Analytics Design Lab

（株）技術情報協会主催セミナー
生成AI・人工知能による特許調査、分析と活用の仕方

今さら聞けない自然言語処理技術の体系的解説と複数のAI技術を応用した特許文書分析

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

2023年12月14日

会社紹介と自己紹介

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

会社概要：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

Analytics Design Lab

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

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

代表略歴：野守耕爾

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

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

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

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

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

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

■ 2017年6月～
株式会社アナリティクスデザインラボ 設立

© 2023 Analytics Design Lab Inc.

コンサルティングサービスの概要

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

分析受託サービス

お客様のデータをお預かりして弊社がデータ分析を実施し、結果をご報告します

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

アドバイザリーサービス

お客様ご自身で実施されるデータ分析・活用のご助言、ご指導をします

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

テーラー研修サービス

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

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

© 2023 Analytics Design Lab Inc.

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

© 2023 Analytics Design Lab Inc.

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

学会での受賞歴

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

© 2023 Analytics Design Lab Inc.

特許文書分析　相談ランキング

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

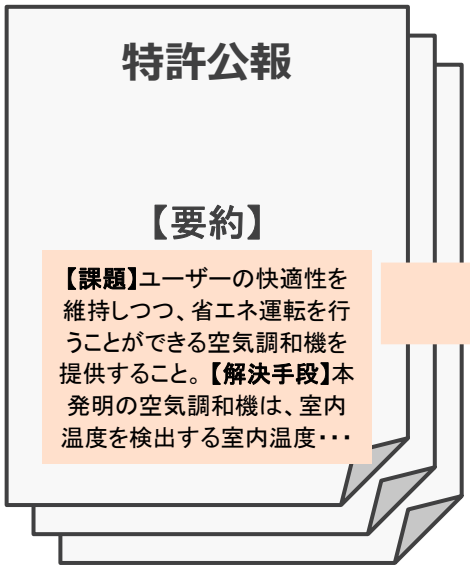
© 2023 Analytics Design Lab Inc.

講演目次

© 2023 Analytics Design Lab Inc.

Nomolyticsによる特許文書分析のアプローチ概要

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、そのトピックをベースに技術戦略に資する特徴を可視化します



文書情報(要約)を分析

テキストマイニング

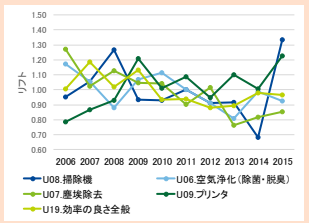
PLSA AI

ベイジアンネットワーク AI

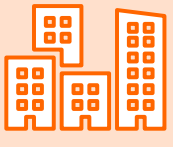
トピックに類型化して母集団の全体像を把握

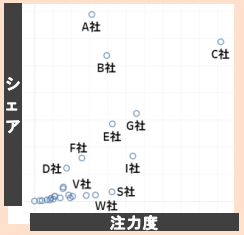
トピックをベースに

トレンドを分析

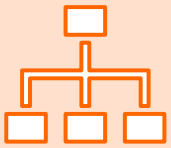


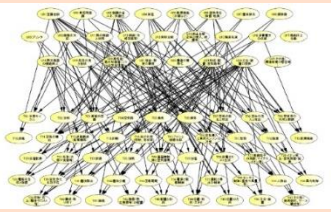
競合他社の動向を分析





用途と技術の関係を分析





★★★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

© 2023 Analytics Design Lab Inc.

1. テキストマイニングと自然言語処理技術

© 2023 Analytics Design Lab Inc.

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

© 2023 Analytics Design Lab Inc.

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**
文章を意味の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

函館できれいな夜景を見た

↓ **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 / きれいな / 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）

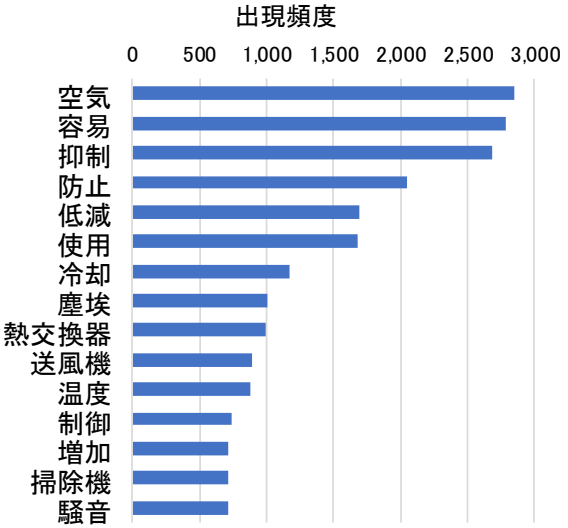
© 2023 Analytics Design Lab Inc.

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

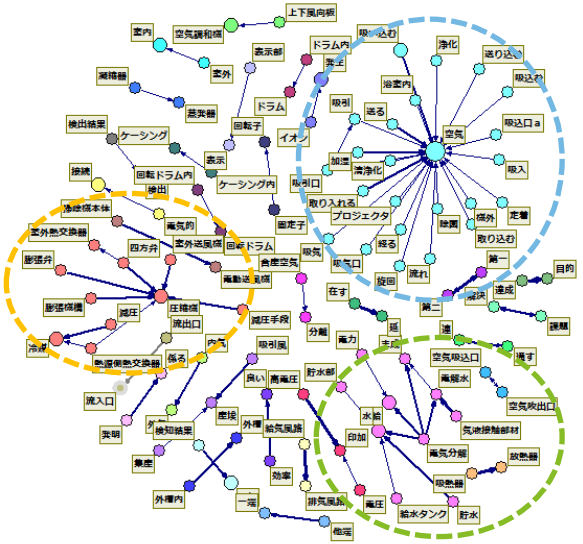
頻度集計

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



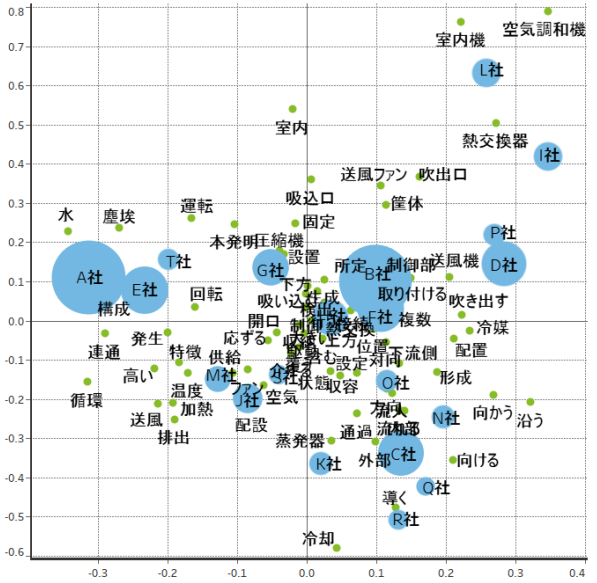
共起ネットワーク

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



コレスポンデンス分析

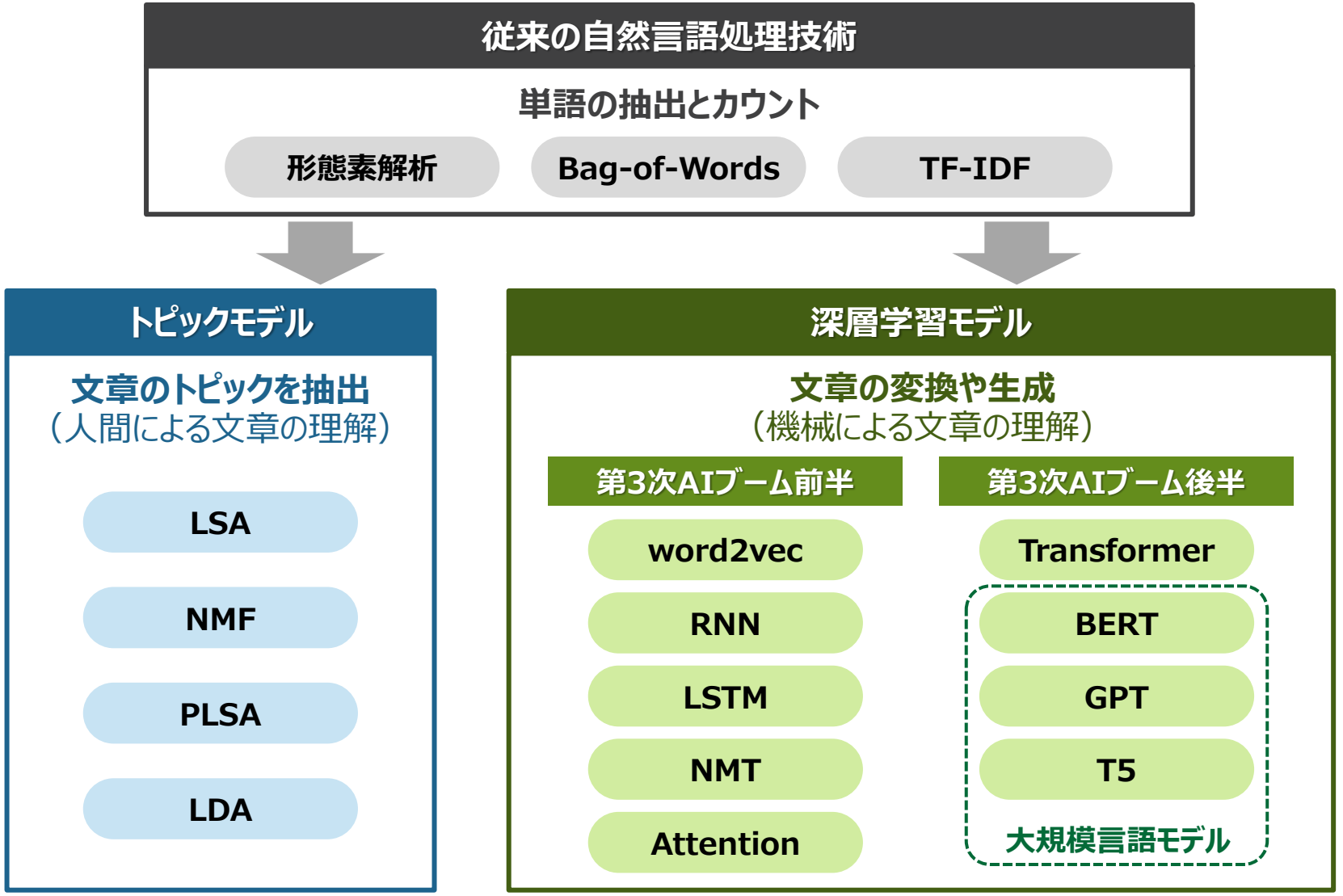
属性と単語の対応関係を同じ平面上にマッピングし、その位置関係から属性（出願人など）の傾向を把握する



© 2023 Analytics Design Lab Inc.

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



© 2023 Analytics Design Lab Inc.

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

© 2023 Analytics Design Lab Inc.

Bag-of-Words (BoW)

Bag-of-Wordsは各文章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

- 各文章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、その単語が全文章で見た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、意味の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
- 単語の数がベクトルの次元数となるため、高次元データとなり複雑で、計算処理が難しくなる

イメージ図

ID	ホテルの口コミ	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

© 2023 Analytics Design Lab Inc.

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、より文章の特徴を捉えた分析が可能になります

概要

- 各文章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)
  - 各文章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
  - 文章の長さの影響を受けるため、各文章の全単語頻度で除すことが多い
- IDF (Inverse Document Frequency)
  - ある文章で頻出する単語でも、それがどの文章でも頻出するなら重要とはいえない
  - IDF=log（全文章数 / 単語wが出現する文章数）
  - 多くの文章で現れる単語は値が小さく、特定の文章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、例えば文章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

© 2023 Analytics Design Lab Inc.

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

© 2023 Analytics Design Lab Inc.

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

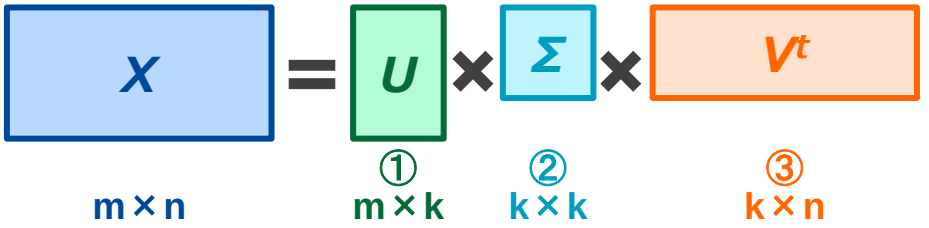
概要

- 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
  - 理論的には特異値分解は主成分分析と同じで、「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 （特異ベクトル）
    ③トピック×単語 （右特異ベクトル）
  - 「トピック×トピック」は対角行列であり、「文章×トピック」「トピック×単語」の行列はそれぞれトピックに対して直交している（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

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



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

Vt	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

© 2023 Analytics Design Lab Inc.

NMF (Non-negative Matrix Factorization)

NMFは「文書×単語」の行列を非負の行列「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

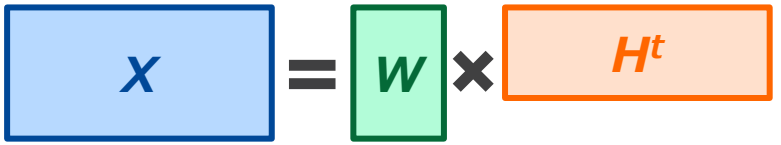
概要

- NMF（非負値行列因子分解）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
- 計算効率は比較的高い
- 分解された「文書×トピック」「トピック×単語」の行列は、トピックに対して直交しておらず、トピックのベクトル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

$$X = WH^t$$



X	W	Ht
m×n	m×k	k×n
文書×単語	文書×トピック	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

© 2023 Analytics Design Lab Inc.

PLSA (Probabilistic Latent Semantic Analysis)

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

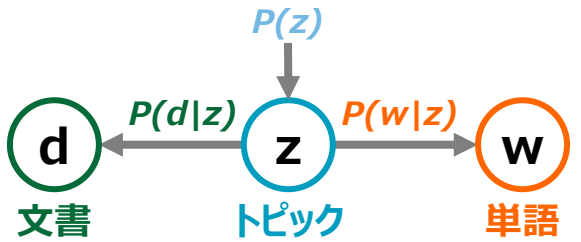
概要

- PLSA（確率的潜在意味解析）は、LSAの特異値分解の表記を確率モデルで表記した手法であり、「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
- 「文書×単語」行列（同時確率P(文書, 単語)）から以下3つの確率分布を推定する
  ①P(文書|トピック), ②P(単語|トピック), ③P(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
- 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
- 各トピックは互いに独立の仮定を置いている
- 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

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



PLSAの結果例

P(文書d | 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(単語w | 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(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

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

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

$\Sigma_z P(z) = 1$

© 2023 Analytics Design Lab Inc.

