トピックモデルとベイジアンネットワークを応用したテキストデータの分析と活用

株式会社アナリティクスデザインラボ　　野守 耕爾

1. はじめに

ビジネスにおけるデータ分析・データ活用の取り組みがますます広がりを見せる中、テキストマイニングを使ったテキストデータの分析と活用は比較的ポピュラーなテーマである。今ではフリーソフトウェアを含む様々なテキストマイニングのツールが登場し、ビジネスの場面でもテキストデータの分析と活用が広く実施されるようになった。テキストマイニングによって、人間ではなかなか読み切れないテキストデータの全体像を把握できるが、一方で各分析アウトプットは基本的に単語をベースに可視化されるため、結果が複雑で解釈がしづらいという課題がある。これは昨今のビッグデータでは特に顕著である。本稿では従来のテキストマイニングに加え、トピックモデルの PLSA とベイジアンネットワークという 2 つの人工知能技術を応用して筆者が開発した Nomolytics という分析手法とその適用分析例を紹介する。

2. Nomolytics: Narrative Orchestration Modeling Analytics

2.1. Nomolytics の概要

Nomolytics はテキストマイニングにトピックモデルの PLSA（確率的潜在意味解析）とベイジアンネットワークを連携させた手法であり、その概要を図 1 に示す。Nomolytics では、まず従来のテキストマイニングでテキストデータの文章に含まれる単語を抽出し、それぞれの単語が同時に出現する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に類型化する。そして分析対象のテキストデータに対してどのトピックがどれくらい該当するのかという重みを計算し、そのトピックのスコアデータを各属性別に集計することで、各属性の傾向をトピックをベースに把握する。最後に、抽出されたトピックや他の属性情報などを確率変数としてベイジアンネットワークを適用することで、トピックや他の属性情報との間に潜む要因関係をモデル化する。さらに、その確率モデルを用いることで、各要因が他の要因に与える影響を確率的にシミュレーションする。

