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

Integrating multiple databases of research papers with the data on the WWW

難波 英嗣 [1)]　奥村 学 [2)]　齋藤 豪 [2)]　阿辺川 武 [3)]
Hidetsugu NANBA　Manabu OKUMURA　Suguru SAITO　Takeshi ABEKAWA

1) 広島市立大学　情報科学部（〒731-3194 広島県広島市安佐南区大塚東 3-4-1
E-mail: nanba@its.hiroshima-cu.ac.jp ）
2) 東京工業大学　精密工学研究所（〒226-8503 神奈川県横浜市緑区長津田町 4259
E-mail: {oku, suguru}@ipa.go.jp ）
3) 東京工業大学大学院 総合理工学研究科（〒226-8503 神奈川県横浜市緑区長津田町 4259
E-mail: abekawa@ipa.go.jp ）

ABSTRACT. We have developed a system that makes it possible to retrieve papers at a time from multiple databases. Our system can show the citation relationships between papers together with their reasons for citations visually. Using this system, researchers can grasp the outline of a specific domain at a glance.

1．背景

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を行うのに，しばしば論文データベースが利用される．しかし，論文データベースが分散して存在していると，データベース毎に検索するのは非効率的である. このため，複数の論文データベースを統合的に利用できる環境の構築が望まれている．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，収集した論文の論文間の関係を解析し，それらを分かりやすく提示する必要があると考えられる.

2．目的

本プロジェクトでは，(1)複数の論文データベースを統合的に利用検索できるシステムの開発を行う．また，(2)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を行う.

（1） 複数論文データベースの統合

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を構築している[1]. しかし，研究者が利用可能な論文データベースは，このような Web 上の論文データ以外にも数多く存在する．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．また，研究者は，それぞれの所属する組織の図書館にある論文データベース等を利用することができる．この他に，学会や出版社の所有する論文データベースも利用できる[1].

[1]以後，個人の所有する CD-ROM，図書館や出版社や学会が所有するデータベースをまとめて，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 中の論文が参照関係にあれば，その参照関係をたどって，効率的に関連論文を集めることができる.

（2） 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

CiteSeer[2]，Cora[3]をはじめとする多くの引用文献データベースの論文検索インタフェースは，検索結果や参照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が，このような表示方法では，より大きな参照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．本プロジェクトでは，論文間の参照関係をグラフで表示し，ユーザがグラフ上の論文アイコンにカーソルを重ねるとその論文の情報が，参照関係を示す矢印にカーソルを重ねると参照個所が提示できるようにする.

3．開発内容

（1） システムの機能

本システムでは下記の機能を提供する.

1. 管理画面認証機能
2. 論文収集機能
3. データソース設定機能
4. 本文データ管理機能
5. 書誌情報・未解決参照情報抽出機能
6. 書誌情報・未解決参照情報収集機能
7. 参照データ生成機能
8. 書誌情報管理機能
9. 参照データ管理機能

[2] http://citeseer.nj.nec.com
[3] http://cora.whizbang.com

10. 論文検索機能
11. 参照関係表示機能

以下，各機能について概説する．

● **管理画面認証機能**
管理権限を有しない利用者がシステムの設定を変更できないようにする機能．アカウント名とパスワードによる認証を基本とする．この機能には，システム管理者の登録，削除等の機能も含まれる．

● **論文収集機能**
インターネット上で公開されている PDF, PS 等の論文ファイルを収集する．検索エンジンなどを利用して公開論文の URL リストを取得し，その結果に基づいて論文ファイルを収集する．

● **データソース設定機能**
大量の論文を保持しているネットワーク上のサーバ及び，ローカルのファイルシステムに保存されている論文集等のデータソース設定を管理する．データソース設定の作成，変更，削除機能を提供する．また，データソースごとの書誌情報更新履歴等も管理する．

