



Analytics Design Lab

2018年度 人工知能学会全国大会

確率的因果意味解析(PCSA)

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

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

2018年6月7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	5,000,000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

PLSAを用いた新しいテキストデータの分析手法の提案

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

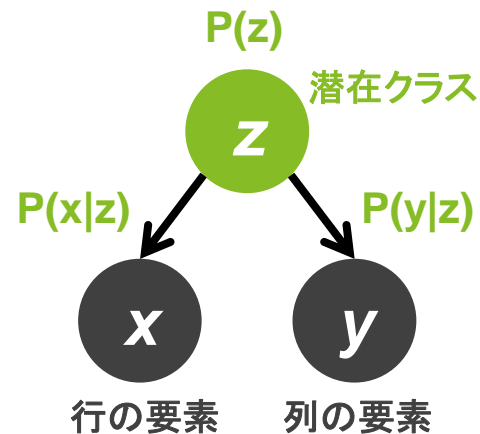
文書ID	単語 1	単語 2	単語 3	...	単語 5,014	単語 5,015
1	0	0	1		1	0
2	1	0	1		0	1
...						



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

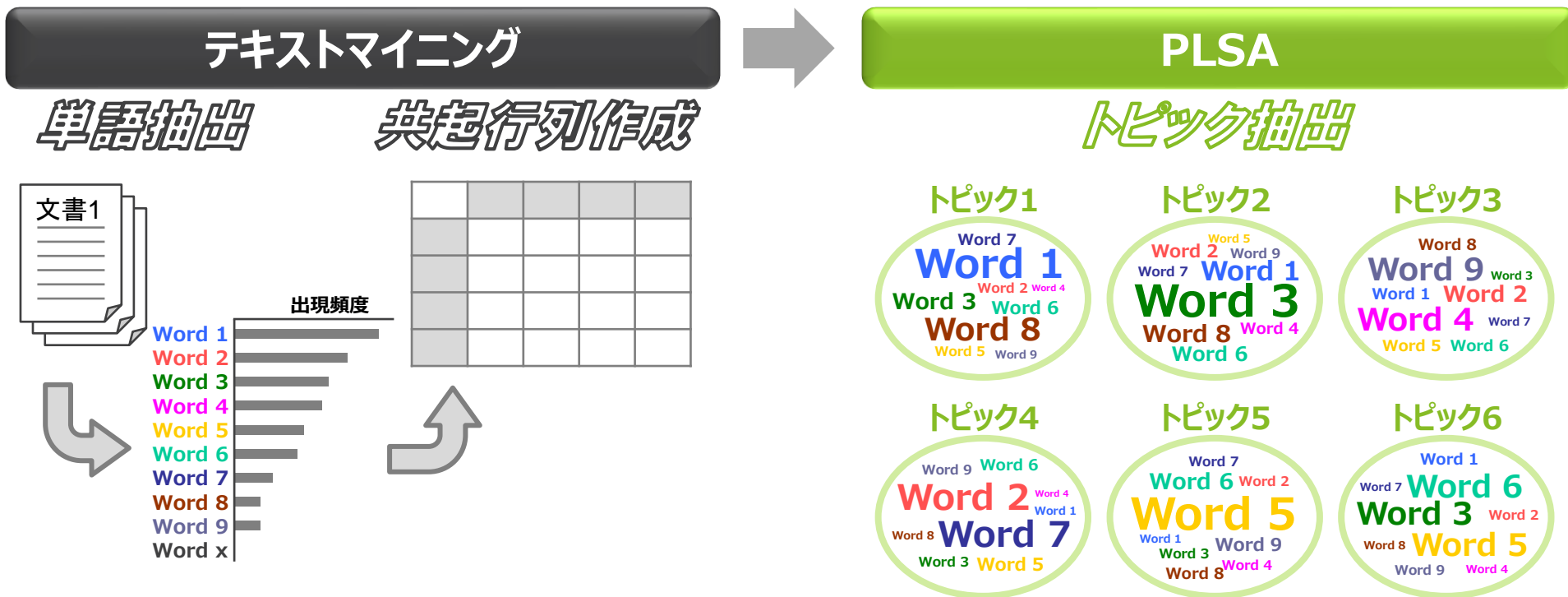
潜在クラスは行の要素と列の要素の2つの軸の変動量に基づいて抽出され、結果も2つの軸の情報から潜在クラスの意味を解釈することができる

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

【従来手法】PLSAを用いたトピックの抽出

テキストマイニングで抽出した単語で構成された共起行列にPLSAを適用して単語をトピックに集約することで、テキストデータの全体像をシンプルに把握します



テキストデータにテキストマイニングを実行して単語を抽出し、その単語の共起頻度を集計した共起行列を作成する

作成した共起行列にPLSAを適用し、単語をトピックに集約する(使われ方の似ている単語をその重みと共にまとめる)

PLSAで抽出したトピックを変数として扱い、他の属性情報との関係をベイジアンネットワークなどでモデリングすることで、その属性(ターゲット事象)に関連のあるトピックを探索できます

Nomolytics®: Narrative Orchestration Modeling Analytics

テキストマイニング

単語抽出



共起行列作成

PLSA

トピック抽出



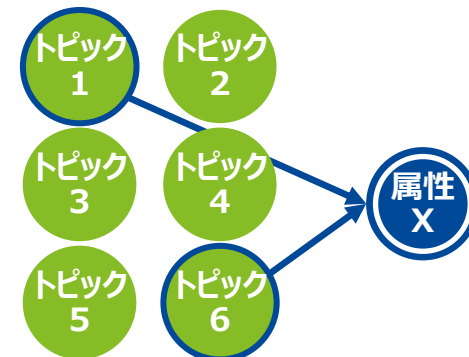
ベイジアンネットワーク

トピックの変数化

データID	属性 X	トピック1	トピック2	トピック3	...
1	有	1	0	1	
2	無	0	0	0	
3	無	0	1	1	
...					

