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

難波 英嗣[1]　阿辺川 武[2]　奥村 学[3]　齋藤 豪[4]
1 広島市立大学　情報科学部
2 東京工業大学大学院　総合理工学研究科
3 東京工業大学　精密工学研究所
4 東京工業大学大学院　情報理工学研究科

1. はじめに

特定分野の研究動向を知るためには，その分野の論文を網羅的に収集する必要がある．このような文献調査においては，しばしば論文データベースが利用される．しかし，分散した論文データベースを一つずつ検索するのは非効率的である．また, 特定分野の研究動向を効率的に知るためには, 単にその分野の論文を収集して列挙するだけでなく，収集した論文同士の関係を解析し，それらを分かりやすく提示する必要がある．

そこで，我々は，複数の論文データベースを統合的に一つのデータベースとして検索できるシステム PRESRI の開発を行っている．また，特定分野の論文の論文間の関係をわかりやすく提示するインタフェースの作成にも取り組んでいる．本稿では，まず PRESRI の特徴やシステム構成を説明し，次にデータベース管理機能や検索機能を，動作例と共に紹介する．

2. PRESRI を構築する上でのポイント

2.1. 網羅的なデータベースの構築

網羅的な論文データベースを構築するためには, 複数の言語で記述された論文データを対象にする必要がある．そこで，我々は，これまでに Web 上に存在する Postscript および PDF 形式の日英論文データを収集して論文データベース(以後，‘WEB-DB’)を構築してきた[難波 2002].

しかし，研究者が利用可能な論文データベースとしては，Web 上に載らないものも数多く存在する．例えば，近年では，国際会議や学会の全国大会では予稿集の代わりに CD-ROM が配付されることが多い. このように個人の所有する CD-ROM や，所属組織の図書館，学会，出版社が所有するデータベースをここではローカルな論文データベースと呼ぶ.

このようなローカルに存在する論文データを，PRESRIと統合的に利用できれば，非常に便利である．例えば，CD-ROM中の論文とPRESRI中の論文が引用関係にあれば，その引用関係をたどって，効率的に関連論文を集めることができる．また, 大学図書館等の蔵書データと統合することで, 図書館所蔵の資料を引用する論文を検索するといったことも可能になる[1].

2.2. 検索結果のわかりやすい提示

CiteSeer[2][Lawrence 1999]やGoogle Scholar[3]をはじめとする多くの引用論文データベースの論文検索インタフェースは，検索結果や引用関係にある論文をリスト形式で表示するのが一般的である．しかし，このような表示方法では，より大きな引用構造の中での個々の論文の位置付け(関係)がわかりにくいという問題点がある．そこで，本研究では，検索結果のわかりやすい提示を目指している．以下，まず検索結果をわかりやすく提示するのに有用な引用情報について説明し，次に引用情報に基づく検索結果の可視化について述べる.

2.2.1. 引用情報

学術論文は先行研究について言及する際，引用を行う．このとき，当該論文と被引用論文との関係について記述される個所(引用個所)ができる.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，引用個所を読むことで，著者がその論文を引用した理由について知る手がかりが得られる.

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．(例えば，本論文における[Lawrence:1999])
- 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．(例えば，本論文における[難波 1999])
- type O (その他型)
  type B にも type C にも当てはまらない引用.

これまで，我々は，手がかり語に基づいたルールを用いて引用個所の抽出や引用タイプの決定を行う手法を開発してきた[難波 1999]. 本論文中の論文[Lawrence 1999]に対する引用を用いてこれ

[1] 広島市立大学では附属図書館の蔵書データとPRESRIを統合した検索サービスを 2004 年 12 月から開始している.

[2] http://citeseer.ist.psu.edu/
[3] http://scholar.google.com/

らの手法を説明する．

この引用が出現する文(2.2 節 1 文目)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[Lawrence 1999]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，論文中の引用と，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，引用タイプを自動的に決定できる．また，引用個所の抽出では，「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，引用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．

2.2.2. 引用情報を用いた検索結果の可視化

前節で述べた引用情報を考慮し，検索結果の可視化を行う．ひとつの方法は，論文をドット，引用関係を矢印で表現し，論文間の引用関係をグラフとして提示することである．

ここで, 図 1 に示すように, type C を直線, type B を点線，type O を破線で示すことにより，ユーザには引用タイプの区別ができる．また，グラフ中の各論文の書誌情報等は，図 2 に示すように，ポップアップウィンドウ内に提示することができる．例えば，ユーザがグラフ内に表示されたドットにカーソルを重ねれば論文の書誌情報が，矢印上に重ねれば引用個所が提示されるようなインタフェースが考えられる．本研究では，このようなグラフィカルなインタフェースの開発を行っている．

