様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　３４３１５
研究種目：　挑戦的萌芽研究
研究期間：　2013～2016
課題番号：　２５５４０１５１

研究課題名（和文）機関リポジトリを活用した潜在的研究クラスタの創出

研究課題名（英文）Constructing Potential Research Clusters Using an Institutional Repository

研究代表者
田中　省作（TANAKA, Shosaku）

立命館大学・文学部・教授

研究者番号：００３２５５４９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究は，自組織の研究者が執筆した論文などの著作物を電子的に蓄積・公開しているデータベース（機関リポジトリ）とLDAベースのトピック分析（オーサトピックモデル）を活用し，研究者間の潜在的な関係を明らかにする方法を提案した．本手法は，まず各研究者の論文の本文まで活用し，各研究者の主たるトピックを推定する．そして，1．研究者の主副トピックに基づき研究者間を関係づける，2．論文内のトピックを媒介して関係づける．実際に，九州大学に対して研究者マップを描き，その有効性を確認した．

研究成果の概要（英文）：This research proposed some methods of constructing potential research clusters using an institutional repository, which is an online archiving system for publications authored by members of an institution, and a type of topic model (the author topic model) based on latent Dirichlet allocation. These methods estimate the research topics of authors using whole sentences in their papers. Then, to construct research clusters, they link these authors on the basis of their main topics or subtopics or via the other papers as pivot topics. Some of the experiments conducted in this research were successful in obtaining networks of researchers using the institutional repository of Kyushu University as a practical case.

研究分野：　情報学

キーワード：　機関リポジトリ　トピックモデル　研究者ネットワーク　研究情報データベース

１．研究開始当初の背景

科学の急速な進歩と総合・学際化に伴い，近年の科学技術基本計画にもあるように，伝統的な学術分野を確実に深めつつも，分野・部局を超えた研究者らの連携が求められつつある．実際，大学内の所属部局が異なっていても，近接した対象を異なる視点から取り扱う研究者は少なくなく，学内の潜在的・超域的な研究者間の関係（研究クラスタ）を顕在化させることは重要な課題の一つである．

このような課題に，近年，多くの研究機関で整備されている所属研究者の研究情報等のデータベース（以後「研究者 DB」と略記する）を活用することがある．しかし，研究者 DB には，クラスタ構成の照合法をはじめ，次のような問題がある．

問題 1．登録情報の浅さと緩やかな定型性

登録されている研究情報が自由記述を除き，キーワードや論文題目・学会名程度の浅い表層情報である．また，研究者の専門分野などの各種情報が，従来の分野分類に基づいていることが多い．

問題 2．研究者への負荷

自由記述の充実が前項の解消に寄与する可能性があるものの，研究者自身の自主的努力が求められる上に，情報の粒度がまちまちとなる懸念がある．

問題 3．単純かつ良くない意味でのリジッドな照合

研究者間の関係性を探るのに関心領域や論文題目等での単に共通のキーワードや部分文字列の有無を頼りにするような照合では，たとえば「言語処理」と「図書館」といった潜在的な連携可能性を見逃してしまう．

２．研究の目的

目的 1．機関リポジトリに基づいた潜在的研究クラスタ創出法の確立と実践

本研究は，自組織内での研究者の潜在的な研究クラスタの創出を目的とし，その方法論の確立と実践を行う．その過程で，前節で述べたような課題を念頭に，「機関リポジトリ」（Institutional Repository; 以後，適宜“IR”と略記する）を導入した方法を検討する．IR は研究機関などが自組織の研究者が執筆した論文・記事などの著作物を電子的に蓄積・公開しているデータベースで，近年，多くの研究機関で整備されている．

目的 2．機関リポジトリの新たな意義の示唆

IR の良い意味での本来的ではない活用を提案し，単なる電子アーカイブとしての役割とは別の意義と有効性を示す．

３．研究の方法

本研究では，次のような流れで各種研究クラスタの構成を図った．

手順 1．研究者のトピックの同定

機関リポジトリの学術論文を得，トピックモデルに基づいた分析を行い，著者（研究者）らのトピックを同定する．

手順 2．研究者の関係付け

手順 1 で求めたトピックと目的に応じて，研究者同士を関係付ける．本課題では，研究者の間で直接共有されないトピックを論文で媒介して関係付ける方法と，各研究者のトピックを主副に分けた上で関係付ける方法を試行した．