LDA (Latent Dirichlet Allocation)

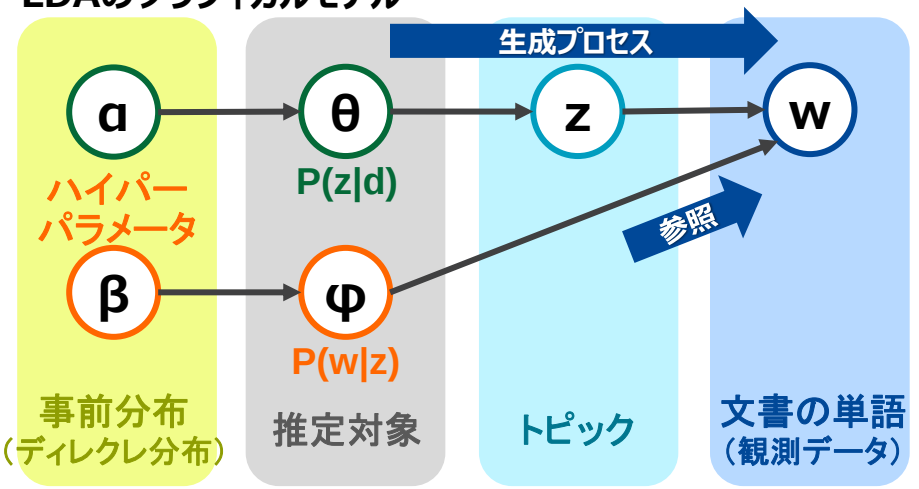
LDAはPLSAをベイズ的に拡張させて事前分布を導入した手法で、汎化性能が高く過学習を抑制でき、新しい文書のトピックも推定できますが、抽象度の高い結果になりえます

概要

- LDA（潜在ディリクレ配分法）は、PLSAをベイズ的に拡張させた手法であり、ディレ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θ=P(トピック|文書)とφ=P(単語|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
- 汎化性能が高い一方で、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。特定のトピックが各単語を生成する確率分布P(w|z)はφで表し、そのトピックの下で単語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を持つディレクレ分布に従う。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、観測データ（単語w）からθとφを推定する。

ハイパーパラメータα,βの大きさとディレクレ分布の特徴

α,β>1	事前分布は平滑化された形となり、多様な要素を含む
α,β=1	事前分布は一様分布となり、PLSAと近い振る舞いをする
1>α,β>0	事前分布は集中的な形となり、確率が一部の要素に偏る

© 2023 Analytics Design Lab Inc.

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）

© 2023 Analytics Design Lab Inc.

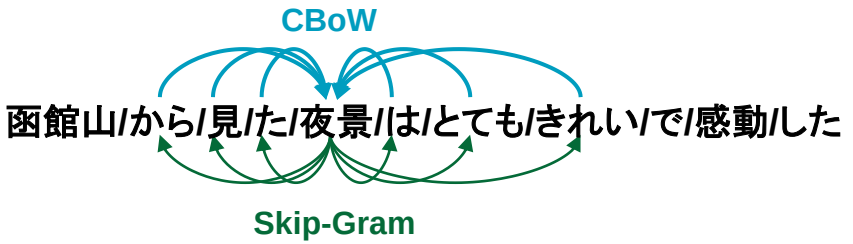
word2vecは単語の意味や類似性を捉えたベクトル表現を得ることができ、単語間の類似度や関係性を捉えた自然言語処理が可能となります

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が線形なモデル構造と学習をしているため単語ベクトルの演算が成立する

© 2023 Analytics Design Lab Inc.

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を処理することで、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1986年に発表された

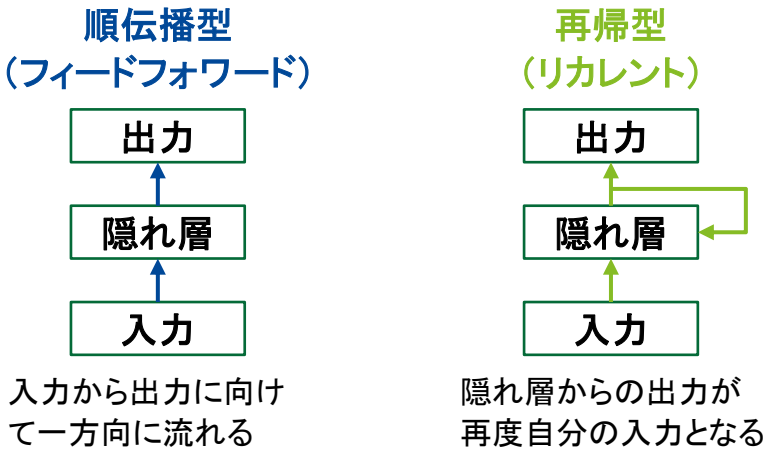
- ➢ 文章を単語の系列データとして捉え、単語を一つずつ順番に逐次的に処理をする
- ➢ 各ステップで新しい隠れ状態を生成し、それが次の隠れ状態の入力にもなる再帰的な処理をする
- ➢ 過去の処理情報を保持して、それが現在の入力に影響を与える構造により、単語の順序（文脈）を考慮した処理ができる

■ 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

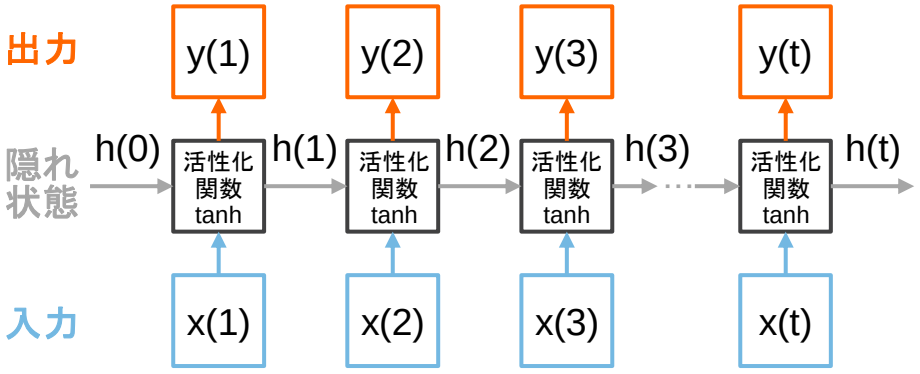
- ➢ 再帰的処理により隠れ状態が次の隠れ状態に連携し、長い系列でネットワークが深くなると、誤差逆伝播法において勾配爆発や勾配消失が起き、過去の遠い情報を最適に保持できなくなる
- ➢ 勾配爆発は、再帰処理により、同じ重みが繰り返し乗算され、勾配が指数関数的に増加する
- ➢ 勾配消失は、再帰処理により、微分が0に近い活性化関数を繰り返し通過し、勾配が急速に小さくなる

イメージ図

ニューラルネットワークの構造の種類



RNNの処理



※h(0)は初期化された隠れ状態（要素ゼロ）

© 2023 Analytics Design Lab Inc.

LSTM (Long Short-Term Memory)

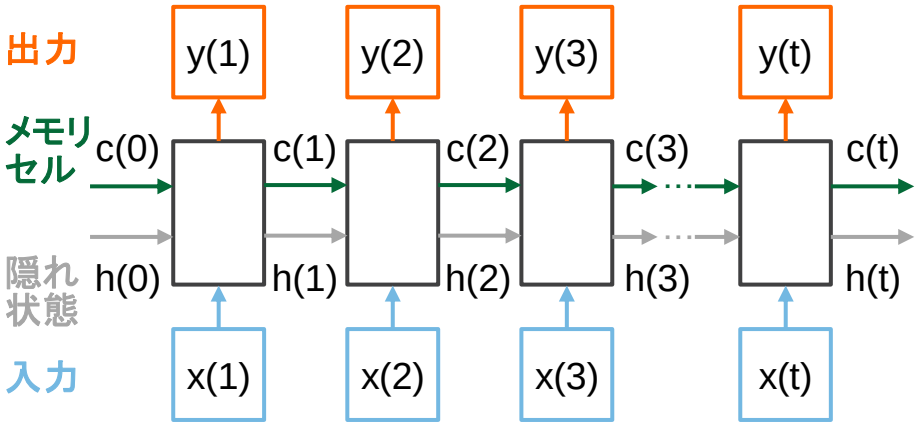
LSTMはRNNの長期依存性の問題を解決したネットワーク構造を持ち、過去の重要な情報を保持し、不要な情報を忘れる能力を有し、長い系列のデータの処理ができます

概要

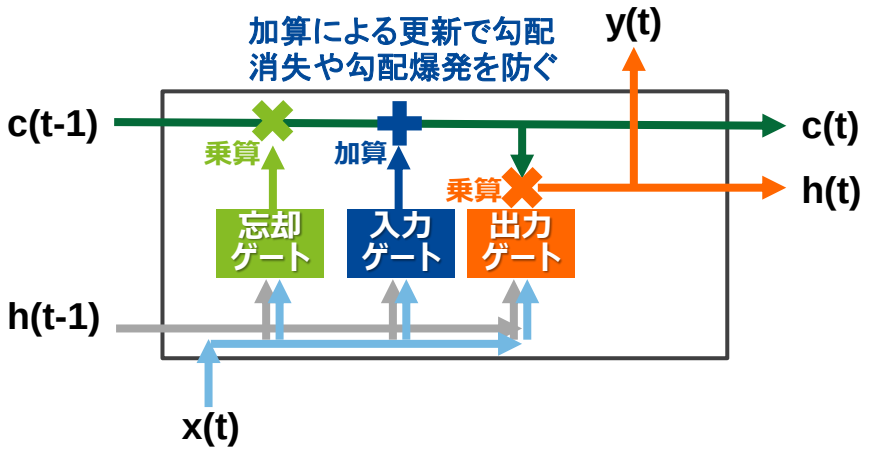
- LSTMはRNNと同じく系列データの処理モデルで1997年に発表されており、長い系列の処理ができる
  - RNNの隠れ状態h（短期記憶）に加え、メモリセルc（長期記憶）の情報がセルからセルに受け継がれる
  - 忘却ゲート、入力ゲート、出力ゲートという3つのゲート構造を持ち、各ゲートが情報の流れを制御し、長期記憶が加算によって更新されることで（RNNの更新は乗算のみ）、長期依存性の問題を解決する
  - 忘却ゲートは、受け継がれたメモリセルcの長期記憶の情報のうちどれを消去するか制御する
  - 入力ゲートは、現在の入力xと前の隠れ状態hを使って、新しく記憶するべき情報の候補を生成し、どの情報を入力し記憶させするか制御する
  - 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



© 2023 Analytics Design Lab Inc.

NMT (Neural Machine Translation)

NMTは機械翻訳など文章を文章に変換するseq2seqのモデルで、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

■ NMTは系列データを系列データに変換するseq2seqモデル（Encoder-Decoderモデル）で、機械翻訳を主な用途とし、2014年に発表された

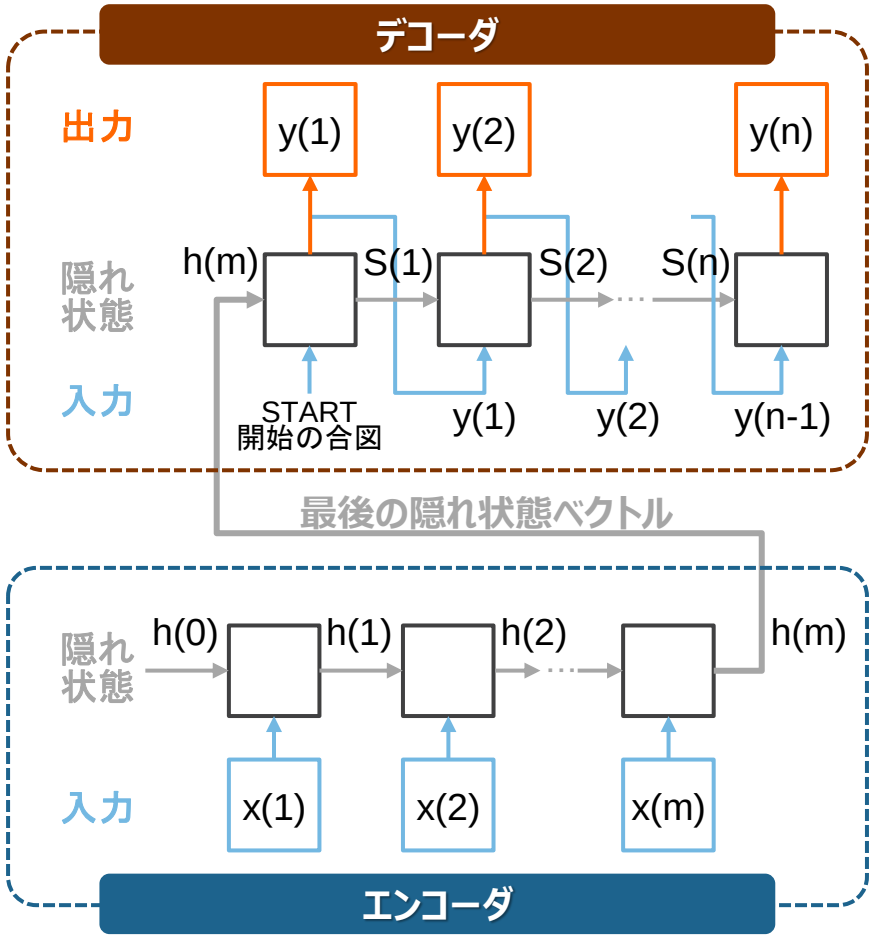
- 文章をベクトルに変換する**エンコーダ**（seq2vec）と、ベクトルを文章に変換する**デコーダ**（vec2seq）が連結した構造を持つ

■ 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
- veq2seq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
- seq2seqは、RNNやLSTMは入力の度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れば処理できるが、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

NMTの構造（RNNを用いた場合）



※h(0)は初期化された隠れ状態（要素ゼロ）

© 2023 Analytics Design Lab Inc.

Attention

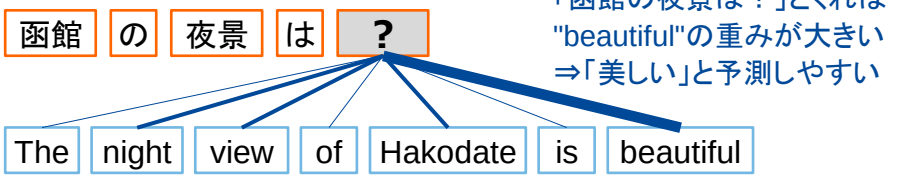
Attentionは入力文を出力文に変換する際に、入力のどの単語に注目すべきかを考慮する仕組みで、入力と出力の依存関係を捉えることができ、精度の高い結果を生成できます

概要

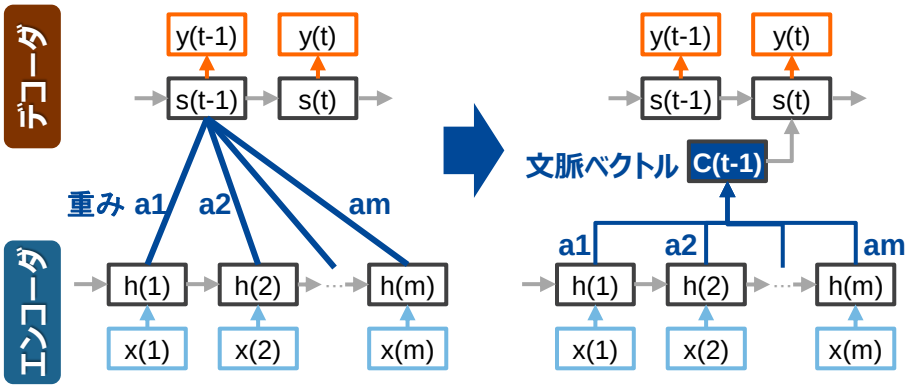
- Attentionは、文章の中でどの単語に注目すべきかを判断する仕組みであり、2014年に発表された
  - 従来のseq2seqは、エンコーダの最後の隠れ状態だけがデコーダに連携されるため、受け渡しできる情報量に限界があり、長い系列ほど精度が落ちた
  - Attentionでは、入力文章を出力文章に変換する際に、入力のどの単語に注目すべきか、あるいは無視すべきかを考慮できるため、出力精度が上がる
  - 遠い所にある単語も含め、入力文章に含まれる全ての単語を参照できるため、RNNやLSTMで課題となっていた長期依存性の問題を補完できる
- Attentionでは、デコーダの今の隠れ状態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

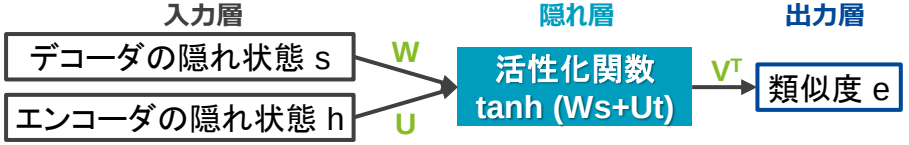
Attentionによる翻訳のイメージ



Attentionの重み計算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



このニューラルネットで各類似度eが計算され、それらにsoftmax関数を適用して合計が1となる重みaを計算する。重み行列W,U,Vはソース文とターゲット文のペアを教師とした学習過程で誤差逆伝播で最適化される。

© 2023 Analytics Design Lab Inc.

