サブテーマ１：
大規模・異種情報の収集・解析・結合・分類の手法および知識基盤の構築

研究代表者：

[国立情報学研究所] 高野 明彦

共同研究者:

[国立情報学研究所] 西岡 真吾、佐藤 真一、相澤 彰子、根岸 正光、安達 淳、孫 媛、
西澤 正己、高須 淳宏

[国立遺伝学研究所] 大久保 公策

[統計数理研究所] 馬場 康維、石黒 真木夫、土屋 隆裕

[新領域融合研究センター] 高久 雅生（国立情報学研究所在勤）

１．研究目標

本サブテーマでは、分野横断型融合研究を実効的に推進するための情報空間・情報基盤構築を目指して研究を進める。このための主たる要素技術である①異種情報の結合・分類手法、②大規模リンケージ情報の収集・分析手法、の二つの研究項目を中心に取り組む。

研究項目①では、分野横断型融合研究のための実用的な情報空間を構築する方法について実証的に追及する。特に、論文や専門辞書、教科書に記載されている専門的な知識記述と、一般の科学雑誌や新聞などの非専門家向けの知識記述を横断的に大規模に収集して、それらを研究者の自由な発想で動的に結合･統合する手段を実現する。3 年後には、下記のような大規模な異種データ群に対して、情報の類似性に基づく連想的な結合・融合と自動分類を可能とする。

論文（学術論文 500 万件のフルテキスト等)、
専門辞典（理化学辞典、数学辞典、情報科学辞典等)、
教科書（大学講座シリーズ等)、新聞データ（600 万記事)、
書籍（日英書籍 900 万タイトルの目次・概要)、特許情報（100 万件）

研究項目②では、このような情報空間の中核となっている要素間の参照やつながりを表わす情報を「リンケージ情報」と総称し、このリンケージ情報を収集・解析し、活用するための横断的な研究を行う。まず、機械学習や情報検索の最新の成果に統計分析的な観点を導入することで、効率的で対象データに依存しないリンケージ情報収集・処理技術の開発を目指す。また、引用文献によるリンク構造や研究者同士の関係ネットワークに注目して、情報や統計をはじめとする各学問領域の研究者と協力してビブリオメトリックス分析を行う。これにより、融合分野における学問分野の構造的変化、研究コミュニケーションネットワークの形成過程, 研究の国際連携・セクター間の連携の実態などの解明を図る。

２．年次研究計画

研究項目①では、分野を横断して存在する異種データ（例えば、論文や専門書等の専門知識と科学雑誌や新聞等の一般知識）を、研究者の自由な発想によって、動的に結合・統合すること

を可能にする手段の実現に向けた研究を進める。その際、異種データの情報内容の類似性に基づく連想的な結合・統合と自動分類などの手法に重点を置いて、手法の提案と、そのシステム実装による実証を行う。

平成１８年度に、連想的に結合された異種データの有用性を示すため、環境問題を例題に取り、論文、専門辞典、教科書、新聞記事、書籍、特許など多様な情報を、テーマ別に自動分類する方式について研究する。また、研究者間のコミュニケーションがますます困難になりつつある生命科学分野で、困難さの原因が固有名称の多用や遺伝子の機能構造に関する自然言語表現にあることに注目して、それらを自動的にアイコン化して分野間のギャップを埋めるジーンアイコン（遺伝子象形文字）プロジェクトを推進する。平成１９～２０年度には、環境関連情報と生命科学分野の研究情報を例にとり、異種情報源から情報内容の類似性に基づいて関連情報を収集し、それらを概観しやすい形で提示する情報システムを試作する。専門辞典などを軸に関連情報を動的に整理して提示する。また、ジーンアイコンを活用して、文献の深い理解に必要となる遺伝子等の情報について、最新の関連データの内容を略図表示するシステムも試作する。

研究項目②では、わが国の学術活動に焦点をあてて、リンケージ情報の収集・分析を実践的な側面から評価する。研究者や研究機関同士の連携を共著・引用関係に基づき数量化することや、引用数やインパクトファクタにより研究費配分の効果を数値的に分析することによって、わが国の学術活動の構造を解明し、学術政策の提言を目指す。このため、既存のデータベース資源として別個に存在する研究費データベースや引用索引データベース等の多様な情報源のリンケージによって、情報の付加価値を生み出し、これを分析対象として研究を進める。また、リンケージに必要な多くの分野固有知識と人手による作業に伴うサンプル数の限界を克服し、既存のデータベースに留まらないリンケージ対象の拡大ため、機械学習、マイニング手法、文字列データの解析技術や高速検索技術を活用する。

