地理情報科学における論文情報および用語集を用いた主題の特徴化

小野雅史・柴崎亮介

Topic Characterization of Geographic Information Science Researches by using Research Articles and Glossary Information

Masafumi ONO and Ryosuke SHIBASAKI

Abstract: Geographic information science is an interdisciplinary research field applicable for natural science such as meteorology, hydrography or ecology and also applicable for social science such as economics, urban planning or emergency management. “GIS Association of Japan (GISA)” was founded in 1991, then various researches in geographic information science were activated through the academic journal “Theory and applications of GIS” published by GISA in 1993. Also, after education programs were enhanced since the latter of 1990s, the glossary of geographic information science was edited by GISA subcommittee. The journal and glossary are valuable historical heritages. So, we try to extract the domain specific topics and terms in geographic information science from these resources by using natural language and text mining technics.

Keywords: 論文分析(article analysis), 分野特有語(domain specific term), 用語集(glossary), 自然言語処理(natural language processing), テキストマイニング(text mining),

1. はじめに

統計によれば、日本の大学等の研究関係従業者数は毎年増加しているが（総務省 2012)、科学論文の発表論文数は主要国で唯一日本のみ伸び悩みをみせていることが報告されている（文部科学省 2013)。こうした事態もあり、日本の研究活動全般における生産性の低下が一部で危惧されている（Fuyuno 2012)。

一方、かつて主流だった紙媒体の論文誌の刊行は、2000 年頃からの多くのジャーナルによる電子投稿審査システムの採用によって変化し、今では電子ジャーナルによる刊行が大多数を占めている（林 2007)。さらに近年では、各ジャーナル運営者自体がシステム開発を行う必要もなく、CiNii 等の請負サービスも軌道に乗ってきている（大向 2012)。こうした、電子ジャーナル化の進行によって先行研究へのアクセシビリティは確実に向上しているはずである。

それにも拘わらず、こうした過去の研究へのアクセシビリティの向上が全体として新規の発表論文数の増加に結び付いていない。この原因には様々なものが挙げられるが、論文のデジタルライブラリ化は進んだものの、まだ多くの分野でそのライブラリを十分に活用しきれていないことにも一因があると考えられる。

日本の地理情報科学の歴史において、1991 年に地理情報システム学会が発足し、1993 年から論文誌「GIS-理論と応用」vol.1 が発刊され、GIS に関する研究活動が活発となった。1995 年以後、「GIS-理論と応用」は年 2 回の刊行ペースを維持しており、2013 年 8 月現在では vol.21 に及ぶ。これらの

小野雅史 〒153-8505 東京都目黒区駒場 4-6-1
東京大学　生産技術研究所 Cw503　柴崎研究室
Phone: 03-5452-6417
E-mail: maono@iis.u-tokyo.ac.jp

論文誌はデジタルライブラリ化され、そのうち近過去2年分は学会員専用ページを通して学会員に公開され、それ以前の版は J-Stage を通して一般向けにも公開されている。

この「GIS-理論と応用」を通して発表された研究の中で、過去に教育カリキュラムの作成を目的とした GIS の教科書・教材の分析は数多く行われている。しかし、「GIS-理論と応用」の論文全体を対象とした分析はまだ報告されていない。

地理情報科学は学際的な分野であり、気象・農業・生態系・環境などの自然科学から経済・都市計画・防災などの社会科学まで、応用分野は多岐にわたる。その展開には年代毎のトレンドや、新たに提案されたトピックもあるだろう。そうした知見を論文から得られる統計情報をもとに、明らかにすることには意義があると考えらえる。

著者らは、過去に、都市計画分野の論文データを対象にして、トピックモデル(Blei 2012)を用いて得られる統計・学習情報をもとに、各論文に含まれる明示・暗示トピックの再解析や、論文毎の類似度を計算する手法を提案した（Ono 2013)。こうした経験を下敷きに、本研究では、データマイニングや自然言語処理の手法を活用して、地理情報科学における特徴的なトピックの抽出と整理を試みる。

2. 研究の方法

2.1 概要

「GIS-理論と応用」の 1993 年から 2013 年に発刊された「GIS-理論と応用」vol.1 から vol.21 に収録された原著論文、展望論文、研究・技術ノート、データ論文を対象として分析を行う。