全データに対する各トピックの該当有無{0,1}を計算する

モデリング



トピックと他の属性情報(ターゲット事象)との関係をモデル化する

例えば、口コミのコメントから得られたトピックと口コミの得点との関係をモデル化することで、口コミトピックの中から満足度向上に効果のあるものを探索できる

PLSAで抽出したトピックを変数として扱い、他の属性情報との関係をベイジアンネットワークなどでモデリングすることで、その属性(ターゲット事象)に関連のあるトピックを探索できます

Nomolytics®: Narrative Orchestration Modeling Analytics

テキストマイニング

単語抽出



共起行列作成

PLSA

トピック抽出



教師なし学習

ベイジアンネットワーク

トピックの変数化

データID	属性 X	トピック1	トピック2	トピック3	...
1	有	1	0	1	
2	無	0	0	0	
3	無	0	1	1	
...					

全データに対する各トピックの該当有無{0,1}を計算する

モデリング



トピックと他の属性情報(ターゲット事象)との関係をモデル化する

例えば、口コミのコメントから得られたトピックと口コミの得点との関係をモデル化することで、口コミトピックの中から満足度向上に効果のあるものを探索できる

ターゲット事象の該当データと非該当データからそれぞれ構築した2つの共起行列の差分にPLSAを適用することで、ターゲット事象に影響を与えるトピックに限定して抽出します

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

テキストマイニング

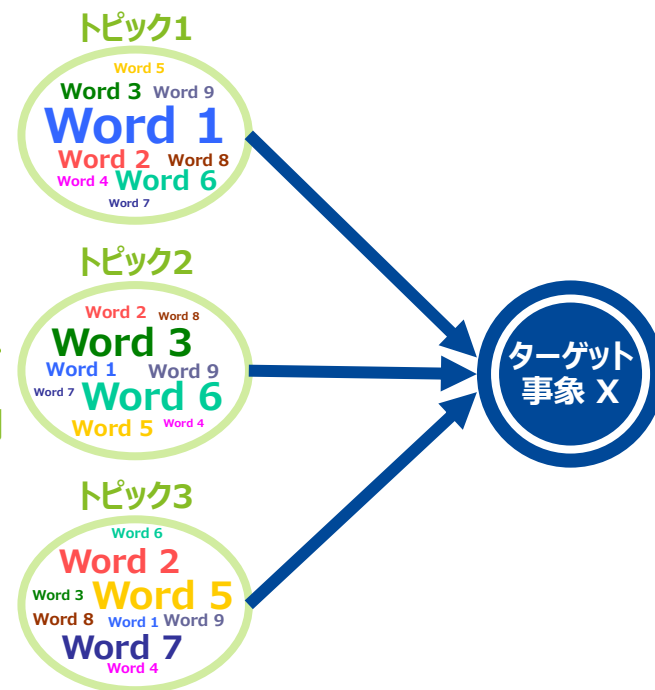
PLSA

単語抽出

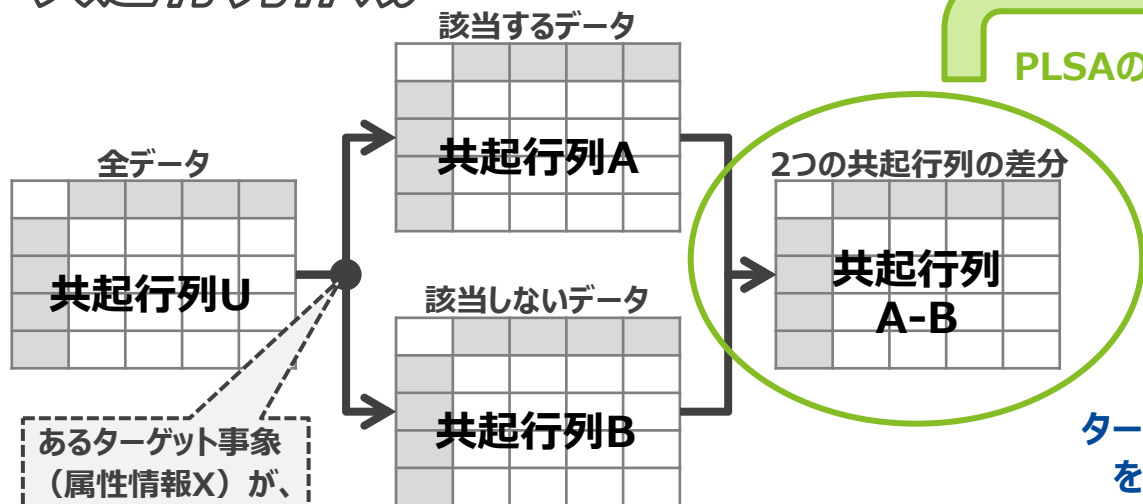


全データから構築した共起行列Uに対して、あるターゲット事象が該当するデータから構築した共起行列Aと、該当しないデータから構築した共起行列Bに分割し、その2つの共起行列の差分を計算した共起行列(A-B)に対してPLSAを適用する

