Vol.2014-CH-102 No.5
2014/5/31

MIMA Search を用いた修士論文とシラバスのテキスト分析 「文化資源学の射程」研究プロジェクト報告

中村雄祐[†1]　美馬秀樹[†2]　増田勝也[†3]　鈴木親彦[†4]

人文学・社会学における学際研究とはどのようなものなのか，これからどのように発展させていくのか．我々は，東京大学大学院人文社会系研究科に 2000 年に設立された人文社会系の学際的研究・教育プログラム，「文化資源学研究専攻」を対象に研究領域の形成過程を分析している．具体的には，文化資源学研究専攻の授業シラバス（2013 年時点で約 600 件），および修士論文（2013 年時点で約 80 本）の「論文要旨」「参考文献一覧」を対象として，報告者の一人である美馬（東京大学知の構造化センター准教授）が中心となり開発したテキスト分析システム MIMA SEARCH を用いて解析し，中村（東京大学大学院人文社会系研究科准教授）・鈴木（同博士課程）がその解釈を行っている．今回は，シラバスと修士論文要旨についての分析・解釈を報告する．

Text Analysis of Master Theses and Syllabi with MIMA Search Scope of Cultural Resources Studies

YUSUKE NAKAMURA[†1]　HIDEKI MIMA[†2]
KATSUYA MASUDA[†3]　CHIKAHIKO SUZUKI[†4]

The Department of Cultural Resources Studies is a graduate program created at the Graduate School of Humanities and Sociology of the University of Tokyo in 2000. The objective of the department is a creation of interdisciplinary education/research program on the basis of the existing humanities and social sciences. Since its launch more than 600 lectures and seminars have been offered and about ninety students have been granted master degrees of Cultural Resources Studies. In this presentation we analyze the syllabi and summaries of master theses using MIMA Search.

1. 学際研究・教育のテキストマイニング

人文学・社会科学における学際研究とはどのようなものなのか, これからどのように発展させていくのか. 我々は, 東京大学大学院人文社会系研究科の学際研究・教育プログラム,「文化資源学研究専攻」を対象に研究領域の形成過程を分析している．具体的には，文化資源学研究専攻の授業シラバス（2013年時点で約600件），および執筆者の使用許諾を得た修士論文（2013年時点で76本）のうち「論文要旨」「参考文献一覧」を対象として，美馬が中心となり開発したテキスト分析システムMIMA (Mining Information for Management and Acquisition) SEARCHを増田が解析し，中村，鈴木が解釈を担当した．報告者のうち，中村は教員，鈴木は大学院生として同専攻に所属する当事者でもある．まず本研究「文化資源学の射程」で設定した課題の説明から始める．

1.1 文化資源学研究専攻の概要と本研究の課題

文化資源学研究専攻は，2000 年に東京大学大学院に既存の人文学・社会科学の基礎研究を踏まえた新しい学問領域として構想された研究・教育プログラムであり，同時期に文化資源学会も設立されている．文化資源学研究専攻の特徴は，人文学・社会科学に軸足を置いた学際性，社会連携，そして共同作業の重視である．そのための基本概念として「ある時代の社会と文化を知るための手がかりとなる貴重な資料の総体」を「文化資源」と呼んでいる[1].

長い蓄積を持つ学問分野にはテーマ，方法，形式の踏まえるべき伝統が存在するのに比べて，今のところ本専攻には明示的な伝統は希薄である．その理由として，一つには発足以来まだそれほど時間が経っていないことがあるが，あえて学問の体系化・細分化以前への原点回帰を重視する学際的領域として構想されていることも大きい[2]．また，一般に基礎研究が実践や応用と距離を取りつつ進められるのに対して，発足以来，社会人を積極的に受け入れ，実践を重視した授業も開講されてきた．さらに，小規模で院生間の共同作業を重視していることも教育プログラムとして重要な特徴である．2014 年度現在，入進学者の累積数は修士課程 139 名，博士課程 49 名，年平均はそれぞれ 9.3 名，3.8 名である．専攻は文化経営学，形態資料学，文字資料