(1) 研究者の情報資源としての機関リポジトリ

機関リポジトリは，近年，多くの研究機関で整備されているデータベースである．その機関の主に研究者らの各種著作物がアーカイブされており，その大半を占めるのが学術論文である．

論文には，その著者たる研究者の関心事・研究テーマに関わる表現が多く含まれる．そのなかには，研究者の実施した事柄のほかに，自身が直接扱うことはできないかもしれない，しかし研究推進上重要な分野や課題に関する記述も含まれることがある．たとえば，言語学者や情報学者らにとっての「コーパス（大規模な電子化用例集）構築」は，「どのような言語資料を対象とするか」「どのように集めてくるか」「どのような形式で電子化するか」といったことが主な論述となるであろう．一方，著作権などのコーパス構築に関する社会的事項については直接的な対処はなくとも，課題として記述される可能性もあり，知的財産を扱う法学者らとの協働が想起し得る．このように，各研究者の論文に含まれる記述を丁寧に，そして適当なレベルで抽象化できれば，研究者らの潜在的な研究ニーズや関係性を見出せる可能性がある．そこで，本課題では，IR 内にアーカイブされている各研究者の論文をトピックモデルとよばれる方法で分析し，研究者に関連ある記述を抽象化して抽出することを考える．

(2) オーサトピックモデル

トピックモデルは，文書の生成を複数の「トピック」の混合分布の結果として考える確率的言語モデルである．ここでいうトピックは潜在トピック(latent topic)ともよばれ，適当な語彙の確率分布として与えられるものである．この潜在トピックは，ある程度まとまりがある話題や事柄・出来事という意味での，一般的な「トピック」とは異なり，直感的にわかりやすいまとまりとなるとは限らないことに留意する必要がある．

代表的なトピックモデルに，Latent Dirichlet Allocation (潜在的ディリクレ配分法; LDA)(Blei, et al., 2003; Griffiths, et al., 2004)がある．このモデルでは，文書をその文書で使用されている語の多重集合と捉え，Bag of Words(BOW)表現化し，その発生過程を与える．

本課題では，論文が文書に相当し，論文集合から論文の各著者（研究者）のトピックを求めたい．そこで，上記のトピックモデルを拡張したオーサトピックモデル(Rosen-Zvi et al., 2004)を採用した．オーサトピックモデルでは，著者（オーサ）ごとにトピック分布が規定され，文書はその著者らのトピック分布の混合として産出される，と考える．論文 d の著者集合 A_d としたとき，n 語で構成される d は，語 w とその著者 $a(\in A_d)$ の組(w,a)の列$<(w_1,a_1), (w_2,a_2), ..., (w_n, a_n)>$と捉え，その発生確率は次のように与えられる．

$$P(d) = \frac{1}{|A_d|}\prod_{i=1}^{n} P(w_i \mid z_i)P(z_i \mid a_i)$$

適当なトピック数 K と，著者情報付きの論文集合を与えることで，語彙分布としてのトピック $P(w \mid z)$と著者のトピック分布 $P(z \mid a)$ が推定される．トピックがどういったことを表すものなのかは，高確率の語から人が考えなければならない．著者のトピック分布も同様で，高確率のトピックから著者が関連している事項を読み取る必要がある．このような解釈過程は，トピックモデルの課題の一つであり，得られるトピックの解釈過程まで見越したようなモデルの拡張も検討されている．

(3) 研究者ネットワークの構成

①研究者の主副トピックに基づいた研究者間の関係づけ

当該機関の IR とオーサトピックモデルに基づきトピックと各研究者のトピック分布が得られる．なお，IR に論文登録のない研究者の情報は当然得られない．トピックの全体集合を Z としたとき，研究者 a のトピック分布から a に特徴的なトピック集合を $T(a)(\subset Z)$ とする．$T(a)$をトピックの発生確率に基づき，さらに 2 つに分ける．一つは a のトピック分布において最もよく出るトピック集合 $T_M(a)$ で，a にとって最も関わりの深いトピックである．この $T_M(a)$を a の主トピック群とよぶ．また，$T(a)$の $T_M(a)$以外のトピック群を a の副トピック群とよび，$T_S(a)$と記す．

