【パートII：情報分野研究者のためのオンリーワン共有イノベーションプラットフォーム】

3. アカデミックリンケージ：膨大な学術情報へのアクセスを支援するリンケージ基盤

相澤　彰子[*1]　高須　淳宏[*1]　深川　大路[*1]
高久　雅生[*2]　安達　淳[*1]

＊1　国立情報学研究所
＊2　情報・システム研究機構

情報爆発と情報リンケージ

我々は日々の生活の中で，さまざまな情報を参照し活用する．これらの情報は，個人のディスクやWebなどに格納され，各人の使い勝手に合わせた情報空間を構成している．ところが情報爆発のもとでは，この空間が猛烈な勢いで膨張する．その隙間は複製や無関係な情報などで埋められていき，結果として必要な情報は遠ざかり，情報入手のコストは高くなる．

膨大な情報空間の中に断片化された情報をどのようにしてつなぎ合わせるか，我々の研究ではこの問題に取り組んでいる．その中心となるのは，ヒトやモノなど，現実世界の同じ対象物を参照する記述の同定である．このような同定によって，離れた情報源どうしを結ぶリンクが新たに生成されることから，我々はこれを「情報リンケージ」と呼んでいる．データベースの分野では，重複して登録されたレコードの検出処理をレコードリンケージと呼ぶが，情報リンケージはリンケージの対象を，属性が厳密に定義されたレコードから属性が明示されないテキストへと拡張したものである．なお，ヒトやモノの間には，同一性だけではなく種々の関係を想定することが可能であるが，我々の情報リンケージでは現在，同一性だけをリンクとして扱っている．

情報リンケージは，①候補ペアの抽出とマッチング関数の適用による同一性の判定，および②抽出したリンクに基づく情報の集約・分析の2つのステップから構成される．ここで，我々が特に注目するのは，大規模なデータベースを利用した信頼性の高い情報リンケージ技術の開発である．これに基づき現在，既存のデータベースを中心に他のデータベースやWeb上の情報を結び合わせ，散在する情報を集約するための基盤技術を開発している．本稿では以下，文献（モノ）や研究者（ヒト）を対象とした「アカデミックリンケージ基盤」を中心に，①のリンク抽出処理を支える書誌リンケージエンジン，および②の適用としての研究者情報サーバとコミュニティ分析の例を紹介する．

アカデミック情報のリンケージ

リンケージの中心となるデータベースとして用いるのは，国立情報学研究所が事業サービスとして提供する書誌データベースである☆1．2008年現在，約1100万論文の書誌データが公開されていて，引用同定に用いる非公開データと合わせて5000万件以上の書誌データが登録されている．各書誌データには，論文題目や掲載雑誌に加えて，著者の氏名および所属機関の情報が登録されており，延べ数で1億1千万人以上の著者情報が存在する．これらの情報の属性や表記は必ずしも統一されていない．特に著者について，同一表記だが異なる人物を参照するもの，表記は異なるが同一人物を参照するものなどが混在する．これは，出版社や検索エンジン会社が提供する多くの大規模な電子図書館の現状と同様である[1)]．

アカデミックリンケージの目標は，データベースに登録された論文やその著者に関する情報を，他のデータベースやWeb上の情報に対応づけることである．これにより，著者や所属機関の同定，関連プロジェクトの情報収集等が可能になる．また，書誌データには各々の学会誌の方針に従い，キーワードや抄録，電子図書館へのURLなどが対応づけられており，豊富なテキスト情報に基づく分析が期待される．

書誌リンケージエンジン

異なる情報源からの情報が混在する書誌リンケージでは，入力テキスト中の表記の揺らぎや文字誤りにも柔軟に対応しなければならない．さらに，Web上の文書を

☆1　国立情報学研究所 NII 論文情報ナビゲータ，http://ci.nii.ac.jp/

そのまま入力とする場合などは，情報の欠落や属性の並び順の入れ替えなどに対する配慮も必要である．

我々が提案する高速リンケージ方式では，サフィックスアレイと呼ばれるデータ構造上で，可変長の単語並びに対する仮想的な転置ファイルを操作することで，高速で柔軟性の高い検索を実現する[2]．直観的には，「を利用した物体抽出」のように，テキスト中に埋め込まれた特定力が高い単語列を利用して，検索コストの削減やノイズへの対応を行う．このリンケージ方式を用いることで，入力テキスト全体ではなく一部に照合するレコードが出力され，また，類似レコードのランキングではなく同一と判定されたレコードが出力結果として得られる．

