論文間の参照情報を考慮した関連論文の組織化

難波英嗣[†] 神門典子[††] 奥村学[†††]

本稿では，論文間の参照・被参照関係，および参照の理由を考慮し，関連論文を組織化する手法について述べる．これまで，引用分析研究の分野で，論文間の参照・被参照関係に着目した関連論文を組織化する手法がいくつか提案されてきた．これらの手法はすべての参照を等価に扱っているが，実際には様々な参照の理由が存在するため，既存の手法では必ずしも論文間の類似度を適切に評価できない．そこで，本研究では 2 論文間で同一論文をともに参照しており，かつそれらの参照の理由が一致している結合のみを数えるという方法で，2 論文間の類似度を測る．この手法により，ノイズとなる結合を削減でき，また，従来の引用分析手法と比べ，精度の向上が期待できる．提案手法の有効性を調べるために，実験を行った．実験では，提案手法，引用分析の代表的な手法である書誌結合，語の共出現を用いたより一般的な組織化の手法（ベクトル空間型モデル）を，精度，フォールアウト，計算コストという 3 つの側面から比較した．その結果，提案手法が精度，フォールアウトによる評価で最も優れ，また，計算コストの面でも十分に速い速度で論文を組織化できることが分かった．

Classification of Research Papers Using Citation Links and Citation Types

Hidetsugu Nanba,[†] Noriko Kando[††] and Manabu Okumura[†††]

In this paper, we propose a method for classification of research papers using citation links and citation types that indicate the reasons for citations. Several methods has been proposed for classification of papers using citation links in citation analysis. However, most of them treats all citations equally. We therefore refine citation analysis by taking account of citation types. Our method measures similarity between papers by counting the couplings of the same citation types. We compared our method with bibliographic coupling that is a kind of citation analysis and some word-based approaches (vector space model) using precision, fallout, and computational cost. The results of our experiments showed that our method is more effective than other methods.

1. はじめに

近年，学術情報の爆発的な増加とともに，数多くの電子化された論文がオンラインから入手できるようになった[1),2)]．これらの論文の書誌情報を抽出・蓄積すれば，論文データベースとして論文検索が可能になるが，さらに，論文の内容に基づいてあらかじめトピックごとに分類・整理しておけば，ユーザは必要な論文を効率的に入手できる．本稿では，論文を分類・整理することを関連論文の組織化と呼んでおり，本研究ではトピックごとの関連論文の組織化を目指す．

これまで，クラスタリングやカテゴリゼーションの研究分野で，文書を組織化する様々な手法が提案されてきた[3)]．その中心的な手法は，文書中の語，文書に付与されたディスクリプタ（キーワード）等を用いて，個々の文書をベクトル空間型モデル，確率モデル等で内部表現に変換し，この内部表現により文書間の類似度を測る．

一方，学術論文には論文間に参照・被参照関係☆があり，論文の組織化にはこのような参照構造が利用できる．これまで引用分析研究の分野において，論文間

† 日本学術振興会特別研究員
Research Fellow of the Japan Society for the Promotion of Science
†† 国立情報学研究所
National Institute of Informatics
††† 東京工業大学精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology

☆ 「参照」と「引用」という言葉の使い分けについて Narin[4)] は次のように述べている．「参照という用語は，それから出発して他へ向かう構成ユニットを示すために用いられるのに対し，引用という用語は，他から受ける構成ユニットをさすのに用いられる」．しかし，実際には同じような意味で用いられることが多い．本稿では混乱を避けるため，すでに専門用語として定着している「引用分析」を除き，一貫して「参照」を用いる．

の参照構造を利用し，2 論文間の類似度を測るいくつかの手法が提案されてきた[5),6)]．しかし，これらの手法はすべての参照・被参照関係を等価に扱っているが，実際には Weinstock[7)] や Moravcsik[8)] が述べているように参照には様々な理由が存在するため，必ずしも論文間の類似度を適切に評価できない．

そこで，本研究では被参照論文の参照の理由を考慮し，参照構造を用いて論文間の類似度を測る手法を提案する．難波ら[9)] は，論文中の参照の文脈を解析し，論文間の参照・被参照関係を参照の理由（以後，参照タイプ）を自動的に判定する手法を開発している．本研究では，2 論文が同一論文をともに参照している書誌結合関係にある場合，難波らの手法を用い，それらの参照タイプが一致している結合のみを数えるという方法で，書誌結合の改良を行う．提案手法により，ノイズとなる結合を削減でき，また，従来の引用分析手法と比べ，精度の向上が期待できる．

本研究では，提案手法の有効性を調べるために，実験を行う．提案手法の他に引用分析手法の代表的な手法である書誌結合，語の共出現に基づいた手法（ベクトル空間型モデル）を計算機上に実装し，各手法を精度，フォールアウト，計算コストの 3 つの尺度で比較し，提案手法の有効性を調べる．

本稿の構成は以下のとおりである．次章では関連研究を紹介し，その問題点について述べる．また，本研究で提案する関連論文の組織化手法を説明する．3 章では，実験の手順と評価について述べ，また結果について考察する．

