの一部で共引用関を利用する検索手法も提案されている[4) 5)]．さらに，*CiteSeer* のような実用のデータベースにおいても共引用関係を利用した検索機能が提供されている[6)]など共引用関係を利用した検索には一定の評価があると判断できる．

1.2　語に基づく検索では得られない適合文書を検索できる可能性

引用関係を利用した検索は，異なるアクセスポイントを用いるため，語に基づく検索では得られないような適合文書を検索できる可能性がある．たとえば，Pao は，引用関係のうち，直接引用関係（図 2）を利用した検索をとりあげ，語に基づく検索との比較をおこなっている．そして，語に基づく検索をおこなった後に，直接引用関係を利用した検索をおこなうことで，適合文書を 24%程度多く検索可能なこと報告している[7)]．

このことから，共引用関係を利用した検索も語に基づく検索では得られない適合文書を検索できる可能性を持つと考えられる．ただし，クラスタリング研究において共引用関係と語を比較した例[8)]はあるものの，Pao のように情報検索の観点から直接的に比較するような研究はみられず，その検証はあまりなされていない．

1.3　共引用関係を 2 値的に利用することによる発展性の限界

従来の共引用関係を利用する手法は，一つの引用文書で示される共引用関係は全て同じ強さであると仮定し，それに基づいて文書得点の算出をおこなう．すなわち，従来の手法では，共引用関係は，「共引用関係にある/共引用関係にない」の 2 値情報としてのみ扱われる．

しかし，共引用関係を 2 値情報として処理することは，粗い扱いであるといえる．たとえば，図 1 の文書 A において，文書 B と文書 C が同一箇所で特に関連のある文書同士として引用されていた場合と，それぞれ全く関連のない箇所で引用されていた場合とでは，前者の方がより共引用関係が強いと考えられる．このことは，共引用関係を利用する手法の提案者である Small 自身が「共引用関係が成立するのは，引用文書の著者が被引用文書同士を関連づけさせようと試みた時のみに成立する」と述べていること[9)]，及び江藤が「同一箇所で引用された文書間の類似度と同一ではない箇所で引用された文書間の類似度を比較し，前者の類似度の方が高いことを実験から数値的に示している」[10)]ことからも明らかである．

したがって，共引用関係を 2 値的に扱うことは文書得点の算出が粗いと指摘でき，より精密な文書得点の算出を望むにおいて，手法の発展性に限界があるといえる．

1.4　目的

1.3 で述べた限界を克服するためには，引用文書の文脈情報を利用して共引用関係を精密に扱う検索手法の開発が必要である．共引用関係を精密に扱うことで，共引用回数が少なくても類似性が高いペアの文書得点を高く，共引用回数が多くても類似性が弱いペアの文書得点を低く算出することができる．

本稿の目的は，文脈情報に基づく共引用関係を利用した文書検索手法を提案し，1.2 で述べたことも含めて提案手法の特徴を分析することである．

以下，本稿では，まず 2 章で提案手法を説明をする．3 章で，手法の特徴を分析するためのテストコレクションを作成し，4 章でそれを用いた検索実験をおこなう．5 章で本稿をまとめ，6 章において今後の課題を述べる．

2　提案手法

2.1　引用関係と文脈情報の組み合わせる試み

引用関係を利用する検索手法の開発に関する近年の動きとして，引用関係と文脈情報を組み合わせるものがある．これは，論文の電子化が進んだことにより，機械処理で引用文書の文脈情報をある程度扱えるようになったことが要因の一つと思われる．

引用関係と文脈情報を組み合わせものとして，たとえば，引用文章（引用箇所の周辺の文）に含まれる語を被引用文書の索引語として追加し，文書検索の性能向上をめざす研究がある．初期には，引用文章に含まれる語を単純に被引用論文の索引語

とすることをO'Connor[11]がおこなっている．さらに，Bradshawは様々な引用文書における引用文章に共通して含まれる語の重みを強くすることを試み[12]，Ritchieらは，引用文章に含まれる語の中でも被引用文書に関係のある語と関係のない語があり，両者を区別することを提案している[13]．

