論文間の引用情報を利用した関連用語の自動収集

難波英嗣
広島市立大学　情報科学部

1. はじめに

本研究は，ユーザが専門用語を入力すると，その用語と関連する用語のリストを自動的に提示するシステムの開発を目指し，その前段階として，システムが関連用語を自動収集する方法として，論文間の引用情報を利用することを提案するものである．

適切な検索キーワードの選択は，その検索対象があいまいであるほど難しい．論文をデータベースで検索する場合も，論文のタイトルや著者名などのキーワードが明確であればそれをそのまま利用することができるが，逆に特定の論文や著者を想定していない場合は，たまたま思いついた語を検索キーワードとすることになり，検索にもれを生じることになる．

そこでしばしば利用されるのがシソーラスである．シソーラスが提示する関連語を合わせて用いることで，検索結果に幅をもたせることができる．そしてシステムそのものにこのシソーラスを構築し，提示する機能があれば，ユーザ，とりわけ検索に不慣れなユーザにとって使いやすい情報検索システムとなる．このように関連語を用いた検索システムとしては，例えば「連想検索」を前面に出した Webcat Plus（国立情報学研究所）などがある．

筆者は，Web上のPDFやPostscript形式の論文データを収集し，それらの引用関係を解析して構築される論文データベースPRESRI[1]の開発に携わってきた[難波 2005]．現在，このシステムではWeb上から収集した約 83,000 件の日英フルテキスト論文データ，自然言語分野の主要学会の英語論文を集めたACL Anthology[2]のフルテキストデータ約 8,000 件および，これらのデータから抽出された参考文献が検索できる．

このデータベースに関連用語の収集システムを組み込むことで，ユーザがある分野の論文を効率よく集めることを支援することがこの研究の目的である．この支援システムが実現した場合，ユーザが入力した用語と異なる言語で書かれた論文のみを対象にすれば，関連訳語リストや，特許文書と学術論文のように異なるジャンルの文書間での引用関係をあらかじめ解析しておくことで，ある学術用語に対応する特許用語を得たりすることも可能になるかもしれない．

関連用語を収集する際，本研究では論文間の引用情報に着目する．一般に論文は同一もしくは関連領域の論文と引用関係にある．そこで，ある用語に関する論文を収集し，それらと直接引用関係にある論文から，各論文の研究内容を示す用語を抽出すれば，入力された用語に関連する用語の自動収集が実現できると考えられる．しかし，論文の引用には様々なタイプ(理由)のものがあり, 単純にすべての引用関係を利用すると関連のない用語も収集してしまう可能性がある．そこで本研究では，特定のタイプの引用のみを利用することで，効率的な関連用語の収集を目指す．

本論文の構成は以下のとおりである. 2 節では，関連研究について述べる．3 節では，関連用語を収集する上で重要な要素技術となる引用情報について述べ，引用情報を用いた関連用語の収集方法を説明する．提案手法の有効性を調べるために実験を行った．4 節では，実験方法および結果について述べる. 5 節ではまとめ, 6 節で今後の課題について述べる．

2. 関連研究

近年，Web を利用して関連用語を自動収集する研究が活発に行われている[佐藤 2003, 佐々木 2004, 白井 2004, 小原 2004, 大石 2004]．Web からある専門用語 t に関連する用語を収集するには，まず，ある用語 t に関する記述を収集し，そこから t と関連する用語を抽出するという手順が必要になる．ここで，どのように用語 t に関する適切な記述を収集するのかというのが，ポイントのひとつとなる．例えば，佐藤ら[佐藤 2003]は，次に述べる方法で，用語 t に関する記述を収集している．ある用語 t に対して，まず「t とは」「t という」「t は」「t」の 4 種類のクエリを検索エンジンに入力し, 得られた URL のそれぞれ上位 100 ページを入手する．次に入手したページを整形して文に分割し，用語 t を含む文のみを抽出し，そこから関連用語の収集を行う．

このように，Web 全体の中から，ある用語に関する記述を収集するというのもひとつの方法であるが，本研究のように対象を学術用語に限定する

[1] http://www.presri.com
[2] http://acl.ldc.upenn.edu/

場合，あらかじめ学術用語の記述が多い論文をWebから収集しておき，そこから用語を探すという方法も考えられる．本研究では，1節で述べたPRESRIを用いてWeb上から学術論文をあらかじめ収集しておき，そこから関連用語を収集する．収集対象を学術論文に限定することで，精度の高い用語収集が期待できる．

