4Y-04 WWW 上の多言語論文データを用いたサーベイ支援システムの開発

難波 英嗣 †
日本学術振興会
特別研究員

奥村 学 ‡
東京工業大学
精密工学研究所

1 はじめに

ある分野の研究を網羅的にサーベイするには，特定の言語で書かれた論文だけではなく，多くの言語で書かれた論文を対象にする必要がある．本論文では，多言語論文データベースを用いてサーベイ支援を行うシステムについて述べる．

近年，数多くの電子化された論文が WWW 上から入手可能になってきている．WWW 上の論文データを用いた論文データベースとしては，ResearchIndex[1] や Cora[2] がある．しかし，これらは英語で書かれた論文のみを対象にしているため，網羅的なサーベイを行うのに十分であるとは言えない．本研究では，WWW 上にある日英論文データを収集し，より網羅的な論文データベースを自動的に構築する．

また，既存の論文データベース [1, 2] では，ある論文と参照・被参照関係にある論文を調べることができるが，本研究で開発するシステムは，さらに，ある論文の関連論文の中での位置付けを明確にする情報 (**参照情報**) を抽出し，ユーザに提示することでサーベイの支援を行う．

本論文では，次節で，まずサーベイ支援の枠組について説明する．3 節では，2 節の枠組に基づいたサーベイ支援システムについて述べる．また，4 節ではシステムの動作例を示す．

2 論文間の参照情報を用いたサーベイ支援

本研究では，サーベイ支援の際，論文間の**参照情報**に着目している．学術論文中には，当該論文と被参照論文との関係について記述されている個所 (**参照個所**) がある．参照個所から得られる情報を，本研究では参照情報と呼んでいる．参照個所からは，被参照論文の重要点や当該論文と被参照論文との相違点を明示する有用な情報が得られる．また，参照個所を読めば参照の理由が分かる．本研究では，参照の理由を**参照タイプ**として以下の 3 種類に分類し，また，参照タイプの決定を自動的に行う．

- **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参照．(例えば，本論文における [1] や [2])
- **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，システムを構築する場合の参照．(例えば，本論文における [3])
- **type O (その他型)**
  type B にも type C にも当てはまらない参照．

ある論文に関する複数の参照情報を集めれば，その論文の関連論文の中での位置付けが明らかになるため，特定分野の研究動向の把握に有用である．

これまで，われわれは，手がかり語に基づいたルールを用いて参照個所の抽出や参照タイプの決定を行う手法を開発してきた [3]．本論文中の論文 [1] に対する参照を用いてこれらの手法を説明する．この参照が出現する文 (1 節 2 段落目) の次の文は逆接の接続詞「しかし」で始まっていることから，論文 [1] に対して本論文中では何らかの問題点が指摘されている (type C) と判断できる．このように，論文中の参照と，その周辺の「しかし」のような手がかり語との前後関係を考慮することで，参照タイプを自動的に決定できる．また，参照個所の抽出では,「しかし」「そこで」といった文間のつながりを示す手がかり語を用いることで，参照の出現する文とつながりの深い文を抽出することができる．

3 サーベイ支援システムの構築

本節では，2 節で説明した参照情報を用い，サーベイ支援を行うシステムの構築手順について述べる．手順は次の 5 つのステップからなる．

(1) 論文データの収集: WWW 検索エンジン[1] [2] を用いて，キーワードの 6 種類の組み合せ ((“業績” or “研究” or “publications”) and (“postscript” or “pdf”)) で検索する．検索結果の URL を 2 階層までたどり，Postscript と PDF ファイルを収集する．

(2) テキストファイルへの変換: Postscript は prescript[3] を PDF は pdftotext[4] をそれぞれ用いて，

[1] http://www.google.com
[2] http://www.goo.ne.jp
[3] http://www.nzdl.org/html/prescript.html
[4] http://www.foolabs.com/xpdf/

A Survey Support System using Multilingual Scientific Papers on the World Wide Web
†Hidetsugu Nanba (nanba@lr.pi.titech.ac.jp)
JSPS Research Fellow
‡Manabu Okumura (oku@pi.titech.ac.jp)
Precision & Intelligence Laboratory,
Tokyo Institute of Technology

