Analytics Design Lab

人工知能技術コンソーシアム セミナー
【自然言語処理の潮流を一挙マスター！】
AI技術で進化するテキストアナリティクス最前線
<第2部>

大規模テキストデータから要因関係を発見する三位一体分析

テキストマイニング×PLSA×ベイジアンネットワーク

株式会社アナリティクスデザインラボ
代表取締役　野守耕爾

2024年10月23日

第2部の講演内容

第2部では、テキストマイニングにPLSAとベイジアンネットワークを組み合わせて開発した、テキストデータの新しい分析手法とその分析事例を紹介します

0. 会社紹介

1. テキストマイニングの課題

2. PLSAとベイジアンネットワークの応用

3. Nomolytics
   テキストマイニングにPLSAとベイジアンネットワークを応用した新手法

4. Nomolyticsの分析事例①
   旅行口コミデータの分析による地域の観光マーケティングの検討

5. Nomolyticsの分析事例②
   特許文書データの分析による企業の技術戦略の検討

6. まとめ

0. 会社紹介

会社紹介と代表紹介

企業様のデータ分析・活用を支援させて頂くコンサルティング会社で、これまでのアカデミックな研究とビジネスコンサルティングの両方の経験を活かして2017年6月に設立しました

会社概要

企業様のデータ分析・活用の支援をさせて頂くコンサルティング会社です

Analytics Design Lab

データというスタートから課題の解決というゴールまでをいかにつなげばよいのか、どのようなデータ処理、分析手法、考察、アクションを検討していけばよいのか、というデータ分析を活用するプロセスを企業様の抱える課題や思惑・事情などに応じてしっかりとデザインし、それを実行することで企業様の課題解決を支援します。

設立	2017年6月1日
事業内容	● 企業におけるデータ活用のコンサルティング ● 新しいデータ分析技術の研究開発
資本金	5,000,000円
所在地	東京都中野区東中野1-58-8-204
URL	http://www.analyticsdlab.co.jp/

代表略歴：野守耕爾

■ 2012年3月
早稲田大学大学院 創造理工学研究科
経営システム工学専攻　博士課程修了
博士（工学）

- 人間行動の計算モデルの開発を研究
（専門領域：人間工学）
- 2010年4月～2012年3月
独立行政法人日本学術振興会 特別研究員に採用

■ 2012年4月～（技術研修生としては2008年～）
独立行政法人産業技術総合研究所
デジタルヒューマン工学研究センター　入所

- センシング技術を応用した子どもの行動計測と人工知能技術を応用した行動の確率モデルの開発を研究

■ 2012年12月～
デロイトトーマツグループ 有限責任監査法人トーマツ
デロイトアナリティクス　入所

- データサイエンティストとしてビッグデータを活用したビジネスコンサルティング及び分析技術の研究開発に従事

■ 2017年6月～
株式会社アナリティクスデザインラボ 設立

コンサルティングサービスの概要

弊社が分析を実施しご提供する「分析受託」、お客様が実施される分析を助言する「アドバイザリー」、弊社実施の分析をお客様にトランスファーする「テーラー研修」がございます

分析受託サービス

お客様のデータをお預かりして弊社がデータ分析を実施し、結果をご報告します

- お客様の業務課題とご提供頂くデータに応じて、弊社がデータ分析の設計を行い、実行します
- 弊社による分析の実施結果をご報告し、その報告書を成果物としてご納品します
- 分析の実施にかかる期間（作業工数）から費用をお見積りします

アドバイザリーサービス

お客様ご自身で実施されるデータ分析・活用のご助言、ご指導をします

- お客様の業務課題の解決に効果的なデータ分析・活用についてご助言します
- お客様が実施される具体的なデータ分析の作業についてもご指導します
- 弊社がご納品する成果物はございません
- １回〇時間の訪問助言を何回ご提供するのかによって費用をお見積りします

テーラー研修サービス

弊社が実施した分析の内容をお客様で実施できるように、その手順を全てレクチャーします

- 「分析受託サービス」で弊社が実施した分析について、実施手順マニュアルや分析のプログラムファイルのご提供とともに解説し、お客様で同様の分析を実行できるように技術トランスファーします
- 「分析受託サービス」の費用に加え、マニュアルの作成や研修の実施などにかかる工数から費用をお見積りします

過去の実績（１）

テキストデータの分析を強みに、Web上の口コミやアンケートの自由記述、コールセンターの問い合わせ履歴、特許文書など、様々なテキストデータの分析を提供してきました

テキストデータを対象とした過去のコンサルティング実績　（デロイトトーマツでの実施案件を含む）

データの種類	クライアント業種	プロジェクト概要	期間
Web口コミ	地方自治体	観光口コミデータを活用した温泉観光地のニーズ分析	2ヶ月
	地方自治体	観光口コミデータを活用した広域観光圏の特徴分析	2ヶ月
	地方自治体	観光口コミデータを活用した観光テーマ抽出と広域ルート検討	4ヶ月
	家電メーカー	製品口コミデータを用いた機能と満足度との関係モデル構築	2ヶ月
	住宅メーカー	戸建て住宅の口コミデータを用いた顧客評価の把握と競合他社分析	3ヶ月
アンケート	地方自治体	市民意識調査の自由記述データを用いた市民ニーズの抽出と効果的な行政施策の検討	2ヶ月
	住宅メーカー	研修受講者アンケートの自由記述データを用いた新規システムのニーズ抽出と改善検討	2ヶ月
	公共事業会社	社内従業員アンケートの自由記述データを用いた従業員のモチベーション傾向分析	1ヶ月
	公益事業会社	消費者アンケートの自由記述データを用いた消費者意見の特徴とその要因関係の分析	2ヶ月
コールセンター	公共事業会社	問い合わせデータを用いた顧客接点業務の課題抽出	2ヶ月
	プリンタメーカー	問い合わせデータを用いた製品の不具合傾向の分析	1ヶ月
	食品メーカー	問い合わせデータを用いた顧客別・商品別の問い合わせ傾向およびニーズの分析	2ヶ月
	ポンプメーカー	ITヘルプデスクの問い合わせデータを用いた質問傾向の分析	1ヶ月
	住宅メーカー	メンテナンス問い合わせデータを用いた不具合問い合わせの類型化と発生傾向の分析	3ヶ月
	ビルメンテナンス会社	メンテナンス問い合わせデータを用いた不具合傾向の分析と業務改善の検討	1ヶ月
特許文書	化学メーカー	特許文書データを用いた技術分類と特許データの整理	1ヶ月
	鉄鋼メーカー	特許文書データを用いた保有技術のターゲット市場探索	3ヶ月
	精密機器メーカー	特許文書データを用いた競合他社の技術動向の把握と自社技術の新規用途探索	3ヶ月
	化学メーカー	国際特許文書データを用いた競合他社の動向把握と用途を実現する重要技術の把握	4ヶ月
	家電メーカー	国際特許文書データを用いた競合会社の技術開発動向の把握と技術戦略の検討	3ヶ月
	家電メーカー	国際特許文書データを用いた競合他社との関係把握と類似特許の探索	2ヶ月

過去の実績（２）

テキストデータに限らずデータ分析全般でサービスを提供しており、また分析の技術的・学術的な観点において学会での受賞実績も複数あります

テキストデータ以外を対象とした過去のコンサルティング実績　（デロイトトーマツでの実施案件を含む）

データの種類	クライアント業種	プロジェクト概要	期間
建築作業記録	住宅メーカー	建築作業の実績データを用いた建築物件の工程日数予測	3ヶ月
機械稼働記録	プリンタメーカー	プリンタの稼働ログデータを用いた故障予測とサービス効率化の検討	2ヶ月
顧客利用情報	保険会社	旅行保険データを用いた事故発生確率および損失額の予測	3ヶ月
顧客利用情報	カード事業会社	ポイントカードデータを用いた顧客セグメンテーションと顧客の来店確率の評価	4ヶ月
アンケート	放送事業会社	アンケートデータを用いた放送局の評価分析	2ヶ月
アンケート	自動車メーカー	ブランド調査データを用いたブランド価値向上の要因構造分析	3ヶ月
ー	システム会社	データサイエンティストの人材育成のための技術指導	6年

学会での受賞歴

受賞年月	学会	受賞内容	発表タイトル
2018年7月	人工知能学会	2018年度全国大会　優秀賞	確率的因果意味解析（PCSA）ーテキストデータを用いたターゲット事象の要因トピックの抽出ー
2018年3月	経営情報学会	2018年春季全国研究発表大会 優秀報告賞	人工知能技術を応用した特許文書分析が生み出す新たな技術戦略の検討
2015年11月	日本マーケティング学会	マーケティングカンファレンス2015 ベストペーパー賞	口コミビッグデータを活用した観光客目線によるテーマ性を持つ広域観光ルートの検討
2015年4月	サービス学会	第2回国内大会 Best Paper Award	観光クチコミデータを用いた類似観光地の発見と満足形成要素の分析
2013年3月	ヒューマンインタフェース学会	学術奨励賞	製品のデザインに関係づけられた乳幼児のよじ登り行動の計算モデル構築と分析
2011年6月	日本人間工学会	大島正光賞（最優秀論文賞）	乳幼児の環境誘発行動を予測する計算モデルの開発

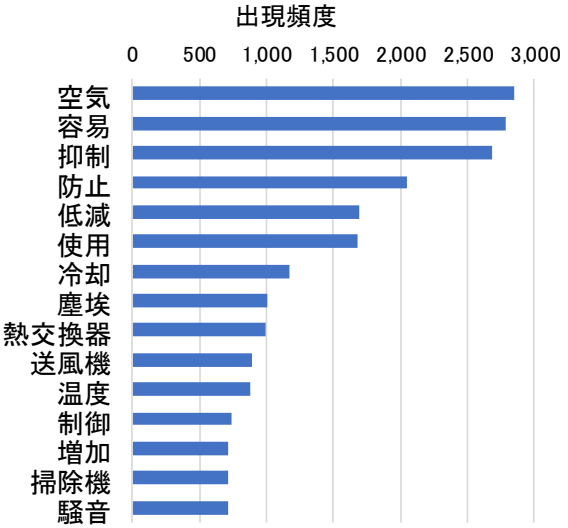
1. テキストマイニングの課題

文章に含まれる単語や係り受け表現をベースとした集計・統計分析を実行することで、文章の特徴を可視化して全体像を把握します

テキストマイニングの可視化例

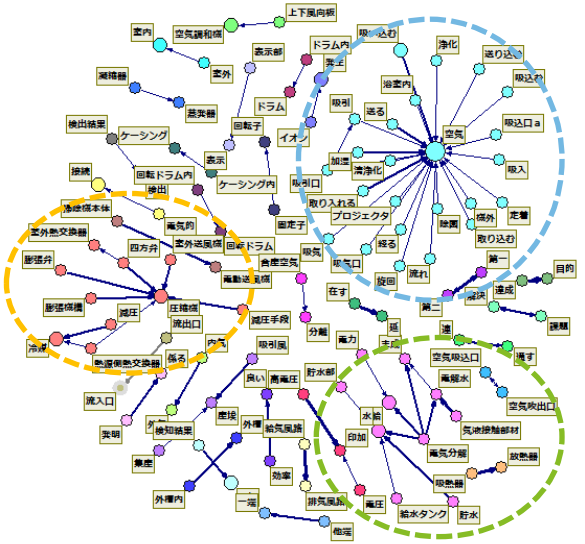
頻度集計

単語や係り受け表現の出現頻度を集計して、どのような記述が多いのか、おおまかな全体像を把握する



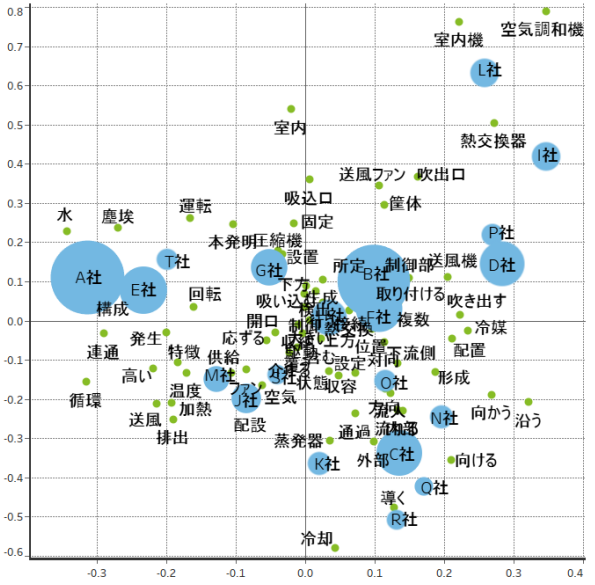
共起ネットワーク

同時に出現しやすい単語同士をネットワークでつなぎ、そのかたまりからどのような話題があるか考察する



コレスポンデンス分析

属性情報と出現単語との対応関係を同じ平面上にマッピングし、その位置関係から属性の傾向を把握する



※特許データをテキストマイニングした例を掲載

従来のテキストマイニングの課題とその解決技術

従来のテキストマイニングは単語ベースの結果で解釈が難しく、特徴がモデル化されていないので状況の変化に応じたシミュレーションができませんが、これをAI技術で解決します

課題①　解釈がしづらい

- 単語をベースにした特徴の可視化は複雑であり、結果の解釈が困難になる
- 単語を人がある程度グルーピングして分析することもあるが、その作業は主観的で負荷が大きい

解決技術

単語をクラスタリングする人工知能技術

PLSA
確率的潜在意味解析

- 単語の出現状況を学習することで、使われ方の似ている単語をまとめ上げ、トピックを抽出する
- 膨大な「単語」ではなく、いくつかの「トピック」をベースに特徴を可視化し、解釈を容易にする

課題②　現状把握に留まる

- 記述傾向の現状把握はできるが、その特徴が一般化（モデル化）されていない
- 現状から状況が変化したときに、それに伴って結果がどう影響を受けるのかシミュレーションできない

解決技術

現象をモデリングする人工知能技術

ベイジアンネットワーク

- PLSAで抽出されたトピックや、その他属性情報などの要因間に潜む要因関係をモデル化する
- 各要因の条件を変化させると他の要因がどのような挙動をとるのか確率的にシミュレーションする

2. PLSAとベイジアンネットワークの応用

PLSA（確率的潜在意味解析）

PLSAは、トピックモデルと呼ばれる人工知能技術で、複雑なデータをいくつかの潜在クラスで説明するクラスタリング手法として用いられています

PLSAの概要

- 行列データの行の要素xと列の要素yの背後にある共通特徴となる潜在クラスzを抽出する手法である
- 元々は文書分類のための手法として開発されている(Hofman, 1999)
- 各文書の出現単語を記録した文書（行）×単語（列）という高次元（列数の多い）共起行列データに適用して複数の潜在トピックを抽出し、文書（行）×トピック（列）という低次元データに変換して文書を分類する

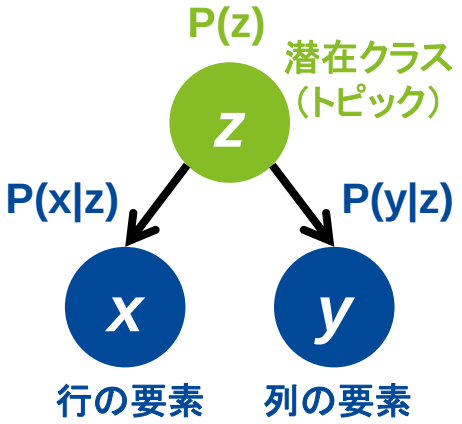
「文書×単語」行列（共起行列）

文書ID	単語1	単語2	単語3	…	単語5,014	単語5,015
1	0	0	1		1	0
2	1	0	2		0	1
…						

文書ID	トピック1	トピック2	…	トピック15
1	0.09%	0.03%		0.04%
2	0.01%	0.12%		0.06%
…				

例えば数千列ある高次元のデータでも十数個の潜在トピックで説明することができる

PLSAのグラフィカルモデル



- P(x|z), P(y|z), P(z)の3つの確率が計算される
- 潜在クラスzの数はあらかじめ設定する

※条件付確率P(A | B)
事象Bが起こる条件下で事象Aの起こる確率

xとyの共起確率を潜在クラスzを使って表現する	$P(x,y) = \sum_z P(x\|z)P(y\|z)P(z)$

PLSAのメリット

行の要素と列の要素を同時にクラスタリングできる

潜在クラスは行の要素と列の要素の2つの軸の変動量に基づいて抽出され、結果も2つの軸の情報から潜在クラスの意味を解釈することができる

ソフトクラスタリングできる

全ての変数が全てのクラスに所属し、その各所属度合いが確率で計算されるため、複数の意味を持つ変数がある場合でも自然と表現できる

複雑な観測情報をシンプルにかつ忠実に把握するため、PLSAを選択します

階層型クラスター分析

- Ward法など
- 要素間の距離を計算し、距離の近い要素同士を結合してクラスタを構成していく
- 結合の過程が樹形図で表され、結果を見てからクラスタ数を決められる（ボトムアップ的なクラスター分析）
- データ数が多くなると計算が膨大となる

非階層型クラスター分析

- k-means法など
- あらかじめクラスタ数を決め、そのクラスタ数に全要素を一回でグルーピングする
- 各クラスタ（の重心）に対して要素の距離を計算し、距離の近い要素で集められたクラスタとなるように分類結果を調整する
- 階層型クラスター分析よりも計算量が抑えられる

LSA (Latent Semantic Analysis)

- 特異値分解と呼ばれる
- (m×n)の行列を、(m×k),(k×k),(k×n)に分解する
- m個のデータとn個の変数を、k個の潜在クラスで表現する（クラス数はあらかじめ設定）
- 大きな値をとりやすいクラスが残る傾向にあるため、各要素は事前にTF-IDFなどで重み付けする必要がある

PLSA (Probabilistic Latent Semantic Analysis)

