



Analytics Design Lab

一般社団法人 企業研究会 主催セミナー

AIの自然言語処理技術の基礎理解と特許文書データの分析

トピックモデルの応用によるインサイトの獲得

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4月11日

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：株式会社アナリティクスデザインラボ

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



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



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「アドバイザー」、弊社実施の分析をお客様にトランスファーする「テラー研修」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ありません
- 1回〇時間の訪問助言を何回ご提供するかによって費用をお見積りします

テラー研修 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、特許文書やコールセンターの問い合わせ履歴、Web上の口コミ、アンケートの自由記述など、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

※デロイト在籍時に提供した案件を含む

データの種類	クライアント業種	プロジェクト概要	期間
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
特許文書	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
特許文書	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
特許文書	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
特許文書	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
特許文書	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
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
コールセンター	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
コールセンター	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
コールセンター	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
コールセンター	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
コールセンター	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
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
口コミ	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
口コミ	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
口コミ	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
口コミ	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
アンケート	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
アンケート	公益事業会社	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
アンケート	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

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

※デロイト在籍時に提供した案件を含む

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
—	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	ロ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

特許文書分析の過去のコンサルティング実績では、特許の“記述内容”に基づいて、客観的に技術を整理し、技術戦略に資する気づきを獲得したいという相談が多く寄せられます

1

客観的な技術分類

- 特許の記述内容から客観的な視点で技術を分類したい
- 自社技術と関連する技術領域の全体像を俯瞰したい

2

競合他社の動向把握

- 競合他社の特徴や棲み分け、自社との関係性を把握したい
- 他社との協業やM&Aの可能性を検討したい

3

保有技術の新規用途探索

- 自社技術を有効活用できる新しい用途を検討したい
- 自社技術を応用したイノベーションのヒントを得たい

4

事業化の技術シーズ把握

- 事業実現のための重要な技術、代替技術を把握したい
- 事業展開における競合他社の存在を把握したい

5

権利侵害リスクの把握

- 権利侵害になり得る類似特許を調べたい
- 従来のキーワード検索では拾えない類似特許を調べたい

1. テキストマイニングと自然言語処理技術

1-1. テキストマイニングと自然言語処理技術の体系

1-2. 従来の自然言語処理技術

1-3. トピックモデル

1-4. 深層学習モデル（第3次AIブーム 前半編）

1-5. 深層学習モデル（第3次AIブーム 後半編）

1-6. テキストマイニングと大規模言語モデル

2. 複数のAI技術を応用した特許文書データの新たな分析手法

2-1. 従来の特許文書分析とその課題

2-2. AI技術の応用: PLSAとベイジアンネットワーク

2-3. Nomolytics: PLSAとベイジアンネットワークを応用した新たなテキスト分析手法

3. Nomolyticsを適用した特許分析事例①

3-1. 「風・空気」に関する特許文書データ

3-2. 分析プロセスの全体像

3-3. トピックの抽出

3-4. トピックのスコアリング

3-5. 出願年×トピックによるトレンド分析

3-6. 出願人×トピックによる競合分析

3-7. 用途×技術の関係分析<その1>
～用途⇒技術の関係～

3-8. 用途×技術の関係分析<その2>
～技術⇒用途の関係～

3-9. Nomolyticsによる特許分析のまとめ

4. Nomolyticsを適用した特許分析事例②

4-1. 電気自動車に関する特許文書データ

4-2. トピックの抽出

4-3. 出願年×トピックによるトレンド分析

4-4. 出願人×トピックによる競合分析

5. インサイト獲得のためのPLSAの新展開技術 PCSAとdifferential PLSA

5-1. インサイト獲得で求められるトピック抽出のあり方

5-2. PCSA:
あるターゲットに特化したトピックの抽出手法

5-3. PCSAを適用した特許分析事例

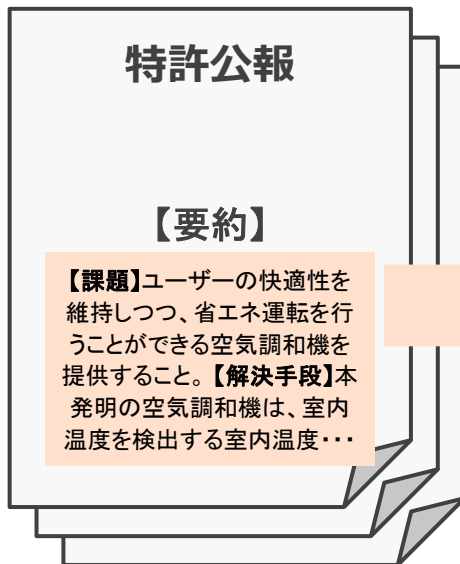
5-4. differential PLSA:
頻度によらない個性的なトピックの抽出手法

5-5. differential PLSAを適用した特許分析事例

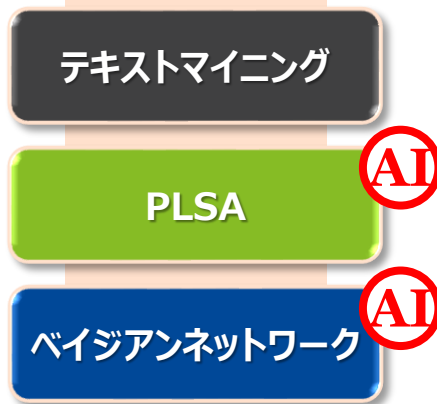
6. まとめ

Nomolyticsによる特許文書分析のアプローチ概要

本日は特許の文書情報(要約文)にテキストマイニングと2つのAI技術を適用して、文書内容をトピックに類型化し、技術戦略に資する特徴を可視化する分析事例を2件ご紹介します



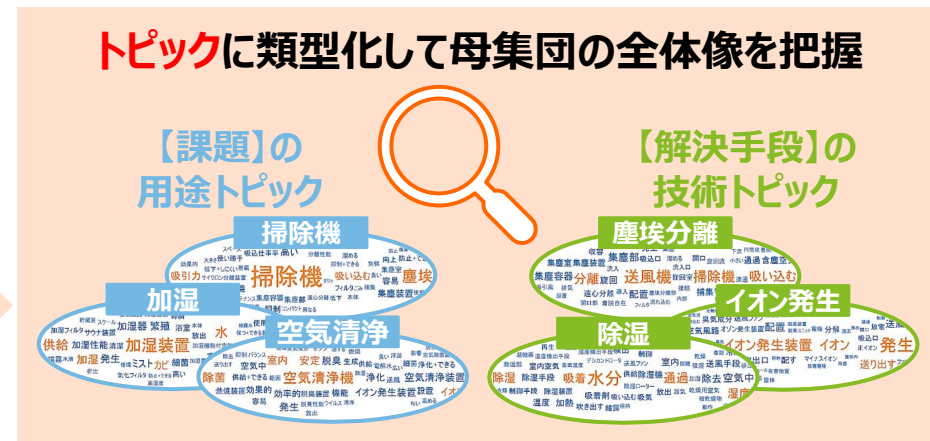
文書情報(要約)を分析



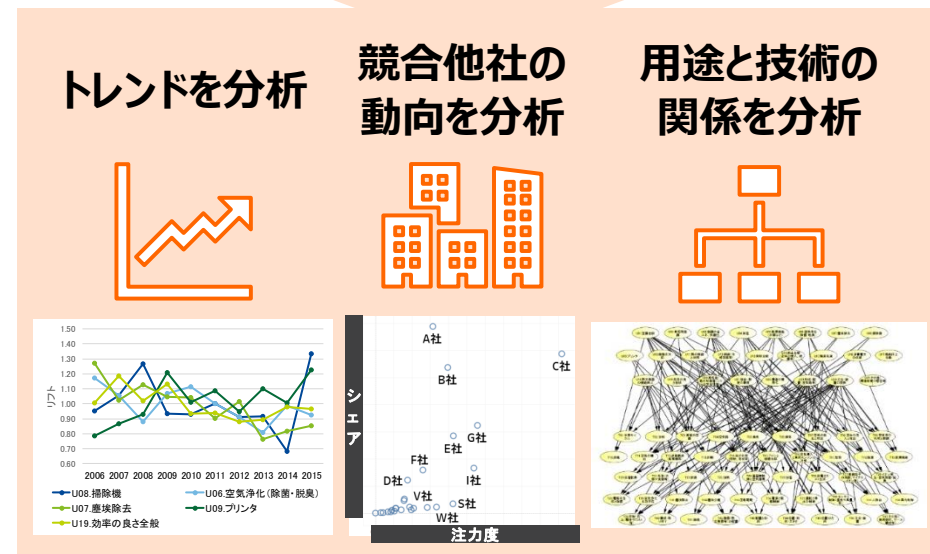
- 2つの分析事例をご紹介します
- ① 風・空気に関する特許 (30,039件)
 - ② 電気自動車に関する特許 (26,419件)

★★★ここがポイント★★★

従来はテキストマイニングで抽出された大量の単語をベースに分析していたが(結果が複雑だった)、それをAIで類型化されたいくつかのトピックをベースに分析することで特許文書に潜む特徴をシンプルに把握できる



トピックをベースに



1. テキストマイニングと自然言語処理技術

1. テキストマイニングと自然言語処理技術

1-1. テキストマイニングと自然言語処理技術の体系

テキストマイニングとは

テキストマイニングは、テキストデータに記述されている内容の特徴や傾向を把握するための分析手法で、ツールが豊富で使いやすく、ビジネスでもよく活用されています

概要

- 大量のテキストデータから、その文章に含まれる単語や係り受け表現を抽出し、その出現頻度を集計したり、その出現関係を可視化することで、文章の記述傾向を把握する手法
- テキストという定性データを統計的に分析可能にする
- テキストマイニングの2つの基本技術
 - **形態素解析**
文章を意味を持つ最小の言語単位(文字列)に分割し、その文法的素性(品詞など)を付与する
 - **構文解析**
形態素を文節にまとめ、文節間の係り受け関係(主語と述語、修飾語と被修飾語など)を抽出する

函館で綺麗な夜景を見た

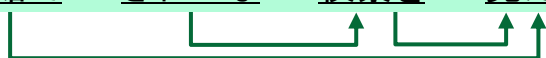
形態素解析

函館 / で / きれい / な / 夜景 / を / 見 / た

(名詞) (助詞) (形容動詞) (助動詞) (名詞) (助詞) (動詞) (助動詞)

構文解析

函館で / 綺麗な / 夜景を / 見た



代表的な公開プログラム

- 形態素解析のプログラム
 - JUMAN (京都大学 黒橋禎夫氏)
 - ChaSen (奈良先端科学技術大学院大学 松本裕治氏)
 - MeCab (京都大学&NTTの共同研究 工藤拓氏)
- 構文解析のプログラム
 - KNP (京都大学JUMANをベースに開発)
 - CaboCha (工藤拓氏&松本裕治氏)

代表的なテキストマイニングツール

- 無償ツール
 - KH Coder (立命館大学 樋口耕一氏)
 - AIテキストマイニング (ユーザーローカル)
- 有償ツール
 - Text Mining Studio (NTTデータ数理システム)
 - 見える化エンジン (プラスアルファ・コンサルティング)
 - TRAINA (野村総合研究所)

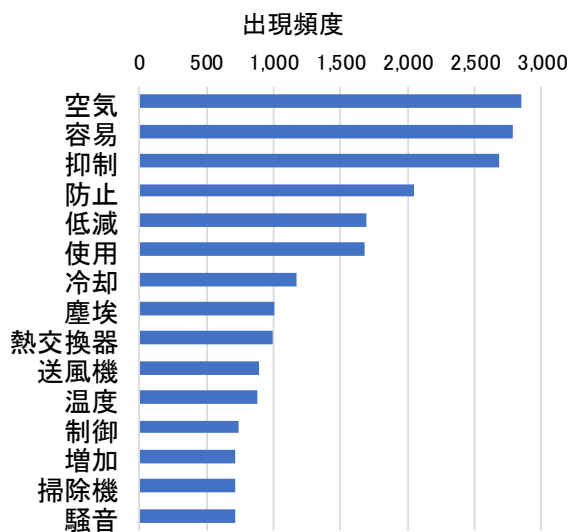
特許文書分析のテキストマイニングでよくあるアウトプット

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、特許文書に記載されている特徴を可視化して全体像の概要を把握します

特許文書のテキストマイニングの可視化例

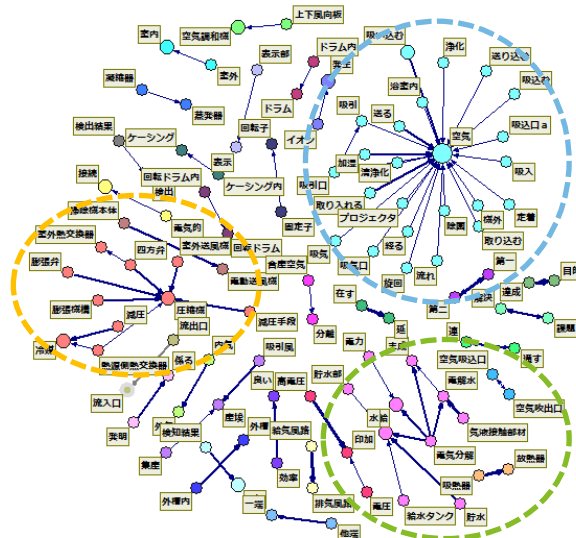
頻度集計

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



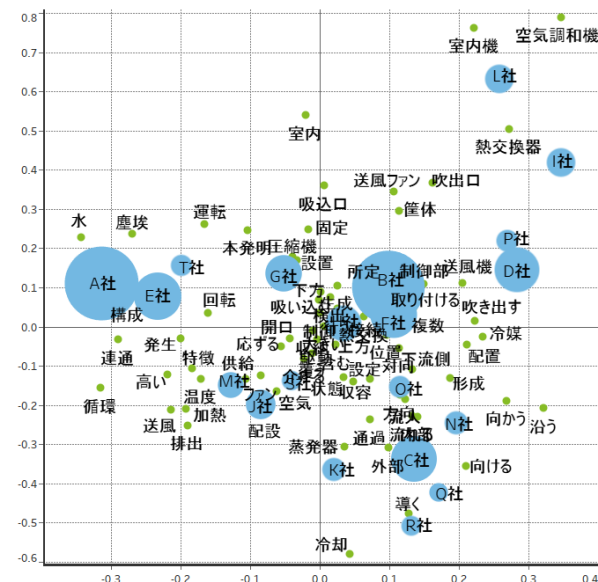
共起ネットワーク

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



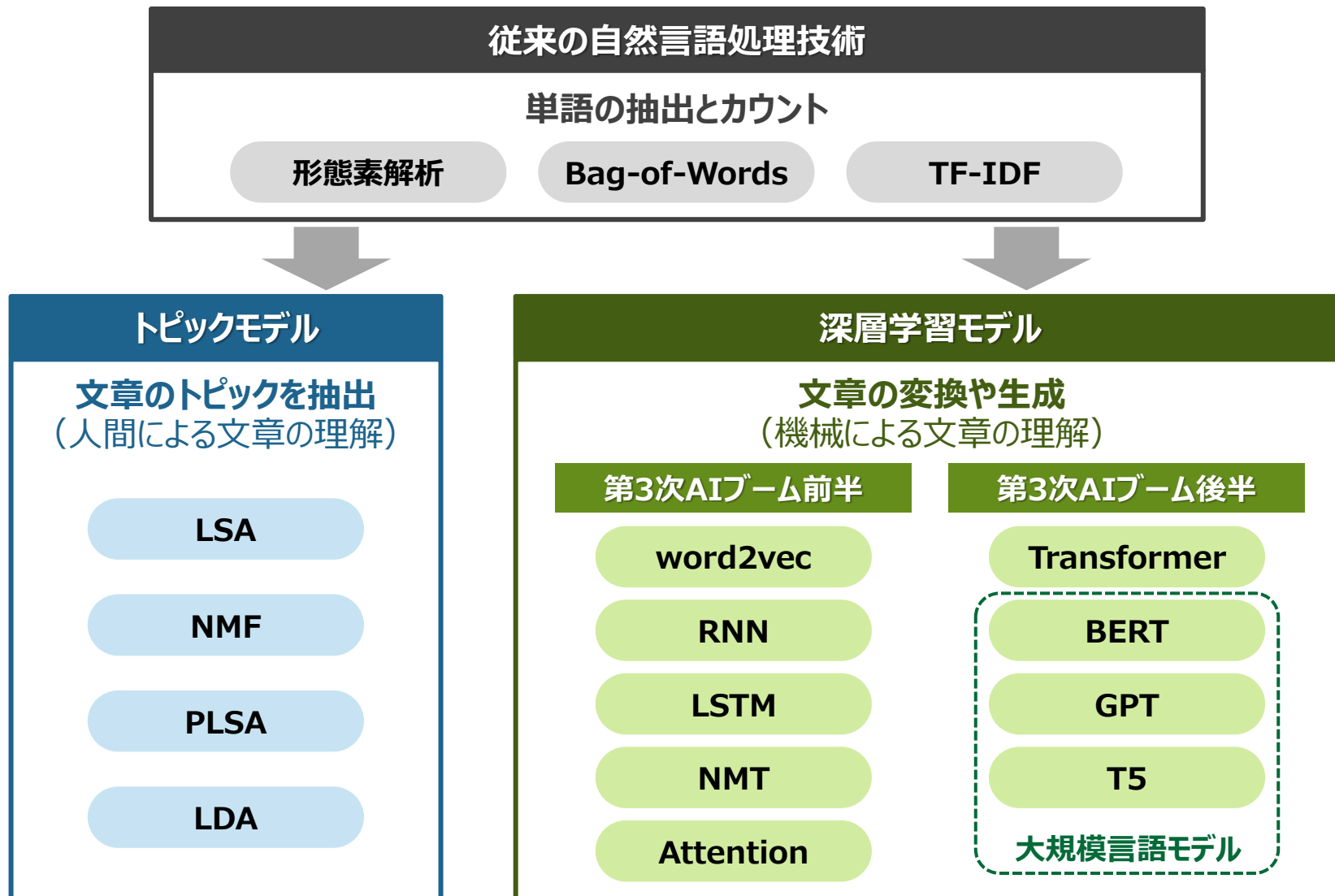
コレスポネンス分析

属性と単語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(出願人など)の傾向を把握する



自然言語処理(NLP: Natural Language Processing)の技術整理

自然言語処理技術は、文章や単語をベクトル表現することで機械で解析可能にする技術であり、トピックモデルで文章を理解したり、深層学習で文章を変換・生成したりできます



1. テキストマイニングと自然言語処理技術

1-2. 従来の自然言語処理技術

Bag-of-Words (BoW)

Bag-of-Wordsは各文書の単語の出現頻度をカウントしたデータで、最もシンプルに単語をベクトル化できる手法ですが、これだけでもテキストデータの様々な定量分析が可能です

概要

- 各文書にどの単語が何回出現したのかをカウントする方法で、出現頻度という数値によって単語一つ一つをベクトルとして表現したもの
- 最もシンプルな単語のベクトル表現だが、これだけでもテキストデータの様々な定量分析が可能で、テキストマイニングのツールで実行される可視化や統計解析は基本的にこのデータ形式をベースにしている

課題

- 全ての単語は同等に扱われ、その単語が全文書で見るときにどれくらい重要であるかは分からない
- 単語の位置や順序、使われ方、文脈上の意味といった情報は保持しておらず、その文字列の出現頻度のみの情報である
- 単語の種類の数ベクトルの次元数となるため、高次元のデータとなり複雑で、計算処理が難しくなる

イメージ図

文書ID	ホテルの口コミコメント	部屋	広い	快適	空調	良い	効く	綺麗	清掃	風呂	駅	近い	便利	人	対応	丁寧	朝食	美味しい	レストラン
1	部屋が広く快適で、部屋の空調も良く効きました。	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	部屋は綺麗に清掃されていて、お風呂も快適でした。	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
3	駅の近くで便利で良く、部屋も広くて良かったです。	1	1	0	0	2	0	0	0	0	1	1	1	0	0	0	0	0	0
4	人の対応が良く、部屋も丁寧に清掃されていました。	1	0	0	0	1	0	0	1	0	0	0	0	1	1	1	0	0	0
5	朝食が美味しく、レストランも広くて綺麗で良かったです。	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1

TF-IDFはBag-of-Wordsで同等に扱われていた各単語の重要度を数値化する前処理であり、文書の特徴をより捉えた分析が可能となります

概要

- 各文書における単語の重要度を数値化する処理であり、TFとIDFを掛け合わせた値を取る
- TF (Term Frequency)
 - 各文書の中における単語の頻度を示す値
 - 頻度が多い単語ほど重要である
 - 文書の長さの影響を受けるため、各文書の全単語頻度で除すことが多い
- IDF (Inverse Document Frequency)
 - ある文書で頻出する単語でも、それがどの文書でも頻出するなら重要とはいえない
 - $IDF = \log(\text{全文書数} / \text{単語}w\text{が出現する文書数})$
 - 多くの文書で現れる単語は値が小さく、特定の文書にしか現れない単語は値が大きくなる
- TF-IDFは、テキストデータの分析でよく用いられる前処理であり、例えば文書間の類似度の分析では、TF-IDFの処理をしたベクトルを使い、そのcos類似度を計算することが多い

イメージ図

Bag-of-Words

ID	部屋	広い	綺麗	スタッフ	丁寧	合計頻度
1	2	1	1	0	0	4
2	1	2	0	0	0	3
3	2	0	1	0	0	3
4	0	0	0	1	1	2



TF-IDF

ID	部屋	広い	綺麗	スタッフ	丁寧
1	0.144	0.173	0.173	0	0
2	0.096	0.462	0	0	0
3	0.192	0	0.231	0	0
4	1	0	0	0.693	0.693

(例) ID=1の「部屋」のTF-IDF

$$TF = 2 / 4 = 0.5$$

$$IDF = \log(4 / 3) = 0.288$$

$$TF-IDF = 0.5 * 0.288 = 0.144$$

1. テキストマイニングと自然言語処理技術

1-3. トピックモデル

LSA (Latent Semantic Analysis)

LSAは「文書×単語」の行列を特異値分解によって「文書×トピック」「トピック×トピック」「トピック×単語」に分解することでトピックを抽出しますが、結果に負の値を含みます

概要

- LSA(潜在意味解析)は、「文書×単語」の行列(Bag-of-Words)を特異値分解によって次元削減することでトピックを抽出する手法で、1990年に発表された
 - 「文書×単語」行列を以下3つの行列に分解する
 - ①文書×トピック (左特異ベクトル)
 - ②トピック×トピック (特異値)
 - ③トピック×単語 (右特異ベクトル)
 - 元の行列と分解後の行列の誤差を最小化する
 - 特異値は対角行列であり、左特異ベクトルと右特異ベクトルの行列は各トピック間で直交している(トピックのベクトルは互いに独立している)
- 大きな値を取るベクトルに引っ張られてトピックが抽出される傾向があるため、Bag-of-WordsにTF-IDFなどで重み付けされた行列を用いられることが多い
- 最適解が数学的に保証されており、計算効率も高い
- 分解する行列の要素に負の値を許容しているため、結果の解釈が難しくなる
- 結果が学習データに完全に依存するため、過学習を起こしやすく、新しい文書のトピックは推定できない

イメージ図

LSAの特異値分解

$$X = U \Sigma V^t$$

① $m \times k$ ② $k \times k$ ③ $k \times n$

LSAの結果例

①文書×トピック

U	トピック1	トピック2	トピック3	トピック4
文書1	-0.47	-0.75	-0.46	0.11
文書2	-0.61	-0.10	0.69	-0.37
文書3	-0.27	0.41	-0.54	-0.68
文書4	-0.58	0.52	-0.10	0.62

②トピック×トピック

Σ	トピック1	トピック2	トピック3	トピック4
トピック1	7.7	0	0	0
トピック2	0	2.8	0	0
トピック3	0	0	2.0	0
トピック4	0	0	0	0.6

③トピック×単語

V^t	単語1	単語2	単語3	単語4	単語5
トピック1	-0.51	-0.49	-0.38	-0.37	-0.47
トピック2	-0.27	0.14	0.80	-0.50	-0.09
トピック3	0.47	0.43	-0.35	-0.67	-0.14
トピック4	0.61	-0.74	0.23	-0.18	0.07

NMF (Non-negative Matrix Factorization)

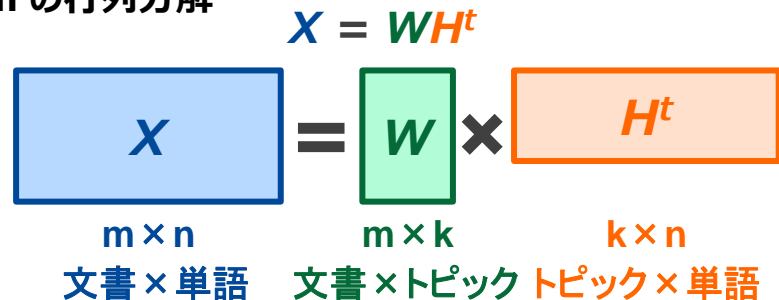
NMFは「文書 × 単語」の行列を2つの非負の行列「文書 × トピック」「トピック × 単語」に分解することでトピックを抽出し、結果が非負となるためトピックの解釈がしやすいです

概要

- NMF(非負行列因子分解)は、「文書 × 単語」の行列 (Bag-of-Words)を2つの非負の行列「文書 × トピック」「トピック × 単語」に分解することでトピックを抽出する手法で、1999年に発表された
- 計算アルゴリズムは、元の行列と分解後の行列の積との誤差を最小化することを目的関数に、初期値を与えた反復計算により最適解を得る
 - 誤差の定義の仕方は、平方ユークリッド距離やKLダイバージェンスなどが使われる
- LSAと比較して、分解後の行列の要素が全て非負であるため、結果の解釈がしやすい
- 分解された「文書 × トピック」「トピック × 単語」の行列は、各トピックのベクトル間で直交しておらず、各トピックは互いに独立していないため、抽出されたトピックの一部は意味が重複する可能性がある
- 結果が学習データに完全に依存するため、過学習を起こしやすく、新しい文書のトピックは推定できない

イメージ図

NMFの行列分解

$$X = WH^t$$


$m \times n$ 文書 × 単語 $m \times k$ 文書 × トピック $k \times n$ トピック × 単語

NMFの結果例

文書 × トピック

W	トピック1	トピック2	トピック3	トピック4
文書1	0.22	0.48	0.16	0.39
文書2	0.35	0.12	0.07	0.88
文書3	0.11	0.58	0.24	0.14
文書4	0.29	0.27	0.60	0.20

トピック × 単語

H ^t	単語1	単語2	単語3	単語4	単語5
トピック1	4.34	0.03	1.21	2.24	1.87
トピック2	3.41	1.89	2.81	0.46	1.03
トピック3	0.38	2.24	1.09	5.62	0.06
トピック4	1.85	3.66	0.08	2.11	4.16

PLSA (Probabilistic Latent Semantic Analysis)

PLSAはLSAを確率的に発展させた手法で、「文書 × 単語」の行列を確率モデルによって $P(\text{文書} | \text{トピック})$ 、 $P(\text{単語} | \text{トピック})$ 、 $P(\text{トピック})$ に分解することでトピックを抽出します

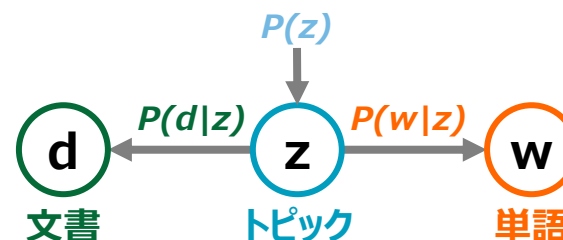
概要

- PLSA(確率的潜在意味解析)は、LSAの処理を確率的な枠組みで発展させたもので、「文書 × 単語」の行列(Bag-of-Words)を確率的に分解することでトピックを抽出する手法で、1999年に発表された
- 同時確率 $P(\text{文書}, \text{単語})$ を以下の3つの確率分布に分解して表現し、「文書 × 単語」データから推定する
 - ① $P(\text{文書} | \text{トピック})$
 - ② $P(\text{単語} | \text{トピック})$
 - ③ $P(\text{トピック})$
 - 対数尤度関数を最大化するEMアルゴリズムを実行し、初期値を与えた反復計算により最適解を得る
- 各トピックは互いに独立の仮定を置いている
- LSAでは「文書 × 単語」の行列に対して事前にTF-IDFなどで重みづけする必要があるが、PLSAでは確率的な処理によりそうした重みづけは不要となる
- 文書および単語のトピックに対する関連度が所属確率として出力されるため、結果が解釈しやすい
- 結果が観測データに完全に依存するため、過学習を起こしやすく、新しい文書のトピックは推定できない
 - 一方で、観測データの再現度が高く、個別性の強い特徴を反映できるモデルと捉えることもできる

イメージ図

PLSAの確率モデル

$$P(d, w) = \sum_z P(d|z)P(w|z)P(z)$$



PLSAの結果例

$P(\text{文書} d | \text{トピック} z)$

$P(d z)$	トピック1	トピック2	トピック3	トピック4
文書1	0.41	0.16	0.06	0.27
文書2	0.29	0.54	0.11	0.06
文書3	0.22	0.13	0.04	0.58
文書4	0.08	0.17	0.79	0.09

$P(\text{単語} w | \text{トピック} z)$

$P(w z)$	トピック1	トピック2	トピック3	トピック4
単語1	0.11	0.09	0.44	0.20
単語2	0.09	0.38	0.04	0.18
単語3	0.50	0.11	0.29	0.09
単語4	0.08	0.14	0.17	0.31
単語5	0.22	0.28	0.06	0.22

$P(\text{トピック} z)$

$P(z)$	トピック1	トピック2	トピック3	トピック4
	0.31	0.27	0.23	0.19

$$\sum_d P(d|z) = 1$$

$$\sum_w P(w|z) = 1$$

$$\sum_z P(z) = 1$$

LDA (Latent Dirichlet Allocation)

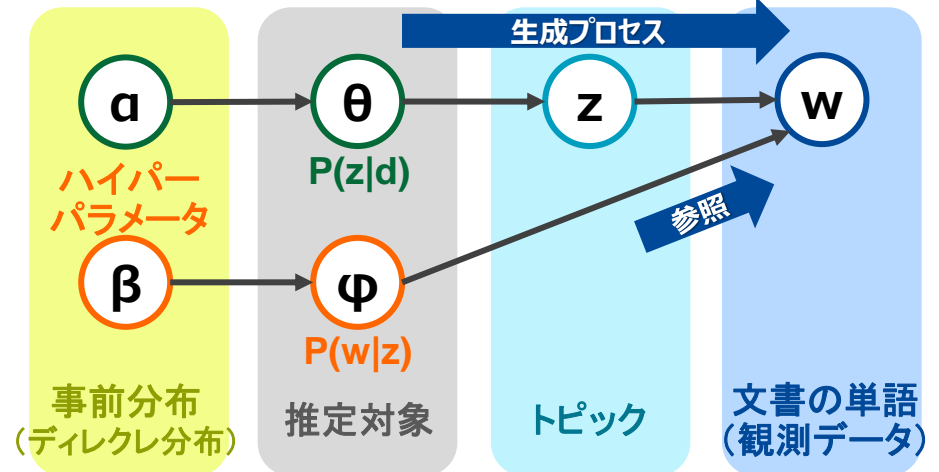
LDAはPLSAをベイズ的に拡張させてディクレ分布を事前分布に導入した手法で、過学習を抑制でき、新しい文書のトピックも推定できる高い汎化性能があります

概要

- LDA(潜在ディクレ配分法)は、PLSAをベイズ的に拡張させた手法であり、ディクレ分布を事前分布に導入しており、2003年に発表され、トピックモデルの中で最もよく使われている手法である
- 推定対象は $\theta=P(\text{トピック}|\text{文書})$ と $\phi=P(\text{単語}|\text{トピック})$ であり、推定アルゴリズムは、ギブスサンプリングや変分ベイズ法などが適用され、反復計算される
- PLSAはパラメータは固定的で、観測データのみから直接推定するため、結果が完全に観測データに依存するが、LDAはパラメータは事前分布に従い変動する確率分布とし、観測データと事前分布から推定する
 - 事前分布を導入することで、確率的なスムージング効果があり、過学習を抑制できる
- LDAは新しい文書についても推定ができ、新しい文書に対応する" θ "を未知とし、" w "(文書に含まれる単語)と学習済みの" ϕ "と" α "から" θ "を推定する
- ハイパーパラメータ α と β によって結果が変動しやすく、この値の設定の仕方、推定の仕方が難しい
- 新しい文書の推定ができる高い汎化性能を持つが、トピックの結果が一般的で抽象度が高くなることがある

