



Analytics Design Lab

特許文書データに人工知能技術を応用した 競合分析と技術の新規用途探索

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

2017年11月29日

人工知能技術を応用したデータ分析の研究開発とビジネスコンサルティングの経験を活かし、2017年6月にデータ活用コンサルティングの新会社を設立しました

株式会社アナリティクスデザインラボ

企業におけるデータ活用を支援する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、というデータ活用するプロセスを企業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の課題解決を支援します。

設立	2017年6月1日
事業内容	<ul style="list-style-type: none">● 企業におけるデータ活用のコンサルティング● データ分析技術の研究開発
資本金	500万円
所在地	東京都中野区東中野1-58-8-204

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
- 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

人工知能技術を応用した特許分析の新たな展開

複数の人工知能技術を組み合わせることで、特許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

単語ベースの分析では
複雑で考察しにくい

カテゴリの設定が主観的で
作業負荷も大きい

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

文脈を考慮した潜在的なトピック
(単語の集合)を抽出する

ベイジアンネットワーク

多様な要因間の確率統計的な
因果関係をモデル化する

PLSAは、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です

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することで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

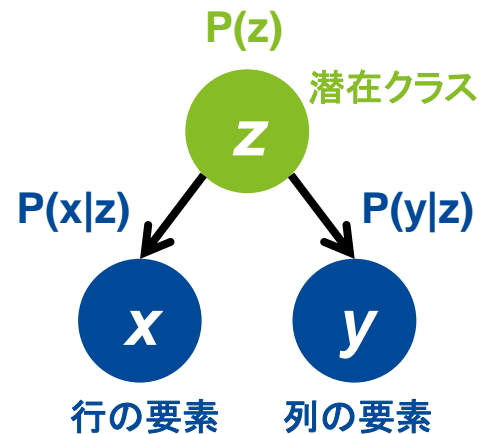
文書ID	単語 1	単語 2	単語 3	...	単語 5,014	単語 5,015
1	0	0	1		1	0
2	1	0	1		0	1
...						



文書ID	トピック 1	トピック 2	...	トピック 11
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(z), P(x|z), P(y|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

$$P(x, y) = \sum_z P(z)P(x|z)P(y|z)$$

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

- 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プリングする
- 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

特異値分解 LSA (Latent Semantic Analysis)

- $(m \times n)$ 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する)
- 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する)
- 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の拡張手法
- PLSA(他左3つの手法も含め)の過学習の問題に対して、LDAではディレクレ分布を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

ハードクラスタリング

- 一つの要素は必ず一つのクラスタに所属する
- 基本的に要素間の距離に基づいて分類を行う
- 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
- 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)

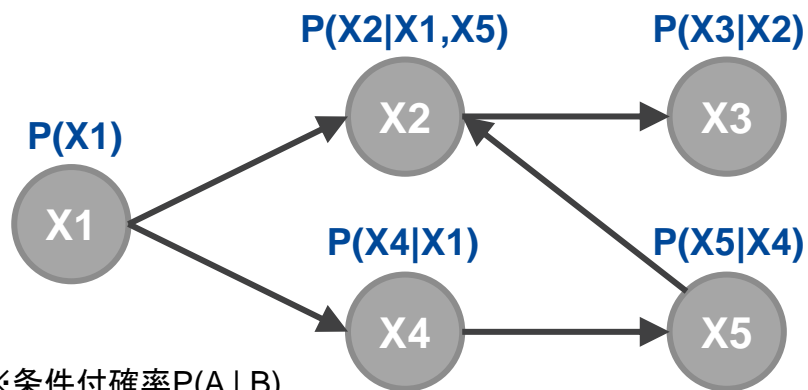
ソフトクラスタリング(潜在クラス分析、トピックモデル、次元圧縮)

- 一つの要素は全てのクラスに所属し、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
- 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
- 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

ベイジアンネットワークは、変数間の確率的な因果関係を探索するモデリング手法です

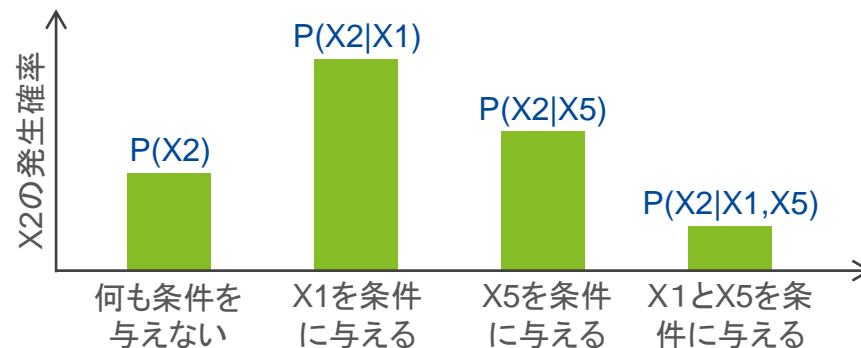
ベイジアンネットワークの概要

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



※条件付確率 $P(A|B)$
事象Bが起こる条件の下で事象Aの起こる確率

確率的因果関係と交互作用



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

ベイジアンネットワークのメリット

現象を理解して柔軟にシミュレーションできる

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

効果を発揮する有用な条件を発見できる

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

テキストで記された現象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、複合条件によって効果が変わ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が変わ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

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

- 文章を単語に分解し、その出現頻度を集計する
- 各文章における出現単語情報のデータ(共起行列)を作成する

単語抽出

Text Mining Studio

PLSA

確率的潜在意味解析

- 単語が出現する文脈を学習し、背後に潜むトピックを抽出する
- 全テキストデータをトピックで説明する(重みを計算する)

トピック抽出

Visual Mining Studio

ベイジアンネットワーク

- トピックを含むテキスト情報内の変数の関係構造をモデル化する
- 各変数が他の変数に与える影響を確率シミュレーションする

モデリング

BayoLink

膨大なテキストデータを人間が理解しやすい形に整理できる

テキストの内容における複雑な要因関係を構造化できる

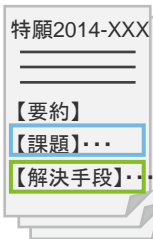
条件を変化させたときの結果の挙動をシミュレーションできる

ある事象の発生確率をコントロールする条件を発見できる

特許要約の【課題】と【解決手段】から用途と技術のトピックを抽出し、トピックのトレンド分析や出願人の特徴分析、また用途と技術の関係分析による新規用途探索を行います

用途と技術のトピック抽出

データの抽出



- 特許文書の要約文の「課題」と「解決手段」のテキストデータを抽出する
- 「課題」からは用途トピックを、「解決手段」からは技術トピックを抽出する

トピックの特徴集計

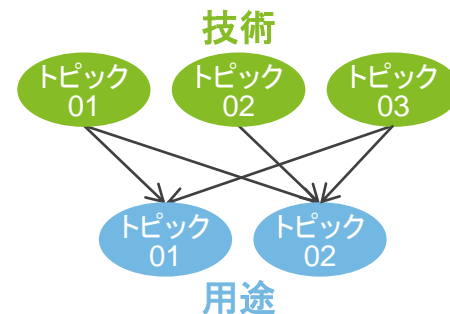
全特許データに各トピックのスコア(該年度)を計算する

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

用途と技術の関係分析

用途トピックと技術トピックの統計的な関係性をベイジアンネットワークでモデル化する

ベイジアンネットワーク



テキストマイニング

テキストマイニングを実行して単語と係り受け表現を抽出する

PLSA

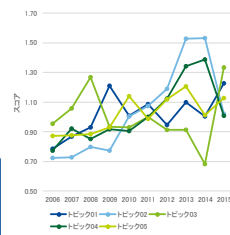
「単語×係り受け」の共起行列を作成し、これにPLSAを適用してトピックを抽出する

係り受け

	機提供	空気調和	効率良い	掃除機提供
単語				
空気調和機	1,578	100	1	
空気	85	144	45	
容易	190	105	67	

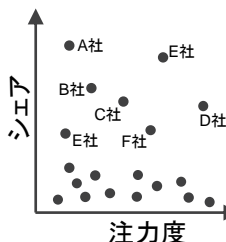
出願年集計

トピックスコアを出願年で集計してトピックのトレンドを把握する

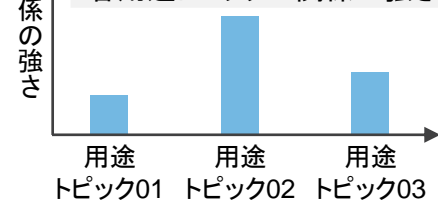


出願人集計

トピックスコアを出願人で集計して、各トピックにおける出願人の特徴を把握する



保有技術トピックに対する各用途トピックの関係の強さ

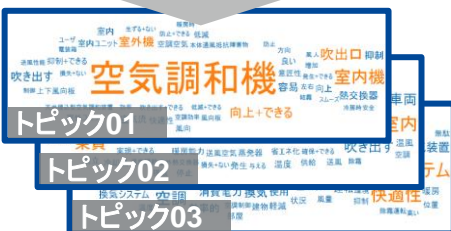


保有技術と関係のある用途トピックのうち、まだ想定していない用途を探索し、それに関連する元の特許文書を確認することで具体的な新規用途を検討する

用途トピックの抽出
技術トピックの抽出

単語	品詞	頻度
空気調和機	名詞	3,106
空気	名詞	2,846
容易	名詞	2,790
抑制	名詞	2,687
...

係り受け表現	頻度
空気調和機-提供	1,575
効率-良い	1,325
掃除機-提供	545
容易-構成	539
...	...



分析事例のご紹介

分析データ

「風」「空気」に関する10年分の特許データ30,039件を分析します

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～2015年

■ 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

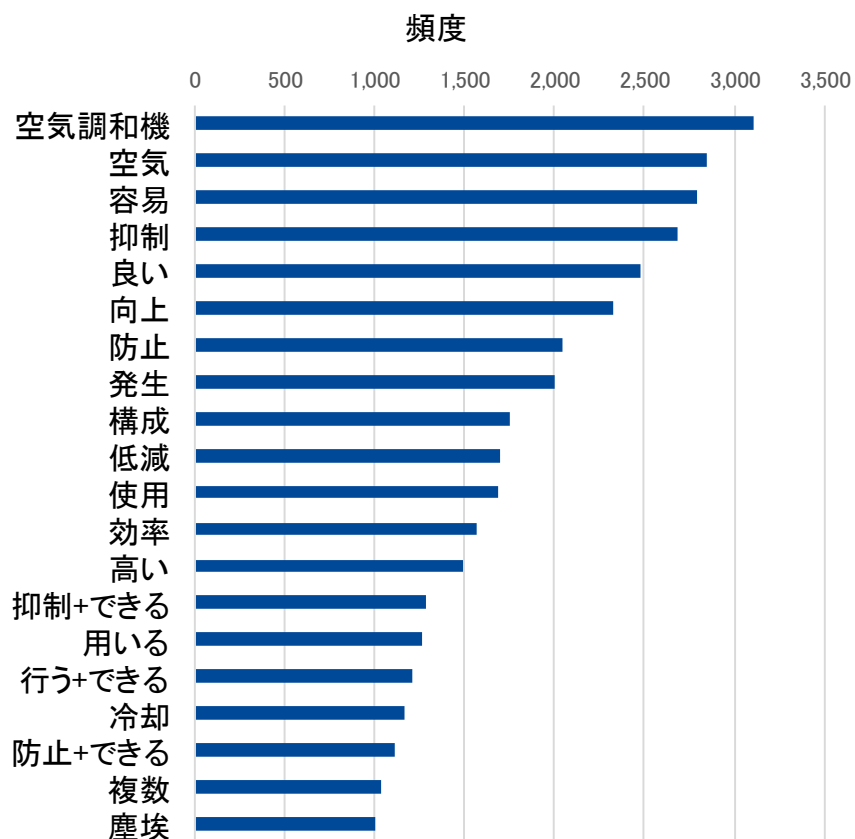
【要約】【課題】ユーザーの快適性を維持しつつ、省エネ運転を行うことができる空気調和機を提供すること。【解決手段】本発明の空気調和機は、室内温度を検出する室内温度検出手段と、人体の活動量を検出する人体検出手段と、基準室内設定温度を設定するリモコン装置30とを備え、室内温度が基準室内設定温度となるように空調制御を行う空気調和機であって、人体検出手段で検出する活動量が所定の活動量以内であるときは、室内温度が、基準室内設定温度を補正した補正室内設定温度となるように空調を行い、補正室内設定温度よりも低い状態を継続すると、圧縮機を停止させ、圧縮機の復帰は、基準室内設定温度に基づいて行う。