2. 関 連 研 究

これまで，学術論文をトピックごとに組織化するいくつかの手法が提案されてきたが，それらは大きく以下の 2 つに分けることができる．

- **アプローチ 1：（引用分析に基づく組織化手法）**
  引用分析とは，論文間の参照・被参照関係を用いて，論文間の関係を分析する方法である．書誌結合（bibliographic coupling）[5)] と共引用分析（co-citation analysis）[6)] は，引用分析の代表的な手法であり，トピックの似た論文を検索できることが知られている[10)~12)]．書誌結合は，論文間の関連度を測るときに，2 論文間でどれだけ同じ論文を引用しているか，という基準に基づいている．一方，共引用分析は，2 論文がどれだけ他の論文でともに引用されているか，という基準に基づいた手法である．
  したがって，発表されてから十分に時間がたっている古い論文を対象にする場合は共引用分析が適しているといえる．これとは逆に，他の多くの論文から引用されていないような新しい論文を組織化するには，書誌結合の方が適している．
  また，難波ら[9)] は，ある論文を参照する複数の論文を，参照の理由（以後，参照タイプ）ごとに自動的に組織化する手法を開発している．難波らは，参照論文中で被参照論文について記述されている個所（以後，参照個所）を自動的に抽出し，さらにその個所を解析することで，以下に示す 3 種類の参照タイプに分類している．
  - **type B（論説根拠型）**
    新しい理論を提唱したり，システムを構築したりする場合，他の研究者の研究の成果を利用する場合がある．たとえば，他の研究者が提唱する理論や手法を用いて新しい理論を提唱する場合等である．このような参照タイプを type B（論説根拠型）と呼ぶ．
  - **type C（問題点指摘型）**
    新しく提案した理論や，構築したシステムの新規性について述べる場合，関連研究との比較，あるいは既存研究の問題点の指摘を行う場合がある．このような目的の参照タイプを type C（問題点指摘型）と呼ぶ．
  - **type O（その他型）**
    type B にも type C にもあてはまらない参照を type O（その他型）と呼ぶ．
- **アプローチ 2：（語の共出現に基づく組織化手法）**
  トピックの似た 2 つの論文間では，多くの語が論文間で共出現する傾向にある．このような共出現する語の数を数えることで論文間の類似度を測る[13)]．

ここで，これらの 2 つのアプローチの問題点を以下に示す．

- **アプローチ 1 の問題点：（引用分析に基づく組織化手法）**
  引用分析に関する大半の研究は，すべての引用を等価に扱っている．しかし，実際は Weinstock[7)] や Moravcsik ら[8)] が示すような様々な参照の理由が存在する．したがって，関連論文をより正確に組織化するためには単純な参照・被参照関係だけでなく，より豊富な参照情報を考慮することが不可欠であると考えられる．
  また，難波らの手法では，直接参照・被参照関係にある論文を組織化することは可能であるが，2 論文が参照・被参照関係にない場合は組織化でき

ない．

- **アプローチ 2 の問題点：（語の共出現に基づく組織化手法）**
  2 論文全体にわたって共出現する語を調べるのは，非常に時間がかかる．したがって，探索時間を削減するために，論文の探索対象を減らす必要がある．

次に，アプローチ 2 の改善手法について述べる．

- **アプローチ 2 の改善手法：（語の共出現に基づく組織化手法）**
  アプローチ 2 の問題点に関する取組みとして，これまで以下のような研究がある．
  Kando は[14]，表層的な言語情報を表す単語の並び（手がかり語）を用いたいくつかのルールにより，日本語論文の構造解析を行っている．ここで述べる手がかり語とは，論文の意味的な構造を示す特徴的な語句のことであり，文間の連接関係を表すもの（「しかし」等），語彙的な手がかり（「本稿では」等），文末表現（「であろう」等）等がある．解析の結果，論文中の各文には意味役割を示すラベルが付与される．そして，特定の意味役割の文（Method and Validity）のみを用いて論文検索を行えば，論文全文を用いた検索と比べ精度が向上し，再現率の低下も最小限にとどめられることを示している．
  三池ら[15] も，Kando と同様に，手がかり語に基づくいくつかのルールにより，日本語技術論文の構造解析を行っている．三池らは「背景」,「話題」,「従来の問題」,「目的」,「特徴」,「結果」,「結論」,「課題」を示す文のみを用いて検索することで，論文全文を検索に用いる場合よりも検索精度が向上することを，実験により示している．
  Kando や三池らの研究で提案されているように，あらかじめ，論文の特徴的な内容を表す文集合（以後，パッセージ）を抽出し，このパッセージを用いて語の共出現を数えれば，アプローチ 2 による計算時間が短縮されると考えられる．
  Kando の研究では，研究の手法について記述された文が検索に有効であるという結果が得られているが，三池らの研究では研究の背景や目的に関する記述が有効であると述べている．そこで，本研究では次節で述べる提案手法の比較手法として以下の 2 つの意味役割を考慮して文の抽出を行い，これらを論文の組織化に利用する．
  - **PURPOSE：**
    研究の目的が書かれてある個所は，研究の分野と密接な関連があると考えられる．
  - **METHOD：**
    研究の背景で用いている理論や手法も，トピックごとの組織化を行う際に有用な指針になりうると思われる．