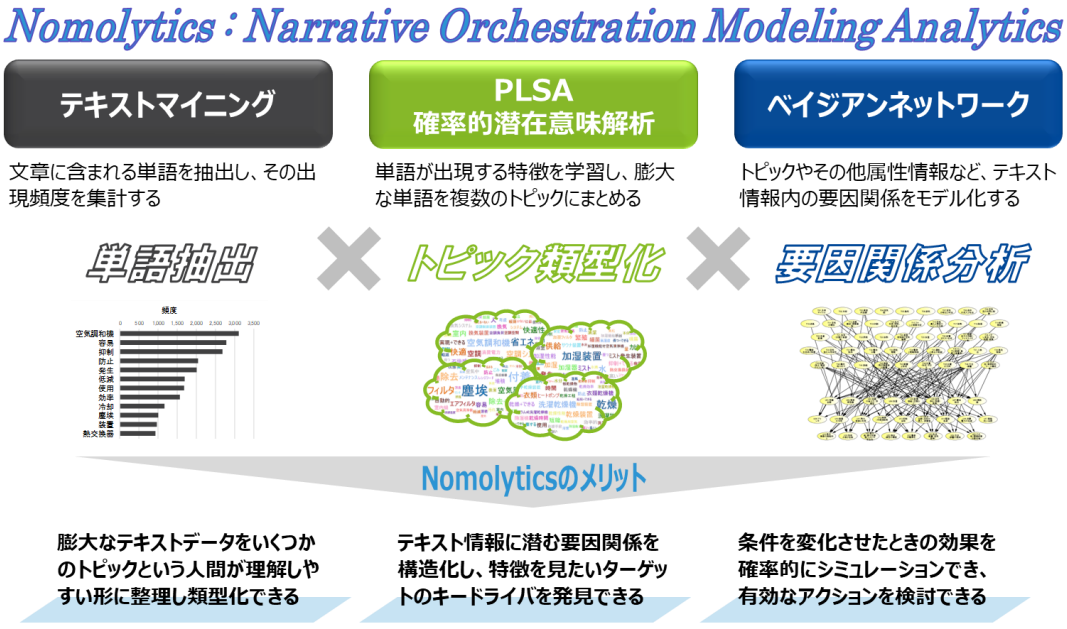

図 1　Nomolytics の概要

2.2. Nomolytics のメリット

Nomolytics では、従来のテキストマイニングのように膨大なテキストデータの傾向を膨大な単語から解釈するのではなく、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して解釈することができる。そしてベイジアンネットワークによってそのトピックや他の属性情報との関係をモデル化することで、テキスト情報内に潜む要因関係を可視化することができる。例えばアンケートの顧客満足度に寄与する自由記述のトピックにはどのようなものがあるのか、コールセンターで抑制したい問い合わせトピックの発生に寄与する属性要因にはどのようなものがあるのかなど、施策に重要なキードライバーの発見も期待できる。

また、従来のテキストマイニングは基本的に現状を把握する手法であり、仮にこの現状から何か条件が変わったらそれに伴って結果がどのように変化するのかというシミュレーションをすることはできない。これに対して Nomolytics では、ベイジアンネットワークで構築した確率モデルを用いることでこの変化に伴うシミュレーションを実行できる。例えば、口コミで話題にされるトピックによって口コミ得点がどのように変化するのか、消費者属性や商品属性によってコールセンターのある問い合わせの起こりやすさがどのように変化するのかといった、各要因の条件を変化させたときの他の要因の挙動を確率的にシミュレーションできる。あるいは、逆にその結果の発生確率を最大化、最小化するような各要因の条件も発見できる。

2.3. Nomolytics の技術的な補足

Nomolytics ではトピックモデルとして PLSA を採用しているが、今ではトピックモデルと言えば LDA を使うことが主流といえる。PLSA はデータだけに基づいてトピックが形成されるので過学習が起こりやすいが、LDA は PLSA をベイズ拡張したトピックモデルであり、ディレクレ分布という確率分布を事前分布に取ることで、過学習が回避され、頑健なモデルを得ることができるとされる。しかし、それは言い返すと、PLSA はデータの個別性をよく表現しているということである。Nomolytics では、データの理解を最も重視した手法であり、過学習をしてデータの特徴をより忠実に表現できる PLSA をあえて採用している。

また、PLSA でトピックを抽出する際にはそのインプットとなる共起行列データの前処理にある工夫を施している。共起行列は一般的には「文書×単語」という構成を取るが、Nomolytics では「名詞の単語×動詞の単語」や「単語×係り受け表現」といった構成を取っている。一般的な共起行列では、各文書にその単語が含まれるか否かという 1,0 で表現されたデータになり、ほとんどが"0"となる疎なデータとなるため、要素間の違いが現れにくく、クリアなトピックが得られにくい。一方で、Nomolytics で採用する共起行列の構成は、単語や係り受けが互いに同時に出現する頻度が入った密なデータとなり、要素間での違いが現れやすく、クリアなトピックを抽出しやすいというメリットがある。

3. Nomolytics を適用したデータ分析例

本稿では Nomolytics を VOC データと特許文書データに適用した場合の分析例を紹介する。

3.1. VOC データの分析例

VOC とは"Voice of Customer"の略で、アンケートの自由記述やコールセンターの問い合わせ履歴、口コミのコメントなど、テキスト情報として得られる顧客の生の声のデータである。顧客の具体的なニーズを把握できるため、商品・サービスの企画やマーケティング戦略を検討する上でとても重要なデータとなる。この VOC のコメントのテキストデータに Nomolytics を適用することを考えた場合、例えば図 2 のような分析が可能である。まず、人間ではなかなか読み切れず、体系的に分類することも難しいような大量のコメント