†1 東京大学大学院人文社会系研究科
University of Tokyo, Graduate School of Humanities and Sociology
†2 東京大学大学院工学系研究科
University of Tokyo, School of Engineering
†3 東京大学知の構造化センター
University of Tokyo, Center for Knowledge Structuring
†4 東京大学大学院人文社会系研究科
University of Tokyo, Graduate School of Humanities and Sociology

学の三コースに分かれているが，全コース共通で開かれる学会発表形式のゼミ，修士・博士課程の一年生全員で外部に向けて文化資源学を発信するフォーラムの企画が必修科目となっており，専攻内の学生の交流も活発である．

以上の３つの特徴のゆえに文化資源学研究専攻は，一般に専門性，基礎研究，個人研究を重視する人文学・社会科学の中でユニークな存在であり，既存の専門分野で研鑽を積んできた教員にとって新たなフロンティアとなる．準拠すべき明確なモデルがない中で教員が開講してきた授業には実験的な試みも少なくない．だが，それゆえにこそ，学術論文の執筆においてはやはり独特の課題も持つことになる．とりわけ，新たに学術の世界に参入してきた人々が最初に取り組む本格的な学術論文である修士論文の場合，その挑戦はさらに根源的なものとなる．それぞれの研究において先行研究として参照すべき隣接領域の研究群はあるものの，それらの枠組みにそのまま準拠することはできない．かといってまだ踏襲すべき蓄積も少ない中で，学術論文として認められるテキストを作り上げることは容易なことではない．修士課程の学生の間では，自分の研究課題の追求と重ねて「文化資源学とは何か？」「文化資源学らしい論文とはいかなるものか？」という問いがつねにある．

修士論文という性格上，個々の論文の完成度は必ずしも高くないかもしれず，現実にも多くの論文は提出後，審査する教員や後輩院生以外に読まれることはあまりない．本研究でも個々の論文の内容は分析の対象としない．しかしながら，準拠すべき明確なモデルがない中で相互に刺激しあいつつ一人一人が文化資源学の名に値する研究を追求する試みは，それらを一つの集合的な営みとして捉えるならば，未来の文化資源学の展開，さらには社会連携を重視した学際研究の試みという点から見ても貴重な知見を与えてくれるはずである．

このような課題に対して，本研究プロジェクトでは，当事者の経験と客観的なデータ分析を有機的に統合するための方法として MIMA Search を援用したのである．

1.2 MIMA Search の概要

文書集合を対象とした分析のためには，その文書集合の全体像の把握を可能にすることが必要である．特に，単純な数値的集計のみではなく，文書の内容に基づいて文書間の関係性を抽出し明示することが重要となる．また膨大な文書集合全体の把握のためには，個々の文書を個別に扱った分析は困難であるため，クラスタリングなどを用いた一定の抽象化が必要となる．これらの関係性の抽出・抽象化を様々な視点・条件からリアルタイムで行えることにすることで，文書集合を対象とした詳細な分析を行うことが容易となる．

これらを実現するためのシステムとして MIMA Search[3] がある．MIMA Search は，用語抽出をはじめとした自然言語処理，テキストマイニング，可視化技術を統合した検索システムであり，東京大学授業カタログ や東京大学工学部シラバスの構造化システム として実用化されている．また，東京大学知の構造化センターにおける「『思想』の構造化プロジェクト」[4] においても，岩波書店の雑誌『思想』に対する，論文集合の俯瞰による全体の把握や新たな知識の発見を促す論文の構造化システムとして利用されている．

MIMA Search は大きく分けて 1)文書からの自動用語抽出 2)文書間の関連度および文書クラスタの生成 3)文書集合の可視化，の三要素からなる．まず前処理として対象文書のテキストから用語抽出エンジン TermEngine[5] により自動的に専門用語の抽出を行う．TermEngine では C-value 手法により，用語をその用語らしさを表すスコアとともに抽出する．具体的には，品詞パターンを用いて用語候補を抽出し，それら候補の出現頻度，長さ，用語候補間の部分文字列関係を基に用語スコアを計算する．TermEngine により抽出された用語および対象文書のテキスト，属性データを対象として，MIMA Search では以下の機能を提供可能である．