用途のトピック抽出

「課題」に記載されている文章に含まれる単語と係り受け表現を抽出します

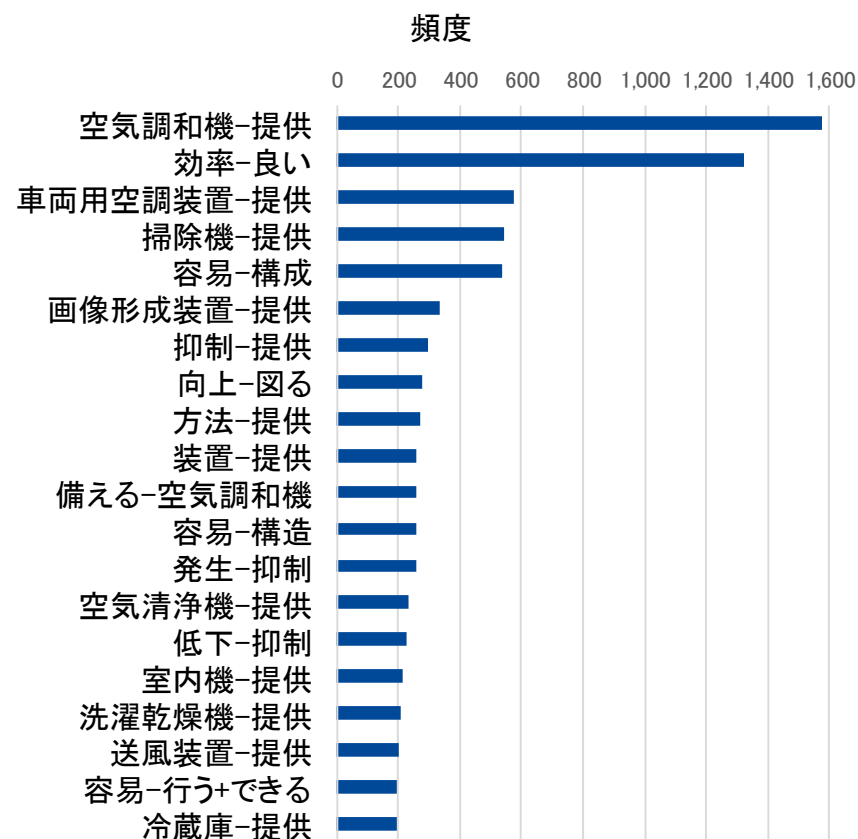
単語の抽出

- 名詞、形容詞、形容動詞、動詞の単語を抽出



係り受け表現の抽出

- 名詞×形容詞、形容動詞、動詞(サ変名詞含む)の係り受けペアを抽出



単語×係り受け表現の共起行列にPLSAを適用し、「課題」の文章で同時に出現しやすい単語と係り受け表現を共通するトピックに集約します

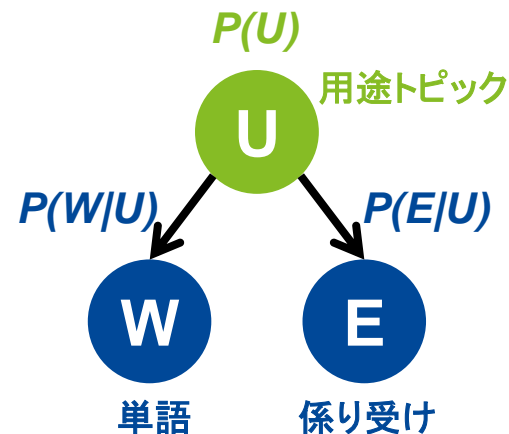
共起行列の作成

- 単語×係り受け表現の共起行列(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:3,256語×係り受け表現:2,084表現(それぞれ頻度10件以上を対象)

		係り受け表現									
		空調和機-提供	効率-良い	車両用空調装置-提供	掃除機-提供	容易-構成	画像形成装置-提供	抑制-提供	向上-図る	方法-提供	装置-提供
単語	空調和機	1578	100	4	1	55	0	39	27	2	1
	空気	85	144	45	45	50	34	31	19	13	22
	容易	190	105	51	67	540	28	14	14	32	25
	抑制	142	95	64	63	36	55	296	27	15	7
	良い	113	1331	12	56	43	31	27	9	19	24
	向上	122	33	24	51	22	8	15	279	11	4
	防止	83	51	23	21	35	34	14	30	6	18
	発生	79	108	40	13	30	43	43	12	13	22
	構成	84	61	42	44	540	31	11	9	7	9
	低減	80	49	44	18	19	10	14	27	16	8

PLSAの適用

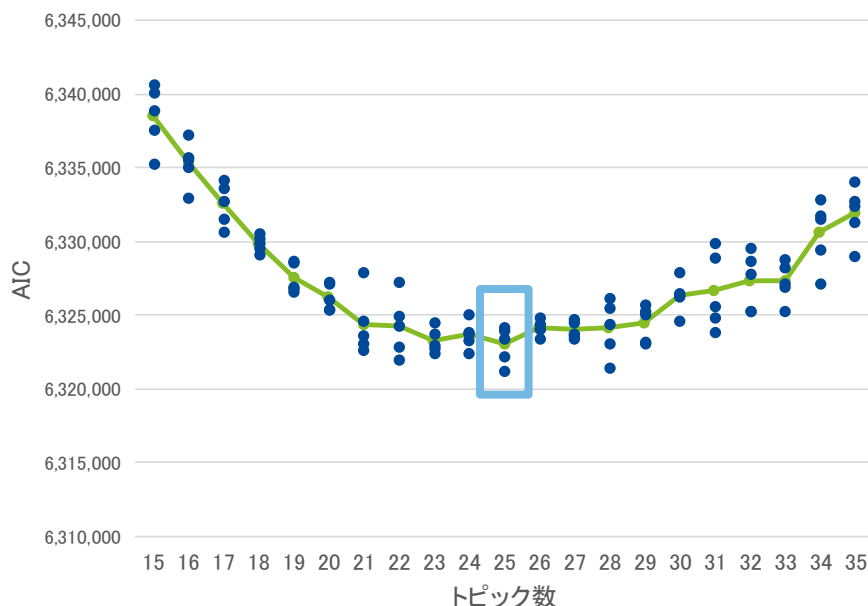
- Visual Mining Studioの二項ソフトクラスタリングを使用する
 - PLSAにNMF(非負値行列分解)を組み合わせた手法で、PLSAの局所最適化を改善する
- クラスタ数と初期値
 - PLSAはあらかじめトピック数を設定する必要があり、初期値により解が異なる初期値依存性がある
 - トピック数を15から35まで1刻みで変化させ、それぞれのトピック数に対して二項ソフトクラスタリングを初期値を変えて5回ずつ実行し、情報量基準AICを計算する



用途では25個のトピックが抽出されました

PLSA実行解の評価

- それぞれの実行解における情報量基準AICを評価
- トピック数=25の実行解の一つがAIC最小となった



- それぞれの実行解のAIC
- 各トピック数における実行解の平均AIC

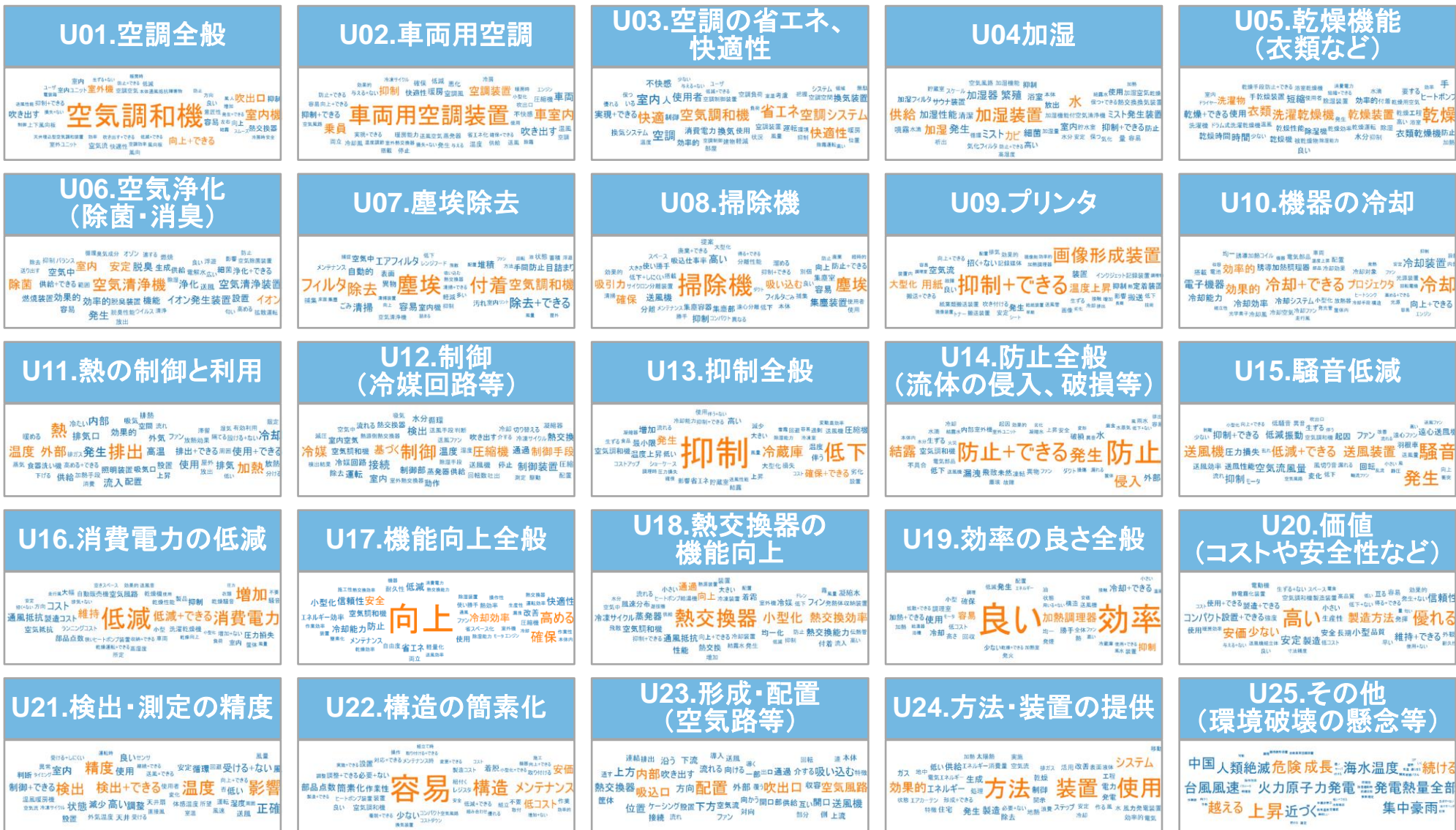
トピックの構成

- PLSAでは以下の3つの確率がアウトプットとなる
 - $P(W|U)$ …トピックにおける単語の所属確率
 - $P(E|U)$ …トピックにおける係り受けの所属確率
 - $P(U)$ …トピックの存在確率
- トピックの構成の例
 - 以下は加湿に関する話題であることがわかる

U04			
P(U)=2.6%			
P(W U)	単語	P(E U)	係り受け
5.5%	加湿装置	6.8%	加湿装置-提供
3.7%	水	3.1%	加湿器-提供
3.3%	供給	2.9%	ミスト発生装置-提供
2.4%	加湿	1.9%	水-供給
2.3%	カビ	1.7%	細菌-繁殖
2.1%	加湿器	1.5%	加湿-行う
2.1%	発生	1.4%	加湿機能付空気清浄機-提供
2.0%	繁殖	1.3%	ミスト-噴霧
1.9%	ミスト	1.3%	繁殖-抑制
1.7%	加湿性能	1.2%	十分-量
1.5%	ミスト発生装置	1.2%	カビ-発生
1.4%	細菌	1.2%	効率-良い
1.3%	室内	1.2%	空気調和機-提供
1.3%	抑制+できる	1.1%	加湿-加湿装置
1.1%	浴室	1.1%	空気-加湿
...

