2022年度　本研究2年目
成果報告会　資料－7

ROIS 戦略的研究プロジェクト 成果報告会
（第3期名称：機構間連携・文理融合プロジェクト）

研究課題名：
機関評価のための書誌ネットワーク推論の研究と人文社会学のための研究IRシステムの開発

2022年10月28日

研究代表者
所属　統計数理研究所
氏名　金藤 浩司

1. 背景と研究目的

◈背景と研究目的

本研究では以下の二点の研究及び、成果の実装（オープンなシステム化）に取り組む

1）『確率モデルを用いた大規模書誌ネットワーク構造の推論アルゴリズムの研究と実装』

研究業績の評価にも使われる書誌データは出版社など主に企業が整備した大規模なデータベースに依拠している。一方、オープンサイエンスの高まりでオープンアクセスのジャーナルや機関リポジトリ等を基盤とした学術論文データベースの整備も進んでいる。このようなオンライン上の分散された書誌情報を研究業績などの評価に用いる上で問題となるのは採録された情報やユーザーインタフェースが提供側の整備、公開状況によって常に変動することである。この、様々な提供元がそれぞれ独自に採取し公開したオープンな情報である学術雑誌、書籍（古典籍も含む）のインターネット上に分散している書誌データは比較的簡単なプログラムでも、断片であれば取得することができる。このようなソフトウェアプログラムのことをクローリングという。評価指標の計算に必要な書誌のネットワーク構造（引用-被引用、共著など）を採取した手元の比較的少数のデータで推測することができれば、それを用いた各種の評価指標の客観性、信頼性の向上につながる。1）では実際にインターネット上から自動的にデータをサンプリングし、全体の構造をできるだけ正確に表現するための確率モデルを用いた推測アルゴリズムを開発することを目標とする。

2）『人文社会学分野の研究機関における評価指標の研究』

学術雑誌の引用数など、ある程度確立した数値指標がある自然科学分野に比べ、人文学（humanities）の分野ではコミュニティ全体に認知された指標は未だ存在しない、またそもそも評価という概念の導入に関しては様々な観点から議論がなされているが、個々の研究者の業績評価とは別に、人文社会学における各専門分野の研究機関においてはその研究活動の実態を把握するための評価項目の重要性は高まっている。ここで、大きな問題となるのは、それぞれが独自の分野をカバーする、比較対象とする機関がほとんどのケースで存在しないことである。また機関間の比較につながるような指標は上記の観点から馴染まない。一方で研究機関が掲げるミッションという観点からみると、人文社会学系の研究機関には基幹とする業務と重要な項目として資料の整備と外部利用の推進がある。研究者の資料利用の変遷を数値化すれば研究所の施策との関係やその分野自体の動向把握に役に立つ。また、古典籍などの文献、博物館の収集物なども資料という観点からみれば基盤整備として各研究所の共通点となる。

今回はとくに古典籍データベースに注目し、その利用状況のデータを用いて時系列情報として可視化する手法の研究を行う。これまで本研究グループでは、トピックモデルという自然言語処理の手法で論文の要旨等の情報を用いて 組織やグループごとの研究動向を把握する分析手法の研究を行ってきておりこの応用の研究と機関のIR室をユーザに想定した分析システムとして実装を行う。

◈国内外の類似・競合する研究との関係

独自の指標を開発し、分析ツールとして公開している例としてはアメリカ国立衛生研究所（NIH）のiCite（*1）がある。オープンな論文検索サービスとしては独自の自然言語処理技術を活用したAI2のSemantic Scholar（*2 ）がある。どちらも書誌データは自組織内に確保したものを用いており、本研究のように完全に外部のオープンデータのみを用いた論文検索・指標サービスを志向したものではない。

*1 https://icite.od.nih.gov/　　*2 https://www.semanticscholar.org/

◈本研究の意義

オープンサイエンスの推進により研究情報のオープン化が国内外で急速に進んでいる（* 1）。研究データ・論文・研究者業績が公開され、オープンな流通体制が充実していくなかで、各サービスを横断的に取得し利活用するアプリケーションの一つとして本研究は独自の位置を占めることができる。特に断片情報から書誌全体の構造を推定する技術は得られる社会的波及成果も大変大きく、特定の有償書誌データベースに依存しがちな研究IRの現状には大きなインパクトとなる。