本研究では，これらの意味役割の文抽出は，以下に述べる手法で “PURPOSE” や “METHOD” の文を抽出する．

本研究では，4 章で説明する E-Print archive という英語論文データベースの論文データを対象にしている．“PURPOSE” の抽出には 5 個の手がかり語 “our work”，“Our work”，“this paper”，“This paper”，“purpose” を用いる．これまで，学術論文の構造解析に関する研究がいくつか行われてきたが[14),16)]，いずれもこれらの 5 句が共通して，論文の研究の目的に関する記述個所を抽出する際の手がかり語として用いられている．そこで，本研究でもこれらの 5 つを含む文を “PURPOSE” として抽出する．

また，“METHOD” の抽出には 84 個の手がかり語を用い（**表 1** にその一部を示す），これらを含む文を抽出する．一般的に，研究で用いる理論やツールは “Introduction” や “Experiment” といった章で述べられることが多い．したがって，“METHOD” の抽出は，“Introduction” や “Experiment” といった語を含んだ章を抜き出せばよいと考えられる．しかし，E-Print archive 上の論文 293 本で調べた結果，“Introduction” を含む論文は全体の 83%（242/293）程度存在したが，“Experiment” は全体の 16%（47/293）しか存在しなかった．そこで，“Introduction” と “Experiment” の章に頻出する特徴的な語句（表現）のリストを作成し，“Introduction” や “Experiment” の章が存在しない論文からリスト中の語句を含んだ文を抽出すれば，それらが “Introduction” や “Experiment” の章と同等の役割をすると考えた．まず，E-Print archive 上の論文の中から “Introduction”（242 論文中）と “Experiment”（47 論文中）を収集し，次にこれらの中から，cost criteria[17] という統計的な手法を用いて，手がかり語の候補を 5,000 語自動的に抽出した．最後に，このリストの中から “METHOD” に関連した表現を人手で 84 個選択した．論文中でこれらの 84 個の手がかり語を含む文を “METHOD” として抽出する．

なお，同一文内に “PURPOSE” 用の手がかり語と “METHOD” 用の手がかり語がともに出現す

表 1 “METHOD” の抽出に用いる手がかり語の例
Table 1 Examples of cue phrases for the extraction of “METHOD”.

based mainly on	based on ... in	is based on
the basic	employ	invoke
assume	underlie	underlain
can use	used as a	by using
Using the	is checked	we use
we used	result	make use of
Making use of	advantage of	we introduce
is given in	are given in	offer
we ... influence	assume	is needed to
are needed to	been given	a given
available for	applied to	application to
We adopted	extend the	we extended
extended to	expands	expanded

る可能性も考えらえるが，このような場合には，その文は “PURPOSE” と “METHOD” の両方に含めている．

ここで提案する手法は，Kando や三池らの手法と比較すると文書構造解析を行っているわけではないので，Kando や三池らの手法と同等の検索精度は望めないが，提案手法と比較するうえで，ある程度の目安にはなると考えている．

次章では，アプローチ 1（引用分析に基づく組織化手法）の問題点を改善する関連論文組織化の手法を提案する．

3. 関連論文の組織化の手法

本研究では，難波らの提案する参照タイプを関連論文の組織化に利用する．引用分析手法として，共引用分析ではなく書誌結合を用いる．なぜならば，本研究で実験に用いる論文データベース（詳細は次章で述べる）は比較的新しい論文から構成されており（1994 年〜1998 年の期間の論文），論文集合全体の被参照数が小さく，共引用分析には向かないからである．

本研究では，2 論文間で同一論文をともに参照しており，かつそれらの参照の理由が一致している場合のみ数えるという方法（以後，BCCT: Bibliographic Coupling using Citation Types）で，2 論文間の類似度を測る．

実験では “BCCT-C” と “BCCT-BCO” という 2 種類の手法を用いる．“BCCT-C” は論文が同一論文をともに参照している書誌結合にあり，かつ，それらが参照タイプ “C” のみで一致している結合のみを数えるという方法である．type C に着目した理由は，type C は，参照論文の問題提起や研究動機を示す参照と考えることができ，2 論文間で多くの type C の参照が一致すれば，これらの論文の著者は共通の問題意識を持っていると考えられるからである．

“BCCT-BCO” は，2 論文が同一論文をともに参照している書誌結合にあり，かつ，それらの参照タイプが B，C，O のいずれかで一致している結合のみを数えるという方法である．“BCCT-BCO” は，2 論文が共通に参照する論文が存在しても，異なる参照タイプで参照していれば類似度に反映しないという点で，従来の書誌結合と異なる．

4. 関連論文の組織化手法の評価

4.1 評価方法

論文集合

前章で述べた提案手法の有効性を調べるために実験を行った．近年，NTCIR☆，OHSUMED☆☆等，大規模な論文検索のテスト・コレクションが作成されている．しかし，これらは論文抄録を検索対象にしており，学術論文全文を検索対象にしたテスト・コレクションはこれまで一般的には公開されていない．そこで本研究では，既存のテスト・コレクションに比べると小規模ではあるが，E-Print archive の “The Computation and Language”☆☆☆に関する TEX 形式の 395 論文のデータを用いる．