上記の方式を用いて我々は現在，テキストを書誌データベースのレコード集合に対応づける書誌リンケージエンジン i-linkage を開発している．i-linkage システムへの入力としては，他のデータベースから抽出したレコード，Web やローカルディスク上のテキスト文書，電子図書館の画像を OCR で自動認識した結果[3]，検索エンジンや電子図書館の検索結果など多様な形態が可能である．**図-1** に i-linkage システムの出力例を示す．「坂和正敏，田中雅博」に対する検索エンジン☆2 の出力を表示画面から直接切り貼りして，レコード数が 2～3 千万規模の図書および論文データベースへのリンケージを行ったものである．この程度の規模であれば，ブラウザを介して待ち時間を意識しない速度で応答が得られる．図より，i-linkage システムでは，ノイズに強い柔軟な同定が行われていることが分かる．たとえば 3 番目の例において，入力は論文誌に掲載された書評への参照文字列である．専門書の書誌と論文の書誌が併記される複雑な構造を持つが，i-linkage システムからは同定結果として，図書データベースからは紹介対象となった専門書が，論文データベースからはその専門書を紹介した学会誌記事がそれぞれ出力されている．

現在，国立情報学研究所が提供する論文データベース約 1100 万論文の書誌レコードを読み込んだ試作版サーバを公開しており，さらに検索効率や性能を強化した改良版の開発を進めている．ここで，現在の実装は入力テキストにおける「著者」や「タイトル」などの並び順をあらかじめ指定する形になっている（指定は複数可）．柔軟性を持たせるため別途，木構造データの同定において有効な類似度の計算法について検討を進めており[4]，その適用が今後の課題である．

図-1　i-linkage システムの出力例

著者リンケージと研究者情報サーバ

既存のサーチエンジンを超える機能としてエンティティサーチが注目を集めている．これは実世界に存在するヒトやモノといったエンティティを単位として認識し，検索を行う機能を指す．ここではアカデミックリンケージの実践例として，エキスパートサーチと呼ばれるタスクをとりあげる．たとえば，自らが専門としない領域で疑問があった場合に，その分野にくわしい専門家の意見を聞くことは日常的に行われる．このような専門家探しの支援がエキスパートサーチの目的である．

我々は，このようなタスクに対応するための基盤として，15 万人規模の研究者情報サーバを構築している（**図-2**）．これは，国立情報学研究所が提供する科学研究費補助金データベース☆3 における成果報告書，実績報告書をもとにして，研究者に関する情報を，科研費研究者番号をキーとして抽出し再構成して提示するものである．各研究者についてユニークに割り当てられた研究者番号をキーとすることにより，異なるプロジェクトに関する情報を横断的に提供できるようになり，また，他の情報源とのリンケージも合わせて行うことにより，より効果的な情報提供が行えるものと考えた．

このプラットフォームでは，科研費研究課題に関する情報だけでなく，氏名，所属，職位，研究分野，研究歴，発表文献の一覧やホームページ URL といった当該研究者に関する情報を他の論文データベースや Web 上の情報源とのリンクを通じて，概観できる．また，基本的情報を提供するだけではなく，Web 等に存在する研究者情報とリンケージすることにより，これまでまとまった情報を入手できなかった研究者情報について，さまざま

