論文間の参照タイプの細分化に基づくサーベイ補助システム

小出　寛史　　橋本　陽平　　秦野　福己　　韓　東力

日本大学文理学部　情報システム解析学科

1　はじめに

文献サーベイを行う際に使用される手法の１つに、起点となる論文（以下「起点論文」と呼ぶ）を１つ選定し、その論文が参照している論文や、さらに参照論文が参照している論文という順にサーベイの対象を広げていくという方法がある。しかしながらこの方法では、研究に必要な文献を収集・理解することは非常に時間がかかる。また、集めた論文同士の関係を把握することも容易ではない。そこで、ある１つの論文を選定した時点で、その論文と参照論文の関連性を明らかにすることが出来れば、文献サーベイの効率化が図れ、時間や労力を大幅に減らすことが出来ると考えられる。

既存研究には論文と参照論文の関連性を「論説根拠型」、「問題点指摘型」と「その他型」の３種類に分類しているもの[1]や、起点論文と参照論文の関係を両方の論文に共通して出現する複合語や専門用語などを使用してつかむもの[2]がある。[1]では、論文間の参照タイプの分類が３種類と具体性に欠けており、起点論文から参照論文への参照目的が掴みづらい。[2]では、特徴語のみを頼りにし、文章の意味内容まで細かく踏み込んで論文の関連付けを行っていないという問題点がある。

本研究では、論文間の参照タイプを従来の３種類から、「歴史」・「類似研究」・「理論」・「研究手法」・「実験・データ」・「結果」の６種類に細分化し、さらに意味解析を利用することにより、論文間の参照タイプを明確にする手法を提案する。具体的には、参照個所・論文タイトル・文献種類・位置情報の４つの観点から論文間の参照タイプを判定し、さらにそれぞれの結果を加点法により統合する。システム全体の流れを図１にて示す。

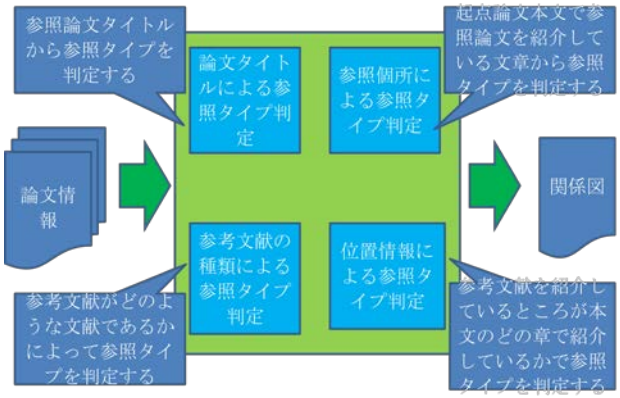

図１　システム全体の流れ

2　参照箇所による参照タイプの同定

参照箇所とは、起点論文において参照論文について記述した場所である。本章では、参照箇所内にある文章に対する意味解析によって論文間の参照タイプを判定するモジュールの説明を行う。参照個所内のテキスト情報から参照タイプを判定するために、本研究では論文情報、文間関係と参照タイプごとの辞書を使用する。参照タイプごとの辞書とは、参照タイプの正解をあらかじめ手動で付与した参照箇所学習データに対し、意味解析器 Aya[3]を適用した結果から抽出した概念識別子（名詞、複合名詞、未定義語）を参照タイプごとにまとめたものである。

参照箇所内のテキスト情報に基づき論文間の参照タイプを判定するのに、まずは参照箇所の範囲を明確に定める必要がある。これには、参照先の論文のメタ情報（著者名や公表年度など）が記述されている文自身だけでなく、同じ段落内で隣

接している他の文との間にも何らかの関係があることから、メタ情報が明記されていない前後の文も参照箇所の 1 部と見なす必要がある場合が多い。本研究では文間関係を使用し参照箇所を定める。具体的には、意味解析器 Aya で使用される 21 種類の文間関係から、学術論文でよく使われるものとして「原因」・「等位」・「逆接」の 3 つを利用して参照箇所の範囲を定める。

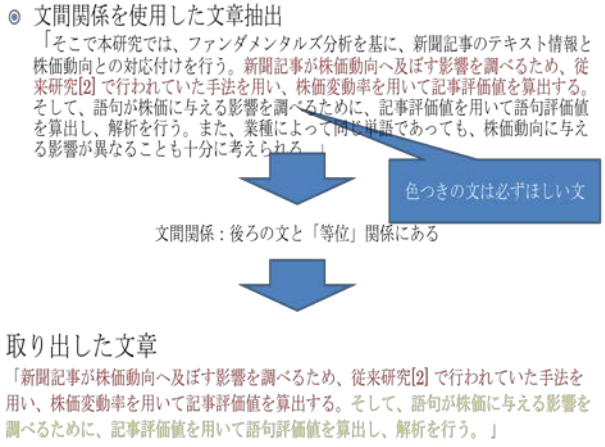

図 2　文間関係を利用した参照箇所抽出の例

図 2 の例では、青文字で書かれている部分が参考文献のメタ情報を記述している。文間関係が等位関係にあることから、「新聞記事が株価動向へ及ぼす影響を調べるため、従来研究[2] で行われていた手法を用い、株価変動率を用いて記事評価値を算出する。」の後文である、「そして、語句が株価に与える影響を調べるために、記事評価値を用いて語句評価値を算出し、解析を行う。」も参照箇所の一部として抽出される。