- LSAを確率的に処理
- LSAのような事前の重み付けは必要がない
- P(x,y)の確率を、P(x|z),P(y|z),P(z)に分解する
- 行要素xと列要素yを、潜在クラスzで表現する（クラス数はあらかじめ設定）
- 結果は観測データのみから定義され、新規データはクラスで表現できない（過学習）

LDA (Latent Dirichlet Allocation)

- PLSAをベイズ拡張した手法
- PLSAの過学習の問題に対して、LDAはディレクレ分布を事前分布に仮定し新規データのクラスを推定できる
- 新規データに対応するため、抽出されるクラスは観測データを忠実に再現するものではなく、クラスの抽象度が高い傾向がある

従来のクラスター分析

- 基本的に要素間の距離に基づいて分類を行う
- 要素数が多くなると要素間の距離が離れていき妥当な結果が得られにくい（次元の呪い）
- 列要素の距離に基づいて行要素を分類するか、行要素の距離に基づいて列要素を分類し、行と列どちらか一方を分類する
- 一つの要素は必ず一つのクラスタに所属し、重複所属を許さないハードクラスタリングとなる

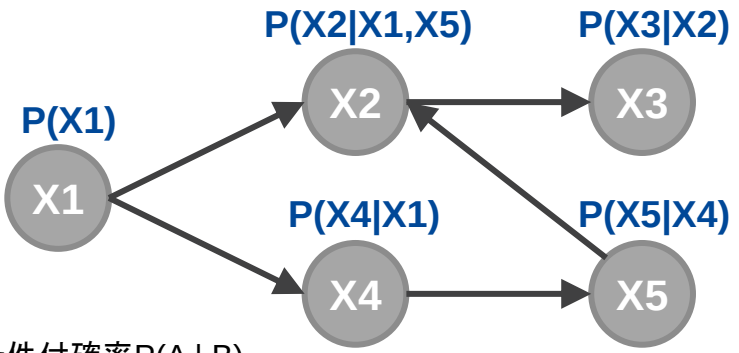
トピックモデル

- 単語一つ一つが列の要素となる超高次元のテキストデータを想定した手法
- 要素間の距離の近さで分類するのではなく、高次元データの情報をできるだけ保存した形で低次元に変換する次元圧縮手法であるため、要素数が多い複雑なデータにも対応できる
- 行の要素と列の要素の背後にある共通する特徴をクラスとして抽出するため、行と列の両方をクラスタリングでき、クラスの持つ情報が多い
- 一つの要素は全てのクラスに所属するソフトクラスタリングで、その所属の重みを計算するため、データが複数の特徴をまたがる場合でも表現できる

ベイジアンネットワークは、ベイズ推論に基づく人工知能技術で、変数間の確率的な因果関係を探索するモデリング手法として用いられます

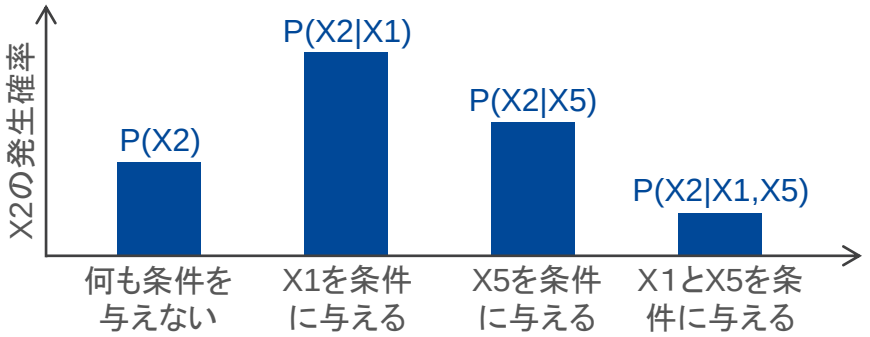
ベイジアンネットワークの概要

- 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条件付確率を推論することができる
- 目的変数と説明変数の区別はなく、様々な方向から変数の確率シミュレーションができる
- 全ての変数は質的変数（カテゴリカル変数）となるため、量的変数の場合は閾値を設けてカテゴリに分割する
- 確率論の非線形処理によるモデル化のため、非線形の関係や交互作用が生じる現象でも記述できる



※条件付確率P(A | B)
事象Bが起こる条件の下で事象Aの起こる確率

確率的因果関係と交互作用



- X2の発生確率は、何も条件を与えない時（事前確率）と比べて、X1やX5を条件に与えると確率が上昇する
  ⇒X1やX5はX2の発生に関して”確率的な”因果関係がある
- しかし、X1とX5の両方を条件に与えると、元々の事前確率よりも確率が下がってしまう
  ⇒X1とX5はX2に対して交互作用がある（X1とX5は相性が悪い）

ベイジアンネットワークのメリット

現象を理解して柔軟にシミュレーションできる

目的変数、説明変数の区別なく変数の関係をモデル化するので、現象の構造を理解でき、推論変数と条件変数を自由に指定して確率推論できる

効果を発揮する有用な条件を発見できる

ある条件のときにだけ効果が現れるといった交互作用がある場合でも、確率的に意味のある関係としてモデル化することができる

テキスト情報内に潜む要因関係を理解するため、ベイジアンネットワークを選択します

ニューラルネットワーク（ディープラーニング）

- 入力（説明変数）と出力（目的変数）の関係（非線形）をモデル化する
- 入力と出力の間に中間層（隠れ層）を設定し、入力情報に重みをつけて出力精度を高める処理を中間層で行う
- 柔軟性が高く複雑な関係もモデル化でき予測精度も高まるが、処理が複雑すぎてモデルの中身がブラックボックス化してしまう

回帰分析・判別分析（数量化Ⅰ類・Ⅱ類）

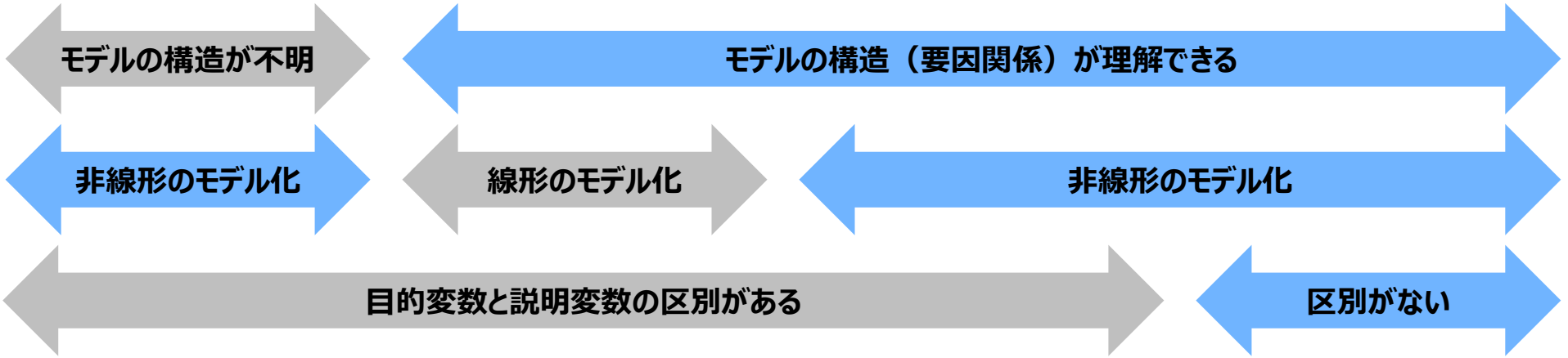
- 目的変数を説明変数の1次結合で定式化する
- 目的変数と説明変数の間に線形関係があるという仮定に基づいている
- 各説明変数の影響は独立しており、複合的な交互作用の影響は表現できない
- 説明変数間で相関が高い場合は解が不安定となり（多重共線性）、変数が多い場合この解消検討の負荷が大きい

決定木

- 目的変数の特徴がよく現れる条件ルールを説明変数とその閾値による分岐で構成する
- ルールがツリー構造で可視化されるため目的変数と各説明変数の関係が分かりやすい
- 目的変数と説明変数の非線形な関係もモデル化でき、複合条件で効果が変化する交互作用を表現しやすい

ベイジアンネットワーク

- 複数の変数の確率的な因果関係をネットワーク構造でモデル化する
- 目的変数と説明変数の区別がないため、それぞれの変数が互いにどのような関係をもってそのデータの現象を構成しているのか理解できる
- 変数間の関係は条件付確率で計算され、複合条件によって効果が変化する交互作用も表現できる



3. Nomolytics

テキストマイニングにPLSAとベイジアンネットワークを応用した新手法

膨大なテキストデータをトピックに変換して解釈を容易にし、テキスト情報内に潜む要因関係をモデル化して、ビジネスアクションに有用な特徴を把握可能にします

Nomolytics : Narrative Orchestration Modeling Analytics

テキストマイニング

文章に含まれる単語を抽出し、その出現頻度を集計する

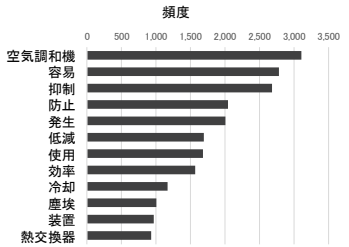
PLSA
確率的潜在意味解析

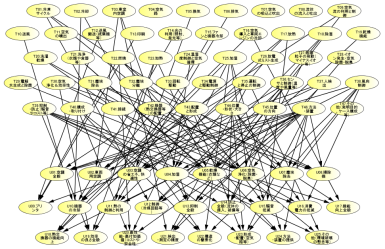
単語が出現する特徴を学習し、膨大な単語を複数のトピックにまとめる

ベイジアンネットワーク

トピックやその他属性情報など、テキスト情報内の要因関係をモデル化する

単語抽出 × トピック類型化 × 要因関係分析





Nomolyticsのメリット

膨大なテキストデータ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し類型化できる

テキスト情報に潜む要因関係を構造化し、特徴を見たいターゲットのキードライバを発見できる

条件を変化させたときの効果を確率的にシミュレーションでき、有効なアクションを検討できる

Nomolyticsの適用対象例

Nomolyticsは様々な業務のテキストデータに適用することができます

口コミ
- 顧客ターゲット別の関心トピックを把握
- 製品・サービス別のトピックを把握
- 口コミ得点に寄与するトピックを把握
- ニーズに応じたマーケティングを検討

アンケート
- 自由記述回答の内容をトピックで把握
- トピック化された自由記述回答と通常の定型設問回答の関係を統計分析
- 顧客満足度に寄与するトピックを把握

コールセンター履歴
- 問い合わせ内容をトピックで把握
- 製品別・顧客別のトピック傾向を把握
- 解約・退会に寄与するトピックを把握
- 満足度向上、顧客離反抑制の施策検討

特許文書
- 特許文書の内容をトピックで把握
- トレンドや競合他社の動向を把握
- 用途と技術の関係分析から用途実現の技術戦略や保有技術の新規用途を検討

営業日報
- 営業活動内容をトピックで把握
- 営業属性別のトピック傾向を把握
- 成約に寄与するトピックを把握
- 成約のための効果的な営業教育を検討

有価証券報告書
- 企業・業界の事業内容をトピックで把握
- 事業内容トピックのトレンドを把握
- 好業績に寄与する事業トピックを把握
- 定性情報から行う企業分析・業界分析

エントリーシート
- 志望動機やPR文のトピックを把握
- 記述トピックに基づいて学生を分類
- トピック傾向から面接の質問内容を検討
- 選考通過に寄与するトピックを把握

診療・看護記録
- 診療記録、看護記録をトピックで把握
- 患者の属性別のトピック傾向を把握
- 検査指標に寄与する定性情報を把握
- 定性情報も用いた診療・助言を検討

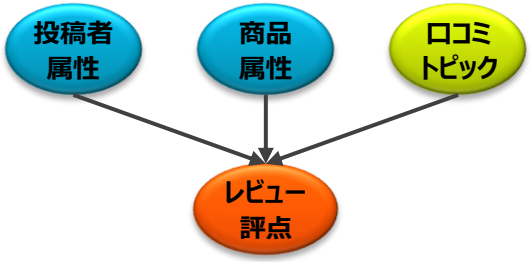
問題発生レポート
- 不具合やヒヤリハットをトピックで整理
- 作業環境別のトピック傾向を把握
- 重大問題に寄与するトピックを把握
- 問題を抑制する作業・環境改善を検討

Nomolyticsによるベイジアンネットワークのモデルの構造例

口コミ、アンケート、コールセンター等VOCデータから、顧客の行動や評価の要因を顧客の声のトピックでモデル化したり、特許文書データから用途と技術の関係をモデル化できます

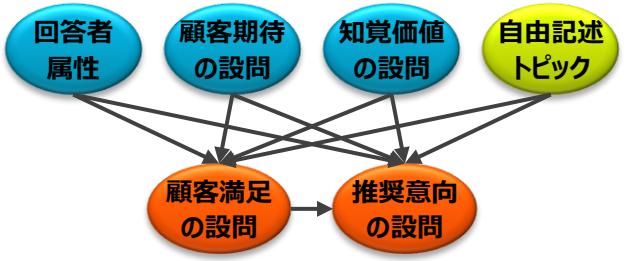
口コミ

口コミのレビュー評点の要因として、投稿者の属性や該当商品の属性、口コミ内容のトピックの効果を把握



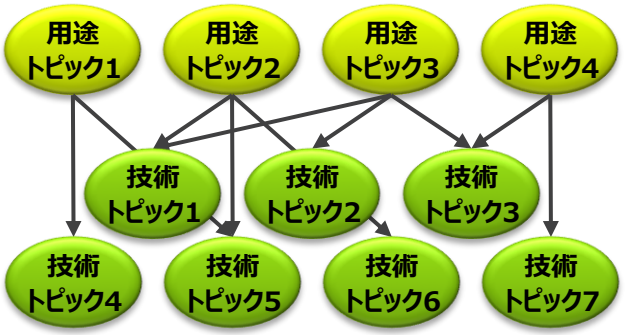
アンケート

顧客満足度の要因として、回答者の属性や選択式設問の回答結果、自由記述内容のトピックの効果を把握



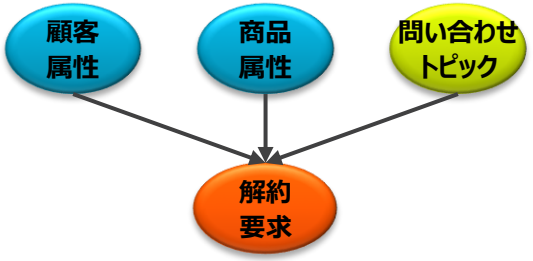
特許文書

特許の用途のトピックと技術のトピックの関係性を把握



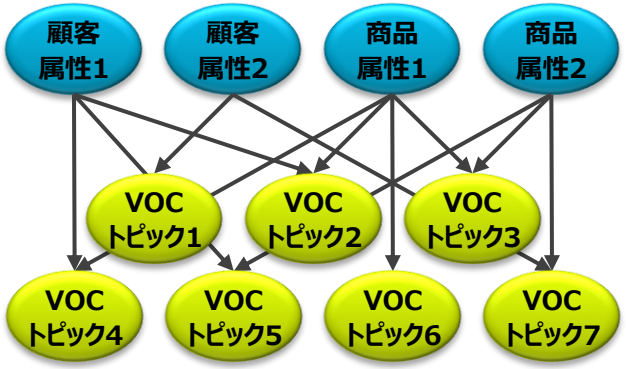
コールセンター問い合わせ履歴

解約などの要求行動の要因として、顧客の属性や該当商品の属性、問い合わせ内容のトピックの効果を把握



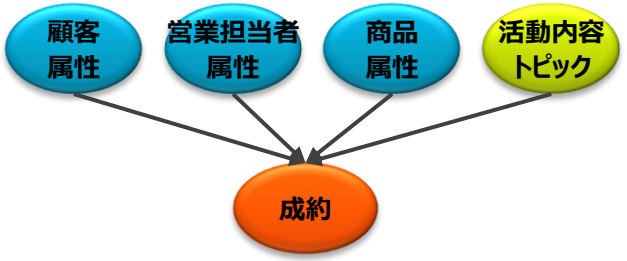
VOC共通

どのような顧客の属性や商品の属性で、どのような声が寄せられるか把握



営業日報

営業の成約の要因として、顧客や営業担当者の属性、商品の属性、営業活動内容のトピックの効果を把握



PLSAを介さないテキストマイニング×ベイジアンネットワークの直接連携の問題

テキストマイニングで抽出された単語一つ一つを変数にベイジアンネットワークでモデル化すると、複雑で解釈が困難なモデルとなり、個別性の強い要因のモデルとなってしまいます

発表論文

人工知能学会論文誌, Vol.25, No.5, 2010

602　　人工知能学会論文誌　25 巻 5 号 SP-A（2010 年）

特集論文　2009 年度 全国大会 近未来チャレンジ

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
–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–

Constructing Model of Relationship among Behaviors and Injuries to Products Based on Large Scale Text Data on Injuries
–Achieving Evidence-Based Risk Assessment–

野守 耕爾　早稲田大学大学院 / 日本学術振興会特別研究員 DC / 産業技術総合研究所
k.nomori@aist.go.jp

北村 光司　産業技術総合研究所 / 科学技術振興機構 CREST
k.kitamura@aist.go.jp

本村 陽一　（同　上）
y.motomura@aist.go.jp

西田 佳史　（同　上）
y.nishida@aist.go.jp

山中 龍宏　緑園こどもクリニック / 産業技術総合研究所 / 科学技術振興機構 CREST
tatsuhiro-yamanaka@nifty.com, http://www.cipec.jp/

小松原 明哲　早稲田大学理工学術院
komatsubara.ak@waseda.jp

keywords: childhood injury prevention, risk assessment, text mining, Bayesian network, knowledge creation