イメージ図

LDAのグラフィカルモデル



LDAは、文書 d 内の各単語 w が特定のトピック z から生成されたと仮定する。このトピック z の分布は文書 d 毎に異なり、その確率分布 $P(z|d)$ は θ で表す。特定のトピック z が各単語 w を生成する確率分布 $P(w|z)$ は ϕ で表し、そのトピック z の下で単語 w を生成するプロセスでは ϕ が参照される。なお、 θ と ϕ はそれぞれハイパーパラメータ α と β を持つディクレ分布に従う。LDAではこれらの生成過程を逆にたどるアルゴリズムにより、単語 w (観測データ)から θ と ϕ を推定する。

ハイパーパラメータ α, β の大きさとディクレ分布の特徴

- $\alpha, \beta > 1$ 分散が小さい集中的な分布となり、一部の要素に偏る
- $\alpha, \beta = 1$ 分布は一様分布となり、PLSAに近い振る舞いをする
- $1 > \alpha, \beta > 0$ 分散が大きく広がった分布となり、多様な要素を含む

1. テキストマイニングと自然言語処理技術

1-4. 深層学習モデル（第3次AIブーム 前半編）

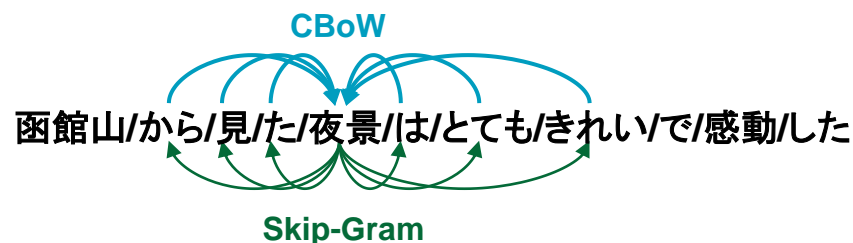
word2vecは大量のコーパスの穴埋め問題を学習したモデルで、単語の意味や類似性を捉えたベクトル表現を得ることができ、単語のベクトル間の演算もできます

概要

- word2vecは2013年に発表された単語のベクトル表現(分散表現)を得る手法であり、これは単語の意味や類似性を捉えており、単語同士の演算もできる
- word2vecは大量のテキストデータ(コーパス)を用いた深層学習モデルで、CBoWとSkip-Gramの2つのアルゴリズムがある
 - CBoW (Continuous Bag of Words)は、ある単語をその周辺単語から予測し、Skip-Gramはある単語からその周辺単語を予測する(穴埋め問題)
- 事前学習されたモデルを使うことで、単語のベクトル表現を容易に得ることができる
 - 例えば「GoogleNews-vectors-negative300」は約300万語に対して300次元のベクトルを得られる
- 各文章に対して単語と同様に「文章ID」の要素を持たせてword2vecの学習を適用することで、単語と同様に「文章ID」のベクトル表現を得ることができ、この手法をdoc2vecという
- word2vecは単語の順序情報が欠落しており、同じ単語でも、文脈で異なる意味を持つ場合は区別できない

イメージ図

CBoWとSkip-Gram



※ウインドウサイズ(周辺単語の数)=3のイメージ

単語の演算

「王」 - 「男」 + 「女」 ≃ 「女王」

(例) word2vecで得られる単語のベクトル表現
(4次元のイメージ)

王 = (1.2, 0.6, -0.8, 2.0)

男 = (1.3, 0.5, -1.0, 1.9)

女 = (1.1, 0.7, -0.9, 2.1)

女王 = (1.0, 0.8, -0.7, 2.2)

word2vecは単語を線形性を持つベクトル空間に配置するため単語間の演算が成立する

RNN (Recurrent Neural Network)

RNNは系列データの処理モデルで、過去の情報を保持して現在の入力进行处理することで、単語の順序を考慮できますが、長い系列は過去の情報を現在に反映することが困難です

概要

- RNNは文章や時系列情報など、系列データを処理するモデルで、考え方の起源は1980年代に発表された
 - 文章を単語の系列データとして捉え、単語を一つずつ順番に逐次的に処理をする
 - 各ステップで生成した隠れ状態ベクトルが次の隠れ状態の入力にもなる再帰的な処理をする
- 過去の処理情報を保持して、それが現在の入力に影響を与える構造により、単語の順序(文脈)を考慮した処理ができる
- 長い系列データの場合、過去の情報を反映させるのが困難となる「長期依存性の問題」がある
 - 再帰的処理により隠れ状態が次の隠れ状態に連携し、長い系列でネットワークが深くなると、誤差逆伝播法において勾配爆発や勾配消失が起き、過去の遠い情報を最適に保持できなくなる
 - 勾配爆発は、再帰処理により、同じ重みが繰り返し乗算され、勾配が指数関数的に増加する
 - 勾配消失は、再帰処理により、微分が0に近い活性化関数を繰り返し通過し、勾配が急速に小さくなる

イメージ図

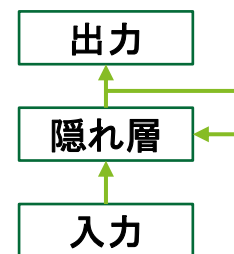
ニューラルネットワークの構造の種類

順伝播型 (フィードフォワード)



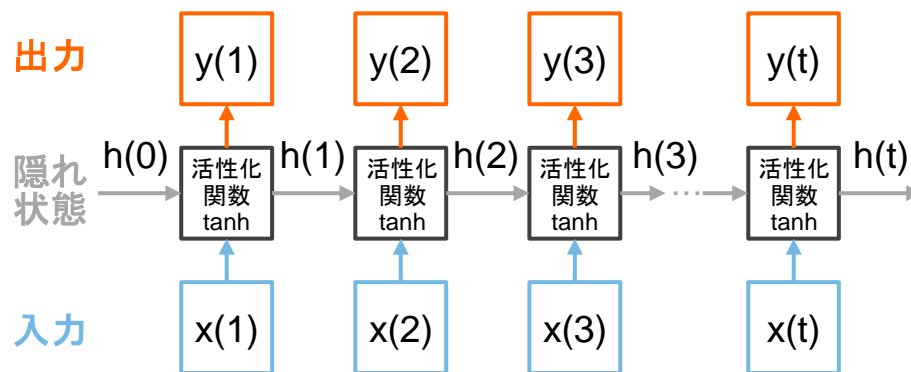
入力から出力に向けて一方向に流れる

再帰型 (リカレント)



隠れ層からの出力が再度自分の入力となる

RNNの処理



※h(0)は初期化された隠れ状態(要素ゼロ)

LSTM (Long Short-Term Memory)

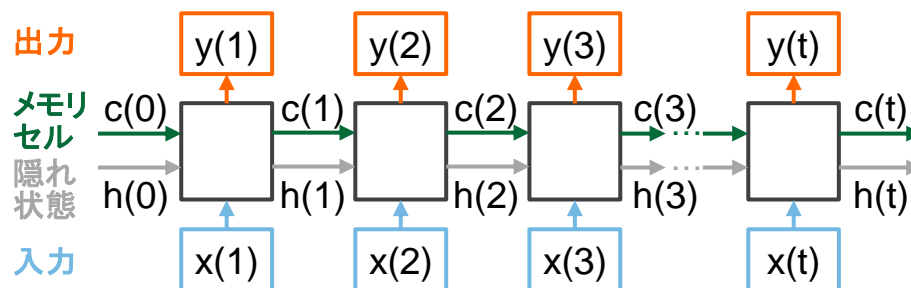
LSTMはRNNの長期依存性の問題を解決したネットワーク構造を持ち、過去の重要な情報を保持し、不要な情報を忘れる能力を有し、長い系列のデータの処理ができます

概要

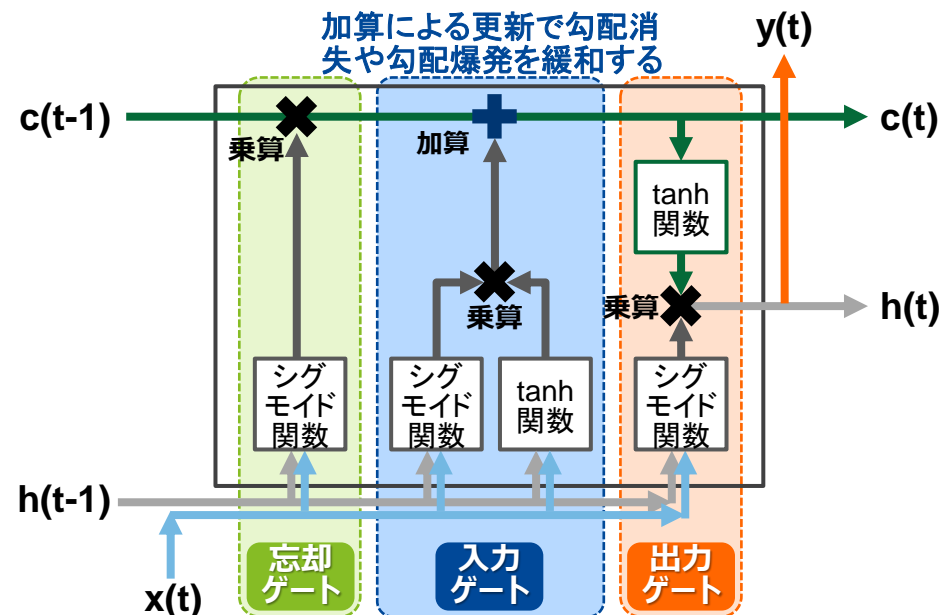
- LSTMはRNNと同じく系列データの処理モデルで1997年に発表されており、長い系列の処理ができる
 - RNNの隠れ状態 h (短期記憶)に加え、メモリセル c (長期記憶)の情報がセルからセルに受け継がれる
- 忘却ゲート、入力ゲート、出力ゲートという3つのゲート構造を持ち、各ゲートが情報の流れを制御し、長期記憶が加算によって更新されることで(RNNの更新は乗算のみ)、長期依存性の問題を解決する
 - 忘却ゲートは、受け継がれたメモリセル c の長期記憶の情報のうちどれを消去するか制御する
 - 入力ゲートは、現在入力 x と前の隠れ状態 h を使って、新しく記憶すべき情報の候補を生成し、どの情報を入力し記憶させるか制御する
 - 出力ゲートは、次の隠れ状態 h と出力 y の情報を制御する
- LSTMは勾配消失や勾配爆発の問題を緩和し、RNNよりも長い系列が扱えるが、完全ではなく、RNNが10語程度に対してLSTMでも20語程度と言われている
- LSTMは計算コストが高いという課題もある

イメージ図

LSTMの処理



LSTMのセルの内部



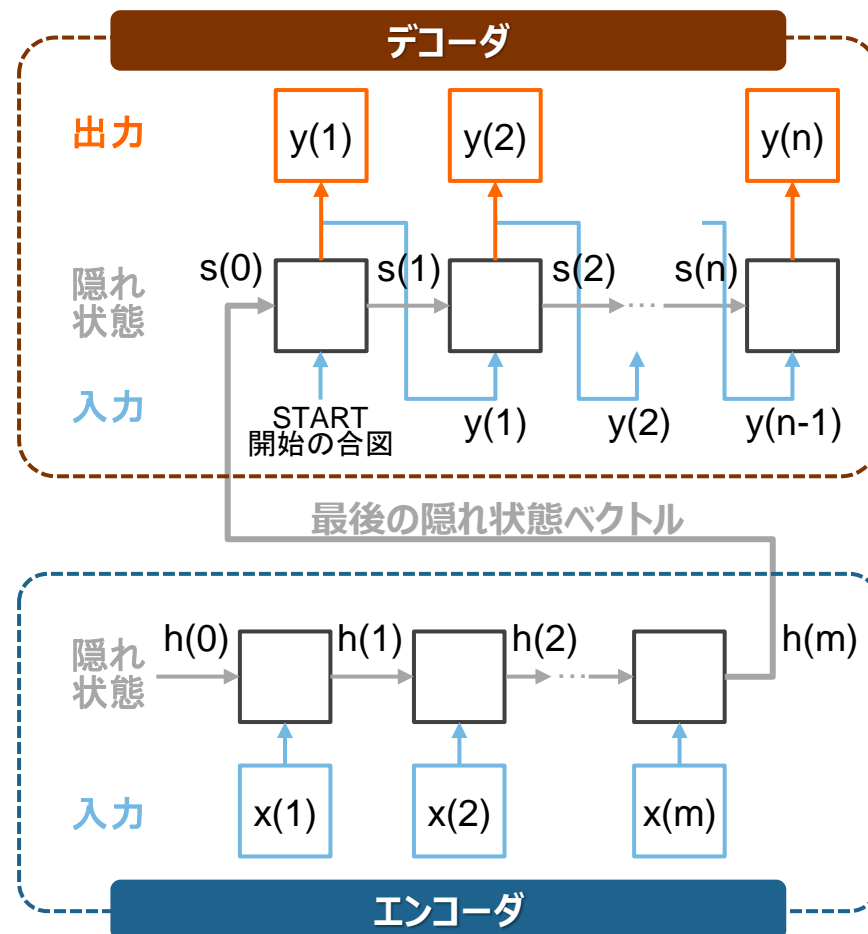
seq2seqは機械翻訳など文章を文章に変換するモデルで、RNNやLSTMで構成されたエンコーダ(文章のベクトル変換)とデコーダ(ベクトルの文章変換)を連結しています

概要

- seq2seqは系列データを系列データに変換するモデルで、Encoder-Decoderモデルとも呼ばれ、機械翻訳を主な用途とし、2014年に発表された
 - 文章をベクトルに変換するエンコーダ(seq2vec)と、ベクトルを文章に変換するデコーダ(vec2seq)が連結した構造を持つ
- RNNやLSTMは単体では、①seq2vec、②vec2seq、③入力・出力の系列数が一致するseq2seqは処理できるが、系列数の一致しないseq2seqは処理できないため、seq2vecとvec2seqを連結して対応したモデル
 - seq2vecは、文章を一つのベクトルに変換する処理で、RNNやLSTMの最後の隠れ状態に該当する
 - vec2seqは、一つのベクトルを文章に変換する処理で、RNNやLSTMでは、一つのベクトルを入力として単語を生成し、その単語と前の隠れ状態を新たな入力として次の単語を生成し、文章を生成する
 - RNNやLSTMは系列ごとに出力を生成するため、単語の品詞割当など入力と出力の系列数が一致するseq2seqは処理できるが、機械翻訳など入力と出力の系列数が一致しないと処理できない

イメージ図

seq2seqの構造 (RNNを用いた場合)



※ $s(0)$ はエンコーダの最後の隠れ状態ベクトル $h(m)$ となる

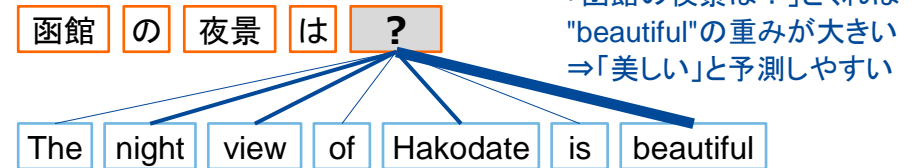
Attentionは入力文を出力文に変換する際に、入力のどの単語に注目すべきかを考慮する仕組みで、入力と出力の依存関係を捉えることができ、精度の高い結果を生成できます

概要

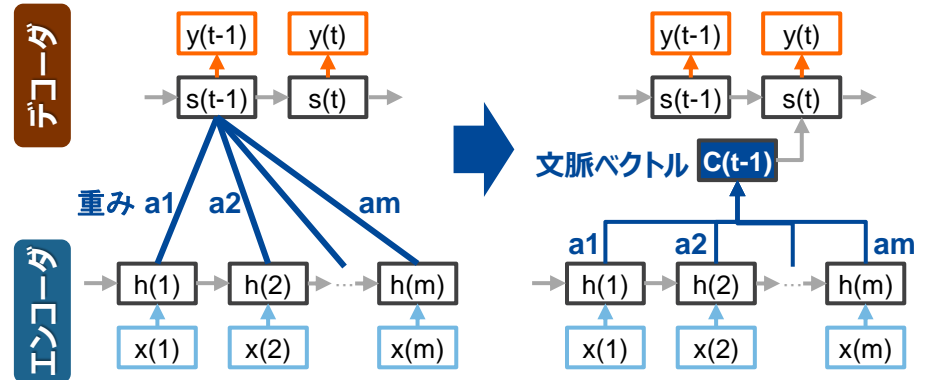
- Attentionは、文章の中でどの単語に注目すべきかを判断する仕組みであり、2014年に発表された
 - 従来のseq2seqは、エンコーダの最後の隠れ状態だけがデコーダに連携されるため、受け渡しできる情報量に限界があり、長い系列ほど精度が落ちた
 - Attentionでは、入力文章を出力文章に変換する際に、入力のどの単語に注目すべきか、あるいは無視すべきかを考慮できるため、出力精度が上がる
 - 遠い所にある単語も含め、入力文章に含まれる全ての単語を参照できるため、RNNやLSTMで課題となっていた長期依存性の問題を補完できる
- Attentionでは、デコーダの今の隠れ状態に対するエンコーダの各隠れ状態の注目度(Attention weight)を計算し、その重み付きの隠れ状態を足し合わせて文脈ベクトルを作成し、これをデコーダに受け渡す
 - この処理をデコーダが新しい単語を生成する度に行い、新たな文脈ベクトルが作成され受け渡される
 - 注目度はデコーダとエンコーダの隠れ状態のペアを入力とする単純なニューラルネットで計算される

イメージ図

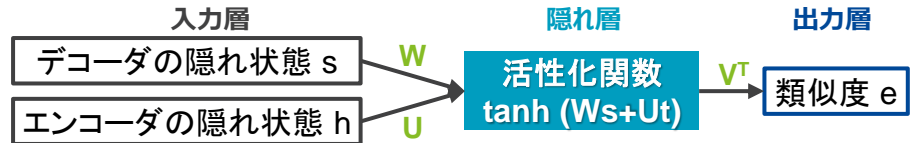
Attentionによる翻訳のイメージ



Attentionの重みと文脈ベクトルの作成



Attentionの重み a を計算するニューラルネットワーク



デコーダの隠れ状態とエンコーダの隠れ状態のペアを入力に、その類似度 e を出力としたニューラルネットを学習してそれぞれの重みを最適化し、各類似度 e_i にsoftmax関数を適用して合計が1となる重み a を計算する

1. テキストマイニングと自然言語処理技術

1-5. 深層学習モデル（第3次AIブーム 後半編）

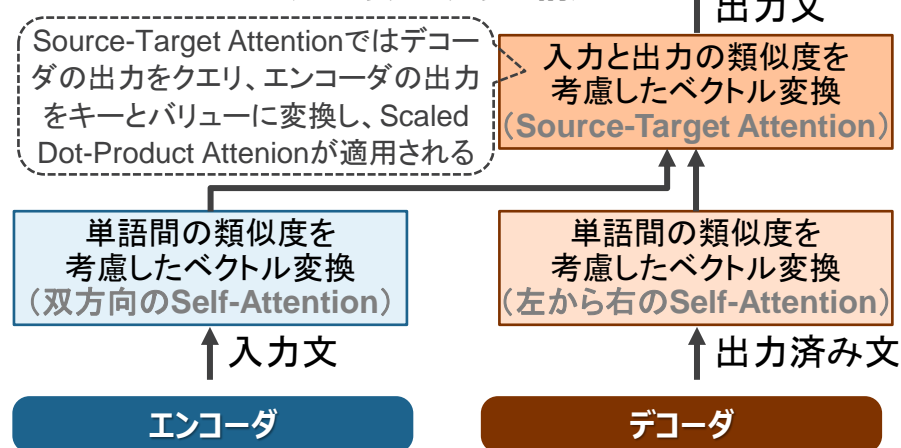
TransformerはAttentionだけを用いたエンコーダ-デコーダ構造の深層学習モデルで、高精度かつ高速化を実現し、自然言語処理の分野で大きなブレイクスルーとなりました

概要

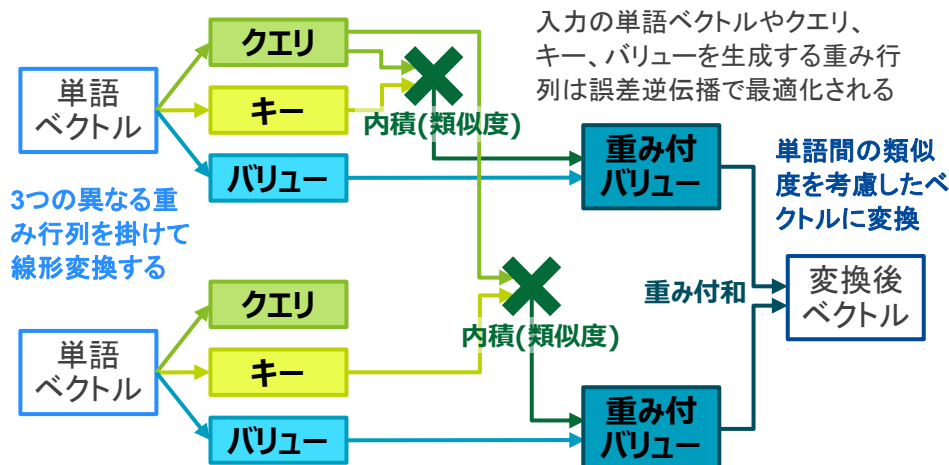
- Transformerは、2017年に"Attention is All You Need"というタイトルで発表された、Attentionだけを用いたエンコーダ-デコーダの構造の深層学習モデル
- 3つのAttentionで、表現力の高いベクトルを獲得する
 - Self-Attention
元々のAttentionは、入力文(Source)の単語と出力文(Target)の単語を対応づけたSource-Target型のAttentionだが、Self-Attentionは、同一文章の中の各単語が他の単語とどの程度関係しているのかを評価し、単語間の依存関係を学習する
 - Scaled Dot-Product Attention
Self-Attentionにおいて、内積の類似度計算で他の単語との関連性を考慮した単語ベクトルを獲得する
 - Multi-Head Attention
Self-Attentionを異なる表現空間(ヘッド)で複数パターン実行し、より柔軟な単語ベクトルを獲得する
- 単語の順序情報は周期性のあるsin,cos関数でベクトル表現され、最初に各単語ベクトルに加算される
- 従来のseq2seqの逐次的処理と異なり、一度に複数の単語を並列化処理でき、計算効率が高い

イメージ図

Transformerのアーキテクチャの構造



Self-AttentionとScaled Dot-Product Attention



BERT (Bidirectional Encoder Representations from Transformers)

BERTは大量のテキストデータから事前学習されたTransformerのエンコーダモデルであり、文章全体の文脈理解に優れており、文章分類や感情分析のタスクに適しています

概要

- BERTはTransformerを使って大量のテキストデータを事前学習した汎用的な大規模言語モデルであり、2018年10月にGoogleによって発表され、2019年10月には検索エンジンに採用している
- Transformerのエンコーダ部分を使用したモデルで、文脈を捉えた単語のベクトル表現あるいは文章のベクトル表現を得ることができる
 - 文章全体の文脈理解において優れており、文章分類や感情分析などのタスクに適していると言われる
- Bidirectional(双方向)の意味は、Self-Attentionを実行するときに、各単語は同一文章中の左右にある他の全ての単語との関係性を捉えるということ
- 事前学習は自己教師あり学習で、一部の単語をマスクしその単語を双方向から予測するMLM(Masked Language Model)と、2つの文が連続するか否かを予測するNSP(Next Sentence Prediction)を実行し、文章全体の文脈理解能力を獲得する
- 事前学習モデルに追加の教師データでファインチューニングすることで、特定のタスクに応じてモデルが微調整され、少量の学習データでも高い精度が得られる

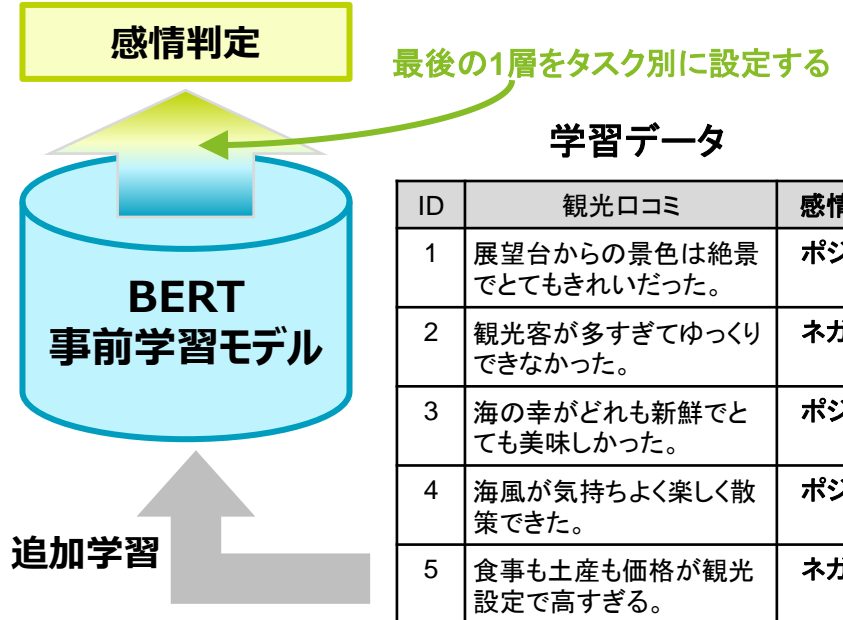
イメージ図

BERTのSelf-Attention

左右双方向の関係性を捉える



BERTのファインチューニング



GPT (Generative Pre-trained Transformer)

GPTは大量のテキストデータから事前学習されたTransformerのデコーダモデルであり、テキストの生成に優れており、文章作成や要約作成、対話生成のタスクに適しています

概要

- GPTはTransformerを使って大量のテキストデータを事前学習した汎用的な大規模言語モデルであり、2018年6月にOpenAIによって発表された
- Transformerのデコーダ部分を使用したモデルで、テキストの生成能力において優れており、文章生成や文章補完、要約作成、対話生成、言語翻訳などのタスクに適していると言われる
- Self-Attentionを実行するときは、左側にある単語のみが使われ、一方向の関係性を捉える
- 事前学習は自己教師あり学習となり、テキストの次にくる単語を予測し、テキストの生成能力を獲得する
- 事前学習モデルを再利用する際は、特定のタスクの説明や指示を入力(プロンプト)として与えるが、shotと呼ばれるタスクを解く例(サンプル)を少数与えることで、ファインチューニングをすることなく様々なタスクに柔軟に対応できる(Few-Shot Learning)
- OpenAIは2019年にGPT-2を、2020年にGPT-3を発表し、2022年11月にはGPT-3.5をベースとした"ChatGPT"をリリースし、利用者は2ヶ月で1億人に達し、その性能の高さに世界は騒然とした

イメージ図

GPTのSelf-Attention

左から右方向への関係性を捉える

函館山/から/見た/夜景/は/とても/きれいで/感動した

GPTのFew-Shot Learningのプロンプト例

Few-Shot

回答>> 感情: ポジティブ

口コミの感情を分析してください。

口コミ: 朝市で食べたうに丼が美味しすぎた。感情: ポジティブ

口コミ: ロープウェイの待ち時間が長すぎる。感情: ネガティブ

口コミ: 赤レンガ倉庫でお土産を買いました。感情: ニュートラル

口コミ: 元町は歴史的な建物が多く散策が楽しかった。

One-Shot

回答>> 英語: Where is the ropeway station?

日本語を英語に翻訳してください。

日本語: 写真を撮っていただけませんか?

英語: Could you take our picture, please?

日本語: ロープウェイ乗り場はどこですか?

Zero-Shot

2泊3日で函館を一人旅するプランを考えて。

T5 (Text-to-Text Transfer Transformer)

T5は大量のテキストデータから事前学習されたTransformerのエンコーダ-デコーダモデルであり、テキストをテキストに変換する処理で自然言語処理タスク全般に対応できます

概要

- T5はTransformerを使って大量のテキストデータを事前学習した汎用的な大規模言語モデルであり、2019年10月にGoogleによって発表された
- Transformerのエンコーダ-デコーダ構造を持つモデルで、従来のように個別タスクごとにモデルを構築するのではなく、あらゆる自然言語処理タスクをテキストからテキストに変換する同一フレームワークで扱える
- 事前学習は自己教師あり学習となり、エンコーダでテキストの一部をトークンで置換し、デコーダでテキスト全体の生成をしてトークンを復元することで、文脈理解能力とテキスト生成能力の両方を身につけさせる
- ファインチューニングでは、特定のタスク(翻訳、要約、分類、感情分析等)を指示する"プレフィクス"というラベルを付けたデータを学習することで、一つのモデルで様々なタスクに高い性能で対応できる
- GPTでもfew-shot learningで様々なタスクに対応可能だが、良質な例(shot)の提供が求められ、またGPTは入力の次にくる単語を予測する処理であるが、T5のように入力と出力の関係を強く捉えた生成ではない
- T5は比較的多くの計算資源が必要となる

イメージ図

T5によるテキストからテキストへの変換

translate English to German:

That is good.

Das ist gut.

sentiment:

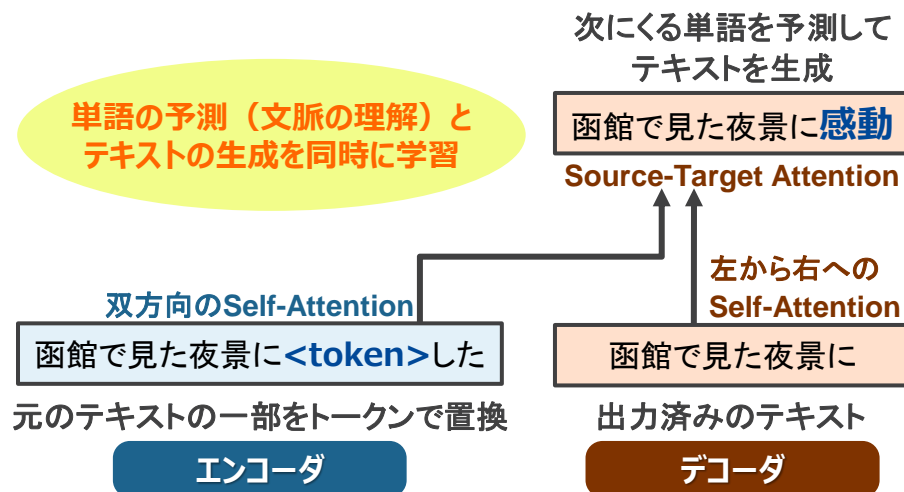
This food is very delicious.