● **本文データ管理機能**
ファイルシステム上に保存されている論文ファイルから本文データのテキストデータを抽出し，保存する．

● **書誌情報・未解決参照情報抽出機能**
本文データを解析し，論文の書誌情報及び論文中で参照されている被参照論文の書誌情報一覧を出力する．

● **書誌情報・未解決参照情報収集機能**
データソースサイトで公開されている，書誌情報・未解決参照情報及び，ローカルに保存されているデータソースの書誌情報・未解決参照情報を収集し，取り纏める．

● **参照データ生成機能**
各データソースから取得した書誌情報と未解決参照情報から，被参照論文の同定処理を行い，参照データを作成する．

● **書誌情報管理機能**
各データソースから取得した書誌情報，未解決参照情報から得られた書誌情報を一括して管理する．

● **参照データ管理機能**
サーバが保持している参照データを管理する．

● **論文検索機能**
キーワード等から，該当する論文を検索する．

● **参照関係表示機能**
論文間の参照関係を分析し，表示する．

（2） データ構成
本システムでは下記のデータを取り扱う．
1. 論文ファイル
2. 本文データ
3. 書誌情報
4. 未解決参照情報
5. 参照関係データ
6. データソース設定

以下に，それぞれのデータの詳細を概説する．

● 論文ファイル
論文ファイルは Web などで公開されている論文そのものである．本システムでは PDF もしくは PostScript で記述されているものを想定していう．論文ファイルはクライアントに論文本文を提示する際や，後述の本文データを抽出する際に利用する．

● 本文データ
本文データは，論文ファイル中に記述されている内容を抽出し，テキストデータに変換したものである．論文１編につき１つずつ作成され，通常はテキストファイルとして保存されている．本文データは書誌情報や参照情報の抽出の際に利用される．また，アブストラクトや引用部分近辺の文書も本文データから抽出される．

● 書誌情報
書誌情報は，論文のタイトル，著者，所属などの情報を取り纏めたものである．また，論文ファイルの所在情報も含んでいる．このデータは本文データ１編につき１つずつ抽出され，データベースに登録される．このデータは主に論文検索や時の際に論文ファイルへのリンク表示の際に利用される．このデータはデータソースからローカルサーバに対して公開する．

● 未解決参照情報
未解決参照情報は，ある論文（参照論文）で参照している論文（被参照論文）のタイトル，著者，掲載誌などの情報である．参照論文中の参考文献一覧から抽出される．参考文献ごとに１つずつ作成されるので，論文１編につき参考文献の本数分だけ生成される．このデータはデータソースからローカルサーバに対して公開する．

● 参照関係データ
参照関係データは，参照論文と被参照論文との対応関係である．未解決参照情報のタイトル，著者などと，システムに登録されている論文の書誌情報とを比較し，同定処理を行う．その結果，未解決参照データと被参照論文との対応が取れたものについて，参照論文と被参照論文の ID の対応関係をデータベースに登録する．このデータは，検索結果から参照関係ツリーを表示する際に利用される．

● データソース設定
データソース設定は，ローカル及びインターネットのデータソースに関する設定である．ローカルデータソースに関しては論文ファイル集合の保存場所，インターネットデータソースでは，書誌情報ファイルの URL 定義などを保持している．

