■Article■

テキストデータを用いた人間行動モデルの構築技術
Technology for Constructing of Human Behavior Model Using Text Data

野守 耕爾
株式会社アナリティクスデザインラボ

Koji Nomori
Analytics Design Lab Inc.

keyword : 行動モデル，テキストマイニング，PLSA，ベイジアンネットワーク

1．はじめに

加速するデジタル化により大量のデータが蓄積され，それを利用する動きが活発化し高い期待が寄せられている．「ビッグデータ」という言葉は，2012年 3 月にアメリカのオバマ大統領が"Big Data Research and Development Initiative"[1]を発表してから一般的に使われ始めたといえるが，科学領域においてはこの大規模データを利用して知識を得る枠組みは第 4 のパラダイムと位置付けられており，"Data Intensive Science（データ中心科学）"と呼ばれ，第 1 の経験科学，第 2 の理論科学，第 3 のシミュレーション科学と並ぶ新たな科学研究の方法の登場として提唱されている[2]．

人間生活工学の領域では，人間の生活の質を向上させるような環境設計の実現のために，人間の行動をモデル化してその特性を理解することは重要であり，ここにもビッグデータを活用することの可能性が期待できる．

ビッグデータと一言でいってもそのデータの種類は様々であるが，人間の行動をモデル化するうえで有用となる情報としては，行動に至った経緯やその周辺状況など，行動発生における具体的な内容が記録されているデータであることが望ましい．それは人間の行動発生という結果に対して，なぜそのような行動に至ったのか，そこにどのような要因が関連しているのかといったことが解釈できなければ，有用なアクションを講じるのが難しいからである．近年 IoT 技術の進展により，人間の移動や購買などの履歴，生体情報といった膨大なセンサデータが収集可能となっているが，こうしたセンサデータは機械が取得した人間の行動の結果の情報であり，その行動に至った原因やプロセスとなる情報が不足しているといえる．一方，人間が自ら文章で記したテキストデータも大規模なものが蓄積されているが，これには人間が自らの行動，あるいは観察した他者の行動をその周辺情報と共に具体的に記録したものもあり，人間行動のモデル化において有用な情報となる可能性がある．

本稿ではこうしたテキストデータから人間の行動をモデル化することを考え，そのモデル化に有用な分析技術を適用事例とともに紹介する．なお本稿では文章で記された情報だけでなく，その文章に紐づく属性情報など非文章情報も含めたデータをテキストデータと呼ぶ．

2．テキストデータの分析技術

テキストデータの分析技術として，テキストデータから文章内の出現単語を抽出するテキストマイニング，その単語間の出現関係に基づいてトピック（単語の集合）を自動構成する PLSA（確率的潜在意味解析：Probabilistic Latent Semantic Analysis），変数間の関係をモデル化するベイジアンネットワークについて紹介する．

2．1．テキストマイニング

テキストマイニングは非構造化データであるテキストデータを統計的に分析可能な形にする自然言語処理技術であり，文章情報に含まれる単語を抽出してその品詞を割り当てる形態素解析と，その単語間

の文法的な係り受け関係を抽出する構文解析を基本技術とする手法である[3]．その単語や係り受けの出現頻度を集計したり，出現関係をネットワークやマップ図で可視化することで，テキストデータの全体像の特徴を単語ベースで把握することができる．

2．2．PLSA

PLSAは，文章分類に用いる次元圧縮手法として提案され，文章における単語の出現関係のデータ（共起行列）を学習し，文章と単語の共通のトピックとなるような特徴を見つける手法である[4]．

テキストマイニングでは文章に含まれる単語を抽出して文章全体の記述傾向を把握できるが，膨大な文章のデータを読み込めば抽出される単語も膨大で大変複雑な分析結果となり，解釈が難しくなる．そこでPLSAを適用することで，出現の仕方が似ている単語をトピックというかたちに集約する．これによりテキストデータの特徴を膨大な単語ベースではなく，いくつかのトピックをベースにシンプルに解釈することができる．