☆2 Google Scholar，http://scholar.google.co.jp/

☆3 科学研究費補助金データベース，http://seika.nii.ac.jp/

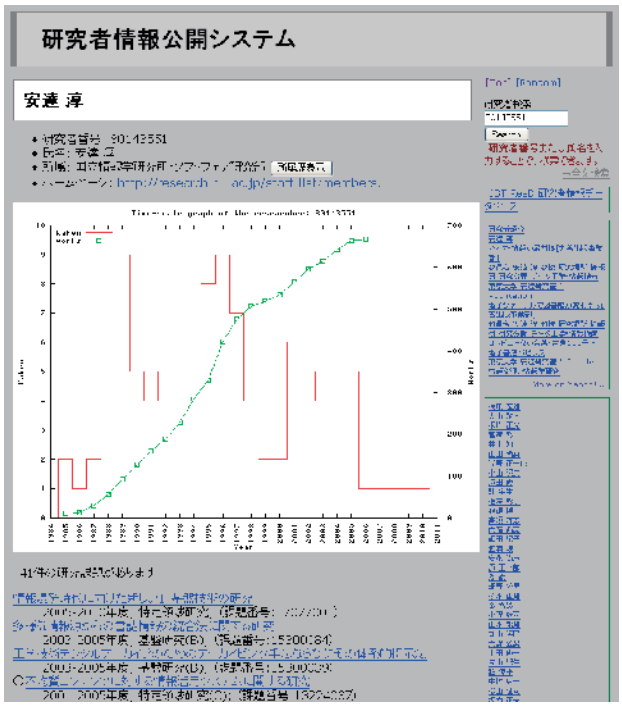


図 -2　研究者情報サーバの出力例

な情報源からの内容を統合し，その全体像が分かりやすいよう提示する，一部の情報については可視化を行うといった機能も盛り込んだ．さらに，科研費報告書に含まれる氏名・所属等の研究者基本情報のブラウジング機能とともに，時系列における課題採択件数や論文報告件数の推移の可視化，共同プロジェクト参画に基づく研究者ネットワーク情報の提示，報告論文の簡易同定，Web 検索エンジンとの連携，Web 外部データベースからの情報取得，といった機能を実装した．

今後，このプラットフォームのもとに，論文書誌情報とのリンケージおよび共同研究者ネットワークの情報を活用した本格的な専門家探しシステムの構築を予定している．

研究者コミュニティ分析への適用例

データベースと Web の間での情報リンケージの例として，特定領域「情報爆発 IT 基盤」の参加研究者 319 名について，領域の Web ページ☆4 に掲載されている氏名と所属機関，および年度報告書に記載されている成果文献を手がかりに，書誌リンケージエンジンによる情報収集を行い，分析を試みた例を紹介する．

まず，研究グループごとに掲載された成果文献を書誌リンケージエンジンでデータベースに対応づけ，さらにデータベースに登録された著者情報を手がかりに各研究者ごとの主著・共著論文を抽出した．次に，抽出した論文リストに基づきクエリを生成して API 経由で検索エンジンに問合せを行い，各研究者の論文を多く含む URL を得た．具体的には，研究室の成果リストや個人の発表文献リスト，会議プログラムなどのページが該当する．これらを書誌リンケージエンジン経由でさらに読み込むことでデータベース上の書誌と対応をとり，合わせてデータベースに登録された引用－被引用関係や共著関係からさらに論文を追加する処理を繰り返した．これによって，当初の 3114 文献（延べ数）から 26264 文献（延べ数）を自動的に獲得した．

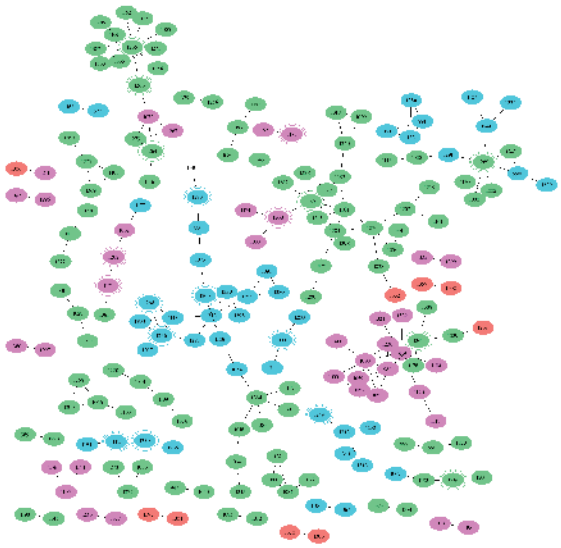
上記において，入力として与えた成果リストは 2006 ～ 07 年の特定領域関連の発表だけで，各研究者との対応づけもなされていなかったが，最終的に得られる文献リストは研究者ごとにまとめられ，1980 年代まで遡るものである．獲得した文献情報から，研究者ごとの所属機関履歴や氏名表記の揺れ，関連 URL などの情報が得られる．さらにデータベースに登録されている抄録や著者キーワードや学会名などを利用すれば，専門分野のキーワード抽出，コレスポンデンス分析による関連研究者や分野の分析など各種分析法の適用が可能になる．

一例として，特定領域参加者の共著関係ネットワークの時間変化を**図 -3** に示す．ノードが研究者，リンクが共著関係，ノードの色が領域内の異なる研究グループを示す．図 -3（a）が，領域発足前の 2005 年，図 -3（b）が 2007 年 12 月時点での結果に基づく．共著関係という切り口から，領域発足後に分野をまたぐ領域内の結びつきが活性化したことがうかがわれる．新たに追加されたリンクを調べると，支援班における共同研究および若手の移動などが要因としてあげられる．