平成１７～１８年度では、リンケージ情報を機械的かつ大規模に収集するための機械学習・マイニングの要素技術を研究するとともに、国立情報学研究所の国内論文誌引用索引データベース(CJP)および科学研究費補助金データベースの各々を対象とした学術動向の調査分析を行う。また、人間の作業者による信頼度の高いサンプル基礎データの構築を進める。特に、平成１８年度には、科学研究費補助金データベースと CJP および Thomson Scientific 社の SCI（国際的な論文誌）の２つの引用索引データベースとの間のリンケージ法の検討に重点を置いて進める。また、数十人（～数百人）程度の規模の研究者をサンプルとして研究者や所属機関履歴に識別子を付与したオーソリティデータを作成し、リンケージ情報抽出技術とリンケージ情報分析技術の融合を試みる。さらに、統計分野の研究者と協力して、統計分野の分析とその実践的な効果について予備的な調査研究を行う。平成１９年度以降には、平成１８年度で得られる知見に基づき、研究者のオーソリティデータの規模や分野を拡大し、分野やセクター間での連携・融合の分析を進める予定である。また、対象を他の情報源に拡大し、リンケージの新たな可能性の検討を行う。なお本研究では、情報・統計分野の関連研究者での情報交換を目的として、「大規模データ・リンケージ、データマイニングと統計手法」と称する研究会を企画して、年１回のペースで開催す

る予定である。

３．平成１７年度の研究進捗

研究項目①の異種情報の結合・分類手法の研究においては、手法の有効性を検証する準備として、商用の環境情報ポータルで提供されている各種最新情報、環境問題資料集成として出版されている条約、法令、政策文書、議会議事録等の基本資料について、研究目的利用の許諾を得て、実証実験に着手した。これらの異なる情報源をユーザの興味により関連づけて提示する利用環境のプロトタイプを開発した。特に、特許情報と論文情報について、本利用環境の有効性を検討した。

また、研究者間のコミュニケーションがますます困難になりつつある生命科学分野で、困難さの原因が固有名称の多用や遺伝子の機能構造に関する自然言語表現にあることに注目して、それらを自動的にアイコン化して分野間のギャップを埋めるジーンアイコン（遺伝子象形文字）プロジェクトをスタートした。文書にデータを盛り込むこのアプローチは、科学的知識の共有に適した新しい文書表現の可能性を示唆している。

研究項目②の大規模リンケージ情報の収集・分析手法の研究では、リンケージ情報抽出のための基盤技術、リンケージ情報の分析を重点に研究を進めた。リンケージ情報抽出のための基盤技術に関しては、多様な記述形式の混在環境における高精度のリンケージの実現に必要な文字列情報の構造化のための学習手法の研究を進め、文字列データの解析と同一ラベルの要素同士の対応比較処理のための近似的な構文解析法の具体的な適用例として、引用文献文字列を分割し、書誌項目に該当する部分文字列をラベリングするための方法を考案した。また、大規模リンケージの取り扱いにおいて必要となる大量の文字情報からリンケージ候補文字列を高速に検出する処理として、複数の情報源による大規模データベース同士での重複エントリの検出に焦点をあて、サフィックスアレイと呼ばれるデータ構造を利用した高速リンケージエンジンの実現方法を検討した。さらに、リンケージ対象として特に重要な「人物名」を対象として、名前表記が等しい複数の人物名を、同一人物ごとにまとめるためのクラスタリング手法を開発した。

抽出したリンケージ情報の分析に関しては、日本の学会誌論文を対象とした「引用文献索引データベース」(CJP)に基づいて、名寄せ作業・所属機関の同定およびセクター分類方法の検討等を行った。具体的には、分析のための基礎データ作成と併行して、このデータを用いた論文の共著関係の定量的解析を行うことによって、研究ネットワークの実態分析に着手した。また、科研費採択課題における総合領域、複合新領域の分野細目間の関連を分析するための研究者履歴調査（500 名弱分）を行い、研究成果データベースからの著者情報の同名他者の識別研究のための基礎データ作成、および、研究者－研究分野の解析に必要な日本統計学会を含む６関連学会の名簿情報のデータベース化に着手した。

