サブテーマ名：　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：　高野明彦　［国立情報学研究所］

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。最終年度の目標として，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どうし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．年度研究計画

<u>平成１７年度</u>

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，このためのシステムや実証実験の設計を行う。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目指す。

また，わが国の学術活動に焦点をあてて，研究者や学術的な成果物を中心とする大規模な情報収集のための情報・統計処理技術について研究を進め，研究者や研究機関同士の連携や研究費配

分効果の実践的な分析を可能にするためのデータ収集，手法の提案と，そのシステム実装による実証を目指す。

平成１８年度

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。

さらに，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科学研究費補助金データベースを利用した研究者基礎データの構築，および，日米の引用索引データベースを利用した学術構造の分析について調査研究を進める

平成１９年度

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。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。

また，平成１９年度に試作した研究者情報サーバを中核として，書誌データベースやWebなど外部の情報源との情報統合について検討を進める。特に，別途開発した書誌同定サーバと連携させ，論文の著者IDを自動認識し統合するための手法の確立を目指す。また，研究課題の代表者と分担者の関係に基づく研究者ネットワークを構成し,研究者コミュニティの抽出や類型化等のネットワーク分析を行う。名簿マッチングに代表されるように人を中心としたコミュニティ相関分析について理論面での検討を行うとともに，引き続き学術構造分析についての研究を進める。

平成２０年度

異種情報の結合・分類手法の研究においては，専門性の極端に異なる情報源の間での連想計算について追求する。これには，専門辞典における用語の説明文を手がかりに，専門性の高い用語の内容を一般的な言葉で表す方法を検討する。また，専門性の極端に異なる情報源の間では用語の違いによる類似性の見落としに関して検討する。さらに，用語集合が極端に異なる例の一つとして，日本語版と英語版のウィキペディアを取り上げ，その間の連想計算の精度向上を目指す。

大規模リンケージ情報の研究では，平成１９年度に統計分野を対象に試作した研究者同定ツールを，他分野に応用させる方策を検討し，分野に関わらず研究者を同定する枠組みを確立する。

リンケージエンジンとあわせて,研究者およびその研究成果としての書籍･論文間の情報リンケージの全体像を確立する。これと並行して，研究者ネットワークと研究者発信コンテンツの内容分析を組み合わせた研究者推薦システムのプロトタイプを試作する。さらに，これまで培った情報リンケージ手法を Web 上の情報源に適用する方策を検討する。また，引き続き分野毎の差異や経年変化についての分析を行うとともに，得られた全体像を俯瞰して分析可能とするツールの試作を検討する。

平成２１年度

大規模な異種データ群に対して，論文（学術論文５００万件のフルテキスト等)，教科書（大学講座シリーズ等)，新聞データ（６００万記事）をはじめとする信頼できるデータベースをコア・データベースとする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。

平成２０年度で試作した推薦システムのプロトタイプを評価し，公開可能，限定的公開可能な情報を選別した上で試行運用を開始する。計算機により自動生成したデータは，コスト／スケール／網羅性の面では圧倒的に優位であるが，情報の誤りや欠落を多く含むことも事実である。品質にばらつきのある大量の情報が流通する今日において，同様の状況は様々な分野・局面において見られ,「少数・高品質」ではなく,「多数・低品質」なデータからの価値創出は，これからの情報社会における重要なポイントの１つである。情報学と統計学の融合領域を目指す本プロジェクトを通して，この問題に踏み込むためのアプローチを模索して行きたい。

平成２２年度以降の展開

本研究で開発した異種情報の結合・分類手法は，実用性の高い汎用技術になると考えられる。特に「想・IMAGINE」の対話環境は，複数の情報源を動的に結合して，ユーザの問題分析に最もふさわしい情報源を作り上げられる点で，分野横断型融合研究者のための情報環境に適している。プロジェクト終了後は，いくつかの研究分野について実際に研究者が利用できる情報サービスの展開を目指す。

また,「論文」や「研究プロジェクト」など，メタデータがデータベースの形で管理されている事物（エンティティ）を対象として，それらを結ぶ関係を自動獲得して分析する手法を検討してきた。これはデータベースと外部情報の対応づけ技術であるといえる。今後は，対象を拡大して，一般のテキストの中で言及される事物と外部情報の対応づけについて検討を進め，テキスト理解に向けた新たな試みとしたい。

