

テキストマイニングを用いた全国河川整備計画の類似性の分析に関する研究

国立研究開発法人 土木研究所 水環境研究グループ河川生態チーム 主任研究員 正会員 ○傳田 正利
同上 上席研究員 正会員 萱場 祐一

1. はじめに

現状の河川生態系の修復、再生及び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「対策的」な取り組みが多い。これに加え、河川事業による人為的インパクト (I) と河川生態系のレスポンス(R)の関係性であるインパクト・レスポンス(以下、「IR」と記述する。)を分析し、河川生態系への負の影響を与える事象を軽減することで、より有効な河川生態系の保全・修復が可能となる。

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分析が必要であるが、「河川整備計画」は、最も適した資料の一つといえる。河川管理の実務では、河川法¹⁾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(河川法 16 条, 16 条の 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との整合を図る等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。

全国の河川整備計画を通読していると、それらの目次構成には類似した傾向があることに気づく。筆者ら所感では、流域や生物の分布特性を整理した後、治水・利水の変遷、河川の治水・利水・環境面の課題整理とよりよい河川管理への方向性、その手段として地域協働の川づくり等が述べられるのが一般的な構成と考えている。

しかし、これらは主観的な印象に過ぎず、より定量的な分析が行えれば、河川管理者の計画に対する試行の流れ、河川環境管理で問題視される主な事象の抽出が可能になる。

このような背景から、本研究では、全国河川の河川整備計画を収集し「テキストマイニング」という近年普及が進むデータマイニング手法を用いて、河川整備計画の構成の共通性を定量的に分析し、河川整備計画の思考の流れを分析することを目的とする。

2. 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、人文・社会科学の分野は工学に比べて定量化されたデータは少ないが、「文章」は大きな可能性を持つ。近年のインターネットや情報システムの普及に伴うテキストの普及が進み、テキストデータの活用や定量的な解析手法が求められている。これらの要請に応えられるのが、「テキストマイニング」の手法である³⁾。上田の定義によれば、「大量のテキストから、新たな事実や傾向を発見することを支持する技術」⁴⁾、金の定義によれば、「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」³⁾をいう。近年、言語文体学の分野等において、急速に発展・普及し、今後の発展が期待される手法である。

3. 研究の方法

(1)テキストデータの作成とテキストマイニングの流れ

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手順を整理しながら、以下に分析手法の概要を示す。

本研究では、全国 109 水系の河川整備計画の PDF が公開されている HP⁵⁾から、2016 年 12 月時点で、河

キーワード 河川整備計画, テキストマイニング, 目次分析, 形態素, 河川環境

連絡先 〒305-8516 茨城県つくば市南原 1-6 国立研究開発法人土木研究所 水環境研究グループ河川生態チーム

TEL. 029-879-6775 E-mail : denda@pwri.go.jp

表-1 形態素の章別の頻度比較

	1章	2章	3章	4章
例：山国川	山国川の概要	山国川の現状	河川整備計画の目標に関する事項	河川整備の実施に関する事項
計画	214		121	
整備	212	123	223	347
概要	199			
目標	183		177	250
課題	144	203		
流域	142	109		
利用	126	139	112	153
治水	120			
沿革	114			
現状	113	200		
環境	105	128	106	153
管理		168		217
川		146		
洪水		113		140
維持		106	105	204
事項		105	155	191
対策		102		122
保全				175
機能				125
災害				100

川整備計画が公開されているもの 100 河川を対象とした。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、画像として公開されている PDF は文字認識ソフト (NTT Data 社, e. typist Ver. 15) によりテキスト化した。テキスト化したデータには、必要としない記号 (ルビ等) 等が含まれていたため、不要な情報を取り除いた。その後、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」⁴⁾⁵⁾、文法規則に基づいて、文の構造を句・文節単位として解析する「構文解析」を行った。「形態素解析」は、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⁶⁾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。

(2)形態素解析後のデータ集計

筆者が、対象とした河川整備計画の内、典型的な章立てと主観的に判断した河川をデータ分析の比較対象として選定した。次に、形態素を各章別に頻度を分析した。これは、節以下のレベルで用いられる形態素の特性から、各章の内容を判断するために行った。1章から4章の優先順位で降順にソートし、全形態素の内、上位10位までの形態素を特定した。多くの河川整備計画は、5章構成であったが、5章以降は、まとめが記述されていることが多いため、解析から除外した。一部の河川では、「編、章、節、項」

に目次立てすることがあったが、編を章として扱った。

3. 結果と考察

表-1 に、筆者が典型的と判断した「山国川」の章立てと対象河川の各章の形態素の頻度と上位 10 個の形態素を示す。1 章においては、「計画」・「整備」・「概要」・「目標」等の総括的な形態素が主であり、その後治水・環境に関する形態素の頻度は類似した。2 章においては、「課題」・「現状」が、3 章・4 章においては、「整備」、4 章においては、「整備」・「目標」・「管理」・「維持」が多かった (ハッチングが上位 10 位の形態素)。

上述の結果から、河川整備計画の思考の流れとしては、1 章において流域の治水・利水・環境に関する概要整理を行い、2 章において、現状の課題の整理を行い、3 章において目標設定、4 章において、その整備・目標の実施 (管理・維持) 方法が示されると考えられる。1 章の記述に基づき各河川の特徴を全国規模で相対的に理解し、2 章で整理された課題の整理し、その課題が生じた要因を河道管理データを対象に河川工学を用いて、定量的に分析をすることで、IR 研究が進展すると考えられる。

4. まとめ

テキストマイニングを用いて、河川整備計画の分析を行った。その結果、管理上の課題は 2 章において指摘されており、インパクト・レスポンスの分析には、2 章の重要性を明らかにした。

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) 金明哲:テキストデータの統計科学入門,岩波書店, 2009
- 4) 上田太郎 (監修)・村田真樹ら:事例で学ぶテキストマイニング, 共立出版, 2008.
- 5) 国土交通省 : http://www.mlit.go.jp/river/basic_info/jigyo_keikaku/gaiyou/seibi/, 2017 年 1 月 10 日確認.
- 6) 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/>, 2017 年 1 月 10 日確認.