Title	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
Sub Title	Multivalued Co-Citation Measure Based on Semantic Distance between Co-Cited Papers in a Citing Paper. A Case Study Focused on Enumeration of Citations
Author	江藤, 正己(Eto, Masaki)
Publisher	三田図書館・情報学会
Publication year	2007
Jtitle	Library and information science No.58 (2007.) ,p.49- 67
JaLC DOI	
Abstract	Purpose: One typical document retrieval method is to use co-citation. The method is based on the premise that the degree of similarity among co-cited papers is equal in a particular paper. The degree is calculated with binary values: “co-cited” or “not co-cited”. To improve upon this method, the author proposes a multivalued co-citation measure based on semantic distance between co-cited papers. Methods: To determine the distance between citations, the author measured two machine parseable relationships (location and citing words) between places where papers are cited. In order to evaluate the proposed method, we identified two categories of co-citation: a group with strong relationships indicating “enumerated co-citation” (papers cited within one statement) and a group with weak relationships showing “non enumerated co-citation”. Similarities within each group were calculated and compared using the CiteSeer dataset and 6 major similarity indicators. Results: All of the similarity indicators showed that the degree of “enumerated co-citation” is higher than “non enumerated co-citation”. Consequently, it became clear that the proposed co-citation measure can be used to distinguish the strength of co-citation more precisely and that it can be applied to large-scale document collections.
Notes	原著論文
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN00003152-00000058-0049

原著論文

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

Multivalued Co-Citation Measure Based on Semantic Distance between Co-Cited Papers in a Citing Paper: A Case Study Focused on Enumeration of Citations

江　藤　正　己
Masaki ETO

Résumé

Purpose: One typical document retrieval method is to use co-citation. The method is based on the premise that the degree of similarity among co-cited papers is equal in a particular paper. The degree is calculated with binary values: "co-cited" or "not co-cited". To improve upon this method, the author proposes a multivalued co-citation measure based on semantic distance between co-cited papers.

Methods: To determine the distance between citations, the author measured two machine parseable relationships (location and citing words) between places where papers are cited. In order to evaluate the proposed method, we identified two categories of co-citation: a group with strong relationships indicating "enumerated co-citation" (papers cited within one statement) and a group with weak relationships showing "non enumerated co-citation". Similarities within each group were calculated and compared using the CiteSeer dataset and 6 major similarity indicators.

Results: All of the similarity indicators showed that the degree of "enumerated co-citation" is higher than "non enumerated co-citation". Consequently, it became clear that the proposed co-citation measure can be used to distinguish the strength of co-citation more precisely and that it can be applied to large-scale document collections.

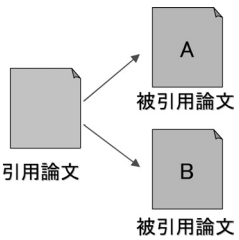
江藤正己：慶應義塾大学大学院文学研究科，東京都港区三田 2–15–45
Masaki ETO: Graduate School of Library and Information Science, Keio University, 2–15–45, Mita, Minato-ku, Tokyo, Japan
e-mail: eto@slis.keio.ac.jp
受付日：2007 年 5 月 19 日　受理日：2007 年 6 月 21 日

I. はじめに

引用は，引用する側の論文（引用論文）と引用される側の論文（被引用論文）の間の何らかの意味的なつながりを示すものと考えられる。そのため，引用は論文を探す際の一つの有力な手がかりとして，古くから論文検索システムへの適用が試みられてきた[1)]。

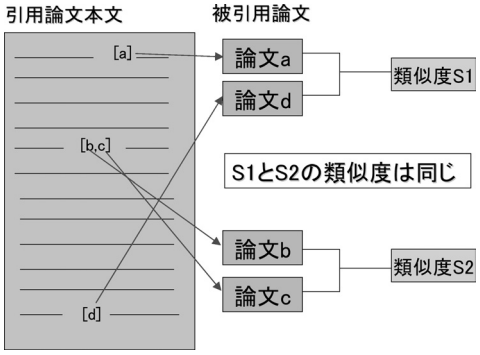
引用を用いた論文検索の方法の一つとして，類似論文検索がある。類似論文検索では，既知論文が検索システムの検索キーとなり，既知論文とシステム内の論文との類似度が計算され，高い類似度であったものが出力される。この類似度の計算の際に，様々な引用関係が用いられる。このうち本稿でとりあげるのは，共引用[2)]と呼ばれる同一の論文から引用された被引用論文同士の関係である。これは，共引用関係にある論文同士には意味的な類似性があるという発想によるもので，この関係を利用した検索手法は類似論文検索の代表的なものの一つになっている。

共引用の関係を利用した論文間の類似度は，一般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をもとに求められる。たとえば第1図の場合，類似度の算出対象は論文Aと論文Bで，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。つまり，共引用関係を利用した類似論文検索では，検索キーの論文と検索システム内の論文との共引用回数が調べられ，それに基づいて計算された類似度が高いものであった論文が出力されることになる。



第1図　共引用

共引用を用いた論文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[3), 4)]，アルゴリズムの一部に共引用を組み込んだ論文検索手法も提案されている[5), 6)]。さらに，*CiteSeer*[7)]のような実用の論文データベースにおいても用いられており，またWebページ[8)]や特許[9)]といった論文以外を対象とした研究にも応用され成果をあげてい



第 2 図　従来の共引用における類似度算出の仮定

る。これらのことから，共引用の有用性は高いと判断できる。

しかし，従来の共引用では，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が共引用関係にあるのかは考慮されていない。つまり，従来の共引用では，一つの引用論文から引用されたすべての被引用論文間の類似度はすべて同じであると仮定されている。たとえば，第 2 図のような引用があった場合，本文中のどのような箇所でどのように引用されているかといったことは全く考慮されず，被引用論文 a と被引用論文 d の間の類似度 S1 と被引用論文 b と被引用論文 c の間にある類似度 S2 は一様に同じものとして扱われる。すなわち従来の共引用は，一つの引用論文から引用された論文同士という漠としたとらえ方により，「共引用関係にある」「共引用関係にない」という 2 値情報を基にして類似度を算出しているのである。

2 値情報としての共引用に基づいて類似度を算出することは，より根本的な「引用–被引用の関係を等価に扱っている」という引用を用いた組織化手法の全てに共通する問題の一部と言える。この指摘は古くからおこなわれてきたが[10]，近年までほぼ放置され続けてきた。その原因は，大規模な論文集合を対象とした場合，「引用–被引用の関係」を一つ一つ解釈し検索に活用することが不可能であったためと考えられる。そうした論文集合を取り扱うには機械的な処理が不可欠であり，そのためには，引用論文の全文が機械可読形式で存在し，一定レベル以上の処理技術があることが求められる。このどちらも近年まで存在しなかったため，論文検索に引用を用いる手法は「引用関係にある」「引用関係にない」の 2 値的情報としての引用情報のみを利用せざるをえなかった。

しかし現在は，引用を機械的に処理際に問題であった上記 2 点の問題は解消されつつある。電子化が進んだことで多くの論文が機械可読形式になり，引用関係を解釈する機械処理技術も登場してきた。実際に「引用–被引用」関係に対して機械処理による解釈をおこなう形で，引用を 2 値ではなく多値の情報として扱う高度な組織化を目指す研究もいくつかおこなわれている（詳しくは II 章で述べる）。

論文の電子化や処理技術の向上の恩恵を受け，共引用も 2 値情報から脱却できると思われる。なぜなら，本文における引用のされ方によって共引用は 2 値以上の情報を持つと考えることが可能であるためである。このことは，引用論文の本文を人間が解釈しながら，共引用関係にある論文間の関係を探る研究（共引用文脈分析[11]）が存在することからもうかがえる（共引用文脈分析については，II 章 C 節で述べる）。

従来の共引用は，媒体と技術の制約により，引用論文の本文における個々の被引用論文同士の関係の違いを考慮せず，2 値情報をもとに粗く大まかに類似度を算出していた。しかし，制約が無くなりつつある現在，個々の共引用の関係をとらえ，多値情報を持つ共引用に基づいて類似度を算出することが不可能でなくなってきている。共引用を多値なものに拡張することで類似度の算出がより精密になり，類似論文検索の性能が向上すると予想される。