抽出された参照箇所内のテキストに対して意味解析を行った結果から、すべての名詞・複合名詞・未定義語の概念識別子を抽出する。これらの概念識別子があらかじめ作成しておいた 6 種類の辞書のどれに含まれているかを判定し、その結果識別子が最も多く含まれている辞書の参照タイプに加点処理を行う。

3　論文タイトルによる参照タイプの同定

本章では論文タイトルの構成とその解析結果により論文間の参照タイプを判定するモジュールの説明を行う。ここでは、村松らの研究[4]で提案された概念語と関係語を利用する。図 3 はこのモジュールの処理フローを示している。

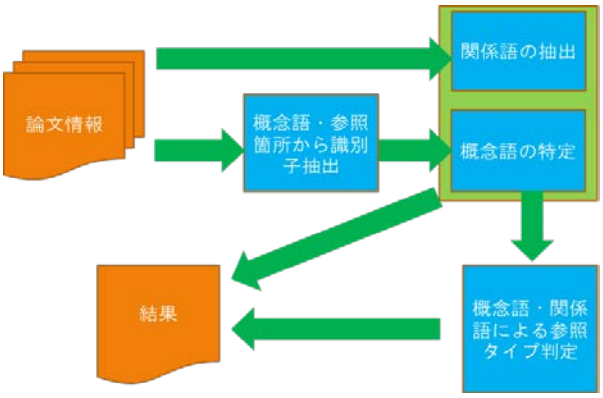

図 3　論文タイトルによる参照タイプ同定の流れ

ここでは、まず論文情報である「起点論文の本文にある参照箇所」、「起点論文のタイトル」と「参照論文タイトル」に対して意味解析を行った結果から概念語の概念識別子を抽出する。次に「参照論文タイトル」に含まれる概念識別子のうち、「起点論文のタイトル」と「起点論文の本文にある参照箇所」において最も多く使用されている識別子のある概念語を特定し、さらにその概念語にかかっている関係語かその概念語がかかっている関係語を抽出する。最後に、これらの結果をもとに表 2 の規則と概念語を構成している識別子により論文間の参照タイプを判定する。

3.1　概念語と関係語の抽出

本節では論文タイトルから概念語・関係語の抽出方法について説明していく。多くの科学技術論文ではタイトルをいくつかの構成で分けることができ、タイトルにはその論文で最も言いたいことが書かれている。そこで我々は参照論文のタイトルを構成ごとに分割できれば、その参照論文に

ついて起点論文において何を1番述べたいかを判定することができると考えた。

論文タイトルを構成ごとに分割するため、既存研究[4]で提案された概念語・関係語を用いることにした。概念語・関係語とは以下のようなものである。

- 概念語（キーワード）
  →名詞・副詞・形容詞

- 関係語（概念語同士の関係）
  →動詞・助詞・助動詞

既存研究では関係語が表1のように分類されている[4]。

カテゴリ名	主な関係語
限定	の、な、された、される
場所	における、での、上の、から見た
手段	による、を用いた、に基づく、を利用した
and	と、および、ならびに、も
目的	のための、を目指した、を指向した
内容	に関する、についての
方向	への、向きの
起点	からの、から
考慮	を考慮した、に着目した、を想定した
対象	に対する、を対象とした
所有	を持つ、を有する。を持った、を備えた
共有	間の、を共有された、間での
同格	としての
補助	を支援する、をサポートした
主格	が、は
適応	に対応した、に適した、に応じた
可能	可能な、を可能とする、が可能な
or	としての

表1　既存研究で提案された関係語のカテゴリ

3.2　最重要概念語の特定

起点論文からある参照論文を参照したとき、参照論文のすべての部分を引用したいわけではなく、参照論文中の手法、定理、あるいはデータなど一部の内容にしか焦点を当てていない場合がほとんどである。つまり参照論文のタイトルを構成している概念語の中のどれか1つについて引用したいと考えられる。したがって、論文間の参照タイプを判定するためには、参照論文タイトルを構成している概念語の中で、起点論文から最も言及したい概念語を特定する必要がある。