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）

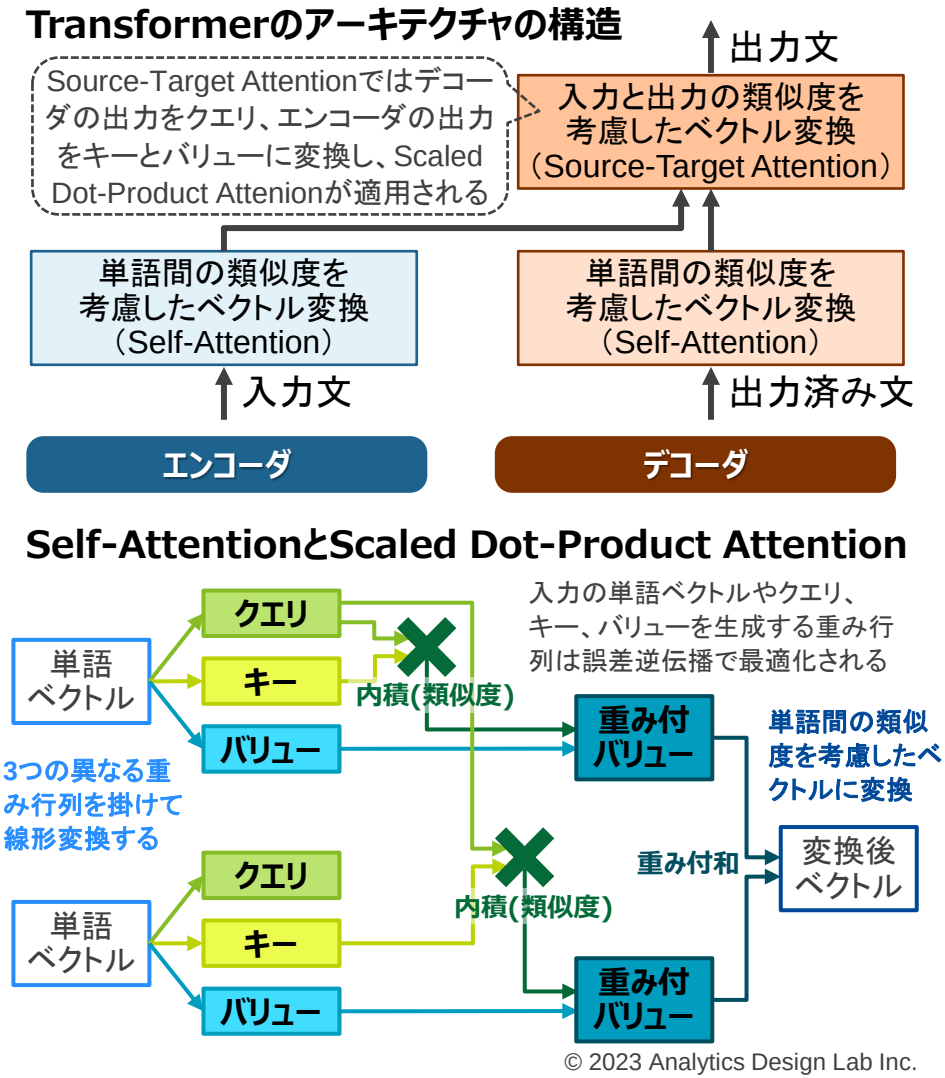
© 2023 Analytics Design Lab Inc.

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



© 2023 Analytics Design Lab Inc.

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

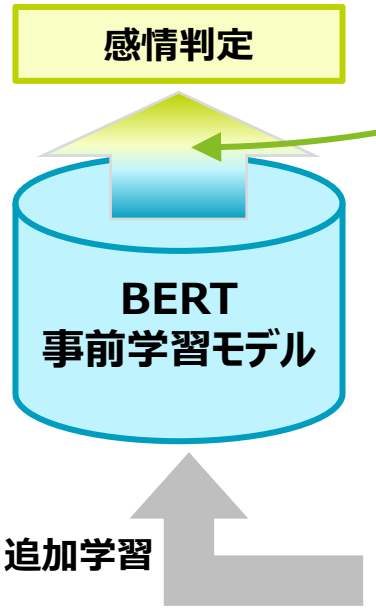
概要

- 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されている
- Transformerのエンコーダ部分に該当するモデルで、文脈を捉えた単語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となり、一部の単語をマスクし、その単語を双方向から予測するタスクMLM（Masked Language Model）を実行することで、文章全体の文脈を捉える能力を獲得する
- 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
- BERT-largeのパラメータ数は3.4億個で学習データは10GB以上となっている

イメージ図

BERTのSelf-Attention

BERTのファインチューニング



学習データ

ID	観光口コミ	感情
1	展望台からの景色は絶景でとてもきれいだった。	ポジ
2	観光客が多すぎてゆっくりできなかった。	ネガ
3	海の幸がどれも新鮮でとても美味しかった。	ポジ
4	海風が気持ちよく楽しく散策できた。	ポジ
5	食事も土産も価格が観光設定で高すぎる。	ネガ

© 2023 Analytics Design Lab Inc.

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

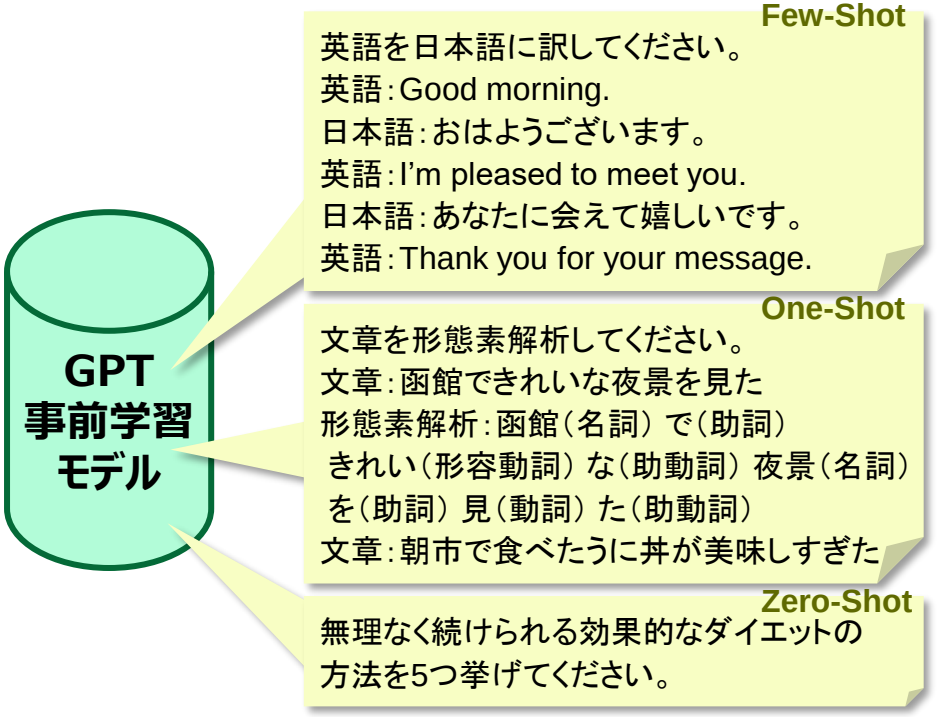
概要

- 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のデコーダ部分に該当するモデルで、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
- 事前学習モデルに、特定のタスクを解く説明や例をプロンプトとして少数与えることで、ファインチューニングなく様々なタスクに対応できる（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
- GPT-3ではパラメータ数は1750億個で学習データは570GBとなっている

イメージ図

GPTのSelf-Attention

GPTのFew-Shot Learningのプロンプト例



© 2023 Analytics Design Lab Inc.

T5 (Text-to-Text Transfer Transformer)

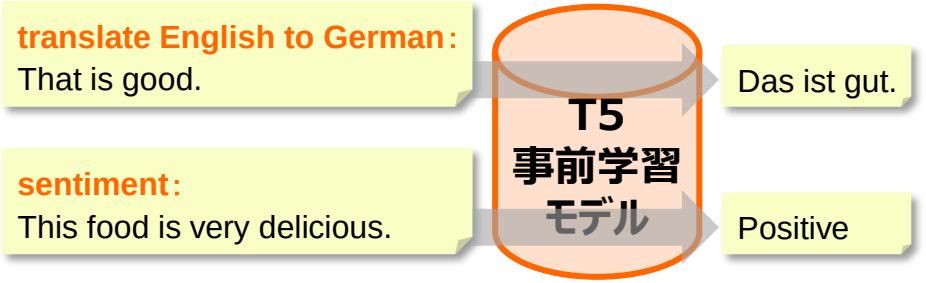
T5は大量のテキストデータから事前学習されたTransformerのエンコーダ-デコーダモデルであり、テキストをテキストに変換する処理で自然言語処理タスク全般に対応できます

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、ノイズを加えた文章を元の正しい文章に復元することで、文脈理解と文章生成において汎化性能も高い能力を獲得する
- 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はプレフィクス付き学習データの作成負荷が大きい
- パラメータ数はT5-Baseで2.2億、T5-11Bで110億

イメージ図

T5によるテキストからテキストへの変換



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による事前学習におけるノイズ除去の処理

エンコーダ

元の文章にノイズを加えた状態でSelf-Attentionを実行

函館山/けど/見た/夜景/は/きれい/とても/で/帰宅/した

誤り　MASK　入れ替え　置換

デコーダ

ノイズを加えた文章全体の特徴

エンコーダから受け渡された文章全体の文脈表現から、ノイズが除去された元の文章を正しく復元する

函館山/から/見た/夜景/は/とても/きれい/で/感動/した

未来の情報を参照しないで左から次の単語を予測する

© 2023 Analytics Design Lab Inc.

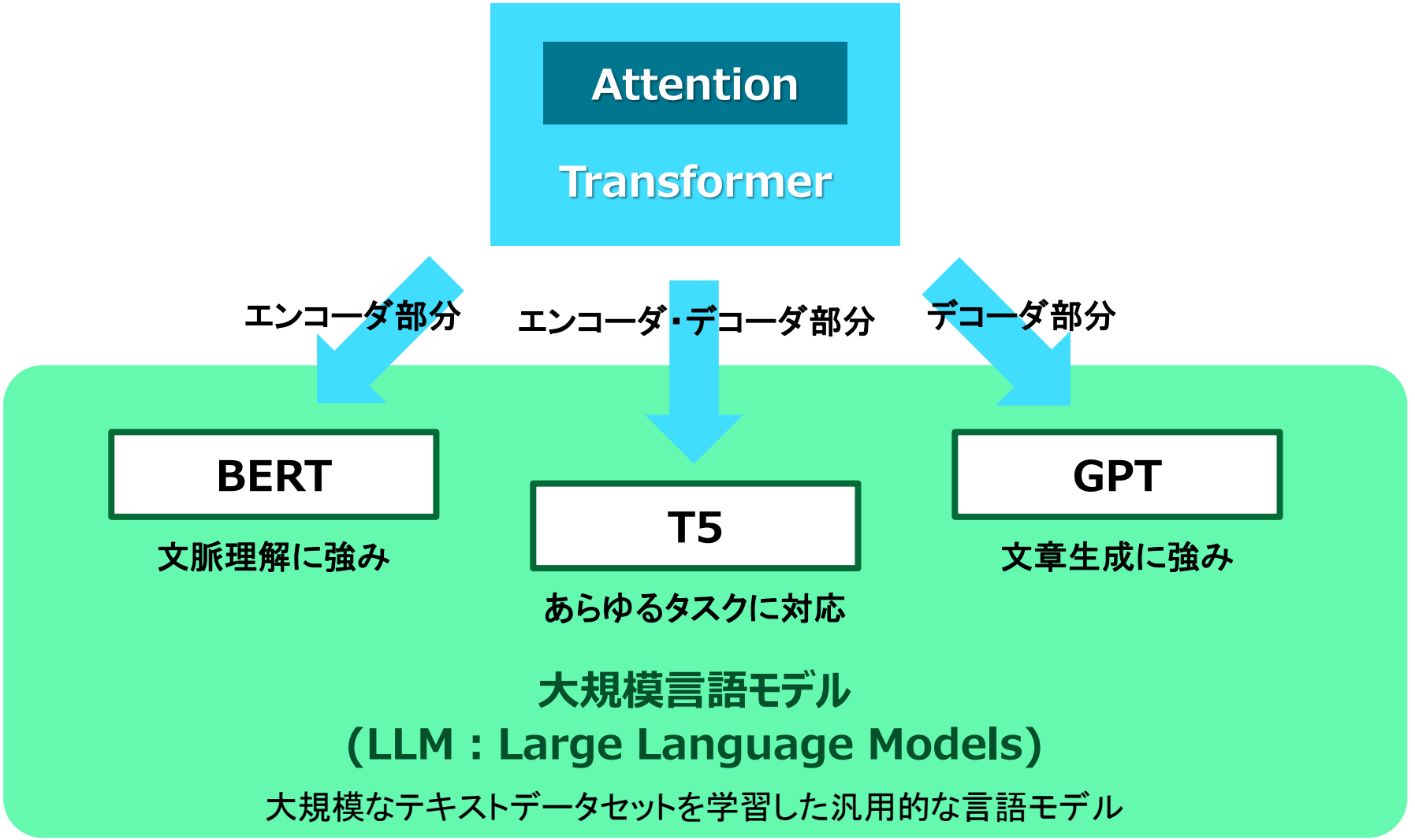
1. テキストマイニングと自然言語処理技術

1-6. テキストマイニングと大規模言語モデル

© 2023 Analytics Design Lab Inc.

大規模言語モデルの位置づけ

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



© 2023 Analytics Design Lab Inc.

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可視化**
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）

© 2023 Analytics Design Lab Inc.

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**

- 分類ラベル付きテキスト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、分析用のテキストデータを分類し、それ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
- 全体のテキストデータに対して大規模言語モデル（GPT）でテーマごとに内容を要約し（特許文書に書かれている用途と技術の内容など）、それぞれのテーマの要約文にテキストマイニングを実行すれば、各テーマ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

© 2023 Analytics Design Lab Inc.

2. 複数のAI技術を応用した新たなテキスト分析手法, Nomolytics

© 2023 Analytics Design Lab Inc.

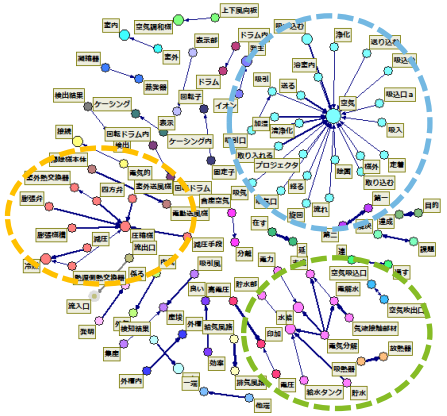
2. 複数のAI技術を応用した新たなテキスト分析手法, Nomolytics

2-1. 従来の特許文書分析とその課題

© 2023 Analytics Design Lab Inc.