そこで，引用論文の本文に依拠して共引用を多値化し，その可能性と有用性を検証することを本稿の目的とする。つまり，引用論文の本文を解析して得た情報に基づいて共引用関係をとらえ，関係の強弱を持つ共引用への拡張を試みる。以下本稿では，まず，II 章で，関連研究をみることによって本文と引用を組み合わせることの可能性や有用性を確認する。そして，III 章で，引用箇所間の意味的な近さに基づいて共引用を多値化する手

法を提案する。IV 章では，提案手法を検証するために，引用論文の本文を機械処理によって解析して被引用論文間の類似度を比較する実験をおこなう。最後に V 章で，実験結果に対する分析・考察をおこない今後の課題について述べる。

II. 関 連 研 究

本章では，本文を解析して得た情報と引用による情報を組み合わせることで成果をあげている関連研究を概観し，共引用の多値化の可能性と有用性を確認する。A 節では，引用–被引用の関係の多値化を目指すものとして，引用と引用文章（引用箇所の周辺の文）に含まれる語を結びつける研究，および引用に対してカテゴリを自動付与する研究についてふれる。そして B 節では，本文を解析することによって書誌結合を多値化した研究について議論する。最後に C 節で，人手による解析によって本文と共引用を組み合わせる共引用文脈分析について述べる。

A. 本文と引用の組み合わせ

引用文章に含まれる語を機械処理によって解析して，様々な目的に利用する研究がおこなわれている。これらは，Web ページのアンカーテキストを利用する研究[12]に近い。

1. 引用文章に含まれる語を利用する研究

引用文章に含まれる語を被引用論文の索引語として追加し，論文検索の性能向上をめざす研究がおこなわれている。この研究のアイデアは“引用文章は被引用論文に関する情報を与える”[13]というものである。この着想に基づく検索システムは，「検索語」と「被引用論文に元々付与されていた索引語，及び引用文章から得られた索引語」との適合度を計算し，結果を出力する。

この種の初期的な研究で，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ものとして，O'Connor[13] のものがある。その後，Bradshaw[14] は引用文章を利用して索引語に対する重み付け（様々な引用論文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）をおこなっている。さらに，引用文章に含まれる語の中でも，被引用論文に関係のある語と関係のない語があるとして，両者を区別することを Ritchie[15] らが試みている。

また，引用文章に含まれる語を検索以外に利用する研究として，自動的なシソーラス構築を目指すもの[16]や，データマイニングに適用するもの[17]もある。

このような研究は，引用を引用論文と被引用論文との単純なつながりとしてだけみるのではなく，そこからより多くの情報を取り出そうとする試みととらえられる。引用を多値化することの可能性と有用性を示す例といえよう。

2. 引用の役割を自動分類する研究

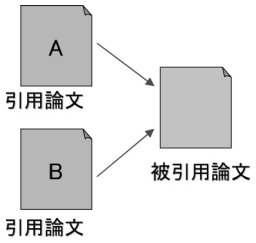
論文間のつながりである引用をいくつかのカテゴリに自動分類する研究がおこなわれている。後述するこれらの研究は，引用文脈分析[18]の成果を利用する試みといえ，引用文脈分析で提案されてきた引用の役割カテゴリ（たとえば Spiegel-Rösing のもの[19]など）に引用を分類することを目的としている。ここで挙げる研究は，より一般的には，教師ありの自動分類の研究に相当し，引用文章の特徴とその引用が属するカテゴリを学習して未知の引用に対して自動的に正解カテゴリを付与する研究としてとらえることができる。

具体的には，引用文章に含まれる語や句を手がかりに引用を分類する試みとして，Garzone ら[20]，難波ら[21]，Teufel ら[22]のものがある。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[23]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[24]によって提案されている。

このような研究は，引用を様々なカテゴリに分類しているものであり，引用を多値化したものととらえることができる。

B. 書誌結合の多値化

A 節 2 項で挙げたもののうち，特に難波らは，引用をカテゴリに分類するだけでなく，それを利



第 3 図　書誌結合

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[25] によって，共引用よりも前に提案された論文間の類似度指標である。書誌結合では，第 3 図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 が共に「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。つまり，この手法は二つの論文間の関係を「書誌結合の関係にある」「書誌結合の関係にない」「同じカテゴリでの書誌結合関係にある」に多値化したものといえる。

難波らの手法は，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。難波らはこの手法により検索性能が向上することを報告しており，二つの論文間の関係を多値化することの有用性を示している。

C.　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，本文を分析して得た情報と引用を組み合わせるものとして共引用文脈分析[11]がある。この研究では，論文集合中の各論文間の関係をとらえるために，その集合中の論文を複数引用している論文の本文が利用される。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るものである。

文脈によって異なる種々の共引用を一つ一つとらえようとする研究があることは，(1) 複数種類の共引用関係が存在するとみなすことができ，(2) 本文の内容を分析することでそれらの被引用論文間の関係をとらえる情報を引き出せることを示唆しており，共引用を多値化しその情報を活用することの有用性がみてとれる。

ただし，共引用文脈分析では人間による作業が想定され，小規模な論文集合のみが対象となる手法である。たとえば，この手法では場合によっては行間を読む作業が必要となる[26]。行間を読むという作業は，現在の機械処理のレベルでは難しい。A 節で述べた研究も，引用文章に含まれる語句そのものを利用したものであり，言外の意味を汲み取るというレベルのものではない。

つまり，共引用文脈分析は有用であるが，この手法をそのまま大規模な論文集合に適用することは難しいものである。引用論文の本文に依拠した共引用を実現するためには，現在の機械処理のレベルに適した形の共引用文脈分析を考える必要がある。本稿の目的は，大規模な論文集合を対象とした機械処理による共引用文脈分析手法を考案することととらえることもできる。

以上本章でみてきたように，引用論文の本文を機械処理を用いた何らかの方法によって分析することが可能であり，それに基づいて共引用を多値化することに有用性があることが確認できた。

III.　引用箇所間の意味的な近さに基づいた共引用

本章では，まず A 節で，引用箇所間の意味的な近さに基づく共引用の多値化手法の提案と手法を成立させるための仮説の提示をおこない，類似論文検索の枠組みの中でその可能性と有用性について論じる。そして B 節で，仮説検証のために引用箇所間の意味的な近さを判別する基準として「列挙形式の引用」をとりあげ，基準として適当である理由を述べる。

A.　共引用を多値化するための手法

本稿では，本文に依拠して共引用を多値化する

手法として，引用箇所間の意味的な近さに基づくことを提案する。この手法は，「引用箇所間が意味的に近ければ共引用関係が強く，遠ければ共引用関係が弱い」という仮説により成り立つものである。II 章で述べた難波らの研究[21]も，引用箇所間の意味的な近さをとらえて書誌結合を多値化したものと考えられ，提案手法の仮説を支持するものといえる。

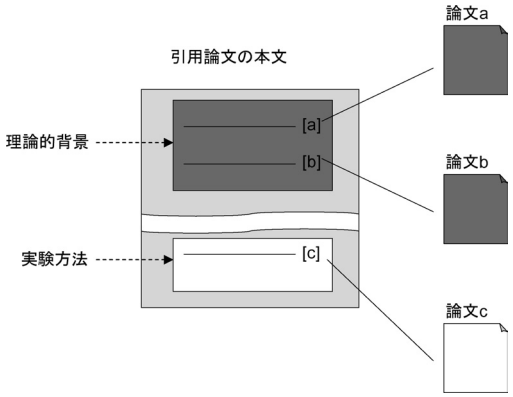
提案手法を具体化するためには，引用箇所間の意味的な近さを機械処理によってとらえることが必要となる。II 章 C 節で述べたように，行間を読むことが必要となる引用箇所間の意味的な近さのとらえ方を考案しても，実用レベルの類似論文検索における共引用の拡張となる可能性は低い。

機械処理によって引用箇所間の意味的な近さをとらえる方法として，「引用箇所の位置」と「引用文章中の語の共起」の二つが考えられる。以下，これら二つについて述べる。

1. 引用箇所の位置からとらえた共引用関係の強弱

論文はある事物について理論的な筋道を立てて説かれた文章であり[27]，論文中の各“パラグラフは内容的に連結されたいくつかの文の集まり”[28]とされている。これらのことから，非常に粗くいえば，引用箇所の位置関係の強さはその箇所間の意味的な近さを反映するものであると思われる。すなわち，意味的に近い内容の記述同士は論文中の近い箇所にある程度あらわれ，意味的に遠い記述同士は離れてあらわれることが多いと考えられる。よって，引用箇所の位置関係に着目することで，引用箇所間の意味的な近さを推測できると考えられる。たとえば論文の理論的背景と実験方法での引用を示した第 4 図の場合，論文 a と論文 b の引用箇所間と論文 b と論文 c の引用箇所間では，前者の方が意味的に近いといえる。

