河北潟研究に対する計量テキスト分析

樽田泰宜・中森義輝

北陸先端科学技術大学院大学
923-1292 石川県能美市旭台 1-1
（連絡先：樽田泰宜　E-mail：y-taruta@jaist.ac.jp）

Quantitative Text Analysis for the Researches Related to Lake Kahokugata

TARUTA, Y., NAKAMORI, Y.

Japan Advanced Institute of Science and Technology
（Corresponding person; Yasuyoshi Taruta. E-mail; y-taruta@jaist.ac.jp)

要約：石川県の中央に位置する河北潟は干拓により広大な面積の農地を有しており県内の農業や酪農において重要な位置を占めている．この河北潟地域の発展に資するためにこれまでの「河北潟」に関する知見や知識の蓄積に注目し、テキストマイニングという言葉に代表されるコンピュータを用いた文章検索や内容分析を実施した. 具体的には, 学術情報等の検索ができるウェブサービスである CiNii を利用した「河北潟」のキーワード検索により研究タイトルを抽出した後，計量的なテキスト分析を実施することで河北潟の研究動向を明らかにした。さらに「干拓地」についても同様の分析を行うことで干拓地の研究動向における河北潟研究の位置づけを明らかにした．CiNii での検索結果は，河北潟が 217 件で干拓地が 987 件であった. 分析では河北潟と干拓地の研究領域の共起関係, 相関関係, 類似関係を明らかにした．結果として「河北潟」研究の中心を為すクラスター内の語は相関と類似性が高いことが明らかとなった．これは河北潟研究が複数の分野間で研究されていると言うよりも各研究領域の深化方向に動いていると解釈された．一方で干拓地研究という枠組みで「河北潟」を見ると，相関は低いが類似性が高いことが明らかとなった．これは分析時に検索キーワードを「干拓地」としたことで，河北潟研究における河北潟干拓地という語を含む論文タイトルが抽出されたことが影響している．干拓地研究における河北潟の研究動向では動植物・農業がよく研究されており，これは干拓地全体の対応分析結果に見られる「土壌・改良・八郎潟」や「干拓・技術・農業」とは相関が低いことを意味しており，タイトルの出現頻度や量によらない類似性に注目した分析では河北潟と他の潟との間には相関があるということである．類似性に関しては分析データを直接参照すると，両者ともに同一タイトルで副題違いの論文が複数あることが結果に影響していることが分かった．河北潟以外でも一部の潟では経年にわたり研究に従事している研究者や研究論文の存在が分かり，深化方向の研究動向を明らかにすることができた.

キーワード：河北潟，干拓地，計量テキスト分析，テキストマイニング，傾向分析

はじめに

河北潟は，石川県のほぼ中央に位置する潟湖である．この河北潟は 1963 年に国営河北潟干拓土地改良区事業として潟の干拓工事が開始され 1971 年には干陸がなされた．干拓前の河北潟の面積である 2248ha から約 600ha になった．また，干拓地の面積は約 1300ha である.

かつて干拓前の河北潟は日本海側と大野川を通じて繋がっており汽水湖であった．そこから，干拓後に防潮水門を設置して日本海と切り離されることで淡水湖へと生まれ変わったのである．同干拓事業の当初の計画では，戦後の食糧不足への対策として水稲を中心とした造成計画であった．しかしながら，

社会情勢の変化や開田抑制措置として所謂減反政策に伴い水稲から畑作へと計画の変更がなされることになった．河北潟の干拓工事の計画はこの変更も含めて合計 5 回実施されていることが『河北潟干拓事業誌』（北陸農政局，1986）から分かる．

潟の生態系に関しては，淡水域から汽水域へと大きく自然環境を変えたことで潟をはじめとした周辺にも大きな影響を与えている．近年の河北潟に関する生態系に関しては，『河北潟レッドデータブック』（河北潟湖沼研究所, 2013）に良くまとめられている．また，日本海からの潟への海水の流入を防ぐために，防潮水門で区切られており潟内の水位は低く保たれている．それにより閉鎖系水域となっており，石川県（2013）の『平成 24 年度公共用水域及び地下水の水質測定結果報告書』を見ても全窒と全リンともに環境基準を超えている．

社会的なジャーナリズムの視点から河北潟干拓事業についてとりまとめられている資料としては『レポート河北潟』（北國新聞社，1985）がある．同書籍によると，当時の入植者及び入植希望者の多く水稲農家であり大面積を誇る干拓地という地で大規模農業を目指していた．しかし，干拓地も用地の利用が水稲から畑作へと転換されたことで農家の入植時に様々な弊害や問題が発生していたことが分かる．つまり畑作地へと変わったことでこれを断念せざるを得なくなり畑作農家への転換を迫られたのである．さらに同書籍から，干陸前後の干拓地や河北潟での農業の様子は次の 4 点にまとめて理解できる．1. 畑作に適さない土地，2. 土地の排水性に関する問題，3. 入植者自身の問題, 4. 農作物の価格の不安定さ（需要と供給のバランス）である．戦後の食糧不足に答えるために，米の生産地としての干拓がはじまったが，後の米余りにより畑作へ移行した点やそもそも土壌が畑作に適していないという点である．干拓地のもとは潟であり，それを埋め立てたことにより土壌の排水性が非常に良くない．加藤ほか（1995）によると軟弱地盤は 30 から 40 メートルに達するとあり，地盤対策工事の重要性が分かる．

また，河北潟と住民との関係では沢野（1995）は「干拓事業は流域住民の意識の中から潟の姿を遠ざける役割を果したことも事実である . 干拓後の河北潟は依然として石川県下で最大の湖沼であるにもかかわらず , 流域住民の中には河北潟は完全に埋め立てられ．既に消滅したものと信じている者や , 若い世代の中には潟の存在そのものを知らない者もいる . このような社会情勢が , 今日河北潟が直面する複合的な環境問題を招いた原因とも考えることができる .」と指摘している．

しかしながら，河北潟地域に関しては総論的な視点での研究はほとんど為されていない．それに関して大串（2002）は，「･･･ しかし私たちが河北潟について，その自然や歴史を総括的に理解し，また人に紹介しようとすると，これらの問題を適切に取りまとめた文献資料が無いことに困惑します．（中略）この河北潟の歴史と自然環境に関する紹介をするのに必要なまとまった文献がまだ我々の手元に無いということは，この河北潟に関心を持った人達にどう説明してよいか判りません .」と述べている．

では，河北潟の干拓の背景や現在の姿から河北潟地域は十分に理解と活用が為されているのだろうか．干拓の歴史も必ずしも順調であったとはいえず，河北潟のよりよい活用とその理解なくして当該地域の発展は得られないのではないかと考える．そのためには，河北潟のこれまでの研究成果を知ることから始めるのがよいと考えた．それに向けた着眼点としてこれまでの「河北潟」に関する学術的研究とその動向に注目する．河北潟の研究を知るためには個別の論文を熟読することや総説論文にあたる必要があるが，大串（2002）が指摘するように総説・総論的な論文は発見することが出来ない．本研究では，CiNii という学術誌や研究論文のウェブ検索サービスを利用して「河北潟」に関した研究論文等の学術資料をできるだけ多く収集して，得られた研究論文に関するデータはテキストマイニングという手法を用いて計量的なテキスト分析を行うことで「河北潟」の研究動向を明らかにすることを目的とする．また，河北潟の研究動向の把握や理解のためには，河北潟や他の潟も含まれる干拓地という視点も重要であると考えた．そこで，干拓地に関しても同様のテキストマイニング手法を用いて研究動向を明らかにし，干拓地研究における河北潟の位置づけを明確化するための分析も実施する．