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

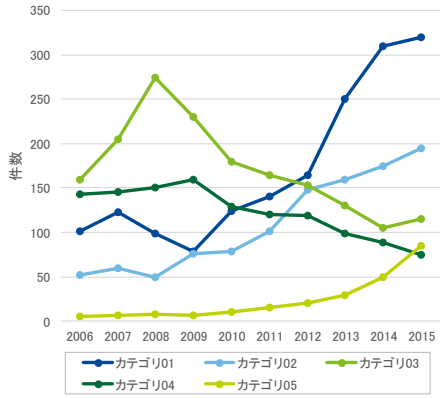
共起ネットワークによる全体像把握



- 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

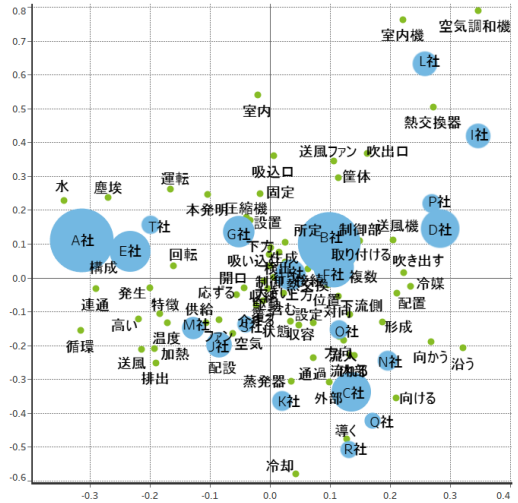
手動作成したカテゴリのトレンド把握

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

コレスポンデ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

- 「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

© 2023 Analytics Design Lab Inc.

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

要因関係をモデリングする
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

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

© 2023 Analytics Design Lab Inc.

2. 複数のAI技術を応用した新たなテキスト分析手法, Nomolytics

2-2. AI技術の応用：PLSAとベイジアンネットワーク

© 2023 Analytics Design Lab Inc.

PLSA（確率的潜在意味解析）

PLSAは、トピックモデルと呼ばれる人工知能技術で、複雑なデータをいくつかの潜在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

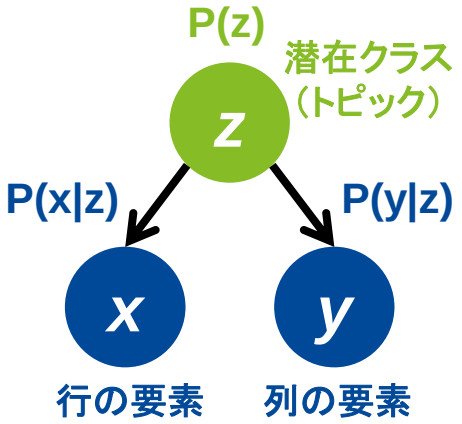
「文書×単語」行列（共起行列）

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

© 2023 Analytics Design Lab Inc.

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

階層型クラスター分析

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

非階層型クラスター分析

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

LSA (Latent Semantic Analysis)

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

PLSA (Probabilistic Latent Semantic Analysis)

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

LDA (Latent Dirichlet Allocation)

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

従来のクラスター分析

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

トピックモデル

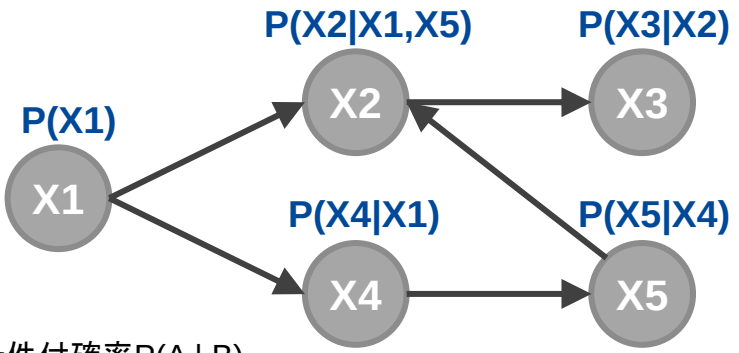
- 単語一つ一つが列の要素となる超高次元のテキストデータを想定した手法
- 要素間の距離の近さで分類するのではなく、高次元データの情報をできるだけ保存した形で低次元に変換する次元圧縮手法であるため、要素数が多い複雑なデータにも対応できる
- 行の要素と列の要素の背後にある共通する特徴をクラスとして抽出するため、行と列の両方をクラスタリングでき、クラスの持つ情報が多い
- 一つの要素は全てのクラスに所属するソフトクラスタリングで、その所属の重みを計算するため、データが複数の特徴をまたがる場合でも表現できる

© 2023 Analytics Design Lab Inc.

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

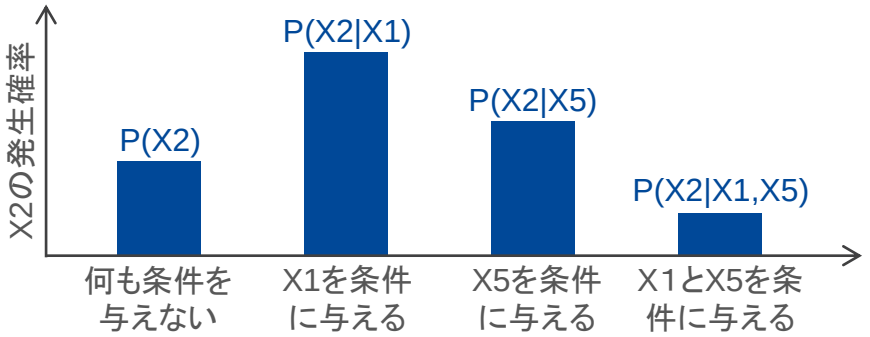
ベイジアンネットワークの概要

- 複数の変数の確率的な因果関係をネットワーク構造で表わし、ある変数の状態を条件として与えたときの他の変数の条件付確率を推論することができる
- 目的変数と説明変数の区別はなく、様々な方向から変数の確率シミュレーションができる
- 全ての変数は質的変数（カテゴリカル変数）となるため、量的変数の場合は閾値を設けてカテゴリに分割する
- 確率論の非線形処理によるモデル化のため、非線形の関係や交互作用が生じる現象でも記述できる



※条件付確率P(A | B)
事象Bが起こる条件の下で事象Aの起こる確率

確率的因果関係と交互作用



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

ベイジアンネットワークのメリット

現象を理解して柔軟にシミュレーションできる

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

効果を発揮する有用な条件を発見できる

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

© 2023 Analytics Design Lab Inc.

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

ニューラルネットワーク（ディープラーニング）

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

回帰分析・判別分析（数量化Ⅰ類・Ⅱ類）

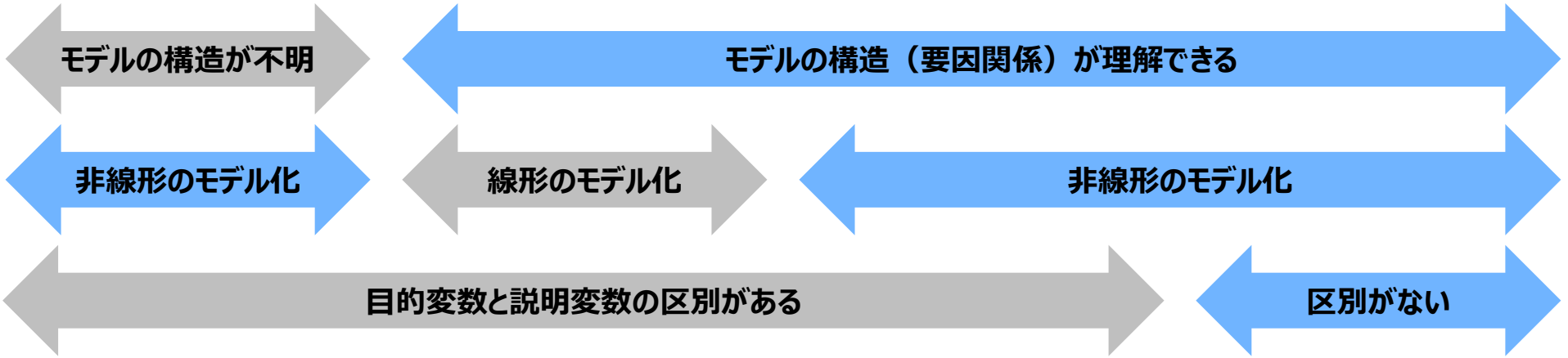
- 目的変数を説明変数の1次結合で定式化する
- 目的変数と説明変数の間に線形関係があるという仮定に基づいている
- 各説明変数の影響は独立しており、複合的な交互作用の影響は表現できない
- 説明変数間で相関が高い場合は解が不安定となり（多重共線性）、変数が多い場合この解消検討の負荷が大きい

決定木

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

ベイジアンネットワーク

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



© 2023 Analytics Design Lab Inc.

AI技術の分類

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

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

- GAN
- GPT
- Diffusion model

※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

© 2023 Analytics Design Lab Inc.

2. 複数のAI技術を応用した新たなテキスト分析手法, Nomolytics

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

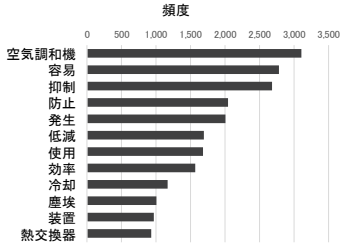
© 2023 Analytics Design Lab Inc.

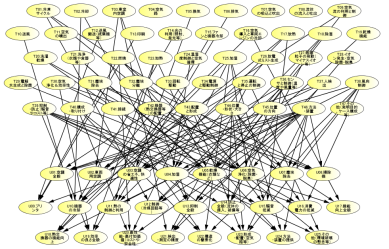
膨大なテキストデータをトピックに変換して解釈を容易にし、テキスト情報内に潜む要因関係をモデル化して、ビジネスアクションに有用な特徴を把握可能にします

Nomolytics：Narrative Orchestration Modeling Analytics

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

単語抽出 × トピック類型化 × 要因関係分析





Nomolyticsのメリット

膨大なテキストデータをいくつかのトピックという人間が理解しやすい形に整理し類型化できる

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

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

© 2023 Analytics Design Lab Inc.

書籍のご紹介

7月にベイジアンネットワークの実践的な解説本がオーム社より出版され、私も一部の章を担当し、テキストデータに適用するベイジアンネットワークについて執筆しました

- 定価：3,740円 （本体3,400円＋税10%）
- 判型：B5変　　頁：352頁
- ISBN：978-4-274-23048-6
- 発売日：2023年7月18日
- 出版社：株式会社オーム社
- https://www.ohmsha.co.jp/book/9784274230486/
- 目次

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

3-1.「風・空気」に関する特許文書データ

© 2023 Analytics Design Lab Inc.

「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

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

3-2. 分析プロセスの全体像

© 2023 Analytics Design Lab Inc.

Nomolyticsを適用した特許分析のプロセス

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

トピックの抽出

特許の要約文にテキストマイニングを実行して単語や係り受け表現を抽出

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

PLSAの適用

【課題】の要約文から用途トピックを抽出

【解決手段】の要約文から技術トピックを抽出

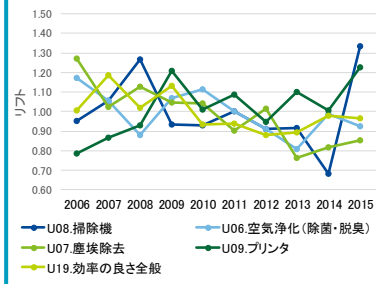
トピックのスコアリング

全特許データに対する各トピックの該当度を計算

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

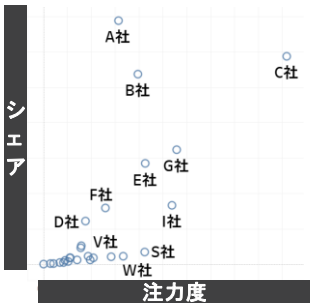
トレンドの分析

トピックのスコアを出願年で集計することでトピックのトレンドを把握



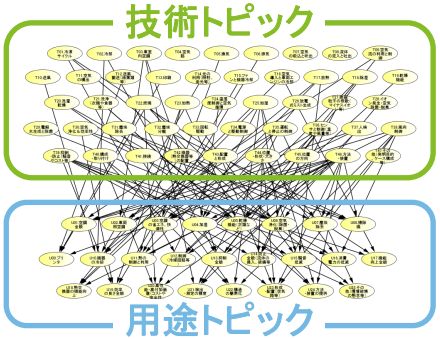
競合他社の分析

トピックのスコアを出願人で集計することで各社の特徴やポジションを把握



用途と技術の関係分析

用途トピックと技術トピックの関係性をベイジアンネットワークでモデル化



用途⇒技術 の分析

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

技術⇒用途 の分析

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

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

3-3. トピックの抽出

© 2023 Analytics Design Lab Inc.

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 → 共起行列の作成 → PLSAの実行 → トピックの抽出

【課題】と【解決手段】の文章に含まれる単語と係り受けを抽出する

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

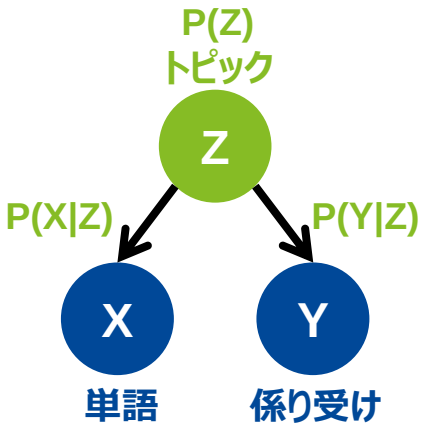
係り受け表現	頻度
空気調和機⇒提供	1,575
効率⇒良い	1,325
車両用空調装置⇒提供	578
掃除機-提供	545
容易-構成	539
画像形成装置-提供	334
抑制-提供	296
向上-図る	279
・・・	…

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

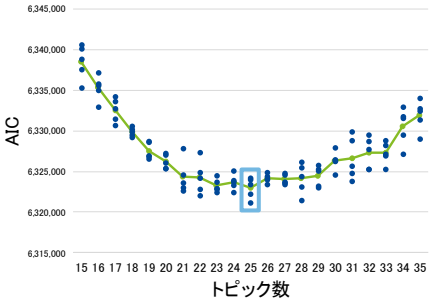
単語 ＼ 係り受け表現	空気調和機⇓提供	効率⇓良い	車両用空調装置⇓提供	掃除機⇓提供	…
空気調和機	1578	100	4	1	
空気	85	144	45	50	
容易	100	105	51	67	
抑制	142	95	64	63	
・・・					

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

共起行列にPLSAを適用する



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



各トピックについて以下の3つの確率が計算される

① P(X|Z)
トピックにおける単語の所属確率

② P(Y|Z)
トピックにおける係り受けの所属確率

③ P(Z)
トピックの存在確率

トピックにおけるP(X|Z)とP(Y|Z)からトピックの意味を解釈する

トピック T32
P(Z) = 2.7%

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

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

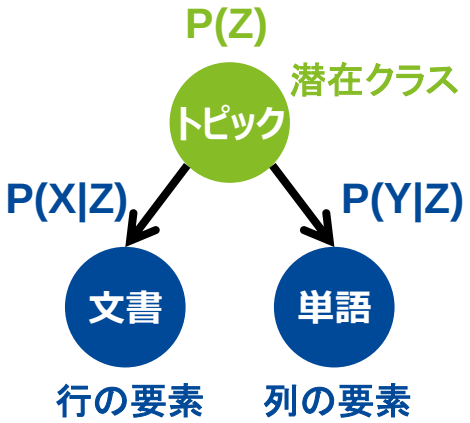
確率の高い構成要素から、トピックT32は「塵埃の分離」に関するトピックと解釈できる

© 2023 Analytics Design Lab Inc.

共起行列の構成の工夫

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

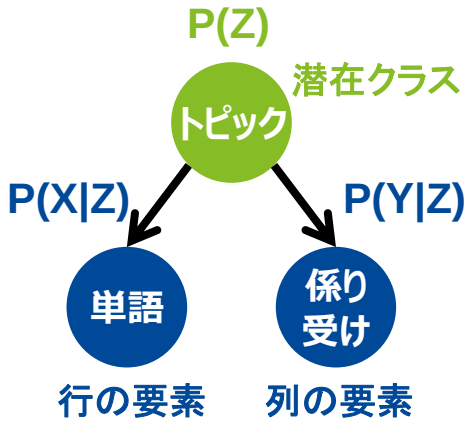
一般的なPLSAの共起行列



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

© 2023 Analytics Design Lab Inc.