2．3．ベイジアンネットワーク

ベイジアンネットワークは，複数の変数の確率的な因果関係をネットワーク構造で表わし，ある変数の状態を条件として与えたときの他の変数の起こりうる確率（条件付確率）を推論することができる確率モデルである[5]．

ベイジアンネットワークは直接テキストデータを分析対象とする技術というわけではないが，これを応用することで，テキストデータの中の要因関係を構造化することも可能である．

3．テキストデータを用いたモデリング技術：Nomolytics

上記のテキストマイニング，PLSA，ベイジアンネットワークを連携させ，テキストデータからそこに潜む要因関係の特徴をモデル化する技術として，Nomolytics（Narrative Orchestration Modeling Analytics）[6]を開発した（特許第6085888号）．Nomolyticsの概要を図1に示す．

本技術では，まずテキストマイニングにより文章から単語を抽出し，文章内で同時に出現する単語間の共起頻度をデータ化した共起行列を作成する．次にその共起行列をインプットにPLSAを適用し，使われ方の似ている単語をトピック（単語の集合）にまとめ上げ，全テキストデータに対して各トピックの該当度を確率的に計算する．最後にベイジアンネットワークによってそのトピックを確率変数として扱い，トピック間あるいは他の属性情報との間の確率的な因果関係をモデル化する．

こうした3つの技術を組み合わせることで，膨大なテキストデータの文章をいくつかのトピックという人間が理解しやすい形に整理でき，ベイジアンネットワークによってそのテキストデータに潜む複雑な要因関係を構造化できる．

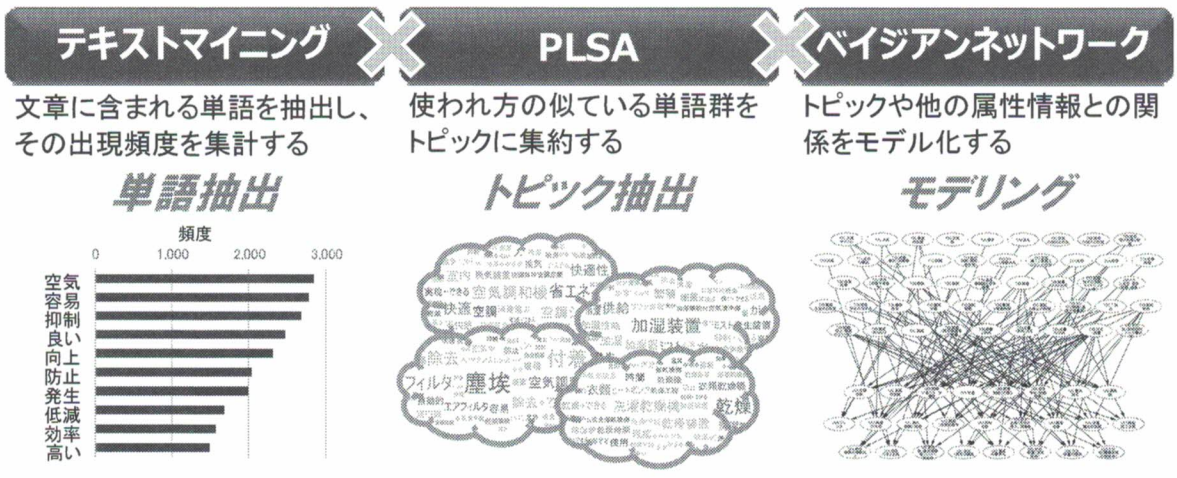

図1　Nomolyticsの手法

4．テキストデータを用いた人間行動のモデル構築

上記の Nomolytics を適用することで，人間の行動情報を含むテキストデータからその行動の発生構造をモデル化するアプローチについて，テキストデータのタイプごとに以下に紹介する．なお，本章の内容は筆者のこれまでの研究活動や企業のコンサルティング業務において適用した事例に基づき紹介したものである．