また，引用文章に含まれる語句を手がかりに引用をカテゴリに分類することも，Garzoneら[14]，難波ら[15]，Teufelら[16]らによって研究されている．これらの研究は，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ものである．特に難波らは，自動分類されたカテゴリに基づいて書誌結合関係（引用関係のとらえ方の一つ）を利用した検索手法の拡張をおこなっている．

2.2 文脈情報に基づく共引用関係の尺度

共引用関係と文脈情報の組み合わせ方として，江藤は，構成単位に基づく共引用関係の尺度を提案している[17]．この尺度は，段落や文などの構成単位から共引用関係の文脈情報をとらえるもので，より小さい構成単位において共引用関係にある文書同士ほど関係の強い共引用と考えるものである．江藤は，この尺度が実際に共引用関係を精密にとらえられることを実験により確認している．

構成単位に基づく共引用関係の尺度を具体的に示したものが，図3である．この尺度は，これまでの単一の共引用関係を，論文の構成単位に応じて，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種類にとらえなおすものである（列挙共引用とは，図3で示したように，同一箇所で並列的に引用された共引用を指す）．この尺度を用いることで，たとえば，非同一段落で共引用されていれば，弱い共引用関係と判断し文書得点算出の際に重みを弱くするというような精密な文書得点の計算が可能となる．

2.3 文脈情報に基づく共引用関係を利用した文書検索手法

本稿では，構成単位に基づく共引用関係の尺度を使い，文脈情報に基づく共引用関係を利用した文書検索手法の提案をおこなう．

提案手法の文書得点の算出に関わる要素は，図4で示すものが基本となる．ここでは，文書a，b

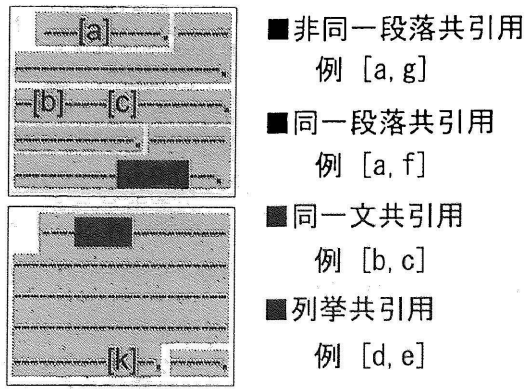

図3 構成単位に基づいた共引用関係の尺度

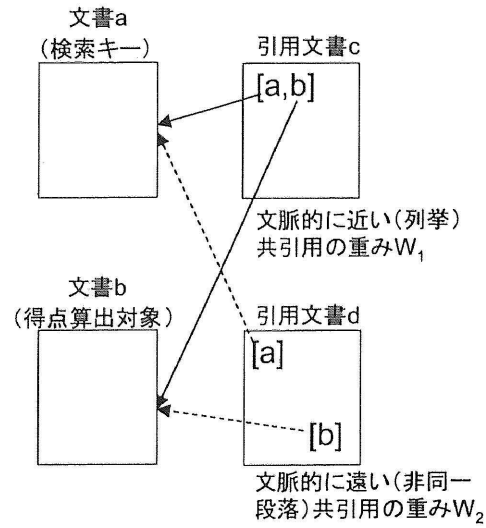

図4 提案手法の文書得点算出に関わる要素

が共引用関係にあるペアで，文書aを検索キーとして，文書bの文書得点を算出をおこなう．その際に利用するのが，引用文書cにおける「文脈的に近い箇所で共引用された場合の重みW_1」と引用文書dにおける「文脈的に遠い箇所で共引用された場合の重みW_2」である．

提案手法は，(1)W_1をW_2より強く設定し，(2)W_1とW_2を集計することで文書得点の算出をおこなう．重みの設定と集計方法にはいくつかの方法が考えられるが[18]，今回は，分析をし易くするため，最も単純なものを採用する．以下，重みと集計方法について説明する．