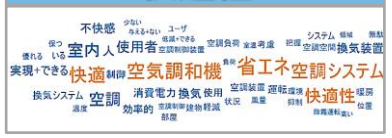
用途トピック25個の一覧

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

U01.空調全般

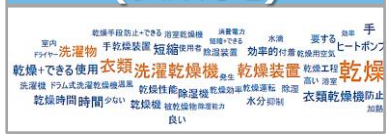
U02.車両用空調

U03.空調の省エネ、快適性



U04加湿

U05.乾燥機能(衣類など)



U06.空気浄化(除菌・消臭)

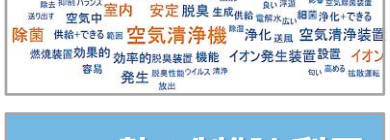
U07.塵埃除去

U08.掃除機

U09.プリンタ

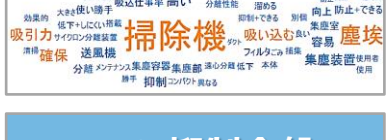
U10.機器の冷却

U11.熱の制御と利用

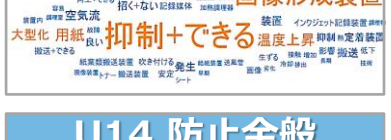


U12.制御(冷媒回路等)

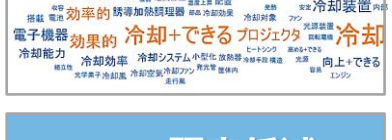
U13.抑制全般



U14.防止全般(流体の侵入、破損等)

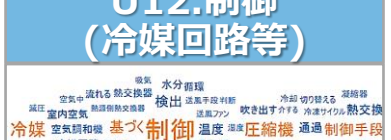


U15.騒音低減



U16.消費電力の低減

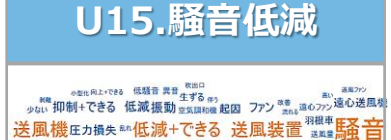
U17.機能向上全般



U18.熱交換器の機能向上

U19.効率の良さ全般

U20.価値(コストや安全性など)



U21.検出・測定の精度

U22.構造の簡素化

U23.形成・配置(空気路等)

U24.方法・装置の提供

U25.その他(環境破壊の懸念等)

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

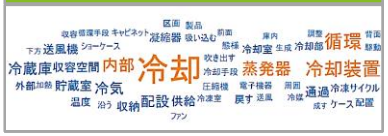
© 2023 Analytics Design Lab Inc.

技術トピック47個の一覧①

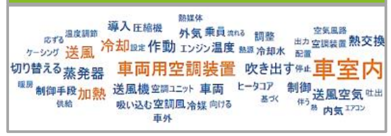
【解決手段】の文章からは、空気の冷却や空気路、換気、放熱、除湿、乾燥、加湿、イオン生成、空気清浄、塵埃分離、センサと制御、構成や配置などの技術が47個抽出されました

T01.冷凍サイクル

T02.冷却

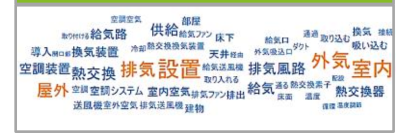


T03.車室内空調

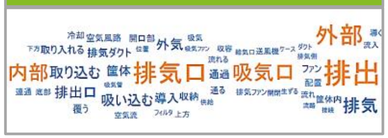


T04.空気路

T05.換気



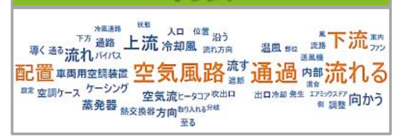
T06.排気



T07.空気の吸込と吹出

T08.流体の流入と吐出

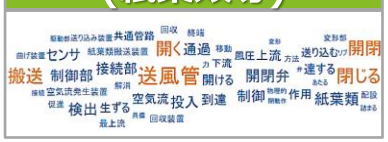
T09.空気流の利用と制御



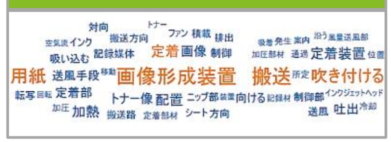
T10.送風

T11.空気の噴出

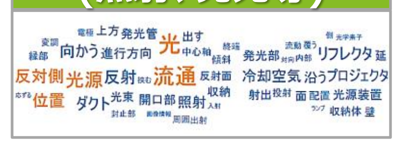
T12.送風搬送(紙葉類等)



T13.印刷

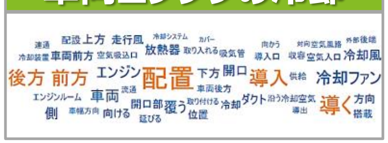


T14.光の利用(照射、発光等)

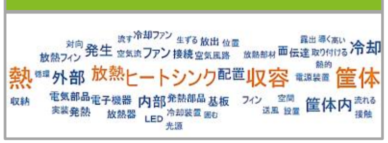


T15.ファンと機器冷却

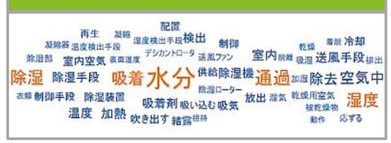
T16.空気導入と車両エンジンの冷却



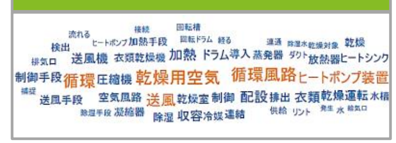
T17.放熱



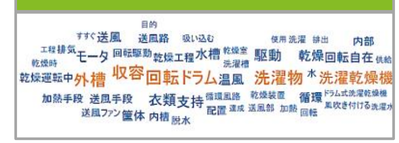
T18.除湿



T19.乾燥機能



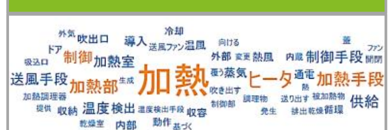
T20.洗濯乾燥



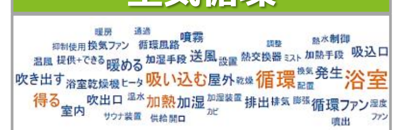
T21.洗浄(衣類や食器等)

T22.燃焼

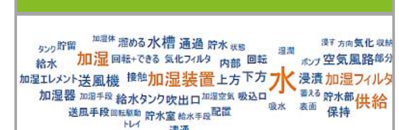
T23.加熱



T24.温湿度制御と空気循環



T25.加湿



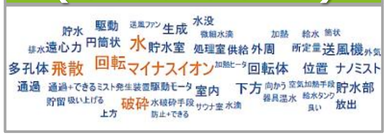
※文字の大きさはトピックに対する関係の強さを表現している
（関係の強い上位5つの単語を赤色で表示している）

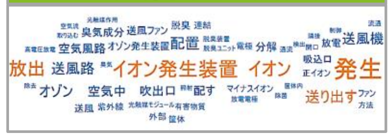
© 2023 Analytics Design Lab Inc.

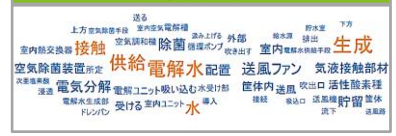
技術トピック47個の一覧②

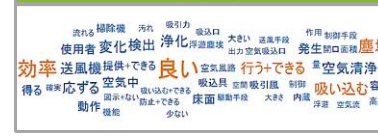
【解決手段】の文章からは、空気の冷却や空気路、換気、放熱、除湿、乾燥、加湿、イオン生成、空気清浄、塵埃分離、センサと制御、構成や配置などの技術が47個抽出されました

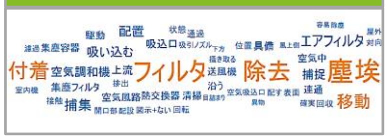
T26.放電式ミスト生成


T27.微細粒子の飛散(マイナスイオン等)


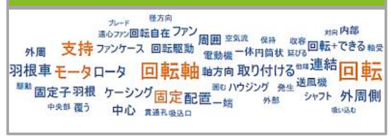
T28.イオン発生・空気除菌・脱臭


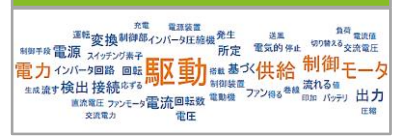
T29.電解水生成と除菌


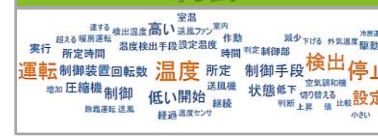
T30.塵埃吸込＆効率性


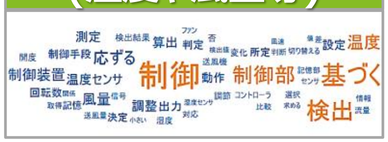
T31.塵埃除去


T32.塵埃分離

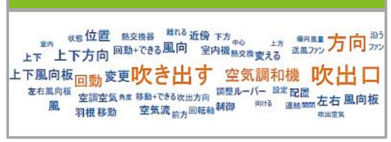

T33.回転駆動


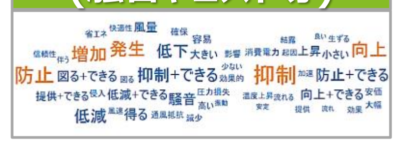
T34.電源と駆動制御


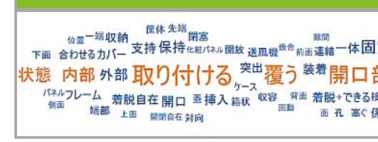
T35.運転と停止の制御


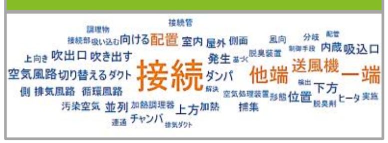
T36.センサと制御(温度や風量等)


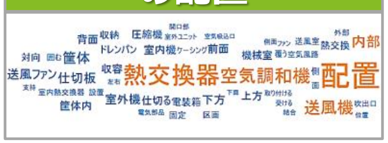
T37.人検出

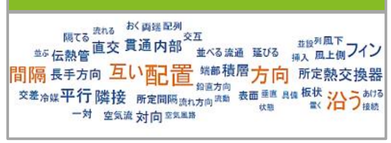

T38.風向制御


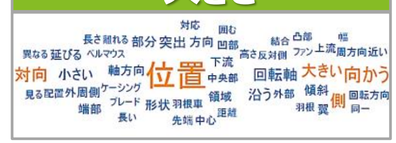
T39.抑制・防止(騒音やコスト等)


T40.構成・取り付け


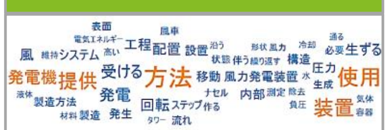
T41.接続


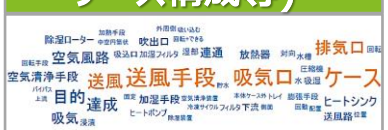
T42.機器(熱交換等)の配置


T43.配置と形成


T44.位置・形状・大きさ


T45.位置の方向


T46.方法・装置


T47.その他(発明目的、ケース構成等)


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

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

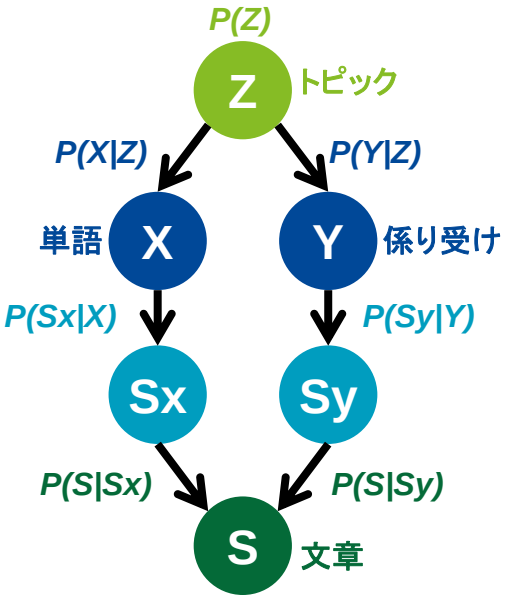
3-4. トピックのスコアリング

© 2023 Analytics Design Lab Inc.

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

文章単位のスコア	$\frac{P(S\|Z)}{P(Z)}$

- リフト値（事後確率÷事前確率）
- トピックを条件とすることで文章の発生確率が何倍になるのかを示す



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

単語X_iで定義される文章Sx_h

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

トピックZ_kを条件とした文章Sx_hの出現確率

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

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

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

係り受けY_jで定義される文章Sy_h

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

トピックZ_kを条件とした文章Sy_hの出現確率

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

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

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

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

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

文章S_hの出現確率

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

トピックスコア算出プロセス

①文章ごとにスコアを計算

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

②特許IDごとに文章スコアを集約

※最大値を採用する

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

③閾値を設定してフラグに変換する

※閾値は3に設定する

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

© 2023 Analytics Design Lab Inc.

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た	通風路を通して回転槽	0	0		1	1	0		0
4	特願2013-XXXX	2013	D社	ウインドシールドの防	車両用空調装置の空	0	1		0	0	1		1
・・・	・・・	・・・	・・・			…	…		…	…	…		…
30039	特願2012-XXXX	2012	Z社	プリ空調時に、除菌ま	冷暖房空調ユニットは	0	1		0	1	1		0

①出願年の分析

②出願人の分析

③用途と技術の関連性の分析

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

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

3-5. 出願年×トピックによるトレンド分析

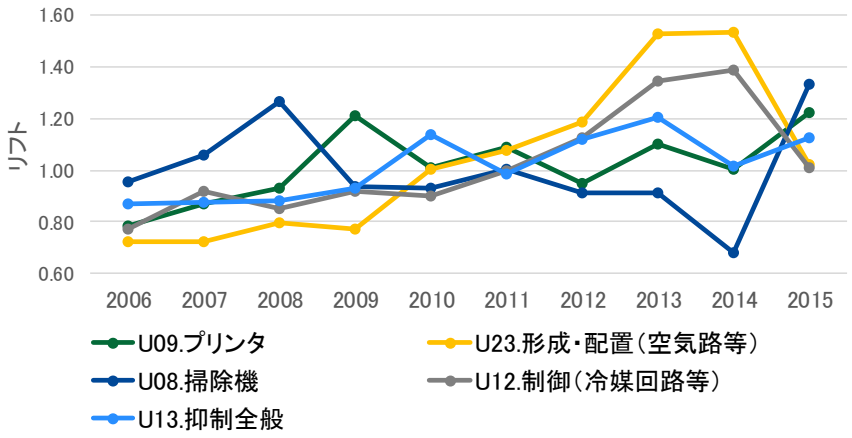
【分析目的】
技術や用途のトレンドを把握し、有望なシーズやニーズを探り、今後の技術開発戦略を検討する

© 2023 Analytics Design Lab Inc.