テキストマイニング手法を用いて研究動向を明らかにした先行研究として，趙ほか（2013）の「介護

福祉学」の研究論文誌を対象とした分析がある．趙らは1994年から2011年の機関誌を対象としたテキストマイニングを実施することで研究方法の変遷や社会ニーズへの応答といった研究姿勢の変化を明らかにしている．さらに別の研究として，林ほか(2009)がある．これは地域ブランドという学際的な研究領域に対して論文書籍の情報データベースを元にした定量分析（テキスト分析）を実施することでその学問領域の構造を明らかにする研究である．林らは研究領域の構造を分析するに当たり，定性的な分析手法として分析者独自の視点で行うものもあるが分析者によって視点に偏りが発生する可能性があり，一般的・客観的事実の分析という点ではやや問題を抱える点もあると指摘している．そこで主観をなるべく排除して定量的に学術領域を分析する手法として定量的なテキスト分析を行っている．それらの一部は計量書籍学的研究として位置づけられている（大和田，1996）．大和田は，計量書籍学的研究では論文数や引用回数などのデータベースがあって始めて可能になる研究であり，これまでは理系の一部に限られた手法であると述べている．そこで科学研究費補助金の採択データを元にして分析を実施することで所謂人文・社会科学の分野という理系以外の研究動向を明らかにしている．

本研究で実施する計量テキスト分析とは，先に挙げた研究と非常に関係性が高く，一般的にテキストマイニングと呼ばれるものに属している．那須川ほか（1999）はテキストマイニングとは，コールセンターの会話文を元に分析を行ったり，医療データから病理の関連性を見いだしたり，特許情報の関連性や類似性を検索したりして利用されていると述べている．他方，樋口（2004）は，単に商業的な側面での利用だけではなくて総合的なアプローチとして計量的な内容分析（Content analysis）というものを提案している．これは，テキストマイニングや文章検索と行った大きな枠組みの中でCorrelationalアプローチとDictionary-basedアプローチの二つの側面から互いに補完して統合的に扱う試みである．樋口（2004）はCorrelationalアプローチとは，「頻繁に同じ文書の中にあらわれる言葉のグループや，あるいは，共通する言葉を多く含む文書のグループを、多変量解析によって自動的に発見・分類するためにコンピュータを用いるアプローチ」とし，Dictionary-basedアプローチを「分析者が作成した基準にしたがって言葉や文書を分類するためにコンピュータを用いるアプローチである.」と説明している．さらに，樋口（2006）は計量テキスト分析を「計量的分析手法を用いてテキスト型データを整理または分析し，内容分析（Content analysis）を行う方法である.」と簡便に定義している．そして，これら計量的テキスト分析を行うためのKH Coderというソフトウェア＜http://khc.sourceforge.net/＞を公開しており，テキスト分析や内容分析など学術分野で広く利用されている．

本研究では，樋口（2006）の計量テキスト分析の定義に従い河北潟に関する研究領域の内容分析から研究動向を明らかにする．先ほども述べた通り，河北潟の干拓の歴史やその道筋は必ずしも順調であったとはいえず，今後，河北潟に関して研究を進めるためには此まで先人たちが築いてきた「河北潟」という学術領域においてどのような知見や知識の蓄積があったのかをまずは注目すべきであると考えたためである．同時に複数の文献を扱う場合には計量的な分析手法は最適であると考えテキストマイニング手法を用いて分析を行う．

方法

本研究では最初にウェブサイト「CiNii」＜http://ci.nii.ac.jp/＞を利用して「河北潟」というフリーワード検索を行う．そして，検索された結果からタイトル，作者，出版年，発行元などのデータを抽出する．今回の研究とその分析では検索で確実に得られるタイトルのみを扱うものとする．分析では，語の出現の量や頻度，共起関係や相関関係を元に研究動向の把握を実施する．そのため，テキストマイニングや計量テキスト分析で一般的に用いられている方法である「共起ネットワーク」「対応分析」「階層的クラスター分析」「多次元尺度構成法」による各分析を行う．分析の詳細については次項にて述べるが，元データを対象に語の頻度，共起関係や類似度に注目をして計量的に分析してマッピング・可視化する手法である．

なお，CiNiiとは，国立情報研究所（NII）が提

表 1.「河北潟」抽出語の頻出語

抽出語	回数	抽出語	回数	抽出語	回数	抽出語	回数	抽出語	回数
河北潟	68	津幡	11	農業	8	活性化	6	考察	5
河北潟干拓地	37	祭神	10	報告	8	堆積物	6	事例	5
研究	20	河北潟周辺	9	計画	7	動物	6	水質浄化	5
石川	18	環境	9	事業	7	利用	6	水生植物	5
神社	13	現状	9	周辺	7	ノネズミ相	5	堆積	5
調査	13	特性	9	植生	7	沿岸帯	5	地区	5
分析	13	河北潟干拓	8	水質	7	検討	5	地盤	5
分布	13	水	8	年	7	湖岸	5	排水	5

供する学術情報として，論文や図書・雑誌などの各種情報を検索できるサービスであり研究者や技術者の間で広く使われている．CiNii を用いた理由として学術的な研究論文や資料の収集が簡便に行える点と誰でも利用できるオープンなデータベースであり，研究者の情報収集におけるバイアスを排除できるためである．CiNii に収録されない論文誌などもあると考えられるが，そういった論文は個別に収集することが求められる．その際，網羅的な収集やどの範囲まで収集するかといった明確な線引きは難しいことから本研究では CiNii を用いることとした．

さらに「河北潟」という語だけではなく「干拓地」という語についても同様の分析を実施する．「はじめに」でも述べたが干拓地から見た河北潟や他の潟と河北潟の関係性を知ることは，河北潟研究の理解には重要であると考えたためである．

分析には樋口（2004）が開発した KH Coder'（Version:2beta32c）を用いる．同分析ソフトは，計量テキスト分析に広く用いられているソフトウェアであり，KH Coder のホームページ＜ http://khc.sourceforge.net/ ＞からダウンロードして自由に利用することが出来る．

「河北潟」の分析結果

1．分析対象の基本情報

CiNii の論文検索のフリーワード検索で「河北潟」と検索した結果は 217 件であった（2015 年 2 月）．これを初期値データとした．この時，CiNii に本文があるか連携サービスがあるかなどは問わずに「すべて」から検索した．さらに分析に使用する文章（タイトル）のスクリーニングを実施したものは分析データとした．河北潟に関係のない論文と出版及び発行が 1950 年以前の論文で旧字体で書かれているものは除外した．さらに英語タイトルは 3 件あったが今回は日本語で表記されたものを対象としてこれを除外した．その結果，1950 年から 2013 年の 195 件となった．また，今回の分析では得られたタイトルのみを用いてその他の要旨，作者，年代，出版元などのデータは扱わない．これはタイトル以外のデータは分析に使用するには欠損が多いためである．次に，分析に使用するタイトルを基にした分析データ中の出現頻度が 5 以上の語を表 1 に示す．

表 1 は出現した語を抽出語としてその出現回数と共に示している．例えば，分析対象中に「河北潟」という語は 68 回出現したということを示す．河北潟に次いで出現回数が多かったのは「河北潟干拓地」の 37 語，「研究」の 20 語となった．

