differential PLSA
－テキスト情報の典型的なトピックではないより個性的なトピックの抽出－

differential PLSA
A Method of Extracting not Representative Topics but More Individual Topics from Text Data

野守　耕爾[*1]
Koji Nomori

[*1] 株式会社アナリティクスデザインラボ
Analytics Design Lab Inc.

Abstract: This study proposes a new method extracting topics from text data named *differential* PLSA. The method applies PLSA to a differential co-matrix computed by logarithm of the ratio of an observed co-matrix to expected one. It enables to extract not representative topics but more individual ones. This paper showed the effectiveness by applying the method to patent document data and comparing with results using normal PLSA. As a result, topics extracted by the method were composed of less frequent and more concrete elements, and they were more individual.

1. はじめに

データ活用がビジネスで進む昨今，テキストデータを分析することで定性情報によるインサイトを獲得して業務に活用しようとする取り組みも増えている．筆者は企業におけるデータ分析・活用のコンサルティング事業を展開しているが，ビジネスにおけるテキストデータの活用では，例えば顧客満足度調査などで実施されるアンケートの自由記述文を分析し，回答者の生の声から潜在的なニーズを抽出することでマーケティングに活用したり，コールセンターの問い合わせ履歴のデータを分析し，消費者ならではの隠れた評価の観点を抽出することで商品・サービスの改善を検討したり，特許公報の文書データを分析し，技術トレンドや他社の得意技術などを抽出することで研究開発戦略や他社との提携戦略を検討するなど，様々な業務領域でテキストデータの活用が進められている．

ビジネスにおけるテキストデータの活用では，新たな気づきとなるインサイトをテキスト情報から抽出することがしばしば求められるが，結果が期待外れということも少なくない．その理由は，業務担当者からすると経験的によく知っている結果であり，目新しさがないということが多い．テキストデータの分析には，テキストマイニング技術を適用することが通例であるが，テキストマイニングは文章に含まれる単語を抽出し，主に文章の全体像を各単語の出現頻度をベースに理解する手法となる．ここで，頻度の多い単語では当たり前の結果となるため，頻度の少ない単語まで確認しようとすると，今度は結果が複雑になり何が重要なのか分からず，結果の理解ができなくなるというジレンマに陥ってしまう．特に対象となる文章がビッグデータになると，抽出される単語も膨大となり，単語ベースの結果では複雑になりすぎるため，文章を単語ではなく意味的なまとまりのあるトピックをベースに理解する手法として，トピックモデルの適用が有効となる．トピックモデルでは単語の出現データを教師なし学習によりトピックに集約するが，これは分析対象のテキストデータ全体をうまく表現するようないわば数学的な最適解となるトピックを抽出するので，どうしても典型的なトピックになりがちになり，ビジネスの業務担当者の視座では満足しがたい結果になってしまう．

連絡先：野守耕爾，株式会社アナリティクスデザインラボ，
koji.nomori@analyticsdlab.co.jp

そこで本稿では，テキストデータにトピックモデルを適用した際に，典型的なトピックだけではなく，より個性的なトピックを抽出する方法を提案し，その適用事例を紹介する．

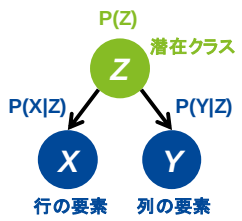
2. PLSA

本稿では従来のトピックモデルとして PLSA（確率的潜在意味解析：Probabilistic Latent Semantic Analysis）を紹介する．

2.1 PLSA の概要

PLSA は，行列データ（共起行列）の行要素 X と列要素 Y の背後にある共通特徴となる潜在クラス Z を抽出する次元圧縮法であり，元々は文書分類のために開発された手法である [Hofmann 1999]．PLSA のグラフィカルモデルを図 1 に示す．PLSA の計算では，X と Y の共起確率を潜在クラス Z を使って式展開し，この対数尤度関数を最大化する EM アルゴリズムを実行することで，3 種類の確率変数 $P(X|Z)$, $P(Y|Z)$, $P(Z)$が計算される．文書分類における PLSA では，文書 X とそこに出現する単語 Y の間には潜在的な意味トピック Z があることを想定し，各文書における単語の出現頻度が記録された「文書」×「単語」の共起行列データを学習し，文書と単語の共通特徴となるトピックを見つける．これにより「文書」×「トピック」という低次元データに変換して文書分類できる．