Positive

T5
事前学習
モデル

プレフィクス(タスクの指示)が与えられたテキストの入力から、そのタスクに対応したテキストを出力できる

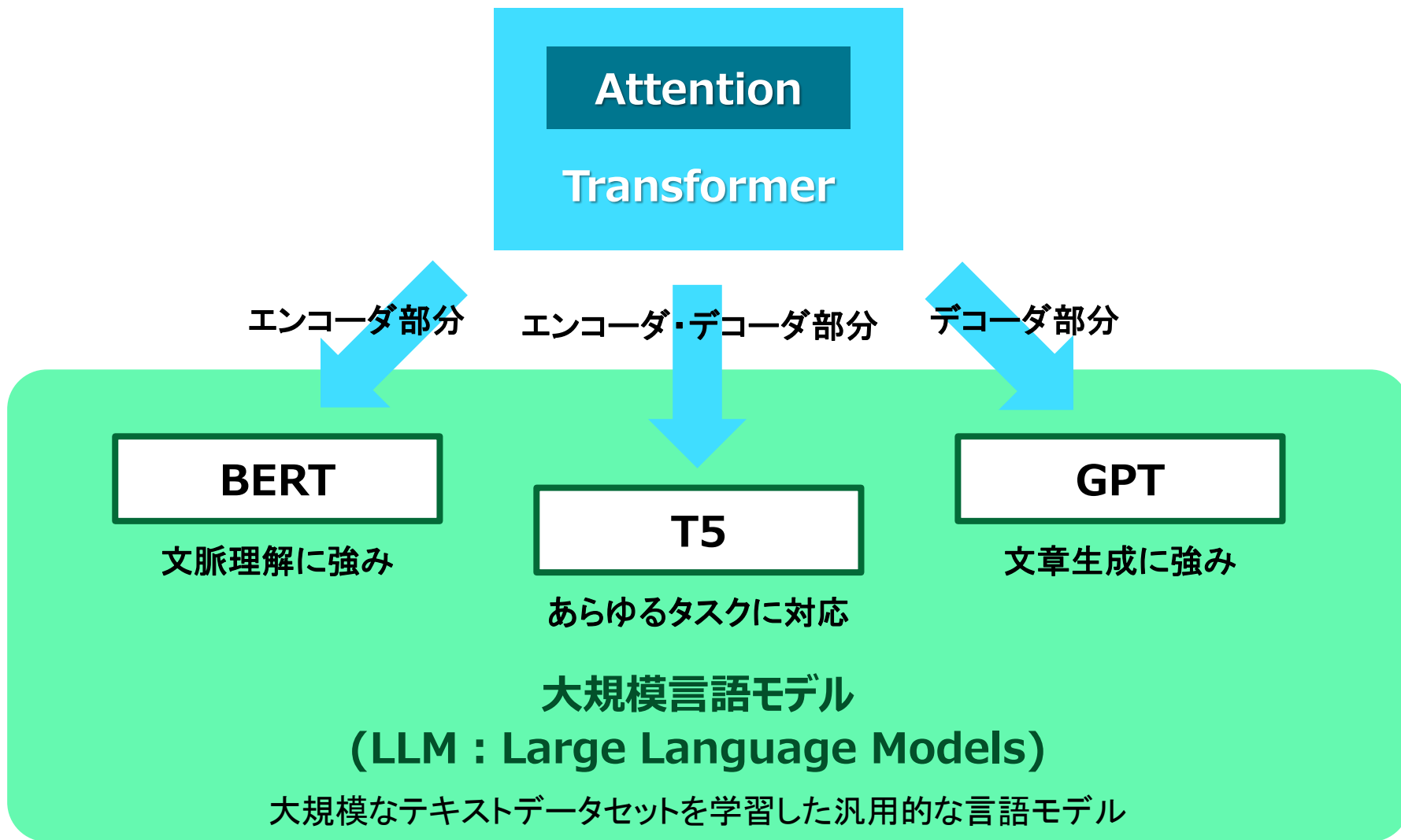
T5の事前学習におけるエンコーダとデコーダの処理



1. テキストマイニングと自然言語処理技術

1-6. テキストマイニングと大規模言語モデル

BERTとGPTとT5は、Attentionの仕組みだけを用いたTransformerというアーキテクチャに基づき、大量のテキストデータから汎用的な言語特徴を学習した大規模言語モデルです



大規模言語モデルとテキストマイニングの対比

大規模言語モデルとテキストマイニングはどちらもテキストデータの分析と活用を目的とした手段という意味では共通しますが、それぞれの特徴や用途には大きな違いがあります

対比の観点	大規模言語モデル	テキストマイニング
用途	文脈の評価や文章の生成	テキストデータの特徴や傾向の把握
活用例・分析例	文章分類、感情分析、文章生成、要約作成、質問応答、対話生成、言語翻訳	単語頻度の集計・推移、単語のネットワーク・マップ化、属性別の特徴語把握
問題解決の方式	直接的 (アウトプットそのものが問題解決において直接的な価値を提供)	間接的 (アウトプットは問題解決のための意思決定に資する価値を提供)
分析の焦点	文脈 (単語間の相互関係性)	単語 (特定の単語の出現有無)
入力データ規模	限定的なテキストデータ (入力単語数に制限あり)	大量のテキストデータ (入力単語数に制限なし)
結果の形式	定性的 (テキスト形式による出力)	定量的 (統計解析による集計と可視化)
適用範囲の性質	汎用性 (言語の汎用的な特徴を事前学習し多様なタスクに適用可能)	個別性 (特定のデータセットに存在する個別の特徴や傾向を把握可能)

大規模言語モデルとテキストマイニングを相互補完的に活用することで、ビジネスの問題解決においてより有用なアプローチを形成できる可能性があります

大規模言語モデルをテキストマイニングの前処理に

大規模言語モデル

テキストマイニング

- 分類ラベル付き文章データでファインチューニングした大規模言語モデル(BERT)によって、分析用の文章データを分類し、そのデータにテキストマイニングを実行すれば各分類の違いを単語ベースに理解できる
- 大規模言語モデル(BERT)は文脈に基づいた文章の特徴ベクトルをエンコードできるため、そのベクトル表現に基づいて文章をクラスタリングし、それぞれのクラスターに属する文章群にテキストマイニングを実行すれば、各クラスターの特徴を単語ベースに解釈できる
- 多言語のテキストデータをまとめてテキストマイニングしたいときに、大規模言語モデル(GPTやT5)で同一言語に翻訳してからテキストマイニングを実行する
- テキストマイニングの辞書作りにおいて、大規模言語モデル(GPT)を使って類義語の候補を生成する
- なお、大規模言語モデルは一度に大量のテキストデータを処理するとは得意ではないため、対象のテキストデータの量が多いときには、事前に分割してから入力するなどの工夫が必要である

テキストマイニングを大規模言語モデルの前処理に

テキストマイニング

大規模言語モデル

- まずは通常通りのテキストマイニングの分析を実行することで、特徴を可視化したり文章のクラスタリングをし、その結果から、さらに深掘りして注目すべきデータ対象を絞り込み、その限定された量のテキストデータに対して大規模言語モデルを適用すれば、その対象の文章理解をより深めることができる
- 絞り込んだテキストデータに対してGPTで要約生成を行えば、テキストマイニングで可視化された定量的な特徴は、どのような文章の内容を含んでいる特徴なのか定性的に理解できる
- 絞り込んだテキストデータに対してBERTで文章分類を行えば、注目した特徴をさらに細分化して分類でき、各分類の細かい傾向を理解できる
- テキストマイニングを大規模言語モデルの前処理として活用する場合は、最初から大量のテキストデータを対象とすることができる

2. 複数のAI技術を応用した特許文書データの新たな分析手法

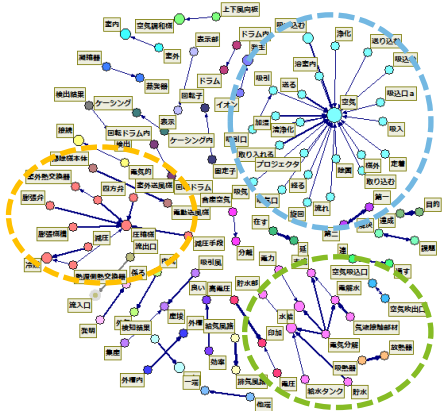
2. 複数のAI技術を応用した特許文書データの新たな分析手法

2-1. 従来の特許文書分析とその課題

これまでの特許文書分析

単語をベースに、あるいは手動でグルーピングしたカテゴリをベースに、全体の出現状況、経年変化、出願人の特徴、課題と解決手段の関係などを把握する分析がよく行われます

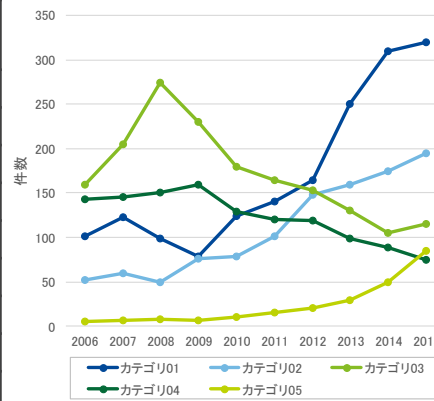
共起ネットワークによる全体像把握



- 単語の共起関係をネットワークで可視化する
- ネットワークのかたまりを見ながら、全体でどのような話題が形成されているのか考察する

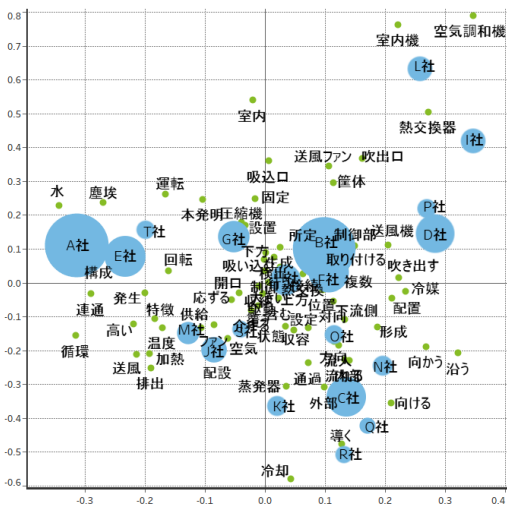
手動設定したカテゴリのトレンド把握

例) 掃除機カテゴリのリスト
掃除機
集塵
集塵容器
吸引力
サイクロン
塵埃->分離
塵埃->吸い込む
塵埃->収容
塵埃->遠心分離



- 抽出した単語を手動でいくつかのカテゴリにグルーピングする
- 各カテゴリの出願年ごとの出現頻度をグラフ化し、トレンドを把握する

コレスポンス分析による出願人の特徴把握



- 単語の出現データから共通して現れる特徴的な軸を2つ抽出する
- その2軸による平面上に単語と出願人を同時にマッピングする
- 出願人の周辺に配置された単語群から各出願人の特徴を考察する

課題と解決手段のクロス集計による関係把握

課題	解決手段カテゴリ						
	カテゴリ01	カテゴリ02	カテゴリ03	カテゴリ04	カテゴリ05	カテゴリ06	カテゴリ07
カテゴリ01	206	80	71	184	26	47	11
カテゴリ02	208	76	87	182	23	48	9
カテゴリ03	172	74	53	57	31	35	10
カテゴリ04	176	54	37	59	26	46	29
カテゴリ05	85	39	13	23	14	16	5
カテゴリ06	87	53	31	33	59	37	15

- 「要約」の【課題】と【解決手段】それぞれに対して出現単語のカテゴリを設定する
- 課題と解決手段のカテゴリのクロス集計をして、用途と技術の関連性を考察する

これまでの特許文書分析の課題と解決アプローチ

複数のAI技術を組み合わせることで、特許文書データを単語ベースではなく、客観的に抽出されるトピックベースで解釈し、そのトピックの統計的な関連性を分析できます

課題①

単語ベースの分析では
複雑で考察しにくい

課題②

カテゴリの設定が主観的で
作業負荷も大きい

課題③

課題と解決手段の統計的な
関係を分析していない

単語を賢く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

使われ方の似ている単語群を
トピックとして集約する

要因関係をモデリングする
人工知能技術

ベイジアンネットワーク

抽出したトピックに関わる要因
関係を統計的にモデル化する

2. 複数のAI技術を応用した特許文書データの新たな分析手法

2-2. AI技術の応用: PLSAとベイジアンネットワーク

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変数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

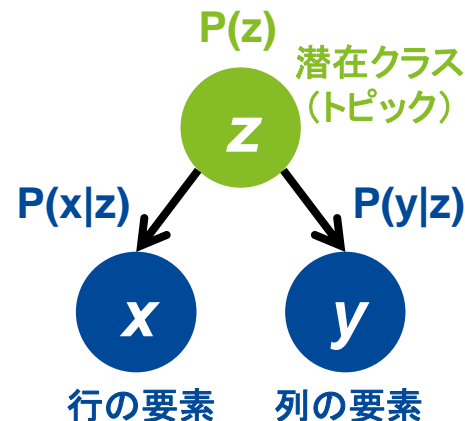
「文書×単語」行列 (共起行列)

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

複雑な観測情報を分かりやすく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
- $(m \times n)$ の行列を、 $(m \times k), (k \times k), (k \times n)$ 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

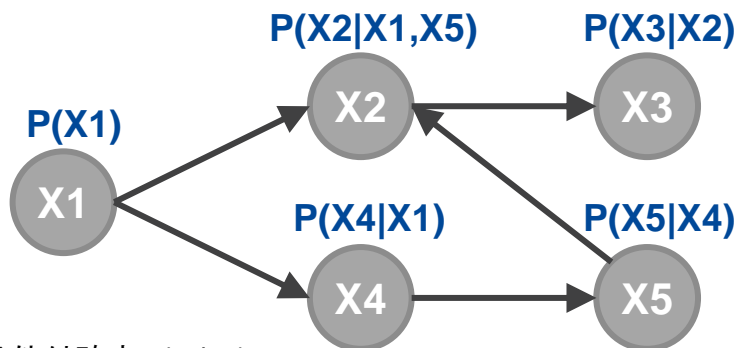
トピックモデル

- 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

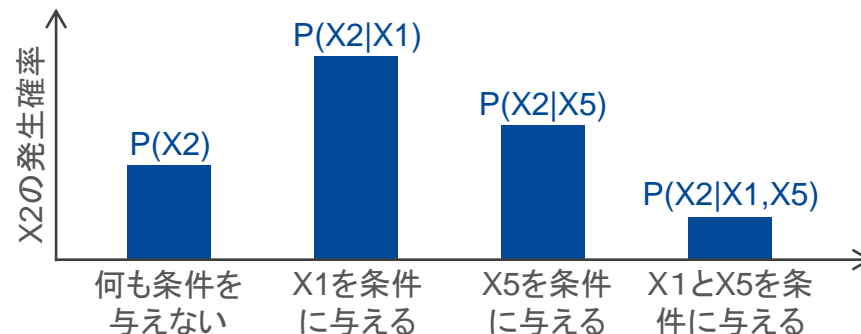
ベイジアンネットワークの概要

- 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、非線形の関係や相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

確率的因果関係と相互作用



- 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に対して相互作用がある(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相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、複合的な相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、複合条件で効果に変化する相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に変化する相互作用も表現できる

モデルの構造が不明

モデルの構造(要因関係)が理解できる

非線形のモデル化

線形のモデル化

非線形のモデル化

目的変数と説明変数の区別がある

区別がない

AI技術と言っても様々あり、例えば「理解系AI」「識別系AI」「生成系AI」に分類することができますが、それぞれの技術を分析目的に応じて賢く使いこなすことが求められます

理解系AI

現状のデータに潜む特徴や要因関係を理解するAIであり、ホワイトボックスのモデルが求められる

PLSA

LDA

決定木

ベイジアンネットワーク

「理解系AI」により、データに潜む傾向を人間が理解し、ビジネスアクションを人間が考え実行する

※私見による分類であり、一般的に定義された分類ではありません

識別系AI

画像判定や文章分類など、新規のデータを識別するAIであり、精度さえ良ければモデルはブラックボックスでもよい

深層学習・CNN

RNN・LSTM・NMT

Transformer・T5

BERT

生成系AI

入力した情報に対して画像や文章を生成するAIであり、精度さえ良ければモデルはブラックボックスでもよい

「識別系AI」や「生成系AI」は人間がデータを理解して業務に活用するのではなく、業務の自動化・省力化を目的としていることが多い

GAN

GPT

Diffusion model

2. 複数のAI技術を応用した特許文書データの新たな分析手法

2-3. Nomolytics:

PLSAとベイジアンネットワークを応用した新たなテキスト分析手法

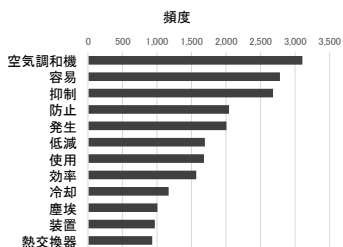
膨大なテキストデータをトピックに変換して解釈を容易にし、テキスト情報内に潜む要因関係をモデル化して、ビジネスアクションに有用な特徴を把握可能にします

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

単語抽出



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

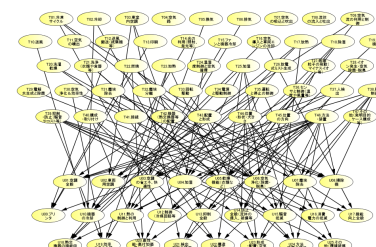
トピック類型化



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

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

3. Nomolyticsを適用した特許分析事例①

3. Nomolyticsを適用した特許分析事例①

3-1. 「風・空気」に関する特許文書データ

「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

3. Nomolyticsを適用した特許分析事例①

3-2. 分析プロセスの全体像

3. Nomolyticsを適用した特許分析事例①

3-3. トピックの抽出

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×係り受けで構成される共起行列に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

【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
...	...

共起行列の作成

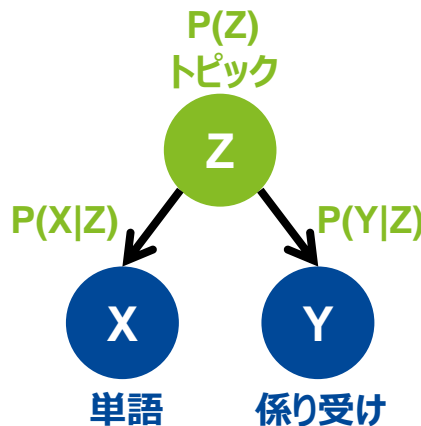
抽出した単語と係り受け表現に基づいて、「単語×係り受け」の共起行列(文章単位で同時に出現する頻度のクロス集計表)を作成する

	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
単語	1578	100	4	1
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
...

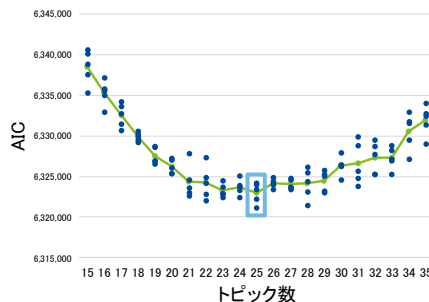
共起行列の構成(それぞれ頻度10件以上を対象)
 課題: 単語(3,256語)×係り受け(2,084表現)
 解決手段: 単語(5,187語)×係り受け(7,174表現)

PLSAの実行

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



トピックの抽出

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

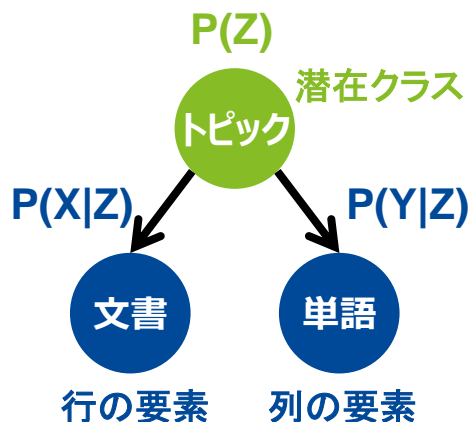
トピックにおける $P(X|Z)$ と $P(Y|Z)$ からトピックの意味を解釈する

トピック T32			
P(Z) = 2.7%			
P(X Z)	単語	P(Y Z)	係り受け
5.5%	送風機	2.1%	塵埃-分離
5.2%	塵埃	1.7%	分離-塵埃
4.1%	掃除機	1.7%	塵埃-含む
3.6%	分離	1.5%	吸い込む-塵埃
3.5%	吸い込む	1.3%	含む-空気
2.3%	集塵部	1.0%	空気-分離
1.9%	配置	1.0%	送風機-吸い込む
1.9%	集塵容器	1.0%	発生-送風機
1.6%	旋回	0.9%	含塵空気-分離
1.5%	含塵空気	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

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

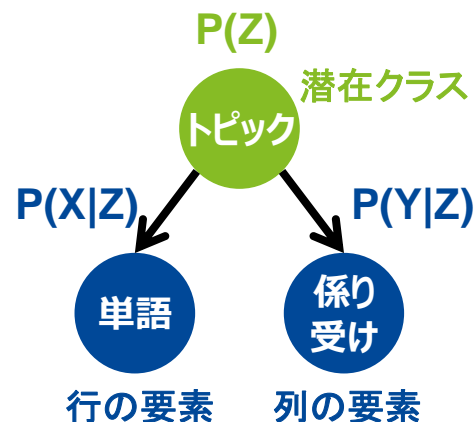
一般的なPLSAの共起行列



	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	0	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

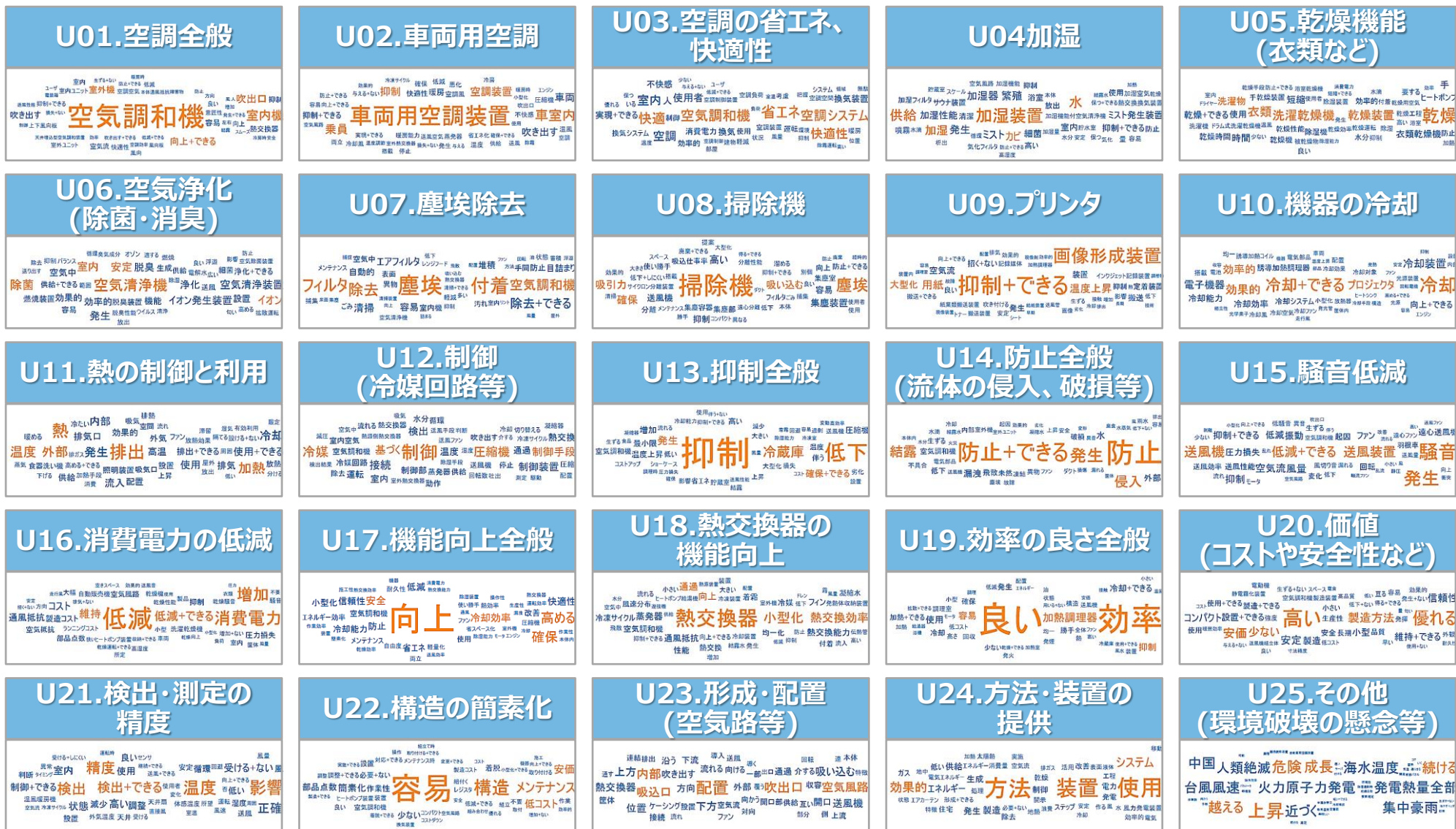


	係り受けa	係り受けb	係り受けc	係り受けd	...
単語1	325	264	11	20	
単語2	241	201	6	8	
単語3	28	41	288	14	
単語4	9	15	4	172	
...					

- 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、どちらもそれぞ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

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



3. Nomolyticsを適用した特許分析事例①

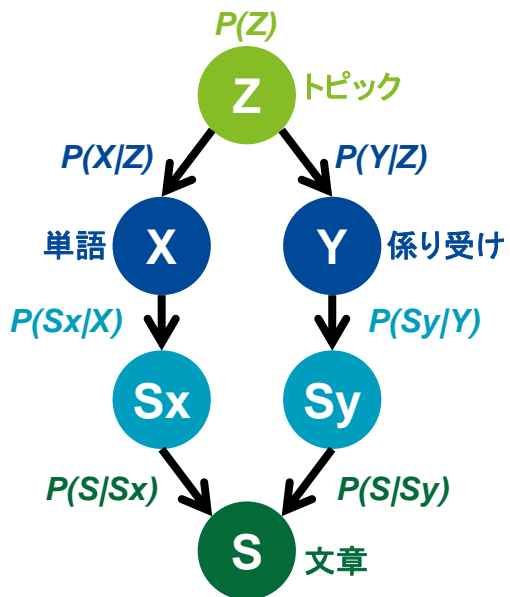
3-4. トピックのスコアリング

トピックのスコアリング

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 $Sx_h = \{X_1, X_2, \dots, X_i\}$ 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 $Sy_h = \{Y_1, Y_2, \dots, Y_j\}$ 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(該当有無のフラグ情報)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	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用途と技術の関連性の分析