この論文データベースは，大部分が 1994 年以降発表された国際会議の論文，博士論文，リサーチレポート，ジャーナル論文等，様々な種類の論文から構成されている．また，E-Print archive は，論文の著者が自発的にデータベースに論文データを登録する形式をとっており，誤ったカテゴリに論文が登録されても，登録した著者本人が気づかない限り，論文が第三者によって削除されることはない．“The Computation and Language” という分野には，一般的に自然言語処理や計算言語学と呼ばれる研究分野の論文を含むと考えられるが，プログラミング言語やコンパイラに関連する論文も若干含まれていた．これらの論文を削除したうえで実験を行うことも考えられたが，検索システムがこのような論文を他の論文と区別することも重要であると考え，実験の対象から削除しなかった．

また，E-Print archive からは，論文間の参照・被参照関係の情報が得られないが，難波ら[9] は，TEX で参考文献を書くために用いられる bibliography というコマンドを自動的に解析し，論文間の参照・被参照関係を明らかにしている．本研究では，これらのデータを実験に用いる．

☆ NTCIR: http://research.nii.ac.jp/ntcir/
☆☆ OHSUMED: http://medir.ohsu.edu/pub/ohsumed/
☆☆☆ E-Print archive: http://xxx.lanl.gov/cmp-lg/

正解セットと検索クエリ

関連論文の組織化システムを評価するために，395論文を用いて正解データセットを作成した．

まず，1つの論文が1つのカテゴリにだけ属するように395論文を分類した．その結果，395論文は58カテゴリに分類された．また，58カテゴリのうち，1カテゴリ中に6件以上の論文を含んでいるものが10カテゴリあった．この10カテゴリは「形態素解析」「構文解析」「意味解析」といった “The Computation and Language” の研究分野における典型的なカテゴリである．他の48カテゴリは，これらの10カテゴリに含まれない，新しい分野の論文であるか，あるいは先にも述べたような，“The Computation and Language” の分野外と考えられる論文のカテゴリである．

次に，実験方法について説明する．まず，10カテゴリに含まれる論文（計330論文）から任意に1論文を選択する．次にこれを検索クエリと見なし，論文集合から検索クエリと同一カテゴリの論文を検索することを試みる．検索システムは入力クエリに関する論文を検索し，クエリに対して適合度の高い順に検索結果として論文の一覧を返す．このような手順を330回繰り返し，330論文それぞれについて関連論文を検索する．これらと人手による組織化との結果を比較し，検索システムの性能を評価する．

実験に，検索クエリとして10カテゴリに含まれる論文しか用いない理由を以下に説明する．一般的に統計的検定を行うには，サンプルが30件以上，当該事象が5回程度以上起こる必要があるといわれている．したがって，情報検索の評価実験の場合，検索課題（query）が30件以上，正解文書が5件以上必要となる[18),19)]．本研究でも正解が5論文以上，すなわち1カテゴリ中に6論文以上（クエリ1論文 + 正解5論文）のものを実験に用いる．

検索エンジン

ベクトル空間型モデルを用いて，検索エンジンを作成した．提案システムは，Brillの品詞タギングツール[20)]を用い，パッセージから名詞のみを抽出しインデックスを作成する．次にコサイン距離で論文間の類似度を計算する．

組織化手法

実験は2種類の提案手法と，比較対象として6種類の手法を用いて行う．語の共出現を用いる手法は，2章で説明した “PURPOSE” と “METHOD” を用いる．また，これらの2つの手法のほかに，論文表題（“TITLE”）や概要（“ABST”）を加えた．これらは論文の著者により作成された，論文の特徴を表すパッセージと考えることができる．また，各パッセージがどの程度，原論文の内容を反映しているのかを調べるため，論文全文（“FULL”）を用いた組織化も行う．

また，“BCCT-C” および “BCCT-BCO” では，参照タイプは難波らの手法により自動的に判定されたものを用いる．

- **“FULL”，“TITLE”，“ABST”：**
  論文全文，タイトル語と概要中の語を用いた語の共出現．
- **“METHOD”，“PURPOSE”：**
  手がかり語により抽出された文中に含まれる語を用いた語の共出現．
- **“NBC”：**
  書誌結合．
- **“BCCT-C”，“BCCT-BCO”（提案手法）：**
  同一論文をともに参照している書誌結合にあり，かつ，それらの参照タイプがCのみで一致している結合のみを数える場合（BCCT-C）と，参照タイプがB，C，Oのいずれかで一致している結合のみを数える場合（BCCT-BCO）．

4.2 評　　価

以下の評価尺度を用いて8種類の組織化手法の有効性を調べた．

- 上位n論文の精度
- フォールアウト
- 計算コスト

再現率 – 精度は，情報検索の分野では最も一般的に用いられている評価尺度である．この尺度は検索エンジンの有効性の全体的なバランスを見るうえでは良い指針となる．しかし，実験で用いる8種類の組織化手法では，収集されるそれぞれの論文数が極端に違う．一般に，情報検索において十分な数の文書（たとえばTREC2の場合，正解文書数の数倍程度）が収集されていなければ，再現率レベルの精度値は正しく計算されず，また再現率 – 精度グラフも正しく書けないといわれている[21)]．したがって，再現率による評価は必要ではあるが，本研究では今述べた理由により評価の対象にはしない．本研究ではフォールアウトと精度で8手法の組織化精度の比較を行う．