4．1．ユーザーレビューデータ

ユーザーレビューは，消費者によって Web 上に投稿された商品やサービスの評価に関する感想文形式の情報であり，口コミとも呼ばれるが，宿泊施設やレストラン，観光地，家電製品，化粧品など，インターネットの普及により様々なジャンルの投稿情報が蓄積されている．レビューには性別・年代といった投稿者の属性情報や評価得点などの情報も投稿されていることが多い．消費者の選択や体験，評価，ニーズに関することを確認できる貴重な行動データといえる．

ユーザーレビューデータから消費者の行動をモデル化する例を図 2 に示す．例えば，レビューの文章をテキストマイニングと PLSA でトピックに変換し，そのレビュートピックに対する投稿者の属性情報の関係をベイジアンネットワークでモデル化することで（図 2 上），どのような属性の人はどのようなコメントを発する傾向にあるのか予測できるため，それぞれの属性の関心対象に応じた商品企画などへの活用が考えられる．また評価得点に対するレビュートピックや投稿者属性の関係をモデル化することで（図 2 下），どのような属性の人がどのようなコメントをすると評価にどの程度影響するのかということを定量的に予測できるため，満足度を向上させるためには属性別にどのような商品・サービスが提供されることが望ましいのか検討したマーケティング戦略などへの活用が考えられる．

実際に，宿泊施設のレビューデータと観光地のレビューデータを用いて上記のようなモデル化を検討した事例を以下に紹介する．

宿泊施設のレビューデータを用いたモデル化では，11,535 件の京都府にある宿泊施設のレビューデータを用いて，評価得点に対するレビュートピックとの関係をモデル化した [7]．例えば総合得点が満点となる確率を向上させる要因の一つとしてスタッフの対応に関するトピックが影響していた．そのトピックの内容にはスタッフの対応が丁寧で親切であるということの他に，笑顔が素敵であること，挨拶が気持ち良いこと，嫌な顔をしないことなどのコメントが反映されており，スタッフの表情や挨拶も宿泊サービスにおける顧客満足に影響するということが示唆された．

観光地のレビューデータを用いたモデル化では，12,564 件の国内の温泉地におけるレビューデータを用いて，評価得点に対するレビュートピックと性別・年代・同行者との関係をモデル化した [8]．例えば男性はカップル・夫婦で行く温泉旅行で高い満足度の確率が上がり，特に温泉で温まることが有効であること，女性はカップル・夫婦では行かない（友達や家族と行く）温泉旅行で高い満足度の確率が上がり，特に砂湯で楽しむことが有効であるという結果となり，属性の条件によって温泉旅行の満足度を高める要因が変化するということが示された．例えばこれを活用し，それぞれの属性によって異なる温泉旅行の魅力をその属性が接触しやすいメディアで紹介するといった，効果的なプロモーション戦略の検討が考えられる．