・キーワード指定や年代等の文書属性の指定による検索
・検索された文書間の関連度の計算（デフォルトでは用語スコアに基づき計算）
・上記により計算された関連度を基にした文書クラスタリング
・上記のクラスタリングの任意の抽象度での実行
・文書間の関連度，クラスタリングを用いたネットワーク表示による文書集合の可視化
・検索結果に対するクロス集計，グラフ表示による可視化

MIMA Search では一般の検索システムと同様に左上のテキストボックスにキーワードを入力し，検索を実行する．検索条件としては，単純なキーワードのみならず，「発行年が 2000 年から 2009 年」のように文書属性を用いた条件指定も可能である．検索結果は左側にリストでの表示，右側には文書をノードとした文書間の関連度に基づくネットワークが表示される．関連度が高い文書ノード間には線が結ばれ，特に関連度が高い部分文書集合はクラスタ化される．各クラスタではクラスタ内の文書中の用語からそのクラスタを代表する用語が自動的に抽出され，クラスタラベルとして付与される．文書ノードはダブルクリックすることで文書の詳細を見ることができる．また，下部ではクロス集計を行うことができ，現在表示されている検索結果について，集計の対象・ベースの属性を選択し集計することができる．対象・ベースの属性はネットワーク表示にも反映され，対象を特徴量として計算した関連度を基に，ベースをノードとしたネットワーク構造を表示する．ネットワーク表示，クロス集計表示はいずれかのみを表示することも可能である．また左下部にはファセット(絞り込み) 検索用のフィールドがあり，登録データの種々の属性を用いた絞り

込みが可能である．

2. データと解析結果

本論の目的である文化資源学の射程を考えるために，研究のインプットとアウトプットをそれぞれMIMA Searchによって解析し，文化資源学に所属する研究者の視点で解釈を行う．研究のインプットとしては研究専攻によって行われる教育，具体的には研究専攻の全講義内容が文字化されているシラバス，アウトプットとしては研究専攻開設以来学位認定されてきた修士論文を解析した．

今回，我々がMIMA Searchで分析したシラバスは約700あるが，修士論文は全部で76本（許諾を受けた本数），シラバスの1/10しかなく，マイニングに使われた要旨の語彙数は6,726である．当事者が直感的に全体像を把握しうるぎりぎりの大きさにして，テキストマイニングが本領を発揮しうるぎりぎりの小ささといえよう．冒頭で，専攻内のコースについて述べたが，十分なデータ数を確保するため専攻レベルでの解析を行っている．

文化資源学専攻の修士論文要旨は，所属する東京大学大学院人文社会系研究科によって以下に様に規定されている．

（1）論文とあわせて3部提出すること．
（2）日本語で，4,000 字以内とする．外国語の場合はそれに相当する長さ．印字する際，読みやすいよう行間は十分にとること．a

これは人文系の研究科では一般的な内容かもしれないが，学会や工学系の研究科で求められる要旨のように400字の前後の短い文章ではなく4,000字である点は重要である．修士論文の内容を要約している上，この文字数もMIMA Searchで分析するに適している．次項「考察」で詳しく述べる文化資源学研究室の特質である研究間の「斥力」の問題からも要旨を利用する妥当性は高い．本文より文字数の少ない要旨を利用した方が，研究に関する用語や固有名詞の出現頻度から考えて斥力の効果が抑えられると判断した．

修士課程に属する学生は開講された講義以外に，独自の調査や方法論研究，学会等に参加しての議論，そして指導教員による直接の指導など，多岐にわたる研究のインプットを活用することで，修士論文と言うアウトプットに至る．その意味では，シラバス情報はインプットの一部分にすぎない．しかしながら，修士論文提出には専攻の講義を受講して一定数の単位を取得すると言う条件が課されており，講義は全ての学生にとって共通のインプットということができる．

a 「平成２５年度 修士論文の提出について」（東京大学大学院人文社会系研究科）http://www.l.u-tokyo.ac.jp/student/master_thesis.html から抜粋．これは2013年時点の規定であるが，基本的に文化資源学専攻設立時より変更はされていない．