（3） システムの構成
本システムの構成図を**エラー! 参照元が見つかりません。**に示す．図中の網掛け部分が今回の実装範囲である．

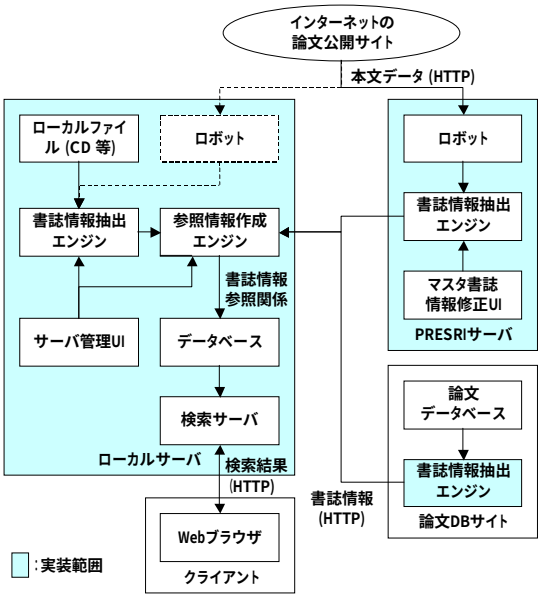

図 1　システム構成

（4） **サーバ構成**

本システムは主に下記のサーバから構成される．

- **ローカルサーバ**
  - 利用者に対して検索機能を提供するサーバである．
  - 大学，研究機関，組織などで別個にサイトを構築し，その組織に所属する利用者にサービスを提供することを想定している．
  - 複数のインターネットデータソースサイトと連携し，外部のサイトが公開する論文を検索することが可能である．
  - CD-ROM などで提供される論文集データをローカルデータソースとして登録し，検索対象とすることが可能である．
  - 利用者からの検索要求をうけて，検索結果や参照情報を送信する．
  - 負荷分散などの目的のため，ローカルサーバを複数台のコンピュータで構成する可能性もある．
  - ローカルサーバの台数構成などについては個別のサイト管理者にゆだねる．

- **PRESRI サーバ**
  - 特定のサイトで動作する論文収集エンジンである．
  - ロボット機能を利用して，インターネット上で公開されている論文を収集し，その情報をまとめて書誌情報として公開する．
  - 論文を収集する際にインターネット上の検索サイトなどを利用する可能性がある．

- **クライアント**
  - 一般的な Web ブラウザが動作するコンピュータである．
  - 利用者が検索機能を利用する際に操作する．
  - またシステム管理者が保守，管理作業を行う際にも利用する．

- **論文 DB サイト**
  - 学会の論文サイトや論文データベースサイトなどである．
  - 大量の論文を保持し公開しているサイトを想定している．
  - ロボットなどが論文を収集することを許容しないサイトである．
  - データソースサイトとしてローカルサーバと連携動作する．
  - ローカルサーバに書誌情報を提供することで，当該サイトが公開している論文書誌情報をローカルサーバから検索することが可能である．

- **インターネット上の論文サイト**
  - 著者などが自著の論文などをインターネットで公開しているサイトである．
  - 少数の論文を公開しているサイトを想定している．
  - 本システムは論文公開サイトから取得した論文を検索対象とすることができる．

（5） **モジュール構成**

本システムを構成するモジュールごとに機能概要を説明する．

- **ロボット**

本モジュールは主に論文収集機能を提供する．
検索エンジンなどを利用してインターネット上で公開されている論文を検索する．検索結果に基づいて，論文ファイルを収集する．このモジュールは原則として PRESRI サーバで動作するが，ローカルサーバで使用することも可能である．

- **書誌情報抽出エンジン**

本モジュールは主に本文データ管理機能及び書誌情報・未解決参照情報抽出機能を提供する．ローカルに保存されている論文ファイルから本文データを抽出し，指定されたディレクトリに保存する．また，抽出した本文データから書誌情報及び未解決参照情報を抽出する．
本モジュールは PRESRI サーバ，論文 DB サイト，ローカルサーバのそれぞれで動作する．PRESRI サーバではロボットが収集してきた論文ファイル集合から書誌情報と未解決参照情報を抽出する．論文 DB サイトではそのサイトが保有する論文ファイルを対象として，書誌情報等を抽出する．PRESRI サーバ及び論文 DB サイトでは抽出した書誌情報及び未解決参照情報をファイルに取り纏め，ローカルサーバに対して公開する．ローカルサーバでは，CD-ROM などの論文集から取得した論文ファイルを対象として，書誌情報等の抽出を行う．取得した書誌情報等はローカルのみで使用する．