このことから，第 4 図における論文 a と論文 b では，引用箇所の位置関係が強く当該箇所間が意味的に近いため，それぞれの引用箇所で引用された論文 a と論文 b の共引用関係も強いことが想定できる。逆に，論文 b と論文 c では，同じ論文から引用された論文同士であっても，引用箇所の位置関係が弱く当該箇所間は意味的に遠いため，それぞれの箇所で引用された論文 b と論文 c の共引用関係も弱いと判断できる。



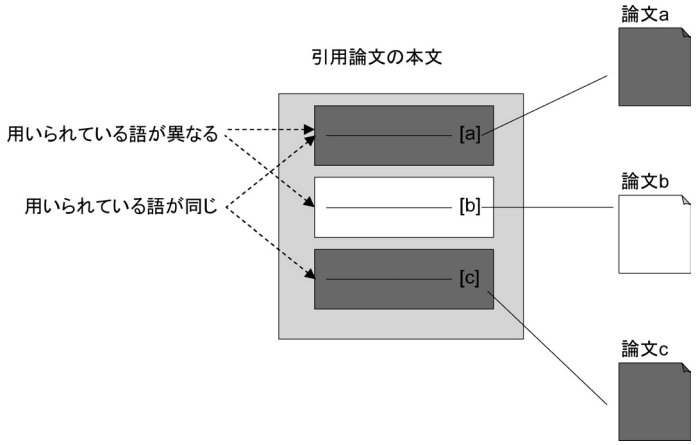
第 4 図　位置からとらえた共引用関係

もちろん，たとえば背景や目的で述べた内容を再び考察で言及することもあり，位置関係の強さが全ての引用箇所間の意味的な近さに対応するとは限らない。しかし，位置関係の強さがある程度引用箇所間の意味的な近さを表すと考えられ，機械処理によって意味的な近さをとらえる方法の一つとして引用箇所の位置関係を利用することは適当と思われる。

2. 引用文章中の語の共起からとらえた共引用関係の強弱

文章中に同じ語が数多く出現するような文章同士は意味的に近いといえる。よって，引用文章中の語が同じようなものであれば引用箇所間は意味的に近く，異なっていれば引用箇所間は意味的に遠いと考えられる。したがって，引用文章中の語の共起関係に着目することで，引用箇所間の意味的な近さを推測できると考えられる。たとえば，第 5 図のような引用文章であった場合，論文 a と論文 c の引用箇所間と論文 a と論文 b の引用箇所間では，前者のほうが意味的に近いといえる。

このことから，第 5 図における論文 a と論文 c では，引用文章中の語の共起関係が強く当該箇所間が意味的に近いため，それぞれの引用箇所で引



第 5 図　語の共起からとらえた共引用関係

用された論文 a と論文 c 間の共引用関係も強いことが想定できる。逆に，論文 a と論文 b では，同じ論文から引用された論文同士であっても，引用文章中の語の共起関係が弱く当該箇所間は意味的に遠いため，それぞれの箇所で引用された論文 a と論文 b 間の共引用関係も弱いと判断できる。

3.　仮説の検証方法

引用箇所間の意味的な近さに基づいて共引用を多値化する手法の可能性と有用性を示すためには，仮説「引用箇所間が意味的に近ければ共引用関係が強く，遠ければ共引用関係が弱い」が成立することを示す必要がある。この仮説中の，引用箇所間の意味的な近さについては，1 項と 2 項の議論より，引用箇所の位置関係と引用文章中の語の共起関係に着目できることが導かれる。また，共引用関係の強弱は被引用論文間の類似性の強弱を意味するといえる。したがって，仮説は「位置関係や引用文章中の語の共起関係からみて引用箇所間が意味的に近い被引用論文の類似性は，意味的に遠い被引用論文間の類似性よりも強い」と言い換えられる。

これらのことから，仮説を検証する方法として，引用論文における引用箇所の位置関係や引用文章中の語の共起関係から，被引用論文のペアを引用箇所間が意味的に近いものと遠いものに分類し，両者の類似度を算出・比較することが考えられる（類似度の算出方法については IV 章 B 節で述べる）。もし，引用箇所間が意味的に近い被引用論文のペアの類似度が，意味的に遠い被引用論文のペアの類似度よりも高ければ，仮説が検証されたといえる。

ただし，検証実験をおこなうためには引用箇所の位置関係が強いものと弱いもの，あるいは引用文章中の語の共起関係が強いものと弱いものを分けるための具体的な基準が必要となる。B 節では，この具体的な基準として「列挙形式の引用」をとりあげ，基準として適当である理由を述べる。

B.　列挙形式の引用を用いた共引用の分類

引用箇所間が意味的に近い被引用論文のペアと意味的に遠い被引用論文のペアを分類するための基準として，本稿では列挙形式の引用に着目した。列挙形式の引用とは，第 6 図で示すような複数の論文を同時に並列列挙する形式の引用のことを指す。本稿では，この形式で引用された論文間の関係を列挙共引用，それ以外の引用で引用された論文間の関係を非列挙共引用と呼ぶ。列挙形式の引用を分類の基準として用いる理由は，以下で述べる三つである。

一つ目の理由は，引用箇所の位置同士が最も近いと判断されることである。列挙共引用の関係にある論文の引用箇所は，同一文の同一箇所である。そのため，引用箇所の位置関係は，すべての

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

第 6 図　列挙形式の引用の例

引用箇所の位置関係の中で最も強いといえる。したがって，列挙共引用を引用箇所の位置関係が強いもの，非列挙共引用を引用箇所の位置関係が弱いものとして分類することができる。

二つ目の理由は，引用文章中の語が全く同じであることである。列挙共引用の関係にある論文の引用箇所は同じであるため，双方の引用箇所で用いられている語は同一である。よって，その引用箇所の引用文章中の語の共起関係は，すべての引用箇所の引用文章中の語の共起関係の中で最も強いといえる。したがって，列挙共引用を引用文章中の語の共起関係が強いもの，非列挙共引用を引用文章中の語の共起関係が弱いものとして分類することができる。

三つ目の理由は，引用が列挙形式であるか否かの判別が機械処理によって容易なことである。大規模論文集合への適用を想定した共引用の拡張であるため，行間を読む作業が求められるような基準は不適当である。列挙形式か否かは，表層的な文字列の解釈のみで判別が可能である。判別するためには，引用文に含まれている語の意味や前後の筋道の流れなどを把握する必要はない。機械処理が得意な表層的な文字列解釈のみで判別可能なため，大規模な論文集合にも適用しやすい基準である。

つまり，列挙共引用は，引用箇所の位置関係と引用文章中の語の共起関係の両方において，最も関係が強いものであるといえる。端的に引用箇所間が意味的に近く，かつ機械処理しやすい特徴も持つため，仮説の検証の際に列挙形式の引用を基準とすることが適当であると考えられる。

これまでの引用研究において，列挙形式の引用に着目しているものとして，Ruff[29] や牛澤[30]のものがある。ただし，両者の研究は，共引用の立場からのものではないため，列挙形式の引用における被引用論文間の関係や類似性については言及していない[31]。

以上のことから，A 節で述べた仮説「引用箇所間が意味的に近ければ共引用関係が強く，遠ければ共引用関係が弱い」ことを検証するためには，具体的には「列挙共引用関係にある論文間の類似度は，非列挙共引用関係にある論文間の類似度よりも高い」ことを確認すれば良いといえる。IV 章では，列挙共引用関係にある論文間の類似度と非列挙共引用関係にある論文間の類似度を比較するためにおこなった実験について述べる。

IV.　類似度の算出・比較実験

本章では，III 章で述べた仮説を検証するために「列挙共引用関係にある論文間の類似度は，非列挙共引用関係にある論文間の類似度よりも高い」ことを確認する実験をおこなった。実験は，列挙共引用と非列挙共引用のそれぞれの関係にある論文を収集し，それらの類似度を比較することによっておこなった。以下，A 節で比較をおこなう論文ペアの収集，B 節で収集した論文ペアの類似度算出方法について述べる。そして，C 節で算出をおこなった結果を比較する。