用途ピック25個の一覧

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、機能向上や構造の簡素化などの用途が抽出されました

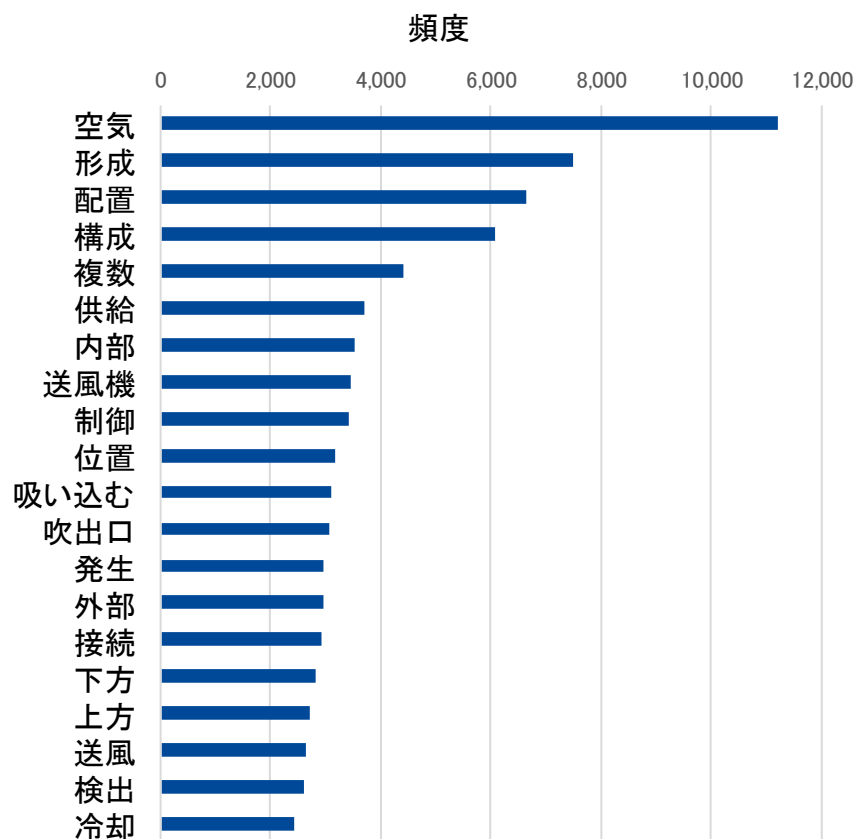


技術のトピック抽出

「解決手段」に記載されている文章に含まれる単語と係り受け表現を抽出します

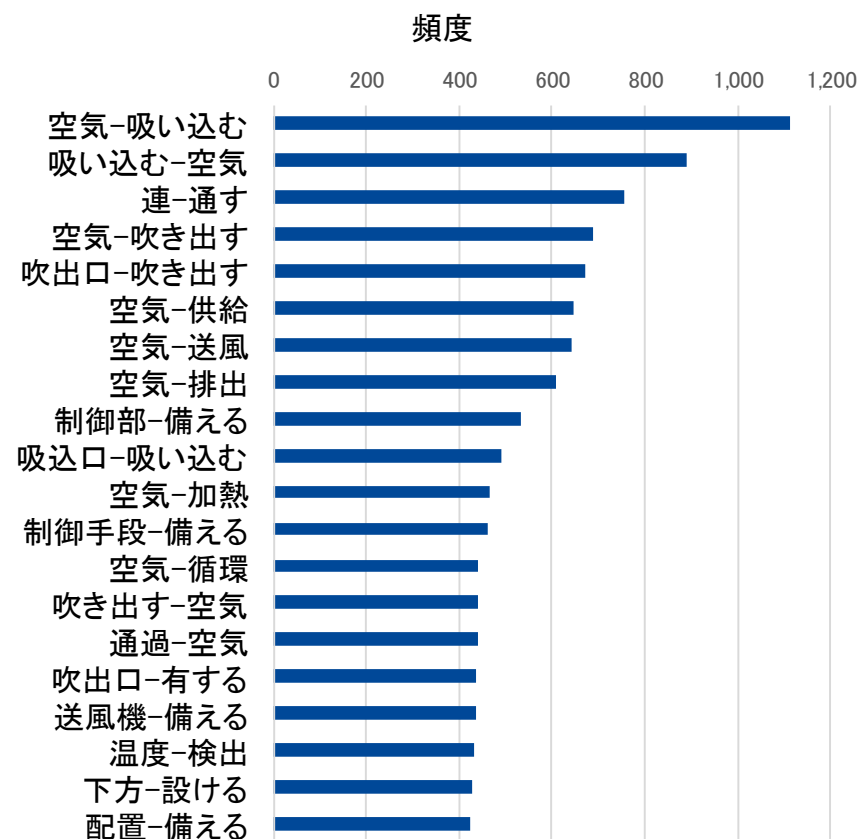
単語の抽出

- 名詞、形容詞、形容動詞、動詞の単語を抽出



係り受け表現の抽出

- 名詞×形容詞、形容動詞、動詞(サ変名詞含む)の係り受けペアを抽出



単語×係り受け表現の共起行列にPLSAを適用し、「解決手段」の文章で同時に出現しやすい単語と係り受け表現を共通するトピックに集約します

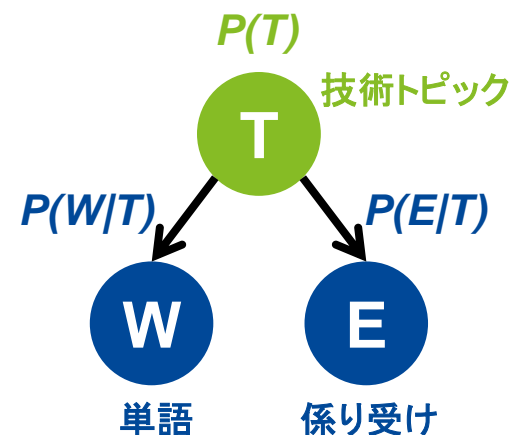
共起行列の作成

- 単語×係り受け表現の共起行列(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:5,187語×係り受け表現:7,174表現(それぞれ頻度10件以上を対象)

		係り受け表現									
		空気-吸い込む	吸い込む-空気	連-通す	空気-吹き出す	吹出口-吹き出す	空気-供給	空気-送風	空気-排出	制御部-備える	吸込口-吸い込む
単語	形成	254	207	219	180	136	124	157	166	65	134
	配置	210	191	150	123	143	100	127	100	59	118
	構成	153	135	142	96	105	117	97	99	49	79
	複数	101	93	88	74	78	71	70	63	50	34
	供給	127	101	67	54	53	675	73	81	65	50
	内部	134	107	126	76	44	84	71	111	39	64
	送風機	225	197	112	113	73	125	109	58	72	137
	制御	81	76	29	72	103	68	92	47	302	50
	位置	82	72	70	68	72	39	44	51	28	43
	吸い込む	1,146	917	99	313	244	68	70	112	58	504

PLSAの適用

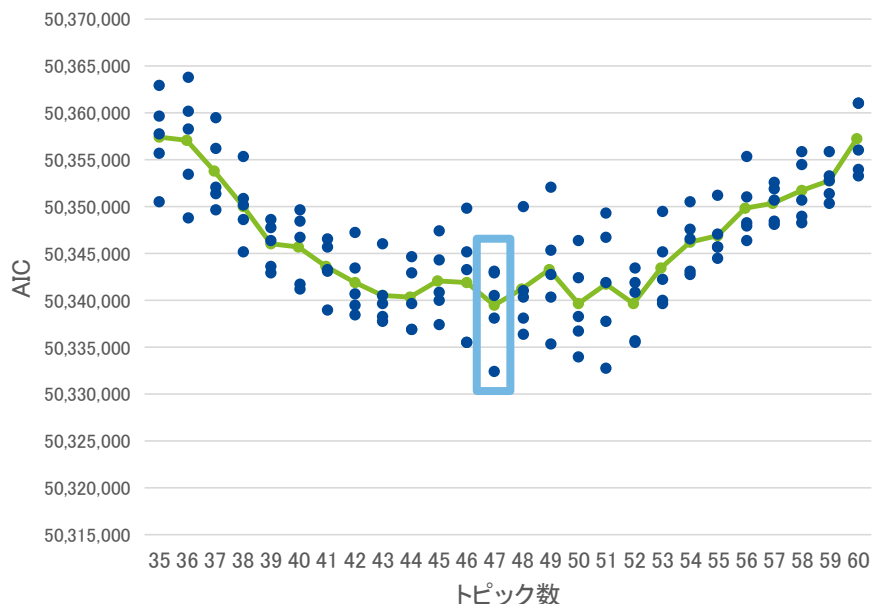
- Visual Mining Studioの二項ソフトクラスタリングを使用する
 - PLSAにNMF(非負値行列分解)を組み合わせた手法で、PLSAの局所最適化を改善する
- クラスタ数と初期値
 - PLSAはあらかじめトピック数を設定する必要があり、初期値により解が異なる初期値依存性がある
 - トピック数を35から60まで1刻みで変化させ、それぞれのトピック数に対して二項ソフトクラスタリングを初期値を変えて5回ずつ実行し、情報量基準AICを計算する



技術では47個のトピックが抽出されました

PLSA実行解の評価

- それぞれの実行解における情報量基準AICを評価
- トピック数=47の実行解の一つがAIC最小となった



トピックの構成

- PLSAでは以下の3つの確率がアウトプットとなる
 - $P(W|T)$ …トピックにおける単語の所属確率
 - $P(E|T)$ …トピックにおける係り受けの所属確率
 - $P(T)$ …トピックの存在確率

トピックの構成の例

- 以下は塵埃の分離に関する話題であることがわかる

T32			
P(T)=2.7%			
P(W T)	単語	P(E T)	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
1.4%	捕集	0.8%	掃除機-設ける
1.3%	集塵室	0.8%	空気-吸い込む
1.3%	通過	0.8%	送風機-備える
1.3%	発生	0.8%	掃除機-備える
1.2%	集塵装置	0.7%	吸い込む-空気
...

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{0:該当無,1:該当有}のデータに変換します

①文章単位にスコアを算出

- 句点(.)などで区切られる文章単位に、各トピックの該当度を計算(計算方法は次のスライドで説明)
- 用途の文章数:37,087件
- 技術の文章数:61,611件

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		1.6
...						

②特許IDごとに文章スコアを集約

- 特許IDごとにその中の文章スコアの最大値を採用する
- 特許ID数:30,039件

特許ID	T01	T02	T03	...	T47
1	3.1	0.9	5.5		2.4
2	1.2	5.8	2.6		1.6
...					

③閾値を設定して{0,1}のデータに変換

- スコアと実際の文章の内容を確認しながら、スコアの閾値を設定し、{0:該当無,1:該当有}のデータに変換する
- 今回は閾値を3.0に設定

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

文章単位のトピックのスコアを単語と係り受けの両方の軸から確率的に計算します

文章単位のスコア

$$\frac{P(S|T)}{P(S)}$$

- 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

単語 W_i で定義される文章 Sw_h

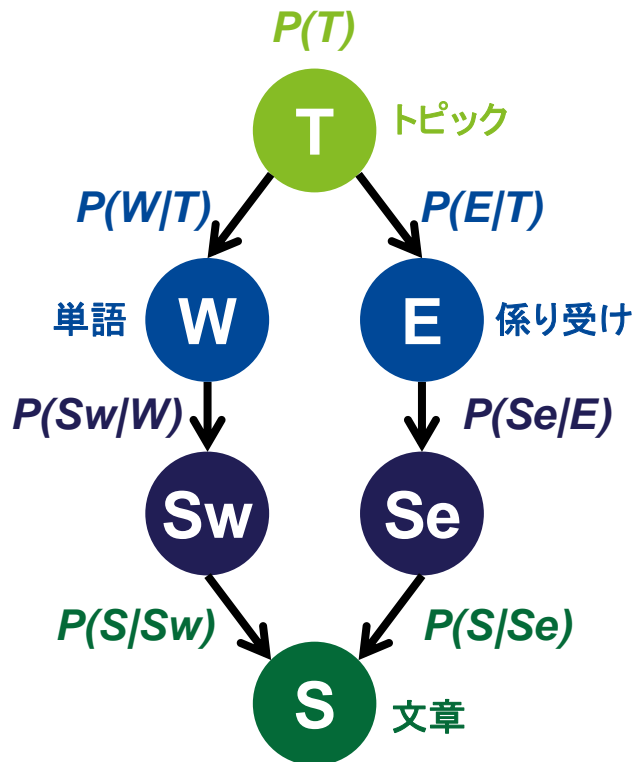
$$Sw_h = \{W_1, W_2, \dots, W_i\}$$

トピック T_k を条件とした文章 Sw_h の出現確率

$$P(Sw_h|T_k) = \sum_i P(Sw_h|W_i)P(W_i|T_k)$$

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

$$P(Sw_h|W_i) = 1/n(W_i)$$



文章を単語で定義される文章 Sw と係り受けで定義される文章 Se 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

係り受け E_j で定義される文章 Se_h

$$Se_h = \{E_1, E_2, \dots, E_j\}$$

トピック T_k を条件とした文章 Se_h の出現確率

$$P(Se_h|T_k) = \sum_j P(Se_h|E_j)P(E_j|T_k)$$

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

$$P(Se_h|E_j) = 1/n(E_j)$$

$$P(S_h|T_k) = P(S_h|Sw_h)P(Sw_h|T_k) + P(S_h|Se_h)P(Se_h|T_k)$$

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

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

文章 S_h の出現確率

特許データとトピックのスコアデータを結合することで、トピックをベースとした様々な集計・分析を実行することができます

トピックスコアを紐づけた特許データ

特許ID	出願番号	出願年	出願人	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
1	特願2006-XXXX	2006	A社	1	1		0	1	0		0
2	特願2009-XXXX	2009	B社	0	1		1	0	1		0
3	特願2012-XXXX	2012	C社	0	1		1	1	0		0
4	特願2013-XXXX	2013	D社	1	0		0	1	0		1
...
30039	特願2015-XXXX	2015	X社	1	0		1	0	0		1

①出願年の集計

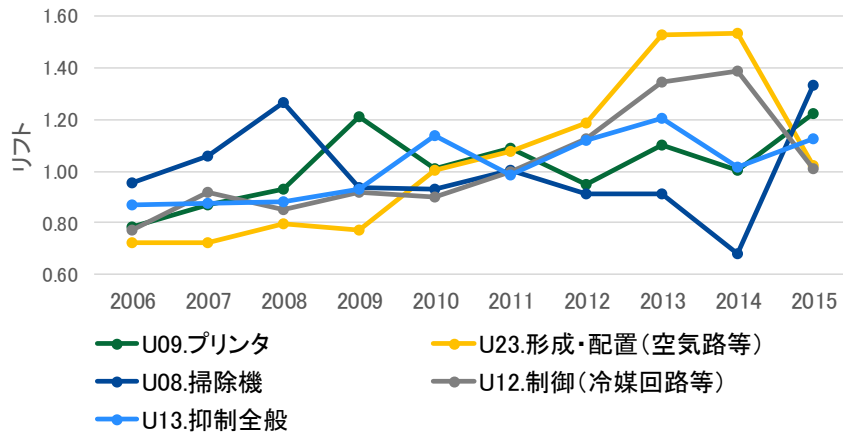
②出願人の集計

③用途と技術の関連性の分析

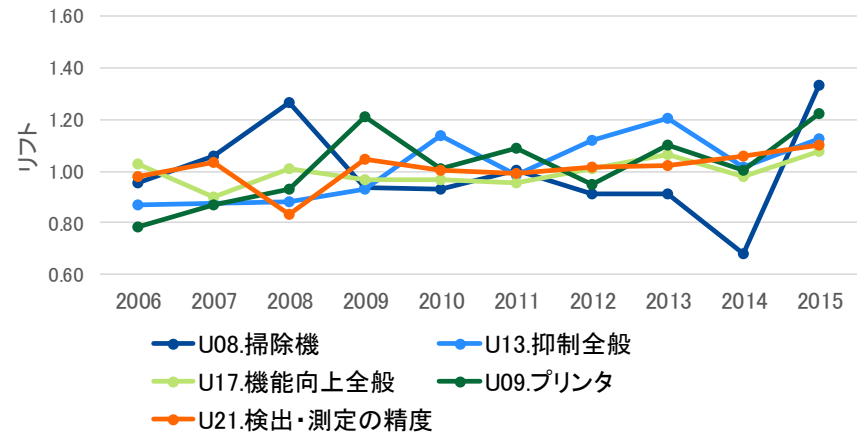
出願年のトピック集計によるトレンド分析

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

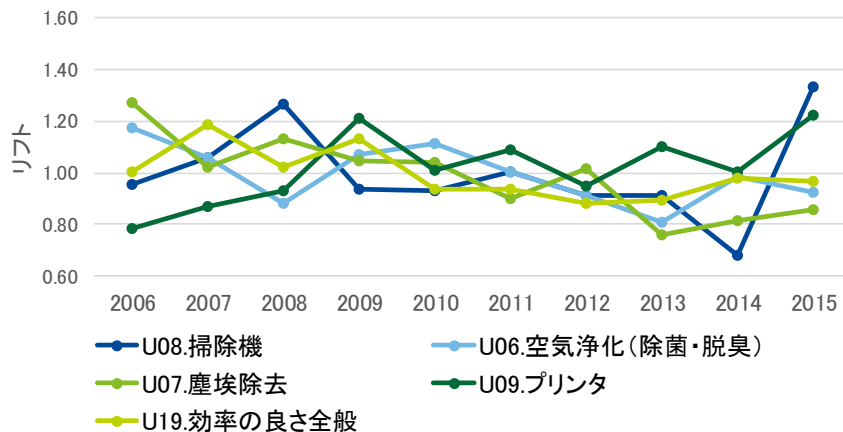
2006年からの上昇率 best5



2011年からの上昇率 best5



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

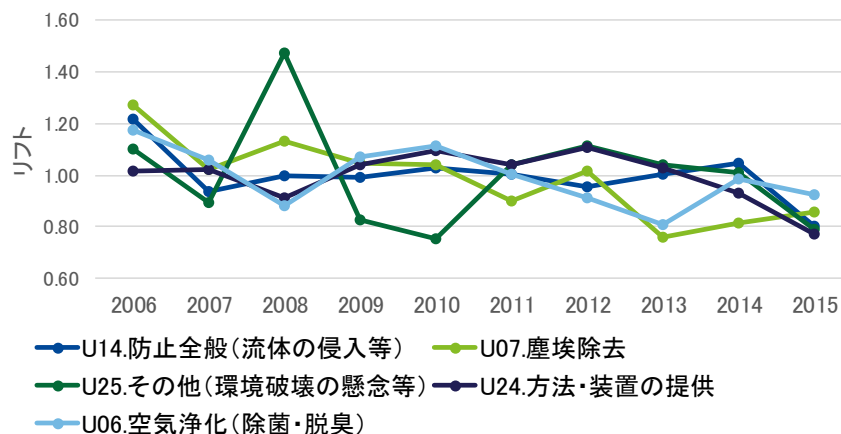
$$P(\text{出願年} | \text{トピック } U_x = 1)$$

