社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

文書ベクトルを用いた中国共産党のイデオロギーの分析

御器谷　裕樹†　　持橋　大地††

† 慶應義塾大学 法学研究科　〒108–8345 東京都港区三田 2–15–45
†† 統計数理研究所 数理・推論研究系　〒190–8562 東京都立川市緑町 10–3
E-mail: †yukimikiya@keio.jp, ††daichi@ism.ac.jp

あらまし　多数のテキスト間の意味上の差異を把握することは, 社会科学・人文科学の領域においても重要な意味を持つ. 本研究は中国共産党のイデオロギーの長期的な変化を把握するために, 中国語で書かれた共産党機関紙のテキストデータを対象に計量テキスト分析を行った. 具体的には文書ベクトル, *t*-SNE, Fisher の線形判別分析を用いた結果, これまで見落とされがちだったが意味を持つ, イデオロギーや政策上の差異をより明示的に判別することができた. さらに, 確率的潜在意味スケーリング (PLSS) を用いてイデオロギーや政策上の差異の妥当性を検証した. こうした手法は, イデオロギーなどの既知のラベル情報の先入観を排除した状態で, データ構造を連続的に理解することを可能にするため, 探索的なデータ分析において有用である.
キーワード　政治学研究, 中国共産党の政治, 埋め込み, DocVec, 線形判別分析, *t*-SNE, PLSS

An Exploratory Text Analysis for the Ideologies of the Chinese Communist Party

Yuki MIKIYA† and Daichi MOCHIHASHI††

† Political Science, Graduate School of Law, Keio University　2–15–45 Mita, Minato-ku, Tokyo, 108–8345 Japan
†† The Institute of Statistical Mathematics　10-3 Midori-cho, Tachikawa Tokyo 190–8562, Japan
E-mail: †yukimikiya@keio.jp, ††daichi@ism.ac.jp

Abstract　Understanding the semantic differences among texts has important implications in the social sciences and humanities. In order to understand the long-term changes in the ideology of the Communist Party of China, this study conducted quantitative text analyses on the official newspaper written in Chinese. Specifically, I explicitly extracted meaningful differences in the policies and the ideologies by applying DocVec, *t*-SNE, and Fisher's linear discriminant analysis. These methods are practical in exploratory data analysis because they allow us to continuously understand the data structure while removing label information such as ideology.
Key words　political Science, Politics of the Chinese Communist Party, Embedding, DocVec, Linear Discriminant Analysis, *t*-SNE, PLSS

1. はじめに

政治学研究方法論の分野において, テキストデータ（例えば新聞記事, 政党マニフェストなど）を用いて政治体制, 政治指導者, 有権者がどのような政治的選好を有しているか推論されている [1]. こうした研究の中でまず用いられる手法は, イデオロギー（保守, 革新など）や政党（民主党, 共和党など）などのラベルを用いた教師あり学習である [2] [3] [4] [5] . また, 言説の概要をトピックモデルとして知られる潜在的ディリクレ配分法（Latent Dirichlet Allocation; LDA）の発展形である構造的トピックモデル（Structural Topic Models）[6], 半教師あり学習や単語ベクトルを用いた手法で分類や計測を行っている研究もある [7] [8].

左記の研究においては, 政党やイデオロギーなど明確な規範性を有するラベルが用いられる場合があり, 多くの場合で政治学の先行研究に基づいて仮説が作られている. こうした背景から, 既存の政治学研究が分析対象としていないようなラベルや, 政党制度が明確ではないために, 政党ラベルが機能しにくい体制における探索的な研究が必要とされている.

本研究は民主主義体制のように公正で競争的な選挙制度がない中国[1] におけるイデオロギーの変化を分析する. 党国家体制である中国は, 中国共産党が政府を指導しているため, 本研究は中国共産党内

(注1)：中国には民主諸党派と呼ばれる組織はあるものの, いずれも中国共産党の指導下にあるため, 民主主義体制における政党とは形態が大きく異なる.

のイデオロギーを捕捉することを目的とする. そのために, 以下, 分散表現である文書ベクトル, 次元削減の手法である t-SNE, 線形クラスタ分析である Fisher の線形判別分析を用いた探索的データ解析を行う. これにより, 長期間にわたって変化するテキストデータの傾向を抽出することで, 中国共産党のイデオロギーの変化を把握する[2].

本研究が用いたデータは「伝統」という言葉に言及した, 中国共産党の機関紙である『人民日報』の 1974 年から 1994 年の 39286 記事である[3]. 「伝統」という言葉に言及した記事を選択した理由は, 中国共産党の言語表現の中で, この言葉が体制の正統性を主張する際に頻繁に使用される言葉のためである. こうしたテキストデータを対象に分析を行うことで, 時代によって中国共産党内部にどのようなイデオロギーの変化が潜在しているかを実証することが本研究の目的である.