Summary

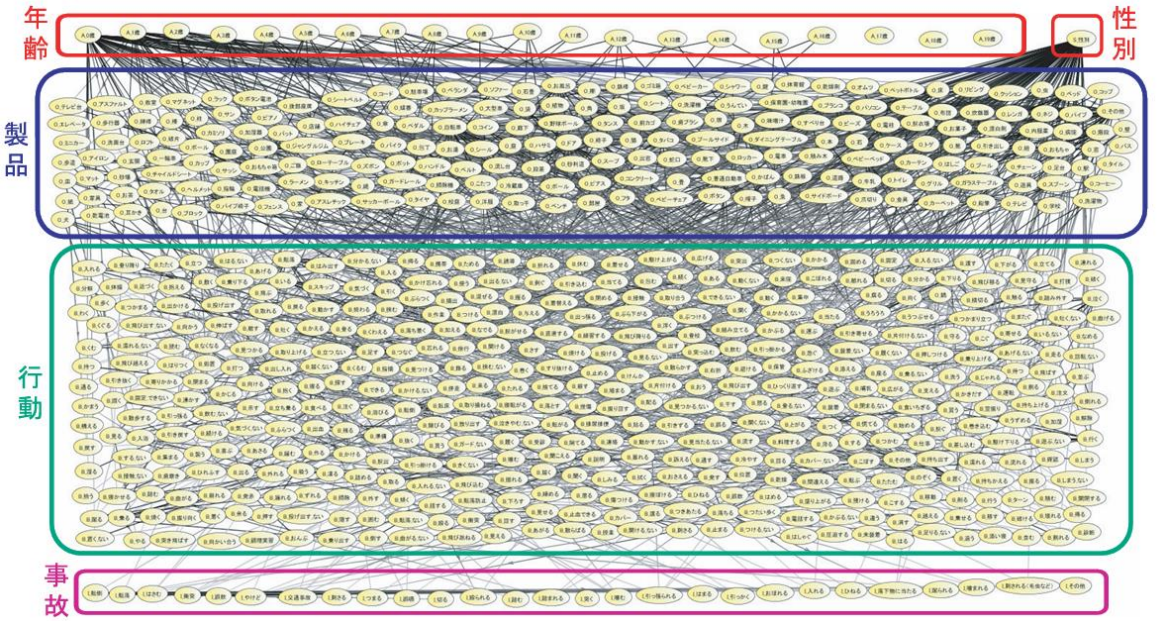
In Japan, childhood injury prevention is urgent issue. Safety measures through creating knowledge of injury data are essential for preventing childhood injuries. Especially the injury prevention approach by product modification is very important. The risk assessment is one of the most fundamental methods to design safety products. The conventional risk assessment has been carried out subjectively because product makers have poor data on injuries. This paper deals with evidence-based risk assessment, in which artificial intelligence technologies are strongly needed. This paper describes a new method of foreseeing usage of products, which is the first step of the evidence-based risk assessment, and presents a retrieval system of injury data. The system enables a product designer to foresee how children use a product and which types of injuries occur due to the product in daily environment. The developed system consists of large scale injury data, text mining technology and probabilistic modeling technology. Large scale text data on childhood injuries was collected from medical institutions by an injury surveillance system. Types of behaviors to a product were derived from the injury text data using text mining technology. The relationship among products, types of behaviors, types of injuries and characteristics of children was modeled by Bayesian Network. The fundamental functions of the developed system and examples of new findings obtained by the system are reported in this paper.

1. は　じ　め　に

人間の日常生活を支援する上で，状況依存性の高い人間の意図や行動を理解するための人工知能技術の発展，さらに日常生活を支援するサービスをどのように構築していくかという具体的な方法論の確立は重要である．近年のセンシング技術，インターネット技術，大規模なデータを扱う人工知能技術の発展によって，日常生活行動の観測データを集積し，データから計算，制御可能なモデルを構築し，モデルを再利用可能なリソースとして新たな研究の基盤として整備し，さらにこれを社会に向けたサービスとして提供するまでを一貫できる，新しい日常生活支援技術の開発が可能になりつつある．これを実現するには，技術開発研究だけで終わることなく，社会に入り込み，実サービスの中で大規模データを収集し，そ

構築したモデル

- 病院で収集された子どもの傷害事故の記録データ4,238件を用いて、事故状況が記されたテキスト情報をテキストマイニングし、事故に関連した製品の単語と行動の単語を抽出した
- データに記録された子どもの性別と年齢（20種）、事故の種類（27種）、マイニングで抽出された製品の単語（208種）と行動の単語（439種）を変数とし、それらの関係をベイジアンネットワークでモデル化した
- 子どもの発達段階に応じた製品に対する行動と事故の発生確率をシミュレーションし、事故の予見に応用できる



4. Nomolyticsの分析事例①

旅行口コミデータの分析による地域の観光マーケティングの検討

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

旅行口コミサイトの口コミデータを使用します

フォートラベル(http://4travel.jp/) の口コミ

旅行時期が2011年1月～2015年12月の直近5年間の口コミデータ総数
⇒**1,072,570件**

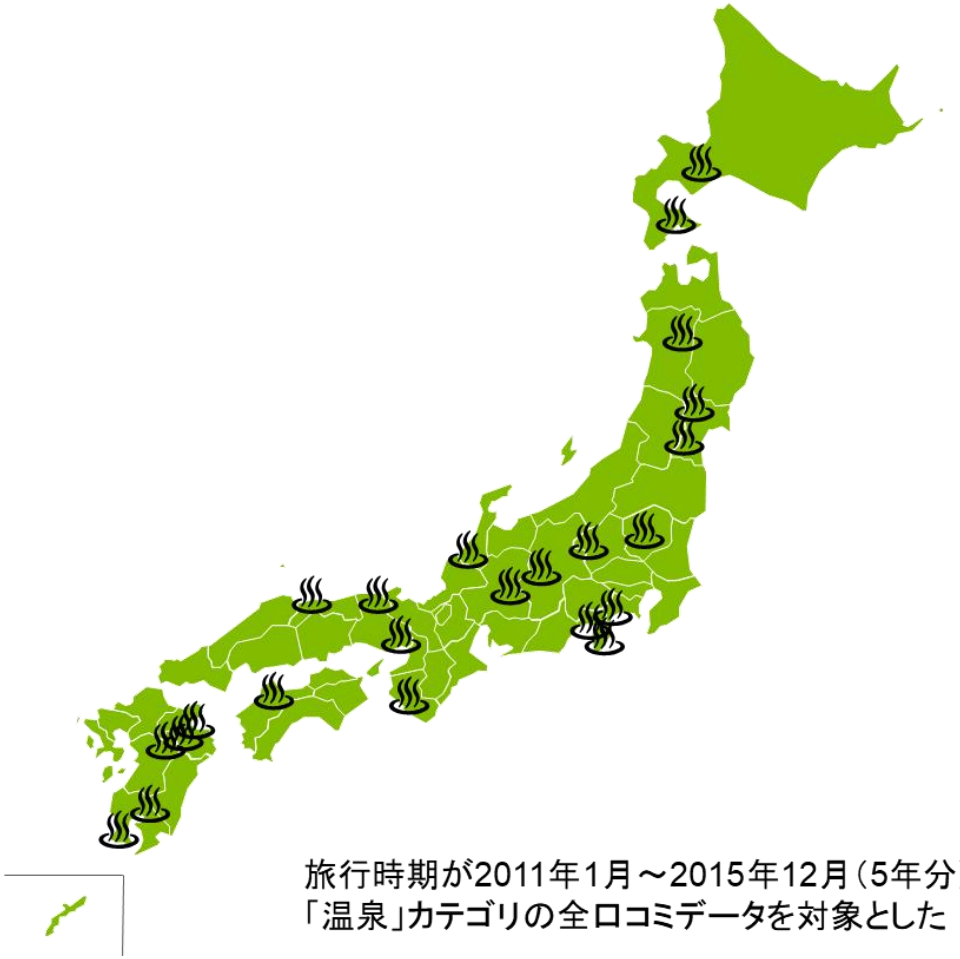
取得できる主な情報

口コミタイプ	施設所在地	施設所在地域
タイトル	コメント	評点（5点満点）
投稿者の性別	投稿者の年代	同行者
旅行時期（年月）	投稿日時	

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

分析の対象

全国2,562スポットの温泉の口コミデータ12,564件を分析します



旅行時期が2011年1月～2015年12月（5年分）で、
「温泉」カテゴリの全口コミデータを対象とした

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

Nomolyticsを適用した口コミ分析のプロセス

口コミのトピックを抽出し、そのトピックをベースに地域や観光客の特徴、満足度の要因を分析することで、地域観光におけるマーケティング戦略を観光客目線で検討します

トピックの抽出

口コミのコメントに**テキストマイニング**を実行して単語や係り受け表現を抽出

係り受け	頻度
泉質⇒良い	318
良い⇒温泉	279
雰囲気⇒良い	263
人⇒多い	221
湯⇒浸かる	159
硫黄⇒匂う	137
・・・	…

PLSAの適用

口コミのトピックを抽出

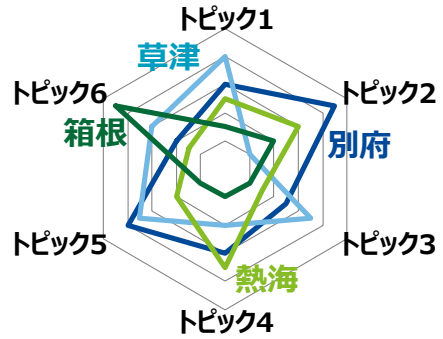
トピックのスコアリング

全口コミデータに対する各トピックの該当度を計算

口コミID	施設	地域	評点	性別	年代	同行者	トピック1	トピック2	トピックK
1	おかだ	箱根	5.0	女性	30代	友人	5.2	0.9	・・・
2	西の河原	草津	4.0	男性	40代	一人	3.6	4.8	・・・
3	吉祥の湯	別府	3.5	男性	50代	家族	1.2	2.3	・・・
n	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・

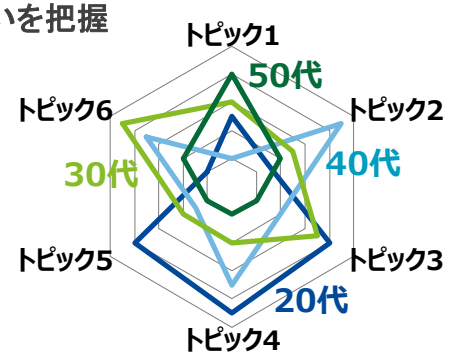
地域の特徴の分析

地域別に各トピックのスコアを集計し、地域の特徴を観光客の目線で把握



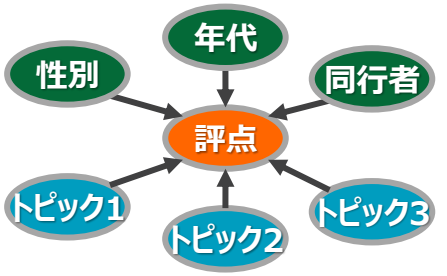
観光客の価値観分析

投稿者の属性別に各トピックのスコアを集計し、観光客の興味の対象の違いを把握



満足度の要因分析

ベイジアンネットワークで投稿者の属性やトピックと評価得点との関係をモデル化し、満足度を高める条件を把握



テキストマイニングとPLSAを応用して口コミのトピックを抽出します

テキストマイニングの実行 → 共起行列の作成 → PLSAの実行 → トピックの抽出

係り受け分析

（名詞⇔動詞・形容詞）

係り受け表現	頻度
泉質⇒良い	318
良い⇒温泉	279
雰囲気⇒良い	263
人⇒多い	221
湯⇒浸かる	159
硫黄⇒匂う	137
景色⇒良い	126
駐車場⇒広い	122

抽出した係り受け表現とそれを構成する単語に基づいて、

「単語×係り受け表現」

の共起行列を作成する

（文章単位の共起頻度を集計）

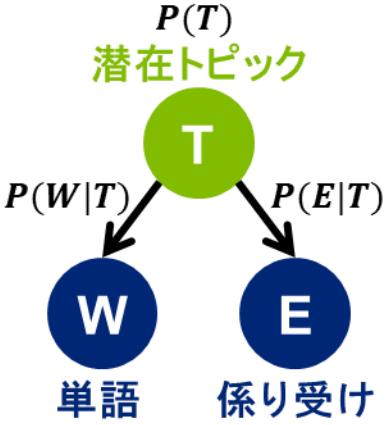
単語＼係り受け表現	泉質⇓良い	良い⇓温泉	雰囲気⇓良い	人⇓多い	…
泉質	318	166	81	22	
良い	318	279	263	36	
多い	45	32	20	221	
景色	85	98	196	14	
…					

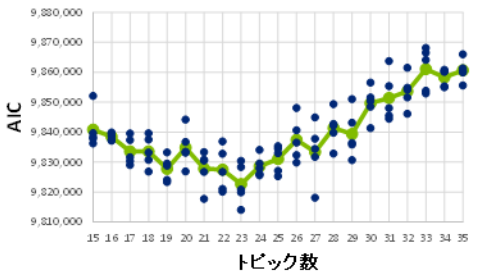
単語：1,648語

係り受け：4,808表現

※係り受けは頻度5件以上を対象

共起行列にPLSAを適用する





情報量基準AICを用いて最適なトピック数を選択する

各トピックについて以下の3つの確率が計算される

①P(T)
潜在トピックの割合

②P(W|T)
各トピックと単語との関係の強さ

③P(E|T)
各トピックと係り受けの関係の強さ

P(T4) = 5.1%

P(W|T)	単語
12.5%	雰囲気
6.8%	建物
6.5%	温泉地
4.1%	きれい
3.0%	古い
2.5%	共同浴場
1.7%	新しい
1.7%	風情
1.4%	銭湯
1.2%	旅館

P(E|T)	係り受け
3.0%	雰囲気⇒良い
1.1%	良い⇒雰囲気
1.0%	落ち着く⇒雰囲気
1.0%	レトロ⇒雰囲気
0.9%	雰囲気⇒味わう+できる
0.8%	清掃⇒行き届く
0.8%	建物⇒古い
0.7%	温泉地⇒ある
0.7%	古い⇒建物
0.7%	野趣⇒あふれる

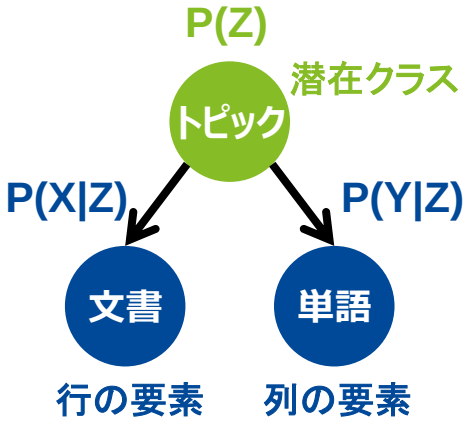
P(W|T)とP(E|T)という2つの軸の要素の重みからトピックの意味を解釈する

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

共起行列の構成の工夫

PLSAのインプットとする共起行列の構成を「文書×単語」ではなく「単語×係り受け」とすることで、要素間の違いが出やすくなり、解釈のしやすいトピックを抽出できます

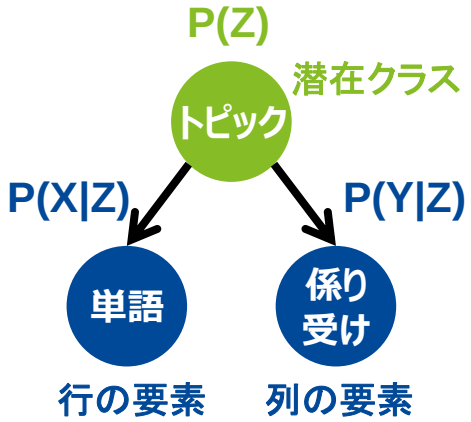
一般的なPLSAの共起行列



	単語1	単語2	単語3	単語4	・・・
文書ID:1	1	1	0	0	
文書ID:2	0	0	2	0	
文書ID:3	0	0	0	1	
文書ID:4	2	0	1	0	
・・・					

- 共起行列はBag-of-Wordsによる単語の頻度で構成され、**ほとんどが"0"となる疎なデータであるため、要素間の違いが現れにくく、クリアなトピックを抽出しにくい**
- PLSAのトピックには行の要素と列の要素が同時に所属し、両方の情報軸からトピックの意味を解釈できるが、一方の軸（行）は文書IDという意味性の低い情報で、トピックの解釈に使用しにくい

NomolyticsでのPLSAの共起行列



	係り受けa	係り受けb	係り受けc	係り受けd	・・・
単語1	325	264	11	20	
単語2	241	201	6	8	
単語3	28	41	288	14	
単語4	9	15	4	172	
・・・					

- 共起行列はクロス集計型の行列で、**単語と係り受けの共起頻度が入った密なデータであるため、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**
- 行と列が単語と係り受けで構成されている共起行列では、どちらもそれ単独で意味を持つ情報となるため、両方の情報軸からトピックの意味を解釈することができ、解釈の容易性が高まる

