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

令和　４ 年　 ６ 月　 ４ 日現在

機関番号：３０１１５
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１４４２

研究課題名（和文）確率的潜在意味解析解の多様性分析方法の確立

研究課題名（英文）Establishment of a diversity analysis method for probabilistic latent semantic analysis solutions

研究代表者
内山　俊郎（Uchiyama, Toshio）

北海道情報大学・経営情報学部・教授

研究者番号：８０７０８６４４

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：確率的潜在意味解析は、文書などを解析する手法であるトピックモデリングとして知られ、大量の文書からトピックを発見する目的で使われている。しかし、用いるアルゴリズムや初期値によりさまざまな解が得られるという問題がある。そこで本研究では、多数の解を求め、解に座標値を与えて解の分布を可視化する方法と、解として取りうる単語分布および典型的な解を見出す方法を提案した。これは「どのような解が存在するのか」を示すことであり、利用者が目的に沿った解を選択するうえで有用な情報となる。

研究成果の学術的意義や社会的意義
確率的潜在意味解析を含むトピックモデルの研究において、さまざまなアルゴリズムが提案され、性能の比較が行われてきた。アルゴリズムによる性能差があまり大きくない一方で、アルゴリズムや初期値によりさまざまな解に到達する問題があり、これらを分析して可視化する研究が学術的に重要であると考え、分析方法についての提案を行い、実験により有用性を確認した。トピックモデルは、大量の文書からトピックを見つける重要な手段であるが、多様な解の中から適切なトピックを選ぶことを可能にする本技術により、その有用性が増すので、本成果は社会的な意義があると考える。

研究成果の概要（英文）：Probabilistic latent semantic analysis, also known as topic modeling, is a method for analyzing documents and other information, and is used for the purpose of discovering topics from a large number of documents. However, there is a problem that various solutions can be obtained depending on the algorithm used and initial values. In this study, we proposed a method to visualize the distribution of solutions by assigning coordinate values to the solutions, and a method to find the word distribution and typical solutions that can be taken as a solution. This is to show "what kind of solutions exist," which is useful information for users to select a solution suitable for the purpose.

研究分野：データマイニング、データ解析

キーワード：トピックモデル　解の多様性　正規化相互情報量　多次元尺度法　単語分布　解の類型化　類似関係ネットワーク

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

確率的潜在意味解析 PLSA などトピックモデルの解を最適化アルゴリズムで求める場合、初期値の違いにより様々な解に到達する。このとき、目的関数の値が同程度であっても、異なる解（局所最適解）が多数存在する。トピックモデルの解は、解析データに対する何らかの知見（解釈）を表しており、異なる解がそれぞれ意味を持ち、モデルとして有用と考えられる。したがって、目的関数の意味で優れている「ある解」を唯一の解として後段の解析を行うことは、重要な観点を見落とす危険性を持つ。

２．研究の目的

本研究の最終的な目的は、トピックモデルの多様な解を積極的に活かすことであり、それに資するため、本研究では「確率的潜在意味解析解の多様性分析方法の確立」を目的とする。

３．研究の方法

トピックモデルにおける解とは、モデルパラメータであるトピック分布と単語分布である。多様な解を分析する方法として、トピック分布を用いる方法と単語分布を用いる方法の２つを提案する。

トピック分布を用いる方法は、2 つの解に対して算出できる正規化相互情報を解の類似度として定義し、多次元尺度法（MDS）により座標値を割り当てるものである。この座標値を用いることで、低次元空間において解の分布を可視化することができる。

単語分布を用いる方法は、クラスタリングと類似関係のネットワーク表現により、互いに類似した単語分布のグループを得て、そのグループの頻度分布に基づいて、各解を類型化するというものである。これにより、いくつかの典型的な解と取りうる単語分布を、人間が理解しやすい形で表現することが可能になる。