2. 文書ベクトルの計算

本研究ではテキストデータを分析可能にするため, 以下の過程を用いて各記事をベクトル化した. 文書ベクトル [10] は単語の分散表現として知られる Word2Vec [11] を文章レベルに応用した手法である.

文書ベクトルは Doc2Vec [11] と数学的に等価であることがわかっており, この意味において両者は同様の計算を行っている. しかし、Doc2Vec はニューラルネットを使用して学習するモデルのため, 解が一定にならない. それに対して文書ベクトルは図 1 のように線形代数で計算されるモデルである. 文章 d に単語 w が登場する頻度を $n(d,w)$ とし, $N=\sum_{d,w} n(d,w)$ をその総和, $n(d)$ を d の単語数, $n(w)$ を w の頻度とすると, 非負の自己相互情報量

$$\mathrm{PPMI}\,(d,w) = \max\left(\log\frac{p(w,d)}{p(w)p(d)}, 0\right) \tag{1}$$

$$= \max\left(\log\frac{n(d,w)\cdot N}{n(d)n(w)}, 0\right) \tag{2}$$

を要素とする行列 Y を図 1 のように $Y = DW^T$ と特異値分解することで, 行列 D と W の各行として文書ベクトル $\vec{d}$ と単語ベクトル $\vec{w}$ が解析的に求められる. なお, ここで K とは圧縮する次元数を示す.

上記が示すように文書ベクトルは解析解を持つ完全に客観的なモデルであり, 再現性を担保する点で利点があるうえ, Doc2Vec より性能が良いことが実験的に示されている [10]. なお, 本研究は単語分割には, spaCy [12] を用いた.

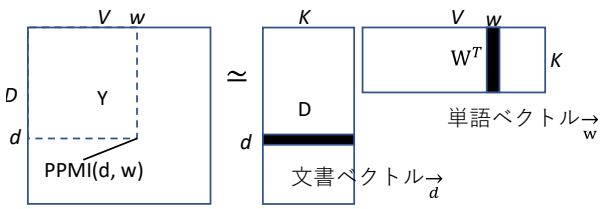
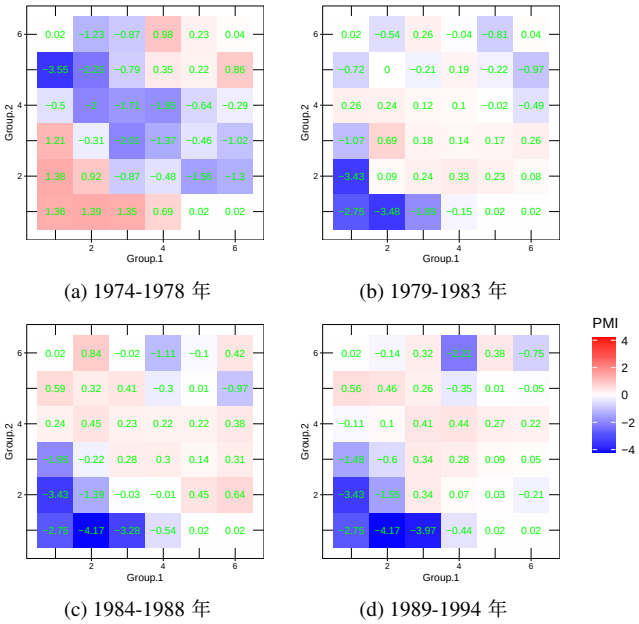


図 1: 文書ベクトルの仕組み. PPMI を要素とする文書–単語の共起行列 Y を SVD で $Y \simeq DW^T$ と分解することで、D の各行として文書ベクトル $\vec{d}$ が得られる. ここで V は語彙数, K は圧縮する次元数である.

(注2): ここでいうところのイデオロギーとは, 公に表明される公定イデオロギーであり, 内なる思想のことを示すものではない.

(注3): データは CDROM 版の『人民日報』データベース [9] を用いた.



(a) 1974-1978 年 (b) 1979-1983 年

(c) 1984-1988 年 (d) 1989-1994 年

図 2: 1974 年以降の記事の自己相互情報量 (PMI). 升目の色が赤ければ赤いほど, 記事の出現とその升目の場所との相関を表す PMI が高く, 青ければ低いことを表す.