なお、ここで、既存研究である教科書・教材リソースを対象としたカリキュラム作成のためのアプローチ（岡部 2007）と、論文リソースを対象とした本研究のアプローチとの基本的な違いを整理しておく。まず、前者は国内外のシラバスや既存の複数の教材から共通の要素を抽出して整理することに主眼がある。これは限られた時間で最大限の教育効果を上げるために、特に重視すべき教育項目を選別するのが目的だからである。それに対して本研究では、共通する要素よりもむしろ例外的だが地理情報科学分野にとって重要かつ特徴的なトピックを抽出することを目的とする。なぜなら、学術論文とは原則として新規性・独創性を競って出版されるものだからであり、共通の要素よりむしろ特殊な要素に焦点を当てることに意義があると我々は考えるからである。

2.2 手法と手順

本研究の手法と手順について以下で述べる。抽出や集計の方法はプログラムによる自動処理を基本とするが、例外やエラーの多発によりプログラムで処理するのが難しいケースについては、手動による処理を組み合わせて実施する。

1) 論文ファイルの収集
   地理情報システム学会デジタルライブラリを通して、PDF フォーマットで提供されている論文ファイルを収集する。
2) テキストデータの抽出
   収集した PDF ファイルからテキストデータを抽出する。各論文の中には日本語、英語、数値、記号が混在しており、その種類によって異なる処理が必要になるため、同時にそれらを分類する。
3) 日本語形態素、英単語毎の解析
   上記で分類した文字データをもとに、日本語は形態素毎に、英語は単語毎に分割し、それらの種類・頻度を算出する。
4) キーワードの情報の整理
   著者により付与されたキーワードの情報を整理し、それらの重複を調べる。
5) 一般的な言語空間との比較
   オンライン百科事典 Wikipedia の全テキストデータの解析により得られる語彙情報との比較を行い、「GIS-理論と応用」のみに出現する

分野特有語を抽出する。

6) 地理情報科学用語集 (GISA,2000)との比較

地理情報科学用語集で定義された語句の論文データ内での出現頻度を調べる。また、出現頻度を年代別に可視化することによって推移を確認できるようにする。

3. 結果

2.2 で示した手法に対する結果を以下に順次示す。

1) 収集した論文数

収集した論文ファイルの数を表 1 に示す。大部分を日本語論文が占めており、英語論文は全体の約 6%程度であることがわかる。

表 1:「GIS-理論と応用」論文数

	日本語	英語	総計
本数	331	22	353

2) 論文の文字構成

331 の日本語論文を対象として、一論文あたりの文字数を日本語・英語・数字・記号別に解析した結果を表 2 に示す。この結果から平均的な日本語論文には、英語文字が 21.8%、数字が 3.6%含まれていることがわかる。

表 2: 日本語記述による一論文あたりの文字数と構成

	日本語	英語	数字	記号など
最小値	2582	564	64	0
最大値	15389	7317	1370	1755
中央値	5068	1100	108	122
平均（文字数）	6021	1773	289	30
（割合）	74.2%	21.8%	3.6%	0.4%
標準偏差	1830.7	1070.4	193.3	179.8

3) 日本語形態素および英単語の種類と頻度

353 の全論文テキストを対象として、日本語形態素および英単語別に分割し、その種類と、一種あたりの出現頻度の最大値および総数（種類×頻度の合計）を算出した結果を表 3 に示す。結果から、日本語形態素と英単語とで総数は一桁違うが、種類として大きな開きは無いことがわかる。

表 3: 日本語形態素および英単語の種類と頻度

	日本語形態素	英単語
種類	20,028	16,780
出現頻度の最大値	53,714	9,097
総数	1,006,088	197,260

4) キーワード情報

著者により論文に付与されたキーワードの情報を収集し、その種類と重複を調べた。重複数の上位 9 ランクまでを表 4 に、全体の分布を図 1 に示す。これらの結果から、全キーワード 2181（日本語:1030, 英語:1151）のうち 1919(日本語:1022, 英語:897)すなわち約 88%が重複のないユニークなキーワードであることがわかる。

表 4: 重複キーワードの上位ランキング

ランク	キーワード	重複数
1	gis	38
2	地理情報システム	24
3	数値標高モデル	13
4	dem, geographic information system, gps	8
7	地方自治体	7
8	リモートセンシング	6
9	local government, remote sensing, spatial autocorrelation, urban planning, 都市計画, 最尤法, 土地利用	5