表 1　重みの具体値

種類	w_{t_1} 非同一段落	w_{t_2} 同一段落	w_{t_3} 同一文	w_{t_4} 列挙
重み 1	1	2	3	4
重み 2	1	4	9	16

2.3.1　重みの具体値

重みは，構成単位に基づく尺度を等間隔にみなす場合と間隔を大きくとる場合の 2 種類のものを設定した．等間隔にみなす場合では，非同一段落の重みの具体値を 1 とし，同一段落を 2，同一文を 3，列挙を 4 とした (重み 1)．さらに，尺度の差を大きくとる場合として，重み 1 を二乗した値を設定した (重み 2)．共引用の種類を t_i とした場合 (i=1,...,4 で，t_1=「非同一段落」，t_2=「同一段落」，t_3=「同一文」，t_4=「列挙」)，各種類ごとの重み w_{t_i} は表 1 のようになる．

2.3.2　集計方法

重みを集計して文書得点を算出する方法については，従来手法に準じた加算法を用いる．文書得点は，各種類の共引用回数にそれぞれの重みを乗じて，それを加算したものとなる．したがって，f_{t_i} を種類 t_i の共引用回数とした場合，文書得点 ($Score$) は式 (1) の形で求まる．

$$Score = \sum_{i=1}^{4} f_{t_i} \times w_{t_i} \qquad (1)$$

なお，従来手法も式 (1) で説明することができ，式 (1) における w_{t_1}〜w_{t_4} を全て 1 とすることで従来手法の文書得点算出式となる．

以下，本稿では，式 (1) における重み w に表 1 中の重み 1 を設定した場合を提案手法 (軽)，重み 2 を設定した場合を提案手法 (重) と表す．

2.4　提案手法の意義

提案手法と従来手法は，双方とも共引用関係を利用するものであるため，検索される文書群自体は同じものとなる．提案手法の意義は，文脈情報を導入したことにより，検索結果の順位をより適切に出力することである．

提案手法の予想される成果は，全体的に適切な検索結果の順位が出力されることであるが，特に共引用回数の少ない適合文書を上位に順位付けられる可能性がある．

従来の検索手法では，共引用回数に基づいて文書得点を算出する．この方法は，検索ノイズとなる文書（共引用関係が成立するが不適合であるような文書）を下位に順位付けるために，共引用された回数を用いるといえる．すなわち，関連のない文書が同一文書で引用されることは偶然的にしか起きないことを想定し，共引用される回数に着目することで，検索ノイズを除外しようとするものである．

しかし，従来手法におけるノイズの除外方法では，共引用回数が少ない適合文書までも下位に順位付けてしまう問題がある．すなわち，検索キーとの類似性が高く将来的には共引用回数が増える可能性がある文書であっても，発行されてから新しい，多くの人に知られていない等の理由により検索がおこなわれた時点で共引用回数が少なければ，文書得点が高くならず，検索結果の下位に順位付けられてしまう．

一方，提案手法は共引用関係が強ければ，共引用回数が少なくとも検索結果の上位に適合文書を出力することができる．したがって，提案手法の意義として，従来手法よりも共引用回数が少ない適合文書に対して特に効果的に機能することが予想される．

3　テストコレクションの作成

「文書集合の中から課題文書と関連した文書を検索すること」を検索課題とするテストコレクションを作成する．作成には，*CiteSeer* が公開しているデータセット *CiteSeer Metadata* [19] 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，(データセット内の論文との) 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．

作成は次の手順でおこなった．まず，*CiteSeer* のデータセットの中から，タイトルかディスクリプタに “database” を含むものを選び，URL 情報を用いて，該当文書群の全文のダウンロードを試みた．その結果，13,551 件の文書集合を得た．これをベース文書集合とする．