3. t-SNE を用いた可視化と自己相互情報量

前章のデータを対象に, 次元削減の手法の一つである t-SNE(t-Distributed Stochastic Neighbor Embedding) [13] を用いて可視化を行った. t-SNE は, SNE の課題であったコスト関数の最適化と混雑問題に対応するために対称なコスト関数と Student-t 分布を内部に組み込んだモデルである.

t-SNE は観測ベクトルの類似度と, 可視化した際の座標の類似度を近くするように学習する[4]. 本研究は Python の scikit-learn [14] に実装されているモデルを利用した.

3.1 t-SNE を用いた 1974 年以降のデータの分析

本項では計算時間を短縮するために, 元々のデータセットである 39286 記事から, 1974 年以降の 8000 記事を無作為抽出したうえで実験を行った. 本研究は, 前節で文書ベクトルを用いて 50 次元に圧縮したテキストデータに対して, 図 3 では t-SNE を用いて次元削減を行った.

本項では各テキストデータの文脈を把握しやすいように単語も同時に埋め込んだ. 図 1 で得られた文書ベクトルと単語ベクトルは次元数 50 で共通のため, これらを連結し, 1 つの長い行列に格納することができる. この行列に対して t-SNE を実行することで, 同一空間に文書と単語を埋め込むことが可能になる. なお, 埋め込む単語の選択基準は, 分布する二次元空間を格子状に区切り, 各格子内で tf-idf 値が高い順に上位 2 語とした.

図 3 のように, 緑色で示した 1974 年の分布は, 図内中央左下や中

(注4): 上記の過程を経て計算されるため, t-SNE の実行結果は試行の度に異なる結果を示すことにも注意が必要である.

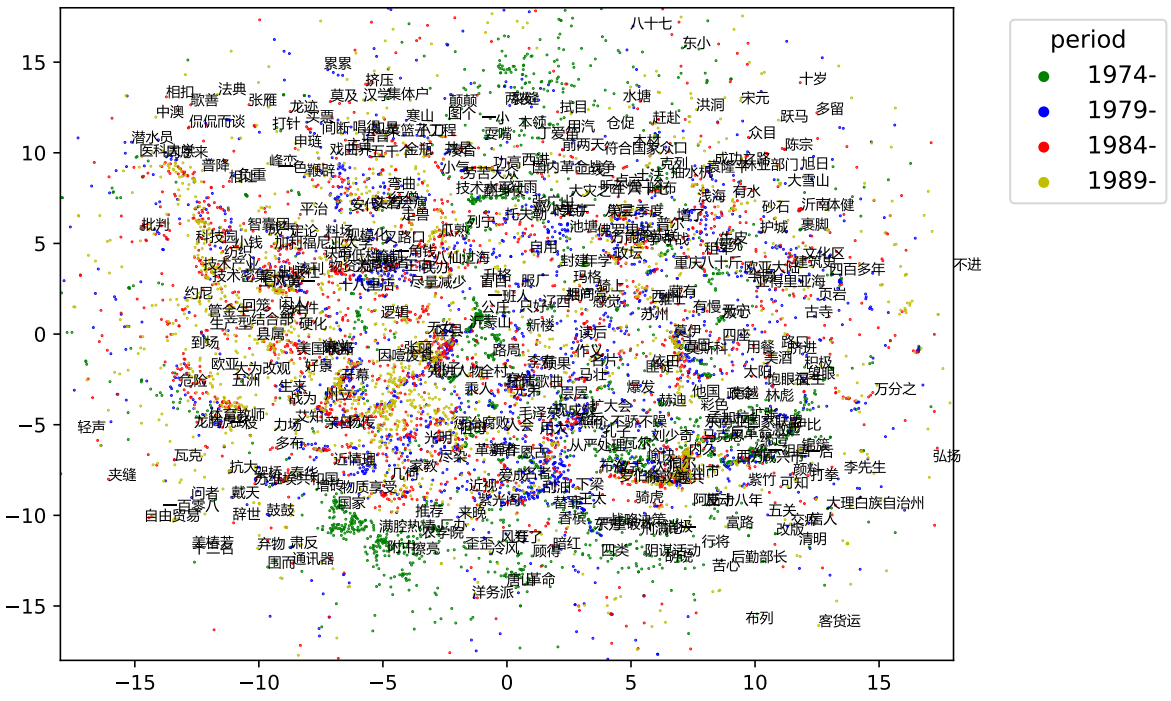


図 3: 1974 年以降の各 5 年間の記事の分布（単語も埋め込んだ場合の図）

央上部に集中していることがわかる．それに対して 1979 年以降のテキストデータの分布には，大きな差異が見られない．

3.2 *t*-SNE と自己相互情報量を用いた 1974 年以降のデータの分析

t-SNE による分析だけでは，各年代の記事がどのような文脈に集中していたかという分散情報をやや把握しにくい状態にあった．そこで追加的に，自己相互情報量（Pointwise Mutual Information, PMI）を算出することで，記事の文脈の変化を可視化する．図 3 を縦横 6 分割の升目に区切り，各セル c の中で，$p(a|c)$ を 1974 年から 1994 年までを 4 分割した各 5 年の特定の年代の記事 a がセルに現れる確率とすると，a と c の自己相互情報量 $\mathrm{PMI}(a, c)$ は以下のように表すことができる．