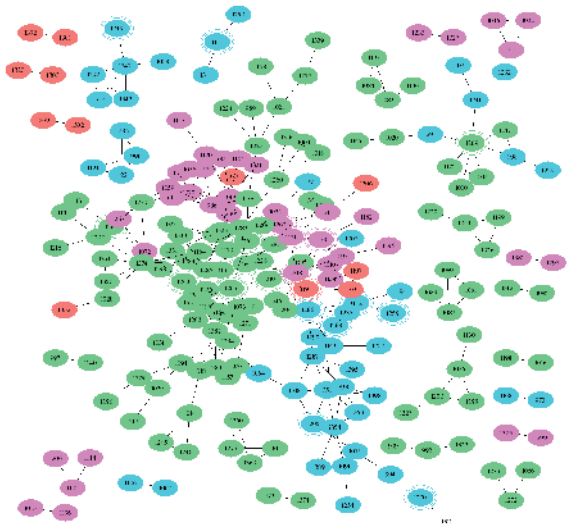
単純に氏名表記を用いた検索では，同姓同名や異体字・表記揺れの問題に対応できないため，何らかの形であいまい性解消の手段が必要である．ここで適用した手法は，データベースと Web 上の情報をうまく組み合わせて，この問題に対応するものといえる．科学計量学の分野では共著や引用関係等を利用した科学的生産活動の計量的な分析が行われるが，分析に必要なデータを自動生成することは容易ではなく，本手法はこのような分析においても人手によるクリーニング作業を大幅に削減することが期待できる．なお，このシステムにおける著者同定の処理は，本特集でも紹介がある共有計算機環境 InTrigger☆5 上で分散実行可能であり，東京大学田浦研究室の協力を得て GXP や dds ライブラリ利用実験も行っている．

☆4 http://www.infoplosion.nii.ac.jp/info-plosion/

☆5 https://www.logos.ic.i.u-tokyo.ac.jp/intrigger/



(a) 2005 年時点での共著関係ネットワーク



(b) 2007 年時点での共著関係ネットワーク

図 -3　共著関係ネットワークの時間変化

今後の展開

本稿では，ヒトやモノなどの同定を軸に断片化された情報を結ぶ情報リンケージの要素技術について述べ，研究者や論文情報を対象としたアカデミックリンケージ基盤の構築に向けた実践の試みを紹介した．本研究による成果は，国立情報学研究所学術コンテンツサービス研究開発センター，および情報・システム研究機構新領域融合研究センターを通して，より実証的なシステムへと展開をはかる予定である．

最後に，日頃から貴重なご議論とコメントをいただいている領域代表の喜連川優教授，爆発 NLP/IR の参加メンバ，国立情報学研究所の大山敬三教授，井上雅史助教，InTrigger の利用について多大な支援をいただいた東京大学の倉沢央氏，弘中健氏，および Web ページの管理に日々貢献されている情報爆発 IT 基盤事務局の皆様に感謝の意を表する．

参考文献

1) Lee, D., Kang, J., Mitra, P., Giles, C. L. and On, B. : 2007 : Are Your Citations Clean?, Commun. ACM 50, 12, pp.33-38 (2007).
2) 相澤彰子，高久雅生，大山敬三：大規模データベースを利用したリンケージシステムの提案と実装，DBSJ Letters, Vol.6, No.4, pp.17-20 (2008).
3) Takasu, A. : Bibliographic Attribute Extraction from Erroneous References Based on a Statistical Model, Proc. of ACM & IEEE Joint Conference on Digital Libraries, pp.49-60 (2003).
4) Takasu, A., Fukagawa, D. and Akutsu, T. : Statistical Learning Algorithm for Tree Similarity, IEEE ICDM, pp.67-72 (2007).

（平成 20 年 5 月 13 日受付）

相澤　彰子（正会員）：aizawa@nii.ac.jp
国立情報学研究所教授．言語コンテンツを中心に情報検索，テキスト処理，知識処理などの研究に従事．

高須　淳宏（正会員）：takasu@nii.ac.jp
国立情報学研究所教授．データ工学，テキストマイニングなどの研究に従事．

深川　大路（正会員）：daiji@nii.ac.jp
国立情報学研究所特任研究員．木構造データに対するマッチングアルゴリズムや確率モデルなどの研究に従事．

高久　雅生（正会員）：masao@nii.ac.jp
情報・システム研究機構新領域融合研究センター融合プロジェクト特任研究員．情報検索や電子図書館システムの研究に従事．

安達　　淳（正会員）：adachi@nii.ac.jp
国立情報学研究所教授ならびに学術基盤推進部長．情報検索，電子図書館システムなどの研究開発に従事．