PLSA はクラスタリングの手法としても用いられるが，クラスタリングという観点で PLSA が他のクラスタリング手法と異なる特徴の一つは，行と列を同時にクラスタリングできることである．一般的なクラスタリング手法は，列をベースに行をクラスタリングする，あるいは行をベースに列をクラスタリングするため，どちらか一方しかクラスタリングできない．一方 PLSA で抽出される潜在クラスには，行の要素と列の要素が同時に所属することができる．



$$P(X,Y) = \sum_Z P(X|Z)P(Y|Z)P(Z)$$

図 1　PLSA のグラフィカルモデル

2.2 PLSA の共起行列構成の工夫

行と列を同時にクラスタリングできる PLSA では，行と列は双方が十分意味を持つ情報で構成すれば，抽出された潜在クラスの意味を行と列の 2 つの情報軸から解釈することができる．本来の PLSA の適用では，「文書」×「単語」という構成の共起行列をインプットとするが，この構成を工夫することで解釈のしやすい潜在クラスを抽出する試みがある．

例えば，「品詞」×「品詞」の共起行列を用いる方法や[Kameya 2005][野守 2014]，「単語」×「係り受け」の共起行列を用いる方法が提案されている[野守 2018]．「単語」×「係り受け」の共起行列を用いる PLSA では，単語と係り受け表現が同時にクラスタリングされ，単語という話題の観点となる軸に基づき，その観点の具体的内容となる係り受け表現をグルーピングできるため，より文脈上近しい言葉・表現でまとめられた解釈のしやすいトピックを潜在クラスとして抽出できるとされている．

3. *differential* PLSA

テキストデータに PLSA を適用したとき，典型的なトピックだけではなく，より個性的なトピックを抽出する方法として，*differential* PLSA（以下 *diff*-PLSA）を提案する（特許出願中，商標出願中）．

diff-PLSA では，テキストデータにおいて行要素 X_i と列要素 Y_j が共起する実測頻度 $n(X_i, Y_j)$ を値として持つ通常の共起行列 M に加え，行要素 X_i と列要素 Y_j が共起する期待頻度 $n'(X_i, Y_j)$ を値として持つ共起行列 M' を構成する．この各期待頻度に対する実測頻度の比率の対数を取った $log\left(n(X_i, Y_j)/n'(X_i, Y_j)\right)$ を値として持つ *differential* 共起行列（以下 *diff*-共起行列）を構築し，これに PLSA を適用する．*diff*-PLSA で適用する共起行列のイメージを図 2 に示す．ここで期待頻度とは，X_i の総頻度（出現文章数）$n(X_i)$，Y_j の総頻度（出現文章数）$n(Y_j)$，総文章数 N から式(1)のように計算される．なお，*diff*-共起行列の値が負数となるものは 0 に置換する．

$$n'(X_i, Y_j) = n(X_i) \cdot n(Y_j)/N \qquad (1)$$

実測頻度の共起行列に適用する通常の PLSA では，その解を求める最適化計算において，どうしても頻度が高い要素に高い確率が割り当てられ，結果として抽出されるトピックは典型的なものになる傾向があり，目新しさに欠けてしまう．一方，*diff*-共起行列では，実測共起頻度を期待共起頻度で除した値を持つが，実測共起頻度が高い共起ペアでも，元々全体の頻度が高い要素が含まれるときには期待共起頻度も高くなるため，実測共起頻度を期待共起頻度で除すことで値の大きさが制限される．逆に実測共起頻度が高くない共起ペアでも，期待共起頻度がそれよりも十分低ければ *diff*-共起行列での値は大きくなり，これに PLSA を適用した解ではこうした要素にも高い確率が割り当てられる可能性がある．つまり，通常の PLSA では頻度が低い要素は高い確率が割り当てられない傾向があるが，*diff*-PLSA ではそうした要素にも高い確率が割り当てられる可能性があり，より個性的なトピックが抽出されることが期待できる．