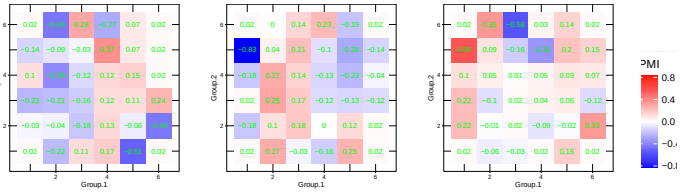
$$\mathrm{PMI}\,(a, c) = \log \frac{p(a, c)}{p(a)p(c)} = \log \frac{p(a|c)}{p(a)} \tag{3}$$

すなわち，PMI は記事 a が平均的な確率と比べて，セル c の中で何倍出現しやすいかの対数となっている．PMI が 0，すなわち対数の中が 1 で確率が同じの場合は a と c が独立であることを意味し，PMI が正の値をとる時は a と c が共起しやすく，負の値をとる時は a と c が共起しにくいことを意味する．

なお，共起頻度が 0 の場合に計算できなくなることを防ぐため，本研究では確率を算出する際にすべての頻度に最初から 1 を加算するラプラス平滑化を用いた．

前節で指摘した傾向は，自己相互情報量を算出するとさらに明確に認識できる．図 2 に示したように 1974 年以降の 5 年間のテキストデータは，他の年代と比較して，左下および上部により多く集中していることがわかる．この部分にあたる文脈は「革命」，「粛反」（反革命者を粛清する）などの単語情報から，急進的な階級闘争を伴う共産主義傾向が強い文脈であると推察される．

しかしながら，上記の分析では 1970 年代のテキストデータの特異性が殊更に際立つため，1980 年代における文脈の変化が把握しにくい状態にある．そこで次節では，1979 年以降にテキストデータを絞って分析を行った．



(a) 1979-1983 年　(b) 1984-1988 年　(c) 1989-1994 年

図 4: 1979 年以降の記事の自己相互情報量 (PMI). 升目の色が赤ければ赤いほど，記事の出現とその升目の場所との相関を表す PMI が高く，青ければ低いことを表す．

3.3 *t*-SNE を用いた 1979 年以降のデータの分析

以下では 1979 年以降のテキストデータを 6000 記事分無作為抽出し，文書ベクトルを計算した．そのテキストデータに対して，*t*-SNE を用いて次元削減を行ったものが付録の図 6 である．

付録図 6 の中央上部より，1979 年以降の 5 年間には国内の汚職摘発や綱紀粛正の観点から「違法乱紀」（法令違反や綱紀の乱れ），「各負其責」（各自が職責を全うする）などの言葉に近い文章が多く登場することがわかる．また，同時代には図中右下部の「革命」など 1976 年以前によく使われた政治的なレッテルも引き続き使われていることがわかる．1984 年以降は図の中央上部「租賃経営」（賃貸経営），「軍事法院」（軍法会議）など経済政策や法律に関する用語が多く観測される．これは，当時計画経済から実質的な市場経済（社会主義市場経済）へ徐々に移行を始めた段階にあって，様々な法整備が進められたためであると解釈できる．また，1989 年以降については，図上部左「法令」，「国民収入」，「資産」，「編制」など国内経済や政治体制に関する議論が多く登場することが明らかになった．

