既存データを活用したインパクト―レスポンスに関する研究

研究予算：運営費交付金（一般勘定）
研究期間：平 27～平 30
担当チーム：水環境研究グループ（河川生態）
研究担当者：萱場祐一、傳田正利

【要旨】
本研究では、情報技術の発展に伴い普及するテキストマイニングの手法を用いて、河川整備計画を分析し、河川管理者の「環境」への意識を分析した。その結果、河川管理者は、「環境」に関して強い関心を示している反面、通常の河川管理とは離れた捉え方をする傾向を推定された。
キーワード：既存データ、インパクトーレスポンス、河川整備計画、テキストマイニング

1. はじめに

現状の河川生態系の修復再生と維持管理を行う場合、過去から減少した河川景観、群集（群落）及び個体群を保全・修復する取り組みが多いが、本質的には人為的インパクトを取り除くことが望ましい。

効果的な河川生態系修復を行うには、河川事業によるインパクトと河川生態系のレスポンス（インパクトーレスポンス、以下、「IR」と記述する。）の関係性を分析することが必要となる。IRを行うためには、過去から現在までの信頼性の高い河川管理に関する資料の分析が必要不可欠である。

河川管理実務では、河川法に基づき「河川整備基本方針」の策定後、「河川整備計画」が策定される[1)2)]。河川整備基本方針には、社会資本整備審議会・都道府県河川審議会による長期的視野での河川整備方針が記され、河川整備計画の策定時には公害防止計画等の他の法令に基づく計画を考慮する等[1)2)]、流域スケールでの河川環境管理に関する重要な項目が網羅されている。この特性に加え、上述の方針・計画の検討・策定時において、河川管理者が正確な河川管理情報を収集し、整備計画の立案に関しての議論を行う「河道技術会議」が開催され、上位計画者から実務者まで幅広い議論とその結果が集積される。

これらの資料を分析し、全国の河川生態系管理に関する問題と修復・再生の事例分析、より詳細な表現を行うには、河川生態系管理に関する技術的課題の類型化とその類型の中の典型事例の抽出、典型的なIRの因果関係分析が可能となれば、より的確なIRの評価が可能となる。

このような背景から、本課題は、全国の河川事業に関する資料を収集し、全国の河川におけるIR分析を通して、河川生態系に劣化を生じさせる河川事業の特徴の把握・類型化を行う。その後、IR分析手法を手続き化し、既存データを活用した河川におけるIRの分析手法の提案を目的とする。

H27年度は、河川整備計画に着目し、河川整備計画の文言から読み取れる河川生態系管理の位置づけ、河川管理者の関心時を整理し、典型事例の抽出に向けた準備を行った。

2. 研究の方法

(1) テキストマイニングの概要

工学が対象とするデータは定型化がされ、定型化されたデータを対象とする定量解析が発展している。一方定型化されていないが重要なデータの一つに「文章」がある。

金の定義によれば，「『文章』とは、何らかの文字が一定の文法規則に基づいている文の集合体」[4)]を指す。同じく金によれば，「インターネットを中心にやりとりされるメール、ブログ文なども上記の『文章』に該当することに懐疑的なことも多くあるため、言語学においては、『文章』と区別するため、記号列が何らかの規則に従って並べられた集合体を『テキスト』と呼び、『文章』と区別することが多い」とされる[4)]。

情報システムの普及により、テキストが急速に普及すると、これらのデータの活用が求められるにいたった[4)]。人によるテキストの分析は多くの労力がかかることに加え、認識や解釈が異なり、定量的な解釈手法が求められ、「テキストマイニング」は大きな可能性を持つ手法である。金の定義によれば，「『テキストマイニング』」とは、蓄積された膨大なテキストデータを何らの単位（文字、単語、フレーズ）に分解し、これらの関係を定量的に分析することをいう。近年、言語文体学の分野等において、急速に発展・普及する手法である[4)]。

⑵ 対象データとテキストデータの分析方法

テキストマイニングの手法は、土木工学・河川生態系管理ではなじみの少ない手法のため、テキストマイニングの手法の詳細を整理しながら、以下に分析手法の概要を示す（**図-1**）。また、**図-1** は、一般的な流れを示すために取りまとめているため、平成 27 年度は実施しなかった構文解析に関してのフロー図、「研究の方法」に入れ、取りまとめた。