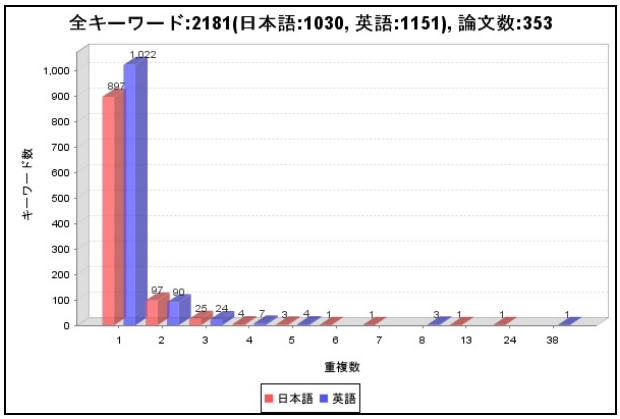

図 1: キーワード情報の重複と分布

5) 分野特有語の抽出

Wikipedia には存在せず、「GIS-理論と応用」の論文データにのみ出現した語句を、地理情報科学

分野における分野特有語として抽出した結果、797 の語句が得られた。そのうち、特に出現頻度が多かった上位 9 ランク(全 10 語)を以下に示す。

表 5: 分野特有語のリスト

ランク	キーワード	頻度
1	ネットワークボロノイ	103
2	スリバーポリゴン	92
3	バリオグラム	87
4	クリギング	79
5	ルルベ	44
6	サークルエリアカルトグラム	36
7	セミバリオグラム	30
8	デジタルオルソフォト	29
9	オブザベーション、コロプレス	28

6) 地理情報科学用語集との照合

地理情報科学用語集に登録された用語1613の論文内での出現状況を調べたところ、部分一致も含めると988（61.2%）の用語が実際に出現することがわかった。なお、最頻出語は「データ」で合計9202回出現していた。

さらに、全論文数あたりでその語句が出現した論文数と、全用語数あたりの出現頻度を年代別に集計し、前者の指標を Generality、後者を Popularity と仮定して、計算した結果の一例を図 2 に示す(なお、図中の値は正規化された数値を採用している)。この図から、「クリギング」に関して言及した論文は、1999 年から現れ始め、2011 年に最も大きなトレンドがあったことがわかる。

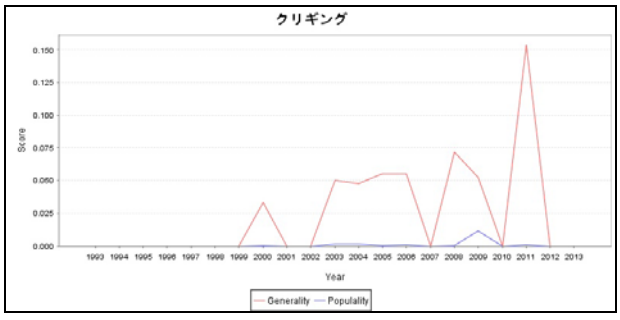

図 2:「クリギング」の年代別の出現状況

さらに、5)で得られた 797 の分野特有語と用語集内の用語を照合したところ、結果は「バリオグラム」、「クリギング」などわずか 7 件であった。これは、別の見方をすれば、5)の分野特有語リストのほとんどは既存の地理情報科学用語集にまだ未収録であるため、用語集の再編集を検討する際、重要な情報になる可能性が期待される。

5. おわりに

「GIS-理論と応用」の全論文から得られる基本的な情報の整理を行った。今回の研究を通して、当初想定していなかった様々な困難があり、残念ながら予定していたモデル計算を行うところまで至らなかったため、より詳細な統計解析やその結果を用いた各論文の再解釈等は今後の課題としたい。

参考文献

総務省統計局（2012）：「統計でみる日本の科学技術研究 平成 24 年科学技術研究調査の結果から」< http://www.stat.go.jp/data/kagaku/ pamphlet/>

文部科学省（2013）：「科学研究のベンチマーキング２０１２-論文分析でみる世界の研究活動の変化と日本の状況-」 <http://www. nistep.go.jp /archives/8865>

Fuyuno, Ichiko (2012)：「Numbers of young scientists declining in Japan」, *Nature*, <http://www. nature.com/news/numbers-of-young-scientists-declining-in-japan-1.10254>

林和弘(2007)：「理工医学系電子ジャーナルの動向―研究情報収集環境と事業の変革―」科学技術動向. 2007, 071, p. 17-29.

大向一輝(2012)：「電子書籍化する学術論文 ： CiNii Articles の展開を中心に」情報処理 53(12) 1282-1286

Blei, David M. (2012). :「Introduction to Probabilistic Topic Models」, Communications of the ACM 55 (4), 77–84.

Ono, M and Shibasaki, R. (2013):「Analysis of Author-Selected Keywords in Urban Planning and Urban Management Papers」, CUPUM 2013.

岡部篤行(2007)「地理情報科学標準カリキュラム・コンテンツの持続協働型ウェブライブラリーの開発研究」,地理情報システム学会 GIS 教育カリキュラム検討ワーキンググループ

地理情報システム学会(GISA) 用語・教育分科会編（2000）：「地理情報科学用語集（第 2 版）」