$$P(\text{出願年})$$

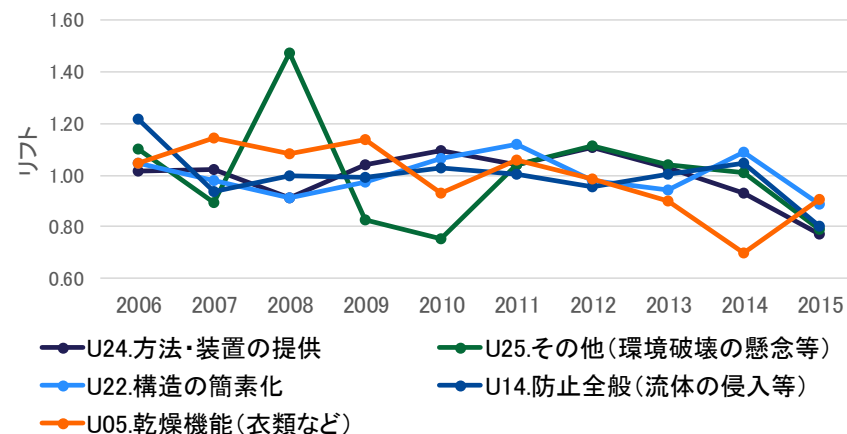
- その出願年の出願件数割合を平均(=1)として標準化した値

長期的には塵埃除去や空気浄化は下降傾向にあり、近年は冷媒回路などの制御に関する用途が下降しています

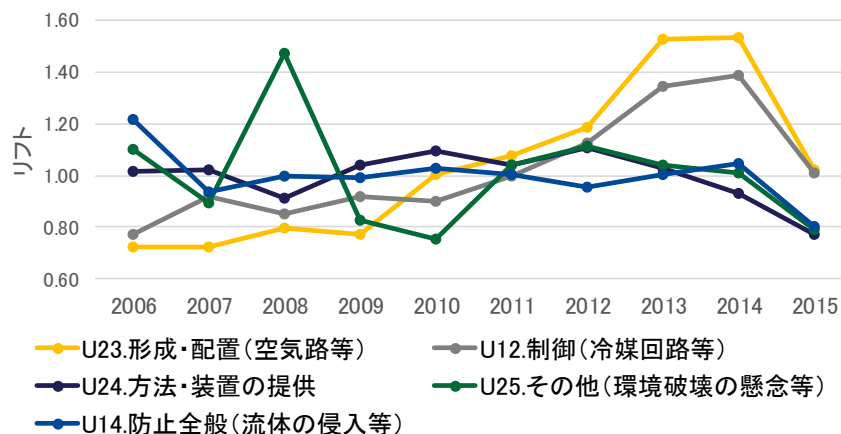
2006年からの下降率 worst5



2011年からの下降率 worst5



2013年からの下降率 worst5



集計の仕方

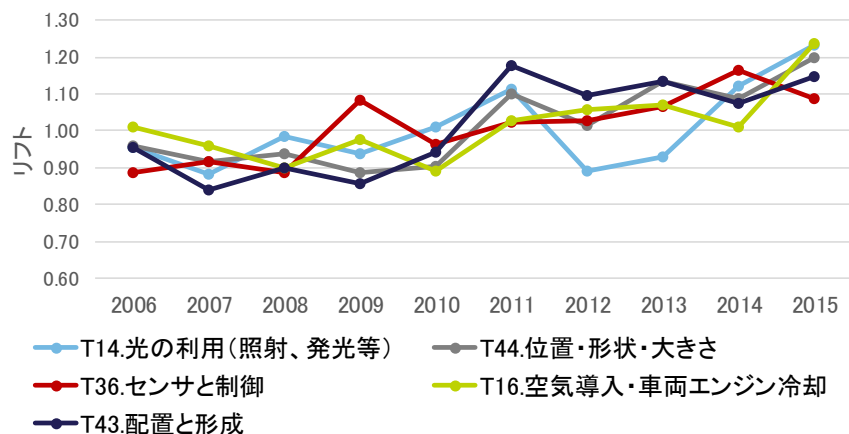
■ リフト値を出願年・トピックごとに集計

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

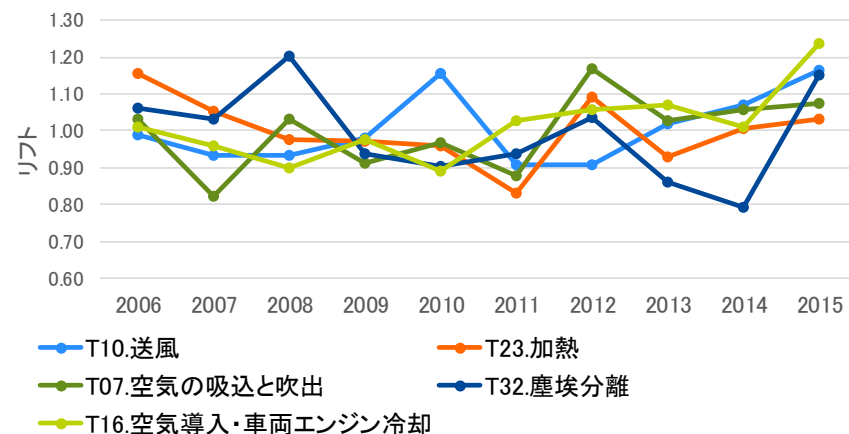
■ その出願年の出願件数割合を平均(=1)として標準化した値

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

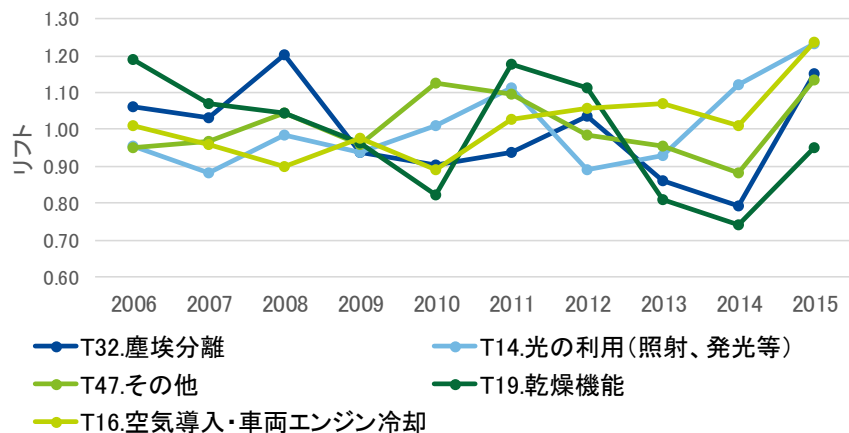
2006年からの上昇率 best5



2011年からの上昇率 best5



2013年からの上昇率 best5



集計の仕方

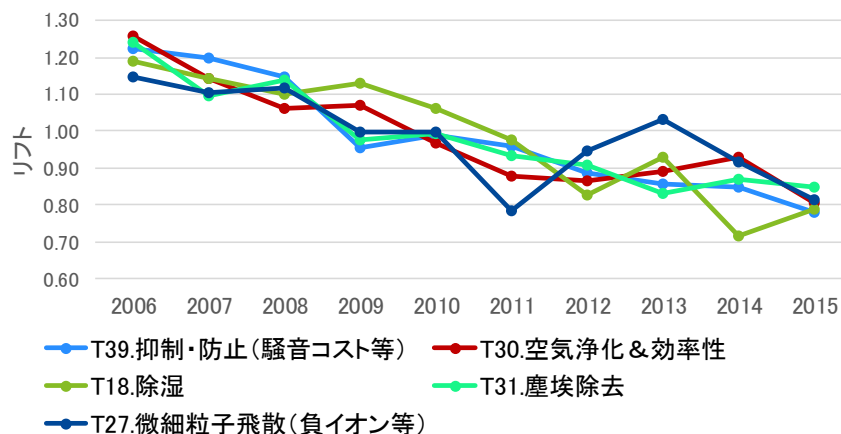
- リフト値を出願年・トピックごとに集計

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

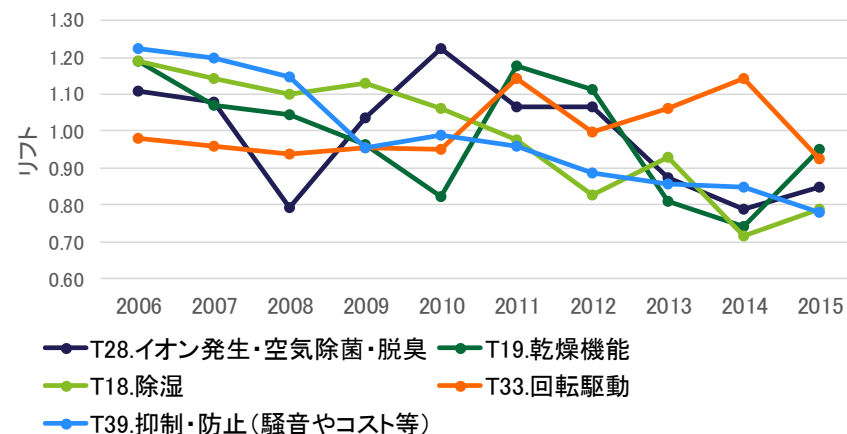
- その出願年の出願件数割合を平均(=1)として標準化した値

中長期的には空気浄化(イオン発生や除菌、脱臭、塵埃除去など)や除湿・乾燥に関する技術が下降しており、短期的には冷凍サイクルや加湿に関する技術も下降傾向にあります

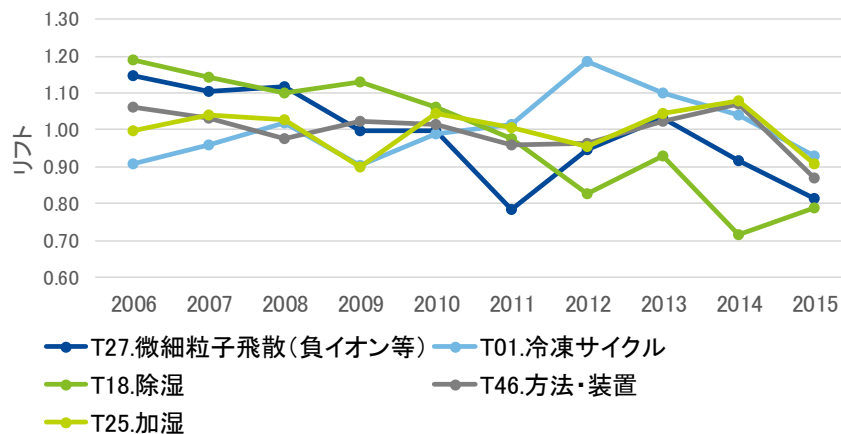
2006年からの下降率 worst5



2011年からの下降率 worst5



2013年からの下降率 worst5



集計の仕方

- リフト値を出願年・トピックごとに集計

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

- その出願年の出願件数割合を平均(=1)として標準化した値

出願人のトピック集計による競合分析

注力度とシェアの2つの軸で各トピックにおける出願人の特徴を把握します

注力度(関心度)

- 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
- 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おり、高い技術を保有している可能性がある

$$P(\text{トピック}T = 1 \mid \text{出願人}X = 1)$$

シェア

- 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
- 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おり、技術市場のシェアともいえる

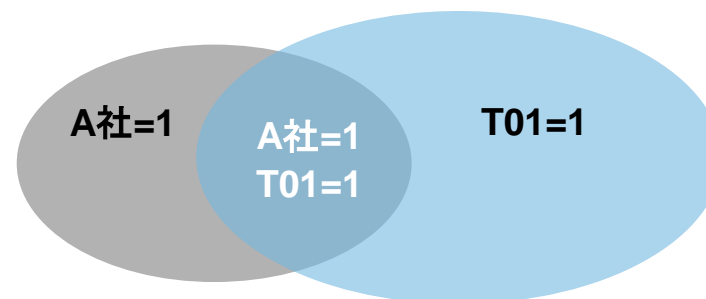
$$P(\text{出願人}X = 1 \mid \text{トピック}T = 1)$$