A.　列挙共引用・非列挙共引用関係にある論文の収集

1.　基礎データ

論文を収集するための基礎データとして，*CiteSeer* で公開されているデータセット *CiteSeer Metadata*[32] を利用した。このデータセットには，約 57 万件の論文書誌情報が含まれており，それぞれの論文には ID 番号が付与されている。書誌情報には，タイトルや著者名に加え，引用情報および，論文全文ファイルを入手するための URL などが含まれている。なお，このデータセットは，*CiteSeer* のシステムによって機械的に作成されており，またデータセット内の論文との引用関係のみが収録対象となっている。そのため，実際の論文本文では引用されていても，作成時にミスがあった場合や収録対象外であるものは，データセットの引用情報には含まれていない。本実験では，データセットに含まれる引用情報を基準として利用し，それ以外の被引用論文については分

析の対象外とした。本研究で対象とする引用論文における被引用論文間の関係は，被引用論文がデータセットに含まれるているか否かによって変化しないと思われ，これらを対象外とすることによる影響はあまりないと考えられる。

2. 論文集合の作成

a. 引用論文集合の作成

共引用が列挙形式の引用によるものか否かを分類するためには，引用論文の全文が必要となる。引用論文の収集は第 7 図のような手順でおこなった。

まず，データセット内の論文から，タイトルかディスクリプタに「database」を含む 29,537 件の論文を選択した。次に，選択した論文の全文のダウンロードを試みた。その結果，13,551 件の論文全文を入手することができた。

全文を入手できた論文の中から，本文の解析をプログラムで処理し易いものとして，引用記号（本文と引用文献リスト中で，被引用論文を示すのに用いられる記号）に

- 大括弧に囲まれているもの
- 数字とアルファベットを含んでいるもの

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

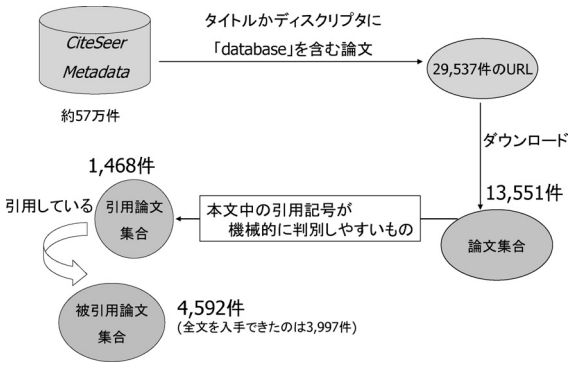
を用いているものを選び，引用論文集合とした。引用論文集合に含まれる論文数は，1,468 件となった。

b. 被引用論文集合の作成

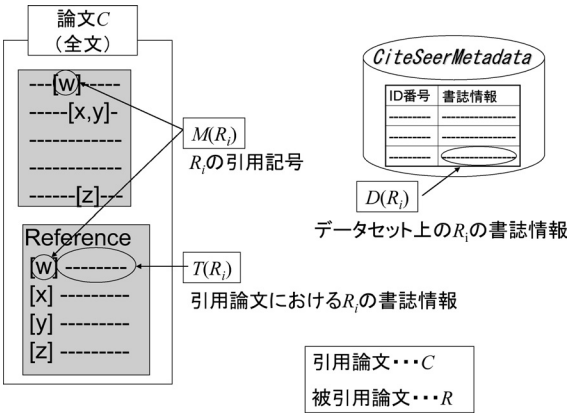
データセットに含まれる引用情報を用いて，引用論文集合の 1,468 件の論文によって引用されている論文を求めた。その結果として得られた 4,592 件の論文を被引用論文集合とした。なお，引用論文集合と被引用論文集合には重なりが存在する。また，詳しくは B 節で述べるが，論文間の類似度指標の一つとして語の共起頻度 (tf*idf/cosine) を用いるため，被引用論文集合についても論文の全文の収集をおこなった。収集作業では，まずデータセットに含まれる全文データの URL を用いてダウンロードを試みた。加えて，語の共起頻度によって類似度が算出可能な論文ペアをできる限り増やすため，*CiteSeer* データベース[7]で入手できる「その他の全文 URL」，「論文キャッシュデータ」を用いて追加収集をおこなった。結果として，3,997 件の全文を入手した。

3. 列挙形式で引用された論文を特定する方法

ここで，引用論文を解析することによって，列挙形式で引用された論文の組を特定する方法についてのべる。まず，解析処理の際に使用するデータを第 8 図のように整理する。解析対象の論文を C（$C \in \mathbb{C}$, $\mathbb{C}$は引用論文集合），C が引用している論文を R_i ($i=1, \ldots, N$)。ここで，N は C が引用している論文総数とする。また，C で用いられる R_i の引用記号を $M(R_i)$，C の引用文献リスト中における R_i の書誌情報を $T(R_i)$ とする。そして，デー



第 7 図　論文収集の流れ



第 8 図　解析に利用するデータ

タセット上での R_i の書誌情報を $D(R_i)$ とする。

引用が列挙形式であるかは，一つの大括弧の中に複数の引用記号が出現するか否かによって判断することができる。よって，列挙共引用の論文の組を特定するためには，「$M(R_i)$ の出現の仕方の分析」，および「$T(R_i)$ と $M(R_i)$ とのマッチング」をおこなわなければならない。加えて，本実験では 1 項でも述べたようにデータセットの引用情報を基準とするため「$D(R_i)$ と $T(R_i)$ とのマッチング」処理もおこなう必要がある。以下，この三つの処理の手順について述べる。

a.　$D(R_i)$ と $T(R_i)$ とのマッチング

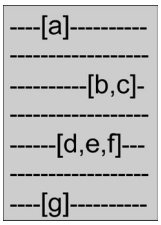
同じ R_i であっても $T(R_i)$ と $D(R_i)$ は，著者名や雑誌名などの省略形式の使用などの影響により，テキスト文字列として完全一致しないことが多い。そこで，論文 ID を用いて $D(R_i)$ を得たあとで，論文のタイトルを手がかりに，$D(R_i)$ と $T(R_i)$ のマッチング作業をおこなった。論文のタイトルは比較的長い文字列でありながら，著者の書誌事項の書き方による差異が少なく，論文を一意に特定するキーとして有効と考えられるためである。

b.　$T(R_i)$ と $M(R_i)$ とのマッチング

$T(R_i)$ を示す $M(R_i)$ は，$T(R_i)$ の文字列が出現する直前の大括弧中の文字列と考えられる。よって，$T(R_i)$ の直前の大括弧に含まれる文字列を $T(R_i)$ に該当する $M(R_i)$ とした。

c.　$M(R_i)$ の出現の仕方の分析

一つの大括弧内に $M(R_\alpha)$ と $M(R_\beta)$ が含まれていれば，$D(R_\alpha)$ と $D(R_\beta)$ は列挙共引用の組と判別することができる。そこで，C の全文中に含まれる全ての大括弧を対象に複数の引用記号が含まれているか否かを分析し，列挙形式で引用された論文の組を特定した。



第 9 図　引用論文本文の例

4.　列挙共引用・非列挙共引用の論文ペアの作成方法

3 項で述べた方法によって列挙形式で引用された被引用論文の組を特定し，類似度比較をおこなうデータを作成する。比較をおこなうデータは論文ペア間の類似度であるため，列挙形式の引用で同時に三つ以上の論文を引用していた場合は，同時に引用された被引用論文の各組み合わせを列挙共引用のペアとした。非列挙共引用のペアは，全被引用論文の組み合わせから，列挙共引用のペアを除くことにより作成できる。

たとえば，a〜g の七つの被引用論文があった場合，全組み合わせは 21 ペアになる。第 9 図で示したような引用が本文でおこなわれていた場

合，(b,c)(d,e)(d,f)(e,f) の四つが列挙共引用のペアであり，残りの 17 ペアが非列挙共引用となる。

5. 比較データの作成

2 項 a 目で述べた 1,468 件の引用論文集合を対象として，3 項および 4 項の方法を用いて列挙共引用のペアおよび非列挙共引用のペアを作成した。その結果，列挙共引用のペアは延べ 1,500 件，非列挙共引用のペアは延べ 38,158 件となった（なお，3 項 a 目の工程で，$T(R_i)$ とマッチングできなかった $D(R_i)$ があったが，そのような被引用論文は引用の形式が不明なため，比較データ作成の対象外とした）。