次に，ベース文書集合のうち，以下の条件を満たすものの中から，無作為に 30 件の文書を抽出し，課題文書とした．

表 2　テストコレクション中の適合・不適合文書

課題数	判定対象文書数	適合 (S,A)	不適合 (B,C)
30	1,766	370	1,396

- 課題文書と共引用関係にある異なり文書がベース文書集合中に 30 以上あるもの
- 5 ページ以上 25 ページ以下であるもの

課題文書と共引用関係にある異なり文書群が，共引用関係を利用した検索によって求められる検索結果文書群であり，これらが適合判定の対象となる．30 課題における適合判定の対象となる文書（検索結果文書）の総数は 1,766 件であり，平均が 58.9 件，最小 30 件，最大 148 件である．課題文書と判定対象文書を共に引用する文書は全部で 477 件であり，引用文書 477 件で示される課題文書と判定対象文書との共引用関係はのべ 3,831 件であった．引用文書の本文を解析し，構成単位に基づく尺度で共引用関係の分類を試みた結果，列挙共引用が 284 件，同一文共引用が 189 件，同一段落共引用が 362 件，非同一段落共引用が 2,996 件となった．なお，プログラムで本文中の引用箇所の解析に失敗したものが 516 件（約 13%）あったが，これについては最も出現する確率が高い非同一段落共引用とした．

適合判定は，課題文書とその課題に対する判定対象文書群を判定者に提示し，対象文書ごとに課題文書との適合度を 4 段階 (S，A，B，C) で判定する形でおこなった．なお，判定対象文書群は，判定者に無作為に並べ替えた状態で提示している．

判定者は，全部で 21 名であり，内訳はデータベースを専門とする理工学部の研究室に所属する学生 18 名 (学部 4 年生 4 名，修士課程 9 名，博士課程 5 名)，図書館・情報学を専攻する学生（博士課程）2 名，情報工学の修士号取得者 1 名である．30 課題を 21 人で分担し，適合判定をおこなった．

適合判定の結果は，S が 112 件，A が 258 件，B が 666 件，C が 730 件となった．以下，本稿では，基本的に S と A を適合，B と C を不適合として分析をおこなう．テストコレクションの内訳を表 2 にまとめる．

4　提案手法の特徴の分析

作成したテストコレクションを用いて，検索実験をおこない，提案手法の特徴を分析する．分析は，(1) 語に基づく検索との比較 (2) 従来手法との全体的な比較 (3) 共引用回数が少ない適合文書に対する効果の検証，からなる．

4.1　語に基づく検索との比較

この実験はベース文書集合 (13,551 件) を対象として，語に基づく検索をおこなった場合，どの程度の適合文書の検索漏れがあるかをみる実験である．テストコレクション中の適合文書は，共引用関係を利用して検索した適合文書である．そのため，この実験で漏れた適合文書は，共引用関係を利用した検索では得ることができるが，語に基づく検索では得ることのできない適合文書といえる．

語に基づく検索は，情報検索用のツールキットである The Lemur Toolkit[20] を使用し，代表的なアルゴリズムである Okapi BM25 と言語モデル (Kullback-Leibler divergence) による全文検索を用いた．また，検索キーとして，課題文書の「タイトル」，「タイトルと抄録」の二つを設定し，語に基づく検索で検索可能とする（人間が実際にみることのできる）文書数の基準として，検索結果の上位 100 件，200 件，300 件の 3 段階を設定した．

30 の検索課題を用いて，12 通りの検索手法（2 種類の検索キー，2 つのアルゴリズム，3 段階の基準）で，検索実験をおこなった．30 課題の平均漏れ率，平均漏れ件数，及び 30 課題の総漏れ数を検索キーごとにまとめたものが，表 3，表 4 である．なお，課題文書の一つに抄録が無く対象から除外したため，表 4 では適合文書数が少なくなっている．

実験結果から，語に基づく検索で検索可能な文書を上位 100 件とした場合，最大 95%もの適合文書が漏れることが分かった．また，検索結果の上位 300 件にまで広げた場合でも，半数以上の適合文書が検索から漏れる結果となった．そして，平均漏れ件数をみた場合，6.6 件以上もの新たな適合文書を検索できることが確認できた．したがって，課題文書と共引用関係にある文書が 30 以上あるという条件で作成したテストコレクションにおいてではあるものの，共引用関係で得られる適合文書の中には，語に基づく検索では得ることのできない適合文書が一定数存在することが確認された．