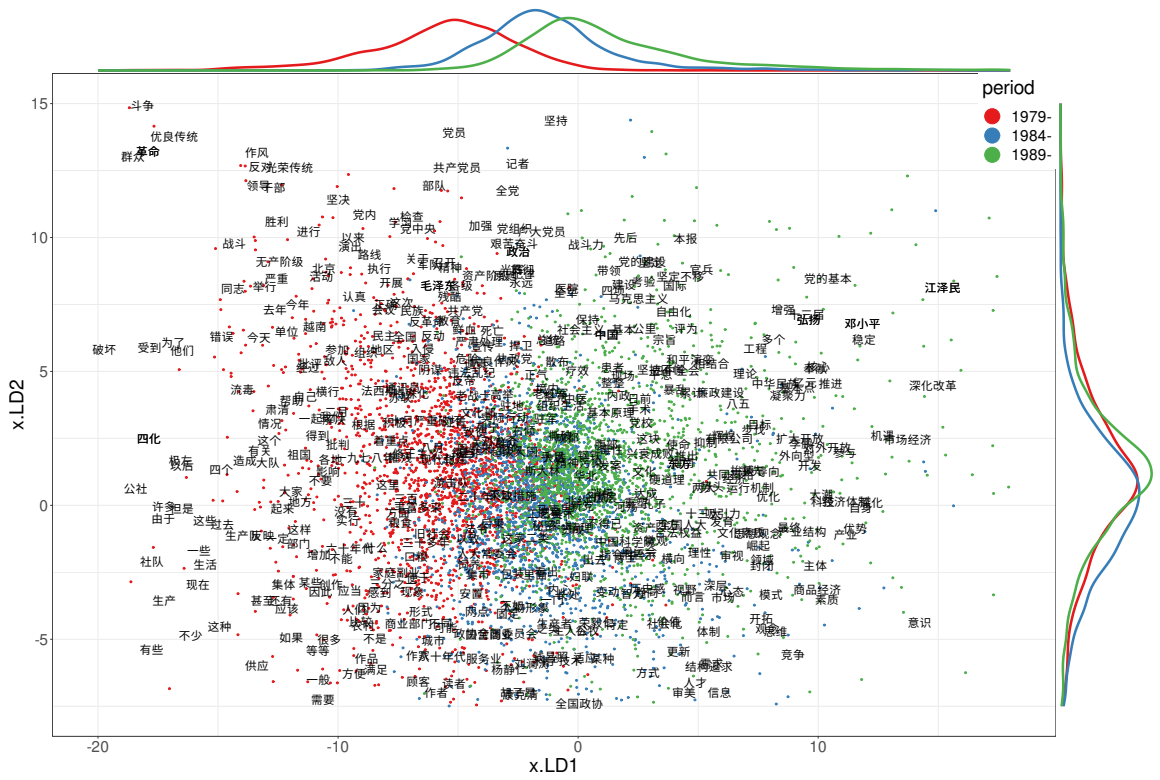


図 5: Fisher の線形判別分析を用いた記事の分析

次にこうした t-SNE を用いて可視化した分布を, 自己相互情報量を算出することでさらに分析した.

3.4 t-SNE と自己相互情報量を用いた 1979 年以降のデータの分析

対象とする記事を 1979 年から 1994 年までに絞り, 5 年ごとに 3 分割した際の自己相互情報量を図 4 に示す.

自己相互情報量の可視化によって, 前節で指摘した傾向はさらに明確になった. 図 4 が示すように, 時代を追うごとに, 徐々に中央上部で観察できる汚職摘発や綱紀粛正に関する言及は減少し, 中央左部で観察できる伝統芸能に関する言説は徐々に増加している. こうして経時的な変化を多様な形で把握することができた一方で, 各時代には重なりも多いため解釈が困難な箇所があった.

そこで, 次節では年代というラベル情報を明示的に与えることで, テキストデータの分布における変化をさらに明確に分析する.

4. Fisher の線形判別分析

本章では文書ベクトルを用いて文書ベクトル化したテキストデータに対して, さらに Fisher の線形判別分析を用いた分析を行った. なお, この分析は 50 次元に圧縮された文書ベクトルに対して直接 Fisher の線形判別分析を適用したものであり, 2 次元での可視化を経ていないという大きな利点がある.

Fisher の線形判別分析は予め付与されたクラスに関する情報（ここでは年代に関する分類情報）をもとに, 別々のクラスのデータをできるだけ分離し, 同じクラスのデータをできるだけ近接させるようにデータを射影する手法である [15].

いま, m を全データ集合の平均とし, m_k をクラス k に属するデータ（C_k）の平均とする. 特徴量を y とし, クラスごとのデータの平均の特徴量を μ_k と表すと、各クラス内の平均である総クラス内共分散行列 S_W, クラス間共分散行列 S_B はそれぞれ次のように表すことができる.

$$
\begin{aligned}
S_B &= \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T \\
S_W &= \sum_{k=1}^{K} \sum_{n \in C_k} (y_n - \mu_k)(y_n - \mu_k)^T
\end{aligned}
\tag{4}
$$

このとき, データを低次元に射影する行列を W とおくと, 線形判別分析では射影した空間においてクラス内の分散を最小化し, クラス間の分散を最大化するように, 次式を最小化する W を求める.

$$
J(W) = \mathrm{Tr}\left((W^T S_W W)^{-1} (W^T S_B W) \right) \tag{5}
$$

この計算は, S_W と S_B の逆行列を用いて, 一般化固有値問題に帰着することで線形射影行列 W を求めることができる.