トピック抽出



共起行列作成



PLSAの適用

ターゲット事象の発生有無に影響を与える潜在トピックを優先的にテキストデータから抽出することができる

ターゲット事象の該当データと非該当データからそれぞれ構築した2つの共起行列の差分にPLSAを適用することで、ターゲット事象に影響を与えるトピックに限定して抽出します

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

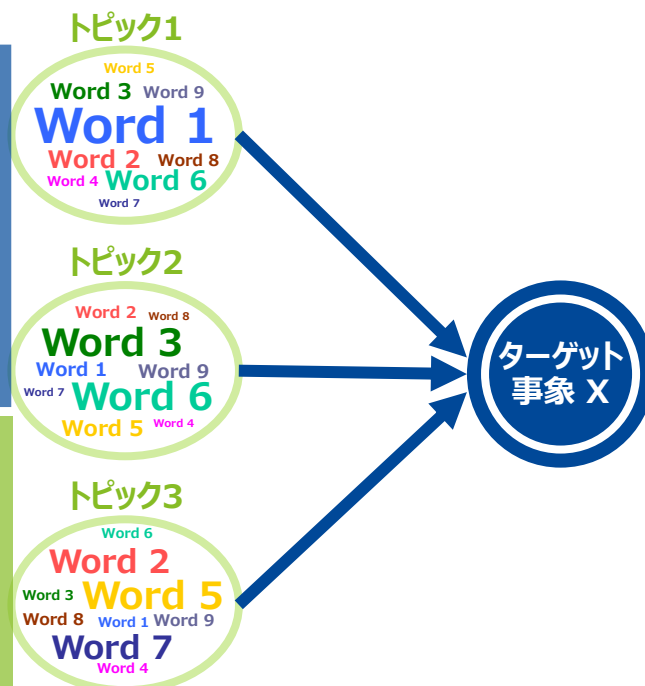
テキストマイニング

PLSA

単語抽出



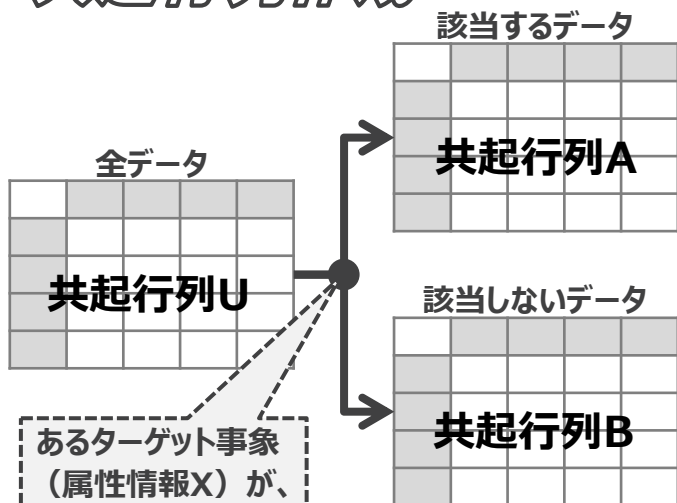
トピック抽出



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

教師あり

共起行列作成

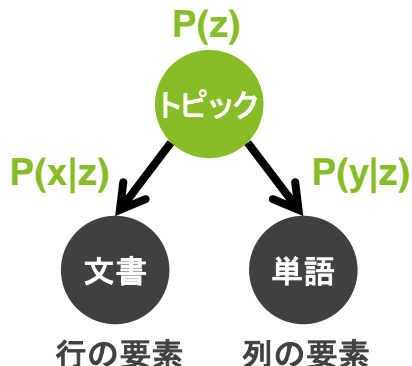


教師なし学習

ターゲット事象の発生有無に影響を与える潜在トピックを優先的にテキストデータから抽出することができる

PLSAのインプットとする共起行列の構成を「文書×単語」ではなく「単語×単語」とすることで、解釈のしやすいトピックが抽出でき、異なるデータの共起行列でも加法性が成り立ちます

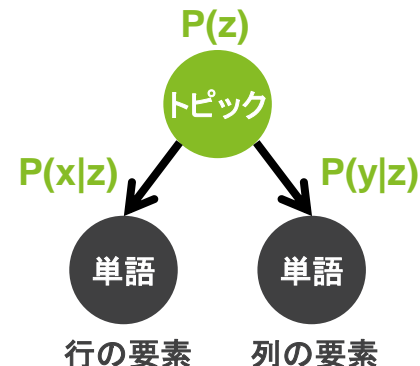
通常のPLSA



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

- 共起行列は「0」か「1」の2値で構成され、ほとんどが「0」となる疎なデータであるため、データ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
- データが異なれば共起行列の構成も異なる

本研究で適用するPLSA



	単語a	単語b	単語c	単語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	
...					