また，本研究では8手法の組織化手法を計算コストの面から比較する．語の共出現に基づく手法において，一般に，手がかり語の語数が異なれば抽出される文の数も異なり，それに比例して計算コストも高くなると推測される．しかし，これまで語の共出現に基づく手法と引用分析に基づく手法を計算コストの面から比較した研究はない．したがって，この比較による知見を

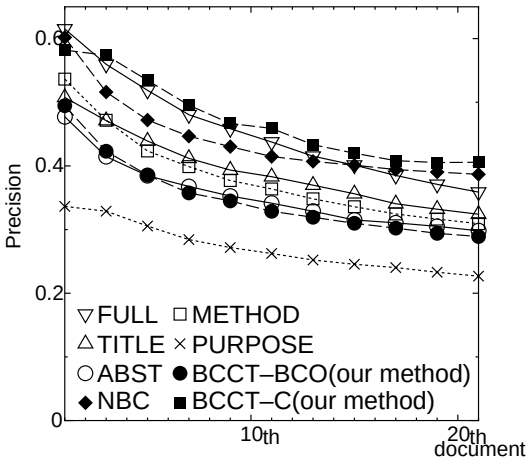

図 1 上位 n 論文の精度の比較
Fig. 1 Precision for each ranking.

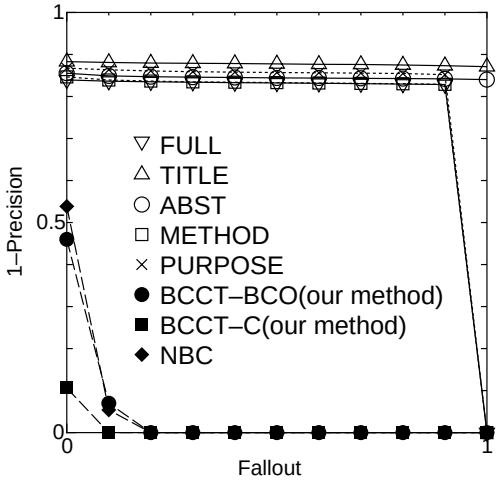

図 2 フォールアウトと精度による組織化手法の比較
Fig. 2 Evaluation by fallout and precision.

得ることも有用であると考えられる．

精度とフォールアウトは次の式で与えられる．

$$\text{精度}\ (Precision) = \frac{\left(\begin{array}{l}\text{検索システムにより}\\ \text{検索された論文の中で}\\ \text{正解の論文数}\end{array}\right)}{\left(\begin{array}{l}\text{検索システムにより}\\ \text{検索された論文総数}\end{array}\right)}$$

$$\text{フォールアウト}\ (Fallout) = \frac{\left(\begin{array}{l}\text{検索システムにより}\\ \text{検索された論文の中}\\ \text{で不正解の論文数}\end{array}\right)}{\left(\begin{array}{l}\text{クエリと異なる}\\ \text{カテゴリの総論文数}\end{array}\right)}$$

フォールアウトは検索エンジンのエラーを測る尺度であり，フォールアウト値が小さいほど良いシステムであるといえる．精度とフォールアウトを算出する際，TREC で使われている trec_eval というツール☆を利用した．通常は，trec_eval に正解文書セットと検索システムの出力結果を与えると，再現率が（0 %, 10 %, 20 %, ..., 100 %）の 11 点における精度が計算される．ここで，正解文書セットの代わりに正解と不正解を反転させた文書セットを trec_eval に与えると，フォールアウトが（0 %, 10 %, 20 %, ..., 100 %）の 11 点における 1-精度の値が計算される．

上位 n 文書の精度による評価の結果を**図 1** に，フォールアウトによる評価の結果を**図 2** に，計算コストによる評価の結果を**表 2** に，それぞれ示す．

4.3 考　　察

上位 n 論文の精度による評価

“FULL” と “NBC” が論文全体の情報を用いた組織化手法，その他が論文の一部の情報を用いた組織化手法であると考えれば，後者の手法で前者の精度を上回っているのは “BCCT-C” だけである．すなわち，“BCCT-C” が論文の中でも特にトピックに関連する情報を高い精度で抽出できていると考えることができる☆☆．

表 2 計算コストによる比較（クエリあたり）
Table 2 Calculation time of eight methods (per query).

手法	計算時間（秒）
FULL	232
TITLE	0.25
ABST	1.2
METHOD	8.1
PURPOSE	0.77
BCCT-C（提案手法）	1.3
BCCT-BCO（提案手法）	14
NBC	14

論文間の参照関係に基づく手法（“BCCT-C”, “BCCT-BCO”）で組織化に失敗した原因は，以下の 3 種類に分類できる．

(1) 参照タイプの決定の誤りによる失敗
(2) type C の参照の一致による失敗
(3) type B の参照の一致による失敗

(1) 参照タイプの決定の誤りによる失敗

難波ら[9] の実験結果によれば，評価用データにおいて参照タイプ決定ルールで type C と判定された 15 個のうち，実際に正解であるものの数は 12 個（精度：80%）である．“BCCT-C” は，2 つの論文が別の論文をともに type C で参照するときの結合の数

☆ ftp://ftp.cs.cornell.edu/pub/smart

☆☆ 今回実験に用いたテスト・コレクションでは，1 論文あたり平均約 18.1 論文を参照している．また，type C での参照は 1 論文あたり約 2.9 論文となっている．

（Scheler, 1996）[22] の（Church, 1988）[24] に関する type C の参照個所
「冠詞を含む名詞句の意味素性の解析，生成，文法チェック」
The different logical forms of the sentences can be represented by a set of sentential operators, which are defined in first-order logic. These sentential operators can be used as atomic semantic features, which are consequently sufficient in representing the logical meaning of a sentence with respect to the chosen semantic dimensions. This approach is significantly different from POS or sense-tagging systems such as (Yarowsky, 1992) (Schmid, 1994) (Jelinek, 1985) **(Church, 1988)** (Brill, 1993).