期待共起頻度に対する実測共起頻度の比率に対数を取る理由としては，極端に高くなる値を制限ためである．特に期待共起頻度は 1 未満となるケースも多く，比率のみでは値が高くなりすぎるものもある．この状態では共起行列全体の値の分布は大きくばらつき，極端な値の開きが生まれてしまうため，PLSA を適用した際の最適化計算において，今度はこの極端に大きな値に引っ張られる結果となり，必要以上にデフォルメされた歪んだトピックとなることが考えられる．そこでこの比率の値の対数を取ることで値の分布をならし，上記の現象を制限する．

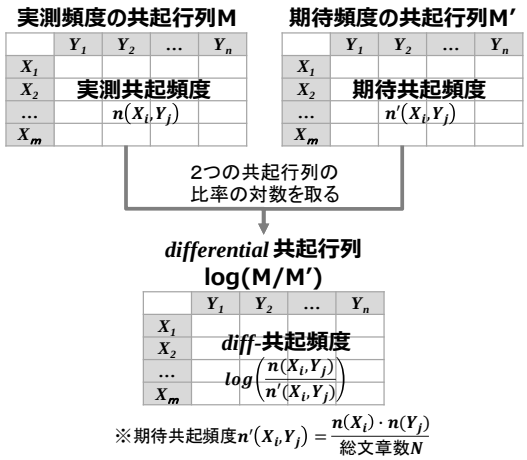


図 2　*diff*-PLSA で適用する共起行列

4. *diff*-PLSA の適用事例

本稿では特許の要約文のテキストデータを対象に，通常の PLSA と *diff*-PLSA を適用した事例を紹介し，その結果を比較することで *diff*-PLSA の有効性を検討する．

4.1 対象データと共起行列の構成

特許の要約文と請求項に「車」「電気」を含む 10 年分（出願日が 2007 年 1 月 1 日～2016 年 12 月 31 日）の国内特許公報 26,419 件を用意し，その要約文の文章データを分析対象とした．これにより電気自動車関連の技術トピックを抽出する．

共起行列の行列構成は，行を「名詞」，列を「係り受け」とした．係り受けは文法的につながりのある単語のペアを抽出するものだが，名詞に対する動詞・形容詞・形容動詞のペアを対象とした．こうした構成の共起行列に PLSA を適用することで，使われ方の似ている名詞と係り受けで構成されるトピックを得ることができる．つまり技術を象徴する名詞とその名詞を修飾する表現でまとめられたトピックを抽出でき，解釈しやすい結果が得られると期待できる．要約文にテキストマイニングを実行し，名詞と係り受けを抽出したところ，データ頻度が 20 件以上となる名詞が 3,020 語，係り受けが 2,128 表現得られた．この「名詞 3,020 語」×「係り受け 2,128 表現」で構成される共起行列を PLSA および *diff*-PLSA で適用するベースとした．

4.2 通常の PLSA の適用

通常の PLSA を適用する共起行列において，名詞と係り受けの各ペアの共起頻度とは，そのペアが同時に出現する文章数とした．ここでいう文章とは，句点で区切られた一文を指し，今回の 26,419 件の特許データに含まれる総文章数は 229,598 件であった．この共起行列に PLSA を適用してトピックを抽出するが，PLSA は予めトピック数を設定する必要があり，また初期値により解が異なる特性がある．そこでトピック数を 1 刻みで変化させ，それぞれのトピック数に対して PLSA を初期値を変えて 5 回ずつ実行し，それぞれの解を情報量基準 AIC で評価して最も評価の良い解を採用した．その結果 34 個のトピックが得られた．