さて，関連用語を収集するためには，用語間の関連度の計算方法ついても検討する必要がある．これまでの研究は，ある用語と与えられた用語tとのテキスト中の共起頻度に基づいて，その用語の関連度を計算するのが一般的であった．これに対し，本研究では論文間の引用情報に着目する．カレント・コンテンツという文献サービスでは，論文間の引用を利用して検索質問拡張(query expansion)を行うキーワード・プラスという機能がある．これは，ユーザがキーワード検索を行う際，著者によって引用された論文の表題から専門用語を抽出し，それらを検索キーワードの候補として，ユーザに提示するというものである．このような検索質問拡張を行う前提には，論文間の引用関係を利用すれば関連度の高い用語をユーザに提示できるという考え方があると思われる．他方，1節でも述べたとおり，論文の引用には様々なタイプのものがあるため，すべての引用関係を使うと，検索質問拡張がうまくいかない場合もありうる．そこで，本研究では次節で述べる引用情報を用いて，この問題点を解決する．

3. 引用情報を利用した関連用語の収集

本節では，まず引用情報について説明し，次に引用情報を用いた用語収集方法について述べる．

3.1. 引用情報の抽出

学術論文中には，当該論文と被引用論文との関係について記述されている個所(引用個所)がある．引用個所から得られる情報を，本研究では引用情報と呼んでいる．引用個所からは，被引用論文の重要点や当該論文と被引用論文との相違点を明示する有用な情報が得られる．また，引用個所を読めば引用の理由が分かる．

本研究では，引用の理由を引用タイプとして以下の3種類に分類し，また，引用タイプの決定を自動的に行っている[難波 1999].

- type C (問題点指摘型)
  他の論文の理論や手法等の問題点を指摘するための引用.
- type B (論説根拠型)
  既存の研究成果を用いて，新しい理論を提案したり，システムを構築したりする場合の引用.
- type O (その他型)
  type Bにもtype Cにも当てはまらない引用.

3.2. 引用情報を利用した関連用語の収集手法

本研究では，以下の手順で関連用語を収集する．

1. ユーザが専門用語を入力する.
2. システムは，入力された用語を論文表題に含む書誌情報を論文データベースから収集する．これらをルートセットと呼ぶ.
3. ルートセットと特定の引用タイプで引用関係にある論文の表題を収集する．これらをベースセットと呼ぶ.
4. ベースセットの論文表題から専門用語を抽出し，用語としての重要度および入力された用語との関連度を考慮して並べ，関連用語リストを出力する.

なお，ステップ4において，論文表題からの専門用語の抽出に，中川らの開発したTermExtract[3]というツールを利用する[Nakagawa 2002]．中川らの抽出手法は，「多くの異なり語と連接する名詞から構成される複合語は重要語である」という考え方に基づいている．このツールをテキスト集合(ある専門分野のコーパス)に適用すると，重要度と共に専門用語リストが出力される．本研究では，この用語リストと重要度を用い，ステップ4において，ベースセット中の専門用語の頻度(ステップ1で入力された専門用語との関連度)とその用語の重要度をかけ，その値の大きい順に関連用語リストとして出力する.

4. 実験

3節で述べた手法の有効性を調べるために実験を行う.

4.1. 実験に用いるデータ

実験には，PostscriptおよびPDF形式の自然言語処理分野を中心とするフルテキスト論文約12,000件を用いる．これらのうち，約8,000件はACLが提供するACL Anthology[4]に含まれるもの，残りの4,000件は，国内外の自然言語処理研究者

[3] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
[4] http://acl.ldc.upenn.edu/

や自然言語処理系研究室のWebページから収集したものや，国際会議の予稿集(CD-ROM)から抽出した論文データから構成される．

PRESRI を用いて，これらのデータから論文データベースを構築する．まず各論文データから，その論文の書誌情報(タイトル，著者名，所属，キーワード，アブストラクト)と，参考文献を抽出する[阿辺川 2003]. 次に引用タイプの自動判定が行われる[難波 1999]. 最後に抽出された書誌情報間で同定処理を行い，すべての書誌情報および引用情報をデータベースに格納する．

4.2. 実験方法

今回の実験では，自然言語処理分野の用語 25 個をシステムの入力に用い, その出力結果を見て, システムを定性的に評価する． 入力の一部を以下に示す．

“summarization”
“information extraction”
“machine translation”
“natural language processing”
“terminology”

なお，3.2 節のステップ 3 に関して，今回は，入力される各用語につき, type B のみを使った場合, type C のみを使った場合, type O のみを使った場合の 3 通りで関連用語の収集を行う．

4.3. 結果

“terminology”と“machine translation”を入力とした時の，各引用タイプにおける上位 5 件の収集結果を以下に示す．

入力 “terminology”

type B:
1. annotation guile line
2. named entity
3. named entity annotation guideline

type C
1. information retrieval
2. language processing
3. term extraction
4. term recognition
5. medical language processing

type O
1. information retrieval
2. language information retrieval
3. language resource
4. natural language
5. language processing

入力：“machine translation”

type B:
1. machine translation
2. language model
3. speech translation system
4. language modeling
5. speech translation

type C:
1. machine translation
2. evaluation of machine
3. evaluation of machine translation
4. MT system
5. parallel corpus

type O:
1. machine translation
2. neural network
3. language resource
4. speech translation
5. natural language processing