研究では多分野の協働が求められることも多い．たとえば，近年の言語研究の一領域では，大量の言語データをコンピュータで分析し，言語学的知見の発見を試みるものがある．このような研究には，少なくとも言語そのものに対する深い理解をもった研究者 a_1 と，言語データを処理することを専門とした研究者 a_2 が求められる．当然，a_1 の主トピックは伝統的な言語学に類したもので，副トピックに言語処理のようなものが存在することが期待される．a_2 はその逆で，主トピックが言語処理，副トピックが言語学である．つまり，現在の伝統的な学問分野体系・研究分野からみると，学際的・総合的研究においては同じ主トピック同士の研究者ではなく，トピックが主副の関係で結びつくような研究者によるクラスタが望ましい．また，自組織で主トピックを共有する研究者を知り得る可能性に比して，自身の副トピックを主トピックとしている研究者や，その逆で自身の主トピックを副トピックとしている研究者を知り得る可能性は低い．そこで，本課題の一つでは，主トピックと副トピックを共有する研究者間を関係付けたようなクラスタ構成を考える．

②第三の論文に基づいた研究者間の関係づけ

共有するトピックがない研究者間でも，両者を媒介するような研究テーマの存在によって研究クラスタを構成し得ることもある．たとえば，両研究者が協働するような研究領域が自覚的でないような場合である．

本課題では，トピックを共有しない研究者間を，論文（第三の論文）を媒介させ，関係づける方法を検討した．たとえば，$T(a_1)\cap T(a_2)=\phi$ である研究者 a_1 のトピック z_1，研究者 a_2 のトピック z_2 を同時に含むような論文 p が存在する場合，a_1 と a_2 を関係づける．ここで，p は a_1,a_2 にとって自身がどういった観点で結びつけられたかを示唆する情報でもあり，p を根拠論文と呼ぶ．

参考文献

Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, pp.993-1022, 2003 年.

Griffiths, T. L. and Steyvers, M.: Finding Scientific Topics, Proceedings of the National Academy of Sciences, 101 Suppl 1, pp.5228-5235, 2004 年.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P.: The Author-Topic Model for Authors and Documentsm Proceedings of Conference on Uncertainty in Artificial Intelligence, UAI, pp.487-494, 2004 年.

４．研究成果

前節で述べたようなアプローチを，次のような形で実装し，実験を行なった．

(1) 対象データ

九州大学を対象機関とし，2014 年 5 月時点の九州大学学術情報リポジトリ（QIR）の日本語の論文 12,593 編を分析した．メタ情報から各論文の著者名や所属部局を，論文の

PDF データをテキスト化し，論文本文を得た．著者の延べ数は 24,647，著者の異なり数は 4,582 で，平均の共著者数は 1.96 であった．同時期の九州大学研究者情報（研究者 DB）によると 2,672 名で，上記論文データの著者に含まれるのは 555 名であった．この 555 名を研究クラスタ構成の対象とした．

オーサトピックモデルを適用するにあたり，文書相当の BOW 表現における“Word”の定義域（語彙空間）を日本語 Wikipedia エントリとした．一般語彙はそもそも研究者の特徴付けにあまり寄与しないことが予想され，その上，一般語彙で語彙空間を構成すると，たとえば「範疇文法」が「範疇」「文法」といった具合に過度に分解されてしまう傾向があるからである．そこで，論文の本文データは，日本語 Wikipedia エントリを辞書に追加した MeCab (http://taku910.github.io/mecab/) を用いていったん形態素解析し，その後，認定された形態素で Wikipedia エントリであるものを計数し，BOW 表現化した．

(2) 主副トピックに基づいた研究者ネットワーク

①主副トピックの認定

オーサトピックモデルに基づいた分析後の，著者（研究者）a に対する主トピック・副トピックの認定法について述べる．次項目(3)で述べる実験とは研究者のトピックの認定法が異なることに留意されたい．

全トピックを a のトピック分布で発生確率 $P(z \mid a)$の高い順に並び替えたものを $z_1, z_2, \ldots, z_K$ とする．つまり，$P(z_i \mid a) \geqq P(z_{i+1} \mid a)$ $(1 \leqq i \leqq K\text{-}1)$ が成り立つ．このとき，最も高い確率となる z_1 を主トピックとする．同値のトピックが複数存在することもあり，次のように主トピック群は与えられる．

$$T_M(a) = \{z : P(z \mid a) = P(z_1 \mid a)\}$$

副トピック群は，適当な閾値 $\beta \in (0,1)$の下，次のように与える．

$$T_S(a) = \{z_i : P(z_i \mid a) > \beta P(z_1 \mid a) \wedge z_i \notin T_M(a)\}$$