トピックをベースにした分析によって読むべき特許文書を効率的に絞り込むことができる

3. Nomolyticsを適用した特許分析事例①

3-5. 出願年×トピックによるトレンド分析

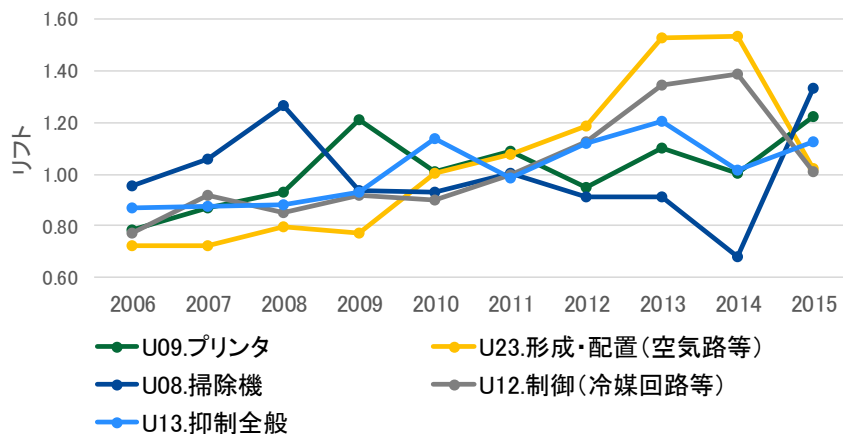
【分析目的】

技術や用途のトレンドを把握し、有望なシーズやニーズを探り、今後の技術開発戦略を検討する

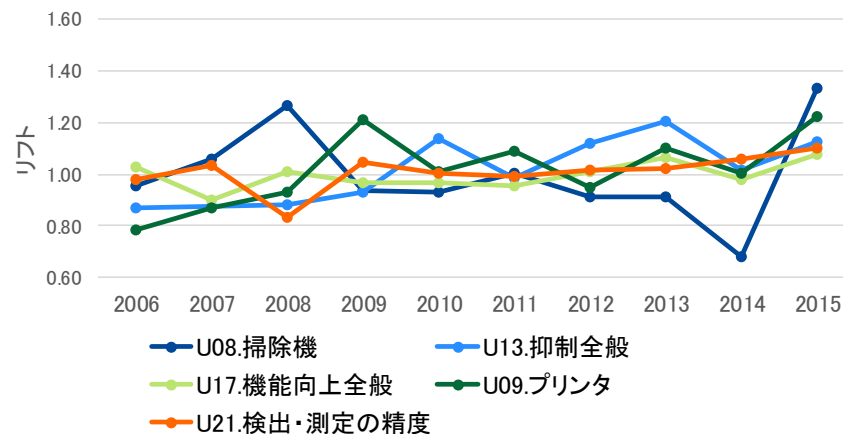
用途トピックの上昇トレンド

近年は掃除機や空気浄化、塵埃除去、プリンタに関する用途が上昇しています

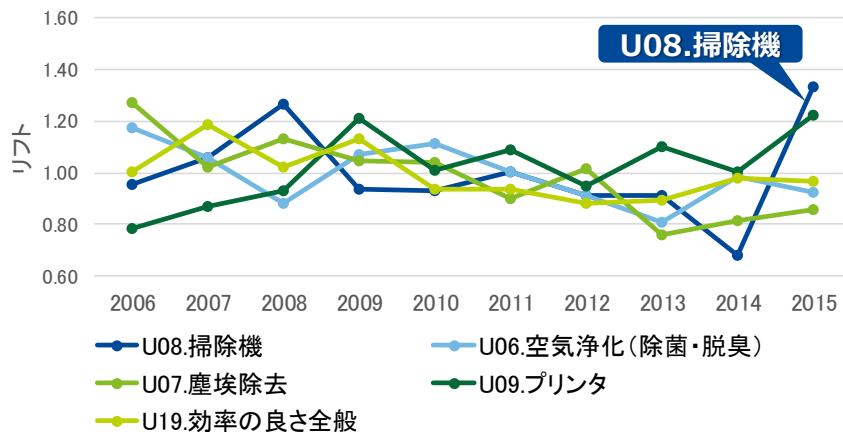
【長期】 2006年からの上昇率 best5



【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

$$P(\text{出願年} | \text{トピック} T_x = 1)$$

$$P(\text{出願年})$$

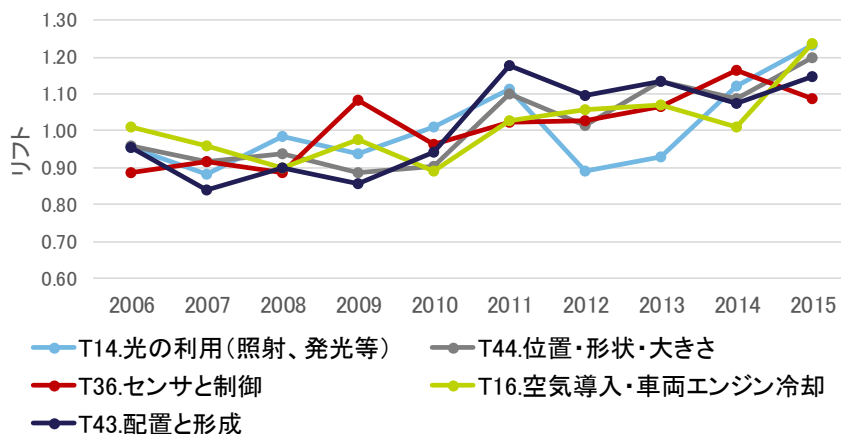
- リフト値は出願年とトピックの関係を示す指標

- トピック毎の各出願年の出願割合を、その出願年の出願割合で正規化した値

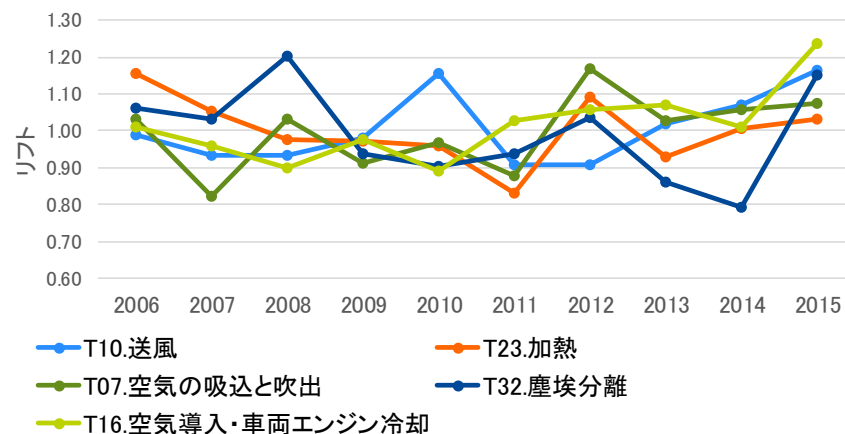
技術トピックの上昇トレンド

近年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

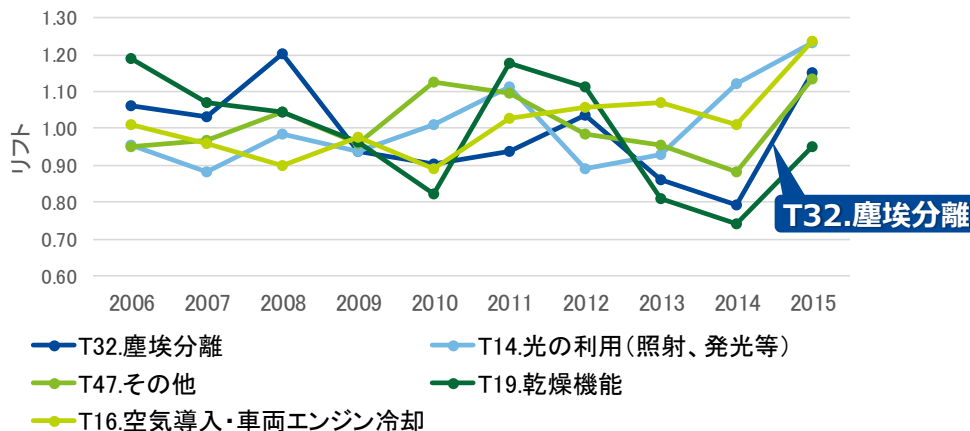
【長期】 2006年からの上昇率 best5



【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

$$P(\text{出願年} | \text{トピック} T_x = 1)$$

$$P(\text{出願年})$$

- リフト値は出願年とトピックの関係を示す指標

- トピック毎の各出願年の出願割合を、その出願年の出願割合で正規化した値

3. Nomolyticsを適用した特許分析事例①

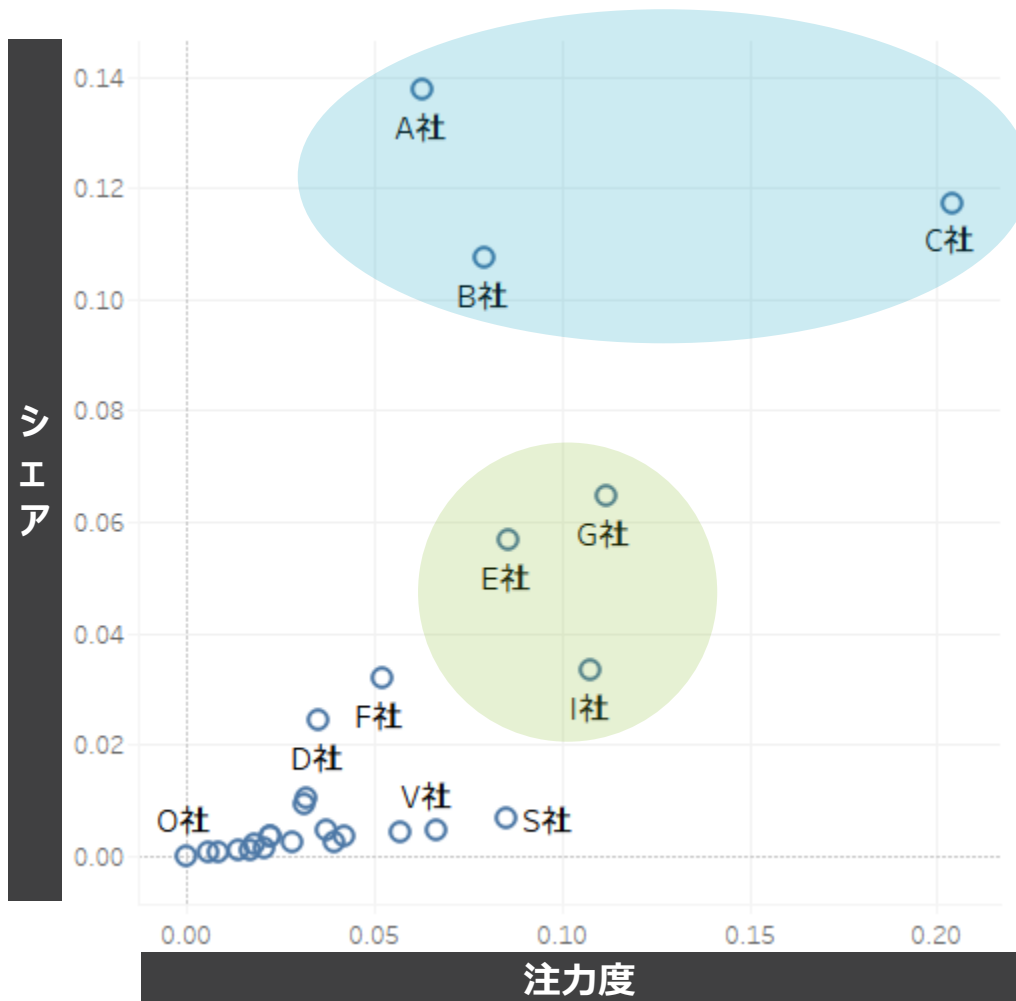
3-6. 出願人×トピックによる競合分析

【分析目的】

各出願人の動向や、業界における棲み分け、競合他社と自社との関係性などを把握し、他社との差別化戦略や協業戦略を検討する

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

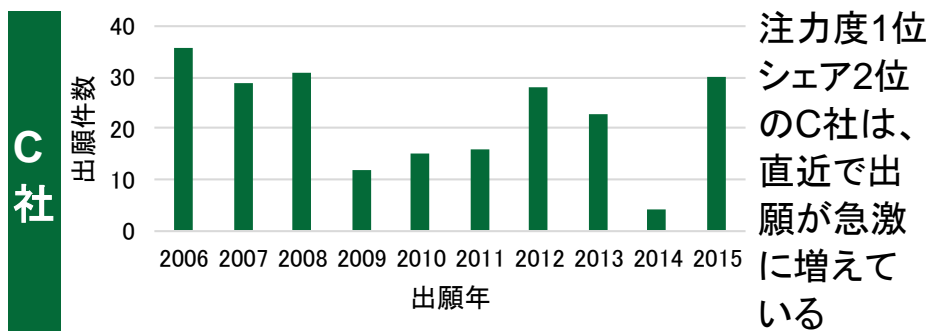
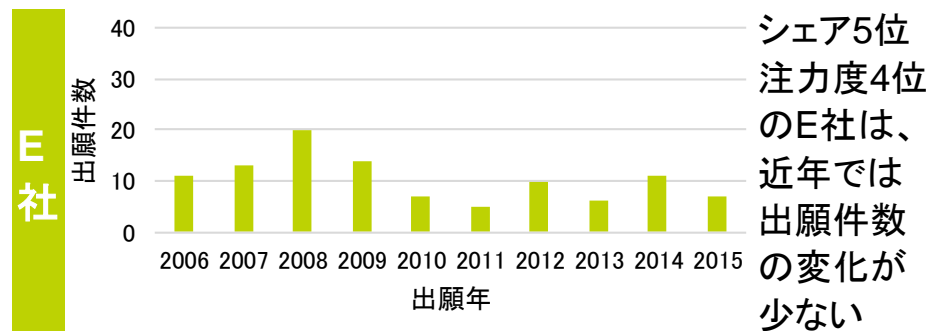
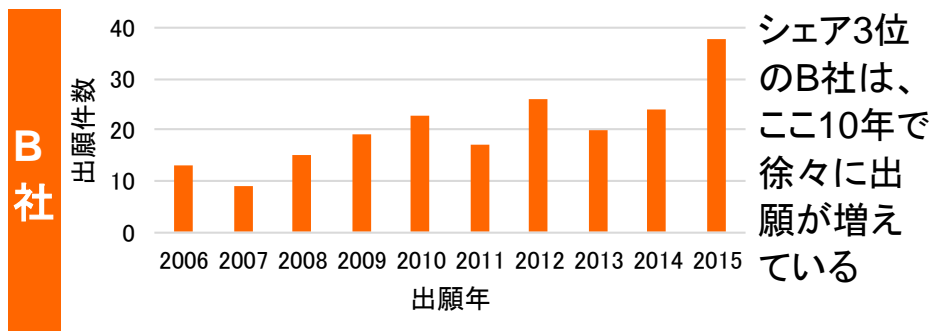
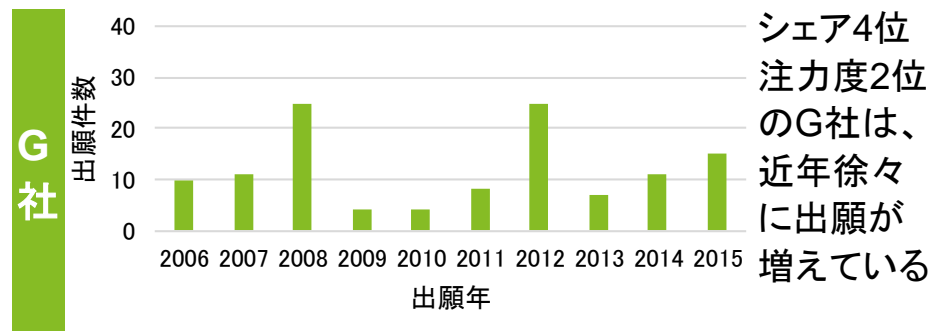
注力度とシェア

- **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

技術「T32.塵埃分離」の各出願人の出願トレンド

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



3. Nomolyticsを適用した特許分析事例①

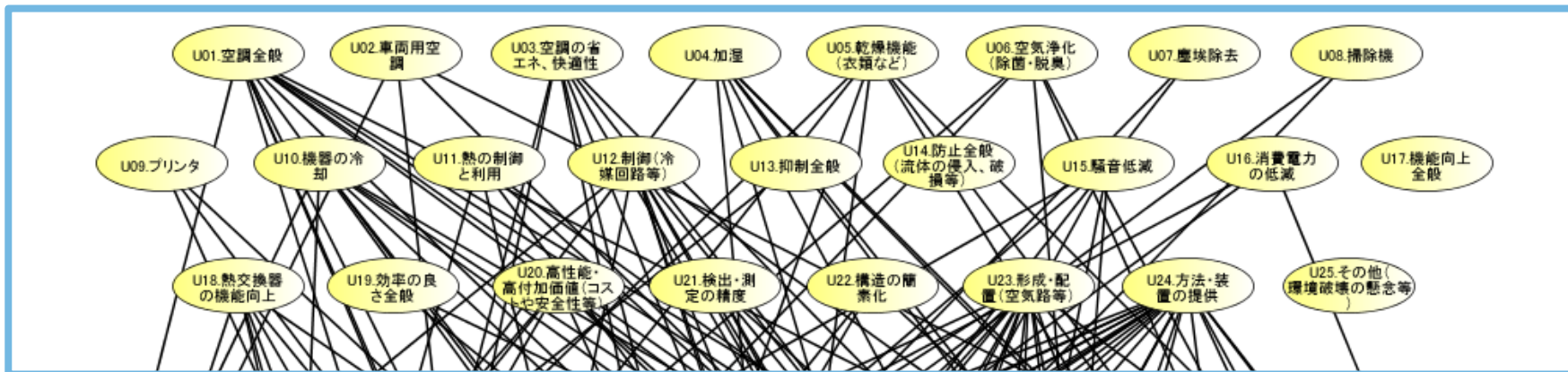
3-7. 用途×技術の関係分析<その1> ～用途⇒技術の関係～

【分析目的】

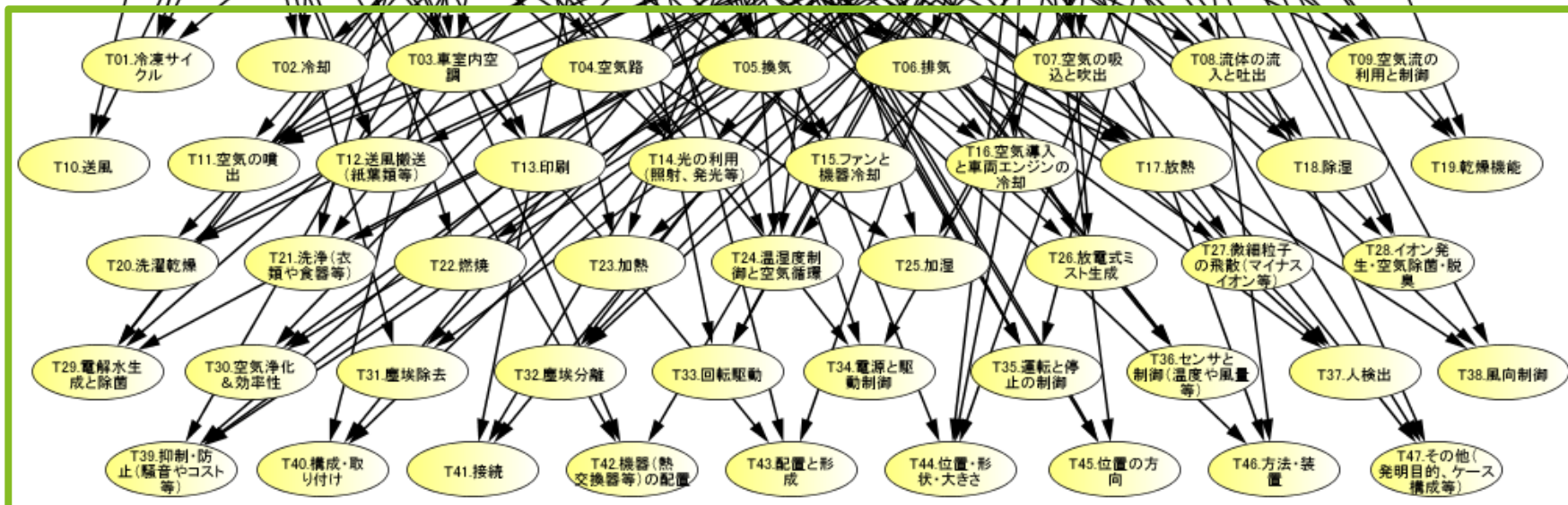
自社で検討中の用途実現のために重要な解決技術や代替技術、競合他社の存在を把握し、用途の事業化のための開発戦略や他社との協業戦略を検討する

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



用途トピック

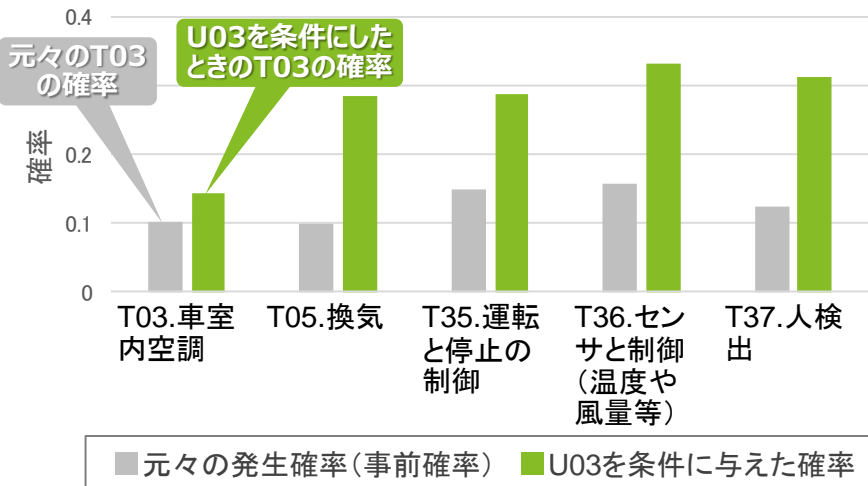


技術トピック

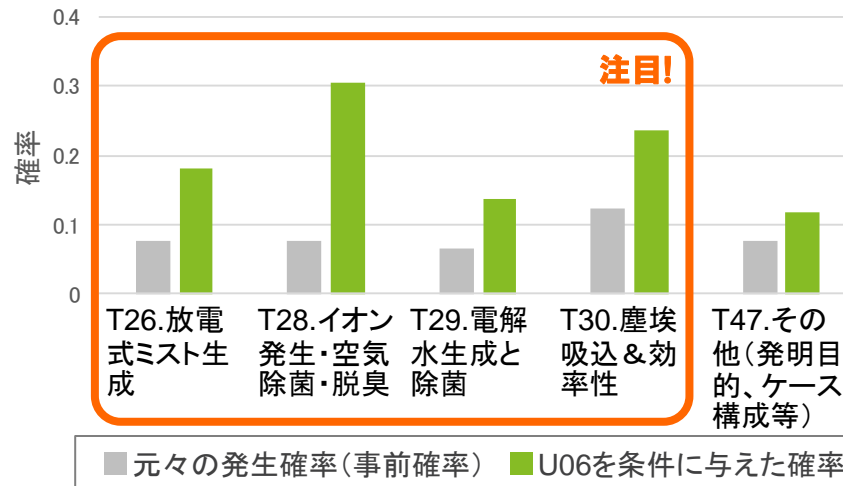
用途と関係のある技術の確認

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

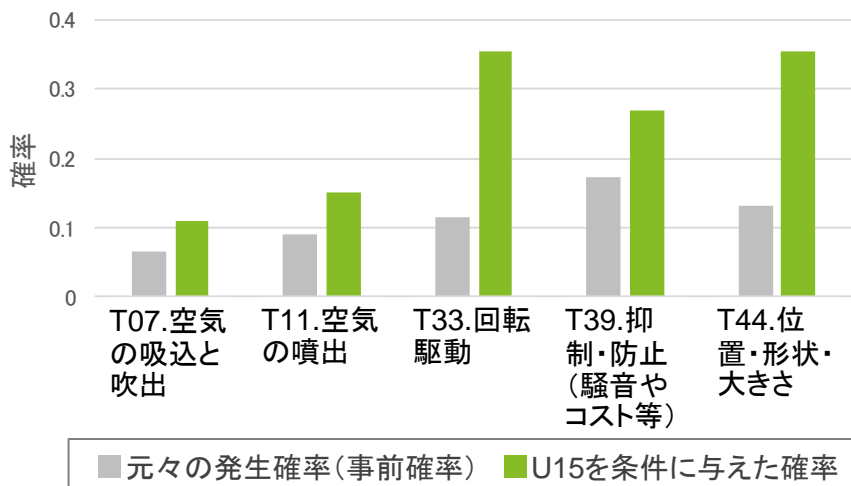
「U03.空調の省エネ、快適性」と関係のある技術



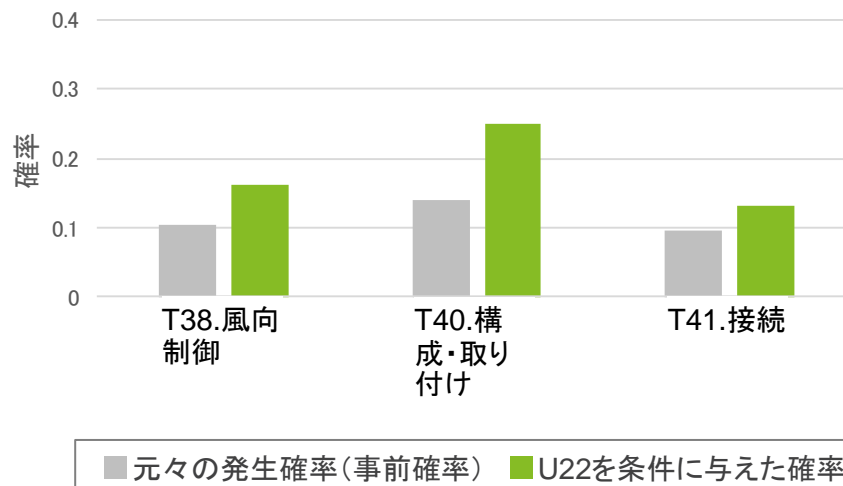
「U06.空気浄化(除菌・脱臭)」と関係のある技術



「U15.騒音低減」と関係のある技術



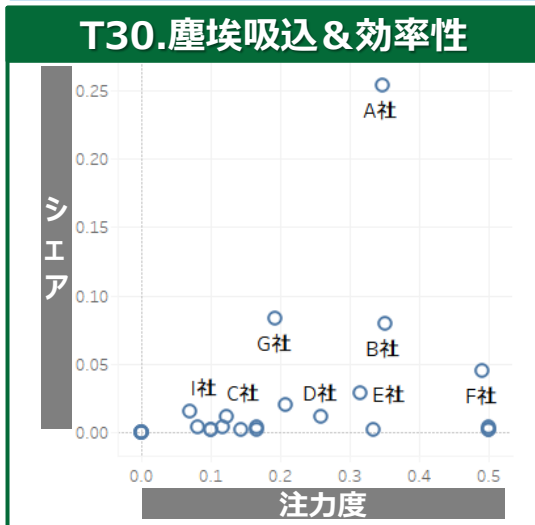
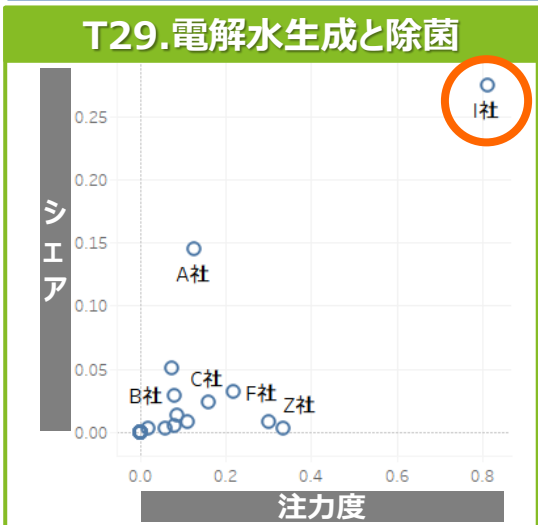
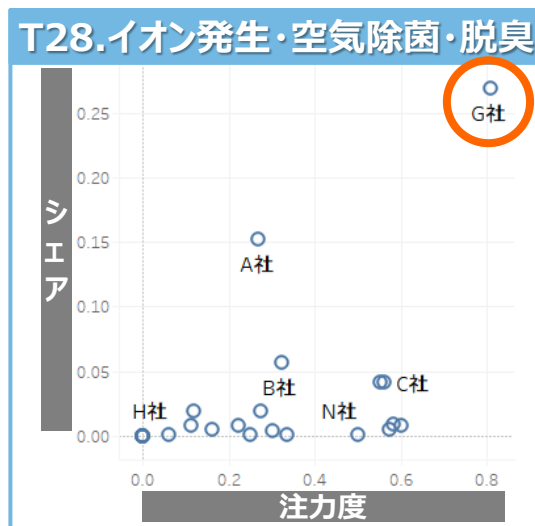
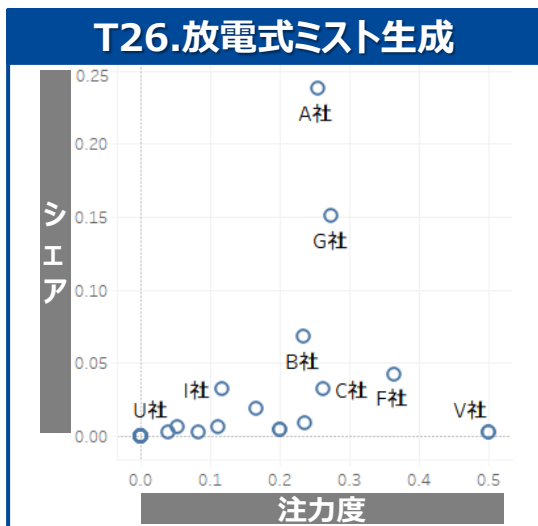
「U22.構造の簡素化」と関係のある技術



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

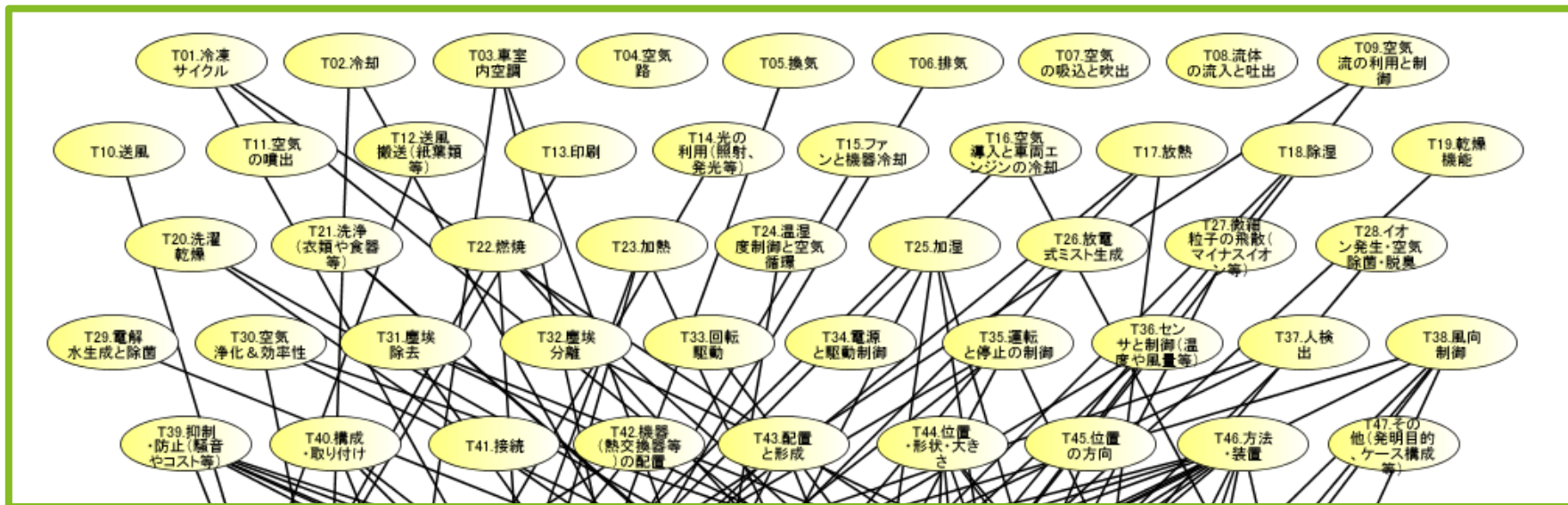
3. Nomolyticsを適用した特許分析事例①

3-8. 用途×技術の関係分析<その2> ～技術⇒用途の関係～

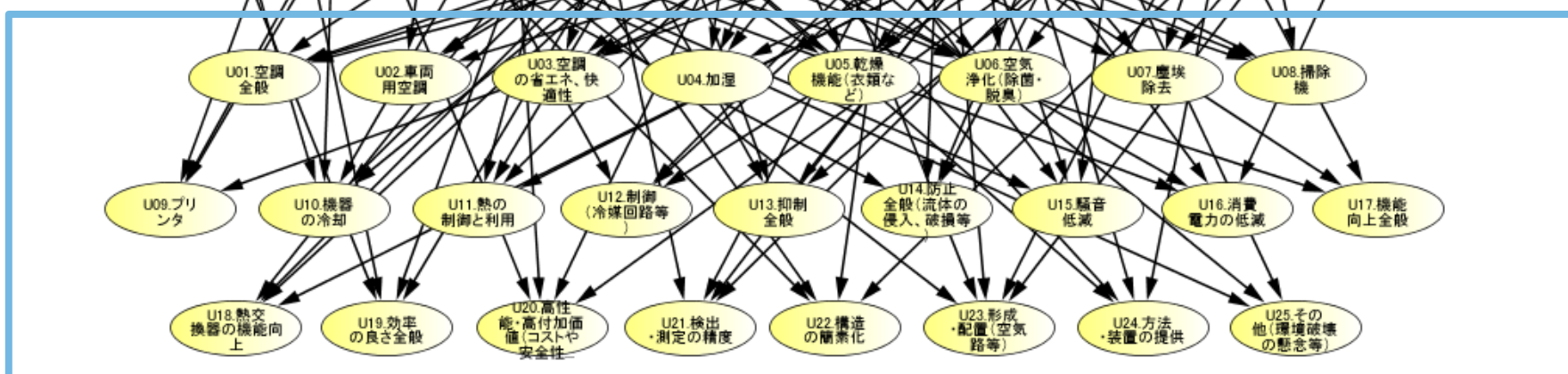
【分析目的】

自社技術と関係のある用途を把握し、まだ自社で想定していない用途を見つけ、保有技術を有効活用できる新しい用途展開のアイデアを創出する

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



技術トピック

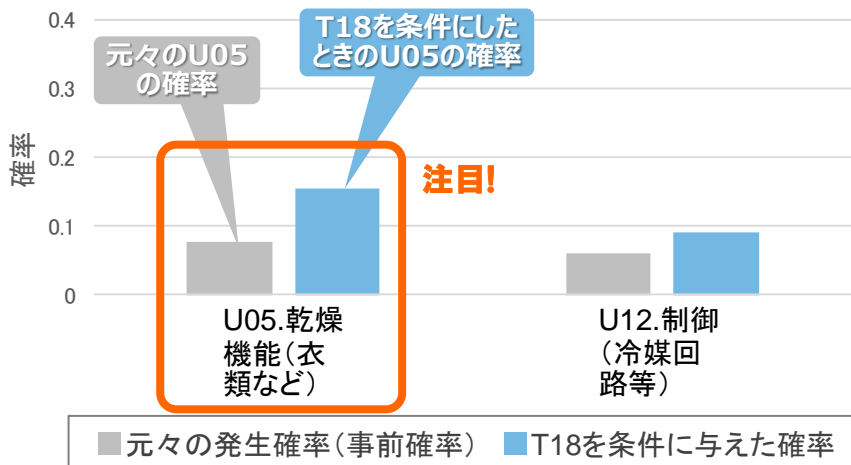


用途トピック

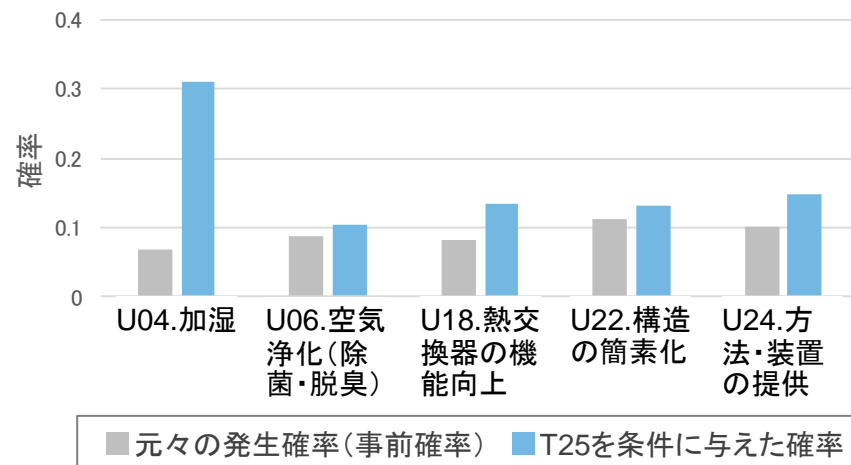
技術と関係のある用途の確認

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

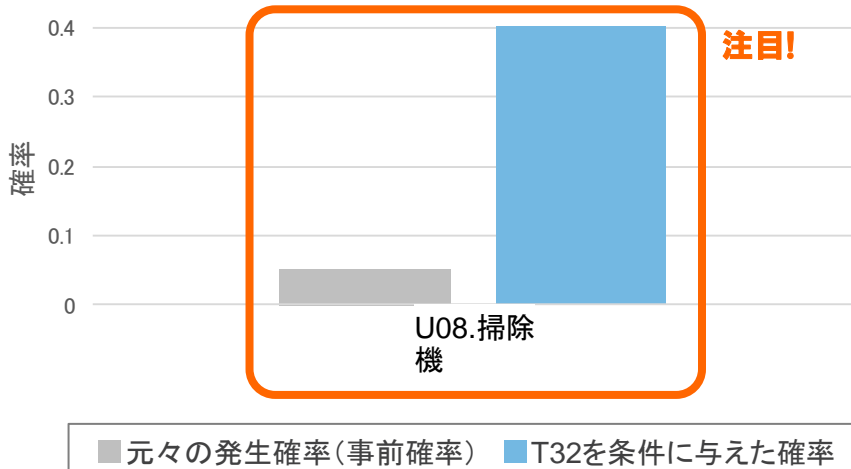
「T18.除湿」と関係のある用途



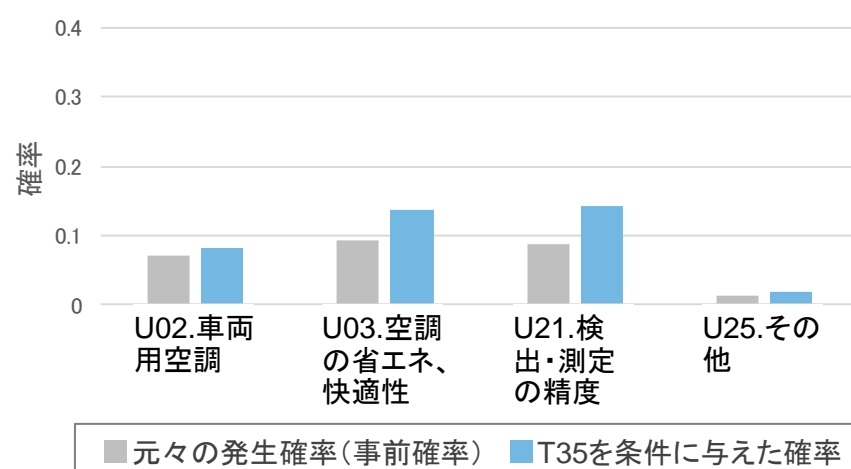
「T25.加湿」と関係のある用途



「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



「T18.除湿」の技術を「U05.乾燥機能」の用途で応用するアイデア創出

印刷機の中でインク液を吸収した用紙の湿気をムラなく取り除く除湿技術は、洗濯乾燥機の中で洗濯物をムラなく効率的に乾燥させることにも応用できるかもしれません

「乾燥機能」を想定した「除湿」の特許例

発明の名称

ドラム式洗濯乾燥機

【課題】

洗濯物を短い時間でムラ無く乾燥させ、乾燥工程の時間を短くすることができるドラム式洗濯乾燥機を提供する。

【解決手段】

送風機に吸い込まれた空気は、風路切替弁の切り替えにより、ドラム開口部に対向する前側吹出口へ流れたり、回転ドラムの後部に設けられた後側吹出口へ流れたりする。制御装置が風路切替弁の切り替えを制御することによって、恒率乾燥過程時、前側吹出口から乾燥用空気が吹き出し、かつ、減率乾燥過程時、後側吹出口から乾燥用空気が吹き出す。これにより、恒率乾燥過程において乾燥用空気が効果的に当たらなかった、回転ドラムの後端壁側の洗濯物に、乾燥用空気が減率乾燥過程で効果的に当たる。

「乾燥機能」を想定していない「除湿」の特許例

発明の名称

インクジェット記録装置及び画像記録方法

【課題】

処理液の厚みムラを低減するとともに処理液による用紙のコックリングを低減することで、高品質かつ高速の画像記録を可能とするインクジェット記録装置及び画像記録方法を提供する。