1「オープンサイエンスの推進について」文科省 科学技術・学術審議会総合政策特別委員会（第15回） 2016/11/24

2. 研究の概要

1) 研究の概要

本研究はROIS第3期概算要求「研究IRハブ」プロジェクトと関連している。プロジェクトのゴールとして新たに開発した指標（多様性指標）の機関への浸透を掲げ、システム開発を実施する計画になっているが、ライセンスフリーなデータでの指標の算出、人文社会学分野への対応がなければ目標の達成は非常に難しい。概算要求プロジェクトを土台に研究的タスクとして、

1）指標計算のオープンデータ化
2）人文社会学分野にも対応可能な評価指標の開発

に地球研、国文研とともに取り組むこととした。
3機関の文理融合研究とすることで、各機関のIR室のスタッフとも連携し大学共同利用機関における自己検証の実践も同時にできる体制とした。
「指標計算のオープンデータ化」は、たんに無料の分析環境を構築するため、というよりも、統計学における高次元データのモデリング手法の一種であるスパースモデリングを大規模書誌データの推測に応用することで、学術論文全体の構造をインターネット上の断片情報から捉えようとする手法の開発が目的である。
「人文社会学分野にも対応可能な評価指標」については、多様性指標のコンセプトとまったく異なる新たな指標を開発する形ではなく、多様性指標の個々のパーツをどのような概念の置き換えをすれば適用できるか、ということ念頭に検討する方法をとった。結果的に機関の強みであり、共同利用の資源である古典籍データベースのアクセス履歴を多様性指標における引用－被引用関係に置き換え、一連の操作履歴にはユーザー（研究者）の学術的興味が背後にあるとみなし、アクセス履歴から書誌を分類（クラスタリング）することで、分類ごとの検索成功の率を指標化することを検討した。これは多様性指標を構成するパーツをほぼ流用でき、博物館等の人文学研究機関がなんらかの形で保有し研究者に提供する「資料」のデータベースにそのまま横展開できる。また少数のアクセス履歴からデータベース全体の構造を推定するアルゴリズムは1）のスパースモデリングを同様に用いている。

2) 実施計画・実績

	2020年度	2021年度	2022年度
	FS (Feasibility Study) →	本研究 →	
	★6/19 FS採択審査会	★3/18 FS評価審査会(本研究採択)	★3/31 最終成果報告書 ★10/28 成果報告会
費用（千円） 予算	950	1,500	
費用（千円） 執行	950	1,500	

実施者（所属機関）

研究代表者：金藤 浩司（統計数理研究所 データ科学研究系）
共同研究者：谷口 真人（総合地球環境学研究所）、神作 研一（国文学研究資料館）、本多 啓介（統数研 国文研）、濵田 ひろか（統数研）

3) 研究成果の概要

以下の2つの課題に取り組んだ。
1）指標計算のオープンデータ化
論文を検索し、引用数等とともに多様性指標のスコアを表示するIR実務者向け分析システムを実装した。このシステムでは論文検索、スコアの計算には特定の商用ライセンスに依存しないようすべてオープンデータのサービスを利用している。今回データ取得元としてNIHの無料検索エンジン、PUBMEDを利用した。このほか開発中であるが、他のアーカイブサービス、arXivやCiNiiも順次取得可能になるよう開発中である。またログイン/ユーザー管理の機能連携としてORCID（*1）を利用している。OpenIDというユーザー認証プロトコルによって、すでにユーザーがORCIDのアカウントを持っていればそのままシステムにログインできる。またシステム側はORCIDと連携し、ユーザーの所属情報などを取得できる。これはユーザーの承認のもと行われる。また多様性指標のスコアの解釈には算出の基になったクラスタリング結果である。潜在的学術分野間同士の学術的距離が重要であるが、この視覚化として3次元空間上に分野を配置した3Dプロットを表示できるようにしている。将来的には論文間の異分野度が直感的に把握できるように、このプロット上に多様性指標のスコアを重ね合わせて表示できるように開発を進めている。
2）人文社会学分野にも対応可能な評価指標の開発
博物館、図書館を母体とする研究機関の保有する「資料」のデータベースのアクセスログからトピックモデルによる文書分類を行った。具体的には国文研の古典籍データベースのログを元に検索キーワードの集合を文書と捉え、そのキーワードの出現頻度によりジャンル（トピック）に分類するいくつかの特徴的なトピックに着目し、アクセス元の（研究グループ）の傾向を調べた。時間的な傾向（科研費、卒論）と地理的傾向（海外の研究機関等）の条件分けでさらに潜在的な利用傾向を把握することを示した。自機関の資料、データベースの整備状況とアクセスログの結果を照らし合わせることでミスマッチが起きていないか、など機関の意思決定に資することができる。