ただしこの中には，(1) 列挙共引用でもあり非列挙共引用でもあるペア（たとえば，論文 a と論文 b のペアが，論文 c においては列挙形式で引用され，かつ論文 d においては非列挙形式で引用される）や (2) 引用論文集合から共引用されている回数が異なるペア（たとえば，列挙形式で 2 回共引用されているペアや非列挙形式で 3 回共引用されているペア）が含まれている。共引用の種類の違いを明確にし，従来の共引用の考え方において類似度が全く同一なもののみを比較するために，次のようなものを比較データとした。列挙共引用のペアは「引用論文集合からの共引用回数が 1 回のみで，それが列挙共引用であったペア」であり，非列挙共引用は「引用論文集合からの共引用回数が 1 回のみで，それが非列挙共引用」である。その結果，比較対象データは列挙共引用のペアが 1,005 件，非列挙共引用のペアが 28,118 件となった。

B. 論文間の類似度指標

ここでは，A 節で求めた列挙共引用のペアおよび非列挙共引用のペアの類似度の算出に用いる指標について述べる。類似度指標として，多くの先行研究あるいは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できる「tf*idf/cosine」「書誌結合」「従来の共引用」「直接引用」を用いる。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[33]。以上示した六つの指標を用い，多様な類似性の観点[34]から類似度を算出し検討をおこなうことで実験の信頼性を高める。なお，この六つの指標のうち，「tf*idf/cosine」「正規化書誌結合」「正規化共引用」は，その類似度が 0〜1 に正規化された値をとる。

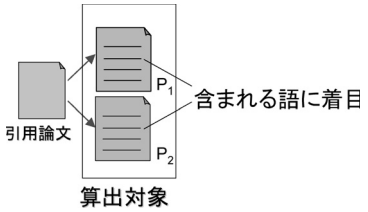
以下，論文間の類似度を算出する指標と，その算出方法について述べる。以下のすべての指標において，求める類似度を S，類似度の算出の対象となる論文を P_1，P_2 とし，P が引用している論文集合を $citing(P)$，その数を $count(citing(P))$，P を引用している論文集合を $cited(P)$，その数を $count(cited(P))$ とする。

1. tf*idf/cosine（第 10 図）

この指標は，論文 P_1 と論文 P_2 の全文における語の共起からみた類似度であり，ベクトル空間モデルによるものである。この指標では，同じような語を用いている論文同士ほど類似していると判断される。類似度の算出には論文の本文が必要なため，ペアのうち一方の本文を入手することができなかったものは対象外とした。

ベクトル空間モデルを使って類似度を算出する方法には様々な種類のものがあるが，今回は，tf*idf 法による語の重み付けをおこない，類似度の算出にはコサイン係数を用いた。式 (1) が算出式である。

$$S = \frac{\langle \vec{P}_1, \vec{P}_2 \rangle}{|\vec{P}_1| \cdot |\vec{P}_2|} \qquad (1)$$



第 10 図 *tf*idf/cosine*

ここで，$\vec{P}_1$ は，$\vec{P}_1=(W_1,W_2,\cdots,W_M)$（$P_1$ を M 次元ベクトルで表したもの，M は全文書の異なり出現語数）であり，$\vec{P}_2$ についても同様である。また，$\langle\vec{P}_1,\vec{P}_2\rangle$ は $\vec{P}_1$ と $\vec{P}_2$ の内積，$|\vec{P}|$ は $\vec{P}$ のノルムである。なお，ベクトル $\vec{P}$ 中の各要素 W（すなわち各語の重み）は以下のように求める。

$$W=\log\left(\frac{\text{その語の出現回数}}{\text{当該論文中の延べ語数}}+1\right)\times\left(\log\left(\frac{\text{総文書数}}{\text{その語の出現文書数}}\right)+1\right)$$

ここでは，全文をダウンロードできたすべての論文（引用論文集合＋被引用論文集合，15,713 件）を総文書とする。なお，ストップワードとして，SMART システムでストップワードとされている語[35]，および本文の内容とは無関係な TeX 用の記号を設定した。また，本文中の各単語を英文 Tagger ソフト *MontyLingua*[36] を用いて原型に変換してから，算出をおこなった。

2. 書誌結合（第 11 図）

書誌結合は，論文 P_1 と論文 P_2 がどれほど同じ論文を引用しているかに基づくものである。この指標では，同じ論文を引用している論文同士ほど類似していると判断される。書誌結合数は，データセットの引用情報を用いて算出した。「書誌結合」は式 (2) によって，「正規化書誌結合」は式 (3) によって求められる。ただし，A 節 1 項でも述べたように，すべての引用情報がデータセットに含まれてはいないため，算出対象の論文の中には引用数が 0 となるものがある。算出対象のペアのうち一方の引用数が 0 である場合，0 で除算することになるため，「正規化書誌結合」の対象から除外した。

$$S=count(citing(P_1)\cap citing(P_2)) \tag{2}$$

$$S=\frac{count(citing(P_1)\cap citing(P_2))}{\sqrt{count(citing(P_1))\times count(citing(P_2))}} \tag{3}$$

3. 従来の共引用（第 12 図）

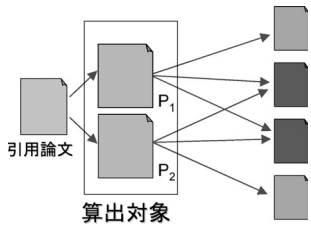
従来の 2 値の共引用による指標である。論文 P_1 と論文 P_2 が，共引用された回数に基づいて類似度を算出する。共引用回数は，データセット全体からの引用を用いて求めた。ここで，従来の共引用を用いることは，「列挙共引用関係・非列挙共引用関係にある論文のペアは，従来の共引用回数に基づく類似度指標の観点から見た場合，どの程度類似しているか」を求めることになる。「従来の共引用」は式 (4) によって，「正規化共引用」は式 (5) によって求められる（正規化書誌結合の場合とは違い，被引用回数が 1 回以上あるもののみが算出対象となるため，0 での除算は生じない）。

$$S=count(cited(P_1)\cap cited(P_2)) \tag{4}$$

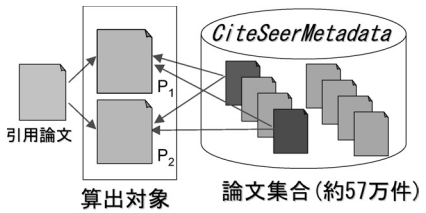
$$S=\frac{count(cited(P_1)\cap cited(P_2))}{\sqrt{count(cited(P_1))\times count(cited(P_2))}} \tag{5}$$

4. 直接引用（第 13 図）

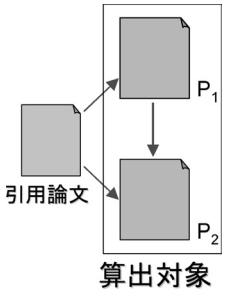
直接引用とは，論文 P_1 と論文 P_2 の間に引用–被引用関係があるか否かに基づく指標である。一方が他方を引用していれば，類似性があるとするもので Small[2] の研究でも用いられている。 直接



第 11 図　書誌結合



第 12 図　従来の共引用



第 13 図　直接引用

引用の算出についても，データセットの引用情報を用いた。式 (6) によって類似度が決まり，一方の論文が他方の論文を引用していれば 1，引用していなければ 0 となる。

$$S=\begin{cases} 1\ (P_1 \text{ と } P_2 \text{ に引用関係があるとき}) \\ 0\ (P_1 \text{ と } P_2 \text{ に引用関係がないとき}) \end{cases} \quad (6)$$

C.　類似度の算出・比較結果

類似度を算出・比較する類似度指標毎のデータ数は第 1 表のようになった。「tf*idf/cosine」と「正規化書誌結合」は，それぞれ B 節の 1 項と 2 項で述べたように，算出対象から除外したものがあるため，データ数が他の指標に比べて少なくなっている。第 1 表のデータを用いて類似度を算出した。その算出した類似度をそれぞれ平均した結果が第 2 表である。また，類似度指標毎にそれぞれの共引用ペアの分布をみたものが，第 14 図である。第 14 図は，横軸が「類似度の値」，縦軸が「ある種類の共引用ペアのうち，当該の類似度の値になった共引用ペアの割合」（たとえば，tf*idf/cosine においては，22,061 件の非列挙共引用のペアのうち類似度が 0〜0.1 であったものはその約 45%）である。