- 共起行列には具体的な頻度が入った密なデータであるため、データ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
- 行も列も単語で構成されている共起行列では、どちらもそれぞれ単独で意味を持つ情報となるため、両方の情報軸からトピックの意味を解釈することができ、解釈の容易性が高まる
- 異なるデータの共起行列でも、行列の単語の構成が共通していれば共起行列の加法・減法が適用できる

分析事例

「風」「空気」に関する10年分の特許データ30,039件を対象に、ターゲット事象を「出願年が2013年以後」に設定し、近年上昇傾向あるいは下降傾向にある技術トピックを抽出します

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



分析対象

■ 要約文の【解決手段】に記載されている文章

※上記書式で記載されていないものは要約文をそのまま使用

【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。

ターゲット事象

出願年が

- ①2012年以前 (22,387件)
- ②2013年以後 (7,652件)

①ターゲット事象との関係を考慮しないPLSAの実行

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、約3万件の特許の全体を構成している技術トピックを抽出します

テキストマイニングの実行

【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
配置	名詞	6,630
供給	名詞	3,692
内部	名詞	3,531
送風機	名詞	3,454
制御	形容詞	3,428
位置	名詞	3,185
吸い込む	名詞	3,099
吹出口	名詞	3,078
...

係り受け表現	頻度
空気⇒吸い込む	1,112
吸い込む⇒空気	889
連⇒通す	754
備える⇒構成	700
空気⇒吹き出す	689
吹出口⇒吹き出す	670
空気⇒供給	648
空気⇒送風	643
...	...

共起行列の作成

抽出した単語と係り受け表現に基づいて、「単語×係り受け表現」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

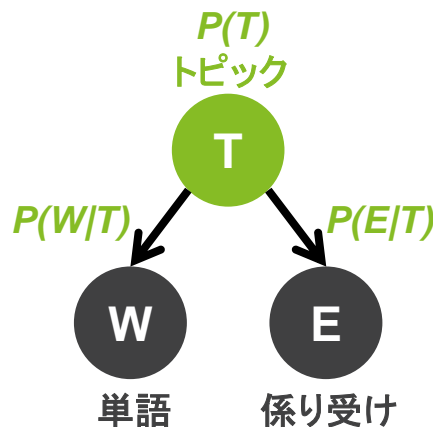
	係り受け表現				
	空気↓吸い込む	吸い込む↓空気	連↓通す	備える↓構成	⋮
単語					
配置	210	191	150	145	
供給	127	101	67	108	
内部	134	107	126	64	
送風機	225	197	112	95	
...					

単語: 5,187語 × 係り受け: 7,174表現 (頻度10件以上を対象)

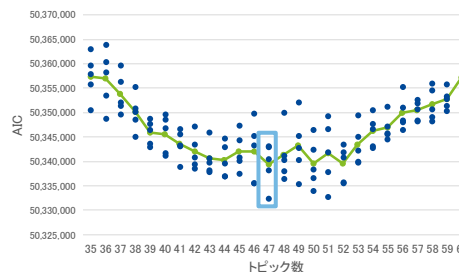
※「空気」「形成」「設ける」など、頻度が突出して多いものはノイズになりやすいため除外している

PLSAの実行

共起行列にPLSAを適用する



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



トピックの抽出

各トピックについて以下の3つの確率が計算される

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

トピックにおける $P(W|T)$ と $P(E|T)$ からトピックの意味を解釈する

トピックt32

$$P(t_{32}) = 2.7\%$$

確率	単語	確率	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
...

確率の高い構成要素から、t32のトピックは「塵埃の分離」に関する技術トピックと解釈できる

技術ピック47個の一覧①

空気の冷却や空気路、換気、機器の冷却、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、人の検出、構成や配置などの技術が47個抽出されました