集計対象

- 名寄済みの出願人情報のうち、出願件数上位26社を集計対象とする

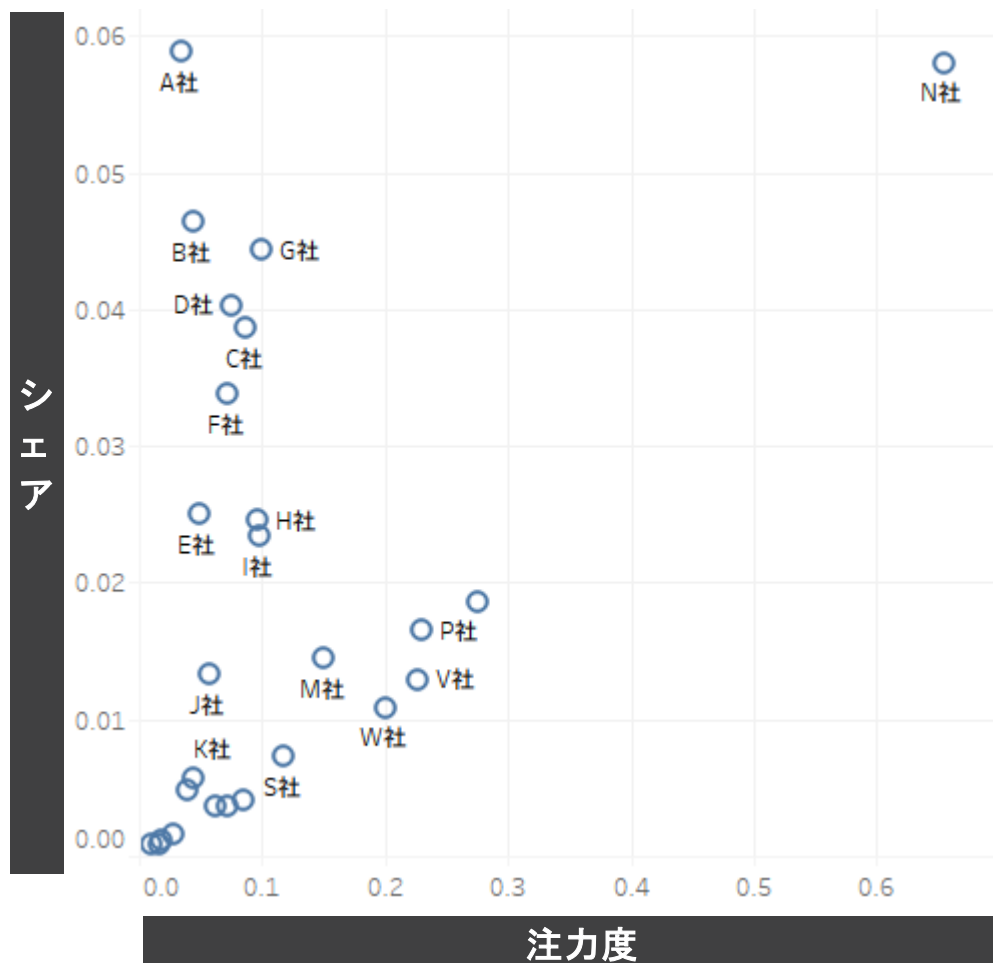
集計データと集計イメージ

特許ID	出願人 A社	出願人 B社	出願人 C社	...	トピック T01	トピック T02	トピック T03	...
1	1	0	0		1	0	0	
2	1	0	0		0	0	1	
3	1	0	0		1	0	0	
4	1	0	0		1	0	0	
5	1	0	0		1	0	0	
6	0	1	0		0	1	0	
7	0	1	0		0	1	1	
8	0	1	0		1	1	0	
9	0	1	1		0	0	0	
10	0	0	1		0	1	0	
...	



光の利用に関する技術は、ある1社が注力度もシェアも他社を大きく突き放しています

注力度とシェアの散布図

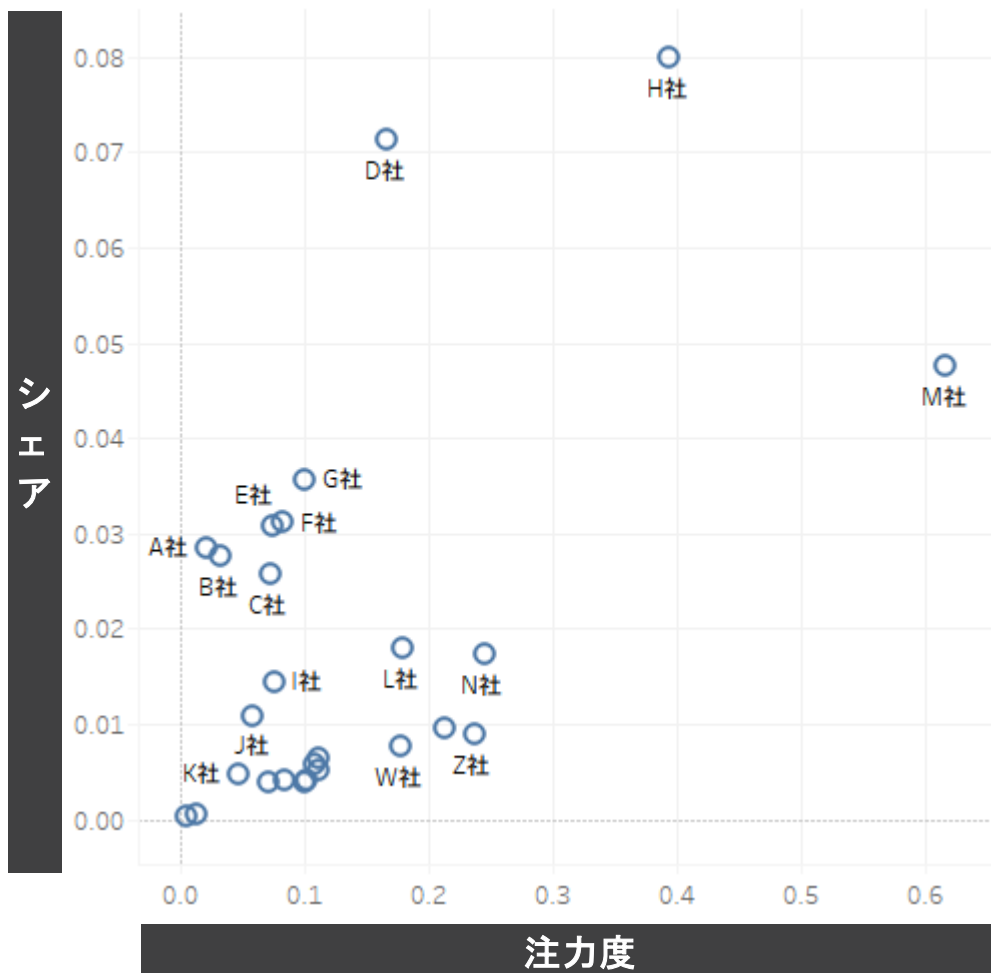


考察と戦略の検討

- 注力度もシェアもN社が他社を圧倒している
- 特に注力度に大きな乖離があり、技術的に突き放されている可能性がある
- 複数社で連携してN社に対抗することも考えられる

車両エンジンの冷却に関する技術は、3社が競合している状態にあるといえます

注力度とシェアの散布図

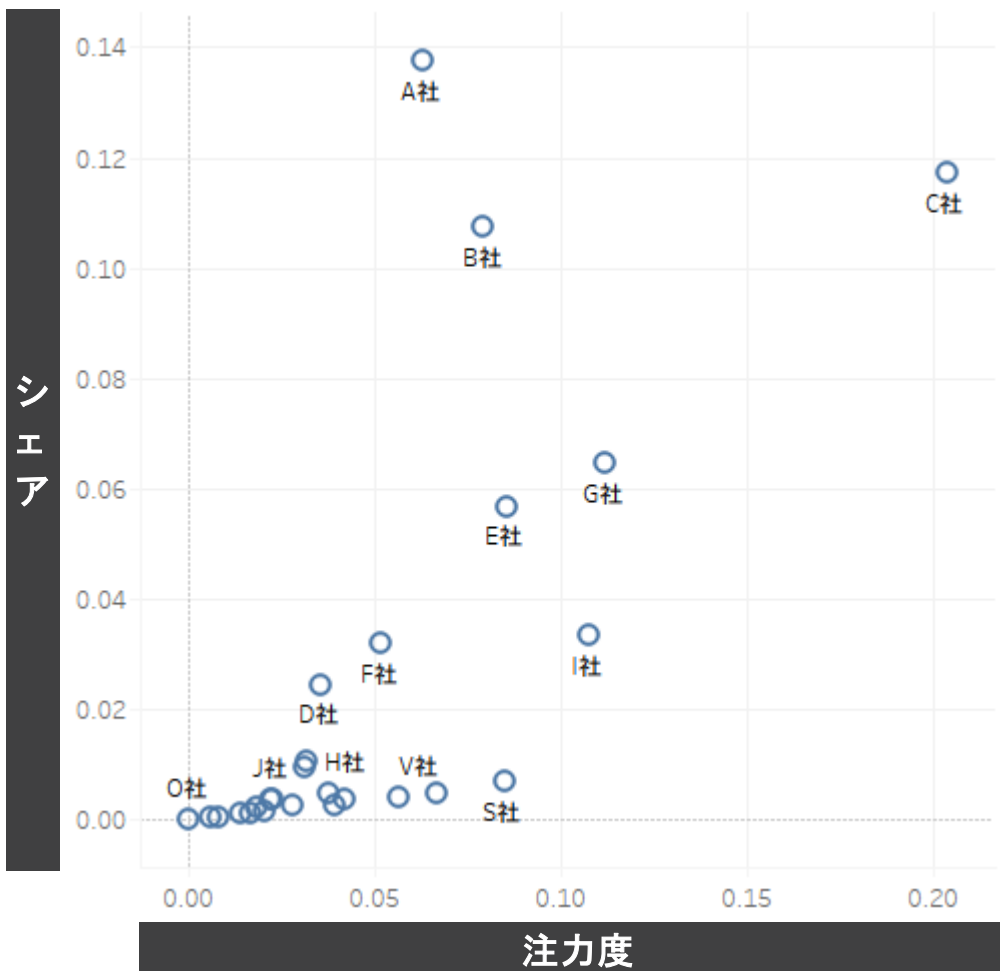


考察と戦略の検討

- H社、D社、M社の3社が競合しているように思われる
- D社もM社もある程度シェアを確保しており、そのなかでもD社はより高いシェアがあり、M社はより高い注力度があるが、この2社が連携することがあれば高シェア高注力度のポジションを獲得できる可能性がある
- H社はそうした他社の連携動向に対抗する手段として、シェアは小さくても注力度が高く技術力の高い小規模の企業と連携することで、より高シェア高注力度のポジションを獲得できると思われる

塵埃分離に関する技術は、1社の注力度が高いものの、他にもある程度のシェア・注力度を保有する企業が何社か存在するため、今後連携などの動きも考えられる領域と思われます

注力度とシェアの散布図



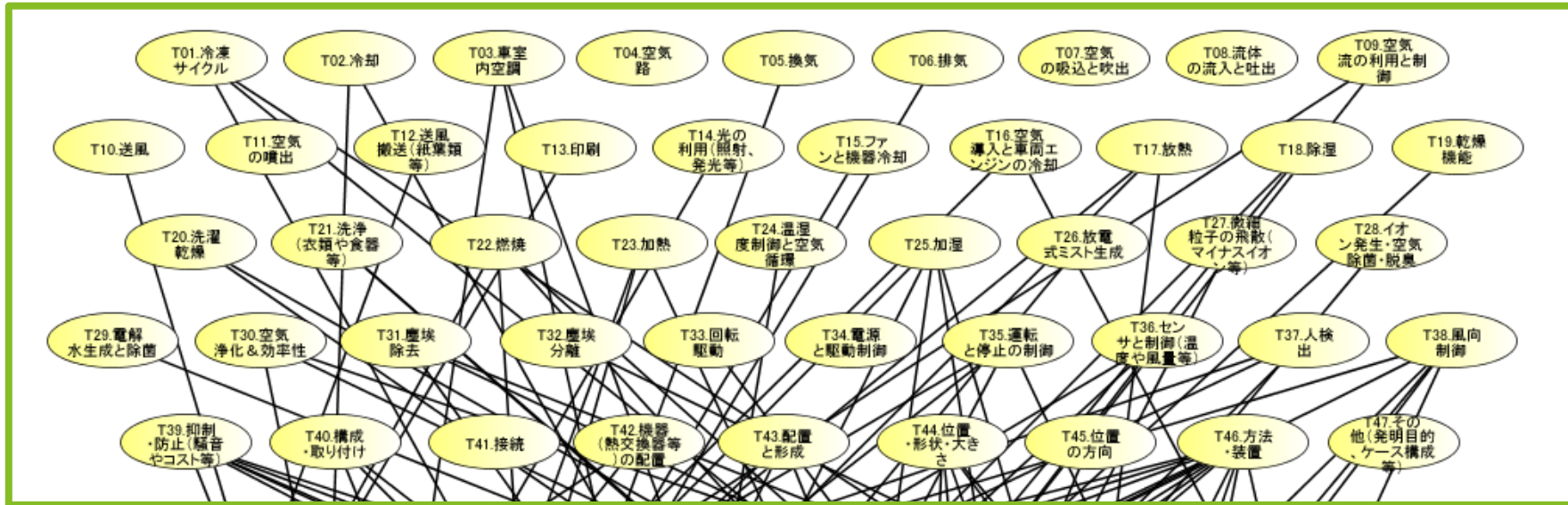
考察と戦略の検討

- C社は、高めのシェアを獲得しつつ、他社と比べて注力度がとて高く、高い技術力を保有していると考えられ、今後はよりシェアを伸ばすことで高シェア高注力度のポジションを確立することができると考えられる
- A社とB社は、シェアは高いがまだC社に注力度で劣っているので、例えば規模は中程度だが注力度は比較的高く、技術力があると思われるE社、G社、I社などと連携することで、C社の上のポジションを狙うことができる可能性がある