PLSA のアウトプットは，①トピック Z における行要素 X（名詞）の所属確率 P(X|Z)，②トピック Z における列要素 Y（係り受け）の所属確率 P(Y|Z)，③トピック Z の存在確率 P(Z)，という 3 つの確率が計算される．特に各トピックにおいて①②の確率が高い名詞，係り受けを確認し，そのトピックの意味を解釈する．抽出されたトピックの内容例を表 1 に示す．表 1 ではトピック Z05 と Z14 について，所属確率の高い上位 7 つの名詞および係り受けと，

それぞれの所属確率 P, 頻度 n(出現文章数)についてまとめたものである. トピック Z05(表 1 上)は,「ブレーキ」や「制動力」に関する表現で確率が高いため, ブレーキに関する技術と解釈できる. トピック Z14(表 1 下)は,「給電」や「電力の供給」に関する表現で確率が高く, また「非接触」という表現も上位にあることから, 非接触受電などの給電装置に関する技術と解釈できる. このように解釈を付けた 34 個のトピックの一覧を表 2 に示す.

表 1 通常の PLSA で抽出されたトピックの内容の例

トピックZ05

P(XZ)	n(X)	X:単語	P(YZ)	n(Y)	Y:係り受け
7.3%	779	ブレーキ	3.2%	92	車両-ブレーキ
5.0%	1,384	作動	2.4%	33	ブレーキ液圧-発生
3.4%	5,188	モータ	2.2%	53	制動力-発生
2.9%	279	制動力	1.8%	49	ブレーキ-備える
2.9%	867	運転者	1.7%	32	操作量-応ずる
2.7%	1,117	車輪	1.6%	82	ブレーキ-提供
2.6%	1,005	操作	1.6%	59	電気信号-基づく
…	…	…	…	…	…

トピックZ14

P(XZ)	n(X)	X:単語	P(YZ)	n(Y)	Y:係り受け
9.7%	1,162	給電	4.3%	1,350	電力-供給
5.1%	3,629	電力	2.7%	115	給電-行う
3.6%	1,140	電源	2.5%	52	電力-受電
3.4%	329	給電装置	2.0%	28	給電-電力
2.4%	4,655	電気自動車	1.6%	124	電源-接続
2.4%	180	非接触	1.5%	20	駐車装置-変化
2.4%	1,052	外部	1.5%	34	非接触-受電
…	…	…	…	…	…

表 2 通常の PLSA で抽出された 34 個のトピック一覧

No.	トピック名	No.	トピック名
Z01	エンジンの始動と停止	Z18	演算・推定
Z02	動力の伝達	Z19	機器の異常検出
Z03	モータ駆動	Z20	操作スイッチ
Z04	ロータ・ステータなど回転部品の構成	Z21	筐体
Z05	ブレーキ装置	Z22	表面の形成
Z06	動作制御	Z23	位置とその移動
Z07	動力伝達の制御	Z24	配置・位置・方向
Z08	スイッチの切り替え	Z25	構成の方位
Z09	交流・直流の変換	Z26	構成
Z10	エネルギーの変換	Z27	接続
Z11	電池モジュールの提供	Z28	方法の提供
Z12	二次電池の構成	Z29	損傷や浸水など不具合の防止
Z13	電気自動車の蓄電池充電	Z30	小型化・簡素化・低コスト化など付加価値
Z14	非接触受電など給電装置	Z31	効率性・安全性の向上
Z15	外部への電力供給	Z32	既存エンジンへの警鐘・樹脂組成物の提供
Z16	空調などの冷却・加熱	Z33	重力発電の活用による地球温暖化防止
Z17	情報通信	Z34	タービン発電の出力向上・燃費低減

4.3 各文章に対するトピックのスコア計算

全特許データの各文章に対して各トピックのスコア(該当度)を計算した. これにより各トピックと関連の強い具体的な文章を確認できるため, トピックを解釈するうえの参考情報となる. 実際に表 2 の解釈においてもこのスコア結果を使用した.