内容を、PLSA によって客観的なトピックに類型化することで VOC の全体像を把握することができる（分析 A)。また、そのトピックを軸にした特徴分析として、各商品やサービスはどのようなトピックの発言がされる傾向にあるのか定量的に把握したり（分析 B)、各顧客はどのようなトピックに関心があるのか定量的に把握することができる（分析 C)。さらに、分析 B と分析 C の結果を組み合わせることで、各商品の特徴を示すトピックに関心が高い顧客はどのような層なのか把握することができるため、どのような価値を誰に提供するべきかというマーケティング戦略を検討できる。また、ベイジアンネットワークによって VOC に潜む要因関係をモデル化することができる。商品やサービスの属性、顧客の属性、抽出したトピックを確率変数としてそれらの関係をモデル化することで、各トピックに影響を与える要因を探索し、そのトピックが発せられやすい要因条件を把握することができる（分析 D)。これにより、どのような属性の商品がどの層の顧客に提供されればどのような評価がされるのか、その商品を市場に投入する前に事前にシミュレーションすることもできる。また VOC データに満足度などの評価項目の情報があれば、評価項目に影響を与える要因をモデル化し、各要因条件からその評価を定量的にシミュレーションすることができるため（分析 E)、顧客の評価を向上させるようなプロモーションなど効果的なアクションを検討できる。

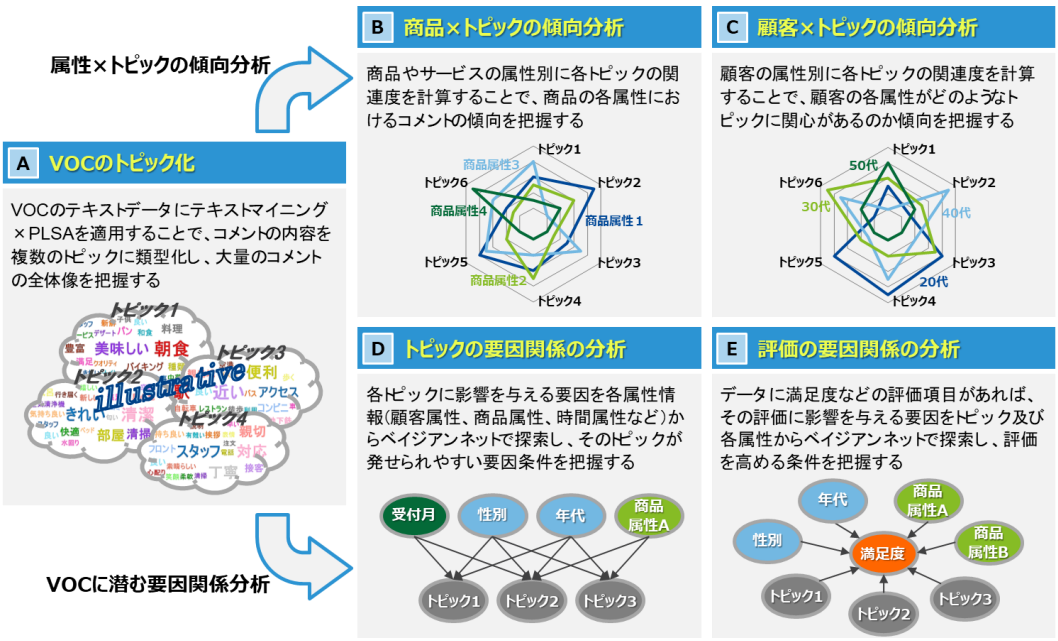

図 2　VOC データに Nomolytics を適用した分析例

3.2. 特許文書データの分析例

特許公報のデータは企業が競合他社に対する技術戦略を検討する上で貴重なオープンデータといえる。この特許広報の文章のデータ（ここでは特許の要約を対象とする）に Nomolytics を適用することを考えた場合、例えば図 3 のような分析が可能である。まず、PLSA によって客観的な視点で特許の記述内容をトピックに類型化することができる。特に要約の「課題」と「解決手段」の文章を対象に、それぞれから用途に関するトピックと技術に関するトピックを抽出することで、分析対象としている特許群における想定用途と解決技術の全体像を把握することができる（分析 A)。また、そのトピックを軸にした傾向分析として、出願年×トピックの関係を分析することで用途や技術のトレンドを把握したり（分析 B)、出願人×トピックの関係を分析することで各トピックにおける競合他社の傾向やポジショニングを把握することができる（分析