具体的な手法として、参照論文のタイトルを構成している各概念語に対して、起点論文のタイトルを構成している概念語と起点論文の本文中にある参照箇所から抽出された概念識別子を用いて頻度情報を加味した類似度計算を繰り返して行う。類似度計算は原田らが提案したもの[5]を利用する。ここで、各概念語を構成している概念識別子数が異なるため、平均をとることによりデータの標準化を図る。その後、平均値が一番高い概念語（最重要概念語）を特定し、その参照論文について最も言いたいこととし、3.3 節に進んで参照タイプを決定するための処理を続行する。もしどの類似度も事前の準備実験によって定められた閾値以下、あるいは標準偏差値以下なら、起点論文からその参照論文の中身について特に何も言及することはないと判断し、この時点で参照タイプを「歴史」と推定し、このモジュールでの処理を終了する。

3.3　最重要概念語による参照タイプの同定

3.2 節で抽出された最重要概念語は、参照タイプを判定するのに重要な手掛かりと認識し、概念語を構成している概念識別子によって参照タイプを判定できるよう2章と類似した手法を考案した。

概念語を構成する概念識別子すべてを抽出し、それぞれの識別子に対し、「理論・研究手法」、「実験手法・データ」と「結果」の3つの辞書と照らし合わせていく。ここでは2章と異なり、辞書を3つしか使用しないのは、論文のタイトルから参照タイプを判定しようとした場合には、タイトルを見ただけでは「理論」と「研究手法」の違いを判別することが不可能と考え、「理論」と「研究手法」を合わせることにしたためである。

照らし合わせた結果、最も参照回数の多い参照タイプにシステムの最終結果の評価に加点する。

また、どの参照タイプも一定以上の点数を得ることが困難な場合には、消去法で「類似研究」にシステムの最終結果の評価に加点する。

3.4　関係語による参照タイプの同定

既存研究[4]で提案された関係語は概念語同士の何らかの関係を表すものである。我々はこの事実をもとに、最重要概念語の前後に現れる関係語を手掛かりに、表 2 に示される規則に従って参照タイプを判定する。

Aは関係語から見て前にある概念語
Bは関係語から見て後ろにある概念語

限定	の	な	された	される		
場所	における	での	上の	から見た		
手段	による	を用いた	に基づく	を利用した		理論A+
and	と	および	ならびに	も		
目的	のための	を目指した	を指向した			結果A+
内容	に関する	についての				
方向	への	向きの				結果A+
起点	からの	から				
考慮	を考慮した	に着目した	を想定した			理論B+
対象	に対する	を対象とした				実験A+
所有	を持つ	を有する	を持った	を備えて		理論B+
共有	間の	で共有された	間での			
同格	として					
補助	を支援する	をサポートした				結果A+
主格	が	は				
適応	に対応した	に適した	に応じた			理論B+
可能	可能な	を可能にする	が可能な			理論B+
or	や					

表 2 関係語の分類による加点基準

表 2 は 18 種類のカテゴリにグルーピングされた関係語を、カテゴリごとに手動で参照タイプを与えた一覧表である。この表に赤字で書かれているものが参照タイプを付与したものである。関係語は必ず概念語と概念語の間にあるものなので、参照タイプを与えた関係語には前後どちらの概念語と掛っているのかも付与している。また「and」・「同格」・「or」の 3 つに対しては、前後の概念語との関係が同じであると判断し、次に出てくる関係語の参照タイプに依存するものとする。表 2 の規則に従い、最重要概念語が加点対象となる関係語と掛っていた場合、与えられた参照タイプに従い、システムの最終評価に加点する。

4　まとめ

本研究では、従来研究よりも論文サーベイの効率が向上することを目標に、論文間の参照タイプを 6 種類に細分化した。また意味解析を導入することにより論文間の参照関係を深層レベルで捉えようとしている。本研究はまだ初期段階にあり、今後は次の課題を逐次的に検討していく。

- 細分化における分類方法に対する見直し
- 加点法における各指標の加重方針の確定
- 加点法の代わりに機械学習を導入する可能性についての検討

また、現時点では、4 つの判定モジュールの中で、参照個所と論文タイトルによる手法しか実装されていないため、評価実験はまだ行われていない。今後は文献種類と位置情報による判定モジュールを完成し、各モジュールおよびシステム全体の評価実験を行い、手法全体の有効性を検討資する予定である。

参考文献

[1]　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，42(11), pp.2640-2649. (2001).

[2]　土橋喜，山内平行，立花隆輝，「キーワードマイニングと関連性の可視化による文献集合からの知識連鎖の発見支援」, 電子情報通信学会技術研究報告, 103(280), pp.55-60. (2003).

[3]　http://www.jsa.co.jp/contents/GG/group/Aya/aya.htm

[4]　松村敦，高須淳宏，安達淳，「単語間の係受け関係を用いた情報検索手法の評価」，情報処理学会論文誌. 41(SIG_1(TOD_5)), pp.22-30. (2000).

[5]　原田実，鈴木亮，南 旭瑞，「意味グラフのマッチングによる事故問い合わせ文からの判例検索システム Jcare」，9(2), pp.3-22. (2002).