全体的な傾向として, type C と O を用いた時に比較的良好な結果が得られた．他方，type B に関しては，関連用語だけでなく，その分野で用いられるデータや手法に関する用語も収集された．

収集された用語の数は, type O がもっとも多く, type B と C は O と比べると少なく，入力によっては，まったく収集できない場合もあった．

4.4. 考察

type C, B, O ごとに，収集された関連用語の傾向をまとめる．

- type C:
  type B や O と比べると, 比較的安定して関連用語が収集できているが，ノイズも多く改善の余地がある．
- type B:
  自然言語処理分野の中でも，“Support Vector Machine”のように要素技術として使われるものの一部では，“coreference resolution” や “named entity extraction” 等，使われている研究分野名が収集されることもある．しかし，すべての要素技術で同様の傾向が見られるわけではない. また, ノイズを収集する割合も高い.

- type O:
  type O は収集される関連用語数が多いため，type C や B に含まれていない用語が収集できることも多いが，ノイズも少なからず含まれている．

今回，どのタイプにおいてもノイズが混入してしまった原因のひとつは，論文表題に含まれるすべての専門用語を用いたことと関係する．例えば，“Automatic Summarization using Support Vector Machine”という論文表題からは，“Automatic Summarization” と “Support Vector Machine” という 2 つの用語が抽出される．前者は研究のトピックを，後者は要素技術を表しているが，今回用いた手法では，両者の違いを区別していない．論文表題はある程度記述形式が決まっているため，一般的なテキストと比べると比較的その構造を解析しやすい．そこで，今後は，例えば佐藤ら[佐藤 1999]や松村ら[松村 2001]が行っているような論文表題解析の技術を用いることで，用語収集の精度が改善できるのではないかと考えている．

5. おわりに

本研究では，論文間の引用関係および引用タイプを利用し，入力された専門用語と関連する用語を自動的に収集する手法を提案した．現状では，まだ定量的な評価を行うまでには至っていないが，出力を見る限りにおいては，比較的良好な結果が得られていると判断される．

6. 今後の課題

今後は，まず，提案手法の定量的な評価を行う．評価には 2 種類の方法を考えている．ひとつは，収集された用語の中にどの程度関連用語が含まれているのかを評価する方法である．もうひとつは，収集された用語を用いて，ユーザの検索効率が向上する度合いにより評価する方法である．後者に関しては，情報検索のテストコレクションを用い，ある用語だけを使って検索した場合と，関連用語を収集の結果を使って検索した場合とを比較し，収集された用語の品質を測ることができるのではないかと考えている．

次に，入力された用語とは異なる言語の関連用語収集にも取り組む予定である．

謝辞

本研究の一部は，平成 16 年度広島市立大学特定研究費による支援を受けて行われました．

参考文献

[阿辺川 2003] 阿辺川 武，難波 英嗣，高村 大也，奥村 学 (2003) “機械学習による科学技術論文からの書誌情報の自動抽出” *情報処理学会研究報告自然言語処理, NL-157*, pp.83-90.

[松村 2001] 松村 敦，高須 淳宏，安達 淳 (2001) “情報検索における単語間の関係の効果” *情報処理学会研究報告データベースシステム Vol.2001, No.70, DBS-125,* pp.257-264.

[Nakagawa 2002] Nakagawa, H., Mori, T. (2002) “A Simple but Powerful Automatic Term Extraction Method” *Computerm2: 2nd International Workshop on Computational Terminology, COLING-2002 Workshop,* pp.29-35.

[難波 1999] 難波 英嗣，奥村 学 (1999) “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.

[難波 2005] 難波 英嗣，阿辺川 武，奥村 学，齋藤 豪 (2005) “Web 上のデータを中心とした複数論文データベースの統合” *言語処理学会第 11 回年次大会*.

[小原 2004] 小原 恭介，山田 剛一，絹川 博之，中川 裕志 (2004) “ウェブを利用した関連用語収集” *第3回情報科学技術フォーラム (FIT2004)*.

[大石 2004] 大石 康智，伊藤 克亘，武田 一哉，藤井 敦，板倉 文忠 (2004) “事典コーパスを用いた単語階層関係の統計的解析” *第 3 回情報科学技術フォーラム (FIT2004)*.

[佐々木 2004] 佐々木 靖弘，佐藤 理史，宇津呂 武仁 (2004) “用語間の関連度を測る指標の提案” *言語処理学会第 10 回年次大会*, pp.25-28.

[佐藤 1999] 佐藤 理史 (1999) “論文表題を言い換える” *情報処理学会論文誌, Vol.40, No.7,* pp.2937-2945.

[佐藤 2003] 佐藤理史，佐々木 靖弘 (2003) “ウェブを利用した関連用語の自動収集” *情報処理学会研究報告 自然言語処理, NL-153*, pp.57-64.

[白井 2004] 白井 清昭，菅井 俊介，平野 健児，星 正人 (2004) “ポータルサイト自動作成の試み” *言語処理学会第 10 回年次大会*, pp.624-627.