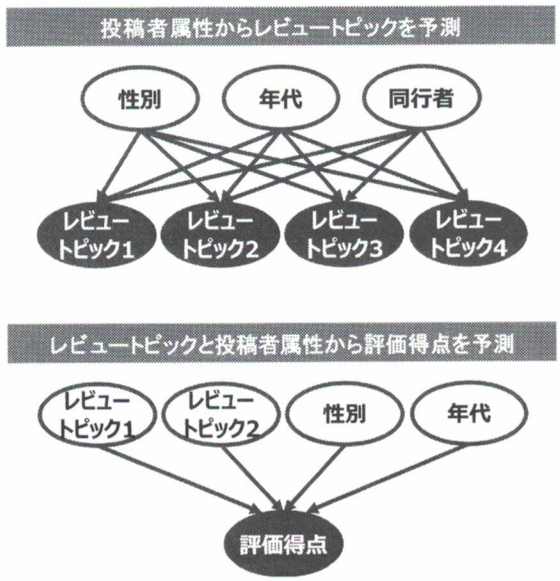

図 2　ユーザーレビューデータを用いた行動のモデル化の例

4．2．コールセンター問い合わせデータ

企業のコールセンターに連絡のあった問い合わせ情報はオペレータがその内容を文章で記録しており，最近では音声から直接テキストに変換されるシステムを導入している企業もある．問い合わせデータの記録対象は企業によって様々だが，その問い合わせが企業のどの商品・サービスに対するものなのか紐づいて記録されていることが多く，取得可能な場合は問い合わせた顧客属性の情報も紐づけがされている．また問い合わせ内容によってエスカレーション（オペレータがその場での対応が困難な問い合わせに対して専任者に引き継ぎをすること）の発生有無が記録されていたり，問い合わせが苦情の場合は，その不満度がオペレータの主観で点数付けがされていたり，保険などの契約商品の場合は解約の有無などの情報も付与されていることがある．

コールセンターの問い合わせデータから顧客の行動をモデル化する例を図3に示す．例えば，問い合わせの文章をテキストマイニングとPLSAでトピックに変換し，その問い合わせトピックに対する顧客属性や商品特徴の関係をベイジアンネットワークでモデル化することで（図3上），どのような属性の顧客はどのような商品でどのような問い合わせをする傾向にあるのか予測できるため，既存商品の改善に活用したり，新規商品を市場に投入する際も，その特徴を有する商品ではどのような問い合わせが発生する可能性が高いのか事前に予測し，問い合わせ対応の準備などに活用することが考えられる．またエスカレーションの発生や不満度，解約の発生などに対する問い合わせトピックや顧客属性，商品特徴の関係をモデル化することで（図3下），どのような属性の顧客がどのような特徴の商品にどのような問い合わせをすると対応困難となるのか，不満度が高まるのか，解約確率が高まるのかといったことを予測できるため，それを抑制するにはどのような内容の問い合わせを解消すべきなのか検討し，顧客離反防止などへの活用が考えられる．

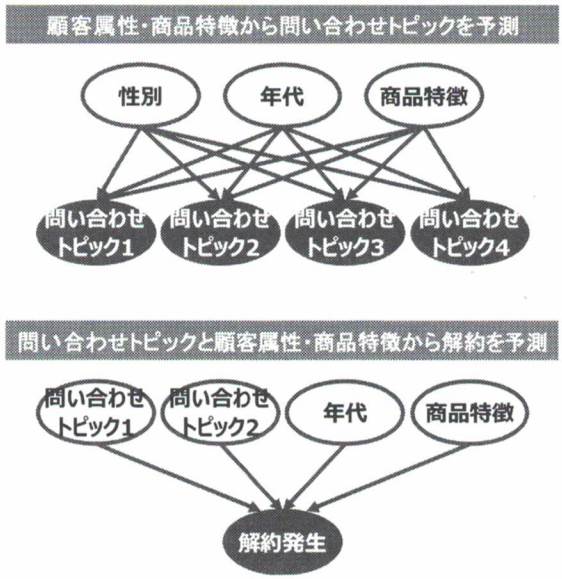

図3　コールセンターの問い合わせデータを用いた行動のモデル化の例

4．3．自由記述付きアンケートデータ

アンケートデータを用いた回答者の行動の特徴分析は従来より多くの取り組みがされているが，その基本は構造化された選択式の回答データに統計解析を実行するものであり，非構造化データである自由記述式の回答がある場合は，その回答データだけ目視で確認したり，テキストマイニングを実行して単語ベースの分析が行われることもあるが，自由記述式の回答は他の選択式の設問回答とは分けた処理がされがちである．

この自由記述付きのアンケートデータから回答者の行動をモデル化する例を図4に示す．自由記述の回答文章をテキストマイニングとPLSAでトピックに変換し，その自由記述トピックの該当度を各アンケートデータに紐づければ，選択式の設問回答と自由記述式の回答を一緒に分析でき，ベイジアンネットワークによって両者の関係をモデル化することもできる．例えば，自由記述トピックに対する他の選択式の設問項目との関係をモデル化することで（図4上），他の設問でどのような選択をした人はどのようなトピックの価値観を有している傾向があるのか予測し，アンケート設計者が用意した設問の回答と回答者の生の声の関係性を把握することができる．また顧客満足度調査に関するアンケートでは，満足度に対する選択式の設問回答や自由記述トピックの