2.1 シラバスの解析

専攻設置以来の講義数は637に上るため，MIMA Searchでシラバス全件を一度に解析すると，かえって構造が読み取りにくくなってしまう．ここでは，3年ごとに区切ってMIMA Searchで構造を解析し，その変化から解釈を行っていく．3年間の講義数は，100件から150件の間に収まり，MIMA Searchで表示するのに適した条件となる．また3年間ごとと言う設定は，社会人学生を多く受け入れており，長期履修制度という特例を利用して3年をかけて修士論文を執筆する学生もいる専攻の状況とも合致している．

3年ごとまとまりで変化を見てった結果を解釈すると，以下のようになる．当初は図1で示されたように，かなりばらつきが多く，明確な構造を見出すことができない．しかし，図2で示した様に，2005年以降「文化政策」に関わるクラスタが拡大し，最新の状況まで大きなグループを形成し続けている．一方で，クラスタを代表する用語に「ミュージアム」「展覧会」と言う揺れはあるが，美術館・博物館における展示に関するクラスタも安定して形成されている．

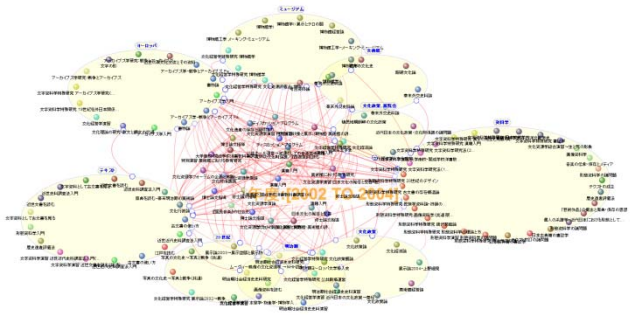

図 1 MIMA Search によるシラバス解析 対象：2002 年-2004 年度分

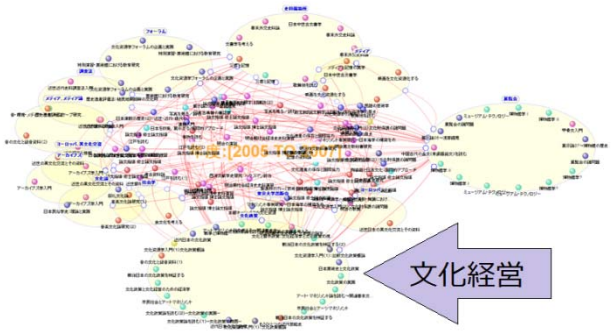

図 2 MIMA Search によるシラバス解析 対象：2005 年-2007 年度分

2.2 修士論文要旨の解析

修士論文の要旨全文をMIMA Searchに投入し，解析を行った結果が図3である．初期状態（MIMA Search機能上では「リンク強度」設定 0）でも一定の構造を読み取ることができるが，より構造を明確化するために，リンク強度を

0.5 に設定した図を提示した．ここでは，修士論文は大きく三つのグループに分かれている．ラベリングされたクラスタを代表する用語に基づき，これらのグループを本論では以降「文化政策クラスタ」「展覧会クラスタ」「その他のクラスタ」として呼ぶこととする．