（Heeman, 1997）[23] の（Church, 1988）[24] に関する type C の参照個所
「品詞タギングと言語モデルの結合」
The final probability distributions are similar to those used for POS tagging of written text (DeRose, 1988:cl) **(Church, 1988)**. However, these approaches simplify the probability distributions as is done by previous attempts to use POS tags in speech recognition language models.

図 3 “BCCT-C” の失敗例
Fig. 3 Example of failure by “BCCT-C”.

を数えるという手法であり，直接 type C の決定精度の影響を受ける．難波らの type C の決定精度をそのまま適用すれば☆“BCCT-C” そのものの精度は，80%× 80%=64%となる．64%という値は概算であり厳密な数字ではないが，図 1 において上位 1 論文における “BCCT-C” の精度が 60%弱であることから，“BCCT-C” における失敗の原因の多くは，難波らの手法による参照タイプの判定の誤りに関するものと推測される．実際に，“BCCT-C” で失敗した事例をいくつか調べたところ，type C ではないものが type C と判定されたために間違って検索されたものが多かった．したがって，参照タイプ決定精度が向上すればそれにともない “BCCT-C” の精度も向上すると考えられるが，図 1 から分かるように，現状でも，他の手法と比較し十分な精度が得られている．

(2) type C の参照の一致による失敗

“BCCT-C” に関して，2 論文が同一論文を type C で参照していても，被参照論文について述べているポイントがずれている場合，異なる分野の論文を検索する場合があった．**図 3** に “BCCT-C” の失敗例を示す．

図 3 は，Scheler[22] と Heeman ら[23] の Church[24] に関する type C の参照個所を示したものである．（Scheler, 1996）と（Heeman, 1997）はそれぞれ，「冠詞を含む名詞句の意味素性の解析，生成，文法チェック」と「品詞タギングと言語モデルの結合」に関する論文である．

Scheler の研究では，名詞句を分類する 5 つの観点（“Generalized quantification”，“Anaphoric relation”，“Reference to discourse objects”，“Boundedness”，“Active involvement”）を設定し，名詞句を自動的に分類する手法を提案している．（Scheler, 1996）の参照個所中の記述（図 3 上）によれば，これらの観点は一階述語論理の形式で定義されているが，同時に文オペレータという異なる形式でも表されており，この文オペレータを意味素性の代わりとして用いる点が（Church, 1988）をはじめとする品詞あるいは意味タグを付与するシステムと異なる，と述べている．

この記述は，参照論文（Scheler, 1996）と被参照論文（Church, 1988）との思想的な違いを述べているので，type C の参照であると考えられるが，（Church, 1998）の問題点を明示的に述べているわけではない．

一方，（Heeman, 1997）では（図 3 下），（Church, 1988）について，「音声認識における品詞タグの取扱いのときと同様，（Church らの研究では）確率分布を単純化しすぎている」と（Church, 1988）の問題点を明示的に述べている．

このように，（Scheler, 1997）と（Heeman, 1997）では，同じ type C でも（Church, 1988）の言及の仕方がまったく異なっており，このような場合，“BCCT-C” では失敗している．

(3) type B の参照の一致による失敗

図 1 において，“BCCT-BCO” は “NBC” よりも精度が低かった．失敗の原因を調査したところ，type B の（書誌）結合が論文をトピックごとに分類する際あまり有効ではなく，また場合によっては組織化を阻害する方向に作用することが分かった．たとえば，“The Computation and Language” の分野において，形態素解析器や構文解析器等のツールは多くの研究で汎用的に使われれる．したがって，2 つの論文がこのようなツールについて書かれた論文をともに type B で参照していても，トピックごとの論文の組織化には有用ではない．

参照タイプを考慮した書誌結合の手法として，“BCCT-C” や “BCCT-BCO” のほかにも “BCCT-B” という手法も事前に考えられた．しかし，予備調

☆ 本研究でも難波らの研究と同じ論文データベース（E-Print archive）を用いているので，本稿における実験においても難波らのものとほぼ同等の精度が得られていると考えられる．

査の段階で，先に述べた理由により type B の（書誌）結合がトピックごとの組織化に向かないことが判明したため，比較手法に “BCCT-B” を入れなかった．しかし，実際には “BCCT-B” ばかりでなく，“BCCT-BCO” においても type B の結合が，その精度を下げる要因になった．

以下では，語の共出現に基づく組織化手法の結果について考察する．あらゆる手法の中で “PURPOSE” の精度が一番低かった．論文から抽出される文数が少なかった☆というのが，その理由の 1 つとしてあげられる．また，“PURPOSE” に含まれる語は，論文の内容をよく表している場合もある．しかし，多くの場合抽象的すぎるか，論文の非常に具体的な記述で，代表語として適切でないものが多く含まれていた．