第 2 表からは，実験で用いた「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」「直接引用」のどの指標においても，列挙共引用の類似度の方が非列挙共引用よりも上回っていることが分かった。また，類似度が 0〜1 に正規化される「tf*idf/cosine」「正規化書誌結合」「正規化共引用」においても，列挙共引用のペアが非列挙共引用よりも 2 倍以上の差を示した。そして，第 14 図の分布からは，列挙共引用は非列挙共引用よりも高い類似度に多く分布していることが確認できた。

ただし，第 14 図でもみられるように，算出した結果の中には非列挙共引用よりも類似度の低い列挙共引用の事例が存在した。そのような事例の分析および比較実験の結果に対する考察を V 章でおこなう。

第 1 表　各類似度指標のデータ数

	列挙共引用	非列挙共引用
*tf*idf/cosine*	791	22,061
正規化書誌結合	804	21,278
正規化共引用	1,005	28,118
書誌結合	1,005	28,118
従来の共引用	1,005	28,118
直接引用	1,005	28,118

第 2 表　類似度算出結果

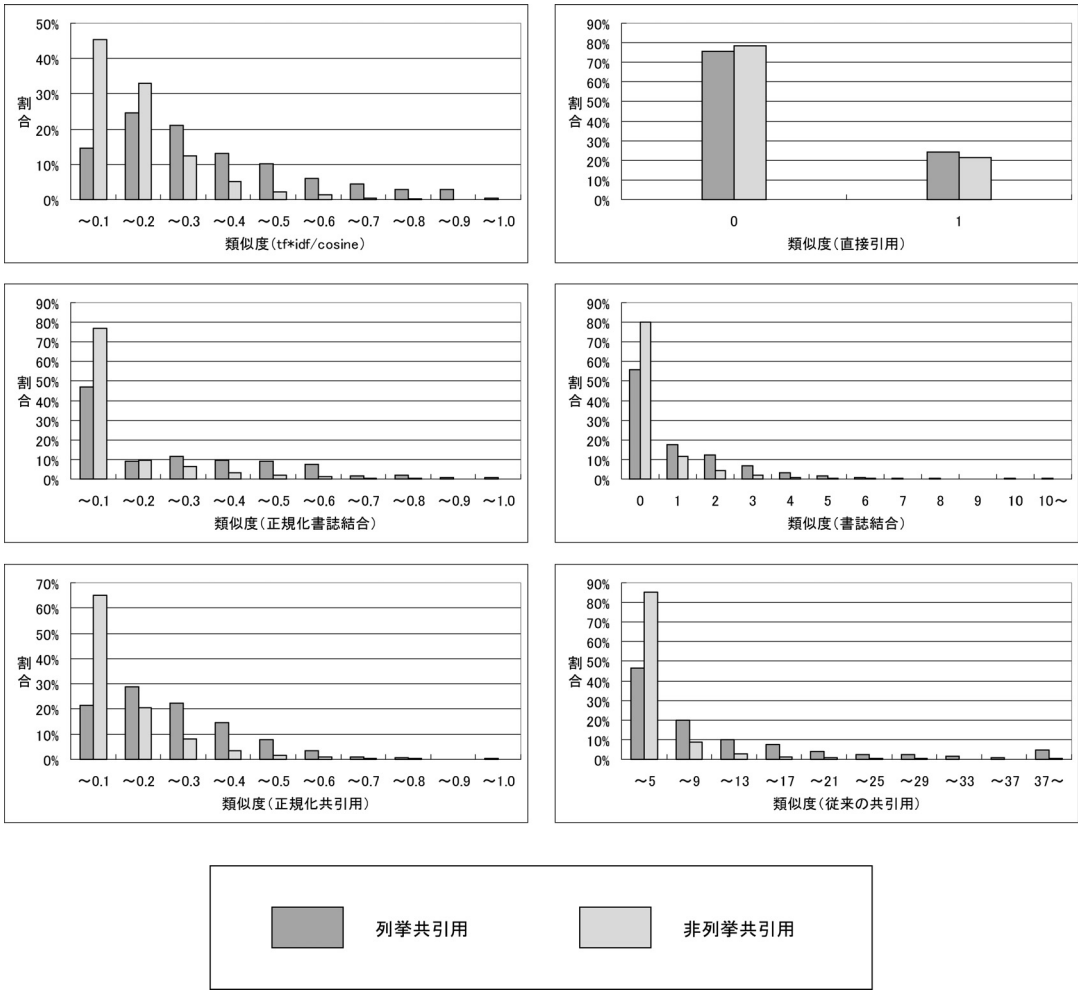
	列挙共引用	非列挙共引用
*tf*idf/cosine*	0.30	0.14
正規化書誌結合	0.20	0.06
正規化共引用	0.23	0.11
書誌結合	1.08	0.37
従来の共引用	9.70	2.86
直接引用	0.24	0.22

V.　分析・考察

本章では，まず A 節で実験において非列挙共引用よりも低い類似度が算出された列挙共引用ペアを分析する。次に B 節で，実験結果に対する考察をおこなう。そして，C 節で類似論文検索システムの実現に向けた今後の課題を挙げ，D 節で多値化された共引用のその他の可能性について述べる。

A.　類似度の低い列挙共引用ペアを引用した箇所の分析

IV 章でおこなった実験の目的は「列挙共引用関係にある論文間の類似度は，非列挙共引用関係にある論文間の類似度よりも高い」ことを確認す



第 14 図　共引用ペアの分布

ることであった。実験の結果，平均値においては，列挙共引用ペアの類似度は非列挙共引用ペアの類似度よりも高い値になった。しかし，算出をおこなったペアの中には，非列挙共引用ペアよりも類似度の低い列挙共引用ペアの事例も存在した。本節では，そのような事例を分析してその原因を探る。

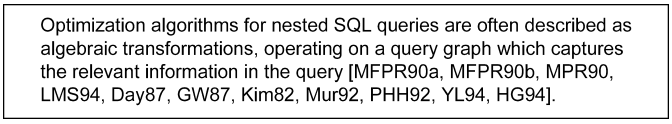
分析対象は，「tf*idf/cosine」「正規化書誌結合」「正規化共引用」の値を足した合計値を順番に並べ，その下位となった列挙共引用ペアを引用している 30 箇所（同一箇所内の複数のペアが下位にある場合は除く）とした。

1.　引用数の分析

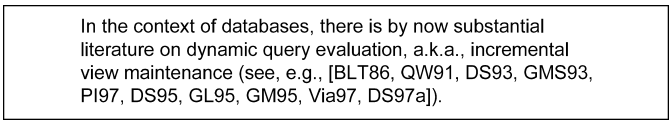
類似度の低い列挙共引用ペアを引用している箇所には，多数の論文を引用しているものが多くみられた。たとえば，第 15 図のようなものである。

下位 30 の箇所が引用している論文数の平均値は，6.2 であった。比較対象として，上位の列挙共引用ペアを引用している 30 箇所の引用数を調査したが，その平均値は 3.0 であった。下位と上位の間で平均値の差をみる検定をおこなった結果，この差は有意（有意水準 1%）であることが分かった。

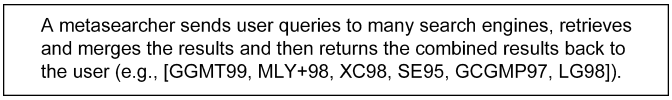
類似度が低い列挙共引用ペアを引用している箇



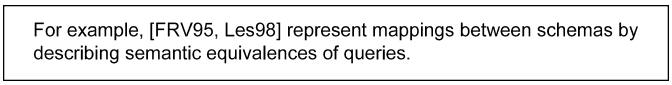
第 15 図　引用数の多い例



第 16 図　例示するための引用 1



第 17 図　例示するための引用 2



第 18 図　例示するための引用 3

所で，多数の論文を引用しているものが多くみられる理由の一つとして，被引用論文の内容にあまり言及しない，おざなりの引用 (perfunctory citation)[37] がされていることが考えられる。このような場合，列挙形式の引用であっても，被引用論文と引用論文との関係は弱いため，被引用間の関係も弱くその類似度が低いと思われる。