表 1 の作成にあたり事前に分析データに対して複合語を調査した．複合語の検出には形態素解析ツールの「茶筌（Chasen）」を使用した．茶筌＜ http://chasen-legacy.osdn.jp/ ＞とは，広く自然言語処理研究に資するため無償のソフトウェアとして開発されたものである．茶筌の著作権は，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座（松本研究室）が保持している．それにより複合語には出現頻度の多い順に，河北潟干拓地，河北潟干拓，河北潟周辺などがあることがわかった．これらの複合語に関しては「河北潟」+「干拓地」と個別に頻度をカウントするのではなく複合語を単体の語としてカウントするようにした．この理由は，次節以降の分析の際に「河北潟干拓地」が河北潟と干拓地で

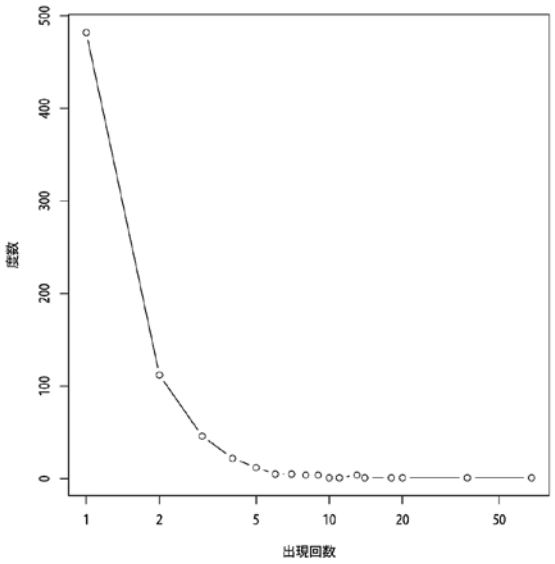

図 1.「河北潟」の出現回数と度数

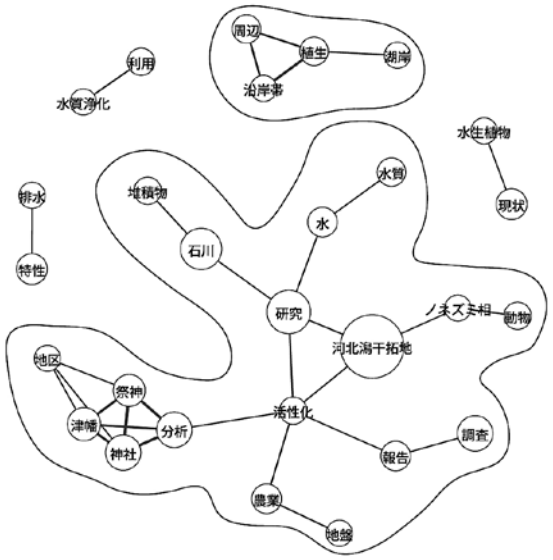

図 2. 共起ネットワーク「河北潟」

個別に処理されると,「河北潟」と「干拓地」の間に強い相関が発生することで本来知りたい事象である「河北潟干拓地」に関した相関に影響が出ると考えたためである. なお, この複合語を単一語とすることで「河北潟干拓地」の中には「河北潟干拓」という語は含まれずに別々の語として扱うことになる.

同様の複合語として, ノネズミ相, 水質浄化, 水生植物などの語を単一の語として扱った. 複合語以外にも「河北潟」の特徴として「河北潟」という接頭語の複合の多さが一つに挙げられる. 表 1 に示した頻出抽出語以外にも「河北潟」と「流域」・「沿岸」・「地方」・「底質」などの語があることが分かった. 次に出現回数とその度数を図 1 に示す. 1 回出現した語が 482 種あり, その出現数の全体の割合は約 68％であった. 出現回数 5 回のものは 12 種類の語がありその割合は全体の 1.71％で累積度数が 673 種類あり累積割合は 95.73％であった. また, 出現回数 5 回以上では度数が平滑化していくことが分かる.

2. 共起ネットワーク分析

ここでは共起ネットワークの結果と解釈について述べる. 共起ネットワークとは, 分析データにおいてよく一緒に共起する語を図示したものであり、図 2 の共起ネットワーク図では共起が強い関係ほど太い線で結び, 出現回数の多い語ほど大きい円で示した. なお, この分析では, 最小出現数 5 以上の語を対象としている. これは, 分析対象となる語を多くすると, たくさんの語が布置されて図が埋まってしまうことで図の解釈が困難になるためである.

共起ネットワーク図を概観すると「河北潟干拓地」が中心となっている大きなグループが一つと「植生」中心の中グループとその他の小グループが 3 つあり, それぞれが独立していることが分かった.

大グループでは「河北潟干拓地」と「研究」「活性化」「計画」が相互に共起している. さらに「研究」には「計画」「活性化」,「水」「石川」という語に共起が見られる. また,「計画」と「活性化」の間には強い共起が見られる. それらに共起している「分析」には,「神社･津幡･祭神」という語がある.「農業」には「地盤」に共起が見られる.「水」に関しては「水質」,「石川」は「堆積物」,「ノネズミ相」は「動物」へ共起が見られ,「活性化」には「報告」「調査」に共起が見られる.

中グループの「植生」に関しては,「沿岸帯」「周辺」などの語に共起が見られる. その他, 小グループには,「水生生物」と「現状」,「排水」と「特性」,「水質浄化」と「利用」に共起が見られる.

これらから「河北潟干拓地」を中心とした大グループには 7 つの研究サブグループ, そのグループとは

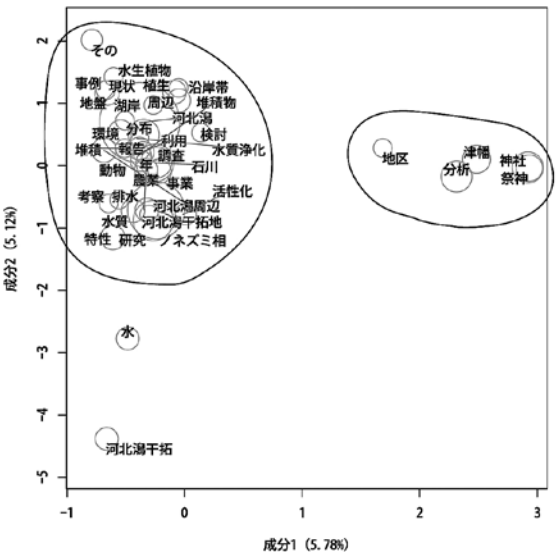

図 3. 対応分析「河北潟」

異なる中グループ 1 つと小グループ 3 つがあると解釈した．大グループは,「活性化計画」「農業地盤」「神社分析」「水と水質」「ノネズミ」「石川･堆積物」「報告・調査」の 7 つの共起関係のある研究サブグループと，中グループに「植生」に関する共起関係の研究グループ, 小グループには,「水生植物」「水質浄化」「排水特性」の 3 つ共起関係の研究グループがあると解釈した.

3. 対応分析

対応分析（コレスポンデンス分析）の結果を図 3 に示す．対応分析とはデータの傾向を図から直感的に理解するのに役立てられる方法である．図では対象となるデータ（行列）の要素に対して相関関係が最大となるようにして，相関関係や同時によく出現しているなど関連の強い語は近くに布置して弱い語は遠くに布置して図として表示するものである．対応分析では，出現回数 5 回以上の語を用いて分析を行った．なお，ここでいう対応分析とは SPSS に代表される統計解析ソフトで用いられているコレスポンデンス分析と同じ分析方法を指す．本研究で用いた KH Coder では対応分析という標記されているので本論文では対応分析に統一した.