文章 S_h におけるトピック Z_k のスコアは $P(S_h|Z_k)/P(S_h)$で定義した. これはトピックを条件とすることで文章の発生確率が何倍になるのかを示し, そのトピックをよく話題にしている文章ほど高くなる. 以下, $P(S_h|Z_k)$と $P(S_h)$の計算について説明する.

$P(S_h|Z_k)$は, 文章 S_h を名詞で定義される文章 Sx_h と係り受けで定義される文章 Sy_h に分解し, それぞれ $P(Sx_h|Z_k)$と $P(Sy_h|Z_k)$を計算し, それらを統合して $P(S_h|Z_k)$を計算する. $P(Sx_h|Z_k)$と $P(Sy_h|Z_k)$は式(2)(3)で計算される. 名詞 X_i と係り受け Y_j の頻度(出現文章数)をそれぞれ $n(X_i)$と $n(Y_j)$とすると, $P(Sx_h|X_i)$は $n(X_i)$の逆数, $P(Sy_h|Y_j)$は $n(Y_j)$の逆数として計算される. $P(X_i|Z_k)$と $P(Y_j|Z_k)$はそれぞれ PLSA の実行結果で得られている. $P(S_h|Z_k)$は式(4)で計算され, 文章 S_h において Sx_h と Sy_h の重みは同じであるため, $P(S_h|Sx_h)$と $P(S_h|Sy_h)$はそれぞれ 0.5 とする. $P(S_h)$は式(5)で計算され, $P(Z_k)$は PLSA の実行結果で得られている.

$$P(Sx_h|Z_k) = \sum_i P(Sx_h|X_i)P(X_i|Z_k) \quad (2)$$

$$P(Sy_h|Z_k) = \sum_i P(Sy_h|Y_i)P(Y_i|Z_k) \quad (3)$$

$$P(S_h|Z_k) = P(S_h|Sx_h)P(Sx_h|Z_k) + P(S_h|Sy_h)P(Sy_h|Z_k) \quad (4)$$

$$P(S_h) = \sum_k P(S_h|Z_k)P(Z_k) \quad (5)$$

上記のトピックのスコアは文章単位に計算したものだが, これを特許単位に集約し, そのスコアを特許の各属性情報(出願年や出願人など)の軸で集計すれば, トピックと属性との関係を分析することもできる. こうした方法で技術のトレンドを可視化したり, 出願人のポジショニングを可視化し, 技術戦略の検討に用いる試みもされている[野守 2018].

4.4 *diff*-PLSA の適用

diff-PLSA で採用する共起行列もその行列構成は 4.2 の通常の PLSA で使用した「名詞 3,020 語」×「係り受け 2,128 表現」の共起行列の構成と同様となるが, 先述の通り *diff*-PLSA では通常の PLSA で構築した実測頻度の共起行列に加え, 期待頻度の共起行列を構築する. 期待頻度の計算は式(1)の通りで, 総文章数 N は 229,598 件である. この期待共起頻度に対する実測共起頻度の比率の対数を取ることで *diff*-共起行列を構築し, これに PLSA を 4.2 と同様の方法で適用したところ, 最も評価の良い解として 50 個のトピックが得られた. 4.2 と同様に, 各トピックにおける名詞および係り受けの所属確率と, 4.3 の方法で計算される各文章のトピックのスコアから, 50 個のトピックに解釈を付けた. その一覧を表 3 に示す.