<h3>t01.冷凍サイクル</h3> <p>凝縮器 冷媒回路 熱交換器 室内熱交換器 冷凍サイクル 蒸発器 圧縮機 冷媒 室外熱交換器 圧縮機 熱交換器</p>	<h3>t02.冷却</h3> <p>送風機 シークラス 冷却 蒸発器 冷却装置 冷却ファン 冷却装置 冷却装置</p>	<h3>t03.車室内空調</h3> <p>送風機 冷却 車室内空調装置 吹き出す 車室内 冷却 送風機 冷却装置</p>	<h3>t04.空気路</h3> <p>空気路 開口 通過 連 空間 開口 通す 連通</p>	<h3>t05.換気</h3> <p>換気 供給 換気装置 換気装置 換気装置 換気装置 換気装置 換気装置</p>
<h3>t06.排気</h3> <p>排気 外部 内部 排気口 排気口 排気口 排気口 排気口 排気口</p>	<h3>t07.空気の吸込と吹出</h3> <p>吸込 吹出 吸込口 吹出口 吸込口 吹出口 吸込口 吹出口</p>	<h3>t08.流体の流入と吐出</h3> <p>流入 吐出 流入 吐出 流入 吐出 流入 吐出</p>	<h3>t09.空気流の利用と制御</h3> <p>空気流 利用 制御 空気流 利用 制御 空気流 利用 制御</p>	<h3>t10.送風</h3> <p>送風 送風機 送風機 送風機 送風機 送風機 送風機</p>
<h3>t11.空気の噴出</h3> <p>噴出 空気流 噴出 噴出 噴出 噴出 噴出</p>	<h3>t12.送風搬送(紙葉類等)</h3> <p>送風 搬送 送風管 搬送 搬送 搬送 搬送</p>	<h3>t13.印刷</h3> <p>印刷 搬送 印刷 印刷 印刷 印刷 印刷</p>	<h3>t14.光の利用(照射、発光等)</h3> <p>光 照射 発光 照射 照射 照射 照射</p>	<h3>t15.ファンと機器冷却</h3> <p>ファン 冷却 機器冷却 機器冷却 機器冷却 機器冷却</p>
<h3>t16.空気導入と車両エンジンの冷却</h3> <p>空気導入 冷却 冷却ファン 冷却ファン 冷却ファン 冷却ファン</p>	<h3>t17.放熱</h3> <p>放熱 放熱 放熱 放熱 放熱 放熱 放熱</p>	<h3>t18.除湿</h3> <p>除湿 除湿 除湿 除湿 除湿 除湿 除湿</p>	<h3>t19.乾燥機能</h3> <p>乾燥 乾燥 乾燥 乾燥 乾燥 乾燥 乾燥</p>	<h3>t20.洗濯乾燥</h3> <p>洗濯 乾燥 洗濯 乾燥 洗濯 乾燥 洗濯</p>
<h3>t21.洗浄(衣類や食器等)</h3> <p>洗浄 洗浄 洗浄 洗浄 洗浄 洗浄 洗浄</p>	<h3>t22.燃焼</h3> <p>燃焼 燃焼 燃焼 燃焼 燃焼 燃焼 燃焼</p>	<h3>t23.加熱</h3> <p>加熱 加熱 加熱 加熱 加熱 加熱 加熱</p>	<h3>t24.温湿度制御と空気循環</h3> <p>温湿度 制御 空気循環 空気循環 空気循環 空気循環</p>	<h3>t25.加湿</h3> <p>加湿 加湿 加湿 加湿 加湿 加湿 加湿</p>

空気の冷却や空気路、換気、機器の冷却、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、人の検出、構成や配置などの技術が47個抽出されました

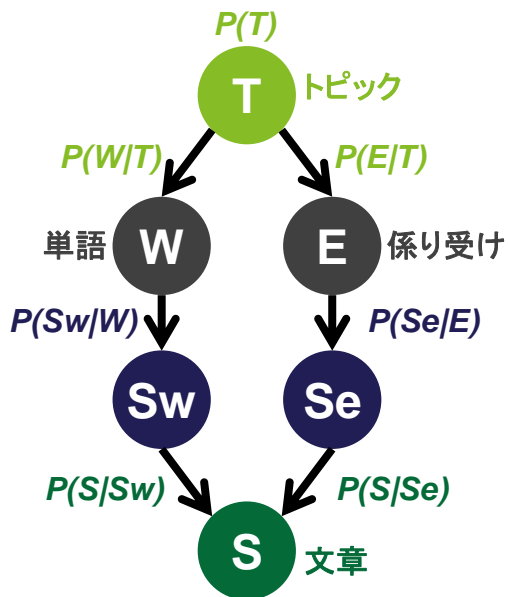
<h3>t26.放電式ミスト生成</h3> <p>発生 高電圧 水 静電霧化装置 水分 印加 放電電極 放出 生成 電極 表面 噴霧</p>	<h3>t27.微細粒子の飛散 (マイナスイオン等)</h3> <p>発生 高電圧 水 静電霧化装置 水分 印加 放電電極 放出 生成 電極 表面 噴霧</p>	<h3>t28.イオン発生・空気除菌・脱臭</h3> <p>放出 送風路 イオン発生装置 イオン発生 送風機</p>	<h3>t29.電解水生成と除菌</h3> <p>電解水生成 電解槽 電極 電圧 電流</p>	<h3>t30.空気清浄 & 効率性</h3> <p>効率 送風機 長い 空気清浄機 吸い込む</p>
<h3>t31.塵埃除去</h3> <p>付着 空気調和機 上流 フィルタ 塵埃 除去 捕集</p>	<h3>t32.塵埃分離</h3> <p>塵埃 分離 送風機 掃除機 吸い込む</p>	<h3>t33.回転駆動</h3> <p>回転 駆動 軸 羽根車</p>	<h3>t34.電源と駆動制御</h3> <p>電力 インバータ 駆動 制御 モータ</p>	<h3>t35.運転と停止の制御</h3> <p>運転 制御 停止 温度 制御</p>
<h3>t36.センサと制御 (温度や風量等)</h3> <p>制御 温度 センサ 制御 動作 制御</p>	<h3>t37.人検出</h3> <p>検出 室内機 空気調和機</p>	<h3>t38.風向制御</h3> <p>風向 制御 吹出口</p>	<h3>t39.抑制・防止 (騒音やコスト等)</h3> <p>抑制 防止 騒音 低減</p>	<h3>t40.構成・取り付け</h3> <p>構成 取り付け 開口部</p>
<h3>t41.接続</h3> <p>接続 送風機 一端</p>	<h3>t42.機器(熱交換等)の配置</h3> <p>配置 熱交換器 空気調和機</p>	<h3>t43.配置と形成</h3> <p>配置 形成 送風機</p>	<h3>t44.位置・形状・大きさ</h3> <p>位置 形状 大きさ 向き</p>	<h3>t45.位置の方向</h3> <p>位置 方向 向き</p>
<h3>t46.方法・装置</h3> <p>方法 装置 送風機</p>	<h3>t47.その他(発明目的、ケース構成等)</h3> <p>目的 構成 ケース</p>			

各データに対するトピックの該当有無の計算

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した

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



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

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)
$P(Sw_h W_i) = 1/n(W_i)$
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)
$P(Se_h E_j) = 1/n(E_j)$
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
$P(S_h T_k) = P(S_h Sw_h)P(Sw_h T_k) + P(S_h Se_h)P(Se_h T_k)$
文章 S_h の出現確率
$P(S_h) = \sum_k P(S_h T_k)P(T_k)$

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		1.6
...						