３．研究経費の推移

平成１７年度実績：　50,381 千円

平成１８年度実績：　48,361 千円

平成１９年度見込：　45,000 千円

４．平成１９年度の研究実施体制

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　　丸川雄三　　相澤彰子　　根岸正光

安達淳　　大山敬三　　孫媛　　西澤正己　　高須淳宏　　市瀬龍太郎　　柿沼澄男

［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　　清水信夫

水田　正弘（北海道大学，統計数理研究所客員教授）

［情報システム研究機構］　高久雅生

５．平成１９年度研究成果

（１）成果物（知見・成果物・知的財産権等）

1. 「2007 年度大規模データ・リンケージ，マイニングと統計手法」研究会予稿集（2008）

（２）成果発表等

＜論文発表＞

〔学術論文〕

〔会議録〕

1. Nobuo Shimizu, Masahiro Mizuta: “Functional clustering and functional principal points”, Lecture Notes in Artificial Intelligence 4693, pp.501-508. Knowledge-Based Intelligent Information and Engineering Systems: KES2007 - WIRN2007. (2007)
2. Nobuo Shimizu: “Local solutions in functional k-means clustering”. Proceedings of the Ninth Japan-China Symposium on Statistics, pp.261-264, (2007)
3. Atsuhiro Takasu, Kenro Aihara: ″A Smoothing Method for a Statistical String Similarity″, Proc. IEEE Intl. Conf. on Information Reuse and Integration (IRI2007), pp.667-672, (2007).
4. 相澤彰子，大山敬三，高久雅生:「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」データベースと Web 情報システムに関するシンポジウム（DBWeb2007）(2007)
5. 孫媛,西澤正己，柿沼澄男，根岸正光:「学術論文の共著関係からみた日本の産学連携」。新領域融合プロジェクトによる研究会「大規模データ・リンケージ，データマイニングと統計手法」予稿集，2008 年 1 月 28・29 日，統計数理研究所，pp.13-22.（2008）
6. 相澤彰子，高久雅生，大山敬三:「書誌リンケージエンジンの開発と著者マッチング問題への適用」2007 年度新領域融合プロジェクトによる研究会「大規模データ・リンケー

ジ，データマイニングと統計手法」予稿集，2008 年 1 月 28・29 日，統計数理研究所，pp.23-30. (2008).

7. 高久雅生，相澤彰子，大山敬三，馬場康維:「統計分野における研究者の氏名同定と応用」2007 年度新領域融合プロジェクトによる研究会「大規模データ・リンケージ，データマイニングと統計手法」予稿集，2008 年 1 月 28・29 日，統計数理研究所，pp.31- 36. (2008).
8. 石黒真木夫:「モデルとプログラムとデータのグラフ構造」，2007 年度新領域融合プロジェクトによる研究会「大規模データ・リンケージ，データマイニングと統計手法」予稿集，2008 年 1 月 28・29 日，統計数理研究所，pp.51-52. (2008).
9. 金城敬太，古川康一，相澤彰子:「帰納論理プログラミングによる定性ネットワーク分析」，2007 年度新領域融合プロジェクトによる研究会「大規模データ・リンケージ，データマイニングと統計手法」予稿集，2008 年 1 月 28・29 日，統計数理研究所，pp.53-58. (2008).
10. Naofumi Sakaguchi, Yasumasa Baba; “Estimation of Number of Common Elements in Several Sets”, 2007 年度新領域融合プロジェクトによる研究会「大規模データ・リンケージ，データマイニングと統計手法」予稿集，2008 年 1 月 28・29 日，統計数理研究所，pp.59-64. (2008).
11. 清水信夫:「関数クラスタ分析における局所解の出現個数に関する分析」，2007 年度新領域融合プロジェクトによる研究会「大規模データ・リンケージ，データマイニングと統計手法」予稿集，2008 年 1 月 28・29 日，統計数理研究所，pp.75-80. (2008).
12. Van B. Dang and Akiko Aizawa: “Multi-class named entity recognition via bootstrapping with dependency tree-based patterns”, the 12nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (2008) (accepted).

〔解説・総説〕

1. 高野明彦：検索から連想へ―情報を発想力に変換する連想エンジン，岩波「科学」，2007.4(Vol.77 No.4).
2. 高野明彦＋高橋真理子：検索から連想へ―ひらめきをもたらす情報技術，NII Today 36 (2007.6).
3. 高野明彦:思考を深めるための情報源を探す―情報を発想力に変換しよう，AURA (2007.6).
4. 高野明彦:「連想の情報学」―思考と響きあう情報空間，月刊「言語」 (2007.7).
5. 高野明彦：人と「知の公共財」を「連想」で結ぶ，ず・ぼん１３ (2007.11).
6. 相澤彰子:「大量の情報から新しい価値を汲み出す ～情報の「検索」から「分析」へのパラダイムシフト～」，情報通信ジャーナル６月号（「情報学探訪」）(2007).