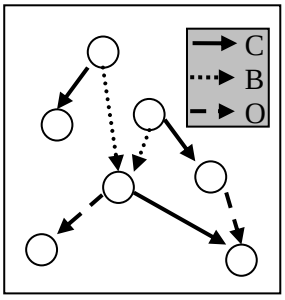

図 1 引用タイプの表示

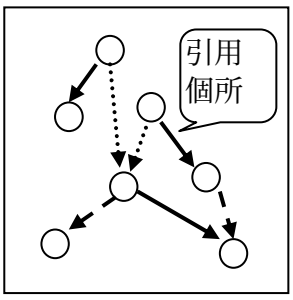

図 2 ポップアップウィンドウによる引用個所の表示

3. PRESRI の構築

3.1. PRESRI の構築手順

PRESRI の構築は, (1)書誌情報の抽出と(2)統合の 2 つのステップから構成される．

ステップ 1 では, Postscript および PDF 形式の論文から, 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．また，日本語論文中で，日英両方のタイトル, 著者名などが記述されている場合には，共にこれを抽出する[阿辺川 2003]. 次に引用タイプの自動判定が行われる[難波 1999]. ステップ 2 では，複数のサーバ上で抽出された書誌情報，アブストラクト，参考文献情報が集められ，統合される．

3.2. システムの構成

システム全体の構成図を図 3 に示す．

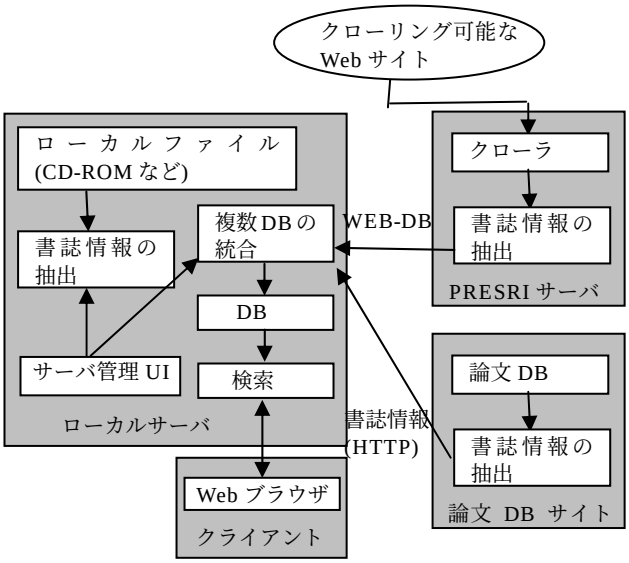

図 3　システム構成

システムは，以下のサーバおよびモジュールから構成される．

サーバ

- ローカルサーバ

ローカルサーバの役割は，ユーザへの検索機能の提供，複数のローカルデータベース(CD-ROM 等)とリモートデータベース(WEB-DB 等)の統合である．このサーバは，大学図書館や研究室等，特定のユーザのみがアクセスできる場所に設置される．

- PRESRI サーバ

PRESRI サーバは，Web 上の Postscript および PDF ファイルを収集し，そこから書誌情報を抽出する(WEB-DB の作成).

- クライアント

クライアントに Web ブラウザがインストールされていれば，ブラウザ経由で論文を検索することができる．

- 論文 DB サイト

論文 DB サイトとは，クローリングが許可されていないが論文データベースを保持するサイトである(例えば，学会や出版社のサイト)．もし，データベース管理者の許諾が得られれば，サイト上で書誌情報および引用情報を抽出し，それらを特定の(許諾が得られた)ローカルサーバ上で統合することができる．

モジュール

- クローラ

wget[4]を用い，ac，eduドメインのサーバからPostscriptおよびPDFファイルを収集する．

- 書誌情報の抽出

PRESRI サーバ，論文 DB サイト，ローカルサーバ上で動く．Postscript および PDF ファイルを XML ファイルに変換し，ファイルのヘッダ部分および参考文献から書誌情報を抽出する．抽出は，字種や手がかり語といった言語的情報と，フォントサイズや文字装飾等の視覚的情報を組み合わせて行う[阿辺川 2003].