【解決手段】

記録媒体に処理液を付与する処理液付与部の後段には、記録媒体表面に残存する溶媒を蒸発させるプレ加熱部が設けられている。プレ加熱部はIRプレヒータにより記録媒体表面を輻射加熱するとともに、吸引ファンにより記録媒体表面の湿り空気を置換する。液状の処理液が不均一にならないように乾燥処理を施すことで、均一な膜厚を持つ固体状の凝集処理層が形成される。その後、本加熱部による熱風噴射加熱により、コックリング量が所定量以下になるように本加熱処理が施される。

※対外説明用のため要約文は一部加工している

「T32.塵埃分離」の技術を「U08.掃除機」の用途で応用するアイデア創出

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】

塵埃を含む空気を回転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



ダイソンの吸引力が落ちないサイクロン掃除機は、製材工場の屋根にあった木くずと空気を分離するサイクロン装置をヒントに生まれた

サイクロン掃除機の技術はダイソンの様々な商品に応用されている

羽のない 扇風機

空気清浄 ファンヒーター

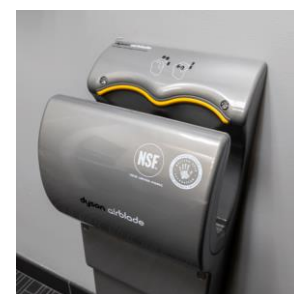
加湿器



ヘアドライヤー

ヘアスタイラー

ハンドドライヤー



3. Nomolyticsを適用した特許分析事例①

3-9. Nomolyticsによる特許文書分析のまとめ

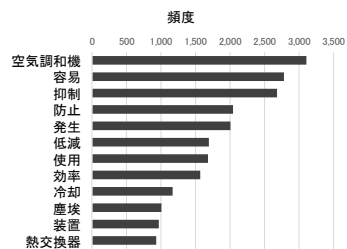
膨大なテキストデータをトピックに変換して解釈を容易にし、テキスト情報内に潜む要因関係をモデル化して、ビジネスアクションに有用な特徴を把握可能にします

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

単語抽出



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

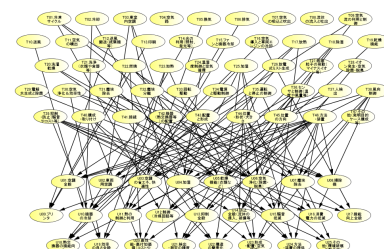
トピック類型化



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

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

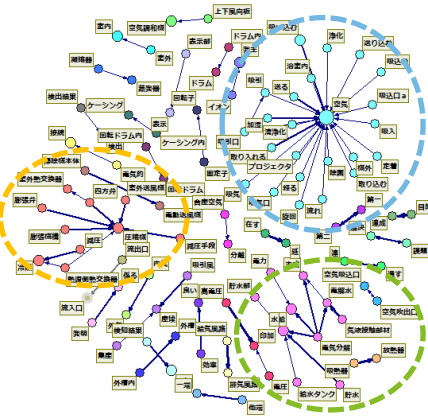
条件を変化させたときの効果を確率的にシミュレーションでき、有効なアクションを検討できる

Nomolyticsを適用した特許分析のメリット①

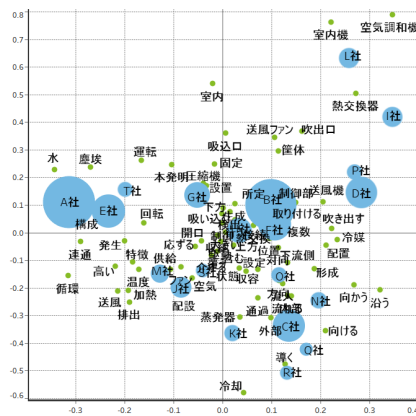
単語ではなく集約されたトピックをベースにした分析を実行することで、膨大な特許情報に潜む特徴を分かりやすく理解することができます

従来の特許分析

単語の共起ネットワーク



単語と出願人の対応マップ



カテゴリのリスト作成

掃除機カテゴリのリスト	
掃除機	塵埃->分離
集塵	塵埃->吸い込む
集塵容器	塵埃->收容
吸引力	塵埃->遠心分離
サイクロン	含塵空気->分離

■ 単語ベースの複雑なアウトプットから、文章全体に存在する話題を解釈したり、各出願人の特徴を把握しなければならない

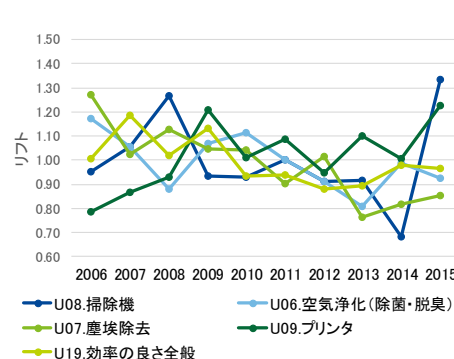
■ 単語を人手でグルーピングしていくつかのカテゴリを作成し、カテゴリベースに分析するものの、そのカテゴリ作成は属人的で作業負荷も大きい

Nomolyticsを適用した特許分析

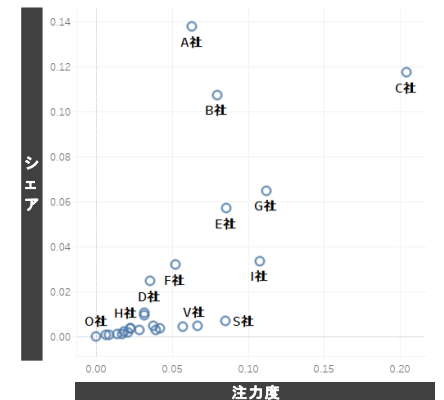
PLSAによって機械的に抽出されたトピック



トピック別のトレンド



トピック別の出願人の配置



■ 文章全体に存在する話題をPLSAで機械的に抽出できる

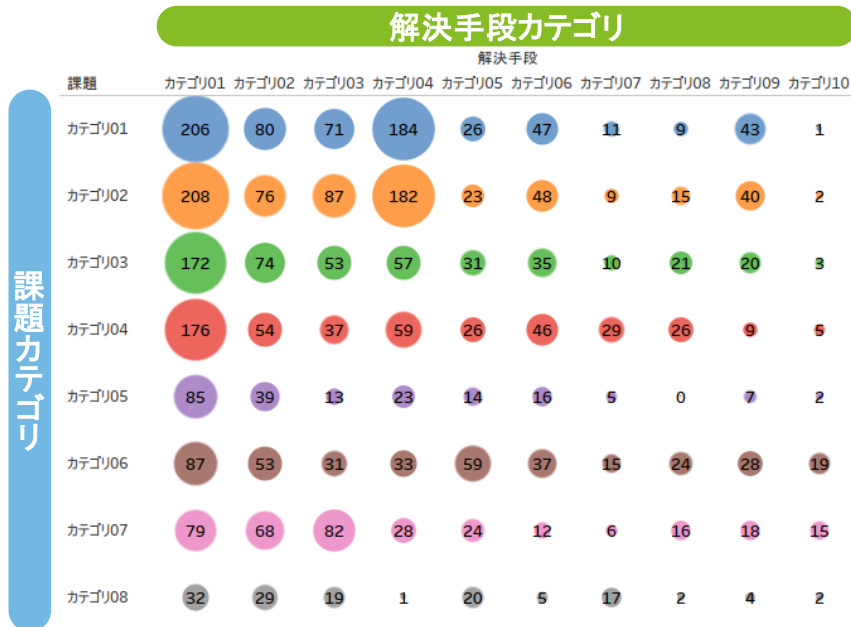
■ 単語ではなくトピックをベースにトレンドや各出願人の特徴を分析し、分かりやすく理解することができる

Nomolyticsを適用した特許分析のメリット②

用途と技術の統計的な関係を把握することで、用途を実現するための重要技術を確認して技術戦略を検討したり、自社技術を有効活用できる新規用途のアイデアを創出できます

従来の特許分析

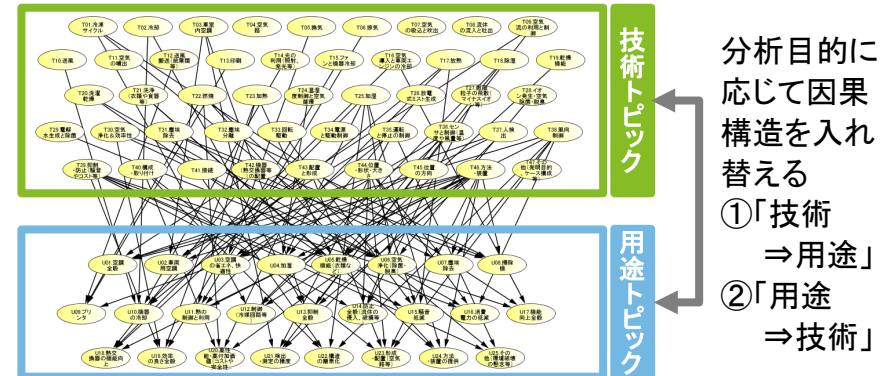
課題と解決手段のカテゴリ間のクロス集計



- 【課題】と【解決手段】それぞれに対して人がグルーピングして作成したカテゴリのクロス集計表を作成し、その対応関係を考察する
- その組み合わせで出願件数が多いからといって、統計的に意味のある関係であるとは限らない(全体的に出願件数が多いだけの可能性もある)

Nomolyticsを適用した特許分析

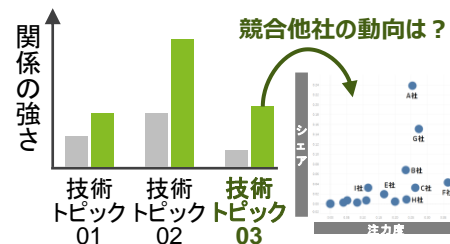
課題と解決手段のトピック間の統計的な因果関係モデル



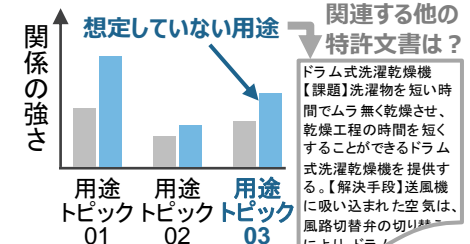
分析目的に応じて因果構造を入れ替える

- ①「技術 ⇒ 用途」
- ②「用途 ⇒ 技術」

想定用途と関係のある技術



自社技術と関係のある用途



- 客観的に抽出されたトピックをベースに課題と解決手段(用途と技術)の統計的な関係をベイジアンネットワークで把握できる
- 検討中の用途に対して、関係の強い技術を確認し、各技術における出願人の動向から自社の技術戦略を検討できる
- 自社技術と関係の強い用途で想定していないものを確認し、その関連特許の探索から技術の新規用途アイデアを創出できる

4. Nomolyticsを適用した特許分析事例②

4. Nomolyticsを適用した特許分析事例②

4-1. 電気自動車に関する特許文書データ

電気自動車関連の10年分の特許データ26,419件の要約全文を対象に、テキストマイニングとPLSAでトピックを抽出し、各トピックの特徴を可視化します

データの抽出条件と分析対象

- 対象
 - 公開特許公報
- キーワード
 - 要約と請求項に「車」と「電気」を含む
- 出願日
 - 2007年1月1日～2016年12月31日
- 抽出方法
 - Patent Integrationを使用
- 抽出件数
 - 26,419件
- 分析対象
 - 要約の全文



分析プロセス

テキストマイニング

要約文から単語と係り受けを抽出

トピック化 (PLSA)

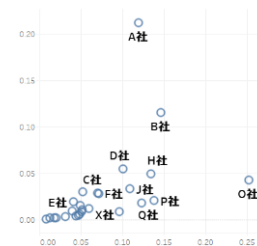
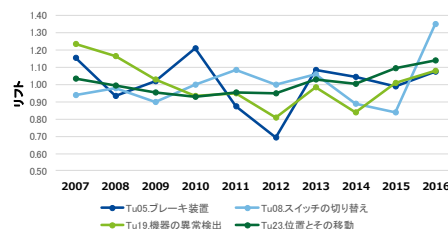
単語と係り受けを複数のトピックに集約

トピックのスコアリング

全データに対する各トピックのスコアを計算

特徴の可視化

トピックのスコアを属性を軸に集計・可視化



4. Nomolyticsを適用した特許分析事例②

4-2. トピックの抽出

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 × 係り受けで構成される共起行列に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

要約の全文章に含まれる「単語（名詞）」と「係り受け」を抽出する

単語	頻度
構成	4,997
制御	4,360
配置	3,895
モータ	3,486
形成	3,459
供給	3,309
検出	3,215
電気自動車	3,181
...	...

係り受け表現	頻度
電力⇒供給	1,208
否⇒判定	517
モータ⇒駆動	460
バッテリー⇒充電	440
効率⇒良い	419
供給⇒電力	332
電気自動車⇒提供	285
充電⇒行う	273
...	...

共起行列の作成

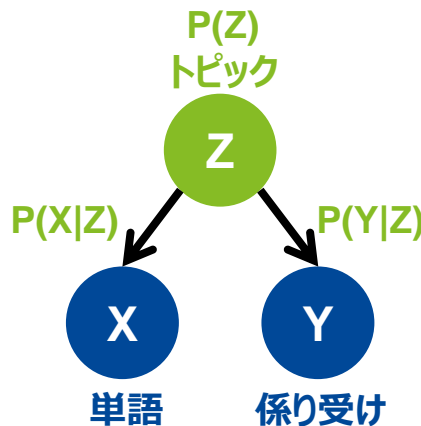
抽出した単語と係り受け表現に基づいて、「単語 × 係り受け」の共起行列（文章単位で同時に出現する頻度のクロス集計表）を作成する

	係り受け表現				
	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	...
構成	118	33	36	33	
制御	268	73	108	85	
配置	69	2	29	8	
モータ	239	61	494	58	
...					

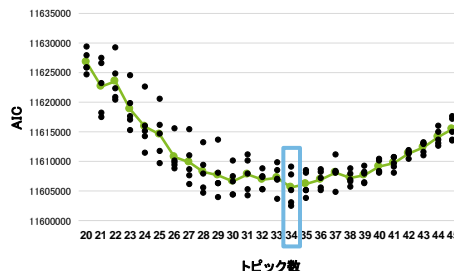
単語（名詞）：3,020語
係り受け：2,128表現
※頻度20件以上を対象

PLSAの実行

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



トピックの抽出

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 Z13	
P(Z) = 5.0%	

P(X Z)	単語	P(Y Z)	係り受け
12.6%	充電	5.1%	バッテリー-充電
8.9%	電気自動車	4.0%	充電-行う
6.5%	蓄電装置	3.9%	電気自動車-充電
3.0%	バッテリー	1.9%	蓄電池-充電
2.0%	充電システム	1.6%	蓄電装置-充電
2.0%	蓄電池	1.6%	電力-供給
1.9%	電力	1.3%	充電-開始
1.7%	制御	1.2%	電気自動車-接続
1.5%	充電スタンド	1.2%	充電-蓄電装置
1.5%	放電	1.1%	用いる-充電
...

確率の高い構成要素から、トピック Z13は「電気自動車の蓄電池充電」に関するトピックと解釈できる

特許トピック34個の一覧①

26,419件の特許は、エンジン、動力伝達、モータ、ブレーキ、電力変換、二次電池、充電、情報通信、異常検出、筐体、構成、小型化、安全性などの34個のトピックに集約されました

Z01.エンジンの始動と停止

検出 充電 駆動
運転蓄電装置 充電状態要求動作 駆動輪 判定 内燃機関
制御手段 クラッチ 発電機 モータ走行 制御

停止 ECU ハイブリッド車両 エンジン

動力 制御装置 開始 モータジェネレータ バッテリ
成立 再始動駆動力 発電

Z02.動力の伝達

動力 回転
制御 制御装置 ハイブリッド車両 車輪 構成 トルク駆動部
モータ クラッチ 駆動輪 エンジン

内燃機関 変速機 トランスミッション 伝達 連結

入力 入力軸駆動装置 モータジェネレータ 駆動 出力軸
電気機械 駆動輪 発電機 駆動力 配置

Z03.モータ駆動

検出 駆動 制御
配置 走行 制御装置 回転 電気自動車 トルク
制御 駆動輪 構成 駆動力
車輪 バッテリ エンジン 駆動
駆動輪 電力 充電機 駆動力
油圧ポンプ 供給 電気エネルギー 車体電源 インバータ
ハイブリッド車両

Z04.ロータ・ステータなど回転部品の構成

一体 同軸対向 回転体 中心 ロータ
ファン 軸方向 配置 固定 固定方向 外周 周囲
回転軸 支持 回転 軸 シャフト 回転+できる
固定子巻線 形成 永久磁石 構成 ステータ 収容 モータ
コイル ハウジング 磁石 反対側

Z05.ブレーキ装置

演算 液圧付与 制御
駆動 マスシリンダ 操作 操作者 駆動部材
電気ブレーキ 制御 トルク 入力 入力装置 電気信号
制動力車輪 ブレーキ プレーキペダル 作動
伝達 モータブレーキ 液圧 操作量 検出 運転者
モータシリンダ装置 プレーキ操作固定 調整

Z06.動作制御

停止 駆動調整 回転速度 供給 トルク 給電 充電電 アクション
入力 制御 制御回路 ECU 速度 制御手段
動作モータ 制御 制御装置
インバータ 検出 制御部 演算 電力

Z07.動力伝達の制御

変速 駆動器 制御 構成 変速時変更
差動状態 燃費 変速部 駆動輪 制御 電気式差動部
エンジン 車両用動力伝達装置 制御装置
動力伝達経路 モータ 車両用駆動装置 運転状態
変化 回転部材 ハイブリッド車両 差動機構 差動部 変速ショック
変速比 ハイブリッド車両 差動機構 差動部 変速ショック
回転速度 連結

Z08.スイッチの切り替え

給電 インバータ 遮断 負荷 電力 制御 直列 直列
制御装置 電流 電気負荷 リレー 検出 車両用電源装置
並列 バッテリ 電圧 スイッチ コンデンサ
モータ DCインバータ オン制御 オフ 放電蓄電装置
スイッチング素子 充電
通電 印加

Z09.交流・直流の変換

モータ制御 電圧 バッテリ 検出 負荷 電気制御装置
電気自動車 直流 電力変換装置 制御装置
直流電力 インバータ 交流 電力 変換
動作 直流電圧 供給 交流電圧 駆動 入力 変換
コンバータ 非接触 電力変換装置 制御部 交流電力
生成

Z10.エネルギーの変換

生成 風車 水車 水車 供給 走行中
設置 回転蓄電 回収 蓄電池
エネルギー 電気エネルギー 発電機
蓄積 発電システム 熱エネルギー エンジン
バッテリー 電力 回転力 充電 走行 発電 運動エネルギー

Z11.電池モジュールの提供

一対 並列 接続 構成 冷却 連結
組合 電気接続 積層 コント 構成 冷却 連結
直列 単電池 バッテリ 電池パック バスバー
装置 バッテリケース 位置 配置
電源 電気自動車 形成 相互 電池モジュール
收容

Z12.二次電池の構成

活物質 收容 積層 電気集電 形成 部構造 電解質
構成 電解液 電池特性 対向表面 形成 負極活物質
正極活物質 二次電池 正極 負極
セパレータ 電極 非水電解質 リチウムイオン電池
非水電解質電池 集電体 配置 含有

Z13.電気自動車の蓄電池充電

蓄電池 充電システム
電力 充電ケーブル 充電+できる 充電電 充電時
放電 検出 情報 構成 充電システム
バッテリー 電気自動車 充電 開始 制御部
充電スタンド 外部電源 制御 コーダ 蓄電装置
充電制御装置 充電制御部 供給

Z14.非接触受電など給電装置

外部 駐車 電圧 プラグ 給電部 送電 受電部
駐車スペース 電気自動車 給電+できる
電力供給システム 制御 受電 給電装置
非接触 電源装置 送電コイル バレット 電力
受電コイル 給電制御部 供給

Z15.外部への電力供給

蓄積 蓄電池 作動 消費
外部電源装置 外部電源負荷 放電 モータ制御装置
発電機 供給+できる 駆動 電源
電気負荷 制御 検出 電力
蓄電 高電圧 バッテリ 電気機器 エンジン 充電 電圧 バッテリ 蓄電装置 制御部

Z16.空調などの冷却・加熱

暖房 構成 熱 電力 冷却水
車室内 燃料冷却装置 空気 冷媒 通電 排出 制御 冷却システム
燃料電池システム 車両用空調装置 燃料電池 温度
放熱 加熱 電気ヒータ 温度センサ 流通 圧縮機 冷却
排気ガス 供給 熱交換 循環 ヒータ

Z17.情報通信

処理 記憶 特定 表示装置 データ 判定 検出 表示
制御 電気自動車 生成 取得 制御装置
信号 制御部 電気信号 情報 受信
入力 表示部 制御信号 記憶部 通信 送信
演算 ユーザ 車載機器 利用者 位置 変換

Z18.演算・推定

走行 情報 モータ 電圧 方法 比較
予測 閾値 制御装置 消費電力 検出 時間 制御 比較
充電状態 推定 判定 電気自動車 演算
補正 取得 値 プログラム 記憶 差 温度 バッテリ
測定 ECU

Z19.機器の異常検出

信号 停止
閾値 電圧 演算 制御部 検出 比較 停止
構成 電流センサ 印加 温度 電流センサ
変化 制御装置 検出 異常 電流 判定
検出部 有無 検出信号 検出結果
検出手段 検出+できる 故障 制御
ECU 回転角度

Z20.操作スイッチ

操作部 安定 破損 形成 調整装置
スイッチ 接点 電気掃除機 ストップボタン 感度 力加へり
車室内 開閉 方向 可動部 検出
自動車 操作部 装着 操作+できる ケース 構成
操作性 検出 固定接点

特許トピック34個の一覧②

26,419件の特許は、エンジン、動力伝達、モータ、ブレーキ、電力変換、二次電池、充電、情報通信、異常検出、筐体、構成、小型化、安全性などの34個のトピックに集約されました

Z21. 筐体

一端カバー 検出 取容部 保持 モータ 外部開口部
 筐体ケース内 形成 電子制御ユニット
 電気部品 開口 ハウジング ケース 収容
 露出 基板 制御回路配置 電気接続部 一体 挿入
 装着 コネクタ貫通孔 構成 対向 固定

Z22. 表面の形成

一対外周面 被覆対向凹部 光 面
 周囲 外側一体 形状先端発光素子
 導電性 突出 形成 接触位置 表面
 構成 基板 本体 端部部分
 配置 反対側 貫通孔 挿入電極 絶縁

Z23. 位置とその移動

直交 方向
 許容形成 ロック 支持 係合 回転 本体
 解除 保持 係 位置 力 連結 駆動 移動
 アクチュエータ ドア軸方向 回転 ハンドル
 反対側 規制 固定 接触 移動+できる
 検出 電気信号

Z24. 配置・位置・方向

供給 方向
 近傍領域 軸方向 モータ
 取容 ハウジング 形成 配列 離間位置
 対向隣接 配置 長手方向 方向
 反対側 垂直 近接 平行
 構成 移動+できる 周囲端部 下方 対向

Z25. 構成の方位

方向 車両前後 下面 配設
 電気自動車空間位置側面構成連結突出 配置
 車輪 車輪 後方 下方 開口 上方 対向 支持
 先端 形成前方 収容上面 バッテリー 電気掃除機
 固定 開口部ケース

Z26. 構成

モータ制御+できる センサ 外部長さ
 自動車 生成 設置 構成
 電気自動車 充電 制御装置 プロセッサ 電圧
 電力 エネルギー貯蔵装置 結合
 電気エネルギーシステム
 電源 供給 制御部 配置 コイル

Z27. 接続

電源ケーブル 車体 固定 検出 保持
 位置 嵌合 電源プラグ 配線構成 基板 供給 接続部
 端子他端 コネクタ 一端 接続+できる
 配置 端部ケーブルバス 外部 形成 ワイヤハーネス
 接地 電線 収容 回路 対向

Z28. 方法の提供

配置監視 段階
 電気自動車 供給測定 実施 自動車 工程生成
 製造 エネルギー
 電気機械調整 方法 バッテリー システム
 駆動 内燃機関 作用 ハイブリッド車再駆動
 存在 電気エネルギー 制御 モータ 分離 車両用
 センサ

Z29. 損傷や浸水など不具合の防止

耐久性 発明
 衝撃 電気機器 不具合 静電気 構造 振動
 電気接続部 製造コスト外部 確保+できる 影響
 損傷 侵入 電動パワーステアリング装置 未然
 外力 電気自動車ノイズ 水 温度変化 信頼性
 起因 衝突 破損 変動 異音 安全 浸入

Z30. 小型化・簡素化・低コスト化など付加価値

リレー 大型化 自動車 リードフレーム 安価 小型 端子材 実現
 確保 低コスト 必要+ない コンパクト 製造コスト 車両用 灯具
 部品 点数 コスト 小型化 信頼性 構造
 簡素化 電気接続箱 耐久性 軽量化 製造方法
 コネクタ ワイヤハーネス 作業性 削減 強度

Z31. 効率性・安全性の向上

燃費 温度上昇 電源システム 安定 実現
 構成 精度 確保 劣化 ハイブリッド車両 バッテリー
 安全 正確 電気自動車 モータ 効率
 短縮 車両制御装置 消費電力 問題 充電+できる 検出+できる
 制御手段 エネルギー効率 走行中 運転者 必要+ない 技術

Z32. 既存エンジンへの警鐘・樹脂組成物の提供

含有成形品 耐熱性成分 組成物
 重合体副トランキング性 成形品外観 金型離型性 電気特性
 電気部品 自動車部品 既存蒸気タービン発電機 重量部
 発電量 理論最良エンジン 電気部品用途
 ポリアリレンスフィド 耐衝撃性 後追いエンジン 発明阻止
 高校大学 機械的強度 既存エンジン
 化合物 溶解流動性

Z33. 重力発電の活用による地球温暖化防止

船舶 電気駆動 既存火力原子力発電全廃 圧縮空気加速 船舶
 落差燃料費ゼロ 垂直下方 全面電化住宅全盛
 工場電化全盛 駆動二酸化炭素排気ゼロ 全地球
 人類総量 大気圧同速度同容積仕事率 既存世界
 海水温度上昇ゼロ 先送り 既存蒸気タービン発電
 安価 重力加速度加速 重力発電運用 水発電量増大
 タービン 重力発電蓄電池駆動 地球温暖化 自動車 発電量

Z34. タービン発電の出力向上・燃費低減

反転 最大速度部水 発電原価 静翼 永遠運用改善
 横軸自動車 軽量蒸気速度 マッハ狙い 容積圧縮仕事率
 安価電気駆動 燃料費ゼロ 太陽光加熱器熱製造
 容積 電気+液体空気+過熱蒸気温熱供給設備D
 宇宙到達費用 空気圧縮液体酸素圧縮駆動 発電量
 軽量物発電日帰り旅行 飛行機 製造物全部
 燃費 既存蒸気タービン発電 自動車 既存 全動翼 船舶
 蒸気速度 出力発電

Z32,Z33,Z34は特定の出願人による重複した要約内容の特許から抽出された

※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 (関係の強い上位5つの単語を赤色で表示している)

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	要約	出願年	出願人	トピック Z01	トピック Z02	...	トピック Z34
1	特願2007-XXXX	【課題】電気式変速操作装...	2007	A社	1	1		0
2	特願2009-XXXX	【課題】従来の電気自動車...	2009	B社	0	1		1
3	特願2012-XXXX	エンジンのための方法及び...	2012	C社	0	1		1
4	特願2013-XXXX	【課題】駐車場に設置された...	2013	D社	1	0		0
...
26,419	特願2016-XXXX	充電ステーションが電気エネ...	2016	X社	1	0		1

トピック×属性の様々な分析が可能に

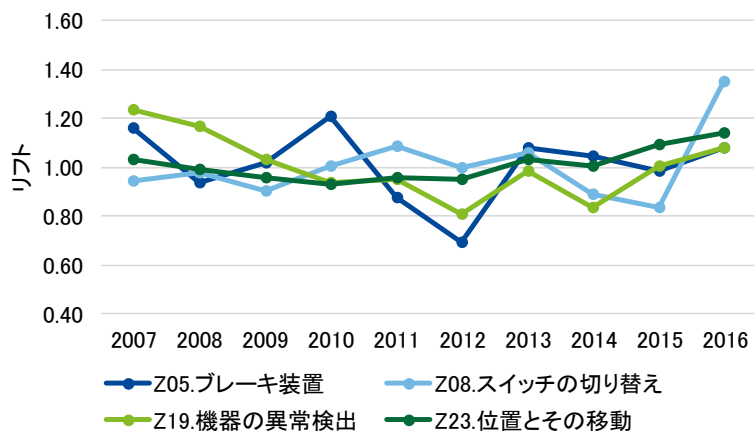
4. Nomolyticsを適用した特許分析事例②

4-3. 出願年×トピックによるトレンド分析

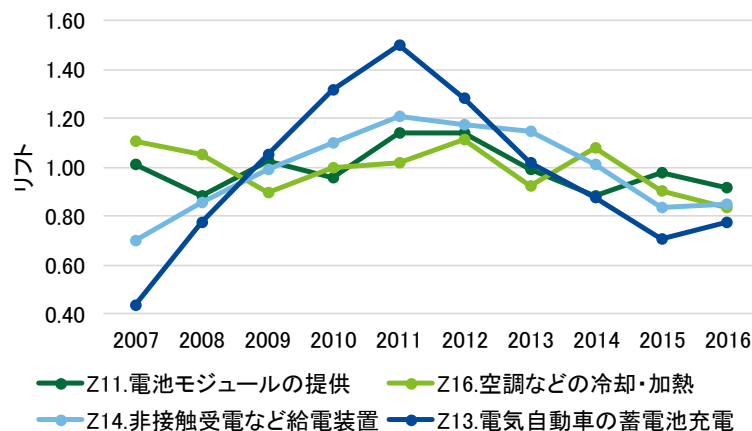
トピックの上昇トレンドと下降トレンド

近年は「Z08.スイッチの切り替え」、「Z01.エンジンの始動と停止」、「Z19.機器の異常検出」などが上昇しており、「Z16.空調などの冷却・加熱」は下降しています