４．研究成果

トピック分布を用いる分析方法の研究では、トピックモデルの２つの解における正規化相互情報量を求める方法を示し、これを両解の類似度として定義した。これまで、クラスタリングの２つ解の類似度を NMI や Rand Index などで求める方法はあったが、本研究ではトピック分布に対して定式化を行った。実際には、トピックモデルの２つの解 A と B を考える。A は J トピック、B は K トピックから成るとし、i 番目のデータについて解 A と B それぞれのトピック分布を θ_j^i及びθ_k^iと表す。するとトピックA^jとB^kが同時にサンプリングされる度数は、$t^i\theta_j^i\theta_k^i$と書ける。ここで、t^iは i 番目のデータに含まれる単語数である。データ全体について同時サンプリングされる度合いを積算すると

$$D_{\mathrm{g}}(A^j, B^k) = \sum_{i=1}^{N} t^i \theta_j^i \theta_k^i,$$

を得る。各セルが上式であるようなクロス表（下記）を考えることができ、

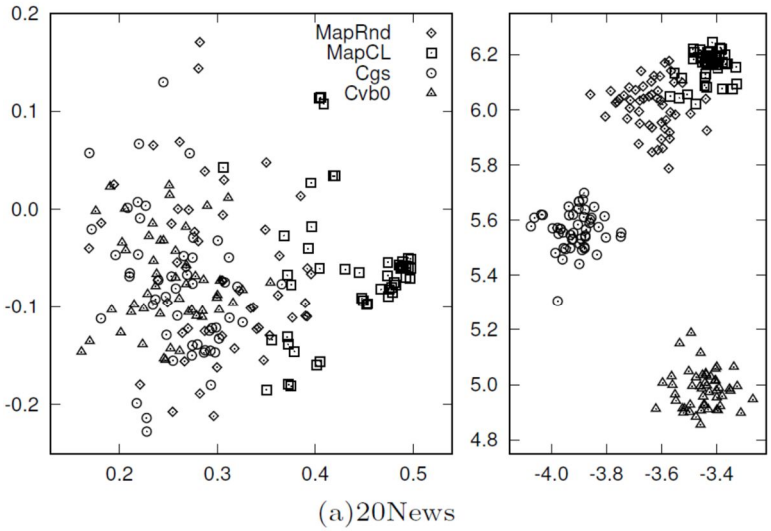
	B^1	B^2	$\cdots$	B^K	Sum
A^1	20.1	1.3	$\cdots$	7.2	302.4
A^2	23.2	0.1	$\cdots$	3.9	311.7
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A^J	1.1	4.5	$\cdots$	19.1	295.8
Sum	210.0	190.8	$\cdots$	240.1	T

ここから同時確率$P(A^j, B^k)$を定義すれば、正規化相互情量

$$\mathrm{NMI}(A, B) = \frac{I(A;B)}{\left(H(A) + H(B)\right)/2},$$

を表すことができる。これを類似度として考え、多次元尺度法を適用することで、解に座標値を付与することができる。

すべての解に座標値を付与して、主成分分析をすることで２次元空間に解を可視化することができる。これを表したのが、以下の図１である。右は、判別分析を適用した結果である。



(a)20News

図１　20News データセットに対する解集合の主成分分析（左）と判別分析（右）による可視化.

このように表すことで、解の分布状況を把握することが可能になる。

単語分布を用いる分析方法の研究の狙いは、人間が理解しやすい形で解を類型化して見せること、また取りうるトピックの単語分布を類似関係も含めて表すことである。トピック分布を用いる分析が、クラスタリング解の分析方法を拡張したものであったのに対し、多様な解の単語分布を分析する方法は、あまり類似の研究が見られなかった。提案した分析方法において重要なことは、「極めて類似する単語分布」と「異なるが似ている単語分布」の関係を区別し、２段階で分析したことである。第１段階は、すべての解に含まれる単語分布（実験では８０００個）を集めて、「極めて類似する関係」に基づき、情報理論的クラスタリングを行って、１００個の代表単語分布に集約した。第２段階では、代表単語分布間にある「異なるが類似する関係」を見出し、これを連結することで、「代表単語分布の類似ネットワーク」として表したが図２である。