口コミトピックの一覧

12,564件の温泉の口コミは23個のトピックに集約されました

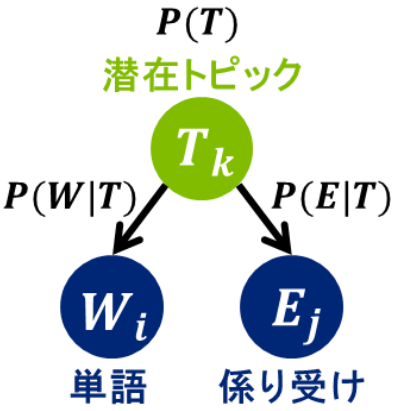
※P(E|T) > 0.05
の係り受け表現を
表示している

※係り受けの文字の大きさはそのトピックに対する関係の強さを表現している （緑色は上位5つの係り受け表現）

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

ロコミごとに23個のトピックのスコアを計算して変数化します

各ロコミデータのトピックスコアの算出



文章単位に各トピックのスコア（関連度）を計算し、ロコミ単位にそれを集約する

文章S_hのトピックT_kのスコア
（事後確率÷事前確率）

$$\frac{P(S_h|T_k)}{P(S_h)}$$

$S_h = \{E_1, E_2, \cdots, E_J\}$	文章S_hを構成する係り受け表現E_j
$P(S_h\|T_k) = \sum_j P(S_h\|E_j)P(E_j\|T_k)$	トピックT_kを条件とした文章S_hの出現確率（トピックT_kと文章S_hの関係の強さ）
$P(S_h\|E_j) = \frac{1}{n(E_j)}$	係り受けE_jが出現する中で文章S_hが出現する確率（E_jの出現文章数の逆数）
$P(S_h) = \sum_k P(S_h\|T_k)P(T_k)$	文章S_hの出現確率

トピックスコア算出プロセス

①文章ごとにスコアを計算

ロコミID	文章ID	T01	T02	T03	・・・	T23
1	1	3.1	0.9	2.0		1.1
1	2	1.4	0.2	5.5		2.4
2	1	0.8	5.8	1.3		0.9
2	2	1.2	3.2	1.7		1.0
2	3	0.6	1.8	2.6		1.6
・・・				※文章数：99,938件		

②ロコミごとに文章スコアを集約（最大値を採用）

ロコミID	T01	T02	T03	・・・	T23
1	3.1	0.9	5.5		2.4
2	1.2	5.8	2.6		1.6
・・・			※ロコミ数：12,564件		

③スコア３以上にフラグを立てる

ロコミID	T01	T02	T03	・・・	T23
1	1	0	1		0
2	0	1	0		0
・・・	※実際のコメントを確認して閾値を設定				

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

トピックのフラグデータの作成

全口コミデータに対して各トピックのスコア(該当有無)を計算することで、トピックをベースとした様々な集計・分析を実行することができます

トピックのスコア（フラグ情報）を紐づけた口コミデータ

口コミID	施設名称	施設所在地域	コメント	評点	性別	年代	同行者	旅行時期	トピック1	トピック2	・・・	トピック23
1	湯の里おかだ	箱根	日帰り温泉で利用しました。広々とした露天風呂は絶景です。	5.0	女性	30代	友人	2012年2月	1	0		1
2	西の河原露天風呂	草津	緑に囲まれた広い湯船の露天風呂で、ゆっくりできます。	4.0	男性	40代	一人旅	2014年8月	1	1		0
3	やまなみの湯	別府	車で立ち寄りやすくて手軽です。別府だけにお湯はいいです。	3.5	男性	50代	家族	2013年6月	0	0		1
4	吉祥の湯	黒川	お風呂は露天風呂が5つ。紅葉の時期でとても綺麗でした。	4.5	女性	40代	カップル・夫婦	2012年11月	1	0		0
・・・												
12,564	家康の湯	熱海	駅前にある足湯ですが、こんなに広い足湯は初めてです。	3.0	女性	20代	一人旅	2015年9月	0	1		1

②観光客の価値観分析

③満足度の要因分析

①地域の特徴分析

各温泉地はどのトピックが多く話題にされているか定量的に把握できます

温泉地別のトピック該当割合

全体の1.1倍以上　全体の1.4倍以上　全体の1.7倍以上　全体の2.0倍以上

全体 (12,564)	トピック		別府 (599)	道後 (449)	城崎 (233)	指宿 (198)	草津 (158)	熱海 (154)	有馬 (124)	山梨 (119)	箱根 (111)	諏訪 (110)	黒川 (109)	由布院 (104)
23%	T01	露天風呂からの眺め	14%	8%	16%	30%	17%	14%	2%	71%	37%	13%	19%	19%
16%	T02	湯船の特徴	14%	12%	17%	5%	32%	5%	13%	18%	19%	28%	8%	12%
10%	T03	入浴受付の説明	12%	12%	11%	11%	8%	5%	8%	16%	11%	11%	6%	12%
26%	T04	雰囲気の良さ	31%	42%	48%	14%	32%	26%	23%	17%	39%	31%	53%	37%
26%	T05	訪問した時間帯	21%	42%	23%	30%	25%	16%	29%	50%	38%	15%	33%	9%
25%	T06	アクセス	21%	12%	24%	23%	21%	30%	16%	20%	38			
28%	T07	良い温泉、疲れの癒し	19%	20%	20%	24%	25%	19%	20%	38%	43			
23%	T08	設備の広さ・充実性	14%	18%	26%	15%	19%	6%	22%	20%	41			
21%	T09	安価、無料入浴	20%	22%	20%	9%	27%	13%	23%	18%	15%	31%	15%	21%
18%	T10	目的地までのルート	18%	12%	9%	15%	16%	28%	10%	20%	25%	12%	20%	23%
23%	T11	泉質・湯量	24%	10%	15%	10%	34%	22%	39%	14%	19%	36%	21%	19%
19%	T12	温泉に入ること	22%	32%	21%	22%	28%	5%	25%	29%				21%
23%	T13	自己・他社の利用	23%	31%	14%	18%	23%	17%	35%	21%				18%
21%	T14	湯の温度	20		%	17%	37%	19%	23%	30%				23%
23%	T15	足湯	2		%	18%	37%	34%	27%	13%	20%	40%	31%	27%
22%	T16	宿泊でも日帰りでも楽しめる	2		%	14%	20%	12%	39%	13%	31%	15%	40%	25%
19%	T17	体を温めること	18%	14%	13%	43%	18%	8%	19%	33%	18%	19%	12%	12%
22%	T18	砂湯、有名な温泉	41%	41%	20%	73%	22%	22%	22%	27%	26%			19%
19%	T19	充実した食事・休憩所	18%	15%	6%	13%	8%	7%	21%	19%	40%			13%
22%	T20	綺麗な施設、家族で楽しめる	15%	11%	20%	9%	29%	7%	24%	18%	36%	25%	21%	13%
18%	T21	人の多さ	20%	26%	20%	16%	20%	16%	29%	22%	14%	18%	24%	26%
17%	T22	アメニティの有無	19%	31%	12%	44%	17%	7%	11%	14%	19%	16%	10%	20%
19%	T23	立ち寄る温泉	16%	11%	15%	27%	15%	12%	15%	18%	30%	18%	14%	14%

雰囲気の良さ

宿泊でも日帰りでも楽しめる

砂湯、有名な温泉

人の多さ

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

【分析②】 観光客の価値観分析

各観光客層はどのトピックに関心が強いのか定量的に把握できます

観光客属性別のトピック該当割合

全体の1.1倍以上　全体の1.2倍以上

全体 (12,564)	トピック		女性 20代 (577)	女性 30代 (1,733)	女性 40代 (1,554)	女性 50代 (762)	女性 60代 (192)	男性 20代 (337)	男性 30代 (1,623)	男性 40代 (2,509)	男性 50代 (2,172)	男性 60代 (903)	同行者 乳幼児連れ家族旅行 (349)	同行者 家族旅行 (1,817)	同行者 カップル・夫婦 (2,436)	同行者 カップル・夫婦(シニア) (543)	同行者 友人 (1,408)	同行者 一人旅 (3,548)
23%	T01	露天風呂からの眺め	24%	25%	26%	25%	28%	20%	23%	20%	20%	24%	26%	25%	24%	28%	27%	20%
16%	T02	湯船の特徴	12%	14%	17%	13%	18%	9%	18%	20%	15%	17%	14%	13%	17%	16%	14%	19%
10%	T03	入浴受付の説明	9%	10%	10%	11%	6%	7%	11%	13%	8%	10%	9%	8%	10%	12%	11%	11%
26%	T04	雰囲気の良さ	27%	27%	26%	26%	31%	24%	27%	24%	24%	25%	30%	27%	28%	33%	25%	25%
26%	T05	訪問した時間帯	32%	30%	28%	29%	21%	31%	24%	24%	24%	18%	31%	27%	27%	23%	29%	23%
25%	T06	アクセス	23%	25%	21%			28%	30%	26%							26%	30%
28%	T07	良い温泉、疲れの癒し	26%	32%	31%			28%	28%	26%							28%	25%
23%	T08	設備の広さ・充実性	27%	29%	23%			22%	23%	20%							26%	21%
21%	T09	安価、無料入浴	20%	22%	22%	20%	19%	18%	21%	23%	17%	23%	23%	18%	20%	27%	21%	22%
18%	T10	目的地までのルート	14%	14%	17%	18%	18%	16%	20%	20%	19%	21%	13%	17%	17%	18%	16%	21%
23%	T11	泉質・湯量	14%	21%	23%	21%	31%	14%	24%	25%	25%	30%	20%	22%	26%	29%	21%	25%
19%	T12	温泉に入ること	18%	20%	21%			14%	19%	20					21%			18%
23%	T13	自己・他社の利用	26%	27%	24%			21%	23%	21					23%			23%
21%	T14	湯の温度	24%	25%	23%			17%	22%	21					23%			21%
23%	T15	足湯	21%	24%	24%	25%	28%	19%	27%	23%	23%	19%	2	26%	24%		23%	24%
22%	T16	宿泊でも日帰りでも楽しめる	24%	24%	22%	21%	29%	17%	23%	20%	20%	25%	26%	25%	24%	32%	22%	19%
19%	T17	体を温めること	22%	22%	21%	23%	18%	15%	18%	20%	17%	16%	20%	20%	21%	21%	21%	18%
22%	T18	砂湯、有名な温泉	27%	25%	25%	24%	24%	23%	21%	21%	19%	18%	23%	24%	23%	27%	23%	20%
19%	T19	充実した食事・休憩所	23%	23%	22%	21%	13%	15%	18%	19%	16%	18%	28%	19%	21%	20%	21%	18%
22%	T20	綺麗な施設、家族で楽しめる	26			23%	13%	18%	21%	23%	19%						%	21%
18%	T21	人の多さ	20			19%	20%	17%	19%	15%	20%						%	18%
17%	T22	アメニティの有無	21			20%	13%	14%	13%	16%	13%						%	15%
19%	T23	立ち寄る温泉	22%	21%	18%	20%	17%	15%	19%	19%	18%	18%	18%	20%	19%	19%	22%	18%

60代女性

乳幼児連れ家族旅行

シニア夫婦

60代女性

60代男性

家族旅行（乳幼児連れ含む）

シニア夫婦

女性全般（特に20代）

家族旅行

シニア夫婦

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

温泉地の特徴トピックに関心を持つ観光客層をターゲットに設定できます

別府・由布院・黒川の温泉地の特徴と関心のある観光客層

※写真は"九州旅ネッ"フォトギャラリーより転載

	別府温泉	由布院温泉	黒川温泉
			
			
温泉地の特徴	■ T18：砂湯、有名な温泉	■ T04：雰囲気の良さ ■ T21：人の多さ	■ T04：雰囲気の良さ ■ T16：宿泊でも日帰りでも楽しめる
ターゲット層	■ 女性（特に20代） ■ カップル・夫婦（シニア） ■ 家族旅行	■ 60代女性 ■ カップル・夫婦（シニア） ■ 乳幼児連れ家族旅行	■ 60代 ■ カップル・夫婦（シニア） ■ 家族旅行（乳幼児連れ含む）

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

満足度に効く要因の構造を口コミトピックや属性情報からモデル化します

ベイジアンネットワークによる満足度モデルの構築

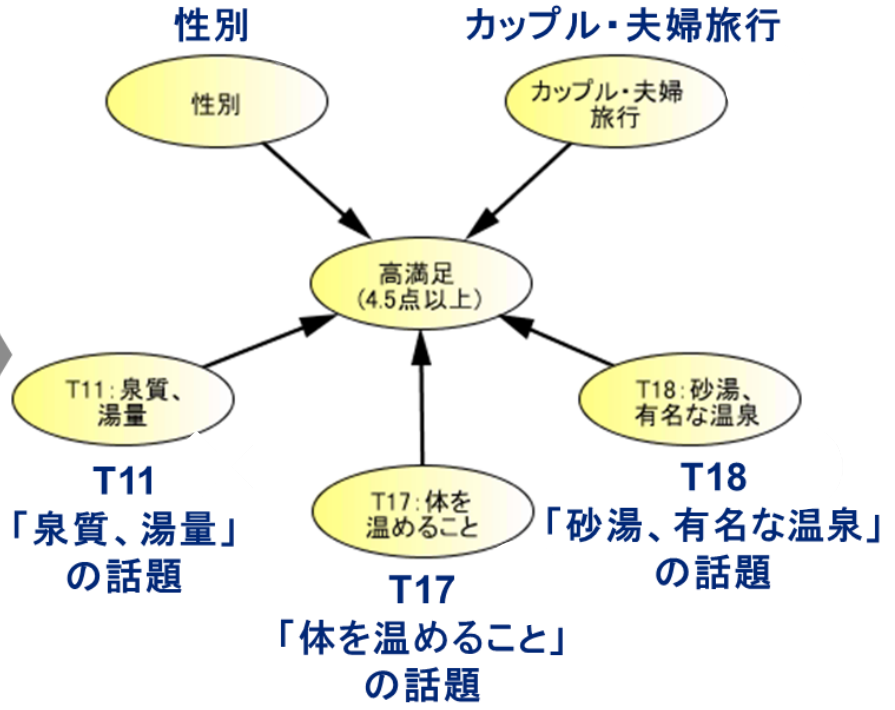
これらの要因の中から高満足に影響を与える要因はどれか？

高満足（4.5点以上）

ベイジアンネットワーク

高満足には5つの要因が影響する

※口コミの評点4.5点以上を高満足とする



条件のセット例

性別	カップル・夫婦旅行	T11:泉質、湯量	T17:体を温めること	T18:砂湯、有名な温泉
女性	Yes	No	Yes	Yes

シミュレーション

結果

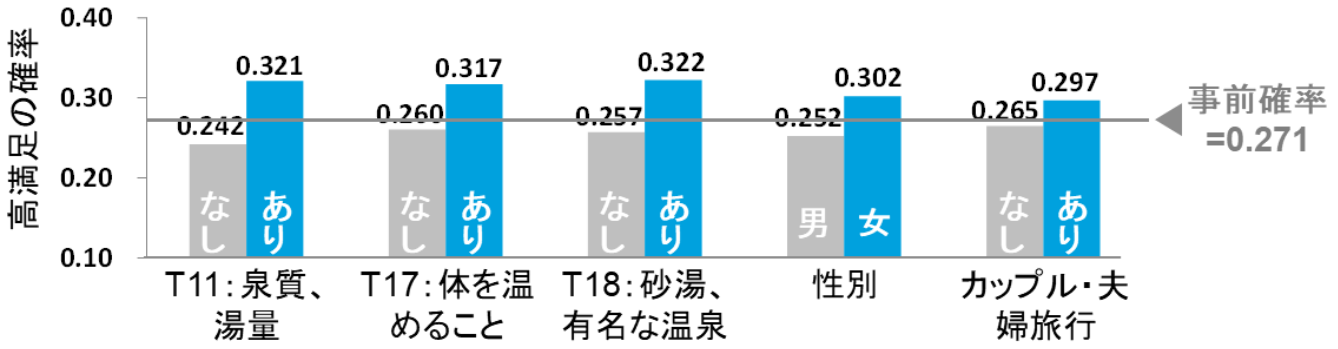
高満足（4.5点以上）
39.6%

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

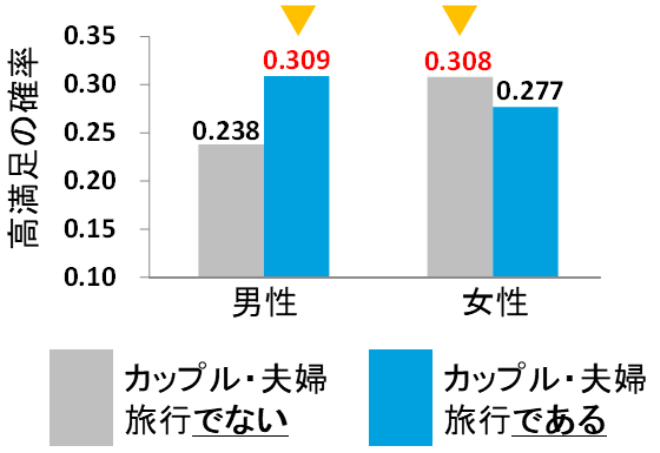
【分析③】 各要因の条件による満足度の確率シミュレーション

観光客の満足度を効果的に高める条件を定量的に把握できます

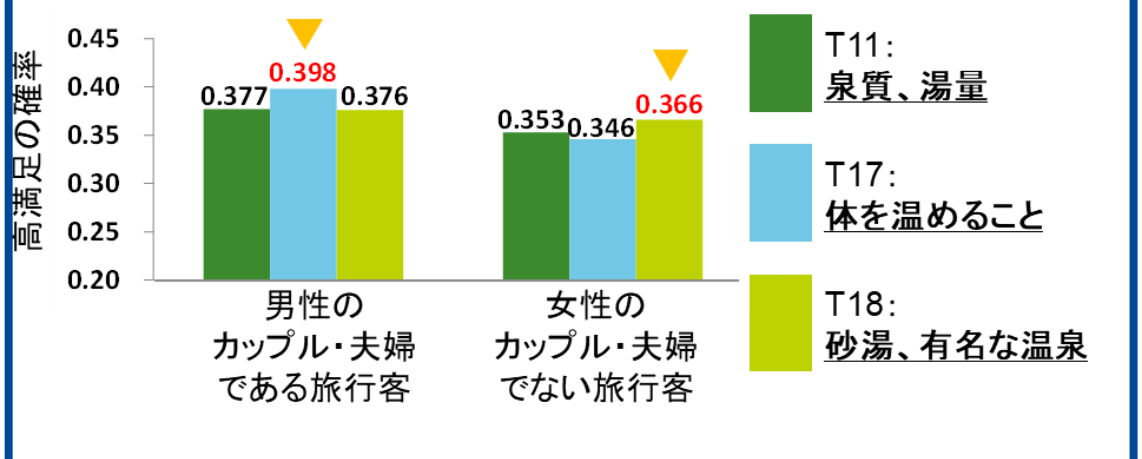
一つの変数を条件にした高満足の確率



属性を条件にした高満足の確率



性別・同行者別でトピックを条件にした高満足の確率



※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

【分析③の活用】 観光プロモーションの検討

ターゲットに応じた効果的なプロモーションを検討できます

※一部写真は"九州旅ネット"フォトギャラリーより転載

男性向けには・・・

女性向けには・・・

※「数理システムユーザーコンファレンス2016（http://www.msi.co.jp/userconf/2016/）」の有限責任監査法人トーマツ（野守耕爾）の講演資料より転載