関係をモデル化することで（図 4 下），選択式と自由記述式の両方の回答内容から顧客満足度を予測し，その要因を考察することができる．

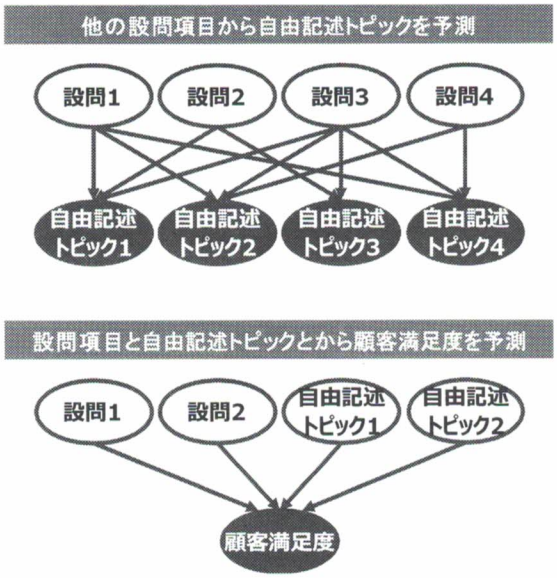

図4　自由記述付きアンケートデータを用いた行動のモデル化の例

4．4．診療データ

怪我や病気で来院した際，あるいは入院している際に，医療従事者によって記録された診療データには患者情報や検査結果だけでなく，専門家の観点によりヒアリングされた患者の行動や患者の状態に関して記録された文章情報も含まれる．

この診療データから患者の行動・状態をモデル化する例を図 5 に示す．例えば，診療記録の文章をテキストマイニングと PLSA でトピックに変換し，その診療記録トピックに対する患者情報や確認項目，検査項目との関係をベイジアンネットワークでモデル化することで（図 5 上），どのような患者はどのような行動をとりやすいのか，どのような状態に陥りやすいのか予測できるため，患者の特徴と医療従事者の観点による定性情報との関係を把握し，今後の診療や人材育成における補助情報としての活用が考えられる．また重症度に関わる情報に対する診療トピックや患者情報，検査項目との関係をモデル化することで（図 5 下），どのような要因が揃うとどれくらい重症となり得るのか予測できるため，重症度を軽減させるような操作可能な要因を探って患者に対する効果的な指導などに活用することが考えられる．

例えば，病院で収集された 4,238 件の子どもの傷害データ（転落や火傷，誤飲といった傷害の情報が受傷した子どもの情報と共に記録されたデータで，傷害発生の詳しい状況や原因となった子どもの行動などは文章で記されたデータ）を対象に，上記のようなモデル化を検討した事例を紹介する [9]．ここでは文章情報はトピック化ではなくテキストマイニングで抽出された単語をそのまま使用しているが，各年齢の子どもはどのような製品でどのような行動をとる傾向があり，それによってどのような傷害に至るのかという関係性をベイジアンネットワークでモデル化した．例えば「コイン」という製品では「飲む」という行動，「誤飲」という傷害の発生確率が高く，「ミニカー」という製品では「押す」「乗る」といった行動，「転倒」「転落」といった傷害の発生確率が特に 1 歳児で高いことが示された．このように子どもの特性や製品の情報から子どもの起こり得る行動と傷害を予測して，子どもの傷害予防を支援するツールとして開発を検討した．