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		1.6
...					

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		0
...					

トピックと出願年の関係

トピックと出願年の関係を示す指標値(リフト値)を①2012年以前②2013年以後で計算し、2013年前後での増減率を集計したところ、高いものから低いものまでばらついています

リフト値の増減率	2012年以前のリフト値	2013年以後のリフト値	トピック
19.3%	0.95	1.14	t44.位置・形状・大きさ
17.3%	0.96	1.12	t09.空気流の利用と制御
16.1%	0.96	1.12	t43.配置と形成
15.0%	0.96	1.11	t36.センサと制御(温度や風量等)
14.0%	0.97	1.10	t38.風向制御
13.5%	0.97	1.10	t16.空気導入と車両エンジンの冷却
12.9%	0.97	1.09	t14.光の利用(照射、発光等)
12.2%	0.97	1.09	t45.位置の方向
11.3%	0.97	1.08	t10.送風
10.6%	0.97	1.08	t03.車室内空調
9.6%	0.98	1.07	t37.人検出
8.9%	0.98	1.07	t41.接続
7.9%	0.98	1.06	t40.構成・取り付け
7.4%	0.98	1.05	t07.空気の吸込と吹出
6.5%	0.98	1.05	t33.回転駆動
6.3%	0.98	1.05	t34.電源と駆動制御
3.9%	0.99	1.03	t04.空気路
3.7%	0.99	1.03	t17.放熱
3.3%	0.99	1.02	t01.冷凍サイクル
3.1%	0.99	1.02	t08.流体の流入と吐出
2.5%	0.99	1.02	t25.加湿
1.9%	1.00	1.01	t11.空気の噴出
0.8%	1.00	1.01	t35.運転と停止の制御
0.4%	1.00	1.00	t42.機器(熱交換器等)の配置
0.0%	1.00	1.00	t06.排気

※リフト値の増減率の高い順に並べている

リフト値の増減率	2012年以前のリフト値	2013年以後のリフト値	トピック
-1.1%	1.00	0.99	t46.方法・装置
-1.3%	1.00	0.99	t22.燃焼
-1.6%	1.00	0.99	t23.加熱
-2.2%	1.01	0.98	t02.冷却
-2.3%	1.01	0.98	t47.その他(発明目的、ケース構成等)
-3.8%	1.01	0.97	t21.洗浄(衣類や食器等)
-4.4%	1.01	0.97	t13.印刷
-4.8%	1.01	0.96	t05.換気
-5.0%	1.01	0.96	t15.ファンと機器冷却
-6.4%	1.02	0.95	t12.送風搬送(紙葉類等)
-7.3%	1.02	0.94	t29.電解水生成と除菌
-9.8%	1.03	0.93	t32.塵埃分離
-9.9%	1.03	0.92	t27.微細粒子の飛散(マイナスイオン等)
-15.5%	1.04	0.88	t24.温湿度制御と空気循環
-15.7%	1.04	0.88	t30.空気浄化&効率性
-17.9%	1.05	0.86	t20.洗濯乾燥
-19.5%	1.05	0.85	t31.塵埃除去
-20.7%	1.06	0.84	t28.イオン発生・空気除菌・脱臭
-21.3%	1.06	0.83	t39.抑制・防止(騒音やコスト等)
-22.0%	1.06	0.83	t19.乾燥機能
-24.0%	1.07	0.81	t18.除湿
-27.4%	1.07	0.78	t26.放電式ミスト生成

トピックと出願年の関係を示すリフト値

※出願年を①2012年以前、②2013年以後として計算

$$P(\text{出願年}Y | \text{トピック}T=1)$$

$$P(\text{出願年}Y)$$

②ターゲット事象との関係を考慮したPCSAの実行

トピック抽出のプロセス

① 2012年以前データ、② 2013年以後データで作成した2つの共起行列の差分にPLSAを適用することで、2013年前後の出願に傾向のある技術トピックを優先的に抽出します

2つの共起行列の作成

① 出願年が**2012年以前**のデータ
22,387件 (文章数: 33,283件)

係り受け表現

	空気↓吸い込む	吸い込む↓空気	連↓通す	備える↓構成	∴
配置	144	139	109	102	
供給	95	65	53	81	
内部	93	64	92	44	
送風機	145	134	79	76	
...					

② 出願年が**2013年以後**のデータ
7,652件 (文章数: 11,831件)

係り受け表現

	空気↓吸い込む	吸い込む↓空気	連↓通す	備える↓構成	∴
配置	66	52	41	43	
供給	32	36	14	27	
内部	41	43	34	20	
送風機	80	63	33	19	
...					

