テキストデータマイニングを用いた河川整備計画分析に関する研究
- 主要直轄河川を対象とした試行的分析 -

傳田　正利[1]・萱場　祐一[2]

[1]正会員　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（〒305-8516 茨城県つくば市南原1－6）
E-mail:denda@pwri.go.jp

[2]正会員　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（〒305-8516 茨城県つくば市南原1－6）
E-mail: y-kayaba@pwri.go.jp

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」への意識を分析すると同時に，主要河川を分類した．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があった．主要河川の分類の結果，北海道・東北北部と本州の河川は異なるグループに分類され，その傾向は，自然度と開発の程度に影響を受けている可能性があった．

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

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，河川事業による人為的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.）の関係性を分析し，河川生態系への負の影響を軽減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，河川整備計画の策定時には，河川管理者自らが，河川管理情報の収集・精査，今後の河川管理の重要事項の選定を行い，上位計画者から実務者までの対象河川の河川整備に関する重要事項の認識が集約される．

また，道路事業と比較して，河川管理は，自然河川を対象とするため，河川管理の課題の個別性が高い特徴がある．河川整備計画の様な，各河川の個別性を凝縮した資料を分析し，直轄河川をその技術課題特性に基づき分類し，各類型に見られる典型的なIR事例を抽出し，詳細な分析を進めることにより，人為的なインパクトの低減による河川生態系の効果的な保全や，一度失われた良好な河川生態系の再生が可能になると考えられる．

このような背景から，本研究では，全国の主要河川の河川整備計画を収集し分析する．分析手法は，「テキストデータマイニング」という近年普及が進むデータマイニング手法を用いて，文章が主となる河川整備計画を定量的に分析する手法を用いて主要河川の河川生態系問題の特徴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，金は．「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．」[4]と指摘し，「『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を指す」と定義する[4].

また，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や，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る現代社会において，これらの要請に応えられるのが，「テキストマイニング」の手法であるとその可能性を示している[4]．「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].

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詳細を整理しながら，以下に分析手法の概要を示す（**図-1**）．

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**）．

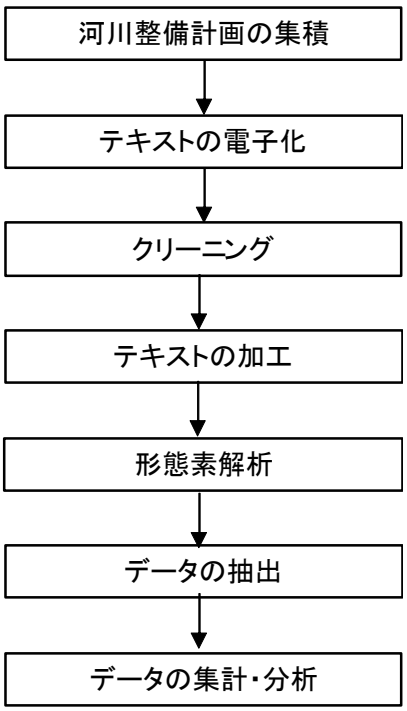

図-1　統計的テキスト解析の過程[4]

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	○

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．

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．

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．

金の定義を要約すれば，「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．

(2)テキストデータの分析

テキストマイニングの基本的分析として，形態素の頻度分析が挙げられる．また，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[4].

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 [4]．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，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）エヌグラム解析を用いた記号列間の関係性分析，（ハ）形態素構成比に基づくクラスター分析を用いた主要河川の分類，以上の分析から，主要河川の治水・利水・環境に関する河川管理者の認識の共通性と個別性を分析した．

3. 結果と考察

(1) 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 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，「ダム」，「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，河川を中心とするネットワークから独立したネットワークとなった．