つまり，主トピックの発生確率の β 倍した確率より大きなトピックを副トピックと考える．$T(a)$をこれらの $T_M(a)$と $T_S(a)$の和集合で与える．

表 1 は，プラズマ核融合と研究戦略に従事していた研究者のトピック分布(K=1,000)で，$P(z \mid a)$の上位 5 位まで示したものである．ここで，1 位と 5 位のトピックは，それぞれの高確率語から，なにかしらの具体的な研究テーマというよりは，論文構成の際に必ず使用される表現に対応するトピックのように推測される．そこで，多くの研究者の $T(a)$に含まれるトピックは関係付けに利用しないこととした．本実験では，全研究者のうち 3 割を超える研究者の $T(a)$に含まれるトピックをそのように扱った．表 1 の例では，1 位と 5 位のトピックが対象から外れ，β=0.1 のとき，当該研究者の主トピックは「計測，遷移，放電」，副トピックは「研究，政策，制度」「温度，触媒，測定」の 2 つとなる．上記条件で，トピック数を 1,000 と 1,500 とした場合で，555 名の九大教員の $T(a)$の平均数はそれぞれ 2.7 と 2.6 であった．

表 1

順位	$P(z \mid a)$	高確率の語
1	0.573	結果，研究，実験
2	0.289	計測，遷移，放電
3	0.036	研究，政策，制度
4	0.013	温度，触媒，測定
5	0.006	問題，関係，意味

本手法では，トピック z を主トピックとしている研究者と副トピックとしている研究者を関係づける．さらにそのような研究者らが異なる部局に所属していることを条件として課した．研究者 a の所属部局を $f(a)$としたとき，次のような研究者対が関係づけられることとなる．

$$R = \{(a_1, a_2) : \exists z[z \in T_M(a_1) \wedge z \in T_S(a_2) \wedge f(a_1) \neq f(a_2)]\}$$

実験の結果，K=1,000 のとき$|R|$=4,934，K=1,500 では$|R|$=3,794 だった．

K=1,000 の場合を Cytoscape (http://www.cytoscape.org/)で可視化したものが図 1 で，ピンクが研究者，グレイの線が主副トピックによる関係付けである．

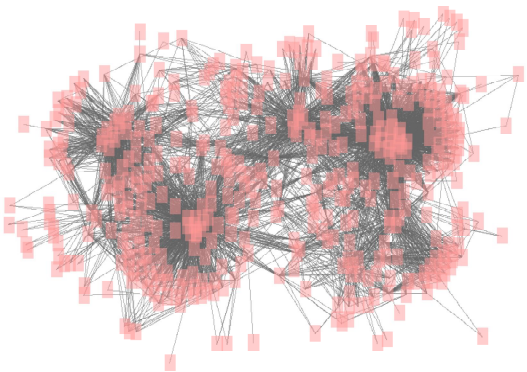

図 1

図 1 から，上述したプラズマ核融合・研究戦略の研究者とつながる研究者群（クラスタ）を抽出したものが図 2 である．このクラスタには当該教員も含め 47 名の教員が含まれ，主トピックの異なり数が 27 となる．

結果の例をもう一つ挙げておく．発見科学や図書館情報学の講座の主宰研究者を中心としたクラスタである．当該講座からは過去，多数の研究者が輩出され，現在，九大のさま

ざまな部局に所属し，新しい研究を展開し，協働している．こういった研究者もきちんとつながったクラスタとなっていた．

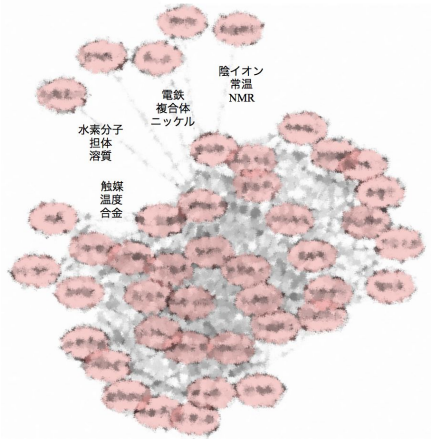

図 2

以上の一つ一つのつながりは必ず有用なものとは限らない．しかし，当該研究者らにとっては，そのつながりの妥当性や協働の可能性を判別することはさほど難しくなく，最後の例のようなものも確認でき，新しい研究クラスタを効率的に試作可能である．