5. Nomolyticsの分析事例②

特許文書データの分析による企業の技術戦略の検討

「風」「空気」に関する10年分の特許データ30,039件の要約に記載されている【課題】と【解決手段】の文章を分析します

データの抽出条件と抽出結果

■ 対象
- 公開特許公報

■ キーワード
- 要約と請求項に「風」と「空気」を含む

■ 出願年
- 2006年1月1日～2015年12月31日

■ 抽出方法
- PatentSQUAREを使用

■ 抽出結果
- 30,039件






分析データの加工

■ 要約文の【課題】と【解決手段】に記載されている文章をそれぞれ抽出する
- このような書式で記載されていないものは要約文をそのまま使用する

■ 出願人情報は名寄せをし、グループ会社などは統一する

課題の文章

【要約】【課題】ユーザーの快適性を維持しつつ、省エネ運転を行うことができる空気調和機を提供すること。【解決手段】本発明の空気調和機は、室内温度を検出する室内温度検出手段と、人体の活動量を検出する人体検出手段と、基準室内設定温度を設定するリモコン装置30とを備え、室内温度が基準室内設定温度となるように空調制御を行う空気調和機であって、人体検出手段で検出する活動量が所定の活動量以内であるときは、室内温度が、基準室内設定温度を補正した補正室内設定温度となるように空調を行い、補正室内設定温度よりも低い状態を継続すると、圧縮機を停止させ、圧縮機の復帰は、基準室内設定温度に基づいて行う。

解決手段の文章

Nomolyticsを適用した特許分析のプロセス

特許要約の【課題】と【解決手段】の文章から用途と技術のトピックを抽出し、各トピックのトレンドや出願人の特徴分析、また用途と技術の関係分析によって技術戦略を検討します

トピックの抽出

特許の要約文にテキストマイニングを実行して単語や係り受け表現を抽出

単語	品詞	頻度
空気調和機	名詞	3,106
空気	名詞	2,846
容易な	形容詞	2,790
抑制する	動詞	2,687
塵埃	名詞	1,687
分離する	動詞	1,231
・・・	・・・	…

PLSAの適用

【課題】の要約文から用途トピックを抽出

【解決手段】の要約文から技術トピックを抽出

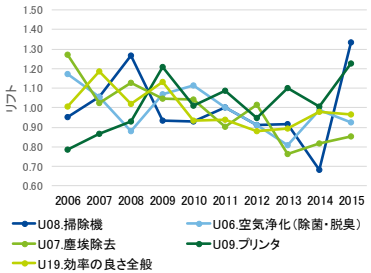
トピックのスコアリング

全特許データに対する各トピックの該当度を計算

ID	出願年	出願人	用途トピック1	用途トピック2	用途トピック**	技術トピック1	技術トピック2	技術トピック**
1	2014	A社	2.1	0.6	…	1.5	5.0	…
2	2013	B社	0.3	3.4	…	4.6	0.9	…
3	2011	C社	4.8	2.2	…	2.7	1.1	…
n	・・・	・・・	…	…	…	…	…	…

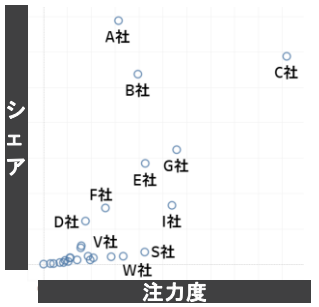
トレンドの分析

トピックのスコアを出願年で集計することでトピックのトレンドを把握



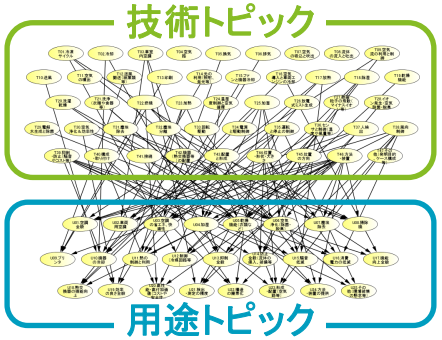
競合他社の分析

トピックのスコアを出願人で集計することで各社の特徴やポジションを把握



用途と技術の関係分析

用途トピックと技術トピックの関係性をベイジアンネットワークでモデル化



用途⇒技術 の分析

検討中の用途を実現する上で重要な技術と各社の動向を把握し、事業化のための技術戦略を検討する

技術⇒用途 の分析

自社技術と関係する用途のうち未想定の用途を発見し、技術の新規用途展開のアイデアを創出する

トピック抽出のアプローチ

テキストマイニングで単語と係り受け表現を抽出し、「単語×係り受け」の共起行列にPLSAを適用することで、単語と係り受けの出現の背後にある潜在トピックを抽出します

テキストマイニングの実行 → 共起行列の作成 → PLSAの実行 → トピックの抽出

【課題】と【解決手段】の文章に含まれる単語と係り受けを抽出する

単語	品詞	頻度
空気調和機	名詞	3,106
空気	名詞	2,846
容易	名詞	2,790
抑制	名詞	2,687
良い	形容詞	2,481
向上	名詞	2,328
防止	名詞	2,047
発生	名詞	2,005
…	…	…

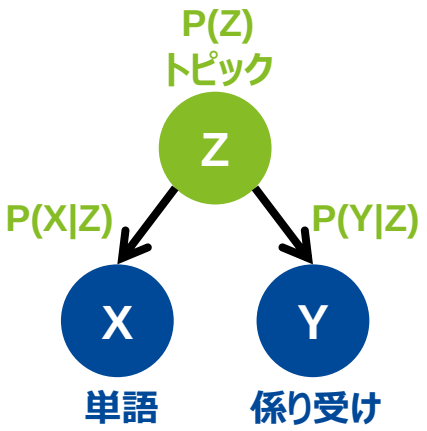
係り受け表現	頻度
空気調和機⇒提供	1,575
効率⇒良い	1,325
車両用空調装置⇒提供	578
掃除機-提供	545
容易-構成	539
画像形成装置-提供	334
抑制-提供	296
向上-図る	279
…	…

抽出した単語と係り受け表現に基づいて、「単語×係り受け」の共起行列（文章単位で同時に出現する頻度のクロス集計表）を作成する

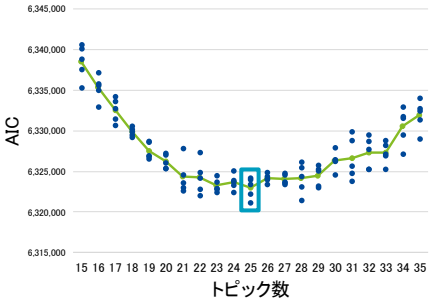
単語 \ 係り受け表現	空気調和機⇓提供	効率⇓良い	車両用空調装置⇓提供	掃除機⇓提供	…
空気調和機	1578	100	4	1	
空気	85	144	45	50	
容易	100	105	51	67	
抑制	142	95	64	63	
…					

共起行列の構成（それぞれ頻度10件以上を対象）
課題：単語(3,256語)×係り受け(2,084表現)
解決手段：単語(5,187語)×係り受け(7,174表現)

共起行列にPLSAを適用する



トピック数を幅を持たせて設定し、各トピック数に対してPLSAを初期値を変えて5回ずつ実行して情報量基準AICを計算し、AIC最小の解を採用する



各トピックについて以下の3つの確率が計算される

① P(X|Z)
トピックにおける単語の所属確率

② P(Y|Z)
トピックにおける係り受けの所属確率

③ P(Z)
トピックの存在確率

トピックにおけるP(X|Z)とP(Y|Z)からトピックの意味を解釈する

トピック T32
P(Z) = 2.7%

P(X|Z)	単語
5.5%	送風機
5.2%	塵埃
4.1%	掃除機
3.6%	分離
3.5%	吸い込む
2.3%	集塵部
1.9%	配置
1.9%	集塵容器
1.6%	旋回
1.5%	含塵空気
…	…

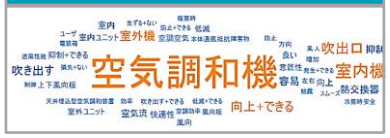
P(Y|Z)	係り受け
2.1%	塵埃-分離
1.7%	分離-塵埃
1.7%	塵埃-含む
1.5%	吸い込む-塵埃
1.3%	含む-空気
1.0%	空気-分離
1.0%	送風機-吸い込む
1.0%	発生-送風機
0.9%	含塵空気-分離
0.9%	備える-掃除機
…	…

確率の高い構成要素から、トピックT32は「塵埃の分離」に関するトピックと解釈できる

用途トピック25個の一覧

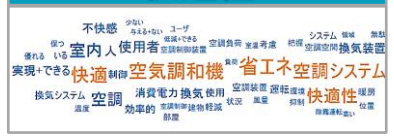
【課題】の文章からは、空調や加湿、空気清浄、掃除機、プリンタ、機器冷却、騒音や消費電力の低減、構造の簡素化などの用途が25個抽出されました

U01.空調全般

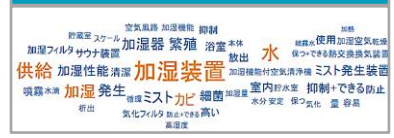


U02.車両用空調

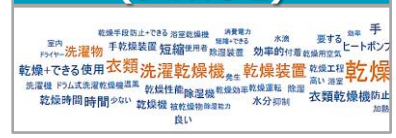
U03.空調の省エネ、快適性



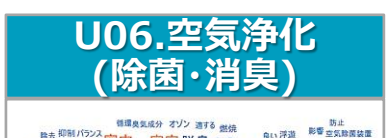
U04加湿



U05.乾燥機能(衣類など)



U06.空気浄化(除菌・消臭)



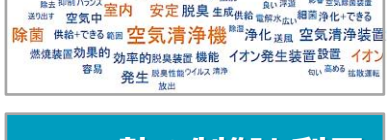
U07.塵埃除去

U08.掃除機

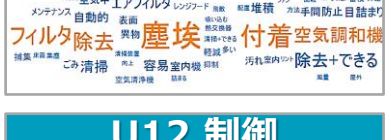
U09.プリンタ

U10.機器の冷却

U11.熱の制御と利用



U12.制御(冷媒回路等)



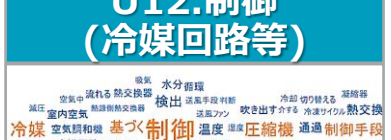
U13.抑制全般

U14.防止全般(流体の侵入、破損等)

U15.騒音低減

U16.消費電力の低減

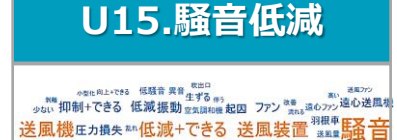
U17.機能向上全般



U18.熱交換器の機能向上

U19.効率の良さ全般

U20.価値(コストや安全性など)



U21.検出・測定の精度

U22.構造の簡素化

U23.形成・配置(空気路等)

U24.方法・装置の提供

U25.その他(環境破壊の懸念等)

※文字の大きさはトピックに対する関係の強さを表現している
（上位5つの単語を赤色で表示している）

技術トピック47個の一覧①

【解決手段】の文章からは、空気の冷却や換気、放熱、除湿、乾燥、加湿、イオン生成、空気清浄、塵埃分離、センサと制御、構成や配置などの技術が47個抽出されました

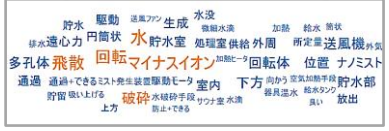
※文字の大きさはトピックに対する関係の強さを表現している
（上位5つの単語を赤色で表示している）

技術トピック47個の一覧②

【解決手段】の文章からは、空気の冷却や換気、放熱、除湿、乾燥、加湿、イオン生成、空気清浄、塵埃分離、センサと制御、構成や配置などの技術が47個抽出されました

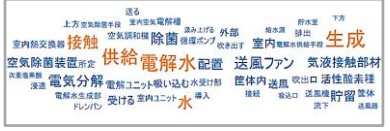
T26.放電式ミスト生成

T27.微細粒子の飛散(マイナスイオン等)

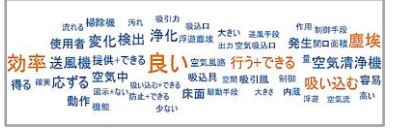


T28.イオン発生・空気除菌・脱臭

T29.電解水生成と除菌

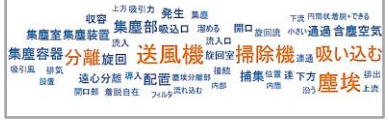


T30.塵埃吸込＆効率性

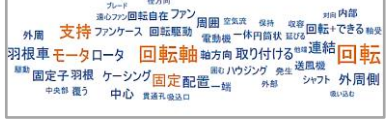


T31.塵埃除去

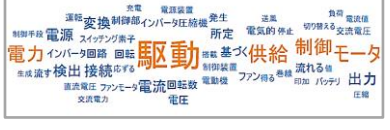
T32.塵埃分離



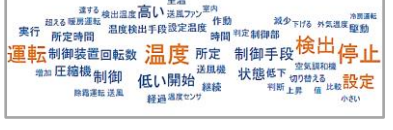
T33.回転駆動



T34.電源と駆動制御

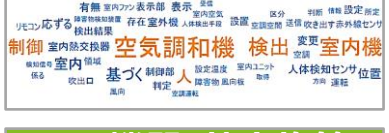


T35.運転と停止の制御

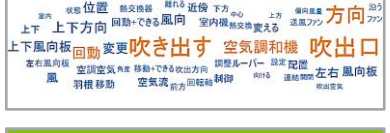


T36.センサと制御(温度や風量等)

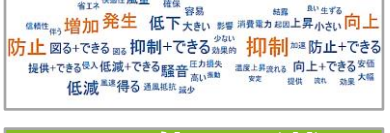
T37.人検出



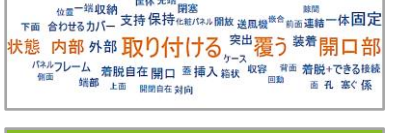
T38.風向制御



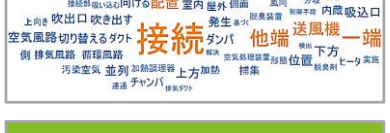
T39.抑制・防止(騒音やコスト等)



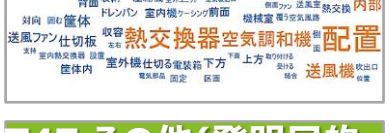
T40.構成・取り付け



T41.接続



T42.機器(熱交換等)の配置

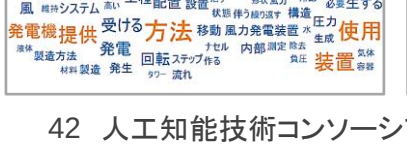


T43.配置と形成

T44.位置・形状・大きさ

T45.位置の方向

T46.方法・装置



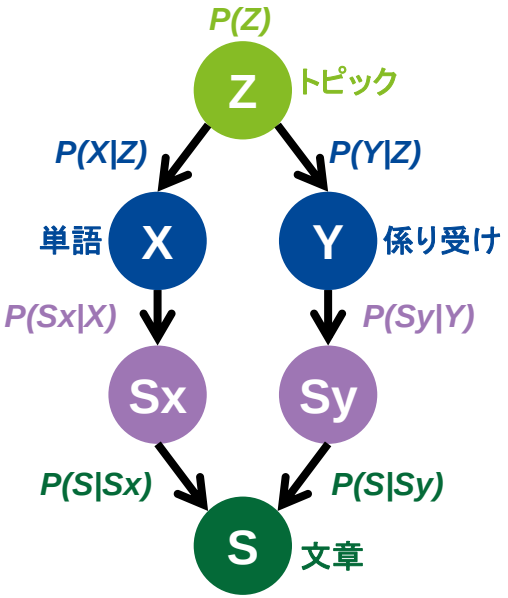
T47.その他(発明目的、ケース構成等)