分析の結果，全体を概観すると原点附近に語のクラスターがあり，その周囲に別のクラスターや語が布置されていることが分かる．原点附近に布置されている語が河北潟研究の中心を為すものであると解釈した．また，右には津幡や神社という語が布置されている．河北潟干拓という語は離れて布置している.

分析結果から分かる全体の特徴として，図中の河北潟研究の中心に集積して布置されている語同士は相関関係が高い（関連性が強い）ものであり，ひとつのまとまった「かたまり」または「クラスター」として認識できる．相関が高くない語には「神社」関係があると解釈した．なお，この図だけでは河北潟干拓という語が離れて布置している理由の説明は困難である.

そこで，分析対象である分析データをコンコーダンスにて「河北潟干拓」という語を検索して前後のつながりの確認を行った．そこから「河北潟干拓」という語は出現回数の低い語と合わせて出現する傾向が高いことが分かった．さらに，論文タイトルを直接参照したところ「河北潟干拓」という語は場所を意味する語として使われており，原点附近の分析方法や研究対象を表す語と比較して相関が低いために離れた位置に布置していると考えることが出来る．対応分析により全体の傾向は把握することが出来たが，この語の「かたまり」だけでは内容を詳しく把握することは困難であるために次に階層的クラスター分析を実施する.

4. 階層的クラスター分析

階層的クラスター分析による結果を図 4 に示す．さらにクラスターの併合水準に関して図 5 に示す．図 4 の縦軸の数値は図 5 の併合水準と同じ測度で表示してある．階層的クラスター分析とは, 出現パターンの似通った語の組み合わせにどういったものがあるのか階層とクラスター（かたまり）で樹形図（デンドログラム）に示したものである．分析の結合は Ward 法で距離は Euclid 距離を用いて分析を実施した．Ward 法とは志津・松田（2011）によると,「クラスターとしてサンプルをまとめるときに生じる, 各サンプルの情報の損失量の増加分をクラスターの距離とする方法である．すべてのクラスター内の偏差平方和の和をできるだけ小さくするように組み合

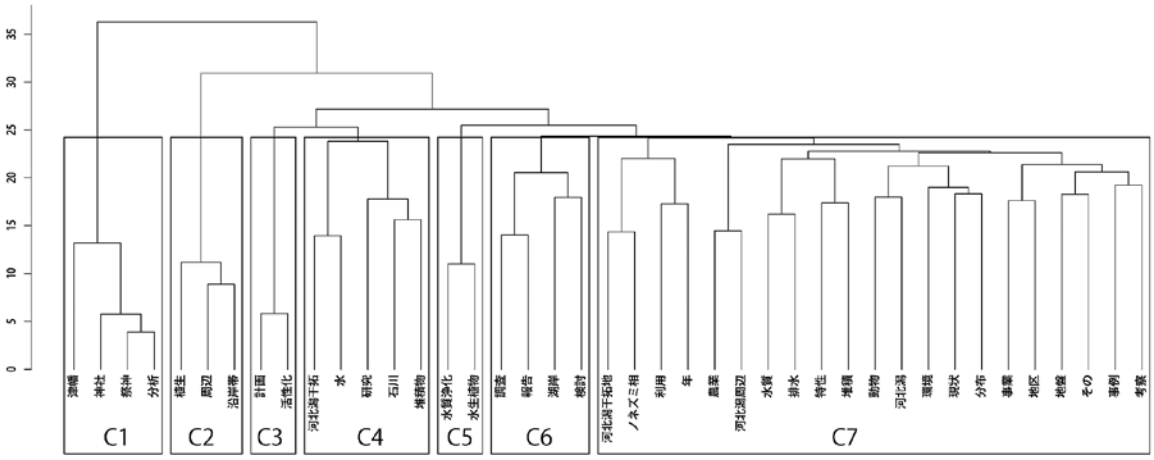

図 4. 階層的クラスター「河北潟」

わせていくので，比較的まとまりのあるクラスターがいくつか得られる」と説明している．そのため，階層的クラスター分析では，比較的良好な結果をえることができる Ward 法が広く使われる傾向にある．

図 4 ではクラスターを Euclid 距離の併合水準が 24 附近を基準に 7 つのクラスターとして C1 から C7 という記号を付加してナンバリングした．図 5 の併合水準はクラスター数が横軸（x 軸）の原点 0 方向に近づくにつれて増加することを示している．図 5 内の数字は縦軸の併合水準での併合されるクラスターの総数を示している．

図 4 では C1 と C2 は併合水準が 13 以下に水準があることが分かる．C3 と C4 は併合水準 25 附近では 1 つのクラスターであるが，それ以下だと 2 に分けられる．C5，C6，C7 は併合水準 27 附近では一つであるが，26 附近で C5 と C6+C7 に分かれる．そして, 24 附近では C6 と C7 に分かれてクラスター化されている．

図 5 の併合水準ではクラスター数が 1 から 4 までは急峻であるが，クラスター数が 5 から 34 までは併合水準が減少するにつれてほぼ線形を示すことが分かる．そこで，ここでは分析結果のから意味を読み取ることのできるクラスターとして併合水準 25 附近が妥当であると判断した．

図 4 からは, C1 と C2 は併合水準が 13 以下であり，他の研究クラスターとは出現パターンが異なることが分かった．さらに各クラスターは次のように解釈した．C1 は津幡・神社のクラスター，C2 は沿岸帯周辺の植生のクラスター，C3 は活性化・計画のクラスター，C4 は河北潟干拓と堆積物のクラスター，C5 は水生植物･水質浄化のクラスター，C6 は調査･報告・検討のクラスター，C7 は動植物・農業・潟の環境に関するクラスターであると解釈した．

5. 多次元尺度構成法の分析

ここでは多次元尺度構成法（MSD）で分析を行いその結果を解釈する．ここまでの分析では語の量や頻度の視点を中心に分析を行ってきた．そこで次に，量や頻度ではない類似性の視点から分析を行う．そのための手法として多次元尺度構成法（MSD）で分析を行いその結果を解釈する．最初に階層的クラスターの理解を補助する分析として量や頻度を表す古典的 MSD を示して次に類似度の MSD を示す．量や頻度の MSD としての結果を図 6 に示す．このとき距離は Euclid 距離にて分析を行った．

図 6 は階層的クラスター分析で解釈を行った 7 つのクラスターを参考に黒丸を付けて示した．図 6 では，原点附近に多く布置している語が中心的なクラスターである．細かく見ていくと，原点附近の多くの語が集積している語には，「ノネズミ相」「河北潟干拓地」などが含まれており階層的クラスターで示された C7 の語が多く布置していることが分かる．原点周囲には，C6「湖岸・検討」などが布置されており, C3 の「活性化･計画」も近くに布置されている．