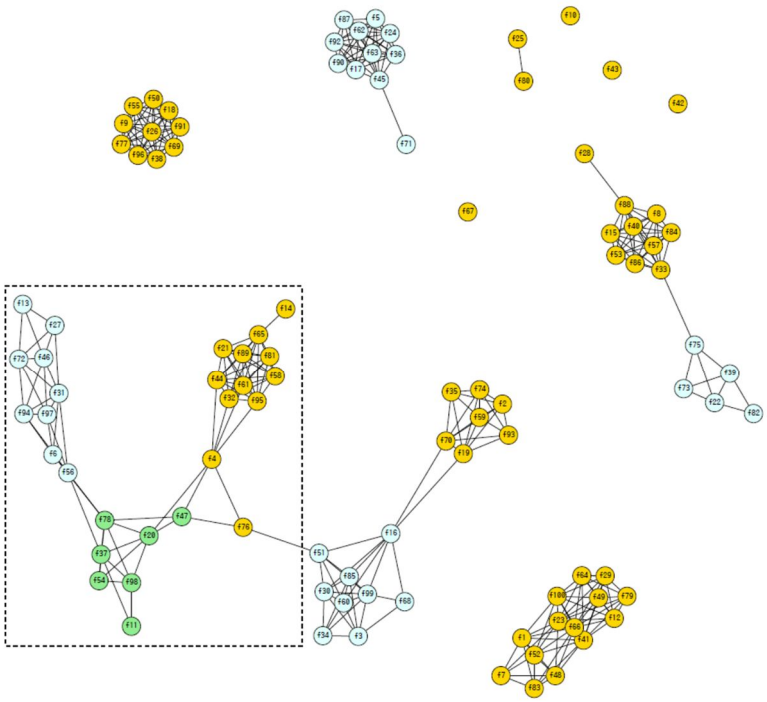


図２　代表単語分布の類似ネットワーク

図２におけるノードは、代表単語分布であり、多様な解において取りうる単語分布を表している。さらに類似ネットワークには、エッジ密度が高い部分があり、これらを「モジュラリティの最大化による分割」手法を適用して分割し、「異なるが似ている」グループを得た。図２ではグループの境界がわかるように色分けした。個々の解を調べると、それぞれのグループに属する単語分布の出現頻度分布がわかる。その頻度分布パターンにより、解を類型化した。よく出現する解は、典型的な解といえる。ニューヨークタイムズのデータセットから、得られた典型的な解を、３パ

ターンについて表したのが、以下の表である。

gn	Note	pn	High freqency words
1	スポーツ	1	season team game run games hit player inning play baseball right home yankees manager left
			team game season player play point games coach win won yard shot final football played
		2	team game season player play games point run coach win football goal shot played guy
		3	run season team game hit inning baseball race player games right home yankees manager start
			team game season player play point games coach win shot won yard played final playing
2	市場	1	company percent million companies market business stock billion industry money firm cost computer web sales
		2	percent company million companies market stock business billion money cost firm industry plan economy pay
		3	percent company million companies market stock business billion money cost industry firm plan economy analyst
3	IT	1	
		2	com web computer information site www mail online .internet internet system technology company newspaper
		3	com web computer site information www car mail system online .internet sites technology internet question
4	大統領選挙	1	.bush president campaign .george_bush .al_gore election political vote republican .white_house bill .congress
		2	.bush campaign president .george_bush .al_gore election political vote republican .white_house bill .congress
		3	.bush president campaign .george_bush .al_gore election political vote republican .white_house .congress bill
5	国際紛争	1	official government .united_states attack .u_s military war leader .american palestinian country terrorist .israel
		2	official government .united_states attack .u_s military war leader palestinian .american terrorist country .israel
		3	official government .united_states .u_s attack military war leader .american palestinian country terrorist .israel
6	薬	1	drug patient doctor percent problem cell research study health test scientist human disease medical care
		2	drug patient doctor health research study scientist cell problem percent human test disease medical care
		3	drug patient doctor scientist disease cell research health study human medical test cancer percent problem
7	学校	1	school student case law court lawyer children police official family com death question told officer
		2	school student law case court lawyer official children police family death told federal officer member
		3	school student children family case law told official home police lawyer court death parent found
8	エンターテイメント	1	show book film movie music look play women friend character love word family young director
		2	show film book movie look women play music friend character love family young word director
		3	show book film movie look music play women character love young friend director word
9	住居	1	car building home room water hour house area air town miles place night feet com
		2	car building home room water hour house town miles area air place night feet small
		3	
10	食事	1	cup food minutes add oil tablespoon wine sugar pepper water restaurant fat chicken teaspoon cooking
		2	cup food minutes add oil wine tablespoon sugar pepper restaurant water fat chicken teaspoon cooking
		3	water food room cup building minutes small hour restaurant large home add town oil house

２番目に多くみられた解（pn=2）は、すべてのグループを１つずつ含む解である。典型的な解と、その解からの変動を表せている。最後に、この類型化した結果を、トピック分布の分析で得られた２次元可視化図に反映させたものが図３である。解の分布状況を表し、加えてそれらの解の単語分布を人間が理解しやすい形で表せている。本研究の成果をすべて反映させた図といえる。