本研究では、全国 109 水系の河川事務所が PDF で公表する最新の河川整備計画を HP からダウンロードしテキスト化し、日本の流域面積が上位の 14 河川に関してのテキストマイニングを行った（**表-1**）。

a) 電子化

PDF で公開される河川整備計画の大半は、テキストデータ化がされているが、一部の PDF は、画像として公開されているものがあった。そのため、画像として公開されている PDF を文字認識ソフト（NTT Data 社、e.typist Ver.15）を用いて、テキスト化した。

b) クリーニング

上記において、テキスト化したデータには、必要としない記号（ルビ等）やソフトウェアの誤認識に伴う誤った文字列などが含まれている場合がある。これらの不要な情報を目視で判読し、取り除いた。

c) 形態素解析

クリーニング後、記号・文字単位でデータを集計することが可能であるが、この解析ではテキストの意味の解釈は出来ない。単語を単位とする分析により、より文章・文書の意味を解析することが出来るため、文を単語単位に分析する「形態素解析」が必要となる。

金の定義によれば，「『形態素解析』は、文を単語単位に分け、品詞の情報を加える等の作業を行うことをいい」[4]、専用のソフトウェアが、言語文体学の研究機関から公開されている。本研究では、京都大学大学院　情報学研究科知能情報学専攻 黒橋・河原研究室が公開する日本語形態素解析システム JUMAN Ver7.0 を用いて、分析を行った[3]。

表-1　河川整備計画の分析対象河川の一覧

順位	河川名	地方整備局	流域面積（km2）	河川整備計画分析状況
1	利根川	関東地方整備局	16,840	○
2	石狩川	北海道開発局	14,300	○
3	信濃川	北陸地方整備局	12,050	○
4	北上川	東北地方整備局	10,250	○
5	木曽川	中部地方整備局	9,100	○
6	十勝川	北海道開発局	8,400	○
7	淀川	近畿地方整備局	8,240	○
8	最上川	東北地方整備局	7,040	○
9	天竜川	中部地方整備局	5,090	○
10	雄物川	東北地方整備局	4,640	○
11	米代川	東北地方整備局	4,100	○
12	富士川	中部地方整備局	3,990	○
13	吉野川	四国地方整備局	3,650	○

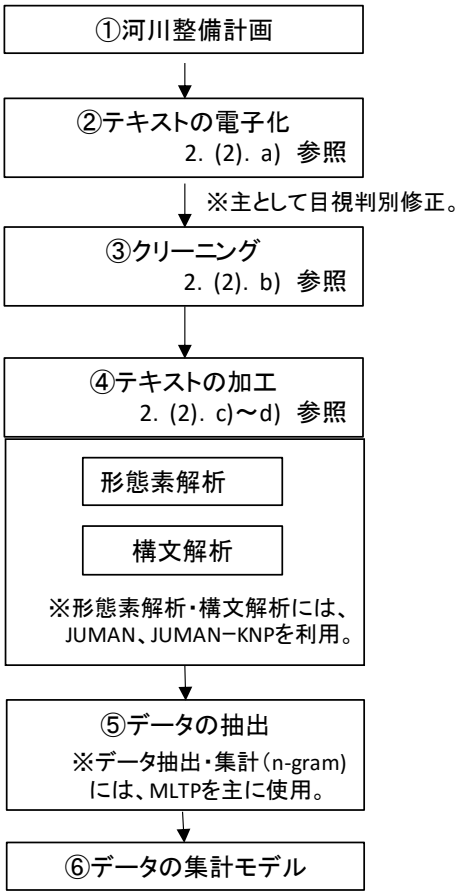

図-1　統計的テキスト解析の過程（金の図を参考に作成）[4]

d) 構文解析

金の定義によれば，「『構文解析（syntactic analysis）』とは文法規則に基づいて、文の構造を句・文節を単位として解析することをいう」[4]。構文解析は、河川整備計画等のより細かな分析の際に用いることを検討しているが、H27 年度の段階では、俯瞰的に河川整備計画を分析することを目的とするため解析は実施しなかった。

e) テキストにおける集計モデル

H27 年度は、テキストマイニングの基礎技術の導入として、記号列がテキストの中に現れる度数を集計した。次に、「その頻度が大きい順に並べるとその順位と頻度の関係には、式(1)のジップ（Zipf）の法則があることが知られている」[4]。

順位　×　頻度　≒　定数　　（式）1

この式(1)で求められる曲線形状の変化は、対象とする記号列の重要性を示すと考え、曲線が変化する順位を重要な記号列の下限として選定した。

次に、頻度分析の拡張的解析手法として、n-gram(エヌグラム)解析を行った。エヌグラム解析とは、n 個の記号列が隣接して出現する度数を集計し、記号列間の関係性を分析することをいう [4]。これらの分析により、記号列間の関係、例えば、複数の記号列が同時にテキスト内に生起する確率、言い換えれば、記号列と記号列の関係性を分析でき、同一の議論に用いられる傾向の分析が可能となる。

隣接する記号列の個数により、Unigram (ユニグラム：単一語)、bigram(バイグラム：2 つの記号列の連結)と解析名が変わる。本研究においては、Bigram 解析を実施するとした。また、エヌグラム解析は、ネットワークをモデル化するグラフ理論との親和性も高いため、記号列間のネットワークを統計ソフト R の i-graph のライブラリを用いて分析した。