なお，この結果は，提案手法だけでなく，従来

表 3　検索キーをタイトルとした場合

検索手法	平均漏れ率	平均漏れ件数	総漏れ件数/全適合文書数
okapi100	90%	11.1	332/370
okapi200	82%	10.1	302/370
okapi300	78%	9.7	290/370
kl100	75%	9.2	276/370
kl200	63%	7.7	232/370
kl300	54%	6.7	201/370

表 4　検索キーをタイトルと抄録とした場合

検索手法	平均漏れ率	平均漏れ件数	総漏れ件数/全適合文書数
okapi100	95%	11.4	330/349
okapi200	90%	10.9	315/349
okapi300	88%	10.6	306/349
kl100	70%	8.4	243/349
kl200	60%	7.2	210/349
kl300	55%	6.6	192/349

手法も含めた共引用関係を利用した検索手法に共通する特徴である．ただし，テストコレクションを用いた検索実験により，共引用関係を利用した検索と語に基づく検索とを比較した例はこれまでにみられず，本稿によって初めて明らかになった点といえる．

4.2　従来手法との比較

文脈情報を導入したことの効果を分析する検索実験をおこなった．提案手法が従来手法よりも，適合文書を検索結果の上位に，不適合文書を下位に出力することができれば，文脈情報を導入を導入することに意義があるといえる．

評価指標としては，MAP(Mean Average Precision) と nDCG(normarized Discounted Cumulative Gain) の二つを使用する．従来手法，提案手法ともに同順位が出現するため，同順位を考慮した MAP，nDCG を用いた [21]．なお，nDCG は，多段階の適合判定をそのまま評価できるため，nDCG に限り，4 段階の適合判定の結果をそのまま用いた．各得点については，S を 7 点 (2^3-1)，A を 3 点 (2^2-1)，B を 1 点 (2^1-1)，C を 0 点 (2^0-1) と設定した．

30 の検索課題を用いて，従来手法と提案手法 (軽)，提案手法 (重) のそれぞれで検索 (検索結果の順位付け) をおこない，MAP と nDCG で評価をおこなった．その結果をまとめたものが表 5 である．

表 5　従来手法との比較

	従来手法	提案手法 (軽)	提案手法 (重)
MAP	0.375	0.397^{*}	0.401
nDCG	0.696	0.721^{**}	0.733^{**}

$*p =< 0.1, **p =< 0.05$

MAP，nDCG ともに，提案手法の方が従来手法よりも高い評価値を得られた．従来手法と提案手法 (軽)，従来手法と提案手法 (重) の間で，それぞれ平均値の差をみる検定をおこなったが，MAP における提案手法 (重) をのぞいて，この差は統計的に有意であることがわかった．

したがって，提案手法の方が従来手法よりも検索結果の順位を適切に出力できることが分かった．すなわち，共引用関係を利用した検索において，文脈情報を導入することの有用性が示されたといえる．

4.3　共引用回数が少ない適合文書に対する効果

提案手法が従来手法に比べて，特に有効に機能することが予想される共引用回数が少ない適合文書についての分析をおこなった．

テストコレクションにおける課題文書と適合判定対象文書の異なり共引用ペア (1,766 件) を共引用回数別にみたものが表 6，そのうちの全適合文書 (370 件) を共引用回数別の割合で示したものが図 5 である．

表 6 で示されるように，共引用回数が 1 回しかなく従来手法では最下位にランク付けされるペアであっても，20%程度の適合文書を含むことが分かる．そして，図 5 からは，適合文書全体のうち，共引用回数が 1 回のペアの適合文書だけで，適合文書全体の半数を占めていることが示された．これらの結果から，共引用回数が少ないペアは，適合文書が含まれる割合は少ないが，総数自体が多いため，含まれる適合文書の絶対数は多いとみることができる．

従来手法は，2.4 で述べたように適合文書を検索結果の下位にしか順位付けられない．しかし，提案手法では，共引用回数が少なくとも，引用文書において強い共引用関係が示されていれば検索結果の上位に適合文書を出力することが可能である．