表 3 *diff*-PLSA で抽出された 50 個のトピック一覧

No.	トピック名	No.	トピック名
Z'01	エンジン制御	Z'26	電流・電圧の検出
Z'02	動力伝達	Z'27	温度、電流、充電量などの検出と制御
Z'03	差動機構などを備えた動力伝達の制御	Z'28	演算や推定、測定などのステップを含む方法
Z'04	回転運動	Z'29	情報の取得・提供(位置情報やバッテリ残量等)
Z'05	ロータ・ステータなどモータの構成	Z'30	スイッチなど操作装置
Z'06	モータ制御(トルク制御や回転数制御等)	Z'31	車両用灯具
Z'07	油圧ポンプなどを利用したモータ駆動	Z'32	掃除機
Z'08	ブレーキ	Z'33	基板の構成
Z'09	状態に応じた制御、運転者の操作補助	Z'34	回路の接続(電力変換回路等)
Z'10	コンバータとバッテリ昇降圧	Z'35	端子接続
Z'11	直流と交流の電力変換	Z'36	部品・装置の収容ケース・筐体
Z'12	回転力などの電気エネルギー変換	Z'37	部品・装置の配置
Z'13	エネルギー効率の向上	Z'38	パーツなどの移動、位置
Z'14	発電と蓄電	Z'39	構造の形成・方位
Z'15	電池モジュールの提供	Z'40	支持構造
Z'16	燃料電池	Z'41	装置やユニットの構成
Z'17	二次電池の構成	Z'42	システム・方法の構成
Z'18	バッテリの充放電	Z'43	その他方法
Z'19	充電システム	Z'44	組成物の製造方法(樹脂や電解液等)
Z'20	充電の接続	Z'45	機能性組成物・成形品(耐熱性や耐衝撃性等)
Z'21	非接触など受給電装置	Z'46	製造の効率化(小型化や低コスト化等)
Z'22	車両用空調など熱交換	Z'47	不具合の防止(損傷、感電、盗難等)
Z'23	冷却装置と放熱	Z'48	その他
Z'24	信号の入出力と検出	Z'49	タービン発電と船舶・飛行機への応用
Z'25	電気信号の取得と変換(センサ検出等)	Z'50	重力発電の活用による地球温暖化防止

4.5 通常の PLSA と *diff*-PLSA の比較

通常の PLSA の結果と *diff*-PLSA の結果を比較したところ, ① *diff*-PLSA の方が通常の PLSA よりも頻度の少ない名詞・係り受けでトピックが構成される傾向があること, ②*diff*-PLSA では通常の PLSA では抽出されないトピックが抽出されること, ③*diff*-PLSA の方が通常の PLSA よりもトピックの解釈が難しいことが考察された. 以下それぞれの考察について説明する.

① 頻度の少ない名詞・係り受けでトピックが構成されること

①の考察に関しては, 通常の PLSA と *diff*-PLSA で同様の解釈ができるトピックの内容を比較すると分かりやすい. 表 1 に示した通常の PLSA による「Z05:ブレーキ装置」「Z14:非接触受電など給電装置」に対応する *diff*-PLSA のトピックはそれぞれ表 3 の Z'08 と Z'21 であり, その内容を表 4 に示す. ブレーキに該当する Z'08 は, 通常の PLSA と比べて,「マスタシリンダ」や「液圧」など頻度は少ないがブレーキに関連するより具体的な表現が上位に位置している. また, 非接触受電などの給電装置に該当する Z'21 は,「非接触」が最も確率が高く, また「送電コイル」や「受電コイル」といった頻度は少ないが具体的な表現が上位に位置している. 電気自動車において非接触の充電システムは

今後重要な技術であり，その仕組みは駐車場の路面に設置した送電コイルと車両に搭載された受電コイルが重なり合うことで電磁誘導が発生し電力が供給されるという技術が代表的であり，*diff*-PLSA ではそうした技術のより具体的な表現がトピックの上位語となっている．つまり同様の意味を示すトピックでも *diff*-PLSA の方がより細かい技術を示すものが抽出されている．