- **参照情報作成エンジン**

本モジュールは書誌情報・未解決参照情報収集機能及び参照データ生成機能を提供する．
ローカルサーバに登録されているデータソース設定に基づいて，各データソースより書誌情報及び未解決参照情報を収集し，ローカルのデータベースに登録する．また，ローカルの DB 上に登録された参照情報について，未解決の被参照論文情報と DB に登録されている書誌情報とを比較し，該当する論文が見つかった場合は該当論文の ID を DB に記録する．

- マスタ書誌情報修正 UI

本モジュールは書誌情報管理機能を提供する．論文 DB サイトや PRESRI サーバがローカルサーバに対して公開している書誌情報の抽出エラーなどを修正するインタフェースである．

- サーバ管理 UI

本モジュールは管理画面認証機能，データソース設定機能及び書誌情報管理機能を提供し，ローカルサーバ上で動作する．ローカルサーバ上のデータソース設定や，各データソースからの書誌情報取得の設定，参照情報作成の設定などを変更することが可能である．また，ローカルデータソースからのデータコピーや書誌情報抽出，インターネットデータソースからの書誌情報取得，参照情報の更新を直接実行できる．ローカルに登録されている論文書誌情報について自動抽出のエラーなどを修正する機能も提供する．
データソース管理画面はユーザ認証によって保護されており，管理者以外は設定の変更ができない．データソース管理画面には管理者情報の管理機能も含まれる．

- **データベース**

本モジュールは管理画面認証機能，データソース設定機能, 書誌情報管理機能及び参照情報管理機能を提供する．また，論文検索機能及び参照データ生成機能の一部も提供する. このモジュールはローカルサーバ上で動作する．ローカルサーバが保持する各データソースの設定や管理者の設定情報を保持している．また，参照情報作成エンジンが収集した書誌情報及び，解決済みの参照関係データを登録する．参照関係の解決や論文検索の際には，論文の書誌情報等をキーとして該当する論文のリストを検索する．また，参照関係の表示等の際には論文 ID をキーとし，参照論文及び被参照を検索する．

- **検索サーバ**

本モジュールは，論文検索機能及び参照関係表示機能を提供する．主にローカルサーバ上で動作する．利用者が Web ブラウザを利用して送信した検索要求に対して，該当する論文一覧を送信する．また，論文参照関係の表示依頼に対して，論文の参照データを整形し表示データを送信する．

（6） ハードウェア・ソフトウェア構成

プラットフォーム

本システムでは下記のプラットフォームを対象として想定する．

アーキテクチャ	OS	ミドルウェア等
PC-AT 互換機	Linux	なし
	FreeBSD	Linux thread
	Solaris	なし
	Windows2000	Cygwin
	WindowsXP	
Sparc	Solaris	なし

ソフトウェア・ツール

本システムは下記のツール・ライブラリを必要とする．

- リレーショナルデータベース
  - MySQL もしくは PostgreSQL
- Web ドキュメント収集
  - wget
- テキスト変換
  - prescript (PDF, PS → text)
  - imdkcv (日本語コード → Unicode)
- Web サーバ
  - Tomcat
  - apache
- プラットフォーム
  - sun jre
  - perl

クライアント環境

本システムは管理 UI，検索 UI の表示用ブラウザとして下記のものを想定する．

Web ブラウザ	バージョン等
Netscape Navigator	4.7.*, 6.0, 6.2, 7.0
Internet Explorer	5.5SP2, 6.0
Opera	6.0

（7） システムの制限

認証機構について

本システムでは認証を必要とする諸学会，論文データベース等へのアクセスの認証代行機能は提供しない．従って，これらのサイトから提供される情報を閲覧するためには，別途，各サイトへの認証を利用者が行う必要がある．