本研究では、(イ) 頻度分析、(ロ) Zipf の法則に基づく記号列の有効性の分析、(ハ) エヌグラム解析を用いた記号列間の関係性分析から、治水・利水・環境に類別される各記号列間の関係性を分析した。

3. 結果と考察

(1) 記号列の頻度分析を通した重要記号列の抽出とその特性の把握

図-1 に記号列の順位と頻度 (対数表示) の関係性を

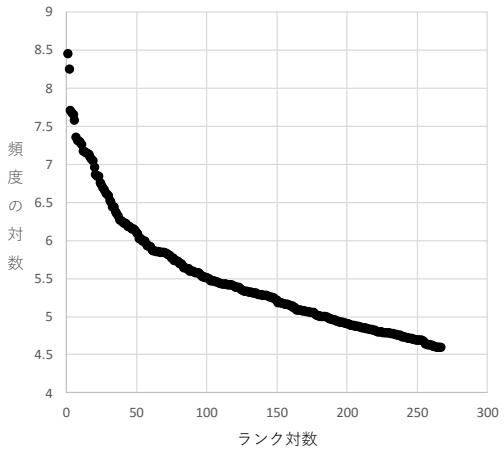

図-1　河川整備計画の分析対象河川

示す。曲線は、順位 40 位付近で傾向が変化した。40 位までは他の記号列に比べて用いられる回数が多かった。

図-2 に河川整備計画に用いられる記号例の頻度分布を示す。「環境」の記号列は、7 番目、環境に関連する記号列、「保全」、「水質」「生息」が 40 位内に入る等、河川環境 (河川生態系管理) が河川整備計画の中で、重要事項であることが明示的に確認された。