※文字の大きさはトピックに対する関係の強さを表現している
（上位5つの単語を赤色で表示している）

文章単位に各トピックのスコア（該当度）を計算し、それを特許ID単位に集約し、最終的には閾値を設定して{1:該当有, 0:該当無}のデータに変換します

文章単位のスコア	$\frac{P(S\|Z)}{P(Z)}$

- リフト値（事後確率÷事前確率）
- 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す



文章を単語で定義される文章Sxと係り受けで定義される文章Syを設定し、それぞれトピックとの関係を計算し、最終的にそれらを一つに統合する

単語X_iで定義される文章Sx_h

$$Sx_h = \{X_1, X_2, \cdots, X_i\}$$

トピックZ_kを条件とした文章Sx_hの出現確率

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

単語X_iが出現する中で文章Sx_hが出現する確率
（X_iの出現文章数の逆数）

$$P(Sx_h|X_i) = 1/n(X_i)$$

係り受けY_jで定義される文章Sy_h

$$Sy_h = \{Y_1, Y_2, \cdots, Y_j\}$$

トピックZ_kを条件とした文章Sy_hの出現確率

$$P(Sy_h|Z_k) = \sum_j P(Sy_h|Y_j)P(Y_j|Z_k)$$

係り受けY_jが出現する中で文章Sy_hが出現する確率
（Y_jの出現文章数の逆数）

$$P(Sy_h|Y_j) = 1/n(Y_j)$$

トピックZ_kを条件とした文章S_hの出現確率
※$P(S_h|Sx_h)$と$P(S_h|Sy_h)$はともに1/2とする

$$P(S_h|Z_k) = P(S_h|Sx_h)P(Sx_h|Z_k) + P(S_h|Sy_h)P(Sy_h|Z_k)$$

文章S_hの出現確率

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

トピックスコア算出プロセス

①文章ごとにスコアを計算

特許ID	文章ID	T01	T02	T03	・・・	T47
1	1	3.1	0.9	2.0		1.1
1	2	1.4	0.2	5.5		2.4
2	1	0.8	5.8	1.3		0.9
2	2	1.2	3.2	1.7		1.0
2	3	0.6	1.8	2.6		3.6
・・・						

②特許IDごとに文章スコアを集約

※最大値を採用する

特許ID	T01	T02	T03	・・・	T47
1	3.1	0.9	5.5		2.4
2	1.2	5.8	2.6		3.6
・・・					

③閾値を設定してフラグに変換する

※閾値は3に設定する

特許ID	T01	T02	T03	・・・	T47
1	1	0	1		0
2	0	1	0		1
・・・					

トピックのフラグデータの作成

全特許データに対して各トピックのスコア(該当有無)を計算することで、トピックをベースとした様々な分析を実行することができます

トピックのスコア(フラグ情報)を紐づけた特許データ

特許ID	出願番号	出願年	出願人	要約文		用途トピックU01	用途トピックU02	・・・	用途トピックU25	技術トピックT01	技術トピックT02	・・・	技術トピックT47
				【課題】	【解決手段】								
1	特願2006-XXXX	2006	A社	空気調和機の高外気	吸気口から導入された	1	0		0	0	1		0
2	特願2009-XXXX	2009	B社	短時間で除霜を行うこ	着霜検出手段が室外	0	1		0	1	0		0
3	特願2011-XXXX	2011	C社	乾燥運転が中断され	通風路を通して回転槽	0	0		1	1	0		0
4	特願2013-XXXX	2013	D社	ウインドシールドの防	車両用空調装置の空	0	1		0	0	1		1
・・・	・・・	・・・	・・・			…	…		…	…	…		…
30039	特願2012-XXXX	2012	Z社	プリ空調時に、除菌ま	冷暖房空調ユニットは	0	1		0	1	1		0

①出願年の分析

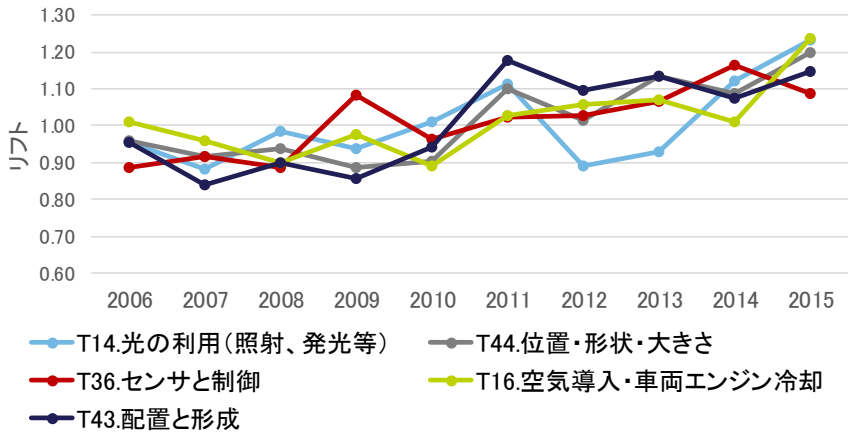
②出願人の分析

③用途と技術の関係分析

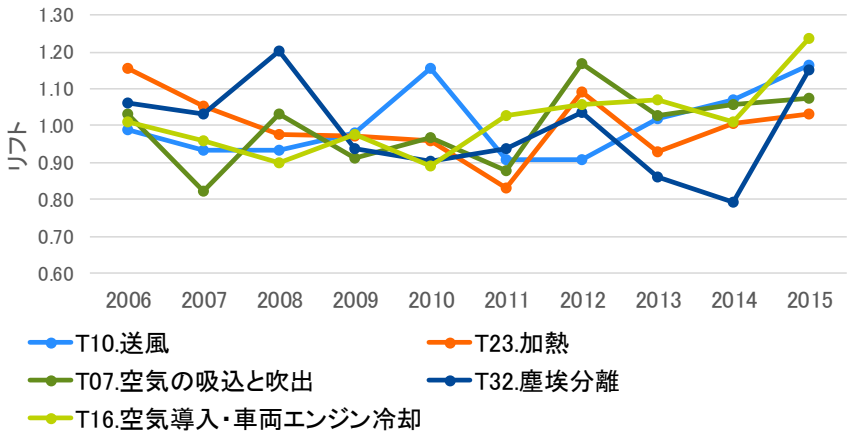
【①出願年の分析】 技術トピックの上昇トレンド

短期的には塵埃分離や車両エンジンの冷却に関する技術が、長期的にはプロジェクタなどの光の利用に関する技術が上昇しています

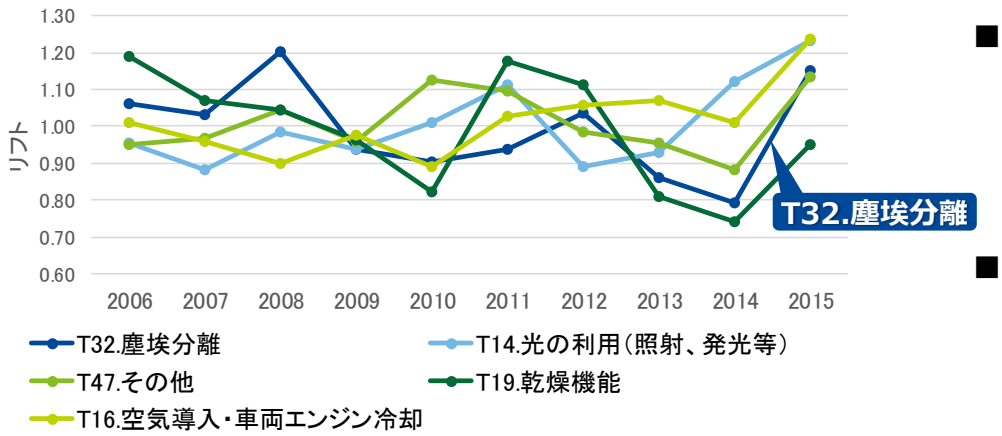
【長期】2006年からの上昇率 best5



【中期】2011年からの上昇率 best5



【短期】2013年からの上昇率 best5



集計の仕方

■ リフト値を出願年・トピックごとに集計

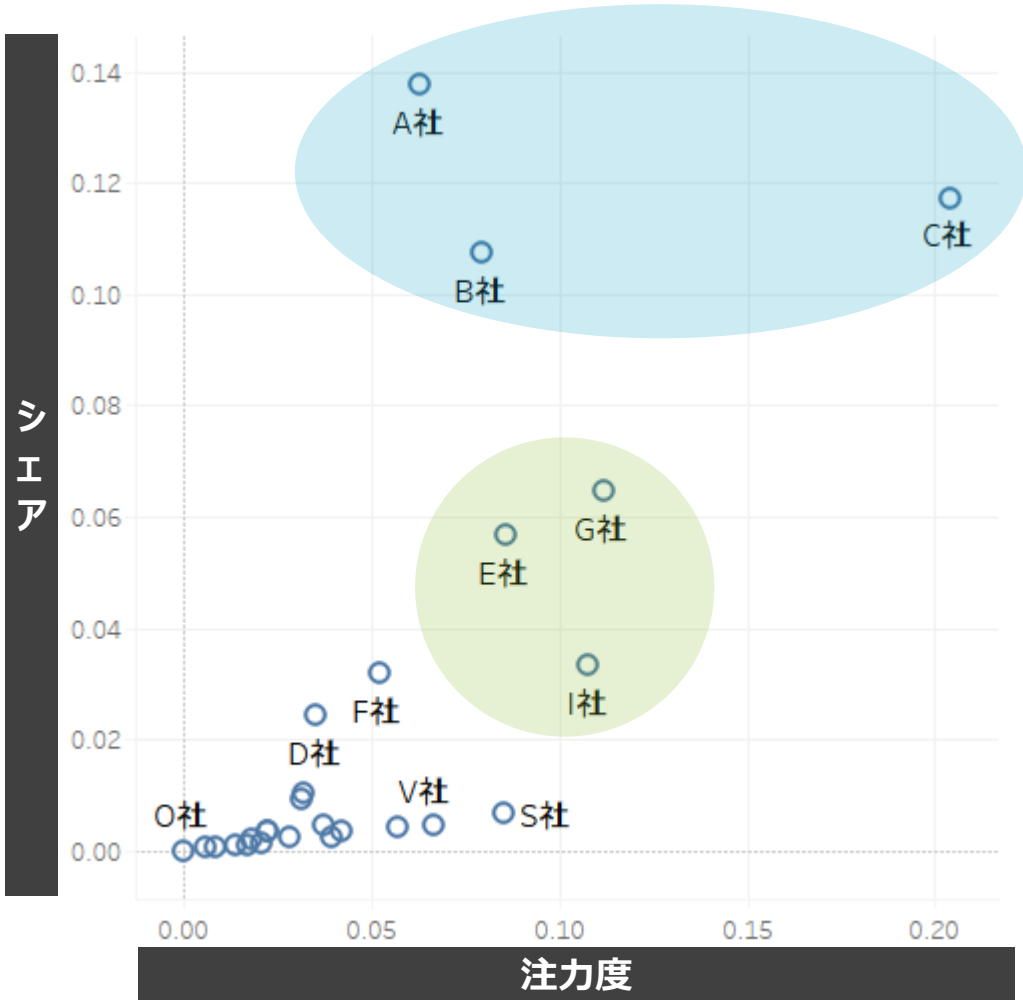
$$\frac{P(\text{出願年} \mid \text{トピックTx=1})}{P(\text{出願年})}$$

■ その出願年の出願件数割合を平均（=1）として標準化した値

【②出願人の分析】 技術「T32.塵埃分離」の各社のポジショニング

塵埃分離に関する技術は、3社のシェアが高いものの、他にもある程度のシェア・注力度を有する企業が何社か存在するため、連携によって競争力を高める動きも考えられます

注力度とシェアの散布図



考察と戦略の検討

- シェアではA社・B社・C社が高いが、特にC社は注力度がとても高く、特有の技術力を保有していると考えられる
- E社・G社・I社はシェアは中程度だが、注力度は比較的高く、技術力もあると思われる
- 高いシェアを持つ企業は、中程度のシェアの企業と連携することで、より技術力を高めながらシェアを伸ばすことが期待できる
- あるいは中程度シェアの企業の間で連携し、高シェアの業界大手に対抗することも考えられる

注力度とシェア

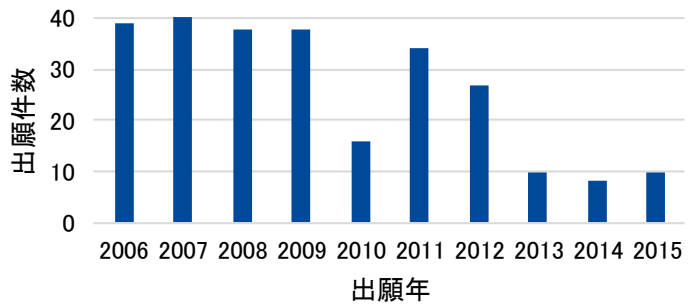
- **注力度**: $P(\text{トピック}T \mid \text{出願人}X)$
  - 出願人Xの出願特許の中で、どれくらいの割合がそのトピックTに該当するものか、つまり出願人がどれくらいそのトピックに注力しているのかを示す
- **シェア**: $P(\text{出願人}X \mid \text{トピック}T)$
  - トピックTが該当する特許の中で、どれくらいの割合がその出願人Xの出願によるものか、つまりトピックの中でどれくらいその出願人が占めているのかを示す

【②出願人の分析】 技術「T32.塵埃分離」の各社の出願推移

高シェアのA社とB社の近年の出願動向は、A社は減少ですがB社は増加し、注力度1位のC社は直近で出願が急増し、シェア4位のG社も出願を伸ばしており、今後に要注目です