また，各トピックにおいて所属確率の高い上位語の頻度 *n* が，通常の PLSA よりも *diff*-PLSA の方が少ないことを確認するため統計的検定を行った．各トピックにおいて所属確率の高い順に名詞および係り受けを並べたとき，累積確率が 50%になるまでの名詞および係り受けの平均頻度を検定用データとし，通常の PLSA と *diff*-PLSA におけるこの値のトピック平均の差について Welch の t 検定を実施した．その結果，通常の PLSA（n=34）では名詞が平均 1288.2 件（SD=686.5），係り受けが平均 85.6 件（SD=26.0），*diff*-PLSA（n=50）では名詞が平均 440.9 件（SD=137.5），係り受けが平均 58.8 件（SD=15.9）となり，名詞および係り受けの両方で有意水準 1%の違いがみられた．

表 4　*diff*-PLSA で抽出されたトピックの内容の例

トピックZ'08

P(XZ)	n(X)	X：単語	P(YZ)	n(Y)	Y：係り受け
2.2%	112	マスタシリンダ	3.2%	32	基づく-発生
2.0%	73	ブレーキ液圧	2.6%	32	操作量-応ずる
1.6%	72	ブレーキ操作	2.6%	33	ブレーキ液圧-発生
1.6%	117	液圧	2.5%	53	制動力-発生
1.4%	779	ブレーキ	2.3%	49	ブレーキ-備える
1.4%	279	制動力	2.1%	20	備える-ブレーキ
1.3%	111	操作量	2.0%	92	車両-ブレーキ
…	…	…	…	…	…

トピックZ'21

P(XZ)	n(X)	X：単語	P(YZ)	n(Y)	Y：係り受け
3.2%	180	非接触	3.0%	52	電力-受電
2.5%	78	送電コイル	2.7%	28	給電-電力
2.4%	160	受電	2.3%	35	受電-電力
2.4%	86	受電コイル	2.1%	25	電力-給電
2.2%	329	給電装置	1.8%	20	駐車装置-変化
2.0%	73	受電装置	1.7%	34	非接触-受電
1.8%	124	受電部	1.6%	21	供給-給電装置
…	…	…	…	…	…

② *diff*-PLSA でのみ抽出されるトピックがあること

②の考察に関しては，今回の方法では *diff*-PLSA の方がトピック数が多かったということもあるが，通常の PLSA のトピック数を 50 個に設定して計算した解も確認したところ，*diff*-PLSA のみに現れたトピックが複数存在した．その例を表 5 に示す．トピック Z'09（表 5 上）は，「シフトレンジ」や「パーキングレンジ」，「検出」，「停止」，「自動的に行う」といった表現で確率が高く，運転者の誤操作を抑制したり自動停止などの運転アシストに関する技術と解釈できる．トピック Z'29（表 5 下）は，「ナビゲーション」や「情報」，「目的地」，「位置情報」といった表現で確率が高く，位置情報を取得してドライバーにナビ情報として提供するなど，情報の取得と提供に関する技術と解釈できる．どちらも近年の自動車業界において付加価値を高める重要な機能といえる．

表 5　*diff*-PLSA でのみ抽出されたトピックの内容の例

トピックZ'09

P(XZ)	n(X)	X：単語	P(YZ)	n(Y)	Y：係り受け
1.4%	50	シフトレンジ	1.9%	23	自動的-行う
1.0%	38	パーキングレンジ	1.7%	38	操作-行う
0.9%	147	検出結果	1.6%	22	駆動-停止
0.7%	1,200	停止	1.6%	26	動作-行う
0.7%	365	解除	1.6%	40	要する-時間
0.6%	34	キースイッチ	1.4%	46	停止-状態
0.6%	3,960	検出	1.2%	26	ブレーキ-作動
…	…	…	…	…	…