2. 引用文の内容分析

次に，下位 30 箇所の引用文に関する簡単な内容分析をおこなった。分析の結果，類似度が低い列挙共引用ペアの引用文の特徴として，以下の二つがみられた。

a. 例示するための引用

類似度が低い列挙共引用ペアの引用文にみられた特徴の一つとして，「例示するための引用」が挙げられる。その例が，第 16 図，第 17 図のような引用である。

例示として引用をおこなう場合，引用論文と被引用論文間の関係が弱くなることが考えられる。たとえば第 16 図の “see, e.g.” のような語があった場合，引用論文の当該箇所の文脈と間接的にしか関係しないような論文を引用することもありうる。このような場合，列挙形式の引用によるものであっても，被引用論文間の類似性が強いとは限らない可能性がある。

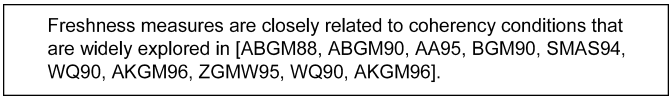
なお，下位 30 箇所の中には，第 18 図のように，同時に引用する論文数が少なく例示的な引用をおこなっているものもあった。同時に引用する論文数にかかわらず，例示的な引用が類似度の低い論文を引用するといったことも予想される。

また，同様の原因が想定されるものとして，引用論文中の脚注における引用があった。脚注は引用論文の当該箇所の文脈と間接的にしか関係しないために本文外で書かれたことが多い。このような場合も，引用論文と被引用論文間の関係が弱いためにその類似度が低いことが考えられる事例といえよう。

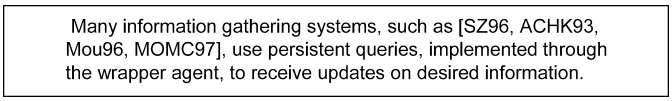
b. 弱い被引用論文間の関係を示す引用

類似度が低い列挙共引用ペアの引用文にみられたもう一つの特徴として，「様々な種類のものが多くある」という主張を意図した引用が挙げられる。第 19 図，第 20 図がその例である。

たとえば，第 19 図の引用は，広く調べられて



第 19 図　弱い被引用論文間の関係を示した引用 1



第 20 図　弱い被引用論文間の関係を示した引用 2

いることを主張する引用と思われる。様々な種類の論文を多くとりあげなければ,「広く」という主張に合致しない。その結果，類似性の弱い論文が列挙形式で引用されてしまっているのでないかと推察される。

B.　考察

IV 章の実験の結果から，列挙共引用関係にある論文ペアの方が非列挙共引用関係にある論文ペアよりも類似度が高く，その差も小さくないことが分かった。よって，III 章 A 節で示した仮説「引用箇所間が意味的に近ければ共引用関係が強く，引用箇所間が意味的に遠ければ共引用関係が弱い」は検証されたといえる。このことから，引用論文の本文に依拠して共引用を多値化する手法として，引用箇所間の意味的な近さに基づくことの可能性と有用性が示唆された。

これまで，一つの引用論文から生じる共引用は，すべて同一の類似度を示すものとして利用され続けてきた。同一に扱うことへの批判は理論的な側面からはこれまでもなされてはいたが，大規模な論文集合において個々の共引用をとらえることは難しかったため，実際には検証されてこなかった。

しかし，引用論文を機械処理によって解析することが可能な時代を迎えており，大規模な論文集合においても本文に基づいた共引用をみていくことが不可能でなくなってきている。引用箇所間の意味的な近さを解析して多値の値を持つ共引用としてとらえることで，それまで大まかで粗く算出される類似度であった共引用を，より精密なものへと拡張していくことができる。IV 章の実験結果は共引用の拡張に関して，次の三つの点で意義があるといえる。

一つ目は，共引用関係にある論文間の類似性には強弱があることを数値的に証明し，共引用を多値化できることを示したことである。共引用関係の強弱を，実証的に数値として明らかにした研究はこれまでになく，本稿で初めて明らかになったといえる。

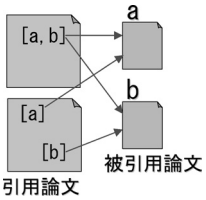
二つ目は，共引用を多値化する手法として，引用箇所間の意味的な近さに基づくことの可能性と有用性を示唆したことである。本文の内容に依拠した共引用を利用することは，引用分析は論文の一部（引用文献リスト）にしか着目していない[38]という批判に積極的に応えることになる。

三つ目は，機械処理による方法を用いたことである。このことは，論文集合が大規模であったとしても，提案手法が適用可能なことを意味する。大規模な論文集合を対象に，本文を解析して得た情報と引用を組み合わせた形で共引用を利用することは，従来の共引用文脈分析では不可能であったものであり，一つの成果と考えられる。

C.　今後の課題

本稿の結果より，引用箇所の位置関係や引用文章中の語の共起関係などから引用箇所間の意味的な近さを調べ，引用箇所間が意味的に近ければその被引用論文間の類似性を強く見積もり，意味的に遠ければ類似性を弱く見積もることが有効であるという知見を得られた。

今後の課題として，引用箇所間の意味的な近さをより細かくとらえていくことがまず挙げられる。本稿では列挙共引用と非列挙共引用のみに着



第 21 図　類似論文検索における類似度の算出

目したが，非列挙共引用の方が数としてはかなり多い。そのため，列挙形式の引用のみに基づいて類似論文検索システムを構築しても，従来の共引用との差がそれほど生じない恐れもある。今後は，引用箇所の位置関係や引用文章の中の語の共起関係を，より細かにとらえる方法を考案していかなければならない。

また，A 節の分析結果より，引用数や引用文に着目することで，引用箇所間の意味的な近さを補正できる可能性があることが分かった。引用数を調べることや，引用文に含まれる “e.g.”（例示するための引用）や “widely”（弱い被引用論文間の関係を示した引用）などの語が含まれているか否かの判別は，機械処理によって扱えるレベルのものである。したがって，このような情報も含めた形で引用箇所間の意味的な近さを解析し，列挙共引用であっても同時に引用している数が多い場合や引用文に特定の語があった場合には，被引用論文間の類似性を弱く見積もるなどの補正手法が実現できると考えられる。このような補正手法の検討も今後の課題である。

さらに，類似論文検索においては，複数の引用論文からみた評価も考えなければならない。類似論文検索においては，第 21 図の論文 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。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列挙共引用されている。このような場合，a と b の間の類似度をどのように計算するのかについても検討していく必要がある。

D. おわりに

ところで，従来の共引用の短所の一つとして，発行されてから一定の時間が経過した論文集合のみが利用対象となることが挙げられる。これは，多数の論文から引用されなければ，各論文間の共引用関係の強弱をとらえられないためである。この短所は，本稿で提案した多値の共引用によって克服できる可能性がある。多値の共引用では，個々の共引用ごとにその強弱の差をとらえていくため，少ない引用論文で論文間の共引用関係の強弱の差を見いだすことができる。したがって，多値の共引用は，即時的な利用という面でも，従来の共引用よりも優れていることが予想される。

即時的に利用できる共引用は，類似論文検索システム以外にも有用であると考えられる。たとえば，共引用マップ[2]がその一つである。対象となる論文があまり引用されていない時点であっても共引用マップを作成することができるため，従来の共引用マップよりも速報性を持たせることができる。速報性が重要視される学問分野においては，このような共引用マップは重要な役割を果たすと思われる。

共引用は，検索だけでなく，共引用マップをはじめとして他の様々な用途に用いられるものである。そのため，共引用を多値なものに拡張することは，それらの用途にも発展をもたらすと想定される。したがって，多値化された共引用を検索以外の用途へどのように適用していくかについても，今後検討すべき事項の一つといえよう。

謝　辞

本稿を執筆する上で，指導していただいた慶應義塾大学文学部の原田隆史准教授，岸田和明教授，田村俊作教授，細野公男名誉教授に感謝いたします。また，同大学理工学部の遠山元道准教授からも貴重なご意見をいただきました。

注・引用文献

1) Salton, G. Automatic indexing using bibliographic citations. Journal of Documentation. 1971, vol. 27, no. 2, p. 98–110.
2) 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. 1973, vol. 24, no. 4, p. 265–269.