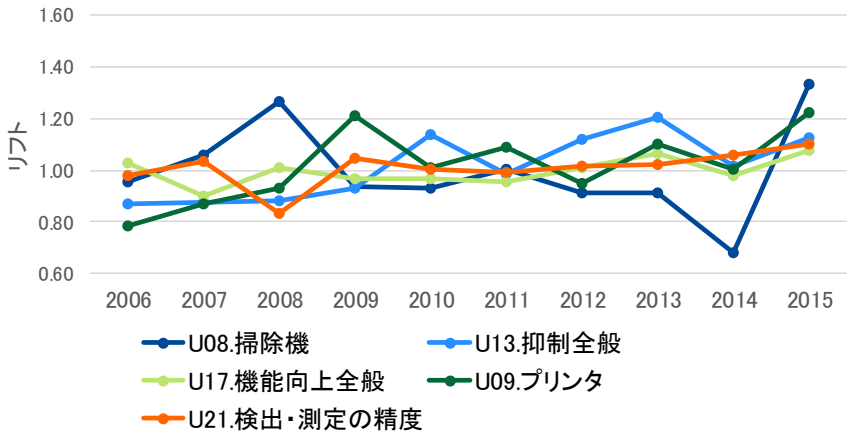
用途トピックの上昇トレンド

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

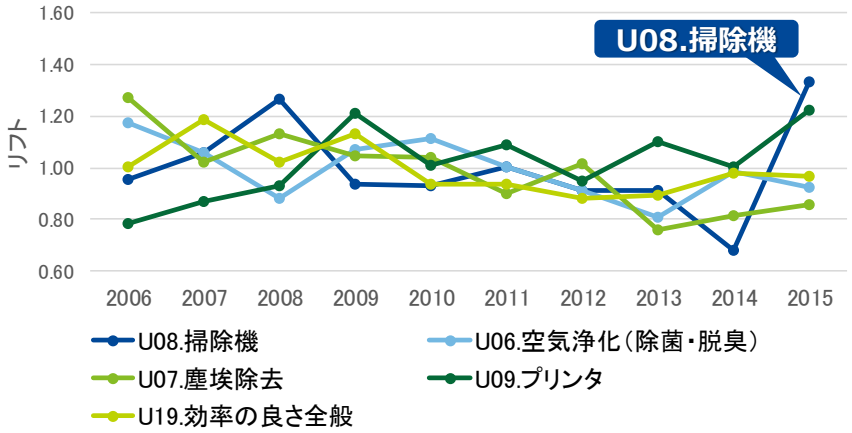
【長期】2006年からの上昇率 best5



【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

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

■リフト値は出願年とトピックの関係を示す指標

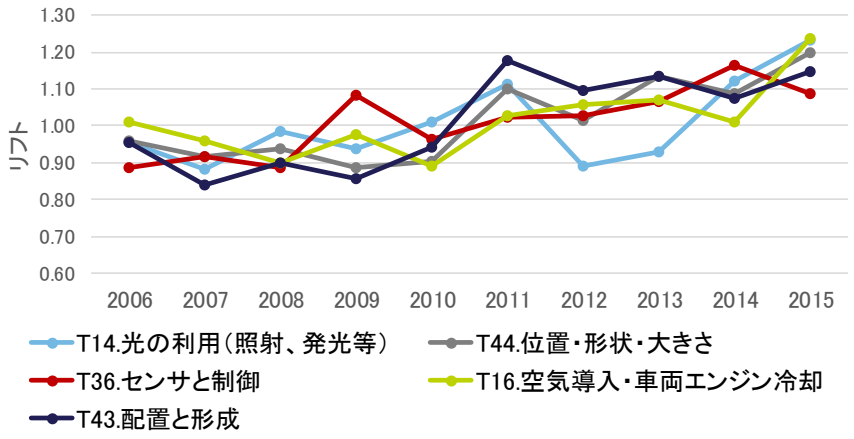
■トピック毎の各出願年の出願割合を、その出願年の出願割合で正規化した値

© 2023 Analytics Design Lab Inc.

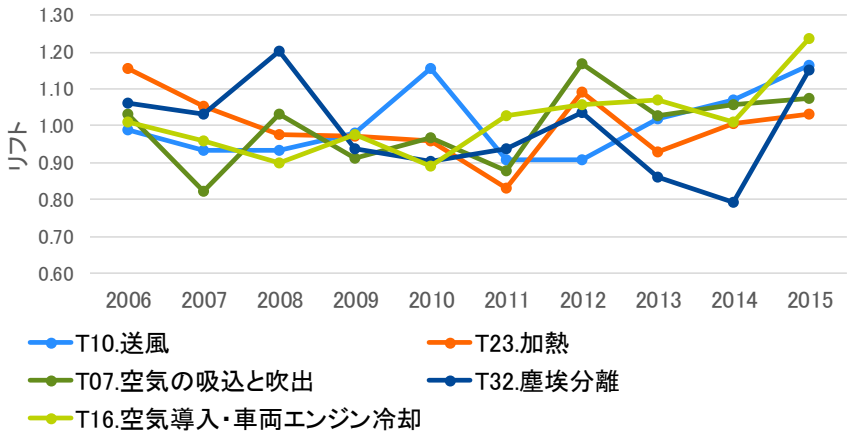
技術トピックの上昇トレンド

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

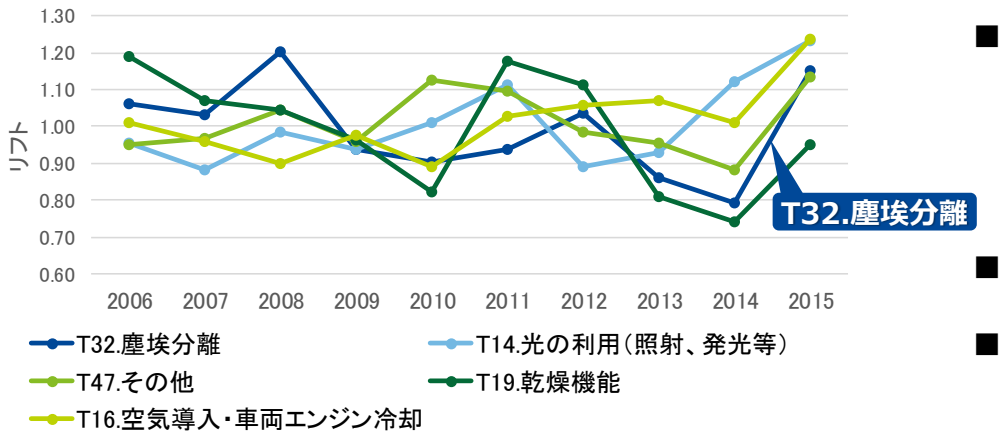
【長期】2006年からの上昇率 best5



【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

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

■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

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

3-6. 出願人×トピックによる競合分析

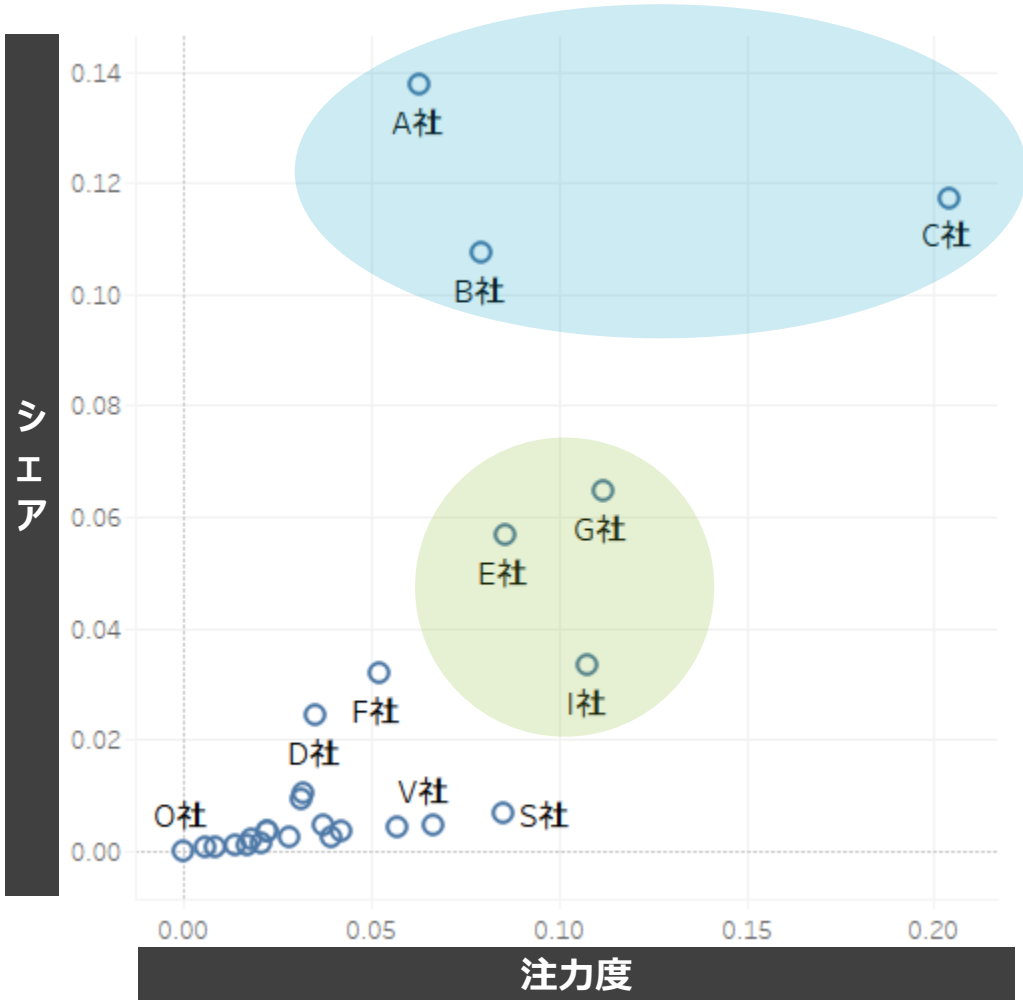
【分析目的】
各出願人の動向や、業界における棲み分け、競合他社と自社との関係性などを把握し、他社との差別化戦略や協業戦略を検討する

© 2023 Analytics Design Lab Inc.

技術「T32.塵埃分離」の出願人のポジショニング

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

注力度とシェアの散布図



考察と戦略の検討

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

注力度とシェア

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

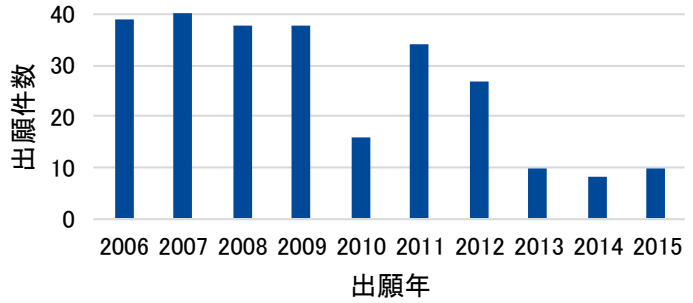
© 2023 Analytics Design Lab Inc.

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

A社

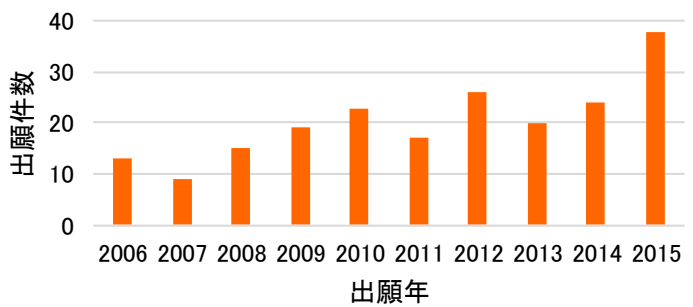


シェア1位のA社は、近年は出願が少ない

G社

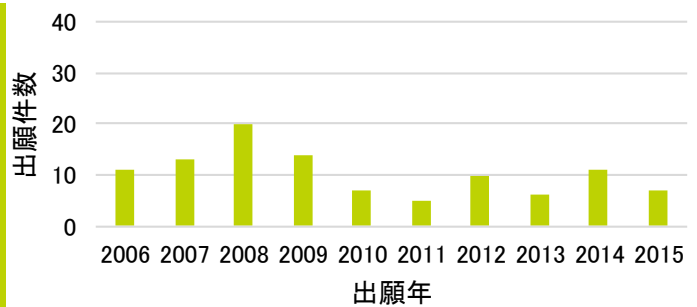
シェア4位注力度2位のG社は、近年徐々に出願が増えている

B社



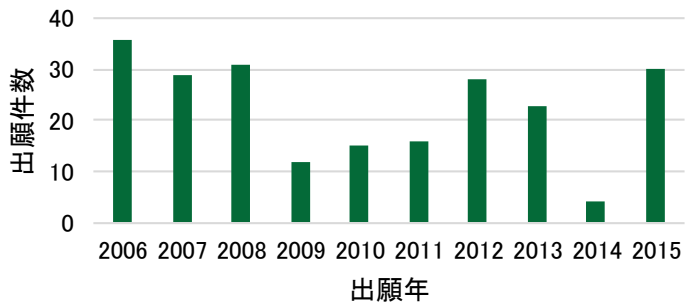
シェア3位のB社は、ここ10年で徐々に出願が増えている

E社



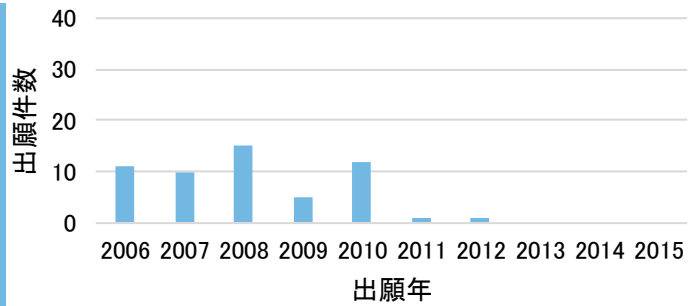
シェア5位注力度4位のE社は、近年では出願件数の変化が少ない

C社



注力度1位シェア2位のC社は、直近で出願が急激に増えている

I社



シェア6位注力度3位のI社は、2013年以降の出願が存在しない

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

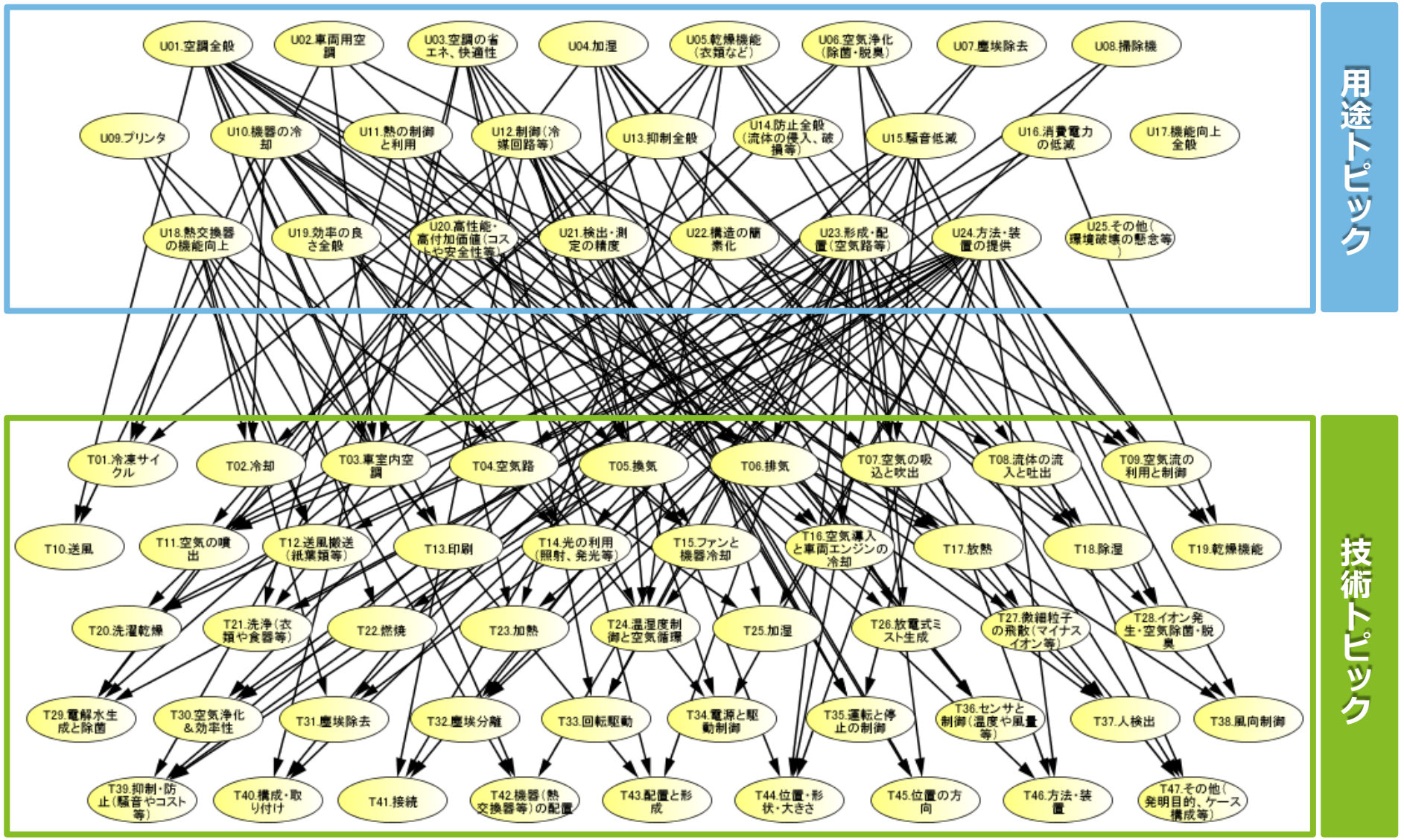
3-7. 用途×技術の関係分析＜その1＞
～用途⇒技術の関係～

【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

© 2023 Analytics Design Lab Inc.