論文本文へのリンクについて

論文検索結果などに出力される論文本文へのリンクが直接論文本文を参照できない可能性がある．具体的には下記のようなケースが想定される．

1. 論文検索結果に対する相対的な ID から論文本文を閲覧するサイト
   - 論文本文への直接のリンクを出力することが困難なため，論文検索結果画面へのリンクになる．
   - この場合，論文の検索結果へのリンクが出力される．
   - NACSIS-ELS などが該当する
2. 認証を経由せずに論文本文 URL をポイントするとエラーになるサイト
   - 論文本文へのリンクを出力しても論文を参照できない．
   - 事前にユーザが該当サイトへの認証を行っておく必要がある．
   - この場合，該当サイトの認証画面などへのリンクを出力する．
   - Europhysics Letters などが該当する
3. その他, 論文本文を表示するために認証が必要なサイト
   - 論文本文へのリンクを辿るとそのサイトの認証画面が表示される．
   - 認証を行うことで論文本文の表示は可能である．
   - 多くの学会サイトが該当する．

アブストラクト及び参照部分近辺の引用の表示について

論文の本文データ取得に認証が必要な論文については，

論文検索結果などの表示の際に参照部分近辺の引用などを表示することができない．また，アブストラクトの取得に認証が必要な場合は，アブストラクトについても同様に表示することが出来ない．

（8） 画面の構成

本システムの管理画面の画面遷移を**エラー! 参照元が見つかりません。**，検索画面の画面遷移を**エラー! 参照元が見つかりません。** に示す．

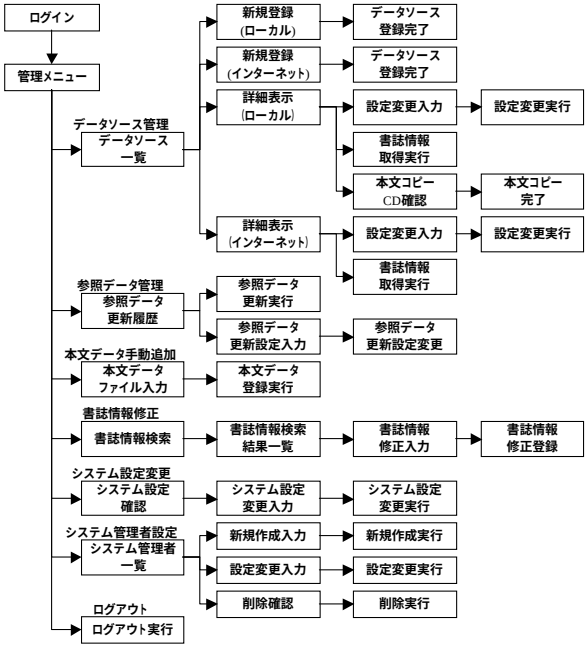

図 2　管理画面遷移図

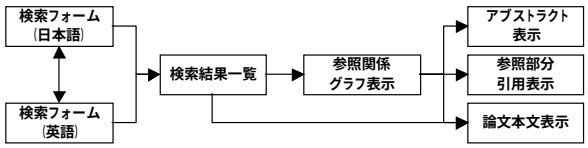

図 3　検索画面遷移図

画面の詳細

各画面の機能について概説する．

管理画面

- **ログイン**
  管理画面のログインフォーム．

- **管理メニュー**
  管理画面のメニューである．メニューとして下記の項目が表示される．
  - データソース管理
  - 参照データ管理
  - 本文データ手動追加
  - 書誌情報修正
  - システム設定変更
  - システム管理者設定
  - ログアウト