〔研究ノート〕

〔その他〕

＜会議発表等＞

〔招待講演〕

1. 高野明彦：Information Access by Association -- From Search to Imagine (招待講演), 日伊国際シンポジウム「創造と再生」, 2007.4.17
2. 高野明彦：検索から連想へ―連想による情報アクセス(基調講演), コンテンツワールド, 2007.9.7.
3. 高野明彦：検索から連想へ―知の創発を促す「想・IMAGINE」(基調講演), デジタルドキュメントシンポジウム 2007, 2007.11.22

〔一般講演〕

1. Akihiko Takano, “Information Access by Association”, 2nd Austria-Japan Summer Workshop on Term Rewriting, Obergurgl, Austria, 24 August, 2007.
2. Hai-Yen Siew, Kunio Shimizu, Asymmetric t-type distribution on circles, The Fifteenth International Conference of Forum for Interdisciplinary Mathematics (FIM), Shanghai, 21 May 2007
3. Hai-Yen Siew and Yasumasa Baba, A case study of the application of directional statistics on wind data, The 2007 IASC－ARS Special Conference, Seoul, Korea, The proceeding of the 2007 IASC－ARS Special Conference, pp.85-88, 7 June 2007
4. Hai-Yen Siew and Yasumasa Baba, Regression analysis on surface ozone by meteorological variables: a case study, Hokkaido, The proceeding of the 9th Japan-China Symposium on Statistics, pp. 277-282, 26 September 2007
5. Naofumi Sakaguchi and Yasumasa Baba, Estimation of number of common elements in several sets, 56th Session of the ISI INTERNATIONAL STATISTICAL INSTITUTE, 29 August 2007.
6. 馬場康維, 坂口尚文。マッチングによる共通メンバー数の推定, 日本計算機統計学会第２１回シンポジウム, 神奈川県鎌倉市, 2007 年 11 月 15 日

＜著書等＞

＜受　賞＞

1. 相澤彰子:「類語関係抽出タスクにおけるコーパス規模拡大の影響」,［2006-FI-84(2006.9.13)］（情報学基礎研究会), 平成 19 年度山下記念研究賞

2. 第6回東京インタラクティブ・アド・アワード(2008.3.18)

「想-IMAGINE」(受賞部門:サイト部門 プロダクトサイト・入賞)

「Powers of Information」(受賞部門:サイト部門 キャンペーンサイト・入賞)

（3）その他の成果発表

1. 「想－IMAGINE Book Search」 http://imagine.bookmap.info/
2. 千代田図書館「新書マップコーナー」公開
3. “本を置くだけで情報検索―千代田図書館で”, ITmedia News, 2007/4/25
4. “千代田図書館「親しみやすく」夜10時まで開館”, MXニュース, MXテレビ, 2007/4/26.
5. “図書館にコンシェルジュ―千代田区 新庁舎内に配置,連想検索システムも導入”, 日本経済新聞 朝刊 37 面, 2007/4/27
6. “新書マップが千代田図書館の顔に―キーボード不要! 新書をのせるだけで連想検索!”, ウェブマガジン風, 2007/4/30.
7. “新書の関連情報　検索システム開発―国立情報学研究所”, NHK 総合・全国ニュース, 5:00am, 2007/5/6
8. “新書の関連情報　検索システム開発―国立情報学研究所”, NHK 総合・おはよう日本, 7:00am, 2007/5/6.
9. “区立図書館 個性で勝負―千代田図書館の目玉「新書マップ」コーナー”, 朝日新聞 東京版, 2007/5/6.
10. “千代田図書館オープン―新書をかざすだけの検索システムを試験的に導入”, NHK 総合・おはよう日本, 7:00am, 2007/5/7.
11. “日本初サービスも, 図書館最前線―千代田図書館・新書マップコーナー”, NHK 総合・ニュースウオッチ９, 9:00pm, 2007/5/7.
12. “サービス&生活 ＩＣタグが変える！―「本の世界を広げる」千代田図書館・ＩＣタグを使った本の情報検索サービスを導入”, テレビ東京・ワールドビジネスサテライト, 11:00pm, 2007/5/7.
13. “公共図書館に新たなサービス”, NHK 総合・スタジオパーク, http://www.nhk.or.jp/kaisetsu-blog/200/2889.html, 2007/5/9.
14. “新書マップコーナー（東京・千代田図書館）―サイトと書棚；融合を体感”, 朝日新聞 夕刊 11 面, 2007/5/29.
15. “「新書マップ」導入（区立千代田図書館）―専用端末に内容表示；古書店の在庫や関

連テーマも”，毎日新聞 朝刊 27 面，2007/6/16.
16. “かがく Café：頭脳と電脳，連想の相互作用”， 日本経済新聞 朝刊 31 面，2007/7/1.
17. “人気図書館の秘密―千代田図書館・新書マップコーナー（千代田図書館よりライブ中継）”，NHK 総合・おはよう日本，7:33-37am，2007/7/2.
18. “「提案型」の「連想検索」 広がるネット検索―関連語ページも表示，利用者の関心を類推”，朝日新聞 朝刊 9 面，2007/8/20.
19. “公共図書館の新しい試み―千代田図書館・豊島区立図書館”，MX ニュース，MX テレビ，18:00，2007/8/21.
20. “千代田図書館訪問―本を置くだけで関連情報一覧”，週刊こどもニュース，NHK 総合，18:10，2007/8/25.
21. “検索エンジンは脳の夢を見る”，爆笑問題のニッポンの教養 FILE30，NHK 総合，2008/3/4.