(2) エヌグラム解析とグラフ理論を用いた記号列間の関係性分析

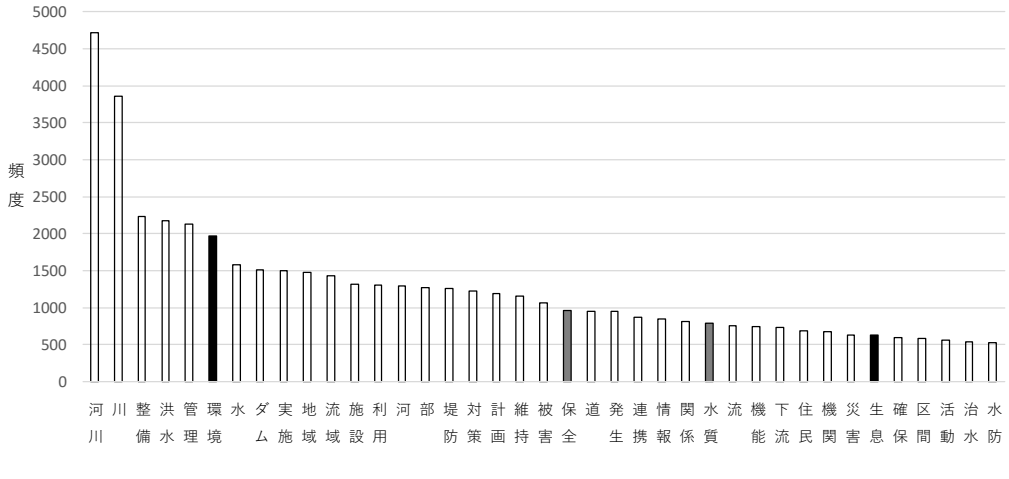

図-2　河川整備計画の分析対象河川

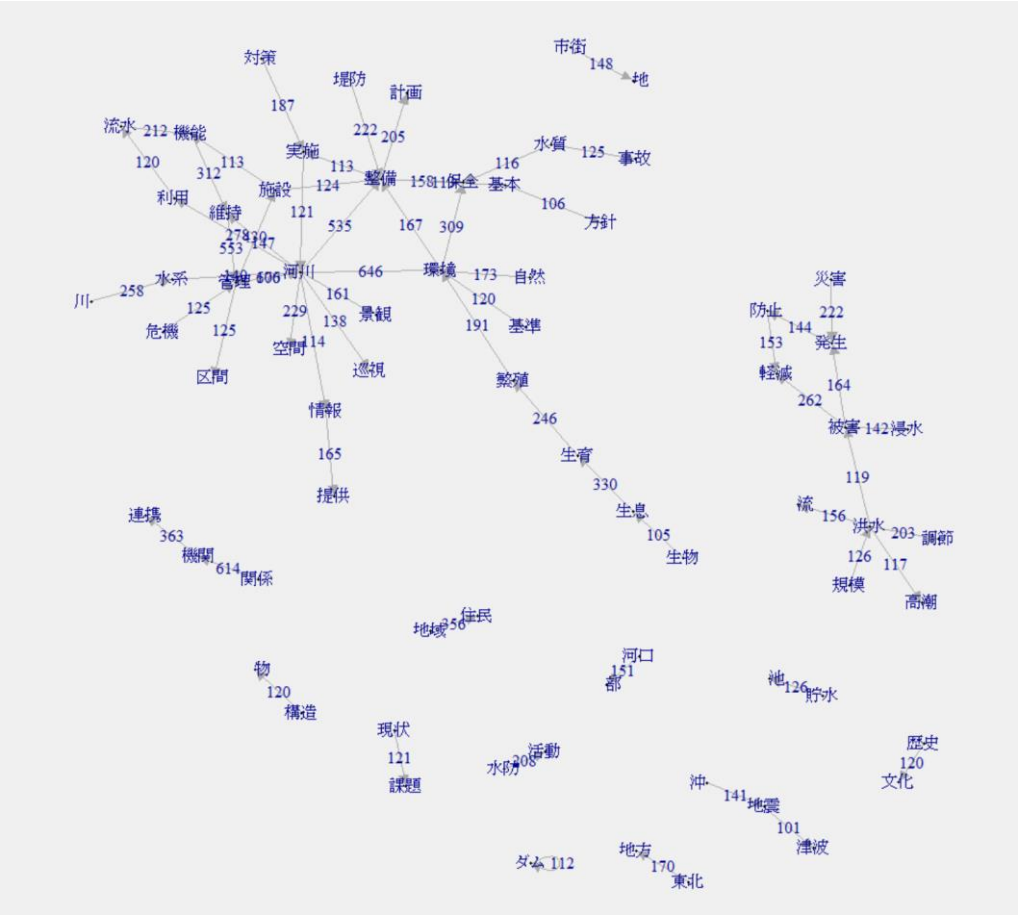

図-3　河川整備計画の分析対象河川

図-3 に Bigram を記号列のネットワークモデルで表示した結果を示す。「河川」を中心として、「環境」、「管理」、「整備」、「空間」などの記号列がネットワークを形成した。また、「環境」は、「保全」、「繁殖」、「生育」、「生息」、「生物」の記号列と明瞭かつ独立したネットワークを形成した。また、「洪水」、「歴史」、「地震」、「機関」を中心とした記号列のネットワークは、河川から独立したネットワークとしなった。

(3) 今後の研究への展望

テキストマイニングにより、河川整備計画における環境の重要性を、河川環境に関連する記号列の頻度、バイグラムのネットワーク分析を通して確認することが出来た。特に、「整備」、「維持」、「管理」を上回る頻度で「環境」の記号列が用いられたことは、河川管理者は、治水・利水等のハード型の河川管理と同程度の重要性を「環境」に感じていることを示す。この反面、「整備」、「維持」、「管理」が複雑なネットワークを形成しているのに対し、「環境」、特に、生息場保全に関連する記号列が単体のネットワークを形成した点が目立った。これは、環境保全に関する意識だけが、独自の発展をしていることに関連すると考えられ、「整備」の項目と連携した「環境」への取り組みの必要性が示唆された。

平成 27 年度の試行的な取り組みにより、テキストマイニングを河川整備計画に適用することにより、河川管理者の河川管理業務時の意識をより定量的分析する手法の有効性を確認した。今後は、テキストマイニングを用いて、直轄河川間の河川整備計画対象の差異等・類型化し、河川生態系管理の代表事例の抽出を行う。

参考文献

1) 電子政府の総合窓口（法令検索）：http://law.e-gov.go.jp/cgi-bin/idxsearch.cgi （2016 年 6 月 13 日確認）
2) 河川法研究会編：平成 24 年度版　河川六法、大成出版社、2011
3) 京都大学 大学院情報学研究科 知能情報学専攻 知能メディア講座 言語メディア分野、黒橋・河原研究室、http://nlp.ist.i.kyoto-u.ac.jp/、2016 年 6 月 13 日確認
4) 金明哲：テキストデータの統計科学入門,岩波書店，2009

RESEARCH ON IMPACT-RESPONSE APPLYING EXISTING DATA

Budged：Grants for operating expenses
General account
Research Period：FY2016-2019
Research Team：Water Environment Research Group
(River restoration Team)
Author：KAYABA Yuichi
DENDA Masatoshi

Abstract：
This study analyzed river improvement plan and consciousness of river managers on river environment using text-mining methods. The results indicated that although river managers have strong consciousness on river environment, and have consciousness differencing from river infrastructure management.

Key words：*Existing data, Impact-Response, River improvement plan , Text mining*