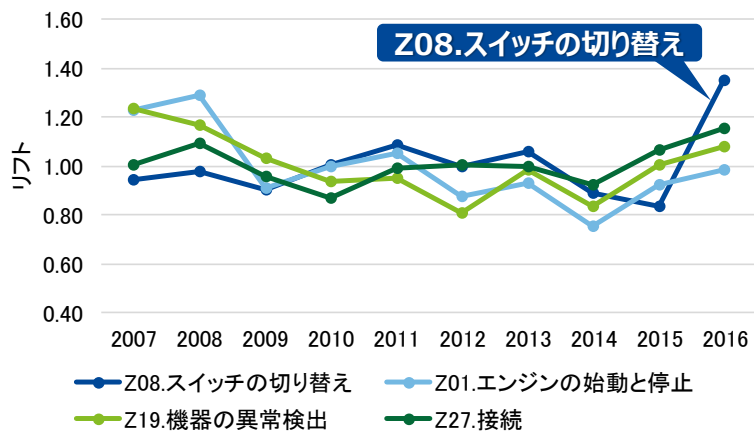
【中期トレンド】2012年からの上昇率 best4



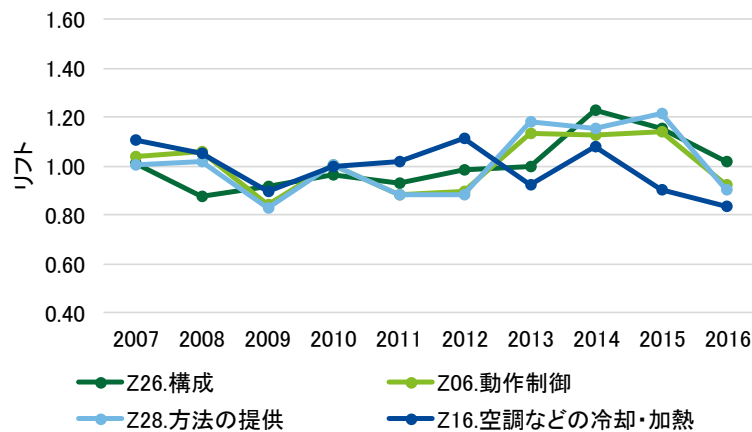
【中期トレンド】2012年からの下降率 worst4



【短期トレンド】2014年からの上昇率 best4



【短期トレンド】2014年からの下降率 worst4



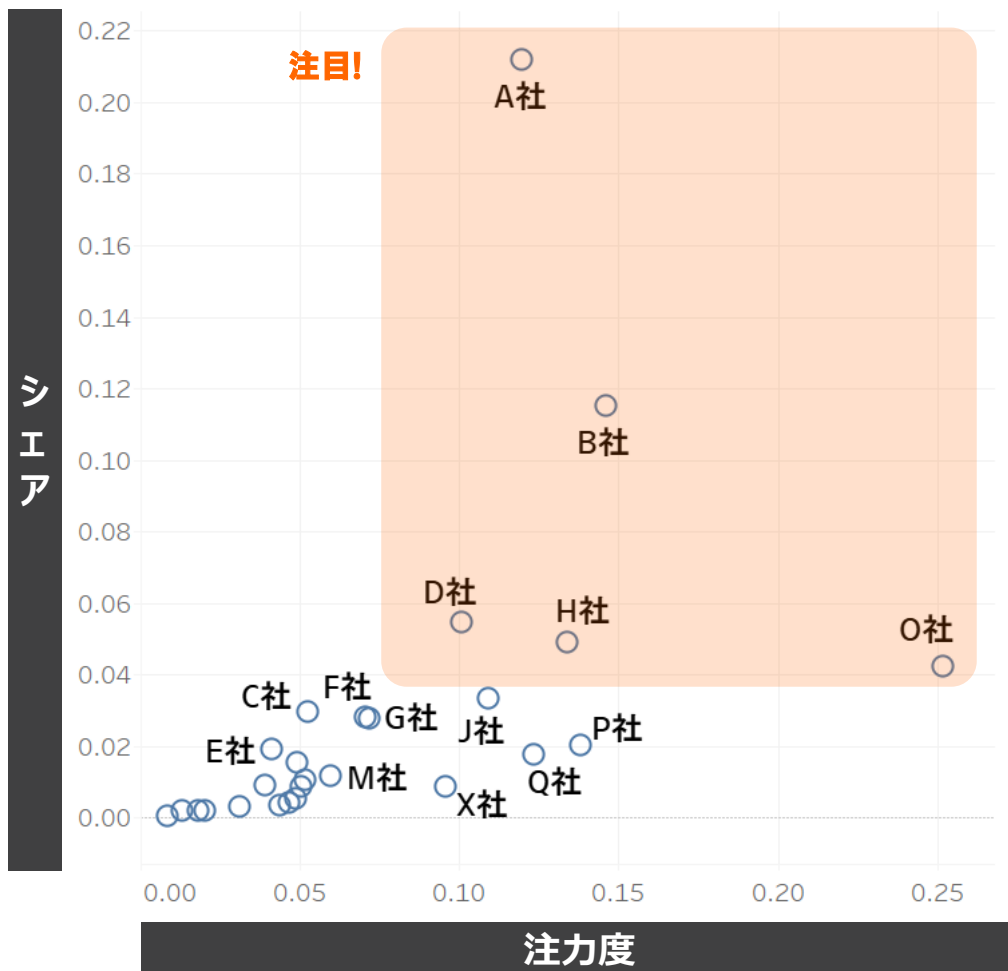
4. Nomolyticsを適用した特許分析事例②

4-4. 出願人×トピックによる競合分析

「Z08.スイッチの切り替え」の出願人のポジショニング

スイッチ切替に関する技術は、シェアではA社が圧倒的に高いですが、ある程度のシェア・注力度を有する企業間で連携すれば、A社という巨人に対抗できることも考えられます

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社が最も高く、他社を突き放しており、この領域で多数の特許を出願している
- 注力度ではO社が最も高く、他社を突き放しており、この領域において全社的に関心が高く、特有の技術力を保有している可能性がある
- 中程度のシェア、注力度にはB社、D社、H社などがあるが、こうした企業で連携することで、より技術力を高めながらシェアを伸ばし、A社という巨人に対抗できることも考えられる

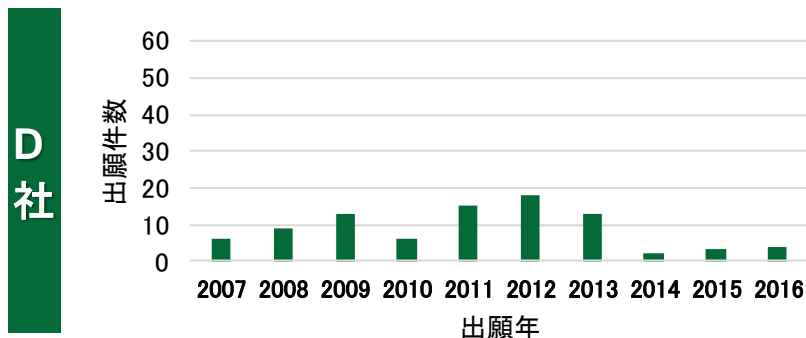
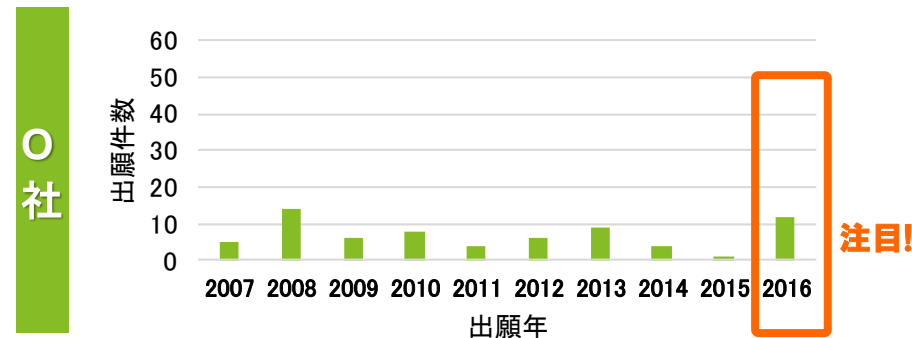
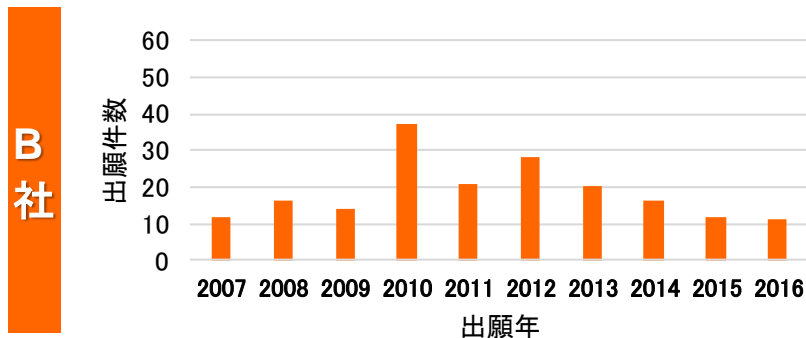
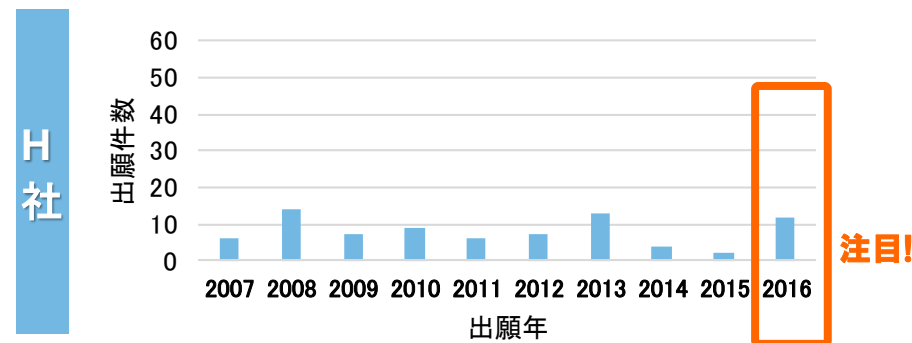
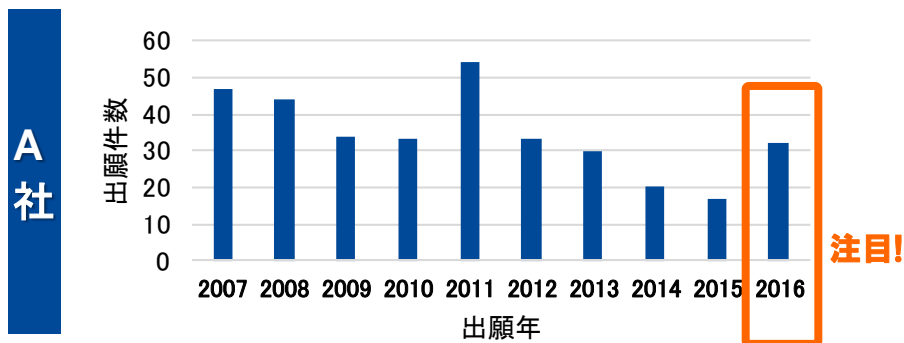
注力度とシェア

- **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している
- **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、つまりトピックのなかでどれくらいその出願人が占めているのかを示している

「Z08.スイッチの切り替え」の各出願人の出願トレンド

シェア1位のA社、注力度1位のO社、またH社も直近で出願件数を急に増やしており、技術戦略の転換の可能性も考えられ、直近での出願内容および今後の出願動向に要注目です

注目企業の出願件数の推移



「Z08.スイッチの切り替え」×「A社」×「2016年」の特許要約文

A社の直近2016年の出願特許には、車の衝突時にそれを検知してコンデンサの放電に切り替える技術や、異常が生じたときにそれを検知して回避する切り替え技術があります

Z08×A社×2016年の該当特許32件から抜粋した8件の要約の課題

発明の名称	電気自動車
要約【課題】	車両の衝突時に平滑化コンデンサを放電する確実性を向上させる。

発明の名称	電気自動車用の電源システム
要約【課題】	車両が衝突したときにより確実に平滑化コンデンサを放電する。

発明の名称	電気自動車
要約【課題】	電気自動車の衝突時に、できるだけパーキングロックを使わずにモータを停止させて平滑化コンデンサを放電させる。

発明の名称	ハイブリッド車両
要約【課題】	車両の衝突時に、モータの回転数を取得できない場合にも、インバータに接続されるコンデンサの電荷の放電を速やかに完了する。

発明の名称	電気自動車
要約【課題】	補機バッテリーが短絡しても、重要な補機への電力供給を継続できる電気自動車を提供する。

発明の名称	電気自動車用の電源システム
要約【課題】	通信不良が生じた場合であってもシステムスイッチを安全に開放することのできる電気自動車用の電源システムを提供する。

発明の名称	電気自動車
要約【課題】	第1インバータ回路と第2インバータ回路のうち一方で異常が生じたときに、両インバータ回路の複数のスイッチング素子を同時にオフするモータ制御ユニットを提供する。

発明の名称	ハイブリッド車
要約【課題】	ハイブリッド車の第1コントローラに異常が生じたときでも、エンジンを始動し得る技術を提供する。

「Z08.スイッチの切り替え」×「H社・O社」×「2016年」の特許要約文

H社とO社の直近2016年の出願は、全てH社とO社の共同出願で、給電が途絶える場合に給電を維持する切り替え技術や、異常発生時の影響を回避する切り替え技術があります

Z08×H社・O社×2016年の該当特許12件から抜粋した6件の要約の課題

発明の名称	給電中継回路、副電池モジュール、電源システム
要約【課題】	一つの車載電源が失陥した場合であっても、車両の電氣的負荷に給電する。

発明の名称	車載用のバックアップ装置
要約【課題】	電源部からの電力供給が途絶えた場合であっても電力供給対象への電力供給を途切れさせることなく供給源を蓄電部に切り替えることが可能な装置を、より簡易な構成で実現する。

発明の名称	車両用電源装置
要約【課題】	第1電源部からの電力に基づく急速充電動作と、充電中に第1電源部の電力供給が遮断されても遮断前後で放電状態を維持し得る充放電動作とを行い得る車両用電源装置を提供する。

発明の名称	車両用電源装置
要約【課題】	第1電源部の電力不足によって特定負荷が駆動できなくなる事態を、第2電源部に及ぼす影響を抑えて回避し得る車両用電源装置を提供することを目的とする。

発明の名称	リレー装置及び車載システム
要約【課題】	少なくとも2つの蓄電部に対して発電機から充電電流を供給することができ、且つ一方の蓄電部側に異常が発生した場合に、その異常が他方の蓄電部側に及ぶことを抑え得るリレー装置を提供する。

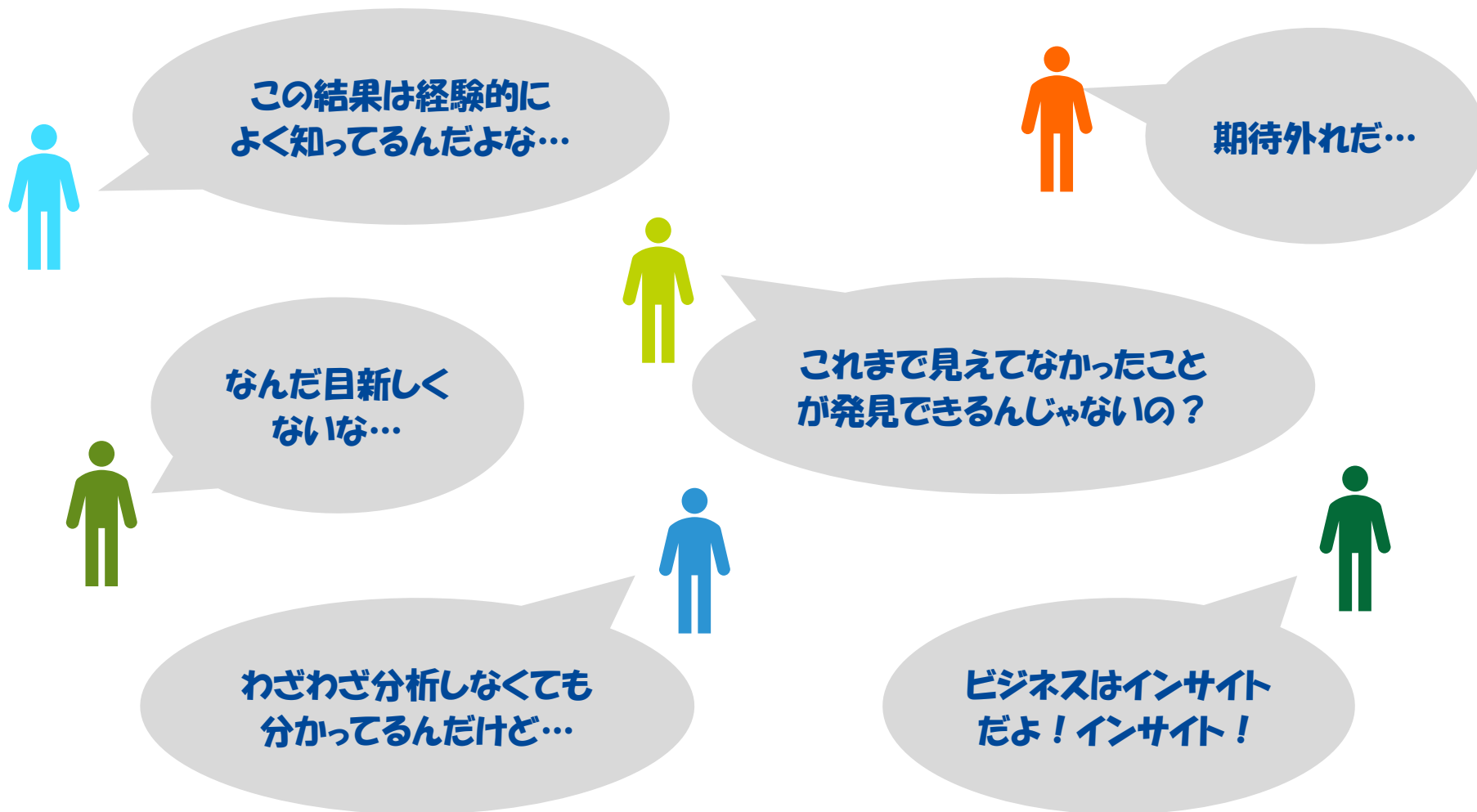
発明の名称	車両用電源装置
要約【課題】	アイドリングストップ状態からの復帰の際に発電機側の蓄電部の電圧が低下しても負荷にその影響が及びにくい車両用電源装置を、電流集中を抑制し得る構成で実現する。

5. インサイト獲得のためのPLSAの新展開技術 PCSAと*differential* PLSA

5. インサイト獲得のためのPLSAの新展開技術 PCSAと*differential* PLSA

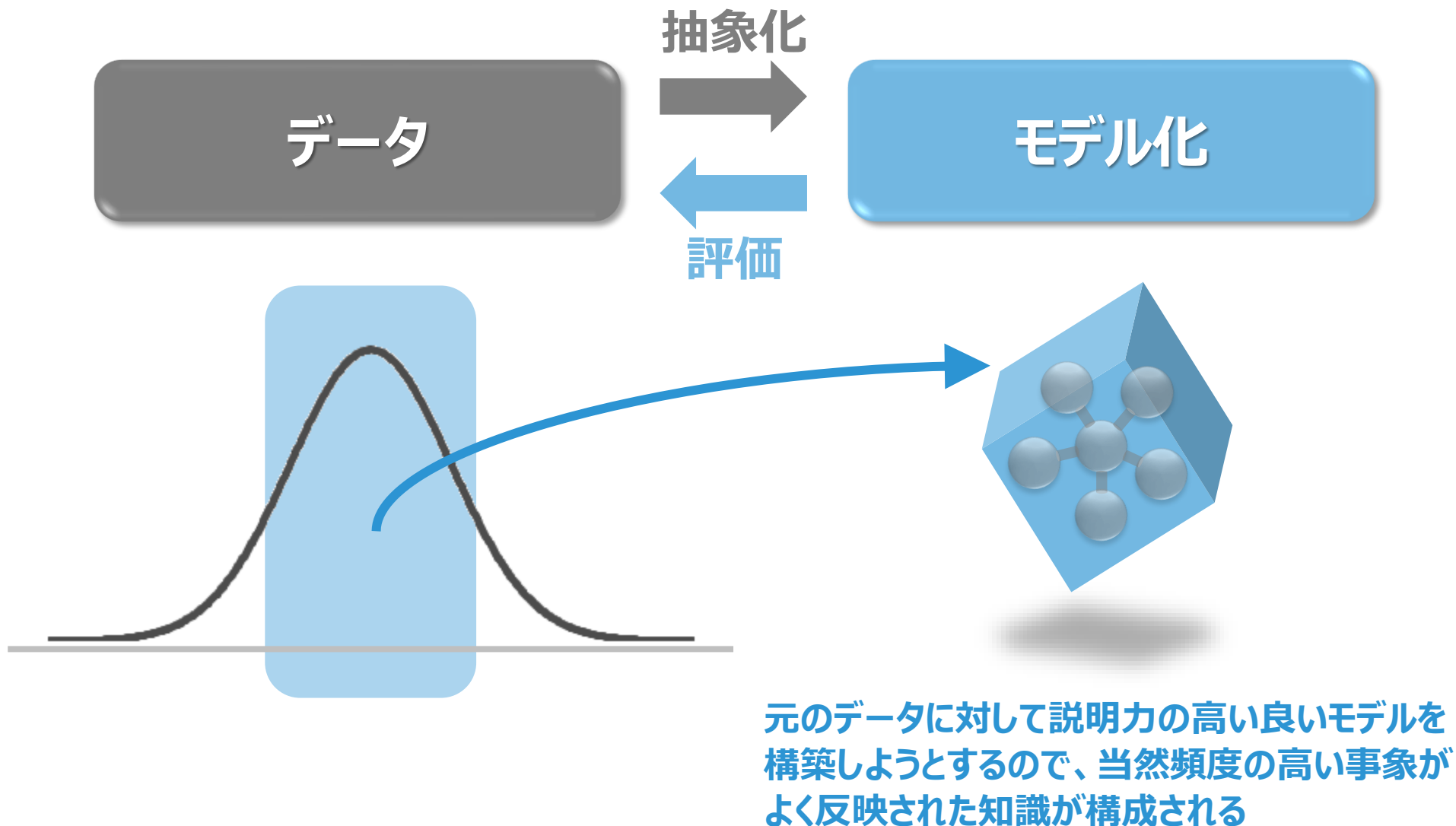
5-1. インサイト獲得で求められるトピック抽出のあり方

特に新たな気づきとなるインサイトの発見が求められるビジネス現場では、データ分析の結果に目新しさがなく、期待外れと評価されることがあります



データモデリングの原理

データをモデル化すると、データに潜む傾向やルールを抽象化することで、説明力の高いモデルでは頻度の高い事象≒経験的によく知っていることが優先的に表現されます



5. インサイト獲得のためのPLSAの新展開技術 PCSAと*differential* PLSA

5-2. PCSA: あるターゲットに特化したトピックの抽出手法

【従来手法】PLSAを用いたトピックの抽出

PLSAでは、全体のテキストデータからテキストマイニングで抽出した単語で構成される共起行列をインプットにすることで、データ全体を表現する代表的なトピックを抽出します



テキストデータにテキストマイニングを実行し、全体のデータに含まれる単語を抽出する

抽出した各単語の全体における共起頻度を集計した共起行列を作成する

作成した共起行列にPLSAを適用し、使われ方の似ている単語をトピックに集約する

あるターゲット事象の「該当データ」と「非該当データ」から構築した2つの共起行列の差分にPLSAを適用することで、そのターゲットに影響する要因トピックを優先して抽出します

PCSA[®] (Probabilistic Causal Semantic Analysis: 確率的因果意味解析)



単語抽出

共起頻度集計

トピック抽出



出現頻度

- Word 1
- Word 2
- Word 3
- Word 4
- Word 5
- Word 6
- Word 7
- Word 8
- Word 9
- Word x

あるターゲット事象 (属性情報) Xが、

全データ

共起行列U

該当するデータ

共起行列A

該当しないデータ

共起行列B

2つの共起行列の差分

共起行列

[A-B]

PLSAの適用

トピック1

Word5
Word3 Word9
Word1
Word2 Word8
Word4 Word6
Word7

トピック2

Word2 Word8
Word3
Word1 Word9
Word7 **Word6**
Word5 Word4

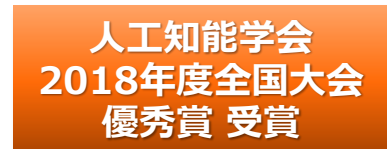
トピック3

Word6
Word2
Word3 **Word5**
Word8 Word1 Word9
Word7
Word4

ターゲット事象Xの
該当有無に影響を
与える要因トピック
を優先的にテキスト
情報から抽出できる

ターゲット
事象 X

全データから構築した共起行列Uを、あるターゲット事象(属性情報)Xが該当するデータから構築した共起行列Aと、該当しないデータから構築した共起行列Bに分割し、その2つの共起行列の差分を取った共起行列(A-B)に対してPLSAを適用する



Nomolyticsは全体を表すトピックを抽出してから属性との関係进行分析しましたが、PCSAはその属性の特徴をよく表すトピックを最初から抽出し、より顕著な特徴の探索を行います

Nomolyticsでは

データ全体を表現するトピックを抽出

してから

属性との関係进行分析

すると

属性の特徴を把握

できた

PCSAでは

属性の特徴をよく表現する

ような

偏ったトピックを抽出

してから

より顕著な属性との関係进行分析

する

5. インサイト獲得のためのPLSAの新展開技術 PCSAと*differential* PLSA

5-3. PCSAを適用した特許分析事例

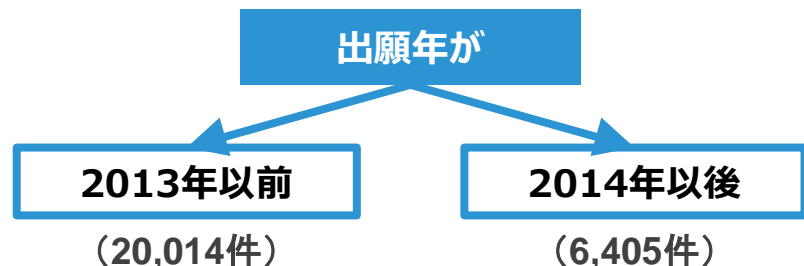
電気自動車関連の特許データ26,419件の要約文全体を対象に、「出願年」をターゲットとしてPCSAを適用することで、近年上昇傾向あるいは下降傾向にあるトピックを抽出します

データの抽出条件と分析対象（先ほどと同様）

- 対象
 - 公開特許公報
- キーワード
 - 要約と請求項に「車」と「電気」を含む
- 出願日
 - 2007年1月1日～2016年12月31日
- 抽出方法
 - Patent Integrationを使用
- 抽出件数
 - 26,419件
- 分析対象
 - 要約の全文



PCSAのターゲット



分析プロセス

テキストマイニング

要約文から単語と係り受け表現を抽出

差分共起行列の作成

出願年が2013年以前の共起行列と2014年以後の共起行列の差分を取った共起行列を作成

トピック化 (PLSA)

差分共起行列にPLSAを適用して2014年前後の出願に特徴があるトピックを抽出

トピックのスコアリング

全データに対する各トピックのスコアを計算

特徴の可視化

トピックのスコアを出願年を軸に集計・可視化

【開発手法】PCSAによるトピック抽出のアプローチ

①2013年以前データ、②2014年以後データで作成した2つの共起行列の差分にPLSAを適用することで、2014年前後の出願に特徴があるトピックを優先的に抽出します

2つの共起行列の作成

①出願年が**2013年以前**のデータ
20,014件(文章数:33,113件)

係り受け表現

	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	...
構成	97	30	27	25	
制御	208	49	80	66	
配置	56	1	23	8	
モータ	192	42	356	46	
...					

単語

②出願年が**2014年以後**のデータ
6,405件(文章数:10,723件)

係り受け表現

	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	...
構成	21	3	9	8	
制御	60	24	28	19	
配置	13	1	6	0	
モータ	47	19	138	12	
...					

単語

差分共起行列の作成

①2013年以前の共起行列と
②2014年以後の共起行列の
差の絶対値を計算した共起行列
を作成する

係り受け表現

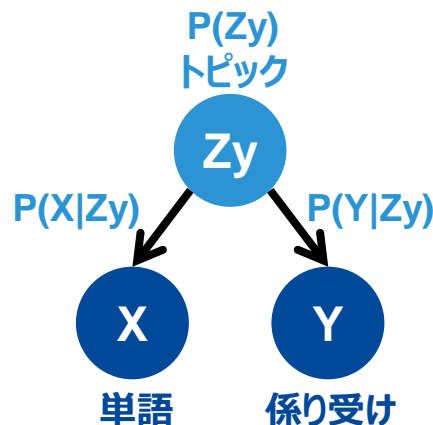
	電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電	...
構成	10.4	6.7	0.3	0.1	
制御	7.4	8.1	2.1	2.4	
配置	5.1	0.7	1.4	2.6	
モータ	15.2	5.4	22.7	2.9	
...					

単語

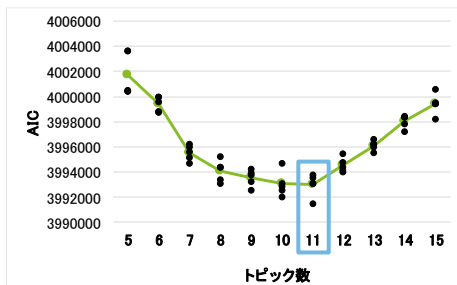
ただし、2つの共起行列は異なる文章数のデータから作成されているので、①2013年以前の共起行列の頻度を2つの文章数の比率(10,723/33,113)で重み調整してから、②2014年以後の共起行列との差を計算する

PLSAの実行

差分共起行列に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



トピックの抽出

各トピックについて以下の3つの確率が計算される

- ① $P(X|Zy)$
トピックにおける単語の所属確率
- ② $P(Y|Zy)$
トピックにおける係り受けの所属確率
- ③ $P(Zy)$
トピックの存在確率

トピックにおける $P(X|Zy)$ と $P(Y|Zy)$ からトピックの意味を解釈する

トピック Zy05

$P(Zy) = 2.5\%$

$P(X Zy)$	単語	$P(Y Zy)$	係り受け
3.1%	製造	4.5%	水素-製造
2.2%	二次電池	3.0%	発電-電気
1.5%	負極	2.6%	製造方法-提供
1.5%	正極	2.6%	製造-方法
1.5%	製造方法	1.6%	含む-リチウムイオン電池
1.2%	エネルギー	1.3%	成形品-提供
1.1%	リチウムイオン電池	1.3%	強度-有する
1.1%	水素	1.3%	表面-形成
1.1%	セパレータ	1.2%	含む-組成物
1.0%	電解液	1.2%	リレースイッチ
...