(3) 第三の論文に基づいた研究者ネットワーク

このアプローチでは，研究者 a_1,a_2 のトピック分布で，それぞれに高確率のトピック z_1,z_2 に強く関わる根拠論文を検索する問題に帰着される．そこで，z_1,z_2 に表すような検索クエリを構成する必要がある．トピック（語彙分布）で高確率の語（本研究では Wikipedia エントリ）には，どのトピックにでも出てくるような語が多数含まれる．たとえば，表 2 は「流体の可視化」に関するトピックの語彙分布であるが，「計測」「速度」といった他のトピックでも頻出する基礎的な語も高確率である．そこで，特定のトピックで出現する語に絞り込むため，次のようなエントロピーを活用する．

$$H(w) = -\sum_{z \in Z} P(z \mid w) \log P(z \mid w)$$

この $H(w)$が一定値以下のものを検索クエリに採用する．たとえば，$H(w) \leqq 3$ の $P(z \mid a)$の上位 10 語は，「乱流」「羽根」「画像」「LDA」「可視化」「シーディング」「マトリックス」「輝度」「気流」「HW」となる．

表 2

順位	w	$P(z \mid a)$	$H(w)$
1	計測	0.030	3.67
2	乱流	0.027	2.30
3	流れ	0.026	3.73
4	速度	0.017	3.79
5	羽根	0.016	1.09

$H(w)$で絞り込んだのちに，$P(z \mid a)$の上位 m 位から n 語の組み合わせで検索クエリを構成する．このようなクエリの集合を $Q_{m,n}(z)$で表す．上述の例では「乱流　羽根」や「乱流　画像」といった具合に 45 個のクエリが生成されることになる．

q_i を a_i のトピック分布を元に生成されたクエリ，$p(q)$はクエリ q としたときの論文データベースにおける検索結果（論文の集合）としたとき，研究者 a_1,a_2 の協働可能性を，次のように判定する．

$$\begin{aligned}&\exists z_1, z_2, q_1, q_2 \\ &\quad [P(z_1 \mid a_1) \geq T \wedge H(z_1) \leq U \wedge \\ &\qquad P(z_2 \mid a_2) \geq T \wedge H(z_2) \leq U \wedge \\ &\qquad q1 \in Q_{m,n}(z_1) \wedge q_2 \in Q_{m,n}(z_2) \wedge \\ &\qquad p(q_1) \cap p(q_2) \neq \emptyset]\end{aligned}$$

ここで，$p(q_1) \cap p(q_2)$が根拠論文の集合となる．

実験では，K=600，m=10，n=2，T=0.2，U=3 という条件で，論文データベースとして CiNii (http://ci.nii.ac.jp/)を API を通して活用した．

①実験 1：分野を限定しなかった場合

(2)と同様の 555 名から，分野を問わずランダムに 30 名を抽出し，本手法を適用した．その結果，145 組が上記条件を満たした．実際にそれらを研究者情報に記載された各研究者の活動状況と照合しつつ，具体的な研究テーマが想起され協働可能性が高いと判断されたものは 4 組だった．

②実験 2：分野を絞った場合

同条件で「認知科学」「情報学」「脳神経科学」に分野を予め限定し，そこから研究者 10 名をランダムに抽出し，本手法を適用した．その結果，3 組が上記条件を満たした．そのうち 2 組は実際に協働可能性が高いと判断された．表 3 にその一例を示す．このとき，根拠論文として「動画像によるオンライン署名認証：～」が挙げられた（クエリは「認証 Web　画像　カメラ」）．

表 3

	研究者 1(a_1)	研究者 2(a_2)
専門分野・活動概要	情報科学，分散システム，情報検索，Web サービス，電子認証，コンテンツ検索，Web マイニング，図書館の電子サービス	コンピュータビジョン，画像処理，並列処理システム，マルチメディア研究
トピックの高確率語	システム，認証，サーバー，コミュニティ，情報サービス，Web	画像，カメラ，システム，画素