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

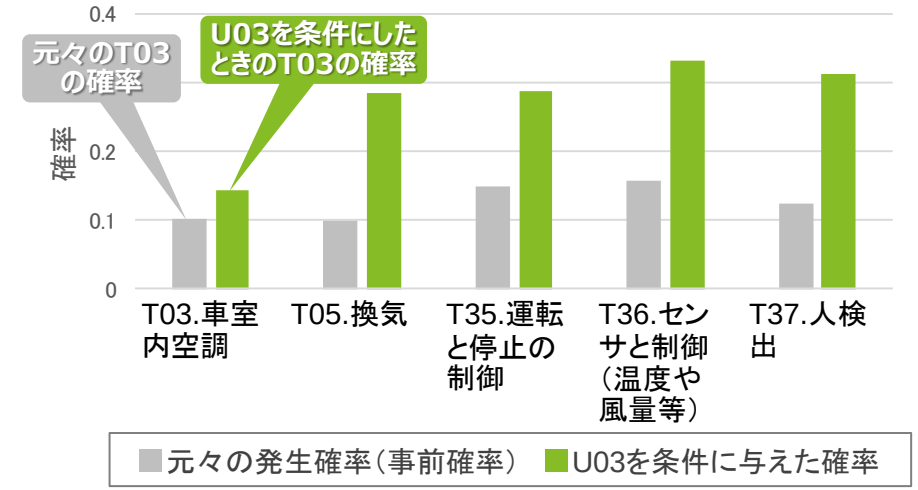


© 2023 Analytics Design Lab Inc.

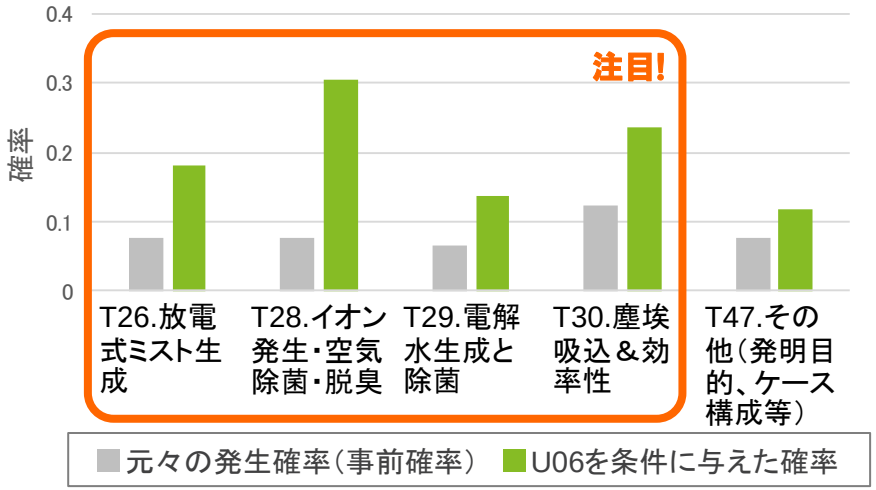
用途と関係のある技術の確認

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

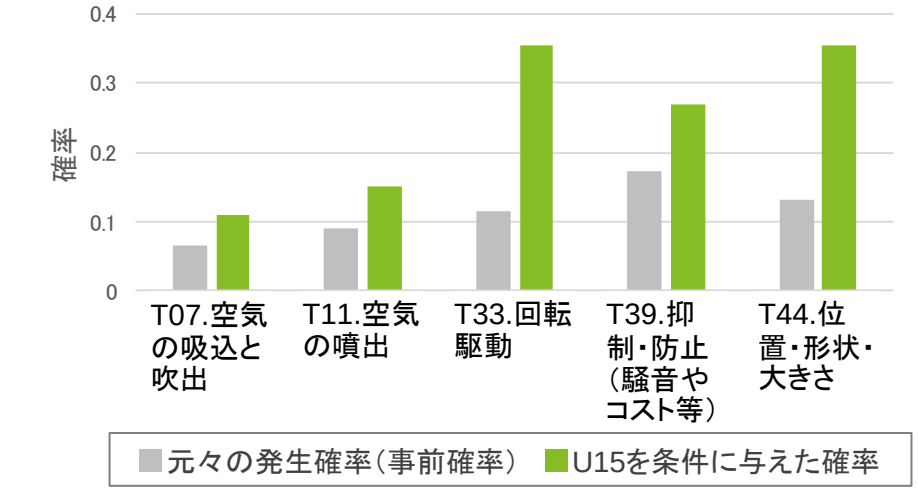
「U03.空調の省エネ、快適性」と関係のある技術



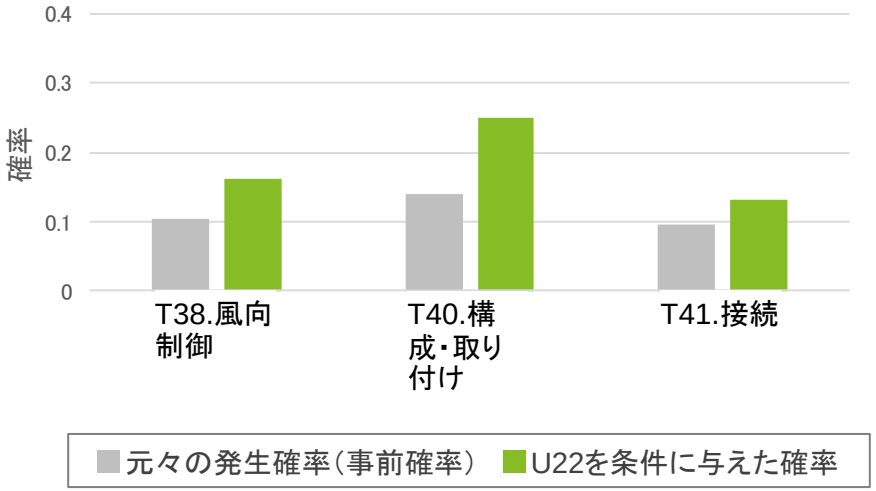
「U06.空気浄化(除菌・脱臭)」と関係のある技術



「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

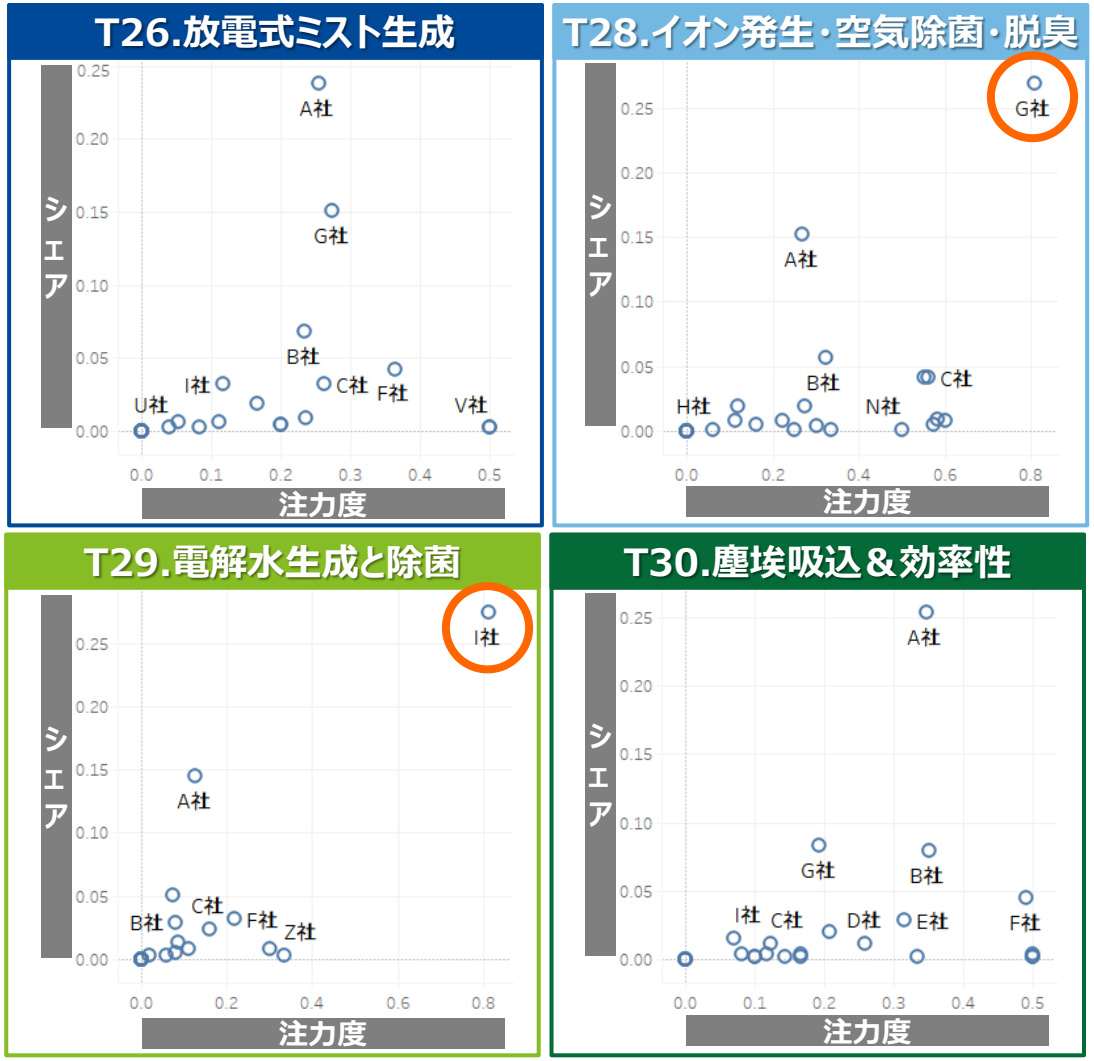


© 2023 Analytics Design Lab Inc.

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社などが高いが、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社など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

© 2023 Analytics Design Lab Inc.

3. Nomolyticsを適用した特許分析事例

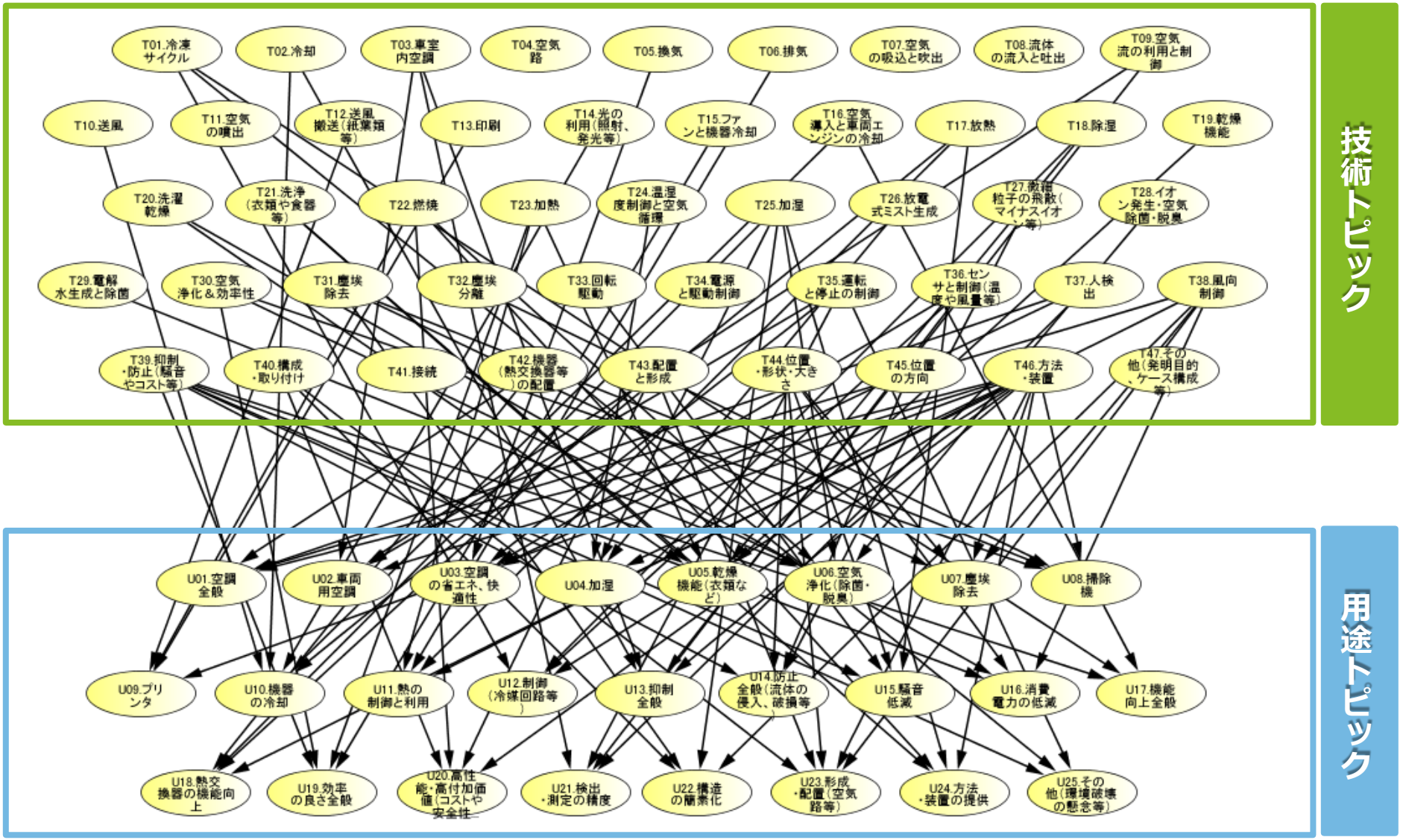
3-8. 用途×技術の関係分析＜その2＞
～技術⇒用途の関係～

【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

© 2023 Analytics Design Lab Inc.

技術⇒用途の関係モデル

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

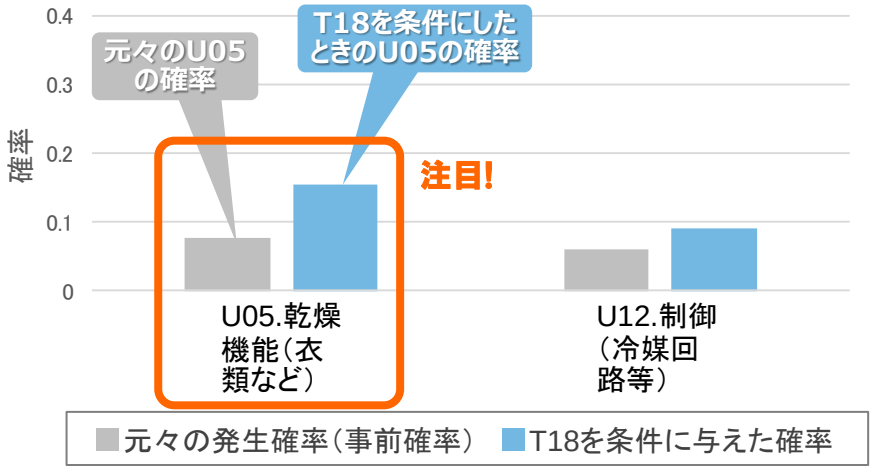


© 2023 Analytics Design Lab Inc.