同様の結果が Kando により報告されている[14]．Kando は，“Method and Validity” と “Evidences” という意味役割が振られた文（本稿の “METHOD” に相当する）と “Research Topic” の意味役割の文（本稿の “PURPOSE” に相当する）を用いて検索を行った結果，どの文書も “Research Topic” の意味役割がふられた文が 1 文以上存在していたにもかかわらず，“Method and Validity” と “Evidences” を用いた解析精度が “Research Topic” の解析精度を上回ると報告している．Kando とは，テスト・コレクションおよび実験条件が異なるので，厳密に比較することはできないが，本研究における “PURPOSE” と “METHOD” はそれぞれ，Kando の “Method and Validity” や “Evidence” とある程度似た傾向のパッセージが抽出されていると考えられる．

フォールアウトによる評価

図 2 において，書誌結合に基づく 3 つの手法（“BCCT-C”，“BCCT-BCO”，“NBC”）でいずれも良い結果が得られている．3 つの中でも特に “BCCT-C” が一番優れている．これは，あらゆる参照の理由の中で，type C が関連論文を検索するうえでは重要な参照の理由であることを示している．また，図 1 では，“NBC” と “TITLE” はほぼ同程度の精度が得られていたが，システム全体で比較した場合，“NBC” の方が “TITLE” よりもトピックの異なる論文を収集しない，という面で優れているといえる．

語の共出現に基づく手法のフォールアウト値が高い理由は，書誌結合に基づく手法に比べ，検索システムが収集する論文数が多いからである．すなわち，関連論文を漏れなく検索するには語の共出現に基づく手法が適しているが，なるべく高い精度で関連論文を検索する場合には書誌結合に基づく手法，特に “BCCT-C” が適しているといえる．

計算コストによる比較

最後に，8 種類の組織化手法の計算コストを比較した．表 2 において，計算時間はクエリごとにトピックの類似度を計算するのに要した時間である．これには品詞タギングや “METHOD”，“PURPOSE” の文抽出に要した時間は含まれていない．

8 手法の中で上位 n 論文の精度（図 1）では “FULL” と “NBC” は比較的良い精度が得られていたが，計算コストの面では “FULL” と “NBC” が最も遅かった．

以上をまとめると，より高い精度でかつ妥当な計算コストで関連論文を検索するためには提案手法である “BCCT-C” が最も適しているといえる．

5. ま と め

本稿では，学術論文をトピックごとに組織化するために，論文間の参照・被参照関係および参照タイプ（参照の理由）を考慮した “BCCT-C” という手法を考案し，「収集された論文の上位 n 件の精度」および「フォールアウト」による評価で，最も優れていることが分かった．また，計算コストの面でも “BCCT-C” は今回比較した 8 手法の中で最速ではなかったものの，十分に速い速度で論文を組織化できることが分かった．これらは小規模なテスト・コレクションにおける実験結果なので，この結果だけから確定的なことを述べることはできないが，有望な方向が示されたと考えられる．

今後は，さらに 2 つの方向で提案手法 “BCCT-C” の有効性を調べる必要がある．1 つは，論文の組織化のスケールの問題である．すなわち，今回設定したカテゴリよりもさらに細かい，あるいは粗いカテゴリで論文を組織化するのに，提案手法がどの程度有効であるかを調査する必要がある．

もう 1 つは，type C 以外の参照タイプの有効性に関する調査である．本研究の結果では，type B の参照は組織化精度を下げる要因になった．しかし，論文中で形態素解析器や構文解析器に関する論文を参照していることは，その論文が “The Computation and Language” の分野の論文であるかどうかを判断するうえでは，逆に重要な情報になると考えられる．これは，先に述べたカテゴリのスケールの問題と関連するが，組織化するカテゴリのスケールに応じて，有効な参照タイプの組合せが変わってくるのではないかと推測される．

謝辞 論文データを提供していただいた E-Print

☆ PURPOSE：平均 4.9 文，METHOD：平均 31.7 文

archive administrator の方々に感謝いたします.