そこで，実際に提案手法が共引用回数が少ない適合文書に対して有効に機能しているか否かを検証した．検証は，従来手法と提案手法のそれぞれ

表 6　テストコレクションの共引用回数別分析

共引用回数	共引用ペア数	適合ペア数	割合
1	1112	193	17%
2	308	62	20%
3	130	45	35%
4	63	18	29%
5	34	13	38%
6 以上	119	39	33%

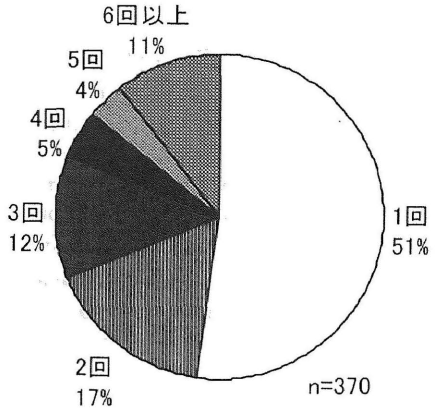

図 5　適合文書の共引用回数別内訳

検索結果上位 10 位以内にある適合文書を共引用回数別に分析することによりおこなった．なお，ここでは 4.2 で評価値の高かった提案手法 (重) を提案手法として用いた．

30 課題で検索結果の上位 10 位以内となった全ての適合文書を共引用回数別にみたものが，表 7 である．なお，同順位があった場合は，その平均順位を当該文書の順位とした．従来手法では共引用回数が 1 回の適合文書は検索結果の上位に全く出現していないが，提案手法では上位に出現することが確認できる．提案手法で検索された共引用回数が 1 回の適合文書の文書得点は，16 点が 9 件，9 点が 2 件，4 点が 3 件であり，1 点のものは存在しなかった．

したがって，従来手法では検索結果の下位に順位付けられてしまう共引用回数が少ない適合文書を，提案手法では上位に出力できることを実際に確認できた．これは，提案手法が共引用関係の文脈情報を導入した効果といえ，提案手法の特徴を示すものである．

表 7　10 位以内の適合文書の共引用回数

共引用回数	従来手法	提案手法 (重)
1	0	14
2	21	20
3	25	26
4	8	10
5	9	8
6 以上	32	30

5　まとめと提案手法の特徴

本稿では，以下のことをおこなった．(1) 引用文書の文脈情報に基づく共引用関係を利用した検索手法を提案した．(2) テストコレクションを作成し，提案手法の特徴を分析する検索実験をおこなった．(3) 実験の結果から提案手法の特徴として，(特徴 1) 語に基づく検索では得られない適合文書を検索できること，(特徴 2) 従来手法より検索結果の順位を適切に出力できること，(特徴 3) 共引用回数が少ない適合文書を検索結果の上位に順位付けられること，を明らかにした．

6　今後の課題

今後の課題として，まず，明らかにした特徴 1〜特徴 3 を定性的な面から深く評価することが挙げられる．特徴をよく示す事例や特徴とは異なる事例になどについて検証する必要がある．

また，4.2 の実験では，尺度の差を大きくとる重みを設定した方が高い評価値となったが，MAP に関しては従来手法との間に有意な差が生じなかった（表 5）．重みに関しては，2.3.1 で設定した具体値の根拠が弱いこともあり，その最適な設定方法を検討していかなければならない．

さらに，特徴 1〜3 をより広い観点から検証することも重要である．特徴 1 については，今回は共引用関係を利用した検索と語に基づく検索の比較のみをおこなった．直接引用関係利用した検索との比較や，語に基づく検索と直接引用関係を利用した検索とを併用した場合との比較などもおこない，共引用関係を利用した手法で特徴的に検索される適合文書についての分析を深める必要がある．

特徴 2 については，提案手法の持つ意義を情報検索手法全体の中で論じる必要があると思われる．提案手法で文書得点が大幅に上昇した適合文書が，他の検索手法（語に基づく検索，直接引用関係を