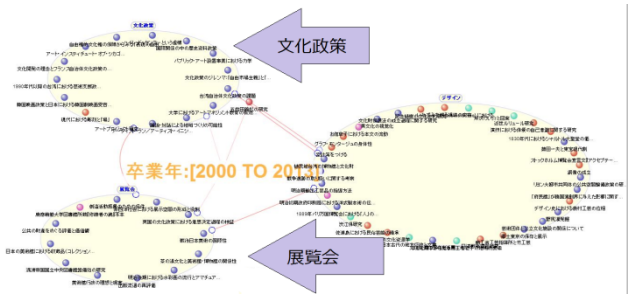

図 3 MIMA Search による修士論文要旨解析　対象：全件
リンク強度：0.5

特徴が明確に示されているのは，「文化政策クラスタ」と「展覧会クラスタ」である．「文化政策クラスタ」では，クラスタを構成する論文 16 本の内，15 本が文化経営学コースに属する学生によって書かれたものである．また「展覧会クラスタ」においても，12 本中 11 本が文化経営学コースに属する学生による修士論文である．最も多くの修士論文で構成される「その他のクラスタ」は，他二つのクラスタに属さなかった修士論文が，1 本を除いて全て集まって作られている．さらに MIMA Search の機能を使って，各クラスタを形作る用語に注目することで，「文化政策クラスタ」および「展覧会クラスタ」の状況がより明確になる．「文化政策クラスタ」でスコアが上位になっている用語は「文化政策」「アーティスト」「センター」，一方の「展覧会クラスタ」では「展覧会」「コレクション」が上位に来ている．

このことから次の構造を読み解くことができる．経営や展覧会などの実践的な研究テーマを持つ学生によって書かれた修士論文は，共通した語彙，知の構造の土台となる要素がすでに形成されている．また「文化政策クラスタ」と「展覧会クラスタ」に分かれている点には，それぞれの内容に近い研究分野を持つ教員の存在も反映されていると解釈出来る．

この二つのクラスタは，インプットであるシラバス解析で示した代表する用語が「文化政策」「ミュージアム・展覧会」であるクラスタと，内容面および教員の研究分野において対応している．このことからは，専攻としてのインプットとアウトプットが合致している状況を確認できる．

ただし，それは「インプットがそのままアウトプットに反映された」というような単純な展開ではない．最大クラスタを構成する修士論文は遺跡，演劇，工芸，雑誌，帳簿など多様なテーマを扱っているが，それらを集約する用語は容易には見つからず，ここではやむなく「その他」というラベルを付けている．つまり，「いかに運営するか，展示するか」については何らかのかたちが作られつつあるが，肝心の「何を」 ― 文化資源 ― については，今回の解析からは未だ明確な射程は見出しがたいということになる．

3. 考察—MIMA Search と文化資源学の視点

3.1 文化資源学の修士論文を生み出す斥力と引力

本研究では，文化資源学という新しい学術的な活動の射程を知るためこれまで書かれたシラバスや修士論文に注目しつつも，あえて個々のテキストの内容には踏み込まず，それらのテキスト群が作るかたちに注目するというアプローチを取ってきた．総じて MIMA Search の解析結果は，専攻に所属する当事者として中村と鈴木が予想していたこととおおむね重なっていたが，いったん当事者の主観から離れ計算機を駆使した解析結果を検討するという経験は修士論文を生み出す過程をより深く考えるための契機となった．

テキストマイニングの結果を踏まえて文化資源学研究専攻の状況を振り返ると，修士論文が生み出されるプロセスに関して改めて，学術研究一般，人文学社会科学一般，文化資源学，そして教育プログラムというレベルを異にするいくつもの特徴を再認識できる．

そもそも学術研究においては，オリジナリティの追求が重視される．剽窃は論外として，「安易な模倣や追従」と見なされる恐れがあるような研究は忌避される傾向にある．さらに，人文学・社会科学では個人研究が一般的で，大プロジェクトの一部を自分の研究として進めることは稀である．それに対して，文化資源学研究専攻では学際性，社会連携，共同作業を重視した教育プログラムが提供されており，これらの諸力の均衡としてアウトプットがかたちづくられていることになる．

MIMA Search の結果を踏まえてまとめるならば，まず，先輩や同期の研究から刺激を受けつつもそれらと重ならないように自分の研究を作っていく，いわば研究同士の「斥力」を高める傾向が存在する．そこに，既存の諸学の伝統に収まらないテーマを求める姿勢が加わった結果が「その他」クラスタの雑多な文化資源群ということになろう．そこでは具体的なテーマを構成する語彙のレベルでの重なりはそもそも低めに抑えられることになる．他方，やはりテキストマイニングの結果からは，この専攻のもう一つの基本姿勢である社会連携という志向に沿って，政策，経営，展示といった実践的な課題へ思考を凝集させる「引力」も存在していることがわかる．この二つの拮抗するベクトルが文化資源学の修士論文を生み出すエンジンなのであろう．