差分の共起行列の作成

① 2012年までの共起行列と
② 2013年からの共起行列の
差の絶対値を計算した共起行列
を作成する

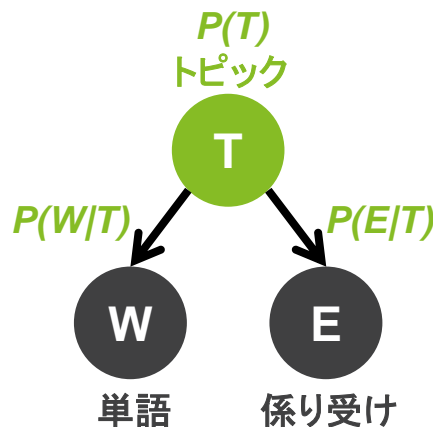
係り受け表現

	空気↓吸い込む	吸い込む↓空気	連↓通す	備える↓構成	∴
配置	14.8	2.6	2.3	6.7	
供給	1.8	12.9	4.8	1.8	
内部	7.9	20.3	1.3	4.4	
送風機	28.5	15.4	4.9	8.0	
...					

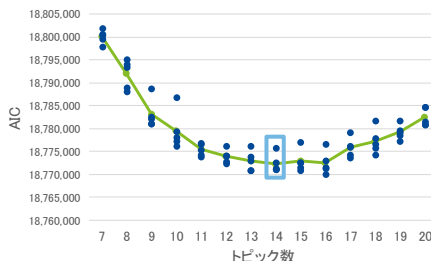
ただし、2つの共起行列は異なる文章数のデータから作成されているので、①2012年までの共起行列の頻度を2つの文章数の比率(11,831/33,283)で重み調整してから、②2013年からの共起行列との差を計算する

PLSAの実行

差分の共起行列にPLSAを適用する



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



トピックの抽出

各トピックについて以下の3つの確率が計算される

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

トピックにおける $P(W|T)$ と $P(E|T)$ からトピックの意味を解釈する

トピックT10

$P(T10) = 6.3\%$

確率	単語	確率	係り受け
3.4%	塵埃	0.9%	付着-塵埃
1.8%	送風機	0.8%	塵埃-除去
1.7%	掃除機	0.8%	吸い込む-塵埃
1.6%	吸い込む	0.7%	塵埃-含む
1.3%	分離	0.7%	塵埃-吸い込む
1.3%	フィルタ	0.7%	塵埃-分離
1.3%	捕集	0.6%	発生-送風機
1.0%	集塵部	0.6%	含む-空気
...