筆者らの既往研究により，「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 [4]，「河川法の改正後，約 20 年が経過しているが，「環境」は，河川管理実務の重要な位置を占めていることが明らかになった．しかし，**図-2** は，環境，特に，生物・生息場に関する記号列が他の記号列と交わらず独自のネットワークを形成していることも示し，河川の施設管理や水質に関する記号列が他の用語と密接なネットワークを形成している点と対照的である．分野の独実性から致し方ない傾向ではあるが，本来，河川生態系の構造は，物理環境，生物群集及び有機物動態（水質）とが一体不可分のシステムとして認識されるべきである．今後は，河川計画，河川管施設理及び水質等の分野との連携を深め，効果的かつ効率的な河川管理を実現する取り組みが重要である．また，「ダム」，「洪水」，「歴史」，「地震」等は，独自の記号列のネットワークを形成したが，ダムは，流域内土砂の連続性に影響を与え，歴史は河川景観や人の河川利用等には重要な観念である．これらの事象を適切に河川整備計画に反映していくためには，河川が気候，流域（人的利用を含む）で構

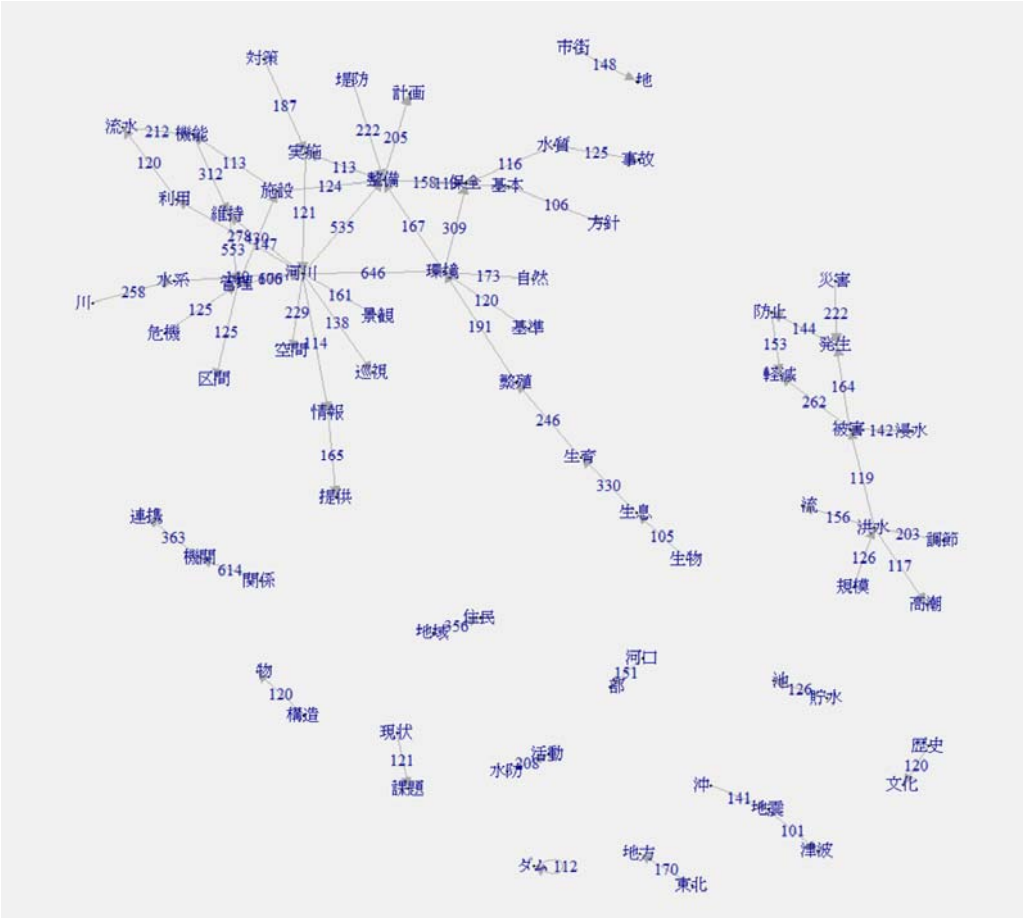

図-2　河川整備計画の分析対象河川 [4]

成される一連のシステムであることを再認識し，治水計画を主とする計画論に関連づけ反映していくことが重要であると考えられる．

⑵ 形態素構成比に基づくクラスター分析を用いた主要河川の分類

図-3 に形態素構成比基づくクラスタ分析を用いた主要河川の分類を示す．クラスタは，グループ１：北海道と東北北部を中心とする河川群，グループ 2：東北南部以南を中心とする河川群に分類された．グループ１内は，天塩川を除く北海道の河川がサブグループを形成した．グループ２内は，淀川，信濃川・天竜川が特徴あるグループを形成した．