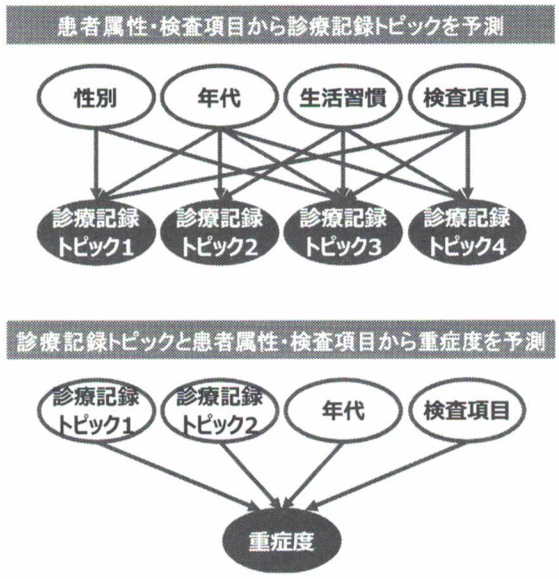

図5　診療データを用いた行動のモデル化の例

5．まとめ

本稿ではテキストデータから人間の行動をモデル化する上で応用可能な分析技術を適用事例とともに紹介した．テキストマイニングに PLSA を適用する

ことで人間の行動情報を含む膨大なテキストデータの文章をいくつかのトピックというシンプルな形に整理し，そのトピックを確率変数として扱いベイジアンネットワークを適用すれば，テキストデータに潜む人間の行動に関連する要因関係をモデル化することができ，人間の行動特性を理解するうえで有用なアプローチとなることが考えられる．

本稿で紹介したようにビッグデータはデータの量では魅力であるが，そこから異なるケースでも適用できるような普遍的な共通モデルを構築するには限界もある．特に容易に収集できる情報がビッグデータとして存在することが多く，また何か特定の分析目的があって収集されたデータではないことが多いため，ビッグデータから得られる行動の情報は限定的であり，そこからモデル化されるのは表面的な行動であることを認識したうえで利用しなければならない．人間の行動の本質となる特徴をモデル化するためには，仮説から計測項目を設計し，それに基づく調査や実験，観察を実施して質の高いデータを収集，利用することが求められる．こうしたデータを産業技術総合研究所の持丸氏はビッグデータに対して「ディープデータ」と呼んでいる[10]．どちらが優れているということではなく，例えばビッグデータで構築したモデルからこうしたディープデータ計測の仮説を生成したり，ディープデータで構築したモデルをビッグデータに適用して，不足する計測項目の情報を推定して補完するなど，両者のデータが連携することでより価値の高いアウトプットが得られるものと思われる．

ビッグデータの活用は新たなパラダイムといえるが，科学研究の方法がそれに置き換わるということでは全くなく，利用可能な方法が一つ増えたということであり，従来の方法と合わせて研究の目的・分析の目的に応じた適切な方法を選択することが重要であることは何も変わらないだろう．

参考文献

1) NITRD (Networking and Information Technology Research and Development) : THE FEDERAL BIG DATA RESEARCH AND DEVELOPMENT STRATEGIC PLAN, 2016
2) Hey, T., Tansley, S. and Tolle, K. : The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009
3) 那須川哲哉：テキストマイニングを使う技術/作る技術: 基礎技術と適用事例から導く本質と活用法，東京電機大学出版局，2006
4) Hofmann, T. : Probabilistic latent semantic analysis, Proc. of Uncertainty in Artificial Intelligence, pp.289-296, 1999
5) 繁桝算男，植野真臣，本村陽一：ベイジアンネットワーク概説，培風館，2006
6) 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.8，pp.32-337，2018
7) 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
8) 野守耕爾，神津友武：口コミビッグデータに人工知能を応用した地域観光の次世代マーケティング，2016 年度人工知能学会全国大会論文集，2016
9) 野守耕爾，北村光司，本村陽一，西田佳史，山中龍宏，小松原明哲：大規模傷害テキストデータに基づいた製品に対する行動と事故の関係モデルの構築ーエビデンスベースド・リスクアセスメントの実現に向けてー，人工知能学会論文誌，Vol.25，No.5，pp.602-612，2010
10) 持丸正明：ビッグデータ時代の歩行データベース，バイオメカニズム学会誌，Vol.40，No.3，pp.147-151，2016