確率の高い構成要素から、トピックZy05は「二次電池の製造方法」に関するトピックと解釈できる

PCSAによる出願年の要因トピック11個の一覧

出願年(2014年前後)で特徴を示すトピックは、エンジン駆動、電力変換、充電、二次電池の製造方法、冷却・加熱、検出判定、小型化・低コスト化など11個抽出されました

Zy01.エンジン駆動・動力伝達の制御

動力 演算 発電機
駆動輪伝達クラッチ 出力軸 開始構成 **モータ**
トランスミッション制御手段駆動 運転者
エンジン 制御装置 ハイブリッド車両
判定**制御**ブレーキ モータジェネレータ トルク
連結 駆動力バッテリー停止 車輪 内燃機関
作動 検出

Zy02.モータの回転構成

位置 伝達 方向形成
中心固定ハウジング自動車 **回転**
収容移動できる 電気機械 駆動部
モータ支持回転+できる **配置 回転軸**
軸方向 ステータ連結駆動周囲 移動口+
構成車輪 発電機 シャフト 結合

Zy03.交流・直流の変換

放電直列直流スイッチ電力 並列 蓄電池
負荷**供給 変換**バッテリー検出 直流電力
交流電力充電電気自動車コンバータ **インバータ**
制御制御部駆動 電力変換装置 蓄電装置
コンデンサ 制御装置 **電圧**電流オフ電源モータ

Zy04.電気自動車への充電、給電装置

検出判定走行 構成 外部
充電ケーブル 制御 ユーザ 演算 取得電源
収容蓄電池 **電力**給電装置 コネクタ 充電+できる
充電システム **電気自動車 蓄電装置**
充電スタンド外部電源 **充電バッテリー**
給電開始供給情報 装置 送信

Zy05.二次電池の製造方法

由来 負極活物質 電解液 成形品
セパレータエネルギー**正極** 電子機器正極活物質
積層エネルギーシステム
製造 負極 二次電池 リチウムイオン電池
再生可能エネルギー電気部品 **製造方法**スイッチ
方法 自動車部品電池特性 水素 形成 安全
電極含有表面発電 活物質

Zy06.空調などの冷却・加熱

制御 燃料 配置 冷媒 排出 空気
変換電気ヒータ **駆動**加熱 構成 冷却水 **発電**
エンジン 電気エネルギー**供給 発電機**
バッテリー 燃料電池 制御装置 温度 **電力**
排気ガス冷却 車室内 作動 熱 循環

Zy07.情報通信、検出判定システム

電流送信自動車 調整 比較 情報
信号受信 **構成**センサ 取得 制御装置
システム**制御 検出**速度 方法 制御部
演算 **判定**生成 測定電圧 電気信号
動作 記憶 入力 変化 モータ 位置

Zy08.形成・配置

開口 方向 下方
他端外部バッテリー 位置端子一端上方**収容**
開口部ケース **形成**面ハウジング **配置**
コネクタ **固定**電気部品 基板 端部 接触 突出 **構成**
保持 支持 対向 筐体 挿入

Zy09.小型化・低コスト化・簡素化・操作性向上

精度 低コスト 効率 バッテリー 安定 **操作**起因 安価
コスト自動車 小型化車室内 損傷 ワイヤハーネス
操作+できる **電気自動車**確保ハイブリッド車両
振動 **構成**信頼性 **構造**電子機器 **スイッチ**
部品点数必要+ない モータ 影響
実現安全

Zy10.重力発電の活用による地球温暖化防止

駆動 タービン 既存火力発電 発電量 安価
重力発電運用 燃料費ゼロ 圧縮空気加速 発電量増大
垂直下方海水温度上昇ゼロ 人類滅亡 発電量増大
海草 重力加速度加速 二酸化炭素排気ゼロ水
落差 大気圧同速度同容積仕事率先送り **船舶**
工場電化全盛 **自動車 既存蒸気タービン発電**
電気駆動 既存世界 全面電化住宅全盛
重力発電蓄電池駆動 既存火力原子力発電全廃 全廃地球
地球温暖化 災類
Top5: 1
順位: 全国
No.2: 0.039

Zy11.既存エンジンへの警鐘、タービン発電・重力発電

自動車 軽量物発電 運用
安価電気駆動 軽量蒸気速度 横軸自動車 反転 永遠
全動翼 太陽光加熱器熱製造 飛行機 液体酸素圧縮駆動
既存エンジン宇宙到達費用 **理論最良エンジン**
船舶 電気+液体空気+過熱蒸気蒸気供給設備 **大学**
発電量既存 既存蒸気タービン発電
容積圧縮仕事率 発電原価 後追いエンジン 発明阻止
製造物全部 日帰り旅行 燃料費ゼロ空気圧縮
高校 静翼 燃費

Zy10,Zy11は特定の出願人による重複した要約内容の特許から抽出された

※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 (関係の強い上位5つの単語を赤色で表示している)

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)を計算し、各トピックが該当するデータのうち、“2014年以後”となる割合について、PLSAの結果とPCSAの結果を比較します

PLSAによるトピックとPCSAによるトピックのスコア(フラグ情報)を紐づけた特許データ

PLSAによる34個の全体トピック

PCSAによる11個の要因トピック

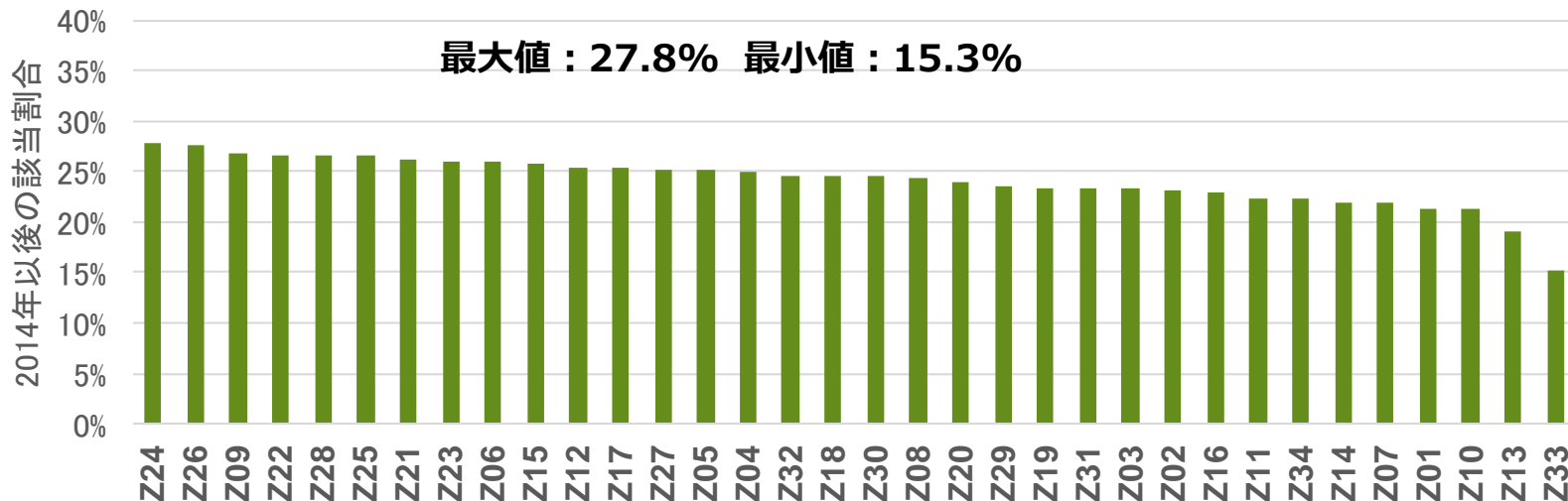
特許ID	出願番号	出願年	出願年グループ	トピック Z01	トピック Z02	...	トピック Z34	トピック Zy01	トピック Zy02	...	トピック Zy11
1	特願2007-XXXX	2007	①2013年以前	1	1		0	0	1		1
2	特願2009-XXXX	2009	①2013年以前	0	1		1	0	0		1
3	特願2012-XXXX	2012	①2013年以前	0	1		1	0	1		0
4	特願2014-XXXX	2014	②2014年以後	1	0		0	1	0		0
...
26,419	特願2016-XXXX	2016	②2014年以後	1	0		1	1	0		0

各トピックが該当する特許のうち、出願年が「②2014年以後」の割合を比較することで、2014年前後においてPCSAではどれくらい偏ったトピックを抽出できているか確認する

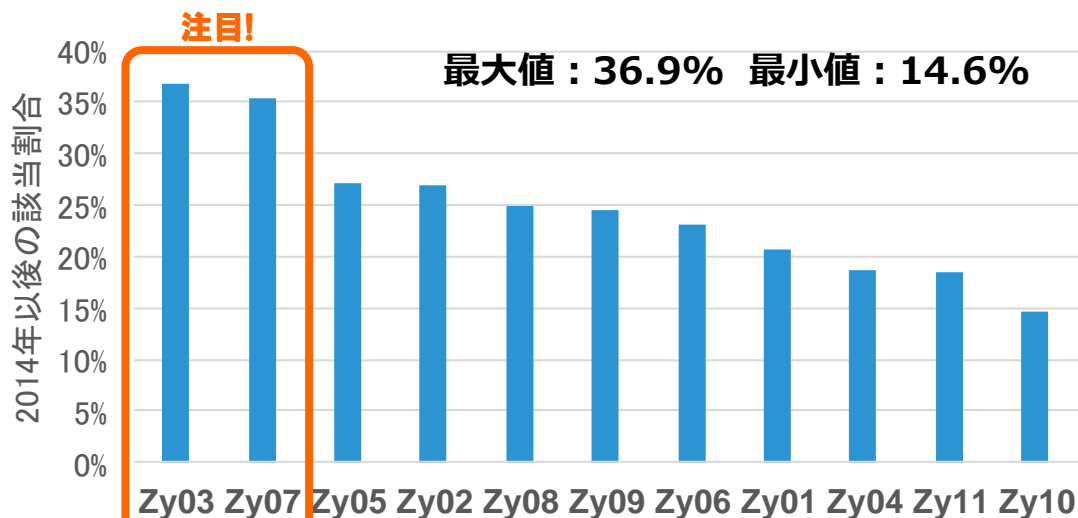
PLSAとPCSAによるトピックの分布の比較

各トピックの2014年以後の割合は、PLSAでおおむね25%前後ですが、PCSAでは割合が高いものと低いものに偏っており、2014年前後に対して特徴的なトピックとなっています

PLSA
による
全体
トピック



PCSA
による
要因
トピック

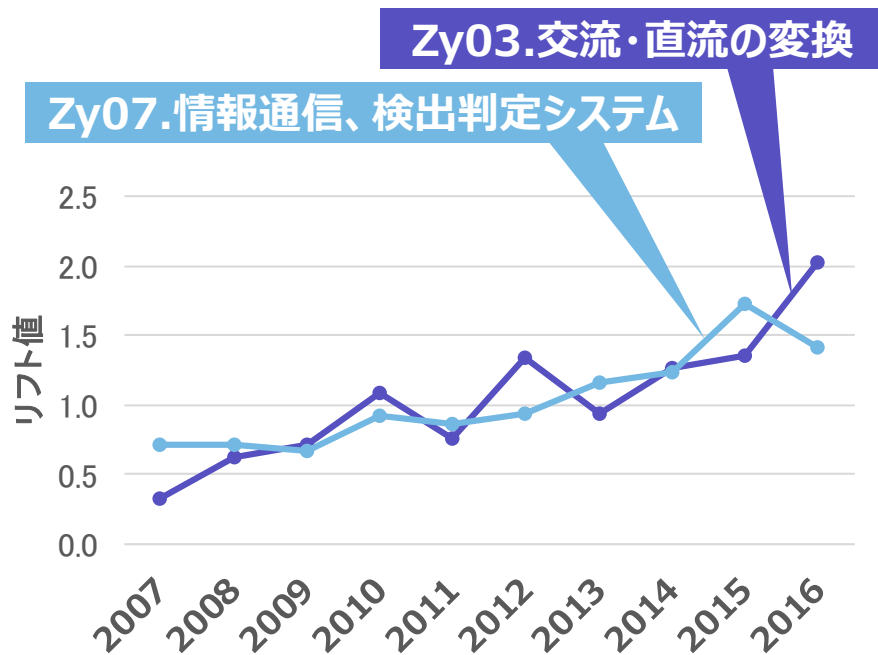


- それぞれ2014年以後の該当割合の高い順にトピックを並べている
- PLSAではおおむね25%前後のトピックが抽出されているが、PCSAのトピックでは割合が高いものから低いものまで抽出されており、最大値も高い
- (参考)元々の2014年以後データの割合は24.2%

近年上昇しているトピックの深堀分析

2014年以後の割合が高いトピックとして、電力変換に関する制御技術(Zy03)、データに基づいた運転者のアシスト技術(Zy07)が近年上昇傾向にあることが分かります

2014年以後の割合がトップ2のトレンド



$$\text{※リフト値} = \frac{P(\text{出願年} | \text{トピックTy}=1)}{P(\text{出願年})}$$

- リフト値は出願年とトピックの関係を示す指標
- トピック毎の各出願年の出願割合に対して、その出願年の出願割合で正規化した値であり、全体における各出願年の出願割合がそのトピックを条件にすることで何倍に変化するかを示す

該当特許の要約の例

Zy03.交流・直流の変換

発明の名称	出願年
電気自動車	2016
要約文 (抜粋)	
<p>車両の衝突時に平滑化コンデンサを放電する確実性を向上させる。衝突検知装置が、衝突を検知したときに第1信号と第1信号に続く第2信号を送信する。第1信号を受信したときに、インバータ制御回路がスイッチング素子駆動回路への電力の供給を停止する。第2信号を受信したときに、インバータ制御回路が平滑化コンデンサを放電する。</p>	

Zy07.情報通信、検出判定システム

発明の名称	出願年
車両速度の制御方法	2016
要約文 (抜粋)	
<p>経路及び交通状況に関する情報の応答として、自動車の速度を制御する方法に関するものである。本方法は、計画経路データ及び/または、繰返行程ロガーデータにより特定された想定経路に基づいて、最適な制動または加速点を決定し、最適な制動または加速点に基づいて、運転者に速度プロファイルを調整するためのサインを送る。</p>	

Nomolyticsはデータ全体を表すトピックを理解し、その特徴を様々な分析軸で探索し、PCSAは探索したい特徴に特化したトピックを抽出し、よりインサイトの獲得を狙います

Nomolytics®

テキストデータ全体を表すトピックを理解し、その代表的なトピックの特徴を様々な分析軸で探索できる

テキスト
マイニング

PLSA

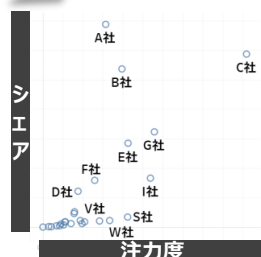
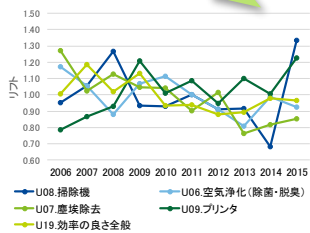
ベイジアン
ネットワーク

単語抽出

トピック化

モデリング

全体を表現するトピックとは？



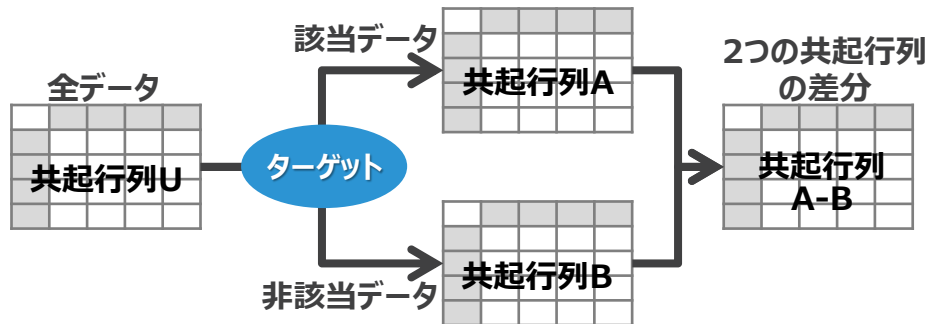
PCSA®

探索したい特徴に特化したトピックを優先的に抽出し、より顕著な要因を深く分析してインサイトを得る

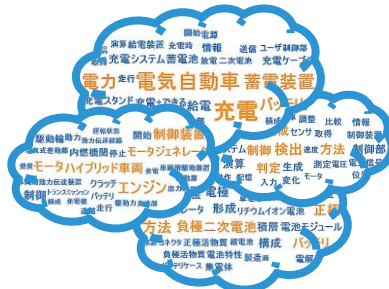
テキスト
マイニング

差分の
共起行列

PLSA



ターゲットの該当有無を左右するトピックとは？

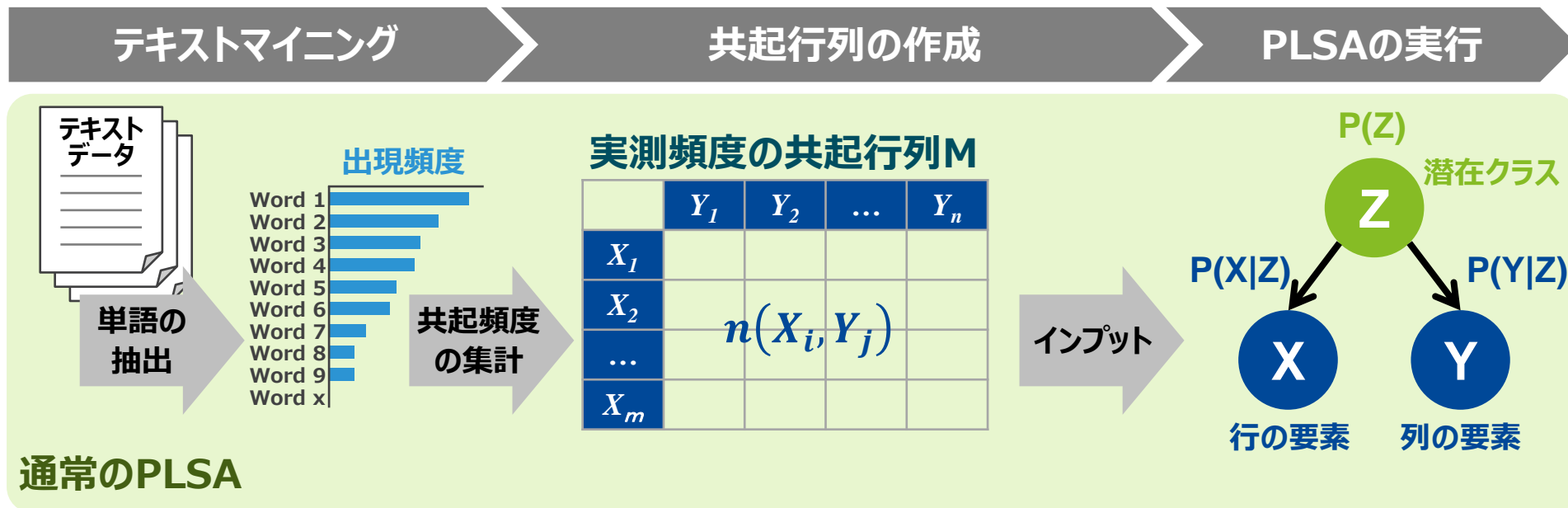


5. インサイト獲得のためのPLSAの新展開技術 PCSAと*differential* PLSA

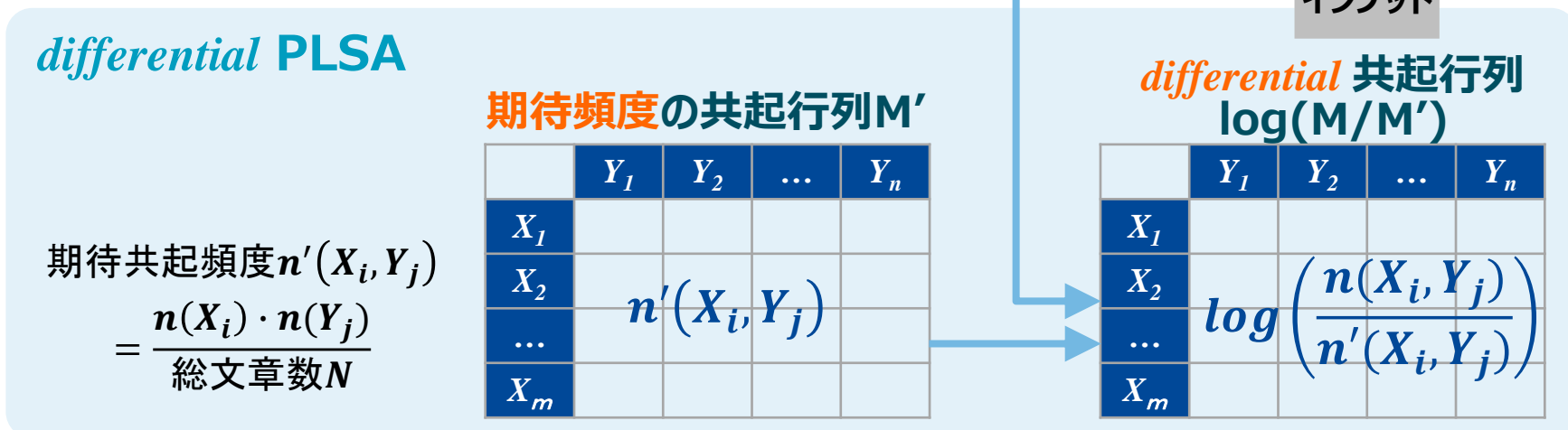
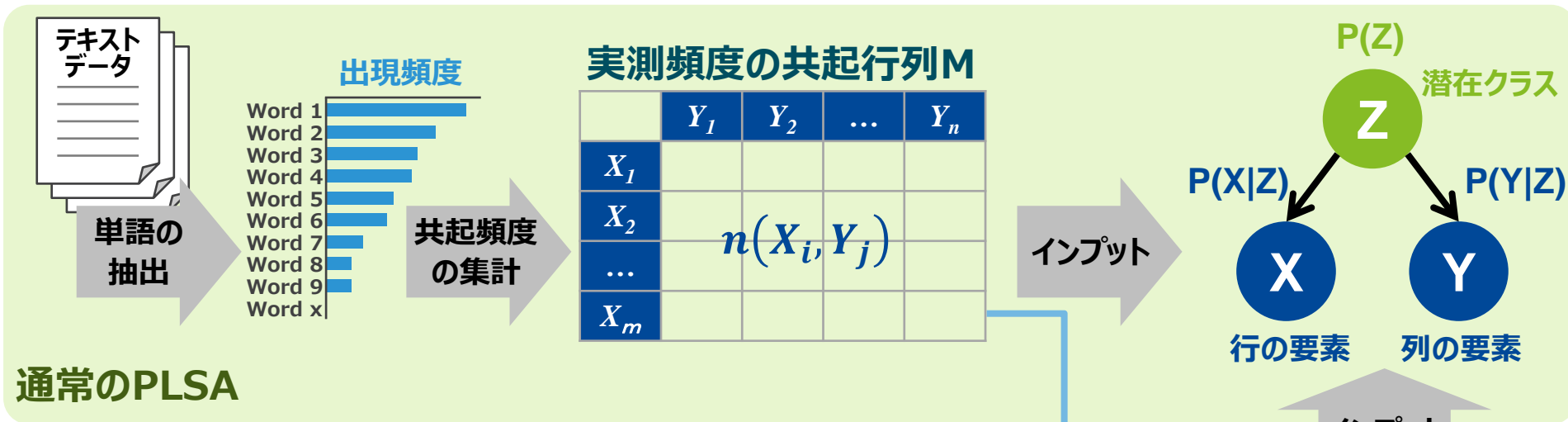
5-4. *differential* PLSA: 頻度によらない個性的なトピックの抽出手法

【従来手法】PLSAを用いたトピックの抽出

通常のPLSAでは、テキストデータに含まれる単語を抽出し、その単語群の共起頻度を集計した共起行列をPLSAのインプットとしてトピックを抽出します



differential PLSAでは実測頻度を集計した共起行列に加え、期待頻度を集計した共起行列を作成し、その比率の対数を取った共起行列をインプットとしてトピックを抽出します



通常のPLSAは、頻度の高い要素が優先的に反映され典型的なトピックになりがちですが、*diff*-PLSAは、頻度が低い要素も存在度が増し、個性的なトピックの抽出が期待できます

通常のPLSA

- PLSAを適用した確率解では、**頻度が高い要素に高い確率が割り当てられる**傾向にある
- 結果として抽出されるトピックは**典型的なもの**になってしまう

differential PLSA

- 実測共起頻度を期待共起頻度で除した値の対数を値とする共起行列をインプットとする
- 実測共起頻度が高い共起ペアでも、元々全体の頻度が高ければ期待共起頻度も高くなり、その比率を取ることで値の大きさが制限される
- 逆に実測共起頻度が低い共起ペアでも、期待共起頻度がそれよりも十分低ければ比率の値は大きくなり、PLSAの解ではこうした**頻度の低い要素にも高い確率が割り当てられる**可能性がある
- 結果として、**より個性的なトピック**が抽出されることが期待できる
- 単純な比率ではなくその対数を取ることは、期待頻度が低すぎることで極端に高くなる値を制限し、必要以上にデフォルトされた歪んだトピックが抽出されることを防ぐ効果がある

5. インサイト獲得のためのPLSAの新展開技術 PCSAと*differential* PLSA

5-5. *differential* PLSAを適用した特許分析事例

先ほどと同様の電気自動車関連の特許データ26,419件の要約全文を対象に、*differential PLSA*を適用することでより個性的なトピックを抽出します

データの抽出条件と分析対象（先ほどと同様）

- 対象
 - 国内の公開特許公報
- キーワード
 - 要約と請求項に「車」と「電気」を含む
- 出願日
 - 2007年1月1日～2016年12月31日
- 抽出方法
 - Patent Integrationを使用
- 抽出件数
 - 26,419件
- 分析対象
 - 要約の全文



分析プロセス

テキストマイニング

要約文から単語と係り受け表現を抽出

差分共起行列の作成

実測頻度の共起行列から期待頻度の共起行列を計算し、両者の比率の対数を取った共起行列を作成

トピック化（PLSA）

差分共起行列にPLSAを適用してより個性的なトピックを抽出

トピックのスコアリング

全データに対する各トピックのスコアを計算

特徴の可視化

トピックのスコアを出願年や出願人を軸に集計・可視化

【開発手法】 differential PLSAによるトピック抽出のアプローチ

①実測頻度の共起行列と②期待頻度の共起行列の比率の対数を値とする差分共起行列にPLSAを適用することで、より个性化的な特許トピックを抽出します

実測と期待の共起行列作成

①実測頻度の共起行列

		係り受け表現				
		総頻度 n(Y)	1,350	539	494	529
			電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電
単語	総頻度 n(X)	5,880	118	33	36	33
	構成	5,092	268	73	108	85
	制御	4,604	69	2	29	8
	配置	5,188	239	61	494	58
	モータ	...				

②期待頻度の共起行列

		係り受け表現				
		総頻度 n(Y)	1,350	539	494	529
			電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電
単語	総頻度 n(X)	5,880	34.6	13.8	12.7	13.5
	構成	5,092	29.9	12.0	11.0	11.7
	制御	4,604	27.1	10.8	9.9	10.6
	配置	5,188	30.5	12.2	11.2	12.0
	モータ	...				

※総文章数: 229,598件

差分共起行列の作成

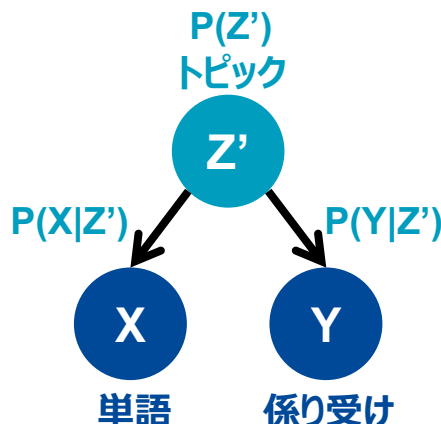
2つの共起行列の各共起ペアにおいて、期待共起頻度に対する実測共起頻度の比率の対数を取った差分共起行列を作成する

		係り受け表現			
		電力↓供給	否↓判定	モータ↓駆動	バッテリー↓充電
単語	構成	1.2	0.9	1.0	0.9
	制御	2.2	1.8	2.3	2.0
	配置	0.9	-1.7	1.1	-0.3
	モータ	2.1	1.6	3.8	1.6
	...				

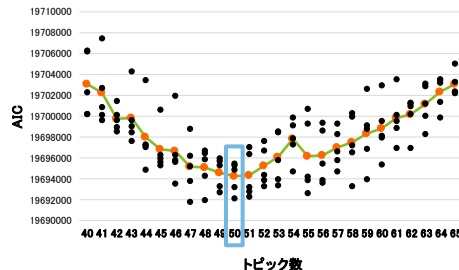
※値が負数となるものは"0"に置換する

PLSAの実行

差分共起行列に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



トピックの抽出

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 Z'20			
P(Z') = 2.3%			
P(X Z')	単語	P(Y Z')	係り受け
2.2%	プラグ	1.3%	充電ケーブル-接続
1.8%	コンセント	1.2%	プラグ-接続
1.7%	インレット	1.2%	外部電源-接続
1.6%	充電ケーブル	1.1%	コネクタ-介する
1.6%	充電プラグ	1.1%	充電コネクタ-接続
1.5%	外部電源	1.1%	充電プラグ-接続
1.3%	充電コネクタ	1.0%	コネクタ-接続
1.3%	電源プラグ	0.9%	接続-状態
1.2%	接続+できる	0.9%	供給-蓄電装置
1.0%	接続部	0.9%	接続+できる-構成
...

確率の高い構成要素から、トピック Z'20は「充電の接続」に関するトピックと解釈できる

diff-PLSAでは、通常のPLSAのトピックに加え、運転者の補助、発電・蓄電、燃料電池、情報の取得・提供、組成物の製造などを含む50個のトピックが抽出されました

Z'01.エンジン制御

オン
始動要求 要求 判定 モード差 時点
オフ 所定条件 成立条件 電気走行モード プログラム
所定時間 解放 ハイブリッドECU所定値モータジェネレータ
経過 再始動 充電状態継続 ハイブリッド車両
エンジン ECU実行中 禁止 NO 停止 開始

Z'02.動力伝達

変速機構
出力要素 出力軸 連結+できる 出力部材 電気機械 適用
動力逆送車装置 駆動軸 遊星歯車機構 キャリア
入力部材 トルク 動力伝達+できる ギヤ トランスミッション
連結 入力軸 回転駆動+できる 伝送 サンギヤ 伝達+できる
リングギヤ クラッチ 内燃機関 変速比 動力伝達機構
動力伝達駆動機

Z'03.差動機構などを備えた動力伝達の制御

変速制御 運転状態 差動部 変速比
駆動系 無段変速機 動力伝達装置 高速回転 自動変速部
駆動輪 車両用駆動装置 回転部材 ハイブリッド走行モード 自動変速機
動力伝達経路 車両用動力伝達装置 エンジン回転速度
エンジン モータジェネレータ 変速ショック 電気式差動部
電気走行モード 差動状態 制御装置 変速部 変速
エンジン始動時 エンジン始動 変速時 ショック

Z'04.回転運動

駆動軸 同軸
伝達機構 軸方向外周面 回転カ一方
ハブ 回転軸 シャフト 羽根車 ロータ 中心 直線運動
クランクシャフト 磁石 回転運動 チェーン 相対回転+できる
角度位置 回転体 回転 回転+できる 回転駆動
回転方向 伝送 ステータ ギヤ軸 反対側

Z'05.ロータ・ステータなどモータの構成

シャフト 巻装
スロット 整流子 回転磁界 外周 径方向
極 界磁巻線 回転+できる コイル ヨーク 電機子 プラシ
永久磁石 巻線 車両用交流発電機 軸方向 固定子 巻線
固定子 鉄心 回転軸 整流器 ステータ プラシホルダ
内周 ロータ 周方向 磁極 摺 整流

Z'06.モータ制御(トルク制御や回転数制御など)

制御トルク制御 回転数トルク推定 モータ一致
車速 回転角度 出力トルク モータ制御装置 電気角
トルク指令値 位相トルク指令 回転位置 モータ制御部
モータトルク 電圧指令 電流指令演算 スイッチング素子
インバータ 回転速度 誘起電圧 加算電流 補正 絶対値
振幅

Z'07.油圧ポンプなどを利用したモータ駆動

電力 供給 車輪 走行用 後輪
駆動 エンジン 動力トルク ハイブリッド車両 再生電力
走行用モータ 発電 油圧ポンプ 発電電力 電気エネルギー
発電機 油圧モータ 駆動輪 回転数 モータジェネレータ
動力源 内燃機関 駆動力伝達 圧油 モータ
前輪

Z'08.ブレーキ

押圧力操作部材 摩擦部材 リンパ 運転者付与
ピストン 制御操作部材 摩擦部材 プレーキ操作 液圧
車輪 モータシンジラダ装置 制動力 プレーキ操作 液圧
スレーブシンジラダ プレーキ液圧 入力装置 マスタシンジラダ
ブレーキ 電気ブレーキ プレーキ液圧 制動力 プレーキペダル
ブレーキアクチュエータ アクセルペダル 操作力 ホールシンジラダ
操作量 回転部材操作者 操作

Z'09.状態に応じた制御、運転者の操作補助

点滅 移行 停止 操作 作動 禁止
運転者閉状態 検出+ない 検出部 検出結果 アンロック
ソフトリバー処理 シフトレンジ パーキングレンジ
ロック 絶縁抵抗 制御手段 運転席オン 解除 キースイッチ
開状態 自動的 検出 警告 切替故障
開始

Z'10.コンバータとバッテリー昇降圧

給電 出力電圧
コンデンサ 主電源 再生電力 並列 昇圧 放電
供給 電気負荷 低電圧 バッテリー 補機 バッテリー サブバッテリー
電圧コンバータ DCコンバータ 降圧 車両用電源装置
高電圧 バッテリー鉛蓄電池 電気二重層コンデンサ
電源システム オルトスタータ 電源装置 電圧コンバータ
バッテリー 発電機 電力 直流

Z'11.直流と交流の電力変換

直流 電力変換部 コンバータ 交流電圧
駆動 主電動機 フィルタコンデンサ 直流電源 架線 スイッチング動作
交流電力 変圧 電気車制御装置 補助電源 直流電力
直流電圧 集電装置 昇圧 電力変換装置 交流電圧
交流電動機 インバータ 再生電力 電圧 交流 供給
変換電力変換

Z'12.回転力などの電気エネルギー変換

海水
水車 太陽光 上下動 流体 蓄電
風 運動エネルギー 風車 カ回転 ナセル 発電システム
機械エネルギー 水 回転エネルギー 回転軸 電気エネルギー
波 回転力 発電機 タービン 風力発電装置 エネルギー
水流 変換 排気ガス 発電 水面 羽根

Z'13.エネルギー効率の向上

快速損失 技術 走行+できる 精度 回避
実現 燃費 電気自動車 無駄 乗員 確保 走行距離
長寿命化 効率 エネルギー 効率劣化 ハイブリッド車両
高速走行時 航続距離 短縮 燃料消費量 消費電力
電力消費 走行中 違和感 起因 性能 最小限

Z'14.発電と蓄電

回収
自由 蓄電 電力量 蓄電量 発光 蓄電池
電力 風力 発電 蓄電+できる 発電夜間 運動エネルギー
電気エネルギー 再生可能エネルギー 発電電流 ソーラーパネル
太陽電池 化石燃料 消費 風力 発電装置 充電機能
エネルギー源 発電+できる 活用 走行 回転力 稼働
照明 車両外部

Z'15.電池モジュールの提供

内蔵 構造 電極 相互
積層方向 単電池 バッテリー ベースプレート 横方向
隣接 スタック 電池システム 電池モジュール 列 充放電+できる
冷却 ユニット 冷却剤 中型 積層 垂直 電池パック 底部
配列 パスバ 下端 片側 構築 直列外形

Z'16.燃料電池

酸素 漏洩 空気排出 反応 電力
排気ガス 燃料タンク 燃料電池 加水 水素
酸化ガス 燃料電池システム 電気化学反応
燃料ガス 電気分解 内燃機関 燃焼室 燃料電池 車両
供給+できる 排気通路 供給ガス 浄化 貯蔵 発電
燃焼 温度

Z'17.二次電池の構成

外装部材
含窒素リチウムイオン バッテリーケース 含有方向 層構造 巻回
電解質 集電体 二次電池 電極 正極 活物質 電解液
非水電解質電池 表面 セパレータ 活物質 リチウムイオン電池
両面 正極 負極 電解要素 積層 負極活物質
電気抵抗 負極集電体 外装体

Z'18.バッテリーの充放電

充 電源 警報建物 放電 給電
給電装置 蓄電池 利用者外部電源
負荷電力変換装置 授受 充放電装置 系統電源
停車時 指示 電力供給システム プレーガ 放電装置
電源電圧 充放電 電盤 双方向車載 電力システム 停電
コネクタ住宅 送受信 電力 開閉器

Z'19.充電システム

充電
充電電流 駐車場 充電動作 駐車 利用者外部電源
記憶 充電電圧 充電+できる 蓄電池 充電制御装置 蓄電装置
充電スタンド 充電時 充電制御部 時間帯 充電システム
電気自動車 商用電源電力量 充電電圧 充電制御
終了 集合住宅 充電電力電気料金情報 ユーザ
管理