図-4 にグループ１とグループ２において頻度高く使われた形態素を示す．グループ１，グループ２ともに形態素の構成に大きな変化はないが，グループ１の方が，「河川」や「整備」といった「治水」に近い語感を持つ語，の形態素の使用頻度が高く「ダム」の形態素が低い傾向があった．これは，グループ１を構成する河川が，北海道を中心とした大河川，かつ，開発され尽くしていない自然度が高い河川である特性に起因すると考えられる．一方，グループ２では，「ダム」をはじめとする人的改変が大きい傾向があるため，使用頻度が高かったと考えられる．

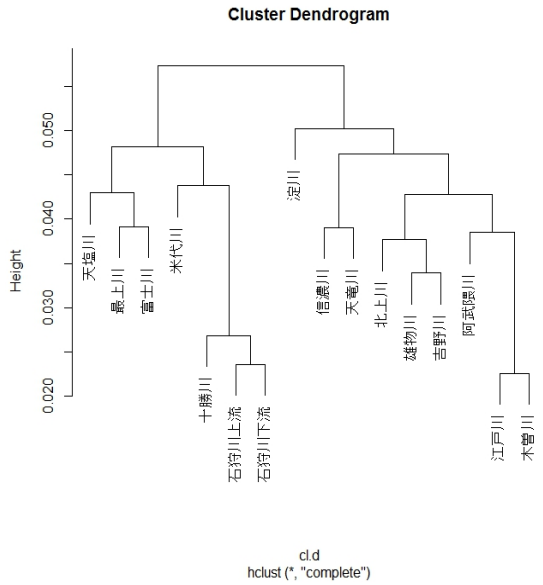

図-3　河川整備計画の分析対象河川

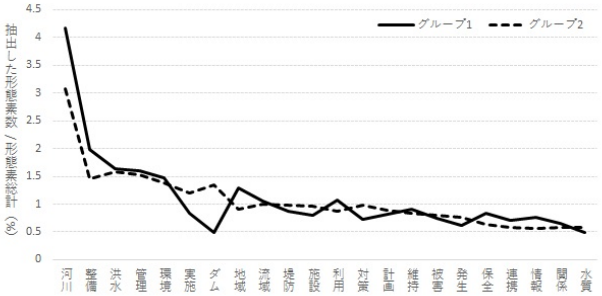

図-4　河川整備計画の分析対象河川

⑶ 今後の研究への展望と課題

本研究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，河川整備計画に共通する形態素，個別河川の特性を検討する可能性を見出した

しかし，テキストマイニングの手法には課題が残る．**図-4** は，個別河川の河川整備計画の特性を相対的には分類できることを示すが，形態素の構成比の差は，%以下の数字に留まる．すなわち，個別の形態素が使われる頻度の差が有効かどうかを見極める解析が必要となる．また，本研究の解析では，形態素を個別に分析したに過ぎない．金は，「語，文節，句などの頻度だけでなく，それらの相互の関連性を分析することが必要である」[4] と指摘している．本研究でも，「河川」，「整備」等の他の語との関係性がある形態素が多く使われた．

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．

4. まとめ

本研究では、テキストマイニングの手法を用いて、主要河川の河川整備計画を分析し、河川管理者の河川計画への意識を分析した．その後，テキストマイニングのデータを用いて，主要河川を分類した．

その結果、河川管理者は、「治水」と同程度に「環境」に関して強い関心を示している反面，環境の内，特に生物生息場保全は，他の河川管理項目とは異なる語のネットワークを形成した．また，主要河川は，地理的位置，開発の程度により分類される傾向があった．

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) 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
5) 傳田正利・萱場祐一：平成 27 年度成果報告書，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，201

(2016.7.31 受付)

RESEARCH ON ANALYSIS OF RIVER MANEGEMENT PLAN USING TEXT DATA MINING

Masatoshi DENDA, and Yuichi KAYABA

This study analyzed river improvement plan and consciousness of river managers on river environment using text-mining methods and classified the rivers based on the mining data. The results indicated that although river managers have strong consciousness on river environment, and have consciousness differencing from river infrastructure management.And the cluster results classified the rivers into two groups, the groups trendo to be separeated based on natural conditions and development conditions of individual river.