- **データソース一覧**
  現在システムに登録されているデータソースの一覧を出力する．管理メニューからデータソース管理を選択すると表示される．

- **データソース新規登録（ローカル）**
  ローカルデータソースを新規に作成する．本文ファイルの保存ディレクトリ等を設定する．

- **データソース新規登録（インターネット）**
  インターネットデータソースを新規に作成する．書誌情報・未解決参照情報の公開 URL 等を設定する．

- **データソース登録完了**
  データソース設定をシステムに登録する．

- **データソース詳細表示（ローカル）**
  ローカルデータソースの現在の設定を表示する．

- **データソース詳細表示（インターネット）**
  インターネットデータソースの現在の設定を表示する．

- **データソース設定変更**
  データソースの現在の設定を変更する．ローカル，インターネットに応じた設定項目のフォームが表示される．

- **データソース設定変更実行**
  データソースの現在の設定変更をシステムに登録する．

- **書誌情報取得実行**
  データソースから最新の書誌情報を取得する．

- **本文コピーCD 確認**
  ローカルデータソースの論文ファイルを再取得する前に CD をマウントするよう促す．

- **本文コピー実行**
  ローカルデータソースの論文ファイルをローカルのディスクにコピーする．

- **参照データ更新履歴**
  直前の参照データの更新履歴及びそれ以降のデータソースの更新履歴を出力する．管理メニューから参照データ管理を選択すると表示される．

- **参照データ更新実行**
  参照データの更新を即時実行する．

- **参照データ更新設定入力**
  参照データの定時更新の設定フォーム．

- **参照データ更新設定実行**
  参照データの定時更新設定を登録する．

- **本文データファイル入力**
  本文データを手動でローカルデータソースに追加する．追加ファイルはローカルのパス，もしくはURL で指定する. 指定された URL が HTML の場合は，その中のリンクをたどる．この画面は管理メニューから本文データ手動追加を選択すると表示される．

- **本文データファイル追加**
  入力された本文ファイルを取得し, ローカルデータソースに追加する．

- **書誌情報検索**
  書誌情報修正を行う論文を検索するフォーム．管理メニューから書誌情報修正を選択すると表示する．

- **書誌情報検索結果**
  論文の検索結果一覧. 論文を選択すると書誌情報修正入力画面が表示される．

- **書誌情報修正入力**
  書誌情報を修正するフォーム. タイトル等の情報を入力できる．

- **書誌情報修正登録**
  修正入力画面に入力した書誌情報を登録する．

- **システム設定確認**
  システムの設定を変更する．書誌情報の保存ディレクトリなどが表示される．

- **システム設定変更入力**
  システム設定を変更する．設定項目の入力フォームが表示される．

- **システム設定変更実行**
  システム設定変更入力画面に入力した設定を反映させる．

- **システム管理者一覧**
  システム管理者を一覧で表示する. 管理メニューからシステム管理者設定を選択すると表示される．

- **新規作成入力**
  システム管理者を新規に追加する．アカウント，ユーザ名，パスワード，表示言語等を選択する．

- **新規作成実行**
  新規作成入力画面に入力されたユーザをシステムに登録する．

- **設定変更入力**
  システム管理者の個別の設定を変更する. ユーザ名，パスワード，表示言語等を選択できる．アカウントは変更できない．

- **設定変更実行**
  設定変更入力画面に入力されたユーザ情報をシステムに登録する．

- **削除確認**
  システム管理者の削除を確認する．

- **参照データ更新実行**
  指定された管理者を削除する．

- **ログアウト実行**
  管理画面からログアウトする．

検索画面

- **検索フォーム**
  タイトル，著者，掲載誌などの検索キーワード入力フォーム．英語と日本語のページがある．

- **検索結果一覧**
  検索フォームに入力された条件に合致する論文を一覧で表示する．

- **参照関係グラフ表示**
  論文間の参照関係をグラフ形式で表示する．個々の論文を年代，類似度に応じて点で表し，その間の参照関係を矢印で表示する．表示範囲については何段階かの切替が可能である．論文や矢印をポイントすると書誌情報等も表示される．

- **アブストラクト表示**
  論文のアブストラクトなど，より詳細な情報を表示する．参照関係グラフからポップアップ的に表示されることを想定する．

- **参照部分引用表示**
  論文の参照部分近辺を引用して表示する．参照関係グラフからポップアップ的に表示されることを想定する．

- **論文本文表示**
  該当する論文ファイルを表示する．ローカルデータソースの論文については直接表示する．また，インターネットデータソースの論文は，論文本文ファイルへ至るリンクを出力する．