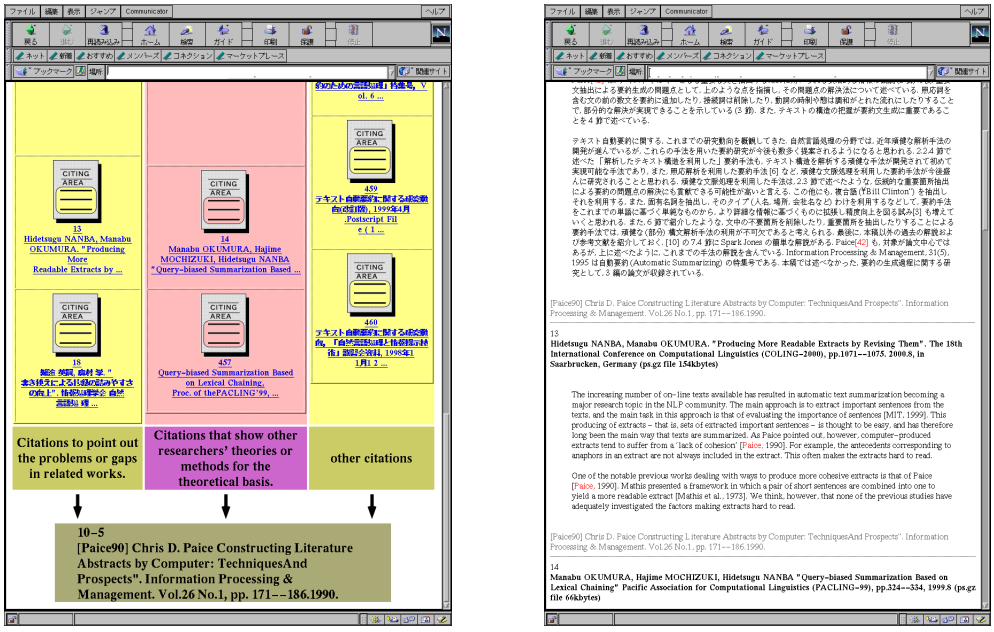

図 1: サーベイ支援システム PRESRI

テキストファイルへの変換を行う．なお，prescript の日本語パッチは国立情報学研究所の片山紀生氏より提供していただいた．

(3) 論文構造の解析: 「参考文献」「References」等の文字列に着目して，個々の論文ファイルが参照している論文を抽出する．次に論文中の参照位置を，1)，(1)，[1] といった参照パターンに基づいて特定する．また，テキストファイルの先頭 5 行以内から論文の書誌情報 (論文表題や著者名など) を抽出する．

(4) 論文間の参照・被参照関係の解析: ステップ (3) で抽出された個々の論文ファイルの書誌情報リストと参考文献リストを比較し，論文間の参照・被参照関係を解析する．

(5) 参照情報の抽出: 2 節で説明した手がかり語に基づくルールを用い，参照個所の抽出および参照タイプの決定を行う．

4 システムの動作例

本節では，システムの操作方法および動作例について述べる．まず，論文表題や著者名をキーワードとして論文検索を行う．検索結果はリスト表示される．リスト中の各論文をマウスで選択することで，その論文と参照・被参照関係にある論文をグラフで表示することができる．このグラフを辿ることで，論文間の参照・被参照関係を用いた検索が可能になる．

図 1 左は，システムが実際に論文間の参照・被参照関係をグラフ表示した動作例である．図は [Paice 90] という論文を 6 つの論文が参照している状態を表している．これらの 6 つは，2 つづつ左から順にそれぞれ type C, B, O で参照している．また，グラフ中の “CITING AREA”(参照個所) をクリックすると，対応する参照個所が表示される (図 1 右)．

図 1 左は，関連 (参照) 論文の中での [Paice90] の位置付けを表していると捉えることができ，また，個々の関連論文との関係は参照個所 (図 1 右) を読めば分かることから，このシステムがサーベイの支援に有用であると考えることができる．

5 おわりに

本論文では，WWW 上の多言語論文データを用いたサーベイ支援システムについて述べた．今後は，さらに多くの日英論文データを収集すると共に，日英以外の言語への適用も目指す．

なお，このシステムは http://presri.pi.titech.ac.jp:8000 から利用可能である．フルテキスト論文が約 30,000 件 (日本語:2790 件, 英語:27823 件) 検索できる．

謝辞

本研究は科学研究費補助金 (特別研究員奨励費) の援助を受けて行われたものである．

参考文献

[1] Lawrence, S., Giles, L., 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”, *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, pp.662–667, 1999.

[3] 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.