図 5 は Fisher の線形判別分析を行った結果の上位 2 次元を, 周辺分布とともに示した図である. なお, 単語は格子状に縦横 60 個に分割した各区切りの中で, tf-idf により上位 2 単語をプロットしたものである.

図 5 上部の周辺分布に示したように, 図内左側から右側へ向けてた左右軸に時代の変化が観測できる. 1980 年代の中国政治は, 70 年代までの文化大革命という急進的な共産主義運動が終焉し, 鄧小平という政治指導者のもとで対外開放政策がとられていた時代であった. そのため, 左側では「革命」という言葉が登場する一方で, 図内右側では「改革開放」という対照的な概念が登場することがわかる.

同様に, 各時代の政治指導者の名前（「毛沢東」,「鄧小平」,「江沢民」など）が各時代に点描されている. こうした実験結果は, Fisher の線形判別分析の結果に妥当性があることを証明している.

また, 判別分析を用いたことでこれまで明示的ではなかった新たな発見もあった. それは党・政府組織の編成や内部統制に関する変化である. 具体的には, 図 5 中央左部の「粛清」（体制内外の反対派に対する弾圧）, 同中央中央部の「整党」（党組織や人員の整理）, 同中央右部の「廉政建設」（廉政[5]を行うための政策執行）という, 各時代において特徴的な言葉が浮かび上がった. こうした側面は, 政策論争や権力闘争に注目する政治学研究や政治史では見落とされがちだった, イデオロギーの（細かい文脈における）変化を示す例である.

さらに, 中央部の上下に目を転じると, 上部では「党員」「共産党員」「全党」「党中央」「党組織」など中国共産党の政治空間が観察できるのに対し, 下部では「全国政協」[6]「服務業」（サービス業）「家庭」「作家」「生産者」など党以外の政治, 経済, 生活空間を示していることが分かり, これが図の上下軸を構成している.

5. Fisher の線形判別分析と確率的潜在意味スケーリングを組み合わせた分析

Fisher の線形判別分析では上述の通り, 1. イデオロギーにおける明示的な変化（急進的な共産主義から対外開放へ）と潜在的な変化（党・政府組織の編成において粛清から廉政建設へ）, 2. 党とそれ以外の領域という 2 つの軸を観測することができた.

本章では前章の探索的な研究を更に発展させるため, これらの軸の情報をもとに言葉の極性を分析した. 本章で用いた手法は確率的潜在意味スケーリング [8] である. 同手法は項目反応理論とニューラル単語ベクトルに基づいて, テキストを連続的な空間に位置づけて測定する統計モデルである.

本研究は Fisher の線形判別分析で導出した対立軸を, 確率的潜在意味スケーリングの対立軸として採用することで, 結果の妥当性を検証する. 具体的には Fisher の線形判別分析で明らかとなった縦, 横それぞれの両極端な単語を, 確率的潜在意味スケーリングで分析する際の正例および負例として極性語辞書とした. それぞれ選択した 5 つの語句は, 上記の Fisher の線形判別分析で文書を単語とともに埋め込んだうえで, 格子状に分割した小区分の中で tf-idf 値が大きい単語の中で上下, 左右の端にある語句を選択した. また, その際に選択した単語を表 1 に示した. 分析結果を単語の極性とともに表 2 と表 3 に掲載する.

表 2, 表 3 が示すように, 前章で行った Fisher の線形判別分析に妥当性があることがわかる.「イデオロギー–対外開放」では外資導入を柱とする対外開放を行う構造改革を推進する政策が明らかに見て取れる.

それと対照的に,「イデオロギー–共産主義」では共産主義や革命に関係する概念や過去の戦争（日中戦争や朝鮮戦争など）での戦没者を烈士として救国のために戦った貢献を讃えている.

2 つ目の軸である「組織–党」では党という単語が付く概念が数多く登場している. また, それらの多くが党の組織や党員, 党員の規

表 1: PLSS で用いた単語の日本語訳.

分類–意味	単語
イデオロギー–対外開放	改革の深化, 江沢民, 市場経済, 鄧小平, 対外開放
イデオロギー–共産主義	公社, 毛沢東, プロレタリアート, 革命, 群衆
組織–党	党員, 共産党員, 全党, 党中央, 党組織
組織–党以外	全国政治協商会議, サービス業, 家庭, 作家, 生産者

表 2: PLSS による計算結果: イデオロギーに関する単語と極性. 正負各上位 20 語を抜粋した.