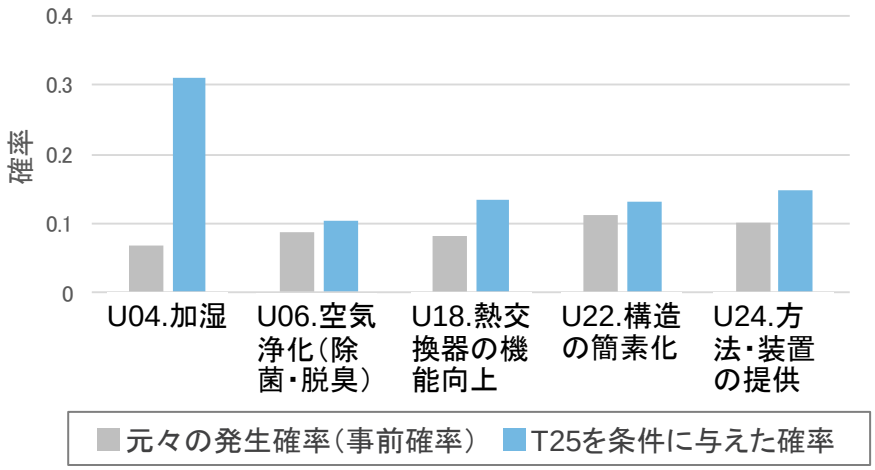
技術と関係のある用途の確認

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

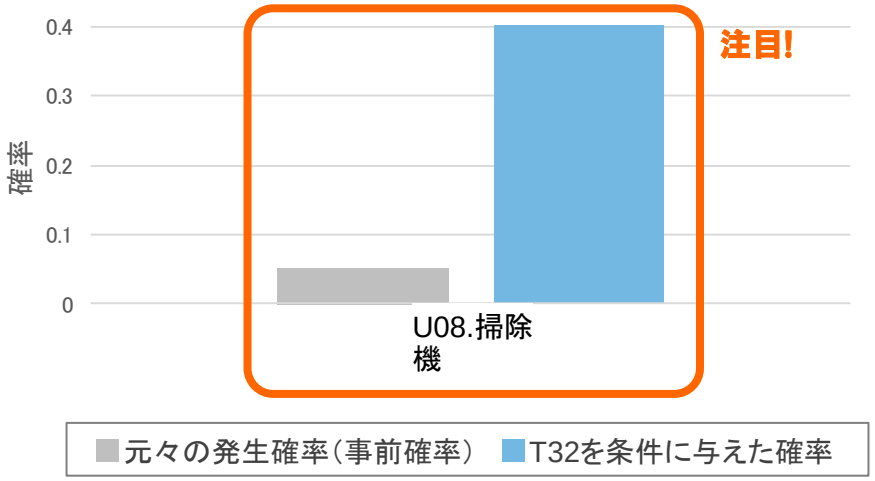
「T18.除湿」と関係のある用途



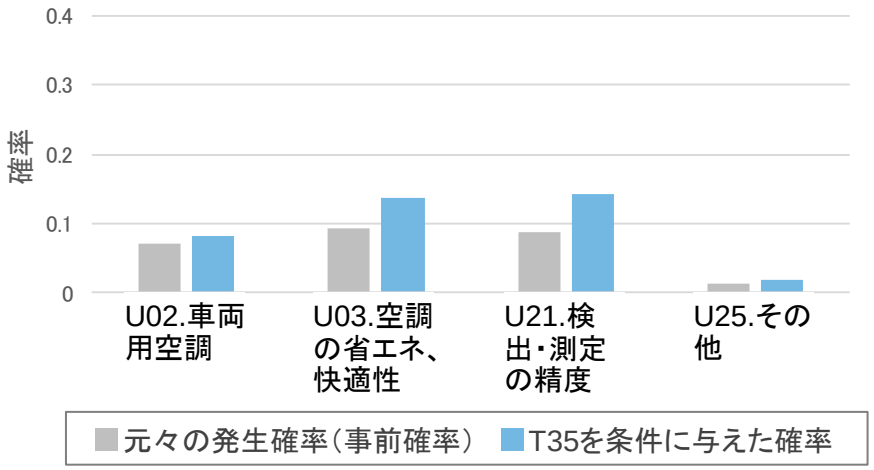
「T25.加湿」と関係のある用途



「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



© 2023 Analytics Design Lab Inc.

「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

© 2023 Analytics Design Lab Inc.

「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】

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

「掃除機」を想定していない「塵埃分離」の特許例

発明の名称

画像形成装置

【課題】

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

【解決手段】

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

※対外説明用のため要約文は一部加工している

© 2023 Analytics Design Lab Inc.

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

© 2023 Analytics Design Lab Inc.

4. Nomolyticsによる特許分析のまとめ

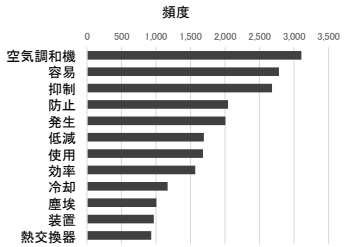
© 2023 Analytics Design Lab Inc.

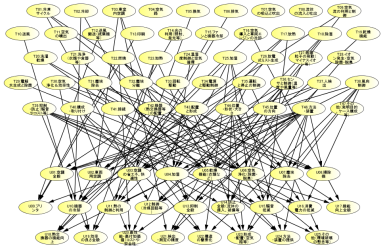
膨大なテキストデータをトピックに変換して解釈を容易にし、テキスト情報内に潜む要因関係をモデル化して、ビジネスアクションに有用な特徴を把握可能にします

Nomolytics：Narrative Orchestration Modeling Analytics

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

単語抽出 × トピック類型化 × 要因関係分析





Nomolyticsのメリット

膨大なテキストデータをいくつかのトピックという人間が理解しやすい形に整理し類型化できる

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

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

© 2023 Analytics Design Lab Inc.

特許文書データへの適用

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

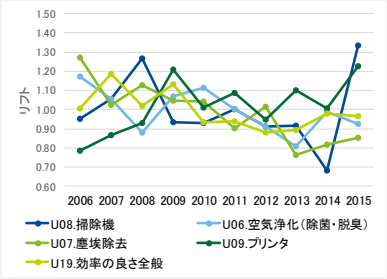
A 特許文書のトピック化

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×トピックモデルPLSAを適用することで、課題からは用途に関するトピックを、解決手段からは技術に関するトピックを機械的に抽出し、大量の特許の全体像を把握する

用途と技術の関係分析

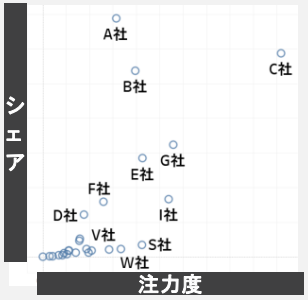
B トレンドの分析

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する



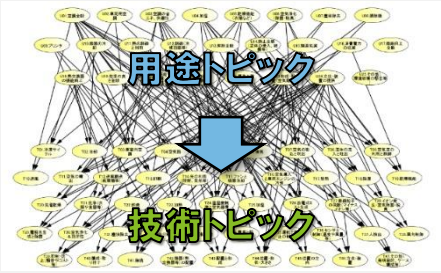
C 競合他社の分析

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する



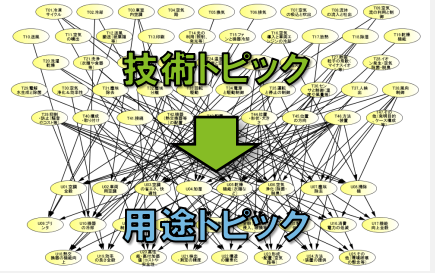
D 用途⇒技術の関係分析

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る



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する



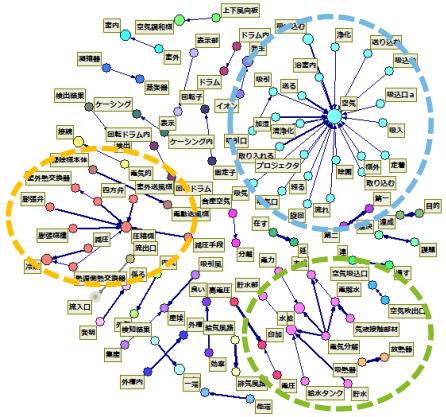
© 2023 Analytics Design Lab Inc.

Nomolyticsを適用した特許分析のメリット①

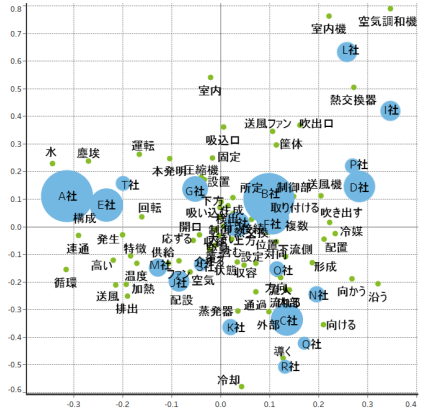
単語ではなく集約されたトピックをベースにした分析を実行することで、膨大な特許情報に潜む特徴を分かりやすく理解することができます

従来の特許分析

単語の共起ネットワーク



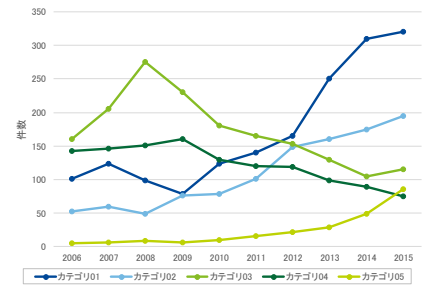
単語と出願人の対応マップ



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->収容
吸引力	塵埃->遠心分離
サイクロン	含塵空気->分離

カテゴリ別のトレンド

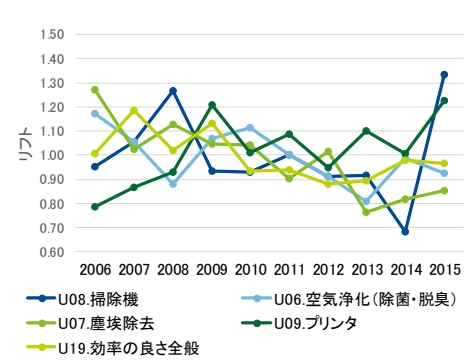


- 単語ベースの複雑なアウトプットから、文章全体に存在する話題を解釈したり、各出願人の特徴を把握しなければならない
- 単語を人手でグルーピングしていくつかのカテゴリを作成し、カテゴリベースに分析するものの、そのカテゴリ作成は属人的で作業負荷も大きい

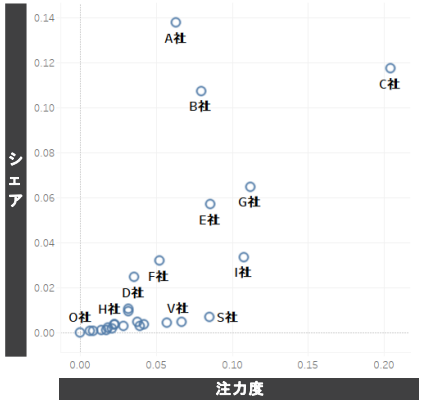
Nomolyticsを適用した特許分析

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

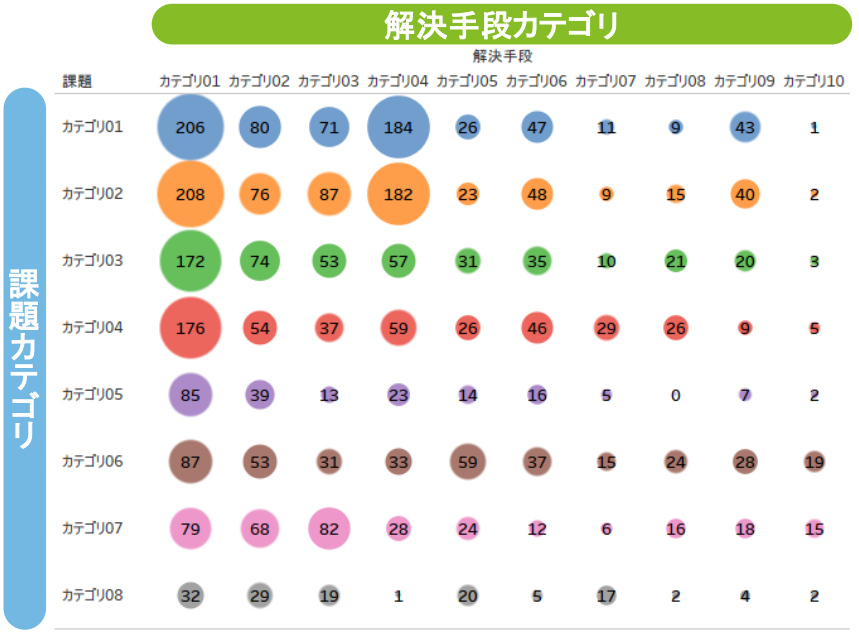
© 2023 Analytics Design Lab Inc.

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

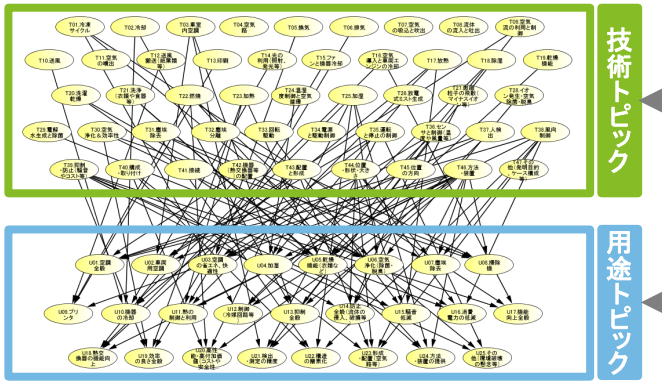
課題と解決手段のカテゴリ間のクロス集計



- 【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）

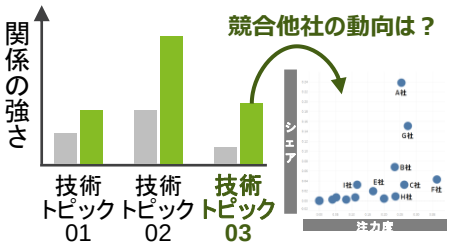
Nomolyticsを適用した特許分析

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

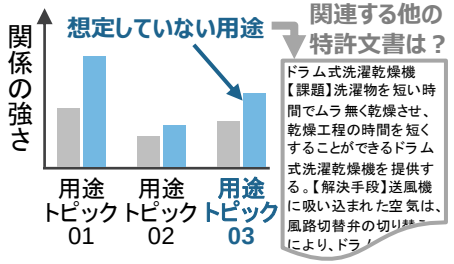


分析目的に応じて因果構造を入れ替える
①「技術⇒用途」
②「用途⇒技術」

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、関係の強い重要技術を確認し、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

© 2023 Analytics Design Lab Inc.

Nomolytics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トピックを把握
- 製品・サービス別のトピックを把握
- 口コミ得点に寄与するトピックを把握
- ニーズに応じたマーケティングを検討

アンケート

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

コールセンター履歴

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

特許文書

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

営業日報

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

有価証券報告書

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

エントリーシート

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

診療・看護記録

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

問題発生レポート

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

© 2023 Analytics Design Lab Inc.

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にお願いします。

株式会社アナリティクスデザインラボ

Analytics Design Lab

© 2023 Analytics Design Lab Inc.