若手研究者海外派遣プログラム
派遣終了報告書

1　派遣者	
所　属　機　関	総合地球環境学研究所
氏　　名	竹村　紫苑

2　派遣計画 概要	
派　遣　国	アメリカ合衆国
派　遣　期　間	平成 28 年 4 月 3 日　～　平成 28 年 7 月 2 日
派遣先機関名	バージン諸島大学
（英語）	University of the Virgin Islands
受入教員名	アレクサンドゥリディス　コスタス
（英語）	Kostas Alexandridis
研究課題名	自然言語処理を用いたテキストデータの知識構造抽出および、共通の知識構造を持つテキスト群のパターン抽出
（英語）	Characterization and pattern extraction of knowledge structures using Natural Language Processing (NLP) of diverse text data

3　派遣による研究実績

（１）調査研究実績　　（研究計画に沿い、実施したことを記載してください。）

1）高精度分析手法の開発

下記の手順により知識構造を可視化する分析手法を開発した。

①TF*IDF（テキストデータ中の単語に関する重要度の指標）に基づき、重要な概念の抽出

②抽出された異なる２つの概念が同じ段落において共起（出現）する頻度に基づいて類似度を算出し、セマンティック・ネットワーク行列の抽出

③ネットワーク分析による概念間のネットワーク構造の抽出と可視化

本手法の開発には、アメリカ合衆国（2 箇所）と日本（1 箇所）で開催したステークホルダー・ワークショップにおいて参加者の発話記録を録音し、その録音データからテープ起こしをおこなったテキストデータ*を使用した。

* i)アメリカ領バージン諸島セント・トーマス島（2012 年 9 月～2013 年 1 月、参加者 31 名、1001 段落）、ii)京都（2015 年 1 月、参加者 45 名、4511 段落）、iii)アメリカ合衆国フロリダ州サラソタ（2015 年 10 月、参加者 26 名、1617 段落）

2）高精度分析手法の有効性および汎用性の評価

アメリカ領バージン諸島（VI）およびアメリカ合衆国フロリダ州（FL）の分析結果を比較した。その結果、テキストデータから抽出した VI と FL のセマンティック・ネットワークは、入次数（ある特定の概念と他の概念との類似度の総和）および TF*IDF の分布が、べき乗則に従い、スケールフリーのネットワークであった。この結果は、ある特定の概念の TF*IDF は、その概念が他の概念との類似度が高くなるほど大きくなることを示し、TF*IDF の大きい概念はネットワークの構造上においても重要であることが示された。この性質は VI と FL の両事例において確認されたことから、地域の様々なステークホルダーの協働により環境問題を解決するために形成されたセマンティック・ネットワークが有する普遍的な特徴であることが示唆される。

また、本手法によって得られたセマンティック・ネットワークは、類似度の低いリンクを除去してもスケールフリー性が失われない性質を有することも明らかとなった。この結果は、類似度の低いリンクは、そのリンクを除去しても異なる２つの概念間を結ぶ経路の最短距離に及ぼす影響が少ないことを示している。つまり、類似度の低いリンクを除去することにより、セマンティック・ネットワークの骨格部分をより鮮明に可視化できる。

また、類似度の低いリンクを除去したセマンティック・ネットワークの構造を可視化することによって、特定の段落において共起する概念から成る知識が抽出できた。バージン諸島とフロリダのデータは、互いに独立したものであるにも関わらず、セマンティック・ネットワークは地域経済・教育・生業（VI：農業、FL：漁業）に関する知識と、それらの知識を繋ぐ核となる知識（VI：環境の持続可能性、FL：環境の管理と再生）が共通していた。一方、バージン諸島におけるワークショップ参加者は、「政策・行動」や「自然環境全般への計画・管理」に関する知識に着目していたのに対して、フロリダにおけるワークショップ参加者は、「海域の生態系管理」に関する知識に着目していた。これらの結果から、バージン諸島とフロリダの両事例において、地域の環境問題の解決に向けて様々なステークホルダーの協働により形成されてきた知識は、知識を構成する概念と概念間のネットワーク構造として捉えられることが示された。さらに、異なる２つの事例間において重要概念のネットワーク構造を可視化し、それらを比較することによって、より抽象的な視点（鳥の目）から事例間の知識の類似性と異質性も比較可

能であることも、バージン諸島とフロリダの分析結果から示された。

最後に、京都で開催したステークホルダー・ワークショップのテープ起こし原稿（日本語）についても同様の方法で分析を行い、バージン諸島とフロリダの場合と同様の結果が日本語でも得られることを確認し、本手法の日本語に対する適用性が検証された。

なお、本手法では、高精度の分析結果を得るため上でノイズとなる言語学的に意味がない単語や、類義語の除去は手作業で行っている（手順①）。そのため、ひとつのテキストデータの分析を行うためには、数日程度の人手と時間が必要となることが課題である。

3）大規模データ分析手法の開発

1）で開発した分析手法では、言語的に意味がない単語および類義語の除去作業（手順①）に人手と時間が必要となるため、大規模データを用いた分析に向けた改良が必要であった。そこで、大規模コーパスからノイズとなる単語の除外リストを構築することによって、手順①を簡素化した大規模データ用の分析手法を開発した。なお、日本語テキストデータに対する除外リストの作成には、国語研究所が公開している『現代日本語書き言葉均衡コーパス』および『日本語話し言葉コーパス』を使用した。英語テキストデータに対する除外リストの作成には、『COCA: Corpus of Contemporary American English』を使用し、IBM SPSS Premium ver.16 に搭載されている英語辞書によってノイズとなる単語が除去されている事を確認した。