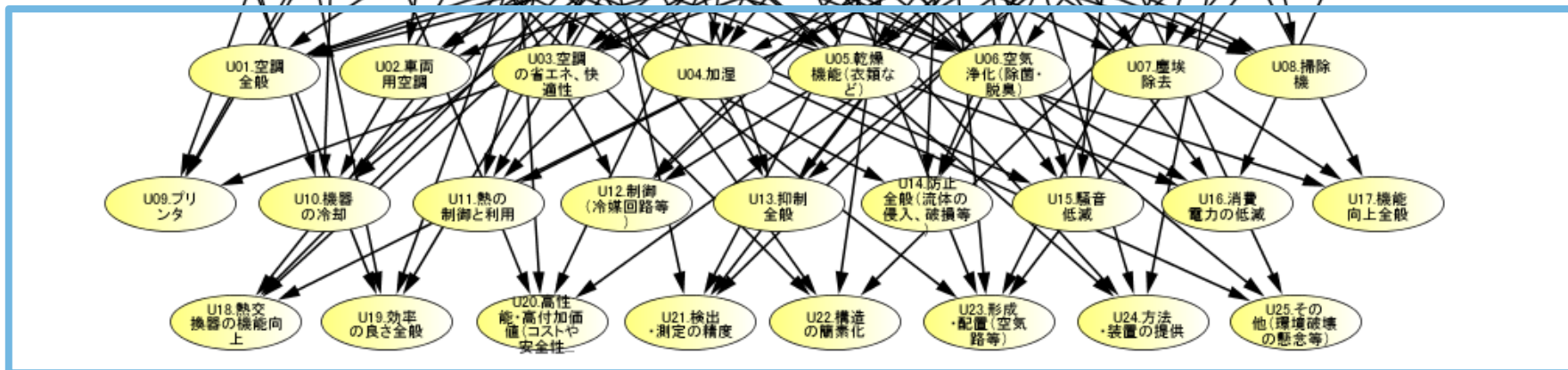
用途と技術の関係分析による技術の新規用途探索

ベイジアンネットワークを適用した技術と用途の関係分析

用途トピックと技術トピックの{0,1}データにベイジアンネットワークを適用して、技術⇒用途の確率的因果関係をモデル化します



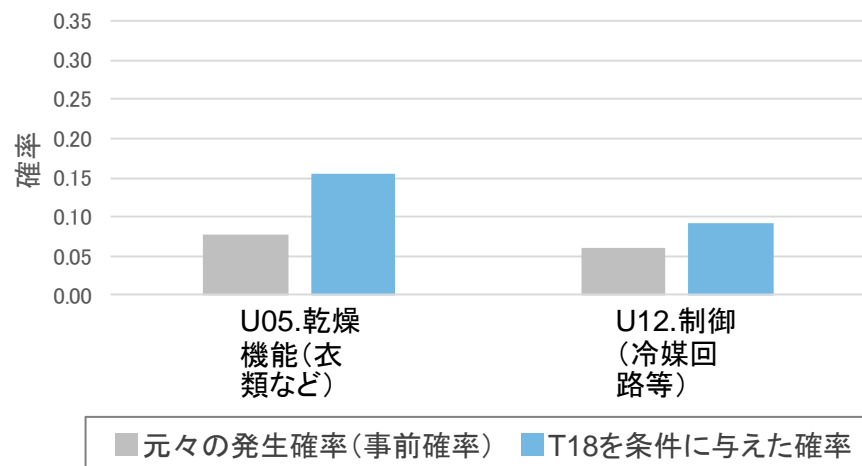
技術トピック



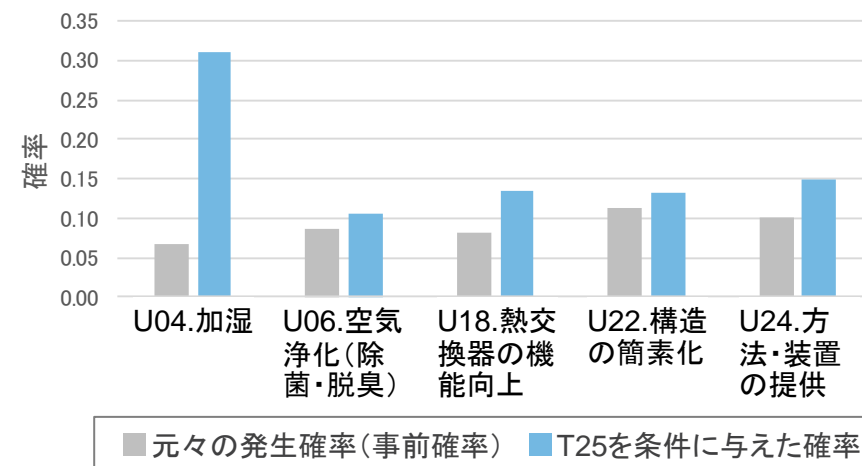
用途トピック

技術トピックを条件に与えたとき、それと確率的因果関係を持つと判定された各用途トピックの確率がどのように変化するのかシミュレーションして、その関連性の強さを確認します

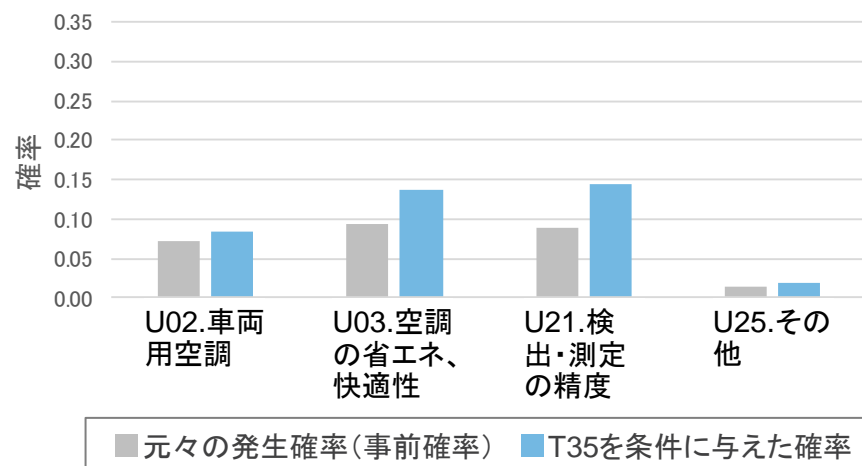
「T18.除湿」を条件に与えた結果



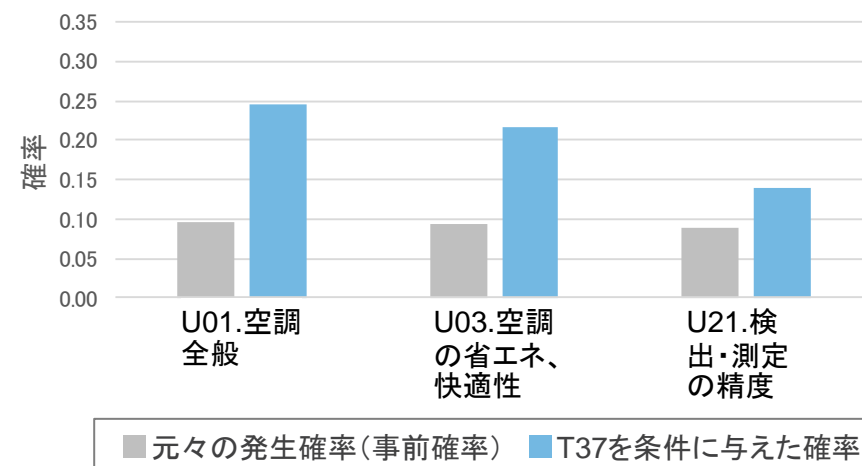
「T25.加湿」を条件に与えた結果



「T35.運転と停止の制御」を条件に与えた結果

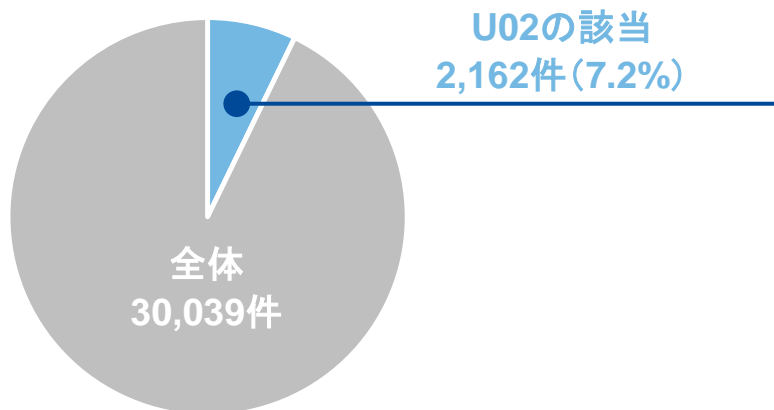


「T37.人検出」を条件に与えた結果

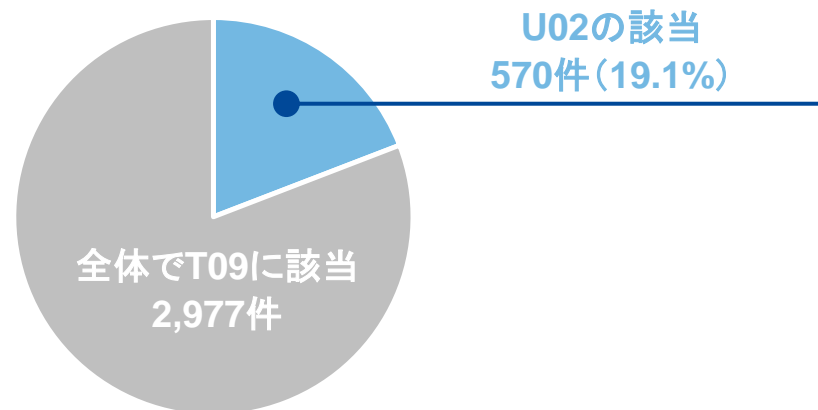


「U02.車両用空調」の用途は、「T09.空気流の利用と制御」の技術の応用先として高い関連性がありますが、出願人Xの保有するT09ではそれがありません

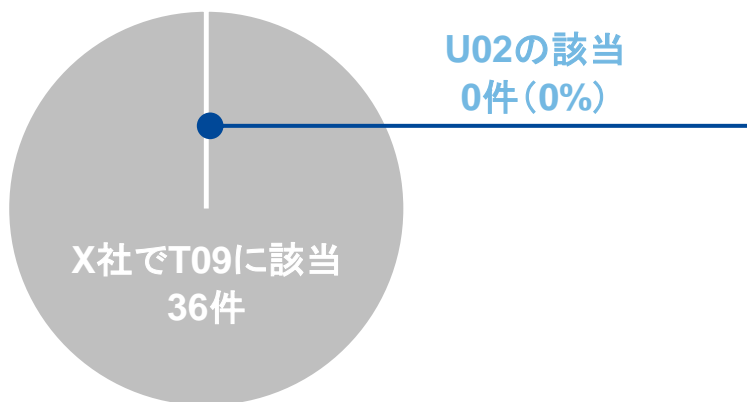
全体でのU02の該当割合



全体でのT09におけるU02の割合



出願人XのT09におけるU02の割合



考察

- ベイジアンネットワークのモデルでは、「T09.空気流の利用と制御」に対する「U02.車両用空調」の関係が見られた
- 全体では、U02の該当は7.2%だが、T09を条件としたときでは、その該当割合が19.1%となり高い関連性が認められる
- しかし、出願人Xでは、T09に該当する特許のうち、U02に該当する特許は1件もない
⇒X社の保有するT09はU02での用途も考えられる

印刷機内で気流を制御して空気量を調整する技術は、車両室内での温度調整のための気流制御にも応用できるかもしれません

T09がU02で応用されている例

発明の名称
車両用空調システム
課題
イグニッションオフ時でも車両内の残熱を利用して室内温度を快適に保つことができる車両用空調システムを提供する。
解決手段
車両用空調システムにおいて、エンジン駆動時に加熱されたヒータコアが、車載空調装置に車載バッテリーから駆動電源を投入することにより、車室内へ向かう空調風を発生させ、エアミックスダンパの開度をヒータコアを通過する空気が多くなるように気流を調整して、車内温度が車外温度に近づくことを妨げる。

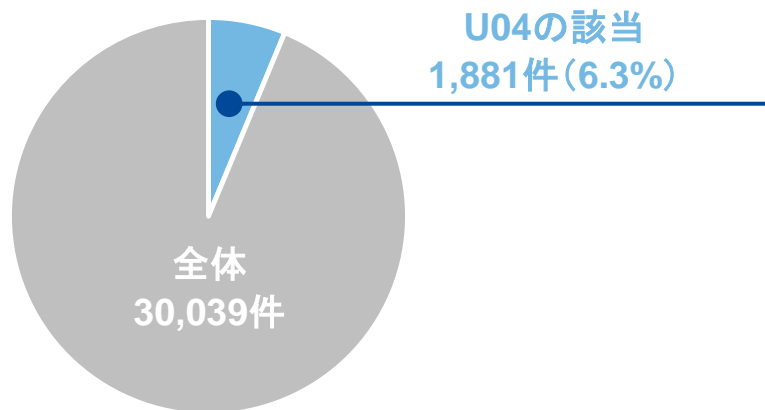
出願人Xの保有するT09の例

発明の名称
画像形成装置及び電子装置
課題
本体内の温度上昇を抑制することができる画像形成装置及び電子装置を提供する。
解決手段
画像形成装置は、画像形成装置本体と、本体に設けられた画像形成部と、本体から排出される空気が流れる第1の流路と、第1の流路から排出される空気の流れと干渉するように排出される空気が流れる第2の流路と、第2の流路を単位時間あたりに流れる空気の量を抑制するリブと、第1の流路を単位時間あたりに流れる空気の量を増加させる導風板を有していて、第2の流路から単位時間あたりに排出される空気の量が、第1の流路から単位時間あたりに排出される空気の量よりも小さくなるように調整されている。

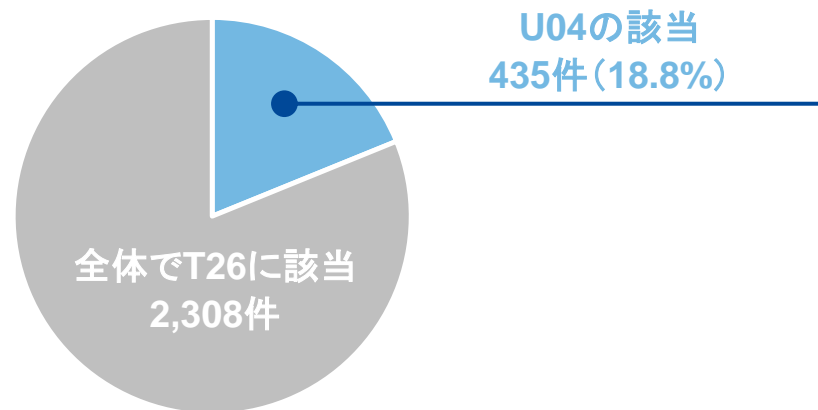
※対外説明用のため要約文は一部加工している

「U04.加湿」の用途は、「T26.放電式ミスト生成」の技術の応用先として高い関連性がありますが、出願人Nの保有するT26ではそれがありません

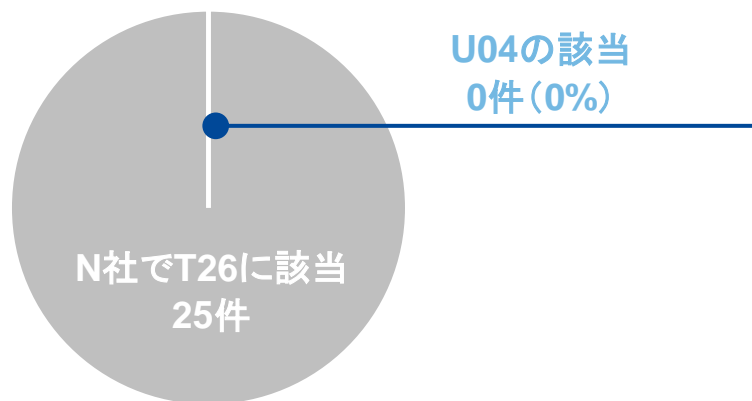
全体でのU04の該当割合



全体でのT26におけるU04の割合



出願人NのT26におけるU04の割合



考察

- ベイジアンネットワークのモデルでは、「T26.放電式ミスト生成」に対する「U04.加湿」の関係が見られた
- 全体では、U04の該当は6.3%だが、T26を条件としたときでは、その該当割合が18.8%となり高い関連性が認められる
- しかし、出願人Nでは、T26に該当する特許のうち、U04に該当する特許は1件もない
⇒N社の保有するT26はU04での用途も考えられる