利用した検索等）においてはどのような意味を持つのか吟味しなければならない．

特徴3として示した提案手法が効果的な「共引用回数が少ない適合文書」は，発行されてから新しいものや多くの人に知られていないものである可能性がある．このような適合文書は検索者にとって新奇な適合文書となりやすいとも予想されるため，新奇性の面からの提案手法評価も今後の課題といえる．

なお，本稿の実験で明らかになった提案手法の特徴は，一つのテストコレクションにおける評価という制限がある．他のテストコレクションを用意し，特徴1〜特徴3についてのさらなる確認作業もおこなうべきと考える．

その他，今回扱えなかったものとして，従来手法の拡張手法で，被引用数によって共引用回数を正規化し，文書得点を算出するものがある．被引用数による正規化の手法は，提案手法にも取り込むことが可能なため，被引用数の考慮という範疇においても提案手法の効果を分析する必要もあろう．

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, pp. 265–269 (1973).
2) Chapman, J. and Subramanyam, K.: Cocitation search strategy, *Proceedings of the 2nd National Online Meeting*, pp. 97–102 (1981).
3) Knapp, S. D.: Cocitation Searching: Some Useful Strategies, *Online*, Vol. 8, No. 4, pp. 43–48 (1984).
4) Bichteler, J. and Eaton III, E. A.: The combined use of bibliographic coupling and cocitation for document retrieval, *Journal of the American Society for Information Science*, Vol. 31, No. 4, pp. 278–282 (1980).
5) Badran, O. M.: An alternative search strategy to improve information retrieval, *Proceedings of the 47th ASIS Annual Meeting*, pp. 137–140 (1984).
6) CiteSeer, http://citeseer.ist.psu.edu/
7) Pao, M. L.: Term and citation retrieval: A field study, *Information Processing and Management*, vol. 29, No. 1, pp. 95–112 (1993).
8) Braam, R. R., Moed, H. F., and Van Raan, A. F. J.: Mapping of science by combined co-citation and word analysis. I. structural aspects, *Journal of the American Society for Information Science*, vol. 42, No. 4, pp. 233–251 (1991).
9) Small, H: Citation context analysis, *Progress in Communication Sciences III*, pp. 287–310 (1982).
10) 江藤正己: 引用箇所間の意味的な近さに基づく共引用の多値化: 列挙形式の引用を例として, *Library and Information Science*, No. 58, pp. 49–67 (2007).
11) O'Connor, J.: Citing statements: computer recognition and use to improve retrieval, *Information Processing and Management*, Vol. 18, No. 3, pp. 125–131 (1982).
12) Bradshaw, S.: Reference directed indexing: redeeming relevance for subject search in citation indexes., *ECDL*, pp. 499–510 (2003).
13) Ritchie, A., Teufel, S. and Robertson, S.: How to find better index terms through citations, *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, Sydney, Australia, Association for Computational Linguistics, pp. 25–32 (2006).
14) Garzone, M. and Mercer, R. E.: Towards an automated citation classifier, *AI '00: Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, London, UK, Springer-Verlag, pp. 337–346 (2000).
15) 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, pp. 2640–2649 (2001).
16) Teufel, S., Siddharthan, A. and Tidhar, D.: Automatic classification of citation function, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, Association for Computational Linguistics, pp. 103–110 (2006).
17) 江藤正己: 論文の構成単位に基づいた共引用関係の尺度, 情報処理学会論文誌：データベース, Vol. 49, No. SIG 7 (TOD 37), pp. 1–15. (2008).
18) 江藤正己: 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, 三田図書館・情報学会研究大会発表論文集, 慶應義塾大学, pp. 17–20 (2007).
19) CiteSeer.PSU OAI. http://citeseer.ist.psu.edu/oai.html
20) The Lemur Toolkit for Language Modeling and Information Retrieval http://www.lemurproject.org/
21) McSherry, F. and Najork, M: Computing information retrieval performance measures efficiently in the presence of tied Scores, *30th European Conference on Information Retrieval*, pp. 414–421 (2008).