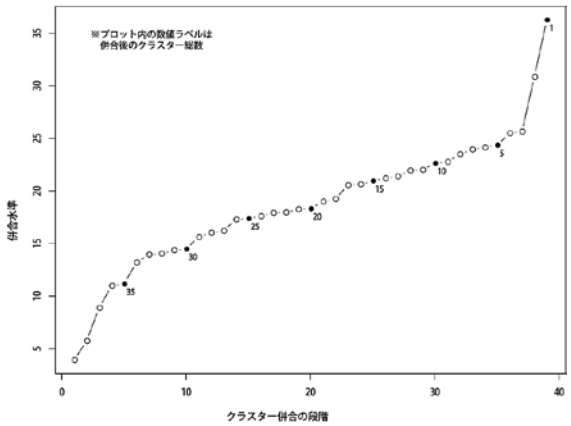

図 5. 階層的クラスターの併合水準「河北潟」

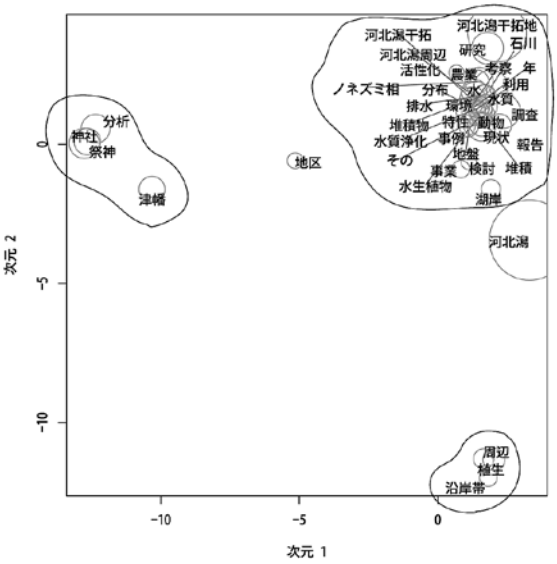

図 6. 多次元尺度構成法（Euclid）「河北潟」

C7 に含まれていた「地区」は原点からやや離れて布置されている.

C1 の「津幡・神社」と C2 の「沿岸帯・植生」は, ここでもやはり離れて布置されていることが分かった.

次に, Jaccard 距離（係数）を用いた MSD 分析とその解釈を行う. 計量テキスト分析において同距離（係数）は出現頻度に左右されずに（非）類似の語のパターンを知りたいときに用いられる係数である. つまり, 同じ語がたくさん出現している場合でも同様の出現パターンが多くても分析結果に影響を与えずに（非）類似度を知ることができるのである. 結果は図 7 に示した.

多次元尺度構成法（Jaccard）「河北潟」（図 7）と多次元尺度構成法（Euclid）「河北潟」（図 6）の結果をどちらか一方を回転すると似た図になるが意図した結果ではない. 分析データが大きく分散している場合, 語は全体的に布置されるはずである. しかし, 図 6 と同様に原点附近に多くの語が布置されていると言うことは分析データの類似度が非常に高い結果であると解釈できる.

なお, 原点附近には,「河北潟･河北潟周辺･水質･特性･環境」などが布置されており周囲には「河北潟･湖岸」「河北潟干拓地・農業」「活性化・研究」が横に広がっている.「植生・沿岸帯・周辺」「地区」「津幡・神社・分析」は原点から離れた位置に布置されていることが分かった.

「干拓地」の分析結果

1. 分析対象の基本情報

「干拓地」に関して CiNii にて検索した結果は初期値データとして 987 件（2015 年 2 月）であった. そこからスクリーニングを行い分析データの 974 件のタイトルを抽出した. 分析データ「干拓地」に含まれる抽出語とその頻出回数を表2に示す.「干拓地」の出現回数は 925 回ともっとも多い. 次いで「土壌」「研究」と続く. 特徴的な語を挙げると, 八郎潟（222 回）, 児島湾（63 回）, 中海（42 回, 河北潟（39 回）, 笠岡湾（39 回）, 有明海（37 回）, 諫早湾（36 回）という各干拓地や干潟の固有名称が挙げられる. () 内はいずれも出現回数である. 出現回数と度数を図 8 に示す. 出現回数では 4 回で累積割合が 80.58％となり 10 回で 90.48％であった. なお抽出結果のスクリーニングでは, 副題が含まれるタイトルはその副題を削除せずにそのまま扱うものとした. 理由は, その副題の重要度に対してバイアスをかげずに一定の基準を設けて自動的に判断できないためである.

2. 共起ネットワーク分析

干拓地の共起ネットワークについて分析とその結

表 2.「干拓地」頻出語

抽出語	回数	抽出語	回数	抽出語	回数	抽出語	回数	抽出語	回数
干拓地	925	整備	58	笠岡湾	39	性質	32	域	27
土壌	404	変化	58	干拓	38	特性	32	技術	27
研究	279	生育	50	事例	37	保全	32	経営	27
八郎潟	222	特集	49	地盤	37	過程	31	営農	26
水田	113	集落	48	有明海	37	利用	31	岡山	26
農業	91	排水	47	諫早湾	36	乾燥	30	窒素	26
土	82	水	46	作物	35	試験	29	畑	26
改良	75	栽培	45	地域	34	生成	29	化学	25
環境	75	調査	45	及ぼす	33	粘土	29	地区	25
水稲	72	影響	43	構造	33	農村	29	地震	25
児島湾	63	中海	42	年	33	効果	28	土地	25
計画	59	河北潟	39	沿岸	32	クリーク	27		

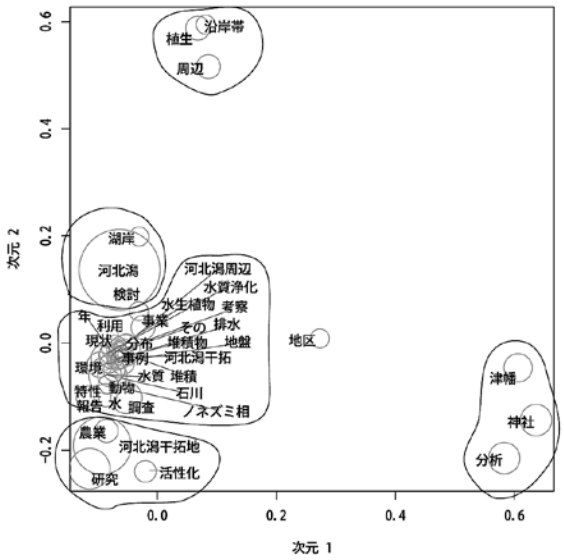

図 7. 多次元尺度構成法（Jaccard）「河北潟」

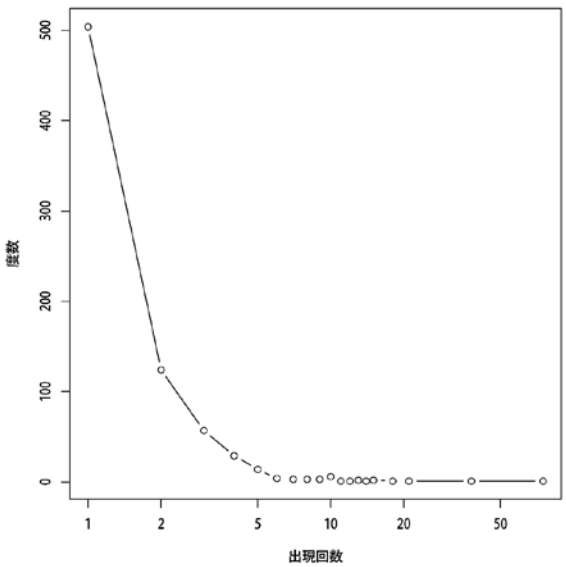

図 8.「干拓地」の出現回数と度数

果について述べる．結果は図 9 に示す．分析には最小出現数 25 以上の語を用いた．干拓地では大きく 3 つのグループと 1 つか 2 つのリンクを持つ小さな 7 つのグループが得られた．大グループには，干拓地に共起するものとして「土壌・改良・八郎潟」に関するグループと，「干拓・技術・農業」のグループと「特集」と「整備」でリンクされた「有明海・クリーク」に関するひとつのグループ，さらに「水稲・生育」に関する別のグループであった．小グループに関しては，「土地・利用」「地震・年」「環境・保全・経営」「地盤・乾燥」「岡山・児島湾」「影響・効果」「生成・過程」という語に共起が見られた．なお，潟の固有名詞に関しては，大グループに八郎潟と有明海が布置されている．また，小グループには，児島湾が布置された．笠岡湾，諫早湾，中海，河北潟は共起ネットワークには現れなかった．