Z'20.充電の接続

充電電圧線
電源ケーブル 商用電源 充電時 電力変換部
車両側コネクタ インレット 接続部 充電口 スタンド電源プラグ
外部電源 コネクタ 充電ケーブル コンセント 充電プラグ
着脱 接続+できる 電力ケーブル 電力 蓄電装置 充電コネクタ
外部機器 プラグ 車両外部ケーブル 接続装置 外部
給電部

diff-PLSAでは、通常のPLSAのトピックに加え、運転者の補助、発電・蓄電、燃料電池、情報の取得・提供、組成物の製造などを含む50個のトピックが抽出されました

Z'21.非接触など受給電装置

電力
電化電磁誘導 給電回路 給電部
駐車装置 非接触給電装置 給電部 パレット 非接触給電システム
給電装置 送電送電コイル 非接触受電コイル
地上 給電+できる受電部 受電部 コイル 給電受電装置
駐車スペース 非接触給電装置 外部

Z'22.車両用空調など熱交換

蒸発 加熱 空気 排出 発熱
圧縮機 ヒータコア 冷媒 熱 冷却 温度冷却装置
ラジエータ 圧縮 車両用空調装置 熱交換 電気ヒータ
流量 車室内蒸発器 熱交換器 熱源 流通 冷却水
熱媒体 暖房 循環 ヒータ 放熱 送風

Z'23.冷却装置と放熱

流入
収容 逆止弁 循環 閉塞 吸気口
筐体 吸込口 撮像装置 収容部 走行風 排出 冷却装置
冷却システム 送風機 流入口 放出 冷却風 排出口
吐出口 所定方向 ケース内 排気口 放射状 仕切壁
排気管 外気 開口部 冷却 送風

Z'24.信号の入出力と検出

指示信号 センサ 入力 オン 検出 情報
操作 マイコン 出力信号 CPU ステアリングシャフト 電気信号
出力+ない 処理部 検出信号 制御部 検出手段
発信 検出結果 制御信号 光信号 送信 スイッチ手段
回転角度 異常 制御回路有無 受信

Z'25.電気信号の取得と変換(センサ検出など)

乗員超音波 撮影演算判定 距離
点灯 検出値 判定結果 物体 電気信号 受光素子
加速度センサ 表示 圧力センサ 合成 撮像素子
検出手段 受信部 受光 圧力 車両制御ユニット
静電容量 衝突 点滅 消灯 検出 記憶 周期 有無

Z'26.電流・電圧の検出

オン 印加 比較 直列
基準電圧 電流値 蓄電素子 地絡スイッチ 電圧検出器
異常平滑コンデンサ 制御回路 電流検出器
電流センサ リレー インバータ回路 電流経路 スイッチング素子
オフ 充放電電流 直流電源 電圧 電圧センサ
半導体スイッチ 短絡故障 電流検出 コンバータ
交流電圧

Z'27.温度、電流、充電量などの検出と制御

推定 取得
検出結果 発電電圧 判定手段 温度 充電量
積算充電状態 検出手段 充電率 演算 電気負荷
判定結果 残量 電圧検出手段 充放電 検出算出手段
放電 蓄電量 電流量 電流値 バッテリ制御装置 記憶手段
制御手段 発電電力 関値 充 二次電池

Z'28.演算や推定、測定などのステップを含む方法

加算段階 時間
演算比較 方法 推定 記憶 閾値 起動 関数 定義
推定値 測定値 パラメータ 差 再充電 充電状態
値 終了測定 記録取得 判定 最大 運転状況
更新 レベル 基準 セル

Z'29.情報の取得・提供(位置情報やバッテリー残量等)

特定
出発地 車載機器 送信 受信 検索ユーザ
残量 地図情報 目的地 記憶部 サーバ 記憶 無線通信部
充電スタンド ナビゲーション装置 表示部 位置情報
無線通信 情報 識別情報 バッテリ残量 車両情報
携帯端末 利用者 現在地表示 通信 取得
通知

Z'30.スイッチなど操作装置

接触 破断
入力装置 装着 保持+できる 当接
ストップランプ 形成+できる 操作部 操作体 スイッチ
固定接点 破損 操作+できる 車室内 スイッチ接点
傾き 電子機器 配線基板 操作段差 作成 可動接点 没入
操作時 変形 安価 勝手 力 誤操作

Z'31.車両用灯具

反射 貫通 近接 LED ねじ クランプ
接続ハウジング 射出 発熱体 車両用灯具 灯室内
ステアリングコラム 半導体発光素子 ステアリングシャフト
基板 光源 内装 給電部材 ランプ 照射 リフレクタ
透過 光 放熱部材 ハルブ 入射 固定部 表面

Z'32.掃除機

車輪連通 前方 開口 下面 前面 転動 上方
床面 本体 ケース 排気口 吸気口 下方 回転 両端 左右一対
吸込口 塵埃 電気掃除機 電動送風機
下ケース 着脱+できる ホース 被清掃面 集塵部
吸引 ハンドル 上面 本体 後方

Z'33.基板の構成

発光部 光源 露出 発光 面 導電 裏側 接合部
配線表示装置 検出部 基板 発光ダイオード 放熱フィン
半導体素子 車両用照明装置 表面 発光素子
光 センサ素子 はんだ付け ケース内 パターン 被検出体
ホールIC 照明装置 厚さ方向 反対側 電極 筐体

Z34.回路の接続(電力変換回路など)

並列
トランジスタ ダイオード 電圧源一端 端子変換器
他端 出力端子 コンデンサコイル 各相 インダクタンス
スイッチング素子 バス 一次コイル 相スイッチ 二次コイル
電流 カソード トランス 整流器 入力端子 アノード
電源スイッチ インダクタ 抵抗拘束 直列

Z'35.端子接続

配線 電装品 嵌合 ホルダ 貫通
絶縁性 接続部材 放射状 本 電線 パスパー
雄端子 端子 嵌合+できる ワイヤハーネス 雌端子
露出 導電部材 金属製 窓ガラス 他端 導体 接地コネクタ
突出 圧接一端 挿通 両端 端部

Z'36.部品・装置の収容ケース・筐体

ケース 挿通 フランジ 基板 露出
対向 電気部品 コネクタ カバー 筐体 突出 収容空間
制御回路 内面 電子制御ユニット 固定 電気接続箱
貫通孔 開口部 ケース内 蓋ハウジング ブラケット
収容部 挿入 収容 形成 一体 開口

Z'37.部品・装置の配置

方向 配表
配置 モータユニット サイドメンバ 前方 車体 上面
左右 車間 前方 較重 型車 両方 上下 左右一対 フロアパネル
車両後方 後輪 パワーコントロールユニット 下方 車両前後
バッテリーケース フレーム部材 車室 車幅方向 電池モジュール
搭載構造 車体 フレーム クロスメンバ 前後 上下方向 後方

Z'38.パーツなどの移動、位置

本体 操作
作業 全開状態 表示部 方向
閉塞 可動部材 所定位置 位置 脚部 アンロック位置
移動方向 係合 移動+できる ロック位置 可動パネル
付勢力 視認+できる 移動 係 押圧 解除 ロックレバー
阻止 操作+できる 移動 許容 回転
車両前方 規制

Z'39.構造の形成・方位

平行 凹部 交差
交互 外周 加熱部 突起 内面 先端 内周面
径方向 軸方向 導電性材料 面 周方向 外周面
突出方向 環状 対向隣接 形成 周縁 外側 間隔
断面 接触 垂直 部分 端面

Z'40.支持構造

変位一端 連結 昇降 固定 回転
コイルばね 保持部材 他端 回転+できる ベース部材
昇降+できる 移動 揺動+できる 係 車体 移動+できる
両端 支持部材 上下方向 支持部 平行 回転+できる
ブラケット 規制 支持 直交 立設 先端 縮

diff-PLSAでは、通常のPLSAのトピックに加え、運転者の補助、発電・蓄電、燃料電池、情報の取得・提供、組成物の製造などを含む50個のトピックが抽出されました

Z'41.装置やユニットの構成

液体
適合車両用ピストン周囲導入 形態
チューブ相互接続シャフト回転+できる チャンバ
長手方向軸受容 取り外し+できるプレート 移動+できる
同軸ハウジング 固定軸方向協働領域 電気機械
モジュール 配置 出口ガイド 車輪入口

Z'42.システム・方法の構成

調整監視 送信信号 ID 推進
無線測定値 応答 構成 伝送測定 データ通信 受信
センサ信号センサ プロセッサ 確立 エネルギー貯蔵装置
送信機 制御信号方法磁界 システム受信機
生成 命令 結合 感知

Z'43.その他方法

形式 熱機関
調整 駆動+できる 作動装置 トルク 車両用駆動
独立ドライブトレイン 連結+できる 歯車列 方法 ブレーキシステム
作動+できる 依存 トランスミッション 駆動部 パワートレイン
運転+できる 内燃機関 自動車 電気機械 電気システム
結合+できる 駆動装置 電気式 車輪 制御 システム

Z'44.組成物の製造方法(樹脂や電解液など)

式重量工程 Cr
質量一種 存在下塗布反応 混合 群ケイ素
化合物組成物 製造方法 ニッケルアルミニウム
錫 乾燥 含有 製造材料 被覆分散式中混合物
銅 金属 MnTi

Z'45.機能性組成物・成形品(耐熱性や耐衝撃性等)

材料 携帯電話 金型 樹脂 性
耐薬品性 電気部品用途 成形品 外觀 耐衝撃性
用途 熱可塑性樹脂 電気部品 ポリアリレンスルフィド 流動性
自動車部品 樹脂 機械的強度 耐ドラッキング性 電気特性
耐熱性 熱伝導性 溶融流動性 透明性 電気絶縁性
機械的特性 成形品 外觀 成形性 難燃性
電子

Z'46.製造の効率化(小型化や低コスト化など)

大型化 効率簡素化 作業性確保 防水性
軽量化対応+できる コンパクト騒音構造 コスト 実現+できる
小型化+できる 軽量 部品点数 小型化 低コスト
作業 信頼性 必要+ない 耐久性 小型 安価 製造コスト
省スペース 放熱性 高効率削減 実現

Z'47.不具合の防止(損傷、感電、盗難など)

侵入 信頼性 衝撃 衝突
劣化 未然 位置ずれ 破壊 異音環境下 異常不具合
意図+ない 感電 確保+できる 漏電 静電気 振動 損傷
起因 過熱 ノイズ 誤動作 ショート 影響 過電流
故障安全 誤検出 盗難

Z'48.その他

変形 静止
メインリレー 閉回路 圧電素子 路面
直列車両駆動用 安全装置 遮断動電 伸長 制御対象
バッテリー制御装置 監視 ハイブリッド トラクタ 空気圧
タイヤ 動作+できる 電池バック バイパス 電気接続
負 一側面 トレーラ バッテリー 電荷 並列 正

Z'49.タービン発電と船舶・飛行機への応用

日帰り旅行 運用 出力発電 燃料費ゼロ
既存 安価電気駆動 製造物全部 発電量 発電原価 宇宙到達費用
液体酸素圧縮駆動 横軸H軸車 太陽光加熱器 製造
改善 永遠 電気+液体空気+過熱蒸気 蒸気供給設備D
空気圧縮 容積圧縮仕事率 マッハ強い 容積 燃費
静翼 全動翼 船舶 最大速度 潜水 自動車 反転

Z'50.重力発電の活用による地球温暖化防止

地球温暖化 燃料費ゼロ 液体金属 電気駆動 先送り
サンゴ工場 電化全盛 発電量増大 水銀 既存火力原子力発電
船舶 重力加速度 加速 全廃 二酸化炭素 排気ゼロ 金属球
安価発電量 大気圧同速度 同容積仕事率 既存世界
全面電化住宅全盛 全廃地球温暖化 既存蒸気タービン発電
水 タービン海水温度上昇ゼロ 落差 重力発電蓄電池駆動
自然現象 高速度圧縮空気加速 垂直下方 重力加速度

※Z'49とZ'50は特定の出願人による重複した要約内容の特許から抽出された

※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 (関係の強い上位5つの単語を赤色で表示している)

【比較1】 differential PLSAは頻度の低い要素でトピックが構成される

通常のPLSAとdiff-PLSAで抽出された同様の意味のトピックを比較すると、diff-PLSAの方がより具体的な表現がトピックの上位語となり、細かい要素技術を抽出できています

「ブレーキ」に関するトピック

「非接触受電などの給電」に関するトピック

各トピックにおいて所属確率Pの高い上位7つの単語・係り受けと、それぞれの出現頻度n(出現文章数)

通常のPLSA

Z05

P(X Z)	n(X)	X: 単語	P(Y Z)	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

Z14

P(X Z)	n(X)	X: 単語	P(Y Z)	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

所属確率の高い単語は全体の出現頻度も高い

全体の出現頻度が低い単語でも所属確率が高い

diff-PLSA

Z08

P(X Z)	n(X)	X: 単語	P(Y Z)	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	2.3%	49	ブレーキ-備える
1.3%	111	操作量	2.0%	92	車両-ブレーキ

Z21

P(X Z)	n(X)	X: 単語	P(Y Z)	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	1.8%	20	非接触-受電
1.8%	124	受電部	1.8%	20	供給-給電装置

※各トピックにおいて所属確率の高い順に名詞・係り受けを並べたとき、累積確率が50%になるまでの名詞・係り受けの平均頻度を対象に統計検定(Welchのt検定)を実施したところ、通常のPLSAとdiff-PLSAで1%有意の違いが見られた

【比較2】 differential PLSAでのみ抽出されるトピックがある

自動車の付加価値を高める重要な技術など、diff-PLSAでのみ抽出されるトピックがあり、頻度の高い要素を中心に典型的なトピックを抽出する通常のPLSAでは拾いきれません

「運転者の操作補助」に関するトピック

各トピックにおいて所属確率Pの高い上位7つの単語・係り受けと、それぞれの出現頻度n(出現文章数)

diff-PLSA

Z'09

P(X Z)	n(X)	X: 単語	P(Y Z)	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
...

【解釈】運転者の操作を補助したり自動停止などの運転アシストに関する技術

特許の原文例

車両停止時の駆動源の切り替えによる車両の動き出しの防止又は車両の動き出しを運転者に知らせることのできる電気自動車を提供すること。アクセルペダル及びブレーキペダルの踏み込みがなく、エンジンの作動中であり、シフト位置が走行レンジである車両停止時に、クリープ走行制御の駆動源がモータからエンジンに切り替わった際のトルク変動により車両が動き出した際には、パーキングブレーキによる制動を自動的に行い、警報器及び警告灯による運転者への注意喚起を行う。

「情報の取得と提供」に関するトピック

Z'29

P(X Z)	n(X)	X: 単語	P(Y Z)	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
...

【解釈】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という技術

特許の原文例

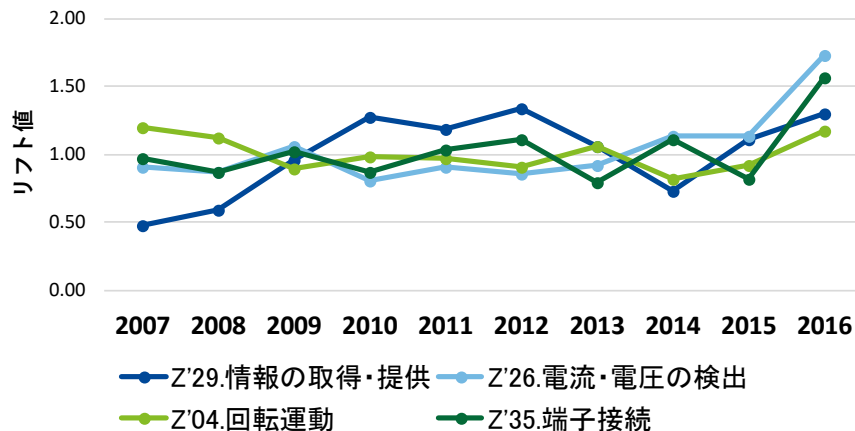
充電スタンドへの電気自動車の有効なナビゲーションを行う。本発明は、複数の充電スタンドの位置情報を記憶するデータベースと、電気自動車と通信して、電気自動車が搭載する電池の種別、エネルギー残量、エネルギー効率、現在位置を受信し、現在位置から電池残量およびエネルギー効率によって到達可能な充電スタンドを見つけ、その位置情報と、その充電スタンドが保持するエネルギー残量の情報を電気自動車に送信する充電スタンドロケータとを備える。

トピックをベースとした特徴の可視化

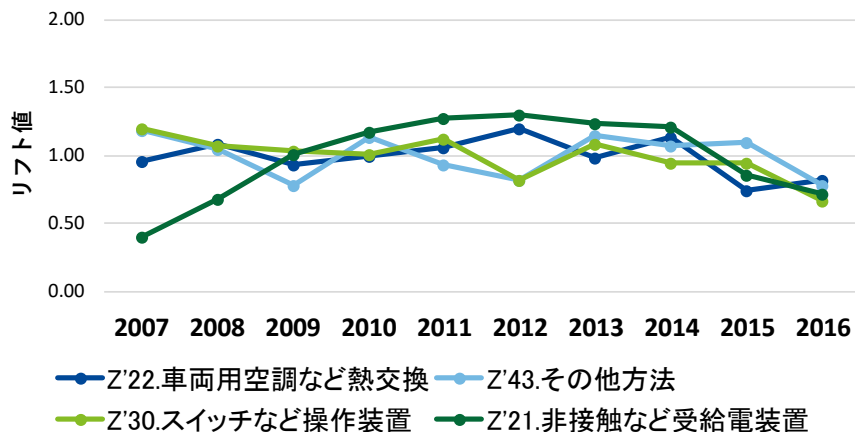
トピックのスコアを出願年や出願人の属性軸で分析することで、技術トレンドや競合他社の動向を可視化でき、技術戦略を検討する上で効率的にインサイト獲得が期待できます

トピック×出願年のトレンド分析

近年上昇傾向にあるトピック



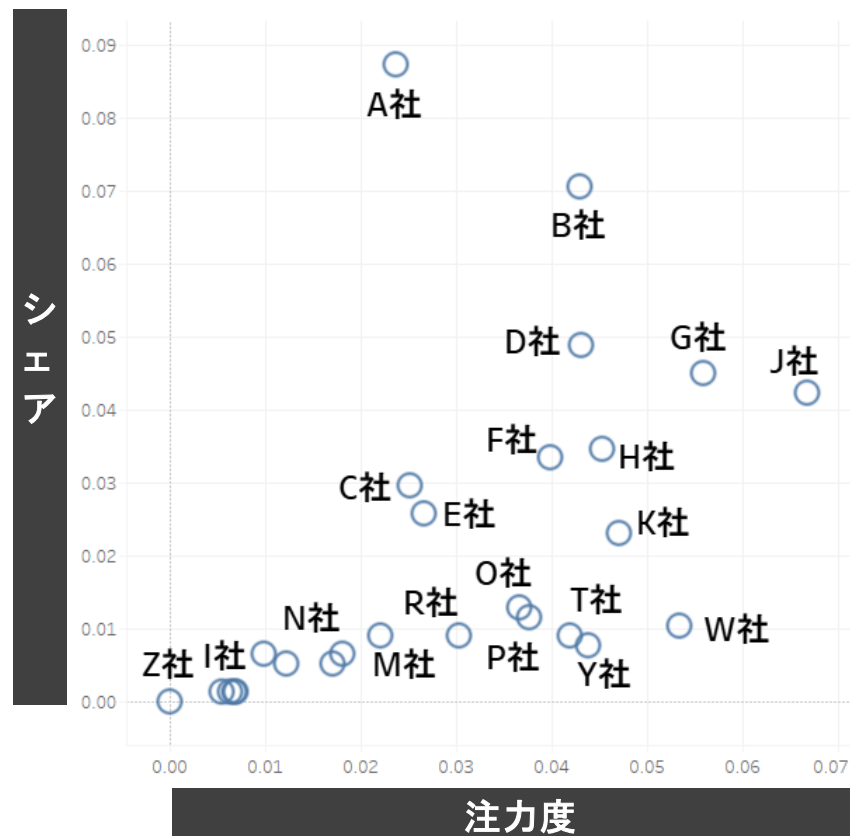
近年下降傾向にあるトピック



トピック×出願人の競合分析

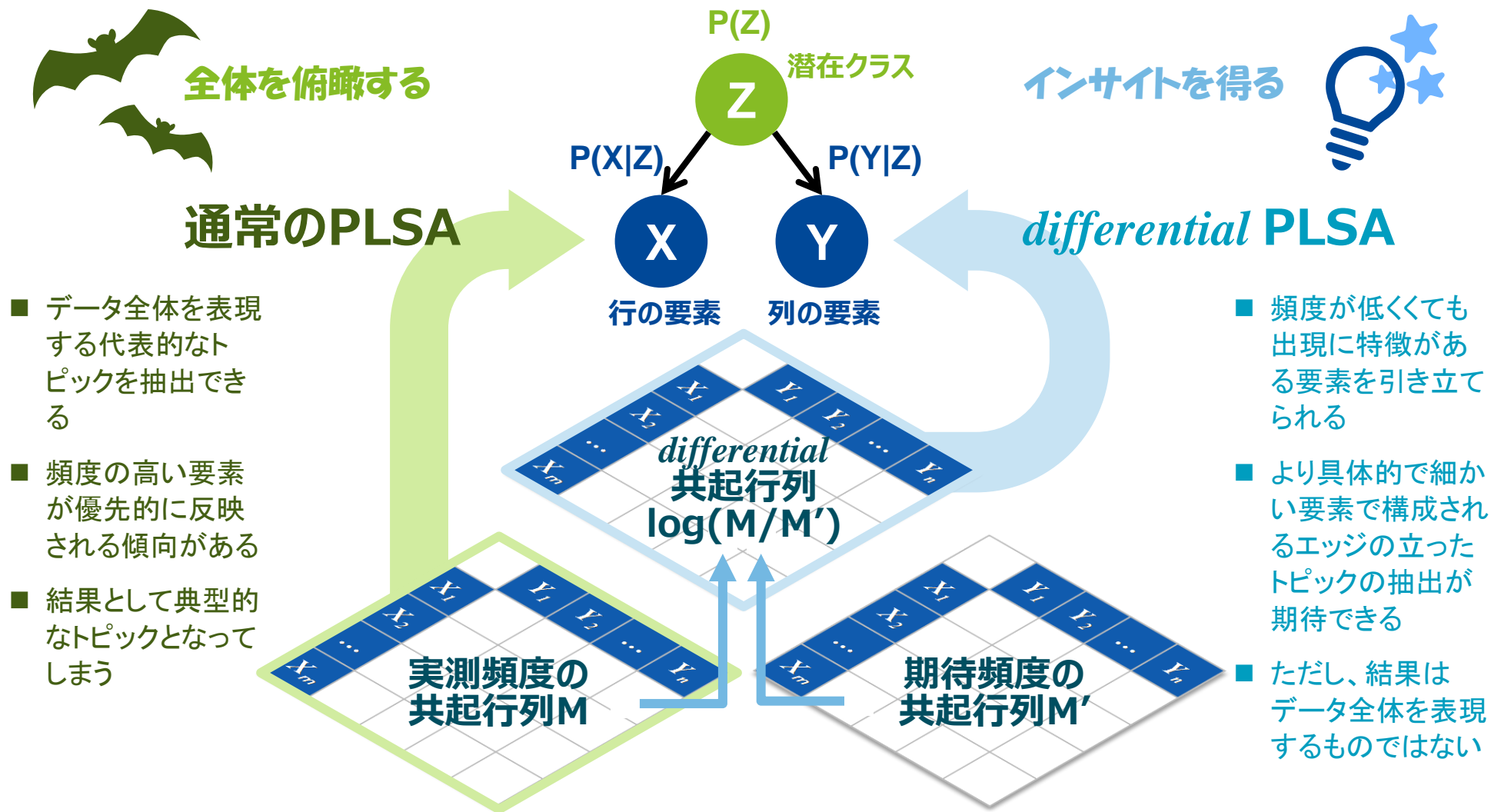
各トピックに対する出願シェアと注力度を軸とした出願人のポジショニングマップ

「Z'29.情報の取得・提供」のポジショニングマップ



通常のPLSAとdifferential PLSAのまとめ

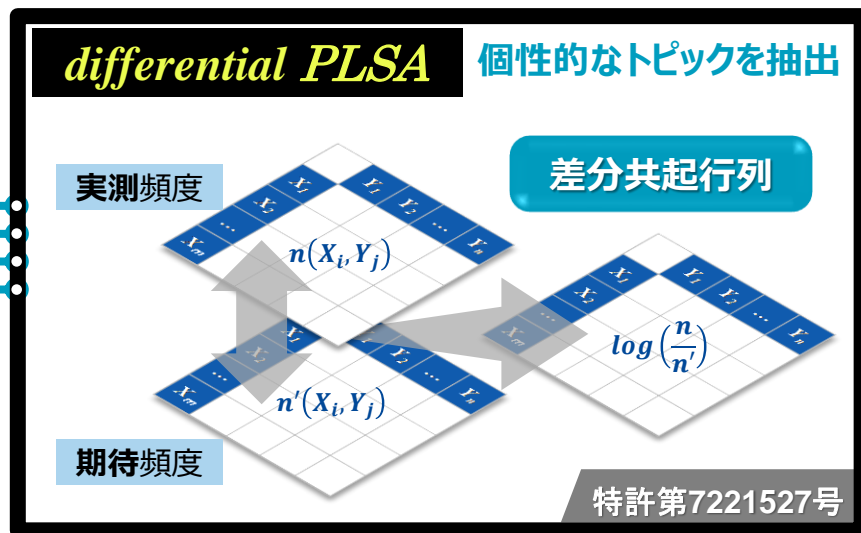
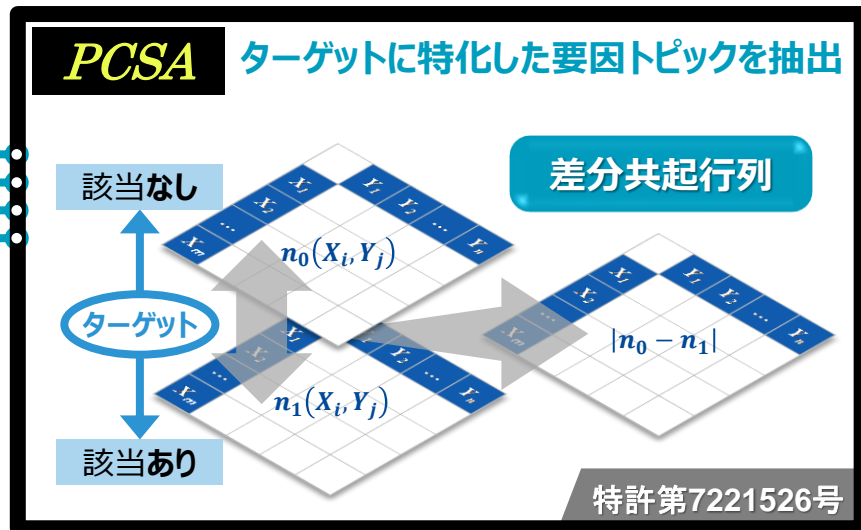
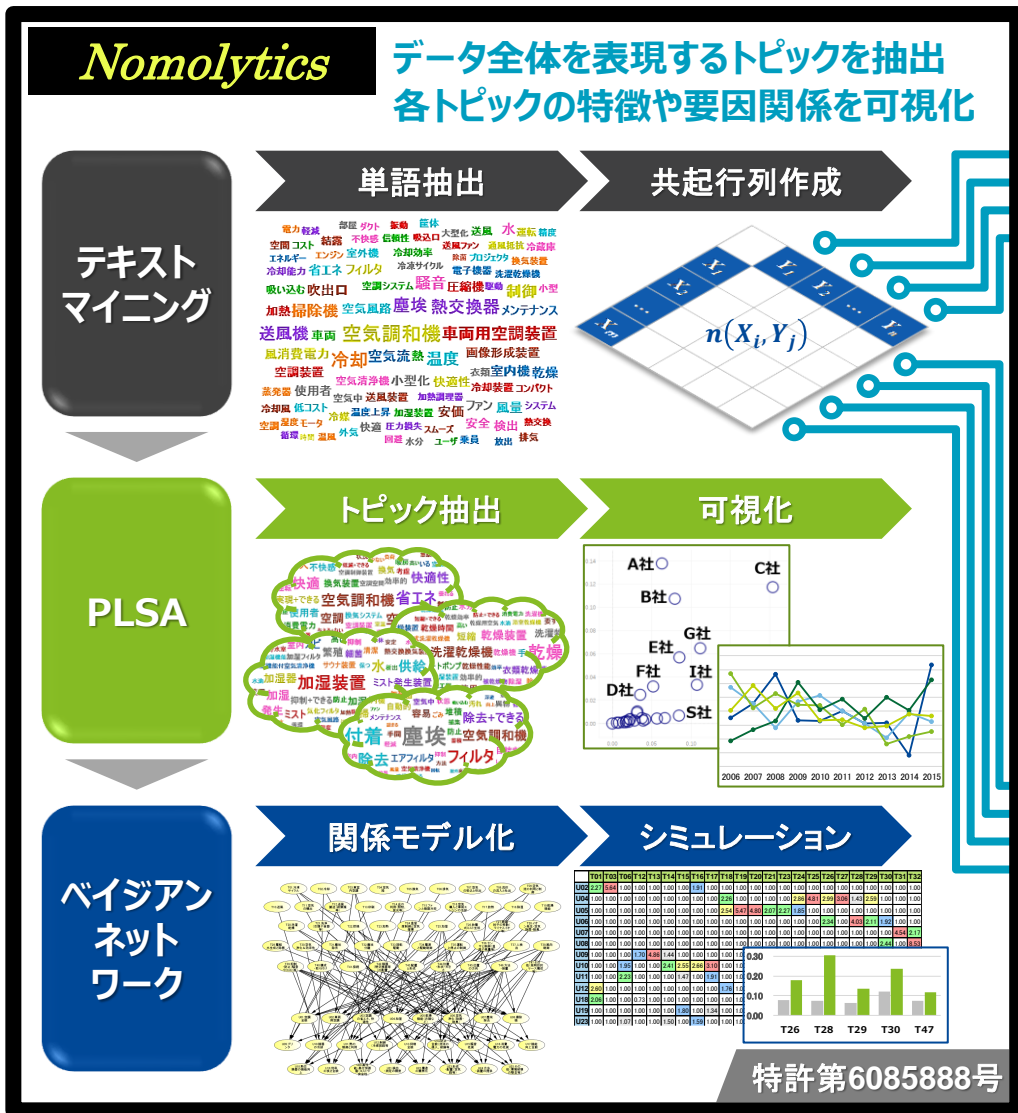
通常のPLSAでは全体の代表的なトピックを把握できますが、膨大で複雑なビッグデータから新たな気づきとなるインサイトを獲得したい場面では、*diff-PLSA*の適用は効果的です



6. まとめ

トピックをベースにテキストデータに潜む特徴を可視化する技術として、PLSAを軸とした、

①Nomolytics、②PCSA、③differential PLSAという3つの技術を開発しました



Nomolyticsを特許文書データに適用することで、特許の要約をトピック化し、トレンドや出願人の競合分析をしたり、用途と技術の関係を分析することで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×AI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

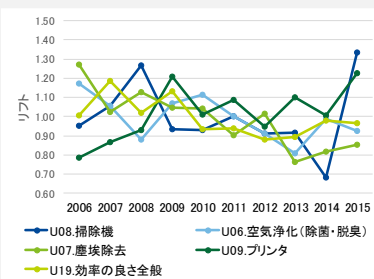
技術のトピック



用途と技術の関係分析

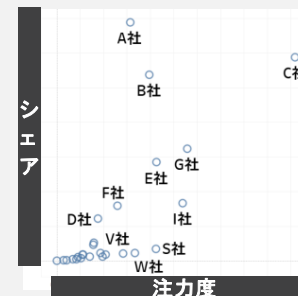
B トレンドの分析

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



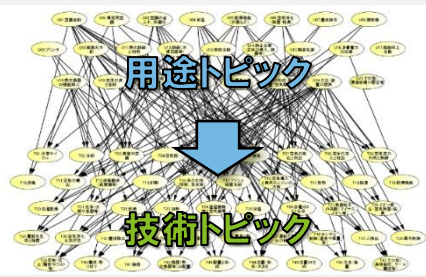
C 競合他社の分析

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



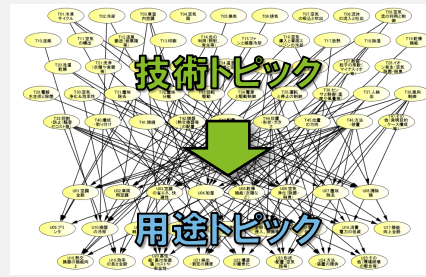
D 用途⇒技術の関係分析

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や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、PCSA、diff-PLSAは様々な業務のテキストデータに適用できます



口コミ

- 顧客ターゲット別の関心事を把握
- 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

ご清聴ありがとうございました

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/>

※ 資料の内容を引用または転載される場合は、必ずその旨を明記いただくようお願いいたします。

株式会社アナリティクスデザインラボ