プロジェクトの中で空気を流動させるためにイオン風を発生させる技術は、加湿装置の中で空気を加湿するために起こすイオン風にも応用できるかもしれません

T26がU04で応用されている例

発明の名称
加湿装置を備えた空気調和機
課題
加湿性能を向上させた加湿装置、及びその加湿装置を備えた空気調和機を提供する。
解決手段
導電性の電極と、電極の対向電極としての機能を有する吸水性加湿材と、電極に電圧を印加する電源と、吸水性加湿材に加湿水を供給する給水手段と、電極と吸水性加湿材との間の空間に形成される風路に空気を流す送風機と、を備え、電極に電圧を印加して電極から吸水性加湿材の面に対して法線方向へ向かうイオン風を発生させながら吸水性加湿材に当てて、風路の空気を加湿するものである。

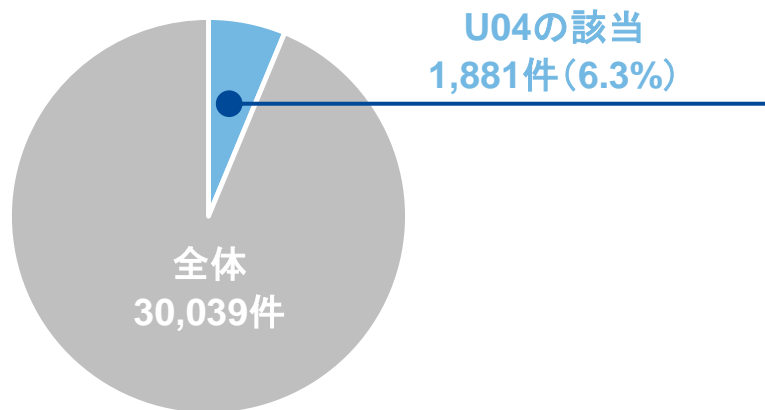
出願人Nの保有するT26の例

発明の名称
光源装置およびプロジェクター
課題
発光管の適切な温度調整を可能とし、構造の簡素化や騒音の発生を抑えた光源装置を提供する。
解決手段
光源装置は、光を射出する発光部と第2封止部と第1封止部を備える発光管と、発光部の一部を覆う副反射面と第1封止部を覆う延伸部を備える副反射鏡と、副反射面で反射した光を反射させる主反射鏡と、副反射鏡を挟んで設けられた第1電極と第2電極を有し、第1電極は、第2電極よりも発光部側にずらして配置され、電圧を印加することでイオン風を起こし、副反射面と発光部との間の空気を流動可能とする。

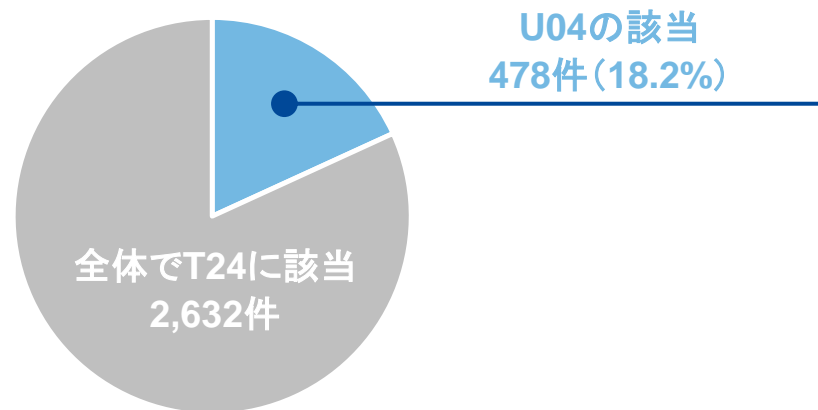
※対外説明用のため要約文は一部加工している

「U04.加湿」の用途は、「T24.温湿度制御と空気循環」の技術の応用先として高い関連性がありますが、出願人Wの保有するT24ではそれがほとんどありません

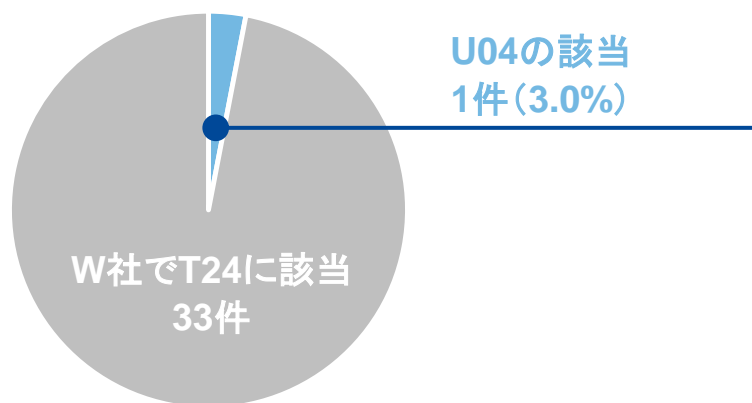
全体でのU04の該当割合



全体でのT24におけるU04の割合



出願人WのT24におけるU04の割合



考察

- ベイジアンネットワークのモデルでは、「T24.温湿度制御と空気循環」に対する「U04.加湿」の関係が見られた
- 全体では、U04の該当は6.3%だが、T24を条件としたときでは、その該当割合が18.2%となり高い関連性が認められる
- しかし、出願人Wでは、T24に該当する特許のうち、U04に該当する特許は1件だけである
⇒W社の保有するT24はU04での用途も考えられる

印刷機内の用紙の水分量を制御する加湿送風技術は、浴室内の快適なサウナ環境の実現にも応用できるかもしれません

T24がU04で応用されている例

発明の名称
空調システム
課題
少ない消費エネルギーで浴室の加熱および加湿を十分にいき、快適なサウナ環境を実現できる空調システムを提供する。
解決手段
冷媒を圧縮する圧縮機と冷媒が放熱する利用側熱交換器と冷媒を膨張させる膨張機構と冷媒が吸熱する熱源側熱交換器を接続した冷媒回路と、浴室に開口した吸込口から空気を吸い込んで浴室に開口した吹出口から空気を吹き出す循環ファンと、ミストを発生させるミスト発生器とを備え、熱源側熱交換器において冷媒が吸熱した熱を利用側熱交換器において循環ファンが送風する空気に対して放熱することによりヒートポンプを作動させて浴室内を加熱するとともにミスト発生器で発生したミストを浴室内に供給することにより浴室内を加湿して高温高湿のサウナ環境を実現する。

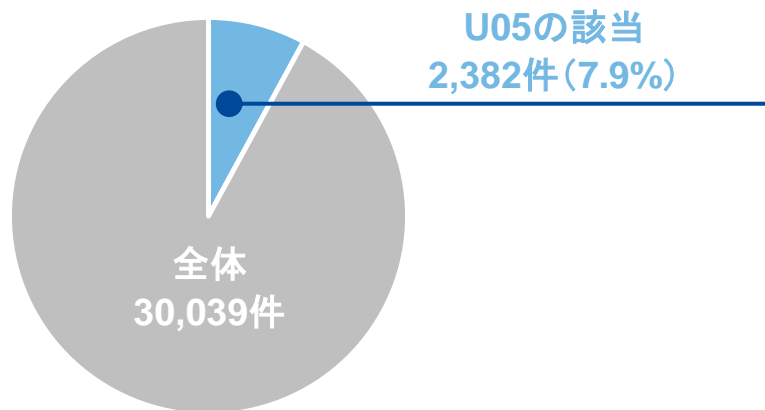
出願人Wの保有するT24の例

発明の名称
シーズニング装置、画像形成装置
課題
画像が形成された記録媒体(用紙)に生じるコックリング(波打ち)を抑制することができるシーズニング装置及び画像形成装置を得る。
解決手段
装置本体から排出された用紙は、吹付装置によって加湿された空気が吹き付けられる(加湿シーズニング)。吹付装置によって吹き付けられた加湿された空気は、用紙間に吹き込まれ、非描画部に水分を供給する。加湿された空気を5分間吹き付けた後は、水管の噴出口が閉止され、加湿されない空気を用紙に10分間吹き付ける(送風シーズニング)。これにより、非描画部の水分量が、用紙が元々含んでいた水分量にまで戻るため、描画部と非描画部との水分量の差は小さくなり、用紙に生じるコックリング(波打ち)を抑制することができる。

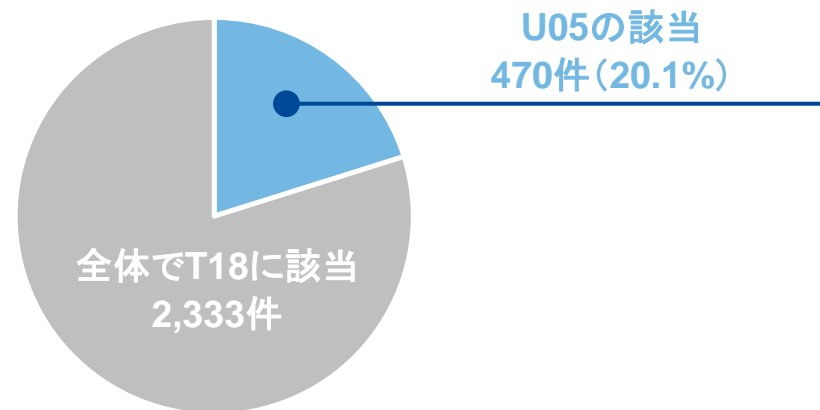
※対外説明用のため要約文は一部加工している

「U05.乾燥機能(衣類など)」の用途は、「T18.除湿」の技術の応用先として高い関連性がありますが、出願人Wの保有するT18ではそれがほとんどありません

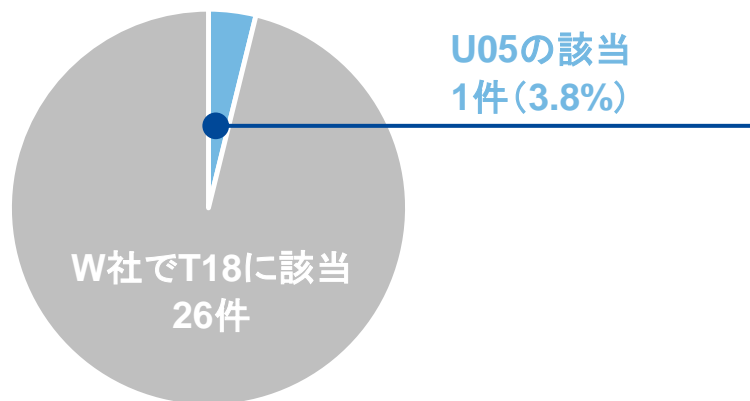
全体でのU05の該当割合



全体でのT18におけるU05の割合



出願人WのT18におけるU05の割合



考察

- ベイジアンネットワークのモデルでは、「T18.除湿」に対する「U05.乾燥機能(衣類など)」の関係が見られた
- 全体では、U05の該当は7.9%だが、T18を条件としたときでは、その該当割合が20.1%となり高い関連性が認められる
- しかし、出願人Wでは、T18に該当する特許のうち、U05に該当する特許は1件だけである
⇒W社の保有するT18はU05での用途も考えられる

印刷機の中でインク液のムラやインク液の吸収による用紙の波うちを防ぐ乾燥処理技術は、洗濯乾燥機の中で洗濯物をムラ無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

T18がU05で応用されている例

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。

出願人Wの保有するT18の例

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

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

- 文章を単語に分解し、その出現頻度を集計する
- 各文章に出現する単語情報をデータ化する(共起行列の作成)

単語抽出

Text Mining Studio

PLSA

確率的潜在意味解析

- 単語が出現する文脈を学習し、背後に潜むトピックを抽出する
- 全テキストデータをトピックで説明する(重みを計算する)

トピック抽出

Visual Mining Studio

ベイジアンネットワーク

- トピックを含むテキスト情報内の変数の関係構造をモデル化する
- 各変数が他の変数に与える影響を確率シミュレーションする

モデリング

BayoLink

膨大なテキストデータを人間が理解しやすい形に整理できる

テキストの内容における複雑な要因関係を構造化できる

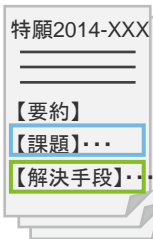
条件を変化させたときの結果の挙動をシミュレーションできる

ある事象の発生確率をコントロールする条件を発見できる

テキストマイニングに加え、トピックを抽出するPLSAと、そのトピックの関係をモデル化するベイジアンネットワークを適用することで、特許情報に潜む特徴を分かりやすく把握できました

用途と技術のトピック抽出

データの抽出



- 特許文書の要約文の「課題」と「解決手段」のテキストデータを抽出する
- 「課題」からは用途トピックを、「解決手段」からは技術トピックを抽出する

トピックの特徴集計

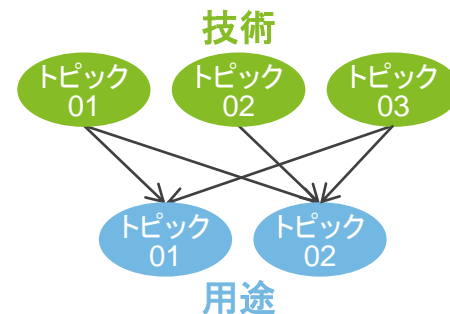
全特許データに各トピックのスコア（該年度）を計算する

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

用途と技術の関係分析

用途トピックと技術トピックの統計的な関係性をベイジアンネットワークでモデル化する

ベイジアンネットワーク



テキストマイニング

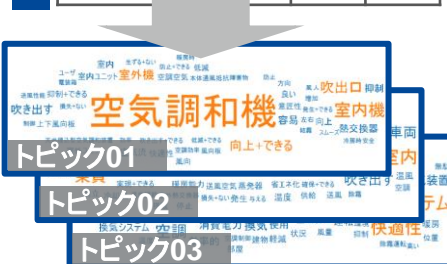
テキストマイニングを実行して単語と係り受け表現を抽出する

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
...