3. 対応分析

干拓地の対応分析の結果について述べる．干拓地の対応分析結果を図 10 に示す．対応分析の結果では共起ネットワークの結果も踏まえて原点附近を中心とした「かたまり」を中心に 5 つのグループに分かれて解釈できると考えた．さらに，原点附近に布置されている語が「干拓地研究」の中心を為すもの

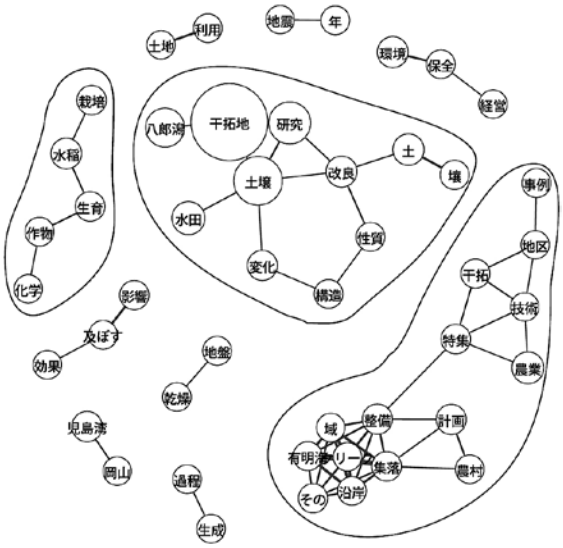

図 9. 共起ネットワーク「干拓地」

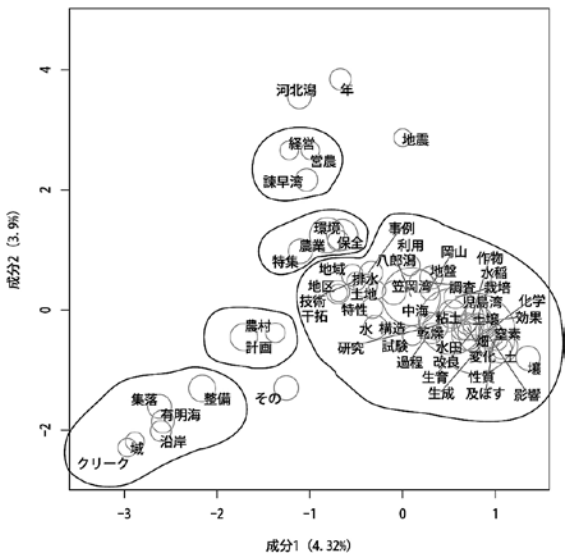

図 10. 対応分析「干拓地」

と考えた．原点附近には語が固まって布置されており解釈が難しいためにその周辺の語について解釈する．周辺には，「農村・計画」「農村・環境・保全・特集」「諫早湾・営農・経営」「有明海・クリーク」などが布置されている．なお，対応分析では「河北潟」という語が上方に布置されているのが確認できる．「八郎潟」の近くには，「児島湾」「笠岡湾」「中海」などの他の干拓地が布置されており，干拓地研究で良く研究されている潟や干拓地がよく分かる．また，この結果だけに注目をすると中心的な干拓地研究と比較して「諫早湾」「有明海」「河北潟」という語は別の研究領域またはグループに分けられると解釈できる．

4. 階層的クラスター分析

階層的クラスター分析の結果とその解釈について述べる．分析の結果として図 11 を得た．クラスターの併合水準は図 12 に示した．ここでは分析結果のから意味を読み取ることのできるクラスターとして併合水準 57 附近が妥当であると判断した．併合水準からクラスターは大きく C1 から C6 に分割した．また，C3 と C6 に関してはさらに 2 分割して C3-1 と C3-2，C6-1 と C6-2 として解釈する．

各クラスターは順に，C1 は「整備・計画とクリーク」に関するクラスター，C2 は「土・壌・研究」に関するクラスター，C3-1 は「児島湾・農村・水田・土壌」に関するクラスター，C3-2 は「畑・窒素・水稲」に関するクラスター，C4 は「諫早湾・営農」に関するクラスター，C5 は「環境・保全」に関するクラスター，C6-1 は「八郎潟・笠岡湾・河北潟」に関するクラスター，C6-2 は「干拓・排水・地盤・有明海・中海」に関するクラスターと解釈した．

5. 多次元尺度構成法の分析

多次元尺度構成法（MSD）の結果とその解釈について述べる．最初に Euclid 距離の MSD の結果を図 13 に示す．次に，Jaccard 距離の結果を図 14 に示す．いずれも階層的クラスターで得た結果をうけて解釈しやすいようにクラスター数を 6 分割して表示した．MSD の Euclid 距離では原点附近に「干拓地」とそれに近い語として「水稲・窒素・効果・作物」などが布置されている．上方には，「土壌・改良・水田・性質」等の語が布置されている．下方には，「笠岡湾・干拓・特性・農村」や「河北潟・諫早湾・八郎潟」などが近くに布置されている．少し離れた位置には，「環境・保全」と「有明海・クリーク」などの語が布置されていることが分かった．

MSD の Jaccard 距離では原点附近に「八郎潟」や

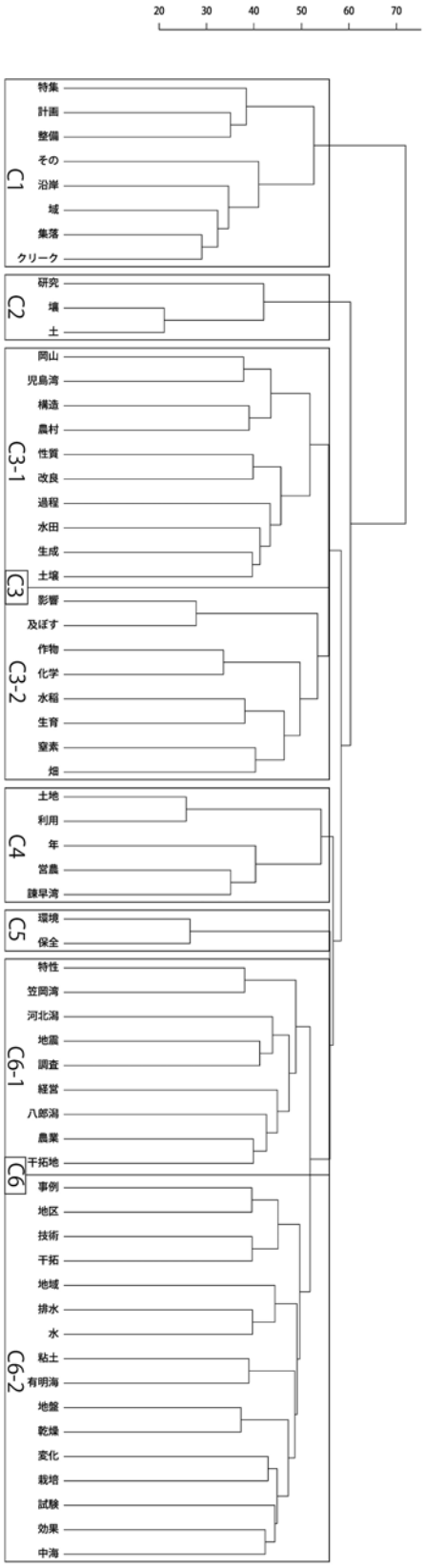

図 11. 階層的クラスター「干拓地」

「畑・調査・栽培」などが布置されそのグループには「中海」「笠岡湾」などが含まれている．その周囲には，クラスターとして「干拓地」「児島湾」「水稲・化学」など，さらにその上方には「土壌・研究・水田」が布置されている．原点下方には「河北潟」「諫早湾」「営農・経営」や「環境・保全・農業・特集」などが布置されており，側方には「農村・計画」がある．一番離れているクラスターは「有明海・集落・クリーク」という語の布置であった．