実験１の精度は必ずしも高くはない．トピックの粒度（K）の問題もあるが，そもそもトピックを共有していない研究者間の協働可能性という観点から考えると，この精度が絶対的に低いかどうかには議論の余地がある．

実験１では多くの組で，各研究者の活動情報を照合するだけで実際の協働可能性が判断される．当該の研究者らにとってはさらに迅速にその判断が下せるものと考え得る．一方，実験２の表３のように一見すると関連しにくい研究者らが検出されることもある．本手法では，根拠論文も併せて提示されるため，どういう方向性の協働が可能か，といったことも示唆され，効率的に解釈することが可能で，その有効性が確認された．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計４件）
①Shirai, T., Tomiura, Y., Tanaka, S., Ono, R.: Mining Latent Research Groups within Institutions Using an Author-Topic Model, Lecture Notes in Computer Science, 9469, pp.318-319, 2015年，査読有．

②宮崎佳典，田中省作，才茂真暉：論文英語要旨に基づいた機関別学術語彙リスト生成プログラムの開発，電子情報通信学会技術研究報告，114(228)，pp.11-16，2014 年，査読無．

③田中省作，冨浦洋一，徳見道夫：機関リポジトリから得られる著者の語彙分布に基づいた部局別重要語彙の選定，じんもんこん2014，2014(3)，pp.207-212，2014年，査読有．

④田中省作：ジニ係数に基づいたランダムフォレストにおける部分木の重要度，統計数理研究所共同研究リポート，321，pp.15-27，2013年，査読無．

〔学会発表〕（計７件）
①田中省作：学術情報マイニング，第４回九州大学異分野融合テキストマイニング研究会シンポジウム，2016年１月30日，九州大学（福岡県福岡市），招待講演．

②田中省作，冨浦洋一,上瀧恵里子：機関リポジトリとトピック分析に基づいた研究者ネットワーク，RA 協議会第 2 回年次大会，2016年９月１日，AOSSA（福井県福井市）．

③小野龍太郎，冨浦洋一，田中省作，上瀧恵里子：オーサートピックモデルを用いた論文分析による潜在的研究グループの発掘，言語処理学会第20回年次大会，2014年３月19日，北海道大学（北海道札幌市）．

④Funatu, T., Tomiura, Y., Ishita, E., Furusawa, K.: Extracting Representative Words of a Topic Determined by Latent Dirichlet Allocation, eKNOW2014 (the 6th International Conference on Information, Process, and Knowledge Mangement), 2014年３月23日，Novotel Barcelona Sant Joan Despi (Barcelona, Spain).

⑤田中省作：分類型ランダムフォレストにおける部分木の重要度，言語研究と統計2014，2014年３月29日，統計数理研究所(東京都立川市）．

⑥田中省作，冨浦洋一，宮崎佳典，徳見道夫：機関リポジトリの言語資源としての活用：大学毎の部局別英語重要語彙の選定，第 62 回日本図書館情報学会研究大会，2014年 11 月 30 日，梅花女子大学（大阪府茨木市）．

⑦田中省作：タスク駆動型のコーパス構築と情報処理技術，英語コーパス学会第 40 回大会，2014年10月５日，熊本学園大学（熊本県熊本市），招待講演．

〔図書〕（計２件）
①石川有香・石川慎一郎・清水裕子・田畑智司・長加奈子・前田忠彦（編）田中省作他著：言語研究と量的アプローチ，金星堂，2016年，307(145-155)ページ．

②岸江信介，田畑智司（編）田中省作他著：テキストマイニングによる言語研究，ひつじ書房，2014年，212(137-151)ページ．

〔その他〕
ホームページ等
http://www.cl.ritsumei.ac.jp/~sho/

６．研究組織
(1) 研究代表者
田中　省作（TANAKA, Shosaku）
立命館大学・文学部・教授
研究者番号：00325549

(2) 研究分担者
冨浦　洋一（TOMIURA, Yoichi）
九州大学・大学院システム情報科学研究院・教授
研究者番号：10217523

上瀧　恵里子（JOTAKI, Eriko）
九州大学・男女共同参画推進室・教授
研究者番号：40211297