3.2 大学院での学びと修士論文

ただし，やはり MIMA Search の結果が示すように，各論文のレベルでも二つの問いがほどよいバランスで拮抗しているというわけではない．ここで我々が注意を向けるべき

は，個々の論文の内容ではなく（本研究がそのようなアプローチを取らないことはすでに述べた通りである），むしろ修士論文というテキストの持つ性格である．

冒頭に述べたとおり，現在，生み出される修士論文の大半は，提出の後，審査する教員や後輩以外に読まれることはあまりない．博士論文と違って公開の義務もなく，本研究も執筆者の許諾を得た修士論文のみを対象としている．修士課程で学ぶのはもっぱら学術論文という独特の様式や制約のある文書の読み方・書き方である．そして，あるアイデアが修士論文に結実する過程は，論文というフォーマットにうまく適合しない多くのアイデアをあきらめる過程と表裏一体である．それらの顕在化しなかった思考群，またそれらの思考の共有の帰結を知るには，今回とは別のアプローチを考える必要がある．

4. 結論と課題

シラバスと修士論文要旨のテキストマイニングを通じて，文化資源学研究専攻の最初の 10 年に修士課程に学んだ人々が文化資源学という新しい学術的な営みにどのような輪郭を与えてきたのかを知ることができた．我々は今後も専攻の修士論文要旨と参考文献一覧のデータを蓄積していく予定である．

他方，今回の研究は，修士課程での学習成果が修士論文という形にすべて顕在化するわけではないという，職業的研究者がつい忘れがちな事実にも目を開かせてくれた．修士論文に結実した学術論文という思考のフォーマットの制約と強さ，論文にはうまく適合しなかったが授業や議論から刺激を得て生まれた無数のアイデア，この双方が学際性と社会連携を指向する文化資源学の展開にとって重要なはずである．そのことは，本論で進めてきたシラバスや修士論文をデータとするアプローチの限界，さらに敷衍すれば大学の外にも視野を広げて文化資源学の射程を捉えることの重要性を確認することにもつながる．

シラバスと修士論文要旨，また，今回は活用できなかった参考文献一覧のデータを，他領域のデータと組み合わせることによって，大学を超えた文化資源学の射程を捉えることが今後の課題である．2013 年に文化資源学会研究会で本研究の経過報告を行った際，フロアから「この研究は文化資源学の鏡のようなものだ」という意見が述べられた[6]．今後，文化資源学の研究がより蓄積されるにつれ，「鏡」の重要性はより高まることになると予想している．

参考文献

[1] 文化資源学会設立趣意書（2002 年 6 月 12 日採択）
http://www.l.u-tokyo.ac.jp/CR/acr/overview/shuisho.html
[2] 人が資源を口にする時，文化資源学，第一号, pp. 1-6.
[3] Hideki Mima, “MIMA Search: Extracting and Visualizing Relationships among Courses using Natural Language Processing “, In Proceedings of OCWC Conference 2008, pp.42-50, Dalian, China, 2008.
[4] 美馬秀樹，丹治信，増田勝也，太田晋．近代文献のデジタルアーカイブ化とテキストマイニング-岩波書店「思想」を題材に．情報処理学会研究報告．人文科学とコンピュータ研究会報告, Vol. 2012, No. 4, pp. 1‐8, 2012.
[5] Hideki Mima and Sophia Ananiadou. An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. Terminology, Vol. 6, No. 2, pp. 175–194, 2001.
[6] 「文化資源学の射程　－　人文情報学のアプローチによる分析」（文化資源学会第 24 回研究会　2013 年 10 月 12 日，東京大学本郷キャンパス，中村雄祐・鈴木親彦共同発表）
https://sites.google.com/site/bunteku2013/home/others/03

謝辞

「文化資源学の射程　－　人文情報学のアプローチによる分析」は科学研究費助成事業「国際連携による仏教学術知識基盤の形成―次世代人文学のモデル構築」（代表者：東京大学大学院人文社会系研究科教授下田正弘，研究課題番号：22242002）の助成を受けています．

本研究に利用したデータベースは，東京大学大学院情報理工学研究科創造情報学専攻の稲葉研究室と共同で開発しました．