日本語 (原語)	極性	日本語 (原語)	極性
輸出入 (进出口)	0.583	革命 (革命)	−0.602
速める (加快)	0.569	戦士 (战士)	−0.589
構造転換 (结构调整)	0.567	革命者 (革命者)	−0.570
外資利用 (利用外资)	0.558	群衆 (群众)	−0.567
拡大し続ける (不断扩大)	0.554	軍人 (军人)	−0.551
情報産業 (信息产业)	0.540	功績 (事迹)	−0.550
対外貿易 (对外贸易)	0.538	英雄 (英雄)	−0.549
技術導入 (引进技术)	0.537	共産党員 (共产党员)	−0.548
来年 (明年)	0.534	烈士 (烈士)	−0.530
対外開放 (对外开放)	0.530	紅軍 (红军)	−0.528
緊縮 (紧缩)	0.524	救う (抢救)	−0.528
政府に回答 (批复)	0.523	犠牲となる (牺牲)	−0.525
速める (加速)	0.523	郭俊卿 (郭俊卿)	−0.516
宇宙戦闘 (星战)	0.517	愛国 (爱国)	−0.511
戦略防御 (战略防御)	0.507	革命闘争 (革命斗争)	−0.508
金融 (金融)	0.507	政府軍 (官兵)	−0.505
信用貸付 (信贷)	0.506	政治工作員 (政治工作者)	−0.502
技術開発区 (技术开发区)	0.504	深い感動 (可歌可泣)	−0.502
予測 (预测)	0.503	一兵 (一兵)	−0.502
制御する (控制)	0.500	言行 (言行)	−0.490

表 3: PLSS による計算結果: 組織に関する単語と極性. 正負各上位 20 語を抜粋した.

日本語 (原語)	極性	日本語 (原語)	極性
全党 (全党)	0.826	象牙 (象牙)	−0.655
党組織 (党组织)	0.795	風味 (风味)	−0.640
党中央 (党中央)	0.761	精美 (精美)	−0.639
党風を整備 (端正党风)	0.757	建築 (建筑)	−0.639
党組織を整備 (整党)	0.750	旅行 (旅游)	−0.636
党員 (党员)	0.748	見とれる (琳琅满目)	−0.663
党 (党)	0.722	服装 (服装)	−0.628
共産党員 (共产党员)	0.718	たくさんの (繁多)	−0.627
高級幹部 (高级干部)	0.708	草花 (花卉)	−0.622
党らしさ (党性)	0.707	おもちゃ (玩具)	−0.621
全ての党員 (全党同志)	0.707	陶磁器 (陶瓷)	−0.616
党風 (党风)	0.701	熱帯魚 (热带鱼)	−0.614
率先垂範 (以身作则)	0.683	色々な色彩 (彩色)	−0.612
各級党委員会 (各级党委)	0.681	人口が多い (众多)	−0.612
党の規律 (党的纪律)	0.676	工芸品 (工艺品)	−0.610
党員幹部 (党员干部)	0.672	東方, 中国 (东方)	−0.606
幅広い党員 (广大党员)	0.670	広告 (广告)	−0.605
党委員会 (党委)	0.669	スーツ (西服)	−0.604
党による指導 (党的领导)	0.667	茶葉 (茶叶)	−0.599
四中全会 (四中全会)	0.659	欧米 (欧美)	−0.598

(注5)：クリーン, 清廉な政治という意味. 政治における腐敗や汚職といった概念の対義語として用いられる

(注6)：党ではなく, 政府における公的な組織

律に関する単語である．「組織–党以外」では伝統芸能や芸術，手工業，文化，食，ファッションなど政治的ではない公共空間に関する概念が多く登場していることが明らかとなった．

6. まとめと展望

本研究は文書ベクトル，*t*-SNE，Fisher の線形判別分析を用いた探索的データ解析により，中国共産党の機関紙である『人民日報』の，約 20 年間にわたるテキストデータの傾向の変化を抽出した．その結果，各時代の政治社会的特徴を反映したような，階級闘争を伴う急進的な共産主義イデオロギーから社会主義市場経済への移行というイデオロギー上の重点の変化を正確に捕捉することができた．こうした明らかな変化と同時に，組織内部の綱紀粛正や組織の再編に係る細かなイデオロギー上の変化や，党とは一線を画した生活・経済などの空間も観察することができた．こうしたイデオロギーや政策領域上の論点は，より大きなイデオロギー上の論点が顕在化する中で潜在しており，こうした細かな変化を連続的な言語空間の中で捕捉できたことは重要な発見であった．

また，本研究は追加的に確率的潜在意味スケーリングを用いることで，Fisher の線形判別分析で探索的に導出した対立軸の妥当性を確認することができた．

その一方で *t*-SNE については 2 次元に可視化したうえで分析しているため，解釈が困難な部分も見られた．また，計算の結果も実行の度に変化しうる性質を持つ．今後は非線形な条件のもと次元削減を行うことで，さらに解釈可能な結果が得られると考えられる．

文　献

[1] John Wilkerson and Andreu Casas. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, Vol. 20, pp. 529–544, 2017.