参 考 文 献

1) Lawrence, S., Giles, L. and Bollacker, K.: Digital Libraries and Autonomous Citation Indexing, *IEEE Computer*, Vol.32, No.6, pp.67–71 (1999).
2) McCallum, A., Nigam, K., Rennie, J. and Seymore, K.: A Machine Learning Approach to Building Domain-Specific Search Engines, *Proc. 16th International Joint Conference on Artificial Intelligence* (*IJCAI-99*), pp.662–667 (1999).
3) 永田昌明，平博　順：テキスト分類 — 学習理論の「見本市」，情報処理，Vol.42, No.1, pp.32–37 (2001).
4) Narin, F., Gabriel, P. and Hofer, G.H.: Structure of the Biomedical Literature, *Journal of the American Society for Information Science*, Vol.27, No.1, pp.25–45 (1976). 神門典子（訳）：生物医学文献の構造，情報学基本論文集 I, pp.189–227, 勁草書房 (1989).
5) Kessler, M.M.: Bibliographic Coupling between Scientific Papers, *American Documentation*, Vol.14, No.1, pp.10–25 (1963).
6) 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, pp.265–269 (1973).
7) Weinstock, N.: Citation indexes, *Encyclopedia of Library and Information Science*, Kent A. (Ed.), Vol.5, pp.16–41, New York, Marcel Dekker (1971).
8) Moravcsik, M.J. and Murugesan, P.: Some Results on the Function and Quality of Citations, *Social Studies of Science*, Vol.5, pp.86–92 (1975).
9) 難波英嗣，奥村　学：論文間の参照情報を考慮したサーベイ論文作成支援システムの開発，自然言語処理，Vol.6, No.5, pp.43–62 (1999).
10) Liu, M.: Progress in Documentation—The Complexities of Citation Practice: A Review of Citation Studies, *Journal of Documentation*, Vol.49, No.4, pp.370–409 (1993).
11) Narin, F., Olivastro, D. and Stevens, K.A.: Bibliometrics/Theory, Practice and Problems, *Evaluation Review*, Vol.18, No.1, pp.65–76 (1994).
12) White, H.D. and Mccain, K.W.: Bibliometrics, *Annual Review of Information Science and Technology* (*ARIST*), Vol.24, pp.119–186 (1989).
13) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, p.448, New York, McGraw-Hill (1983).
14) Kando, N.: Text-level Structure: Implications for Information Retrieval and the Potential for Genre Analysis, *British Computer Society IR SG Annual Colloquium*, pp.68–81 (1997).
15) 三池誠司，住田一男：文書の意味役割解析に基づく全文検索，情報処理学会研究報告—情報学基礎，FI-34-3, pp.17–24 (1994).
16) Teufel, S.: Argumentative Zoning: Information Extraction from Scientific Text, Ph.D. Thesis, University of Edinburgh (1999).
17) Kita, K., Kato, Y.,Omoto, T. and Yano, Y.: A Comparative Study of Automatic Extraction of Collocation from Corpora: Mutual Information vs. Cost Criteria, *Journal of Natural Language Processing*, Vol.1, No.1, pp.21–33 (1994).
18) Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H. and Hidaka, S.: Overview of IR tasks at the First NTCIR Workshop, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.11–44 (1999).
19) Sakai, T., Kitani, T., Ogawa, Y., Ishikawa, T., Kimoto, H., Keshi, I., Toyoura, J., Fukushima, T., Matsui, K., Ueda, Y., Tokunaga, T., Tsuruoka, H., Nakawatase, H., Agata, T. and Kando, N.: BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems, *SIGIR-Forum*, Vol.33, No.1, pp.13–17 (1999).
20) Brill, E.: Some Advances in Rule-based Part of Speech Tagging, *Proc. 12th National Conference on Artificial Intelligence* (*AAAI-94*), pp.722–727 (1994).
21) Harman, D.: Overview of the Second Text Retrieval Conference (TREC-2), *Information Processing & Management*, Vol.31, No.3, pp.271–289 (1995).
22) Scheler, G.: *With Raised Eyebrows or the Eyebrows Raised? A Neural Network Approach to Grammar Checking for Definiteness*, FKI-215-96 (1996).
http://xxx.lanl.gov/ps/cmp-lg/9606017
23) Heeman, P.A. and Allen, J.F.: Incorporating POS Tagging into Language Modeling, *Proc. Eurospeech'97* (1997).
http://xxx.lanl.gov/ps/cmp-lg/9705014
24) Church, K.: A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, *Proc. 2nd Conference on Applied Natural Language Processing*, pp.136–143 (1988).

(平成 13 年 1 月 29 日受付)
(平成 13 年 9 月 12 日採録)

難波 英嗣（正会員）

1996 年東京理科大学理工学部電気工学科卒業．1998 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．2001 年同大学情報科学研究科博士後期課程修了．同年 4 月より日本学術振興会特別研究員．現在に至る．自然言語処理，特にテキスト自動要約の研究に従事．言語処理学会，人工知能学会，ACL，ACM 各会員．nanba@lr.pi.titech.ac.jp, http://oku-gw.pi.titech.ac.jp/~nanba/

神門 典子（正会員）

1994 年慶應義塾大学文学研究科博士課程修了．博士（図書館・情報学）．同年学術情報センター助手．1995 年米国シラキウス大学情報学部客員研究員，1996〜1997 年デンマーク王立図書館情報大学客員研究員．1998 年 4 月学術情報センター助教授．2000 年 4 月国立情報学研究所人間・社会情報研究系助教授．図書館情報大学大学院情報メディア研究科客員助教授．現在に至る．テキスト構造を用いた検索と情報活用支援，言語横断検索，情報検索システムの評価等の研究に従事．ACM-SIGIR，BCS-IRSG，ASIS&T，言語処理学会，日本図書館情報学会各会員．kando@nii.ac.jp, http://research.nii.ac.jp/~kando/index-j.html

奥村 学（正会員）

1984 年東京工業大学工学部情報工学科卒業．1989 年同大学院博士課程修了．同年東京工業大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，2000 年東京工業大学精密工学研究所助教授，現在に至る．工学博士．自然言語処理，知的情報提示技術，語学学習支援，テキストマイニングに関する研究に従事．人工知能学会，AAAI，言語処理学会，ACL，認知科学会，計量国語学会各会員．oku@pi.titech.ac.jp, http://oku-gw.pi.titech.ac.jp/~oku/