４．平成１７年度研究成果

（１）知見・成果物・知的財産権等

- 「想・IMAGINE」のプロトタイプ

  異なる情報源をユーザの興味により自在に関連づけられる新技術「想・IMAGINE」のプロトタイプを作成して、連想計算によるデータベースの動的連携の有効性を確認できた。

- ジーンアイコン（遺伝子象形文字）提案

  ジーンアイコンの効果を検証するため、文書中に自動的にジーンアイコンを挿入・表示する簡易サーバを作成した。タンパク名などの専門用語が正しく認識された例については有効であることが確認できた。しかし、実際の論文では同一概念について多用な表現が使用されているため、現在使用中の専門用語辞書のみでは精度が低い。この実用上の限界の解決を目指し､精度をある程度保証できる遺伝子名称のアクロニムに限定して辞書を整備して、実用水準まで高めることを目指す。

（２）成果発表及び著書執筆等

- Hai Yen Siew and Yasumasa Baba: “Linear regression for a mixture of continuous and binary data,” The 5th IASC Asian Conference on Statistical Computing, 2005.12.17, 香港大, 香港, 中国.
- Hai-Yen Siew and Yasumasa Baba: “Regression with some observations below the threshold,” Second German Japanese Symposium on Classification 2006, 2006.3.7, 日独センター, ベルリン, ドイツ.
- Atsuhiro Takasu, Kenro Aihara: “Bibliographic Component Extraction from References Based on a Text Recognition Error Model,” Systems and Computers in Japan, Vol.36, No. 7, pp.1-12, 2005.
- Atsuhiro Takasu: “A Sequential Labeling Method Using Synactical and Textual Patterns for Record Linkage,” Lecture Notes in Computer Science 3686 (Proc. 3rd ICAPR), pp. 199-208, 2005.

５．その他

（１）実用技術重視の研究推進

特に、研究項目①では、論文、専門辞典、教科書、新聞記事データ、特許情報、条約・法令など多様な情報源の新しい利用環境を追求している。研究成果の実用性を検証しながら研究を進めるため、実際に専門家が業務で参照している高信頼な情報源を導入している。同種の研究においては、論文執筆を優先して、規模や内容が現実の情報源とは大きく異なるものを利用して実験する場合も多いが、本項目は提案技術の実用性を重視して、導入済みの現実の情報源を利用して研究を推進している。

（2）研究会「大規模データ・リンケージ、データマイニングと統計手法」の開催

研究項目②に関して、2006 年 2 月 20 日（月）、21 日（火）統計数理研究所において標記研究会を以下のように開催した。

- 竹田隆治、高須淳宏: 「時系列テキストストリームからの単語共起を使った新情報の検出方法」, 第一回大規模データ・リンケージ、データマイニングと統計手法予稿集, pp.45-50, 2006.
- 孫媛, 西澤正己, 根岸正光: 「日本の引用文献索引データベースを用いた産学連携の現状分析」, 第一回大規模データ・リンケージ、データマイニングと統計手法予稿集, pp.51-55, 2006.
- 西澤正己、孫媛: 「科研費採択課題における総合領域、複合新領域の関連分析」, 第一回大規模データ・リンケージ、データマイニングと統計手法, 2006（口頭発表）.
- 根岸正光: 「発表論文からみた研究分野構成の類似度と引用度指標による大学間競争関係析出の試み」, 第一回大規模データ・リンケージ、データマイニングと統計手法予稿集, pp.57-64, 2006.
- 相澤彰子: 「大規模異種データベース間でのレコード同定手法とその適用例」, 大規模データ・リンケージ、データマイニングと統計手法予稿集, pp.85-90, 2006.
- 馬場康維: 「研究者マップについて」, 第一回大規模データ・リンケージ、データマイニングと統計手法予稿集, pp.91-92, 2006.
- 石黒真木夫: 「情報抽出」, 第一回大規模データ・リンケージ、データマイニングと統計手法予稿集, pp.107-108, 2006.
- 鈴木康平、正田備也、高須淳宏、安達淳: 「書誌情報の共著関係を用いた著者同定に関する研究」, 第一回大規模データ・リンケージ、データマイニングと統計手法予稿集, pp.109-117, 2006.