（9）開発の体制

各モジュールの開発体制は下記の通りである．

- **ロボット**
  - 開発内容: 既存システムを流用
  - 開発言語: perl
  - 使用ツール: wget

- **書誌情報抽出エンジン**
  - 開発内容: 既存システムを修正

タイトル，著者などを別個に抽出
- Unicode 対応
- 開発言語: perl
- 使用ツール: prescript

- **参照情報作成エンジン**
  - 開発内容: 既存システムを修正
    ロジックは流用
    バックエンドに DB を利用する
    論文及び参照の ID の発行方法を変更する
    書誌情報・未解決参照情報の取得部分を新規に実装
  - 開発言語: perl or java
  - 使用ツール: wget (書誌情報の収集に利用する可能性有り)

- **マスタ書誌情報修正 UI**
  - 開発内容: 新規に作成.
  - 開発言語: java
  - 使用ツール: tomcat

- **サーバ管理 UI**
  - 開発内容: 新規に作成
  - 開発言語: java
  - 使用ツール: tomcat

- **データベース**
  - 開発内容: フリーの RDB を利用
  - 開発言語: SQL, java
  - 使用ツール: MySQL

- **検索サーバ**
  - 開発内容: ほぼ新規作成
  - 開発言語: java
  - 使用ツール: tomcat

４．システムの動作例

以下にシステムの動作例を示す.

図 4　検索画面

今回開発したシステムでは，タイトル，著者名，掲載誌にキーワードを分けて論文検索を行うことができる.

図 5　検索結果

図 5 は 1980 年～2003 年の間で著者名に John を含んだ論文を検索した結果を示している．図より，18 件の論文が検索されていることがわかる.

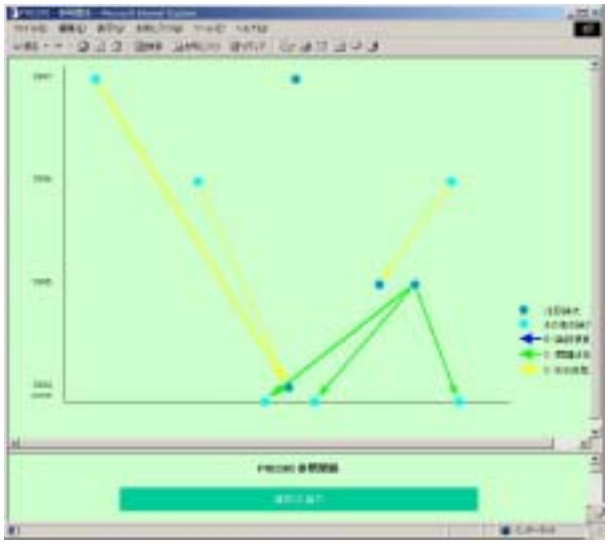

図 6　検索結果のグラフ表示

図 6 は，図 5 の検索結果画面において，チェックボックスでチェックされた論文と，それらに関連する論文を論文データベースから収集し，表示したものである．図において，「注目論文」は，図 5 でチェックされた論文を示している．「その他の論文」は，論文データベースから自動的に収集された，注目論文と関連のある論文を示している．また，図の矢印は論文間の参照関係を示しており，参照タイプ毎に色分けして表示されている.

図 7　グラフ表示における論文情報のポップアップ画面

図 6 中で個々の論文アイコン(ドット)にカーソルを重ねると，図 7 のようにその論文の書誌情報およびアブストラクトがポップアップ表示される．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　参照個所のポップアップ表示

図 6 中で矢印にカーソルを重ねると，図 8 のように参照個所がポップアップ表示される．論文中で複数回参照されている場合には，図に示すように，すべての参照個所が提示される．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 9　システム管理画面(システム設定変更)

システム管理 (システム設定変更) 画面において，PRESRI サーバ上に存在する論文データのディレクトリやログファイルの設定，SQL ドライバ，諸プログラムの所在の設定等を行う．

5．参加企業及び機関

・株式会社デュオシステムズ

6．参考文献

[1] 難波 英嗣，奥村 学: 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(2002)