C)。特に競合他社の分析は、他社との棲み分けや差別化、協業、M&A の可能性を検討できる。さらに、ベイジアンネットワークによって用途のトピックと技術のトピックの関係をモデル化し、用途と技術の統計的な関係性を分析できる。この用途と技術の関係分析は 2 つのパターンがあり、一つは用途に対する技術の関係を確認するパターンである。これにより、ある用途の事業を実現するために重要な技術や代替技術を把握し、またそうした技術における競合他社の動向も把握することで、用途実現の事業化に向けて自社が開発・獲得すべき技術や他社との協業可能性を探ることができる（分析 D)。もう一つのパターンは技術に対する用途の関係を確認するパターンである。これにより、自社技術と関係がある用途のうちまだ想定していない用途を発見し、自社技術を有効活用できる新しい用途展開のアイデアを得ることもできる（分析 E)。

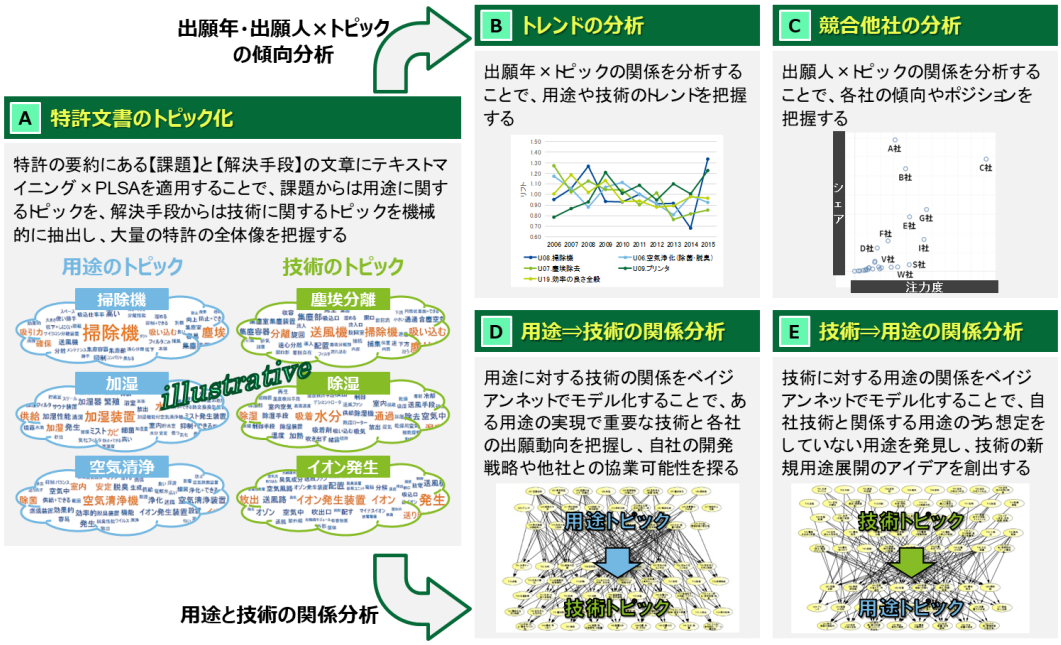

図 3　特許データに Nomolytics を適用した分析例

4. まとめ

テキストマイニングはテキストデータとツールさえあれば、統計や機械学習といった専門知識がなくても誰でも分析がしやすいが、テキストマイニングの結果は単語をベースに可視化されるため、データ規模が大きくなるほど大量の単語が抽出され、結果がとても複雑になり解釈が難しくなる。数千件程度のテキストデータであればあまり気にならないが、数万件以上、数十万件以上のテキストデータが対象になれば、その分析結果を解釈して有効なビジネスアクションを講じるという肝心のステップに辿り着くことが困難となる。

そこで本稿では、テキストマイニングで抽出された単語を PLSA でトピックに類型化し、さらにベイジアンネットワークによってそのトピックの周辺にある要因関係をモデル化する Nomolytics という手法とその分析例を紹介した。Nomolytics では、単語ではなく集約されたトピックをベースに分析を実行することで、膨大なテキストデータに潜む傾向を分かりやすくシンプルに理解することができ、また要因関係のモデルを用いることでビジネスの問題解決における有効なキードライバーの発見にも役立つと期待できる。このように Nomolytics は膨大なテキストデータからビジネスアクションに有用な特徴を発見する武器となり得る。Nomolytics の技術的な解説や具体的な分析事例は弊社ホームページ（http://www.analyticsdlab.co.jp/）を参照されたし。