考　察

河北潟の分析結果から河北潟研究は相関が高く類似性の高い研究動向であることが分かった．対応分析や MSD 分析の結果からも中心を為すクラスター内の語は相関と類似性が高いと判断できる．さらに，「河北潟」で行われている研究領域は原点附近に集中している語であると解釈できる．つまり，研究領域の広さよりも同様の研究領域が深められていると解釈できる．

ただし，対応分析でも見たように原点附近を中心として布置されている語は，タイトルの相関の高いものであり，例えば「河北潟干拓地」と「研究」という共起だけをみると「河北潟干拓地の・・・という研究」のような論文タイトルの骨格が抽出されたと解釈できる．

また，「活性化」「計画」のみに注目をすると活性化研究が盛んであるようなイメージがある．実際に論文を参照すると，これは同一タイトルによる副題違いが 6 件（第一報から第六報）ありこれに強く共起した結果である．この種の副題違いの論文タイトルは「ノネズミ」や「津幡・神社」に関しても同様であった．さらに，分析データからは，この「活性化・計画」「ネズミ」「神社」は同一の研究者が経年に亘り研究を実践していることが明らかとなった．

上記の点と分析データの参照を踏まえると「河北潟」の中心的な研究動向は，動植物の生態関係，水（水質・排水特性），農業・土木分野の研究であるといえる．干拓地の共起関係に注目した結果では，農業関係（土壌・水稲），干拓や整備に関する研究が主流を占めることが分かった．また，研究動向は 6 から 8 クラスターに分割すると解釈しやすいことが

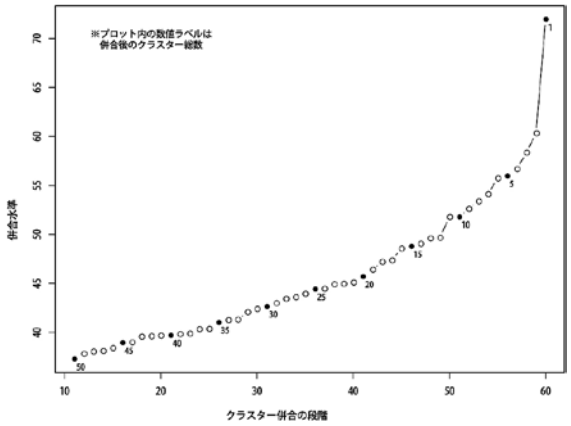

図 12. 階層的クラスターの併合水準「干拓地」

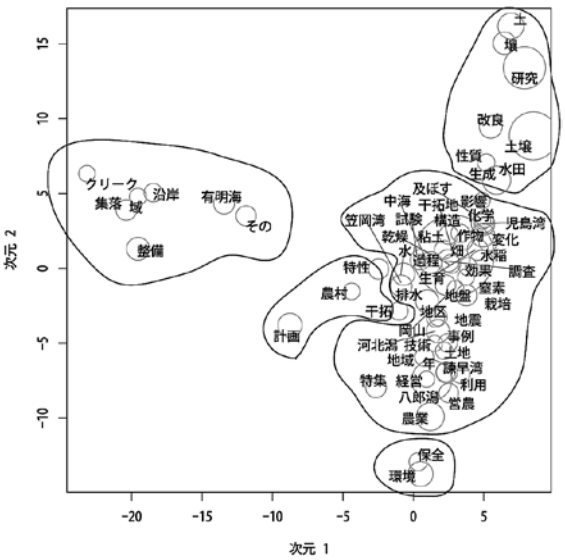

図 13. 多次元尺度構成法（Euclid）「干拓地」

分かった．対応分析の結果では，中心を為す語と「河北潟」は相関が弱いことが明らかとなった．しかし，Jaccard 距離による類似性を基にした分析に注目すると「河北潟」は他の潟や干拓地と類似性を持った研究フィールドを形成している可能性があることが分かる．つまり，「干拓地」研究においては「河北潟」研究は相関が低いが類似度のある研究が行われているといえる．しかしながら MSD の結果では語の関係性や類似度の傾向を知ることは出来たが，語の集中により詳細の把握は困難である．そのため MSD では階層的クラスターの結果をもとにクラスター化した結果に注目することで研究動向を把握することが可能である．

一方，CiNii での「河北潟」という検索結果は 212 件であったが，「干拓地」という検索結果では抽出語を見て分かるように「河北潟」は 34 件であった．分析データ「干拓地」に対して「河北潟」という語でコンコーダンス検索を行うと「河北潟干拓地」という語が抽出される．つまり干拓地研究に含まれる河北潟研究は「河北潟干拓地」に関する研究であるといえる．この結果から「干拓地」の分析データに「河北潟干拓地」以外の「河北潟」という語を含めた分析を実施すべきであったという指摘もあるかもしれない．しかしながら今回の分析ではそれらは含めなかった．その理由としては，「河北潟」のみ河北潟干拓地という語以外の検索結果を含めることは不自然である点，さらにもしもそれらを含めたデータベースを扱うならば八郎潟や児島湾などの他の干拓地も同様に網羅的に扱うべきであるという点，また，干拓地で抽出された各干潟や潟以外の語で干拓地研究に関係している論文はどのように扱うべきなのか判断が難しい点が挙げられる．

しかしながら，「干拓地」研究における「河北潟」研究をより具体的に説明するために「干拓地」における「河北潟」の研究に注目した結果を次に示す．最初に「干拓地」に対して「河北潟」のコンコーダンス結果の 34 件の論文の内容を参照して分析後に分類・コード化する．結果は，小動物関係が 18 件，活性化が 6 件，営農関係が 5 件，水関係が 3 件，その他（土木・土地利用など）が 2 件に分類またはコード化することができた．小動物関係と営農関係は関連性が高いためにグループ化して捉えることもできる．コード化の結果は「河北潟」212 件に見られた分析とは異なる結果である．つまり，単純に検索語「河北潟」と検索語「干拓地」における「河北潟」を比較できないということが分かった．

このことについて CiNii で干拓地という検索で得られた結果は，河北潟干拓地という語を含む論文タイトルであり，「河北潟」研究における河北潟干拓地に関する研究のみが表出されているため得られた結果であると考えられる．さらに「干拓地」研究で