[2] Jonathan B Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, Vol. 52, No. 3, pp. 705–722, 2008.

[3] Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, Vol. 21, No. 3, pp. 267–297, 2013.

[4] Gary King, Jennifer Pan, and Margaret E Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, Vol. 107, No. 2, pp. 326–343, 2013.

[5] Niels D Goet. Measuring polarization with text analysis: Evidence from the uk house of commons, 1811–2015. *Political Analysis*, Vol. 27, No. 4, pp. 518–539, 2019.

[6] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American journal of political science*, Vol. 58, No. 4, pp. 1064–1082, 2014.

[7] Kohei Watanabe. Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, Vol. 15, No. 2, pp. 81–102, 2021.

[8] 持橋大地. 確率的潜在意味スケーリング. 情報処理学会研究報告 自然言語処理 (NL), Vol. 2021, No. 9, pp. 1–16, 2021.

[9] 人民日報社. 人民日報データベース CD-ROM 版. Vol. 1974-94, , 1974-94.

[10] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, Vol. 27, , 2014.

[11] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pp. 1188–1196. PMLR, 2014.

[12] spaCy. `https://spacy.io/`.

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, Vol. 9, No. 11, 2008.

[14] *t*-SNE. `https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html`.

[15] C.M. ビショップ. パターン認識と機械学習：ベイズ理論による統計的予測. 丸善出版, 2012.

付　録

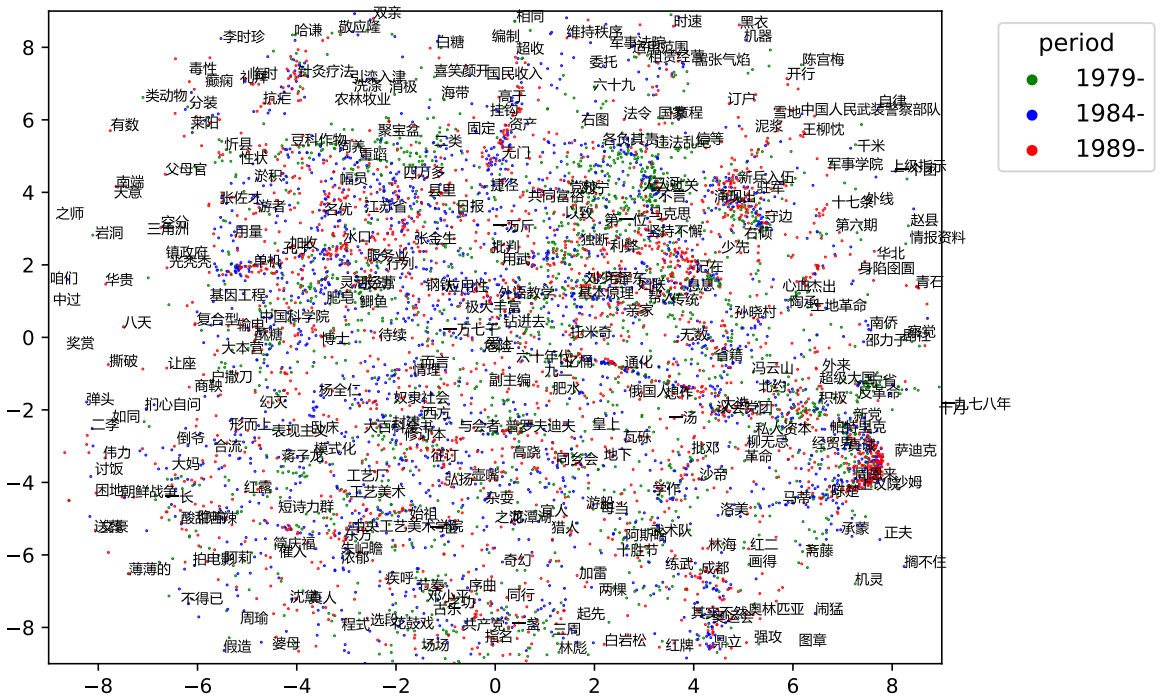


図 6: 1979 年以降の各 5 年間の記事の分布
（単語も同時に埋め込んだ場合）．