4）大規模データ分析手法の有効性および汎用性の評価

3）で開発した大規模データ分析手法を、a）総合地球環境学研究所（以下、地球研）の地域環境知プロジェクト（E-05）のメンバーが石垣島白保集落における取り組みについて執筆した出版物（論文・書籍・報告書・雑誌・新聞等、2004 年～2015 年、221 テキスト、4174 段落）、b）地域環境知プロジェクトが地球研プロジェクトのリーダー（終了プロジェクトを含む）に対して実施したインタビュー調査のテープ起こし原稿（2011 年・2013 年、のべ 15 人、7173 段落）、c）地域環境知プロジェクトのメンバーがカナダにおける生物圏保護区制度を活用した先進的な地域づくりの取り組みについて紹介した書籍（2013 年、68 ページ、604 段落）のテキストデータに適用し、開発した大規模データ分析手法の有効性および汎用性を評価した。

その結果、大規模データ分析手法はテキストデータのサイズが十分に大きく、かつ、テーマが絞られていれば、高精度分析用手法と比べてノイズの除去率は下がるが、知識構造の抽出・可視化に十分耐えうることが明らかとなった。また、テキストデータの種類（日本語／英語、書き言葉／話し言葉）に関係なく知識構造を可視化できたことから、大規模データ分析手法の汎用性も示された。その一方で、テキストデータのサイズが小さい、または、テーマが定まっていないテキストデータは、大規模データ分析手法には適していないことも明らかとなった。

5）英語論文のとりまとめ

開発した高精度分析手法についてその方法論の新規性、有効性、汎用性をアメリカ領バージン諸島セント・トーマス島およびアメリカ合衆国フロリダ州サラソタにおいて開催したステークホルダー・ワークショップのテキストデータを用いて検証し、論文として取りまとめた。現在、投稿準備中である。

（2）基幹研究プロジェクトにおいてこの派遣が果たした役割

地球研は、地域の環境問題を地域住民との協働により解決する研究に世界各地で取り組んでおり、そのような研究アプローチによって実施された研究事例や知見が大量に蓄積されてきた。これらの研究事例や知見から、地球環境問題のホットスポットであるアジアにおいて、地域の様々なステークホルダーの協働による、地球環境問題の解決のための実現可能なオプションを提案するために、これまでは主に研究事例や知見に関する言説を深く・丁寧に分析する質的調査による研究が行われてきた。しかしながら、質的調査の実施には膨大な時間と人手が必要であり、機械処理によって大量の言説を迅速に処理し、定量的に分析・整理することによって、言説構造の全体の傾向や特性を把握する研究アプローチが必要とされている。

本派遣期間に開発した高精度分析手法によって、ステークホルダー・ワークショップにおいて録音したテープ起こし原稿から、ワークショップ参加者が熟議を通して共有した知識を、知識を構成する重要な概念とそれらのネットワーク構造として可視化できた。また、共通の調査票を用いて地球研プロジェクトリーダーに対して実施したインタビュー調査のテープ起こし原稿や、特定の事例に関するテキストデータに対して大規模データ分析手法を適用することにより、地球研プロジェクト間で共通する知識や、特定の事例を特徴付ける知識を可視化することができた。以上によって、地球研に蓄積されてきた膨大なテキストデータの横断的な分析手法をほぼ確立することができ、基幹研究プロジェクトにおける多様な地域からの知見を地球環境問題の解決に向けて統合するという課題に大きく貢献できる分析手法が開発できたと考えている。

（3）所属機関における学術分野に貢献する事項

本派遣中に開発した手法は、テキストデータの中から重要な100概念を抽出し、その抽出された異なる2つの概念が同じ段落において共起するか否かという情報のみを用いて、概念間のネットワーク構造を抽出するという手続きを基本としている。このようなシンプルなアルゴリズムによって、テキストデータとして表現されている知識を、知識を構成する概念とそれらのネットワーク構造として捉えられる点が、本手法の学術的な独創性である。この手法により、大量のテキストデータに対して高速の機械処理による分析が可能になり、地球環境問題に関連する社会現象の基盤となる知識構造を理解するための分析に必要な時間と労力の大幅な低減が可能になった。この大量データの高速処理技術によって、地球環境問題の根幹をなす人と自然の複雑な相互作用環の分析に新たな可能性を開拓し、総合地球環境学に革新をもたらすことができるものと考えられる。さらに重要なことは、テキストデータの品質が分析結果に大きく影響を及ぼすという、機械処理による分析の限界が明らかになったことである。機械処理による分析の限界を十分に理解した上で、様々なデータに対する本手法の応用を工夫することによって、多様で複雑な社会現象を、定量的かつ効率的に把握するための新しい分析手法として成熟させ、総合地球環境学の発展にさらに貢献していくことが課題である。

（4）研究成果（著書、論文及び報告書名・講演題目）

（5）見込まれる研究成果（著書、論文及び報告書名・講演題目）

Alexandridis, K., Takemura, S., Webb, A., Lausche, B., Culter, J., Sato, T., Semantic Knowledge Network Inference Across a Range of Stakeholders and Communities of Practice. Plos One（投稿準備中）

（注意事項）

・本報告書は、帰国後1ヵ月以内に提出して下さい。

・この報告書を、本機構により刊行、Web掲載、広報冊子等として公表することがあります。
この場合、内容に影響しない範囲で修正を行うことがあります。