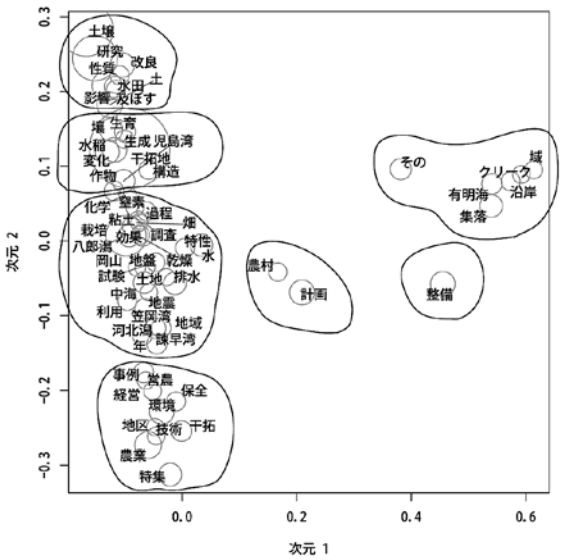

図 14. 多次元尺度構成法（Jaccard）「干拓地」

の他の潟や干拓地も同様に固有名詞＋干拓地における研究結果が抽出されている可能性もあり，この点に関しては今後の課題として研究する必要がある．そこで，本研究では「干拓地」研究における河北潟研究の動向として研究対象に差異は見られるが，他の潟とは類似性に相関があると解釈できる．類似性に関しては分析データを直接参照すると両者ともに同一タイトルで副題違いの論文が複数あることが影響していると考えられる．この点に関しては，分析指針として副題違いも分析データに含めたことで相関関係や経年にわたり研究に従事している研究者や研究論文を発見することが出来た．また，河北潟以外でも一部の潟では経年にわたり研究に従事している研究者や研究論文の存在が分かり深化方向の研究動向を明らかにすることに繋がった．

他方，このようにサンプルが少数である場合は，テキストマイニングなどの計量的な手法を使わなくても，グラウンデット・セオリーや他の社会学的手法を用いることで対象についてよく知ることが出来る．だが，「干拓地」の様に 900 件を越える場合の手作業には限界があり研究者の主観を十分に排除することが困難である．この主観をなるべく排除して正確性や妥当性の高い結果を得ることがテキストマイニングの一つの目標でもあるといえる．しかしなおも，岸田（2003）は「情報検索分野における評価研究に比べて，文書クラスタリングの評価の歴史は浅く，標準化はまだ十分でない（すなわち研究者によって使用する指標が異なることが多い）.」と言及している．この議論については本研究範囲を逸脱するためにこれ以上は言及しないが，分析で得られた結果の妥当性・信頼性・正確性に関してもより追求していくことは必要である．

市村ほか（2001）や那須川（2001）のテキストマイニングの事例ではウェブ上のテキストデータや会話記録，医療分野や特許情報などの分野で研究や実践が行われている．また，情報処理や人工知能などの研究分野では理論的な研究も盛んである．それら既存研究と本研究が異なる点は論文タイトルを元データとして分析を行った点である．会話文や医療情報や特許情報と論文タイトルが異なる点は，ひとつのセンテンスが短く区切られており語の出現頻度や距離の差異が小さい傾向にある．また，あいまいな会話文とは違い論文タイトルは研究内容を端的に示すものである．それとは反対に，通常のタイトルには修飾語や形容詞などがほとんど含まれていない．そのため，結果の解釈には注意が必要である．つまり本研究での分析結果のほとんどが名詞句の列挙であるため，あえて具体的な内容に踏み込むような修飾や形容は控えたのである．同様に，本研究では，論文タイトル中に含まれる語の出現頻度や相関，類似度のみに注目をしているために研究の「質」には触れていない．分析対象としたデータ中には，キーワード，要旨，内容が閲覧できるものも多くあるが，タイトル以外を知ることが出来ない論文や文献も多くあった．テキストマイニングでは出現頻度や文章の長さといった因子が分析に大きく影響を与えるためにデータの誤差や差異を均一化すべく，確実に得ることができるタイトルのみに注目をしたためである．

しかしながら，干潟や干拓地に詳しい研究者や識者であれば単語の列挙だけを概観しただけでどういった研究分野であるか，またどういった研究内容か推測することも可能であると思われる．例えば「畑」という語の近くに「窒素」という語が布置されていれば，畑作の肥料や土壌，作付け等に関する研究であるかもしれないという推論である．さらに

その近くに「水」や「排水」という語が来れば土壌に浸透した肥料分や窒素分に関する研究かもしれないという推論である．テキストマイニングに推論やアブダクションをどのようにして利用するのかは本研究の将来の課題でもある．

他方,「干拓地」の分析結果では「有明海」と「諫早湾」は地理的な近似関係から単純に相似性や類似性が高いのではないかと着想できるが，本研究ではそれらの傾向は見られなかった．このような単純な疑問に答えるとともに今後は干拓地研究の動向に対して個別の潟や干拓地も分析して知見を蓄積すること，干拓地同士の研究動向の差異やその特徴を明らかにするとともに知識の活用方法についても考察を深めたい．そして，より分かりやすい研究の提示を目指してテキストマイニングにコーディングルールの追加を行うことや分析結果と分析データの比較や相互参照が肝要である．

謝　辞

本研究は，平成 25 年度河北潟研究奨励助成を受けて実施したものです．本研究の遂行にあたりまして協力頂いきました皆さまにこの場を借りて改めて御礼申し上げます．ありがとうございます．

引用文献

石川県．平成 24 年度公共用水域及び地下水の水質測定結果報告書．http://www.pref.ishikawa.lg.jp/mizukankyo/shiryo/koukyo/．2015 年 2 月閲覧．

市村由美・長谷川隆明・渡部勇・佐藤光弘．2001．テキストマイニング事例紹介．人工知能学会誌．16（2）：192-200.

大串龍一．2002.「河北潟の自然と文化」編纂事業の提案．河北潟総合研究．5：33-36.

太田和良幸．1996．科学研究費の採択状況に見る法学分野の研究動向分析．学術情報センター紀要．8：337-358.

加藤誠・佐藤典夫・小林文雄．1995．農業土木と軟弱地盤対策（その 7）河北潟粘土地盤の軟弱地盤対策．農業土木学会誌．63（3）：55-62.

河北潟湖沼研究所．2013．河北潟レッドデータブック．橋本確文堂.

岸田和明．2003．文書クラスタリングの技法文献レビュー．Library and information science．49：33-75.

沢野伸浩．1995．河北潟の現状と課題 , 環境技術．24（7）：387-391.

志津綾香・松田眞一．2011．クラスター分析におけるクラスター数自動決定法の比較．アカデミア情報理工学編 11：17-34.

趙敏廷・谷口敏代・原野かおり・松田実樹・谷川和昭．2013．『介護福祉学』誌にみる介護福祉学の研究傾向 : 論文タイトルを用いたテキストマイニングから．介護福祉学．20（2）：152-158.

那須川哲哉・諸橋正幸・長野徹．1999．テキストマイニング : 膨大な文書データの自動分析による知識発見．情報処理 40（4）：358-364.

那須川哲哉．2001．コールセンターにおけるテキストマイニング．人工知能学会誌．16（2）：219-225.

林靖人・中嶋聞多・中嶋聞多．2009．地域ブランド研究における研究領域構造の分析 : 論文書誌情報データベースを活用した定量分析の試み．人文科学論集　人間情報学科編．43：87-109.

樋口耕一．2004．テキスト型データの計量的分析．理論と方法．19（1）：101-115.

樋口耕一．2006．内容分析から計量テキスト分析へ - 継承と発展をめざして．大阪大学大学院人間科学研究科紀要．32：1-27.

北陸農政局．1986．河北潟干拓事業誌．北陸農政局出版.

北國新聞社編集局．1985．レポート河北潟干拓．北國新聞社.