3) Chapman, J.; Subramanyam, K. “Cocitation search strategy”. Proceedings of the 2nd National Online Meeting. New York, USA, 1981-03-24/26. Learned Information, 1981, p. 97–102.
4) Knapp, S. D. Cocitation searching: Some useful strategies. Online. 1984, vol. 8, no. 4, p. 43–48.
5) Bichteler, J.; Eaton III, E. A. The combined use of bibliographic coupling and cocitation for document retrieval. Journal of the American Society for Information Science. 1980, vol. 31, no. 4, p. 278–282.
6) Badran, O. M. “An alternative search strategy to improve information retrieval”. Proceedings of the 47th ASIS Annual Meeting. Philadelphia, USA, 1984-10-21/25. Knowledge Industry Publications, 1984, p. 137–140.
7) CiteSeer, http://citeseer.ist.psu.edu/, (accessed 2006-12-31).
8) Beigbeder, M.; Lafouge, T.; Prime-Claverie, C. Transposition of the cocitation method with a view to classifying web pages. Journal of the American Society for Information Science and Technology. 2004, vol. 55, no. 14, p. 1282–1289.
9) Lai, K.; Wu, S. Using the patent co-citation approach to establish a new patent classification system. Information Processing and Management. 2005, vol. 41, no. 2, p. 313–330.
10) Voos, H.; Dagaev, K. S. Are all citations equal? Or, did we Op. Cit. your idem? Journal of Academic Librarianship. 1976, vol. 1, no. 6, p. 19–21.
11) Small, H. Co-citation context analysis and the structure of paradigms. Journal of Documentation. 1980, vol. 36, no. 3, p. 183–196.
12) McBryan, O. A. “GENVL and WWWW: Tools for taming the Web”. Proceedings of the first International World Wide Web Conference. Geneva, Switzerland, 1994-05-25/27. 1994, Elsevier Science B. V., 1994, p. 79–90.
13) O’Connor, J. Citing statements: Computer recognition and use to improve retrieval. Information Processing and Management. 1982, vol. 18, no. 3, p. 125–131.
14) Bradshaw, S. “Reference directed indexing: Redeeming relevance for subject search in Citation Indexes”. Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries. Trondheim, Norway, 2003-08-17/22. Springer-Verlag, 2003, p. 499–510.
15) Ritchie, A.; Teufel, S.; Robertson, S. “How to find better index terms through citations”. Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?. Sydney, Australia, 2006-07-23, Association for Computational Linguistics. 2006, p. 25–32.
16) Schneider, J. W. Concept symbols revisited: Naming clusters by parsing and filtering of noun phrases from citation contexts of concept symbols. Scientometrics. 2006, vol. 68, no. 3, p. 573–593.
17) Preslav, N. I.; Ariel, S. S.; Marti, H. A. “Citances: Citation sentences for semantic analysis of bioscience text”. Proceedings of the SIGIR 2004 Workshop on Search and Discovery in Bioinformatics. Sheffield, UK. 2004-07-29.
18) Cronin, B. The Citation Process: The Role and Significance of Citations in Scientific Communication. Taylor Graham, 1984, 103p.
19) Spiegel-Rösing, I. Bibliometric and content analysis. Social Studies of Science. 1977, vol. 7, no. 1, p. 97–113.
20) Garzone, M.; Mercer, R. E. “Towards an automated citation classifier”. Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence. Montréal,Canada, 2000-05-14/17. Springer-Verlag, 2000, p. 337–346.
21) 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.
22) Teufel, S.; Siddharthan, A.; Tidhar, D. “Automatic classification of citation function”. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, 2006-07-22/23. Association for Computational Linguistics, 2006, p. 103–110.
23) Pham, S. B.; Hoffmann, A. G. “A new approach for scientific citation classification using cue phrases”. Australian Conference on Artificial Intelligence. Perth, Australia, 2003-12-03/05. Springer-Verlag, 2003, p. 759–771.
24) Le, M.; Ho, T. B.; Nakamori, Y. “Detecting citation types using Finite-State Machines”. Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, 2006-04-09/12, Springer-Verlag, p. 265–274.
25) Kessler, M. M. Bibliographic coupling between scientific papers. American Documentation. 1963, vol. 14, no. 1, p. 10–25.
26) 斎藤泰則. 専門領域の重要概念とその相互関係：共引用文脈分析の内容分析に基づく知識構造の

抽出. Library and Information Science. 1986, no. 24, p. 145–154.
27) 大辞林. 第3版. 三省堂, 2006, 2976p.
28) 木下是雄. 理科系の作文技術. 中央公論新社, 1981, 244p.
29) Ruff, I. Citation Analysis of a scientific career: A case study. Social Studies of Science. 1979, vol. 9, no. 1, p. 81–90.
30) 牛澤典子. 被引用文献の概念シンボル化: 医学雑誌論文を事例として. Library and Information Science. 1992, no. 30, p. 133–146.
31) Ruffは「列挙形式の引用は, 引用論文中の導入や理論の部分によく出現する」ということを経験的な示唆として述べており, 牛澤は「列挙形式の引用と単独の引用との間の伝達される内容の違い」について調査している。
32) CiteSeer. PSU OAI, http://citeseer.ist.psu.edu/oai.html, (accessed 2006-12-31).
33) Couto, T.; Cristo, M.; Goncalves, M. A.; Calado, P.; Ziviani, N.; Moura, E.; Ribeiro-Neto, B. "A comparative study of citations and links in document classification". Proceedings of the 6 th ACM/IEEE-CS joint conference on Digital libraries. Chapel Hill, USA, 2006-06-11/15. ACM Press, 2006, p. 75–84.
34) 観点の違いについて, たとえば次のような研究がおこなわれている。
Pao, M. L. Term and citation retrieval: A field study. Information Processing and Management. 1993, vol. 29, no. 1, p. 95–112.
Jarneving, B. A comparison of two bibliometric methods for mapping of the research front. Scientometrics. 2005, vol. 65, no. 2, p. 245–263.
35) Salton, G.; McGill, M.J. Introduction to Modern Information Retrieval. McGraw-Hill, 1983, 448p.
36) Liu, H. MontyLingua. Version 2.1, 2004. http://web.media.mit.edu/hugo/montylingua/, (accessed 2006-12-31).
37) Krampen, G.; Becker, R.; Wahner, U.; Montada, L. On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. Scientometrics. 2007, vol. 71, no. 2, p. 191–202.
38) Callon, M.; Courtial, J.; Turner, W.A.; Bauin, S. From translations to problematic networks: An introduction to co-word analysis. Social Science Information. 1983, vol. 22, no. 2, p. 191–235.

要　　旨

【目的】類似論文検索の代表的な手法の一つに共引用の関係を利用するものがある。この手法では,「引用論文の本文とは無関係に, 一つの引用論文から引用された被引用論文間の類似度は全て同じ」ことが仮定され,「共引用関係にある」「共引用関係にない」の2値情報を基に類似度が算出される。しかし, 引用論文の本文を解析して被引用論文間の関係を詳細にとらえることで, 共引用関係を多値化し, 類似度の算出をより精密にできると考えられる。本稿では, 引用箇所間の意味的な近さに基づいて共引用を多値化する手法を提案し, その可能性と有用性について検討する。

【方法】提案手法を成立させる仮説「引用箇所間が意味的に近ければ共引用関係が強く, 引用箇所間が意味的に遠ければ共引用関係は弱い」の検証をおこなった。大規模論文集合への適用を想定し, 引用箇所間の意味的な近さを引用箇所の位置関係や引用文章中の語の共起関係によってとらえる方法を採用した。そして, この二つの関係が最も強いものとして, 列挙形式の引用（一つの引用で同時に複数の論文を並列列挙する形式をとる引用）による共引用に着目した。仮説の検証はこれを用い, 列挙形式で引用された論文間の類似度とその他の形式で引用された論文間の類似度とを比較する実験によりおこなった。

【結果】列挙形式で引用された被引用論文間の類似度は, それ以外の形式で引用された被引用論文間の類似度よりも高い値になった。このことにより, 提案手法の可能性と有用性が検証された。したがって, 本文を解析して共引用を多値化することで, 類似度の算出をより精密にできることが明らかになった。また, 提案手法が大規模論文集合へ適用可能なことも確認できた。