トピックZ'29

P(XZ)	n(X)	X：単語	P(YZ)	n(Y)	Y：係り受け
1.2%	122	ナビゲーション装置	2.1%	48	情報-送信
1.1%	809	情報	2.1%	42	情報-含む
1.0%	165	目的地	1.8%	68	情報-取得
1.0%	111	位置情報	1.5%	31	情報-受信
0.9%	786	取得	1.5%	35	示す-情報
0.9%	819	送信	1.5%	126	情報-基づく
0.8%	558	表示	1.4%	25	情報-用いる
…	…	…	…	…	…

③ *diff*-PLSA のトピックの解釈が難しいこと

③の考察に関しては，*diff*-PLSA で抽出されるトピックは，頻度が少ない表現で構成される傾向にあるが故に，トピックの所属確率が頻度の少ない多様な表現に分散しやすく，通常の PLSA で抽出されるトピックよりも意味の解釈が難しい印象があった．実際に表 1 の通常の PLSA の結果と比較し，表 4,5 の *diff*-PLSA の結果は上位語の所属確率が小さい．通常の PLSA は頻度の多い代表的な表現に所属確率が集中するため，解釈はしやすいが，結果として典型的なトピックとなってしまう．この現象を解消する手法として *diff*-PLSA を開発したため，このジレンマは仕方のないものといえる．逆にそれだけより具体的で細かい要素で構成される個性の強いトピックが抽出されていることを意味している．トピックの構成内容だけでは解釈しにくいものは，4.3 で示したように各文章のトピックのスコアを計算し，各トピックと関連の強い具体的な文章を確認することが手掛かりとなる．

5.　まとめ

本稿では，テキストデータから典型的なトピックだけではなく，より個性的なトピックを抽出する手法として，従来の PLSA に対して *diff*-PLSA を提案した．本手法は，実測頻度の共起行列に加え，期待頻度の共起行列を構成し，それぞれの共起ペアにおいて期待頻度に対する実測頻度の比率の対数を取った値を持つ共起行列を構築し，これに PLSA を適用するものである．本稿では *diff*-PLSA の有効性を検討するため，特許の要約文データを例に，PLSA を適用した結果と *diff*-PLSA を適用した結果を比較した．その結果，通常の PLSA では，頻度の多い表現でトピックが構成される傾向にあり，全体を表す代表的なトピックが抽出されやすいが，*diff*-PLSA では，頻度の少ないより具体的で細かい表現もトピックを構成する重要な要素となっており，よりエッジの立ったトピックが抽出されていた．また *diff*-PLSA では通常の PLSA では抽出されないトピックも抽出されていた．

以上より，*diff*-PLSA は個性的なトピックを抽出する方法として適用価値があると思われるが，あくまでも全体を代表する典型的なトピックを理解した上で用いることが望ましく，実際には通常の PLSA と *diff*-PLSA を併用することが有効であると考えられる．

ビジネスにおいてテキストデータを分析し，その膨大な定性情報に潜む特徴を理解して有効なビジネスアクションを検討しようとする際，業務担当者が経験的に獲得するような典型的な特徴を改めて理解することも重要であるが，新たな気づきとなるインサイトを獲得するには，より個性的な特徴を抽出することも求められる．*diff*-PLSA ではそうしたトピックを通常の PLSA よりも効果的に抽出することができ，ビジネスインサイトの獲得に向けて有用な知識を提供することが期待できる．

参考文献

[野守　2014] 野守耕爾，神津友武: 三位一体アプローチによるテキストデータモデリング法の開発―宿泊施設の口コミデータを用いた評価推論モデルの構築―，2014 年度人工知能学会全国大会論文集，2014.

[野守　2018] 野守耕爾: テキストマイニングに複数の人工知能技術を応用した特許文書分析と技術戦略の検討，情報の科学と技術，Vol.68，No.7，pp.332-337，2018.

[Hofman 1999] Hofmann, T.: Probabilistic latent semantic analysis, Proc. of Uncertainty in Artificial Intelligence, pp.289-296, 1999.

[Kameya　2005] Kameya, Y., and Sato, T.: Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora, Proceedings of Symposium on Large-scale Knowledge Resources, pp.65-68, 2005.