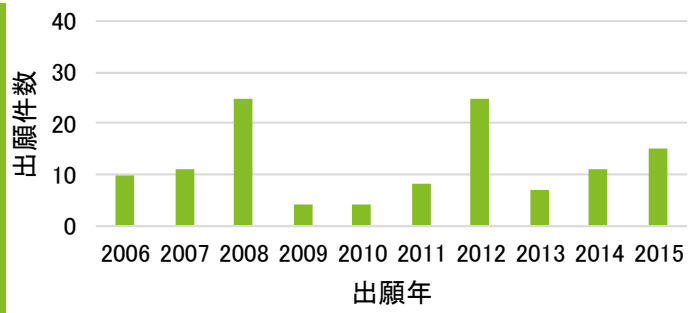
注目企業の出願件数の推移

A社



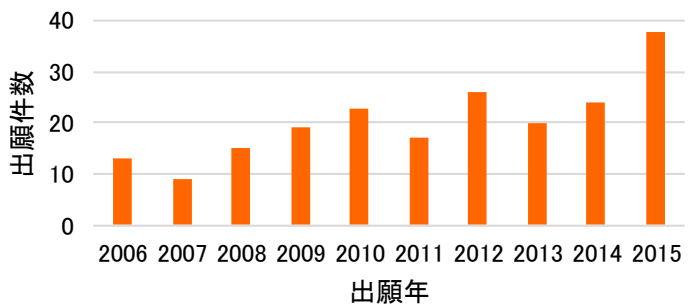
シェア1位のA社は、近年は出願が少ない

G社



シェア4位注力度2位のG社は、近年徐々に出願が増えている

B社

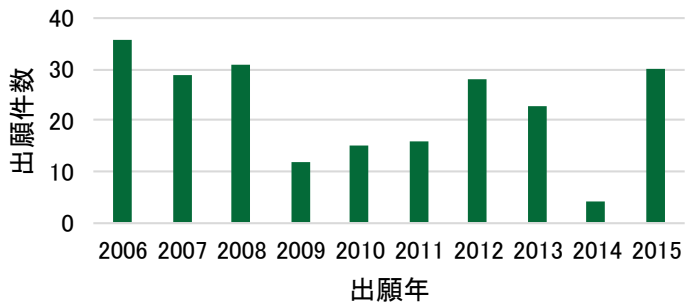


シェア3位のB社は、ここ10年で徐々に出願が増えている

E社

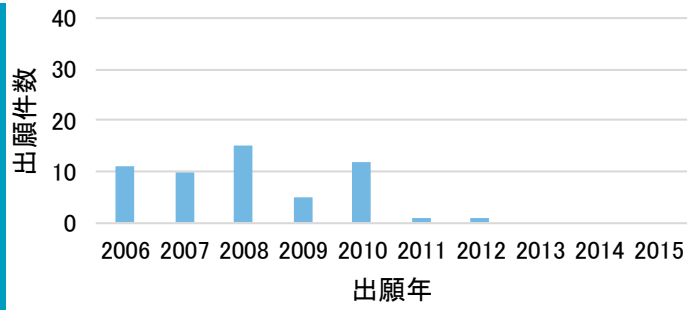
シェア5位注力度4位のE社は、近年では出願件数の変化が少ない

C社



注力度1位シェア2位のC社は、直近で出願が急激に増えている

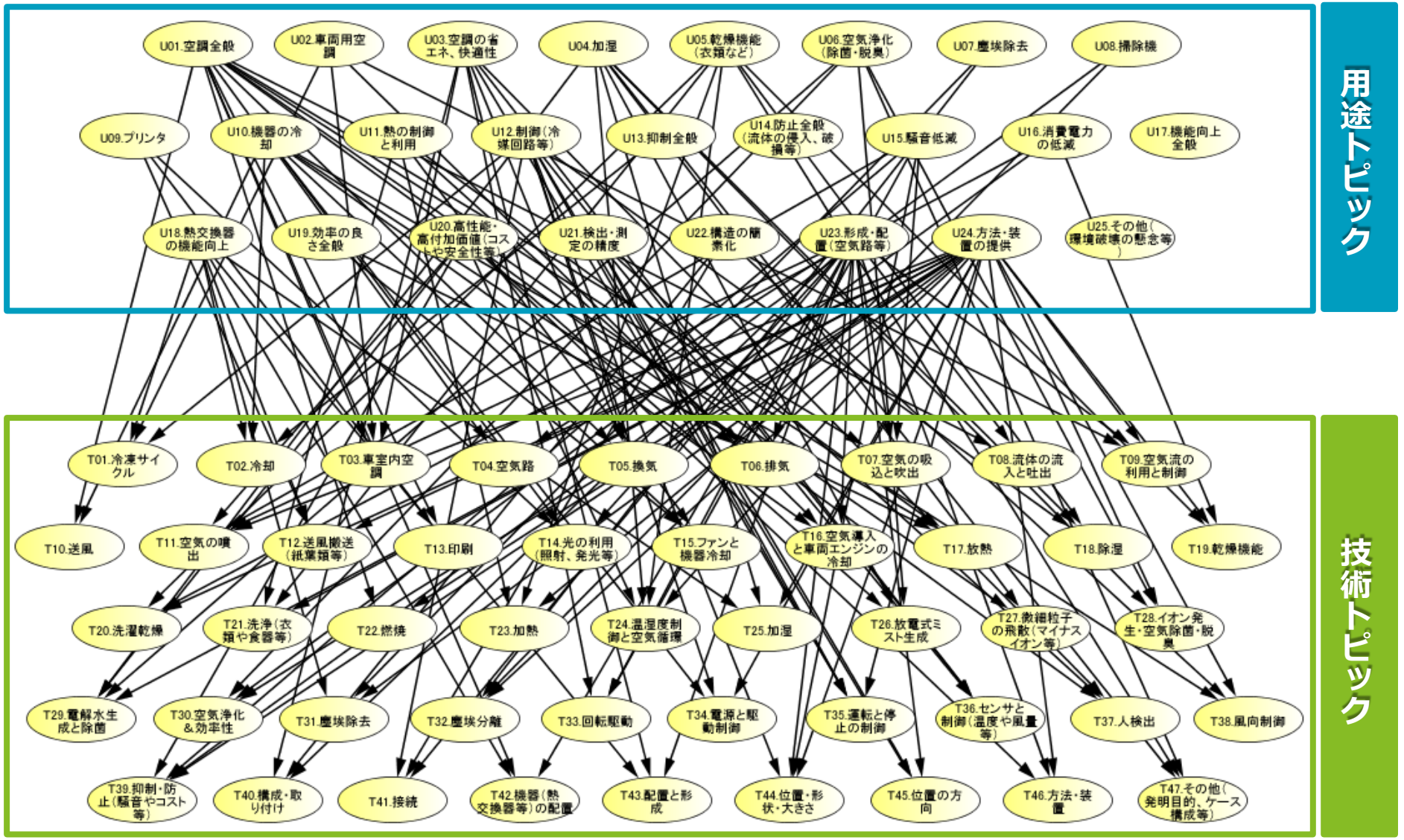
I社



シェア6位注力度3位のI社は、2013年以降の出願が存在しない

【③用途と技術の関係分析(その1)】 用途⇒技術の関係モデル

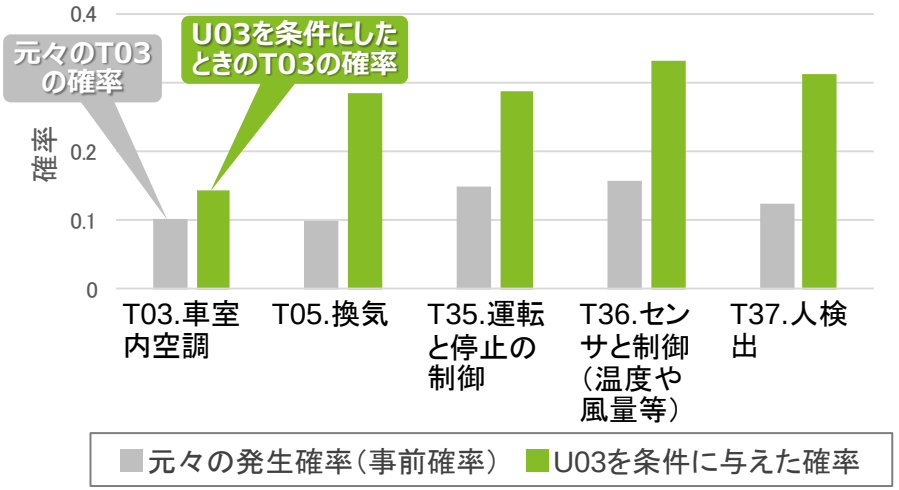
ベイジアンネットワークを適用して、用途トピックに対する技術トピックの確率的因果関係をモデル化します



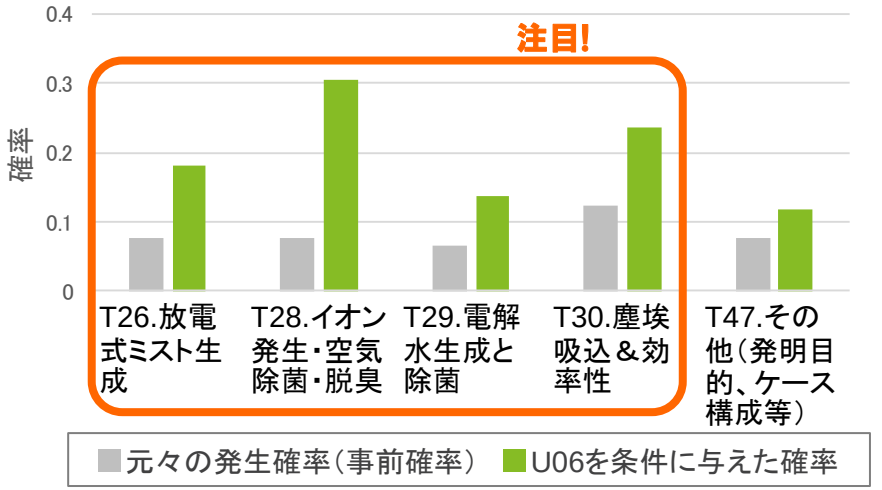
【③用途と技術の関係分析(その1)】 用途と関係のある技術の確認

ベイジアンネットワークによって、1つの用途トピックを条件に与えたときの各技術トピックの確率の変化をシミュレーションし、用途に対する技術の関連性の強さを確認します

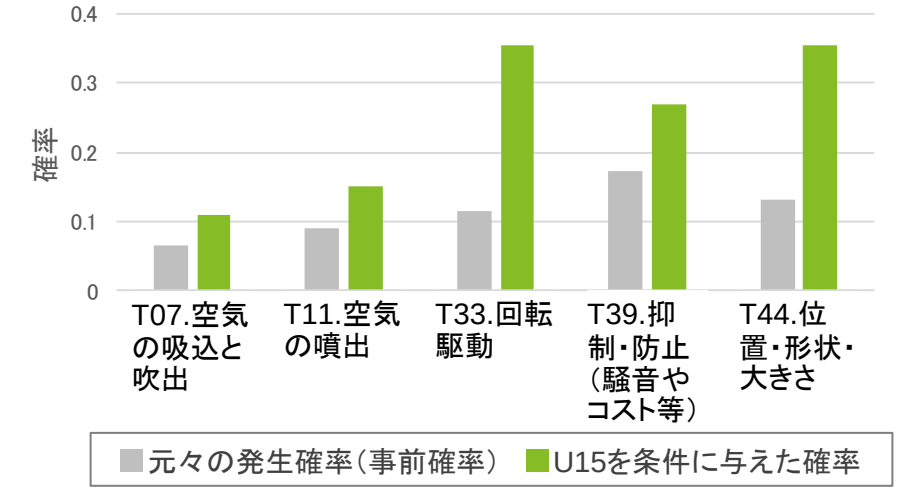
「U03.空調の省エネ、快適性」と関係のある技術



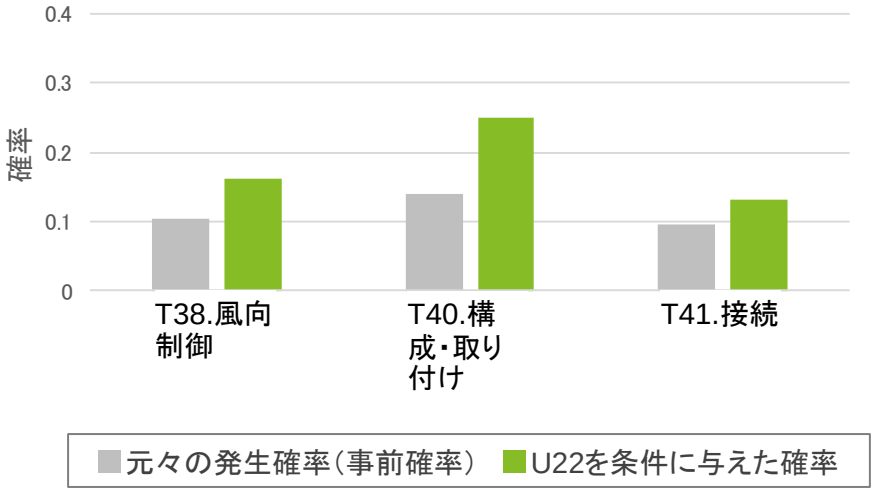
「U06.空気浄化」と関係のある技術



「U15.騒音低減」と関係のある技術



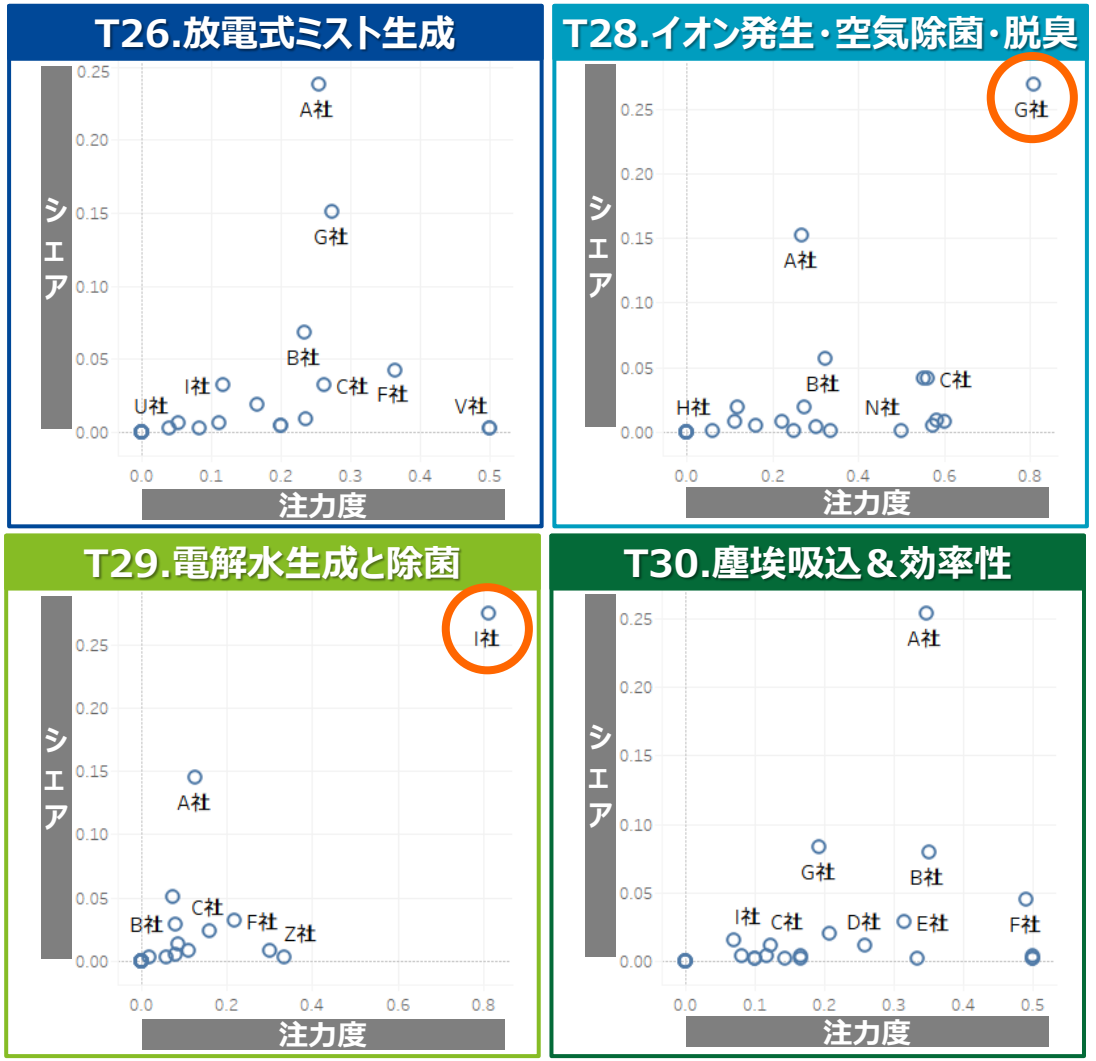
「U22.構造の簡素化」と関係のある技術



【③用途と技術の関係分析(その1)】 用途「U06.空気浄化」と関係する技術トピックの出願人動向

U06の用途と関係する4つの技術のうち2つは一強状態にあり、U06の事業化では、この技術を避けた他の技術の開発を検討する、あるいはその一強企業の買収も考えられます

「U06.空気浄化」の関係技術トピックにおける出願人マップ

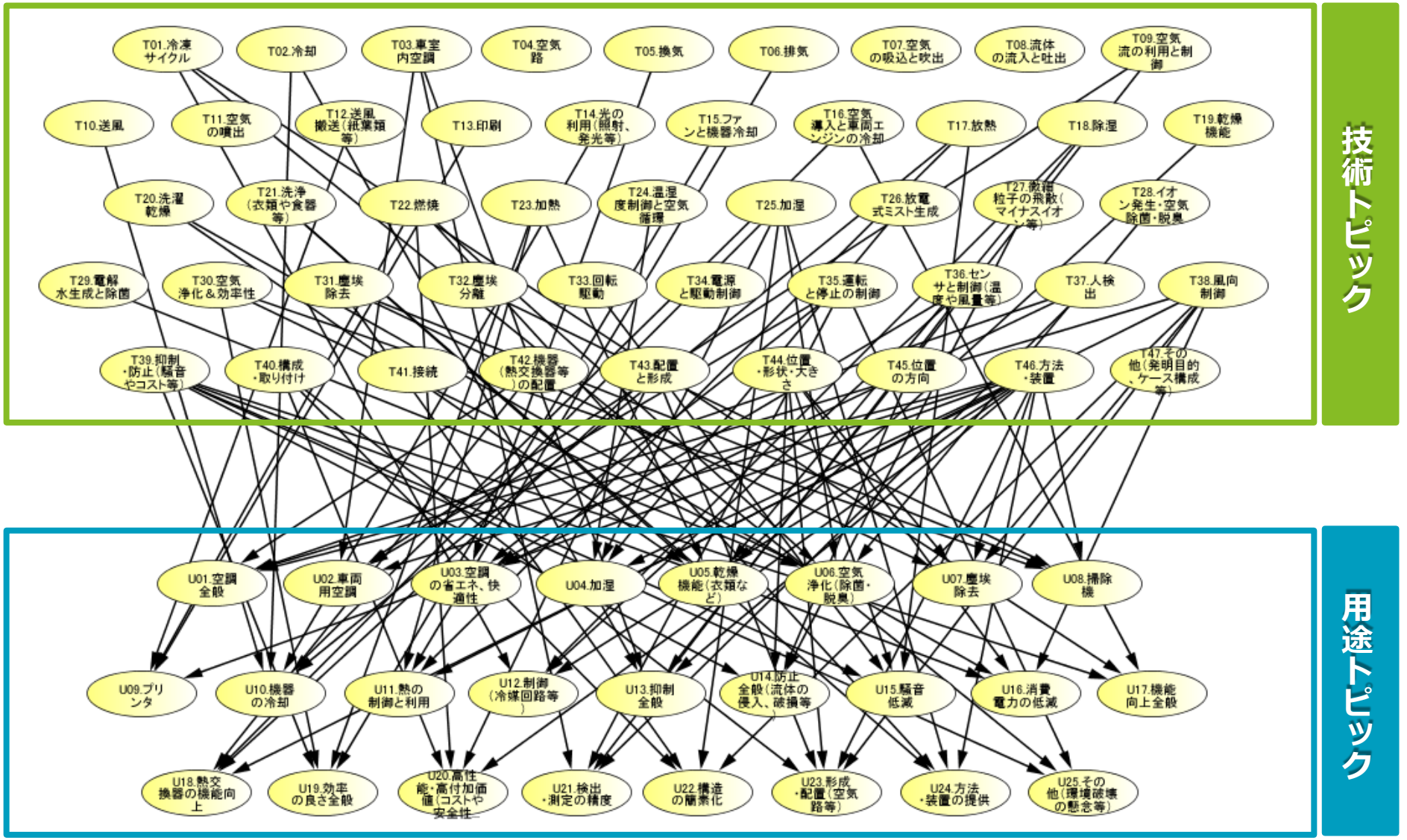


考察と戦略の検討

- 「T28.イオン発生・空気除菌・脱臭と「T29.電解水生成と除菌」は、それぞれG社とI社が高シェア高注力度のポジションを確立した一強状態の技術といえる
- 「T26.放電式ミスト生成」と「T30.塵埃吸込＆効率性」は、シェアではA社が高く、注力度では例えばF社が高いが、高シェア高注力度の右上のポジションは空いている
- 一強状態の技術を避けて「U06.空気浄化」の用途を実現する場合、T26やT30の技術の開発が狙い目といえるが、シェアの高いA社や注力度の高いF社の動向は要注目である
- 一強状態にあるT28やT29の技術において、その一強企業と提携あるいはM&Aを実現すれば、その技術領域ごと獲得できる

【③用途と技術の関係分析(その2)】 技術⇒用途の関係モデル

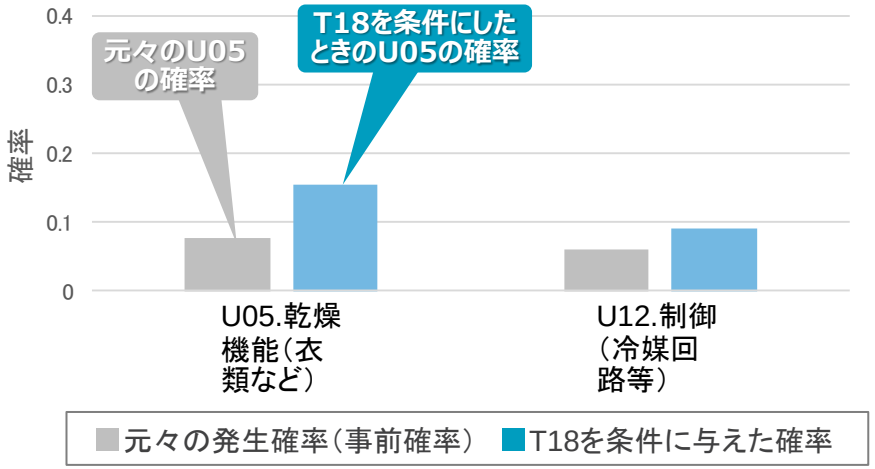
ベイジアンネットワークを適用して、技術トピックに対する用途トピックの確率的因果関係をモデル化します