PLSA

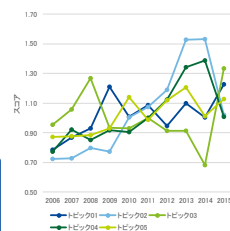
「単語×係り受け」の共起行列を作成し、これにPLSAを適用してトピックを抽出する

単語	係り受け			
	機提供	空気調和	効率-良い	掃除機-提供
空気調和機	1,578	100	1	
空気	85	144	45	
容易	190	105	67	



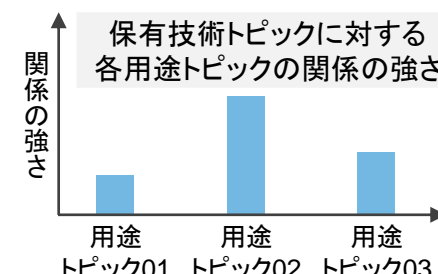
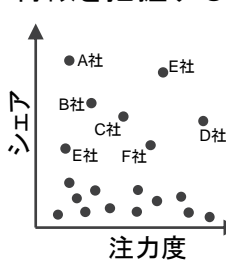
出願年集計

トピックスコアを出願年で集計してトピックのトレンドを把握する



出願人集計

トピックスコアを出願人で集計して、各トピックにおける出願人の特徴を把握する



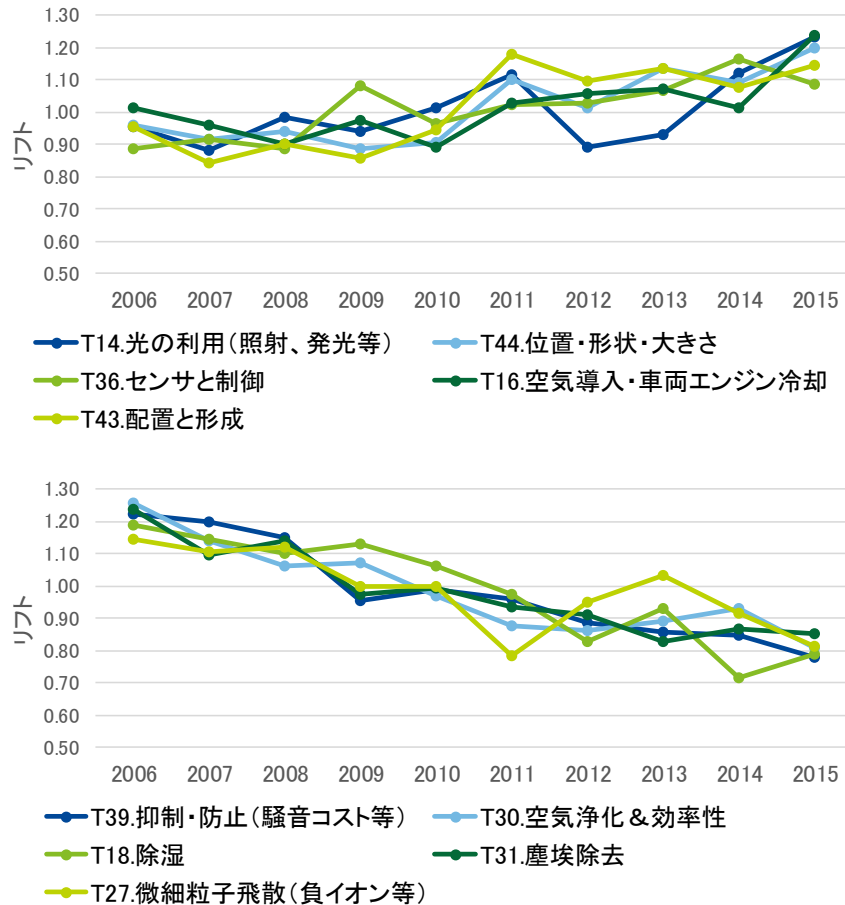
保有技術と関係のある用途トピックのうち、まだ想定していない用途を探索し、それに関連する元の特許文書を確認することで具体的な新規用途を検討する

用途トピックの抽出
技術トピックの抽出

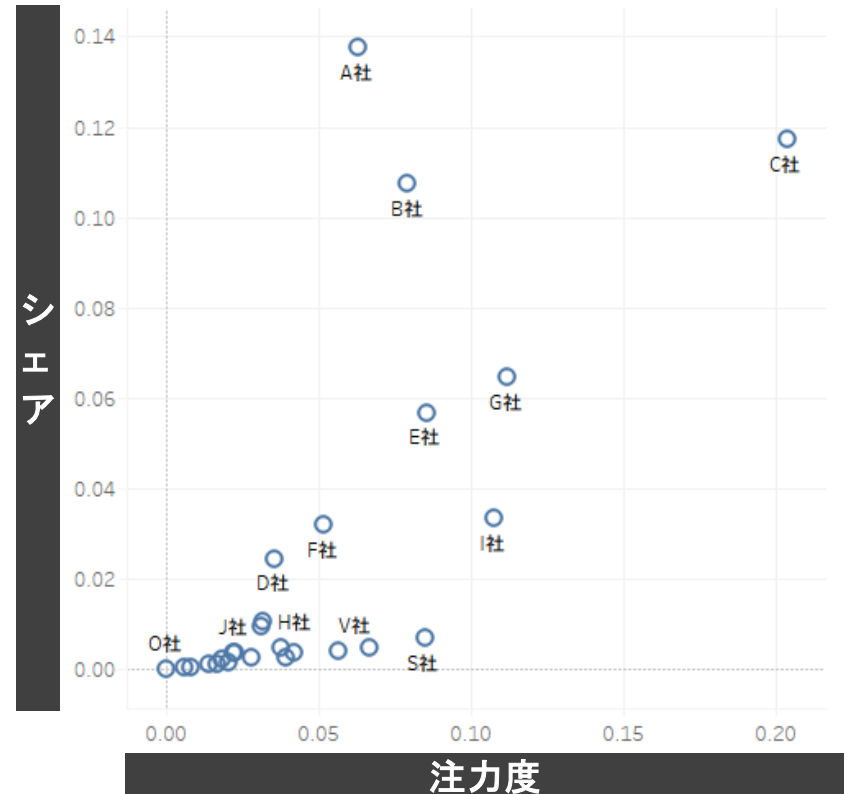
係り受け表現	頻度
空気調和機-提供	1,575
効率-良い	1,325
掃除機-提供	545
容易-構成	539
...	...

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

トピックをベースにしたトレンド分析



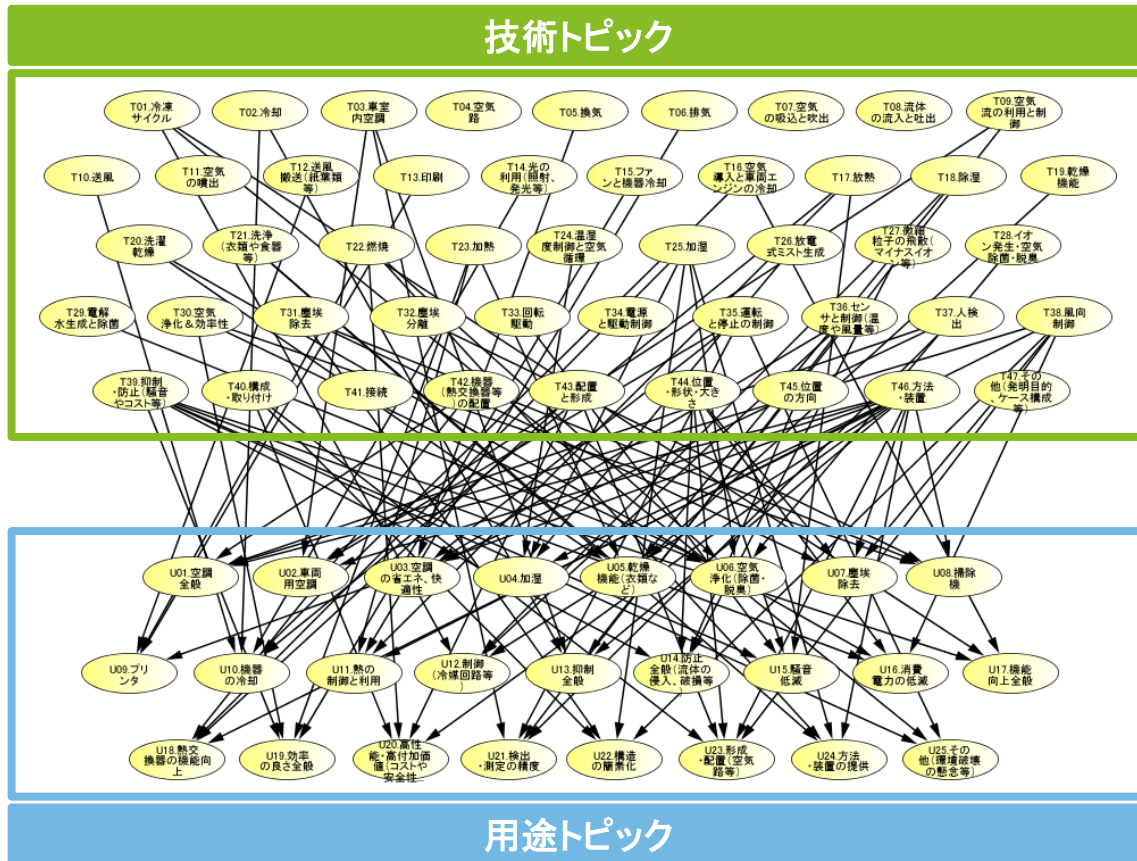
トピックをベースにした競合分析



用途と技術の統計的な関係を把握し、本来関連の強い用途でもそれを想定していないで出願されている特許技術を見つけ、技術の新規用途のアイデアを創出できます

技術トピックに対する用途トピックの関係分析

技術の新規用途探索



発明の名称	
空調システム	
課題	
少ない消費エネルギーで浴室の加熱および加湿を十分にを行い、快適なサウナ環境を実現できる空調システムを提供する。	
解決手段	
冷媒を圧縮する圧縮機と冷媒が放熱する利用側熱交換器と冷媒を膨張させる膨張機構と冷媒が吸熱する熱源側熱交換器を接続した冷媒回路と、浴室に開口した吸込口から空気を吸い込んで浴室に開口した吹出口から空気を吹き出す循環ファンと、ミストを発生させるミスト発生器とを備え、熱源側熱交換器において冷媒が吸熱した熱を利用側熱交換器において循環ファンが送風する空気に対して放熱することによりヒートポンプを動作させて浴室内を加熱するとともにミスト発生器で発生したミストを浴室内に供給することにより浴室内を加湿して高温高湿のサウナ環境を実現する。	

発明の名称	
シーズニング装置、画像形成装置	
課題	
画像が形成された記録媒体(用紙)に生じるコックリング(波打ち)を抑制することができるシーズニング装置及び画像形成装置を得る。	
解決手段	
装置本体から排出された用紙は、吹付装置によって加湿された空気が吹き付けられる(加湿シーズニング)。吹付装置によって吹き付けられた加湿された空気は、用紙間に吹き込まれ、非描画部に水分を供給する。加湿された空気を5分間吹き付けた後は、水管の噴出口が閉止され、加湿されない空気を用紙に10分間吹き付ける(送風シーズニング)。これにより、非描画部の水分量が、用紙が元々含んでいた水分量にまで戻するため、描画部と非描画部との水分量の差は小さくなり、用紙に生じるコックリング(波打ち)を抑制することができる。	

Nomolyticsは様々な業務のテキストデータに適用することができます



口コミ

- 顧客の関心トピックのターゲット別把握
- 顧客目線での製品や競合の比較分析
- 満足度向上の要因の把握
- 価値観を理解したマーケティング検討



アンケート

- 自由記述の内容をトピック化
- 自由記述トピックを変数として扱うことで定型設問回答と一緒に分析可能
- 話題を生む要因の把握



コールセンター履歴

- 問い合わせ内容をトピック化
- 製品別・顧客別の問い合わせ特徴把握
- 問い合わせトピック等の条件から解約確率をシミュレーション



特許文書

- 課題と技術のトピックのトレンド把握
- 競合他社の技術動向把握
- 課題と技術のトピックの関係モデル化による保有技術の新規用途探索



営業日報

- 営業活動内容のトピック化
- 営業活動トピック等の条件から成約確率をシミュレーション
- 成約要因を把握した効果的な営業教育



有価証券報告書

- 各企業の事業内容をトピック化
- 事業トピックとそのトレンド把握
- 各種IR指標と事業トピックの関係分析
- 定性情報からの企業分析、業界分析



エントリーシート

- 志望動機やPR文のトピック抽出
- 記述内容からの学生の分類・振り分け
- 記述内容と入社後成果の関係分析
- 効率的な人材発掘



診療記録

- 診療記録、看護記録のトピック化
- 生活習慣と病状の関係分析
- 治療内容とその経過の関係分析
- 定性情報を用いた効果的な診療支援



問題発生レポート

- 不具合やヒヤリハット等のトピック抽出
- 作業環境等の条件から問題の発生確率をシミュレーション
- 効果的な製品や作業環境の改善支援

資料に関するお問い合わせやコンサルティングのご相談は以下までお願いします。

analytics.office@analyticsdlab.co.jp

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

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

株式会社アナリティクスデザインラボ