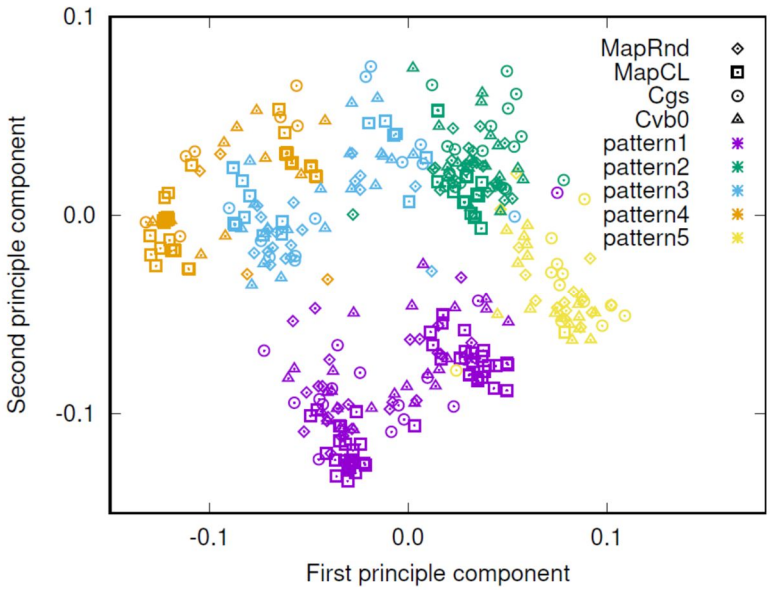


図３　解を可視化した図

なお、本研究では時間的に構造が変化する特徴をもつ文書データの分析法を検討することについても目標の一つとしてきたが、研究の検証を行うために十分な実現象のデータを得ることが難しく、具体的な成果を得るまでには至らなかった。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 内山俊郎、甫喜本司	４．巻 J102-D
２．論文標題 トピックモデルにおける解の多様性の分析と可視化	５．発行年 2019年
３．雑誌名 電子情報通信学会論文誌	６．最初と最後の頁 698-707
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2019JDP7017	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 内山俊郎、甫喜本司	４．巻 J105-D
２．論文標題 トピックモデルにおける多様な解の単語分布に基づく解析	５．発行年 2022年
３．雑誌名 電子情報通信学会論文誌	６．最初と最後の頁 405-415
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2021JDP7053	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 内山俊郎、甫喜本司
２．発表標題 トピックモデルにおけるさまざまな解の特徴分布の違いを明らかにする方法
３．学会等名 電子情報通信学会
４．発表年 2020年

１．発表者名 内山俊郎、甫喜本司
２．発表標題 トピックモデルに関する様々な解の特徴分布に基づく分析
３．学会等名 第20回複雑系マイクロシンポジウム
４．発表年 2021年

１．発表者名
内山俊郎、甫喜本司

２．発表標題
トピックモデルにおける解の多様性の分析方法と結果

３．学会等名
電子情報通信学会

４．発表年
2019年

１．発表者名
甫喜本司、内山俊郎

２．発表標題
スイッチング型マルコフモデルの食品機能性評価への応用

３．学会等名
電子情報通信学会

４．発表年
2019年

１．発表者名
内山俊郎、甫喜本司

２．発表標題
トピックモデルにおける典型的あるいは意外性のある解の探索法

３．学会等名
第19回複雑系マイクロシンポジウム

４．発表年
2020年

１．発表者名
Toshio Uchiyama

２．発表標題
A Method for Analyzing Solution Diversity in Topic Models

３．学会等名
5th International Conference on Business and Industrial Research (ICBIR)（国際学会）

４．発表年
2018年

１．発表者名 内山俊郎
２．発表標題 トピックモデル解の多様性の分析
３．学会等名 電子情報通信学会　信学技報
４．発表年 2018年

１．発表者名 内山俊郎　甫喜本司
２．発表標題 トピックモデルの解の推定方法への依存性
３．学会等名 複雑系マイクロシンポジウム2019 予稿集
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	甫喜本　司 (Hokimoto Tsukasa) (00241373)	北海道情報大学・情報メディア学部・教授 (30115)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関