- 複数 DB の統合

複数 DB を統合し，以下に述べる 3 つの手順で，データをローカルサーバに登録する．(1)サーバ管理者が登録した複数のサイトから書誌情報を収集，(2)DB 内および DB 間のデータの重複を調べる，(3)データを統合し，ローカルサーバに登録する．

- サーバ管理 UI(ユーザインタフェース)

ローカルサーバ管理者は，サーバ管理 UI を通して新規ローカル DB やリモート DB を登録し，それらを統合することができる．

- データベース

データソースのセットアップ，書誌情報の抽出，複数 DB の統合の制御を行う．

- 検索

論文を検索したり，引用情報を提示したりする機能を提供する．

4. システムの動作例

現在，世界中から利用可能な最新版のシステムはhttp://www.presri.comから利用可能である．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，主要学会の英語論文を集めたACL Anthology[5]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．

一方，特定のユーザに限ったシステムでは，広島市立大学で利用しているようにローカルに保持，閲覧可能な CD-ROM データの文献を追加してユーザにサービスすることができる．このようなデータベースの追加統合はすべて Web ブラウザを通して行える．

4.1. 複数データベースの統合

データベースの管理は，すべて Web ブラウザを通して行う．図 4 は，新しいデータベースを登録する画面である．データベースがローカルディレクトリに存在する場合はディレクトリの場所を，リモートサーバに存在する場合は，データベースが置いてある URL を指定する．もしデータベースが定期的に更新されるのであれば，更新日時を指定することで，自動更新できる．

図 4　データベース登録画面

図 5　データベース管理画面

登録されたデータベースは図 5 のように表で一覧表示される．不要になったデータベースは，後から削除することができる．

4.2. 検索インタフェース

以下，システムスナップショットを用いて，検索過程を説明する．PRESRI のトップ画面では，ユーザは，タイトル，著者名，掲載誌の項目ごとにキーワードを入力し論文検索を行う．この時，著作年の範囲の指定もできる．

図 6 は 1980 年～2005 年の間でタイトルに“citation”を含んだ論文を検索した結果を示している．図より，19 件の論文が検索されていることがわかる．検索結果は，デフォルトでは被引用数の多い順にランク付けされるが，この他，著作年順，著者名順等でもランク付けできる．

[4] http://www.gnu.org/software/wget/wget.html
[5] http://acl.ldc.upenn.edu/

この画面上で，ユーザが検索目的に適合する論文を選択し(図 6 の「グラフ」欄のチェックボックスをチェックする),「チェックした論文をグラフ表示」というボタンを押すと，図 7 および図 8 のような論文間の引用情報を示す画面が表示される．

図 6　キーワード検索の結果

図 7　引用情報の提示画面(書誌情報の提示)

図 7 において,「選択論文」は，図 6 でチェックされた論文を示している．被引用回数が 5 回以上の論文はドットが強調表示される．「その他の論文」は，選択論文と直接引用関連にある論文を示している．また，図の矢印は論文間の引用関係を示しており，引用タイプごとに色分けして表示されている．グラフ中で，ユーザがグラフ中のドットにカーソルを重ねると，図のように，論文の書誌情報がポップアップウィンドウ内に表示される．

引用情報の提示画面中で，ユーザが矢印にカーソルを重ねると，図 8 のように引用個所がポップアップ表示される．論文中で複数回引用されている場合には，すべての引用個所が提示される．また，矢印にカーソルを重ねた状態でクリックすると，新しいウィンドウが立ち上がり，ポップアップ画面に表示されている内容がウィンドウ内で閲覧できる．

図 8　引用情報の提示画面(引用個所の提示)

5. おわりに

本稿では，多言語引用論文 PRESRI の特徴，システム構成を説明し，データベース管理機能や検索機能を紹介した．

謝辞

本研究の一部は，平成 14 年度 IPA 未踏ソフトウェア創造事業による支援を受けて行われました．プロジェクトマネージャーの喜連川優先生(東京大学)からは有益なコメントを頂きました．また，現在は，NEDO 平成 16 年度産業技術研究助成事業の支援を受けています．

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 自然言語処理研究会, NL-157*, 83-90.

[Lawrence 1999] Lawrence, S., Giles, L., Bollacker, K. (1999). Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, Vol. 32, No. 6, pp.67-71.

[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 2002] 難波 英嗣, 奥村 学. (2002) “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” *第 64 回 情報処理学会全国大会*.