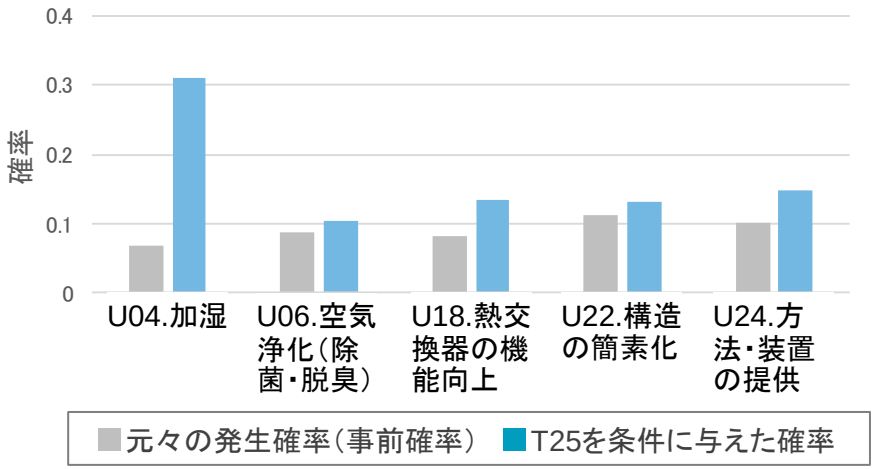
【③用途と技術の関係分析(その2)】 技術と関係のある用途の確認

ベイジアンネットワークによって、1つの技術トピックを条件に与えたときの各用途トピックの確率の変化をシミュレーションし、技術に対する用途の関連性の強さを確認します

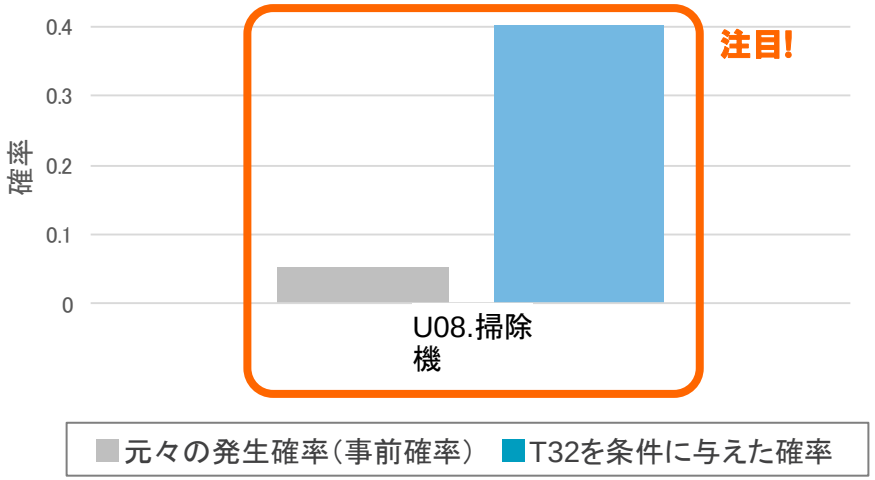
「T18.除湿」と関係のある用途



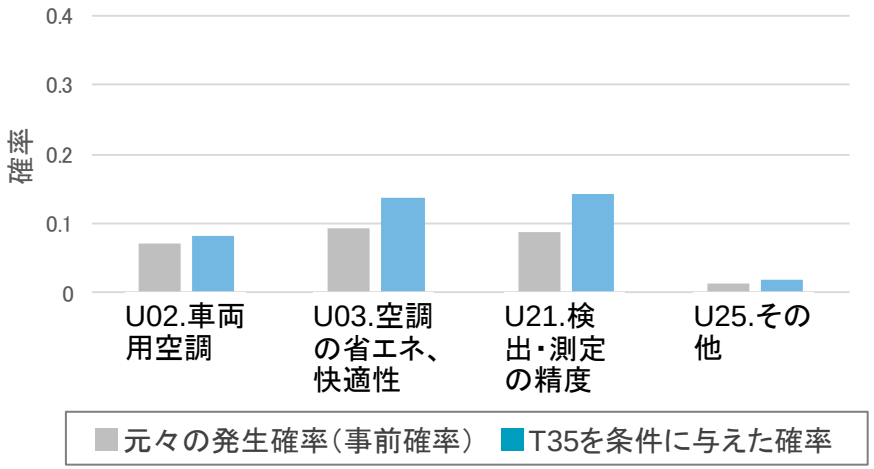
「T25.加湿」と関係のある用途



「T32.塵埃分離」と関係のある用途



「T35.運転と停止の制御」と関係のある用途



【③用途と技術の関係分析(その2)】 技術「T32.塵埃分離」の新規用途アイデア

印刷機でトナーを分離・回収するサイクロン部の清掃時期を判断して分離効率を維持する技術は、サイクロン掃除機の集塵部の集塵性能向上にも応用できるかもしれません

「掃除機」を想定した「塵埃分離」の特許例

発明の名称

電気掃除機

【課題】

集塵性能が向上しメンテナンスの軽減が図れる電気掃除機を提供すること。

【解決手段】

塵埃を含む空気を旋回させ塵埃分離する略円筒状の1次旋回室と、1次旋回室に連通した2次旋回室と、1次旋回室の下方に位置し塵埃を溜める集塵室と、塵埃を圧縮する圧縮板と、塵埃が流入する流入口を有し、圧縮板の底面の一部に突出部を流入口から見て集塵室の奥側に配設する構成としたことより、集塵室内に入った塵埃は、圧縮板の突出部に引っかかり動きが止められ、流れに乗って2次旋回室や1次旋回室側に戻ることが無いため集塵性能が向上し、排気筒の詰まり防止によるメンテナンスの軽減を図ることができる。

「掃除機」を想定していない「塵埃分離」の特許例

発明の名称

画像形成装置

【課題】

サイクロン部の清掃時期を適正に判断して、トナーの分離効率の低下を抑制することが可能な画像形成装置を提供する。

【解決手段】

画像形成装置は、トナー含有空気からトナーを遠心分離するサイクロン部と、サイクロン部によって分離されたトナーを回収する回収部と、サイクロン部によってトナーが分離された空気を通過させ、残留トナーを捕集するフィルタ部と、空気を吸引する送風部と、フィルタの汚れを検知する汚れ検知センサが設けられたトナー捕集部を備え、汚れ検知センサで検知されたフィルタの汚れから推定した風量と、風速センサで取得した風量の実測値の差分が、サイクロン清掃閾値を超えたと判断すると、サイクロン部の清掃モードを実行する。

※対外説明用のため要約文は一部加工している

これまで培ってきた技術や経験と関連のある用途をいかに発想できるかということがイノベーションの鍵になります

サイクロン掃除機

羽のない扇風機

空気清浄ファンヒーター

加湿器

ヘアドライヤー

ヘアスタイラー

ハンドドライヤー

6. まとめ

複数のAI技術を応用したテキストデータの分析手法：Nomolytics®

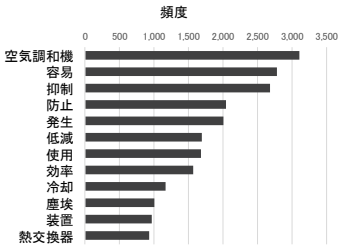
【 特許第6085888号 】

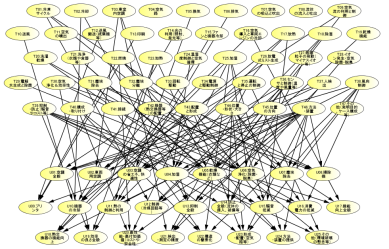
膨大なテキストデータをトピックに変換して解釈を容易にし、テキスト情報内に潜む要因関係をモデル化して、ビジネスアクションに有用な特徴を把握可能にします

Nomolytics : Narrative Orchestration Modeling Analytics

テキストマイニング	PLSA 確率的潜在意味解析	ベイジアンネットワーク
文章に含まれる単語を抽出し、その出現頻度を集計する	単語が出現する特徴を学習し、膨大な単語を複数のトピックにまとめる	トピックやその他属性情報など、テキスト情報内の要因関係をモデル化する

単語抽出 × *トピック類型化* × *要因関係分析*





Nomolyticsのメリット

膨大なテキストデータ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し類型化できる

テキスト情報に潜む要因関係を構造化し、特徴を見たいターゲットのキードライバを発見できる

条件を変化させたときの効果を確率的にシミュレーションでき、有効なアクションを検討できる

VOCデータの分析例　（Web口コミ、アンケートの自由記述、コールセンターの問い合わせなど）

VOCデータにNomolyticsを適用することで、そのコメント内容をトピック化して各属性の傾向や要因関係を分析でき、顧客目線の業務改善やマーケティング戦略を検討できます

属性×トピックの1対1
の関係を集計・可視化

A VOCのトピック化

VOCのテキストデータにテキストマイニング×PLSAを適用することで、コメントの内容を複数のトピックに機械的に類型化し、大量のコメントの全体像を把握します

属性×トピックの複数間
の関係構造をモデル化

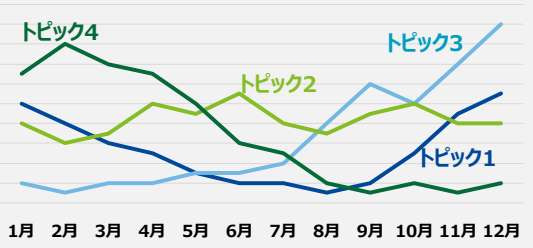
B トピック×属性の傾向分析

顧客属性別、商品属性別に各トピックの関連度を計算することで、各属性におけるコメントの傾向を把握します

	トピック1	トピック2	トピック3	トピック4
20代30代	0.6	1.2	**4.3**	0.9
40代50代	**2.4**	1.5	1.6	**2.3**
60代70代	**3.8**	1.1	0.4	1.2
商品属性1	0.5	**3.4**	1.2	**2.2**
商品属性2	**4.1**	0.8	0.3	1.0

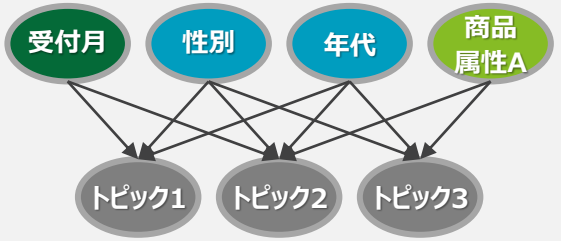
C トピックのトレンド分析

時系列情報と各トピックの関連度を計算することで、トピックのトレンドや季節性を把握します



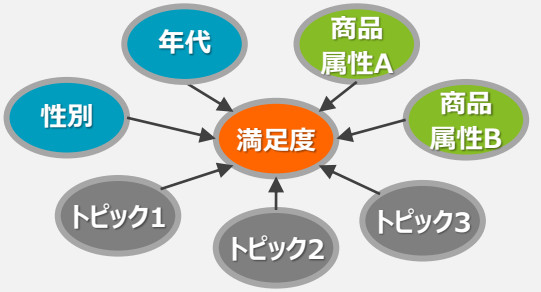
D トピックの要因関係の分析

各トピックに影響を与える要因を各属性情報（顧客属性、商品属性、時間属性など）から探索し、そのトピックが発せられやすい要因条件を把握します



E 評価の要因関係の分析

データに満足度や解約などの評価項目があれば、その評価に影響を与える要因をトピックや属性から探索し、各要因条件からの評価の予測や制御する条件を把握します



特許文書データの分析例

特許文書データにNomolyticsを適用することで、特許文書の内容をトピック化し、トレンドや出願人の動向を分析したり、用途と技術の関係分析によって技術戦略を検討できます

出願年・出願人×トピックの特徴分析

A 特許文書のトピック化

特許の要約にある【課題】と【解決手段】の文章にテキストマイニング×PLSAを適用することで、課題からは用途に関するトピックを、解決手段からは技術に関するトピックを機械的に抽出し、大量の特許の全体像を把握します

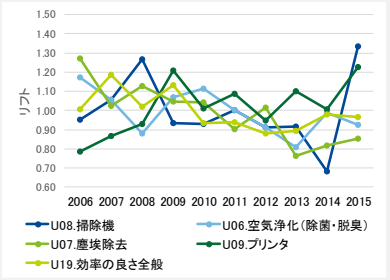
用途のトピック

技術のトピック

用途と技術の関係分析

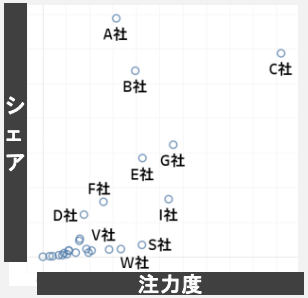
B トレンドの分析

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ます



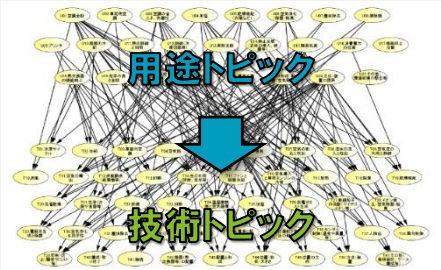
C 競合他社の分析

出願人×トピックの関係を分析することで、各社の特徴やポジションを把握します



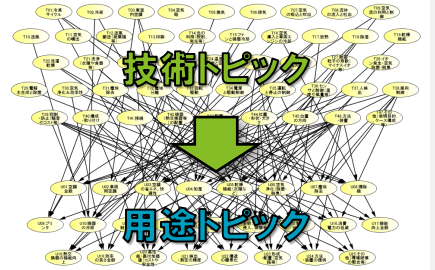
D 用途⇒技術の関係分析

用途に対する技術の関係を分析することで、ある用途を実現する上で重要な技術と各社の出願動向を把握し、自社の開発戦略や他社との協業可能性を探ります



E 技術⇒用途の関係分析

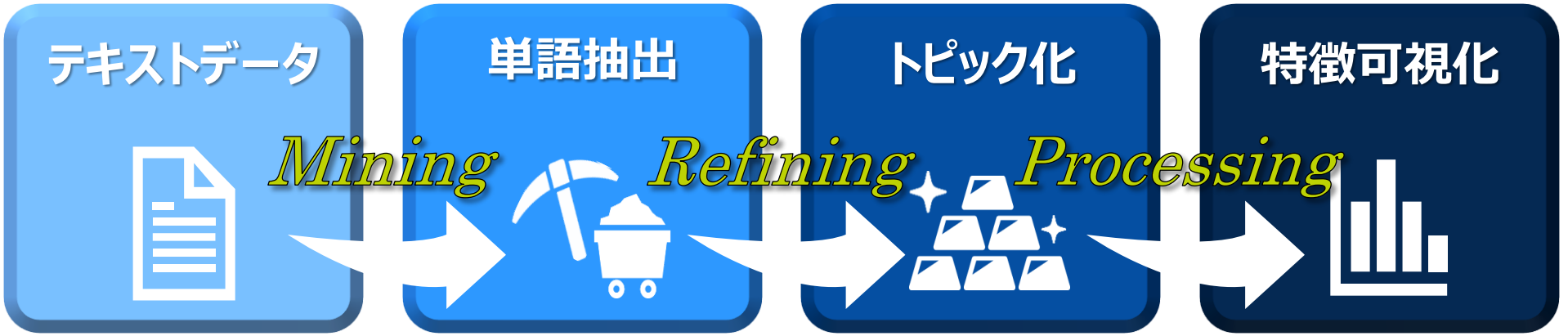
技術に対する用途の関係を分析することで、自社技術と関係がある用途のうち想定をしていない用途を発見し、技術の新規用途展開のアイデアを創出します



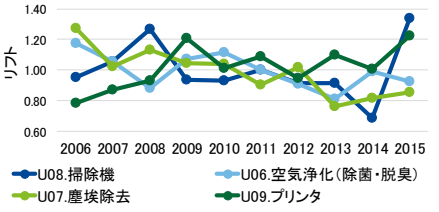
Nomolyticsによるテキストデータ分析の要点

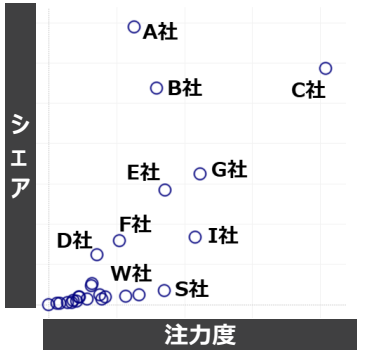
テキストマイニングで抽出した大量の単語をトピックに集約し、そのトピックをベースに特徴を可視化することでテキストデータに潜む特徴をシンプルに理解することができます

テキストリファイニング：トピックをベースにしたテキストデータの分析アプローチ



室内における人の位置を確実に把握して人に直接冷風あるいは温風が当たらないように風の到達位置を制御することで快適空調を達成することができる空気調和機を提供する。室内機に設けられたファンの回転数と、室内機に設けられた上下羽根の角度を制御して、暖房時は人体検知センサにより検知された人の足元近傍に温風を到達させる一方、冷房時は人体検知センサにより検知された人の頭上上方に冷気を到達させるように設定した。





資料に関するお問い合わせやコンサルティングの
ご相談は以下までお願いします。

analytics.office@analyticsdlab.co.jp

会社ホームページもご参考にしてください。
過去の講演・論文資料や技術解説も掲載しています。

http://www.analyticsdlab.co.jp/

株式会社アナリティクスデザインラボ

Analytics Design Lab