確率の高い構成要素から、T10のトピックは「塵埃の分離」に関する技術トピックと解釈できる

技術ピック14個の一覧

冷凍サイクルや冷却ファン、衣類乾燥、燃焼、イオン発生、電解水の生成、塵埃の分離、羽の回転、検出と制御、車内空調の配置などの技術が14個抽出されました

<h3>T01.冷凍サイクル</h3> <p>温度圧縮 送風 吐出熱 膨張弁 接続 切り替える 車室内 冷媒回路 冷却 放熱器 蒸発器 熱交換器 室内熱交換器 圧縮機 減圧室 室外熱交換器 熱交換 吸い込む 冷凍サイクル 冷媒 停止 流れる 凝縮器 循環制御 放熱 加熱 通過</p>	<h3>T02.空気の冷却</h3> <p>通す 室内 屋外熱交換 外部 配設 内部冷却 外気 送風機 空気風路 排気風路 吸い込む 排気口 吸込口 通過 排出 取り込む 冷却装置 ケース排気 吹き出す 吹出口 循環設置 開口 供給 連通</p>	<h3>T03.冷却ファン</h3> <p>光源 加熱 載置 熱 吸い込む 取り込む 下方 ファン 外気筐体内排気口光 収納 下方 ファン 配置 供給 冷却ファン 筐体 吸気口 冷却 電気部品 外部上方 ヒートシンク 冷却風 冷却風 内部送風位置 収容 覆う 排出</p>	<h3>T04.空気流 (吸込と吹出)</h3> <p>風向ける室内 熱交換 回転 下方取り付ける 空気流 送風機 送風機 送風機 送風機 送風機 熱交換器 送風ファン 室内空気 筐体内 吸込口 吸い込む 吹出口 空気調和機 上下風向板 室内機 吹き出す 空気吸込口 空気風路 室内機 吹き出す 風向板 ファン 方向 上下方向 筐体 位置 配置</p>	<h3>T05.紙葉類の搬送</h3> <p>投入下流 開く 移動 到達用紙 センサ送り込む 定着部 制御部 加熱送風部 定着装置 トナー像 シート 送風管 画像形成装置 検出 搬送 空気流 通過 上流 吹き付ける 制御 定着開閉弁 紙葉類生ずる 閉じる 接続部</p>
<h3>T06.衣類乾燥</h3> <p>内部制御 配設 排気口 送風ファン 水槽 収容 ヒートシンク 排出加熱手段 衣類 ヒートポンプ装置 放熱器 回転ドラム 乾燥用空気 送風 循環風路 乾燥運転 洗濯乾燥機 乾燥室 送風手段 衣類乾燥機 洗濯物 加熱外槽 循環除湿 接続ドラム 乾燥</p>	<h3>T07.空気の燃焼</h3> <p>水 排ガス 向上 熱風 温度 生成 外部 発生 使用 提供 圧縮空気 送風機 燃焼 排出燃焼用空気 防止 供給 装置 加熱 噴出 燃焼室 良い 燃焼ガス バーナ 混合 燃料 放出熱 ノズル 乾燥</p>	<h3>T08.イオン発生</h3> <p>付着 放電高電圧 電圧 生成 オゾン帯電 保持 吹出口 送風 対向電極 分解送風路 空気中 電極 発生 対向電極 分解送風路 空気中 空気風路 イオン発生装置 静電霧化装置 臭気成分 印加 供給 放電電極 吸込口 配置 イオン 放出 水分 空気清浄装置 ミスト</p>	<h3>T09.電解水の生成</h3> <p>室内 貯留 貯水室 貯水部 下方 ケース 配置 接触 生成 加湿手段 供給 貯水 電解水 送風手段 送風ファン 加湿 加湿装置 水 給水タンク 加湿フィルタ 空気風路 吹出口 送風機 送風 通過 吸込口 水槽 上方 回転浸漬</p>	<h3>T10.塵埃の分離</h3> <p>開口排出 上流 捕集 流入口 配置 通過 下方 分離 含塵空気 送風機 集塵部 集塵容器 連 塵埃 除去 吸い込む 接続フィルタ 集塵装置 吸込口 掃除機 着脱自在 集塵室 吸引力 付着 旋回流 流入 発生 連通</p>
<h3>T11.羽の回転</h3> <p>連結ファン 空気流 向かう 中心 沿う 吸込口 周囲 一端ケーシング モータ 羽根 羽根車 固定 軸方向 対向 回転軸 ロータ 取り付ける 回転 ファンケース 径方向 外周 外周側 配置 固定子 支持位置 回転+できる 内部</p>	<h3>T12.検出と制御</h3> <p>測定動作 開始 記憶温度センサ 算出 決定 人 駆動 検出 制御手段 温度 運転 判定 空気調和機 基づく 制御部 制御 回転数 制御装置 設定 風量 停止 応ずる 所定 室内出力 湿度 取得 状態</p>	<h3>T13.車両用空調の配置</h3> <p>流れ 下方 後方 冷却風 車両 送風 導入 開口 吹出口 送風機 延びる 配置 吹き出す 車室内 車両用空調装置 側空気風路 導く 向ける 下流ダクト 方向 流れる 沿う 向かう 位置 内部 上流 通過 上方</p>	<h3>T14.配置と形成</h3> <p>挿入 平行 背面 突出 底部 覆う 固定 端部 収納 配置 沿う 方向 傾斜 延びる 上方 位置 熱交換器 内部 取り付け 開口部 下方 下面 互い 仕切板 開口 対向 側面 フィン 間隔 前面</p>	

トピックと出願年の関係

トピックと出願年の関係を示す指標値(リフト値)を①2012年以前②2013年以後で計算し、2013年前後での増減率を集計したところ、増減率の高いものと低いものに集中しています

リフト値の増減率	2012年以前のリフト値	2013年以後のリフト値	トピック
48.5%	0.89	1.32	T13.車両用空調の配置
30.3%	0.93	1.21	T14.配置と形成
23.9%	0.94	1.17	T04.空気流(吸込と吹出)
17.7%	0.96	1.13	T12.検出と制御
14.9%	0.96	1.11	T11.羽の回転
13.1%	0.97	1.09	T03.冷却ファン
11.6%	0.97	1.08	T02.空気の冷却
7.0%	0.98	1.05	T01.冷凍サイクル
-0.3%	1.00	1.00	T05.紙葉類の搬送
-5.9%	1.02	0.96	T07.空気の燃焼
-8.8%	1.02	0.93	T09.電解水の生成
-17.0%	1.05	0.87	T10.塵埃の分離
-24.2%	1.07	0.81	T06.衣類乾燥
-28.7%	1.08	0.77	T08.イオン発生

※リフト値の増減率の高い順に並べている

全体の共起行列で抽出したトピックでは47個あるうち、増減率が高いものから低いものまでばらついていて

リフト値の増減率	2012年以前のリフト値	2013年以後のリフト値	トピック
19.3%	0.95	1.14	t44.位置・形状・大きさ
17.3%	0.96	1.12	t09.空気流の利用と制御
16.1%	0.96	1.12	t43.配置と形成
15.0%	0.96	1.11	t36.センサと制御(温度や風量等)
14.0%	0.97	1.10	t38.風向制御
13.5%	0.97	1.10	t16.空気導入と車両エンジンの冷却
...			
0.8%	1.00	1.01	t35.運転と停止の制御
0.4%	1.00	1.00	t42.機器(熱交換器等)の配置
0.0%	1.00	1.00	t06.排気
-1.1%	1.00	0.99	t46.方法・装置
-1.3%	1.00	0.99	t22.燃焼
-1.6%	1.00	0.99	t23.加熱
...			
-19.5%	1.05	0.85	t31.塵埃除去
-20.7%	1.06	0.84	t28.イオン発生・空気除菌・脱臭
-21.3%	1.06	0.83	t39.抑制・防止(騒音やコスト等)
-22.0%	1.06	0.83	t19.乾燥機能
-24.0%	1.07	0.81	t18.除湿
-27.4%	1.07	0.78	t26.放電式ミスト生成

トピックと出願年の関係を示すリフト値

※出願年を①2012年以前、②2013年以後として計算

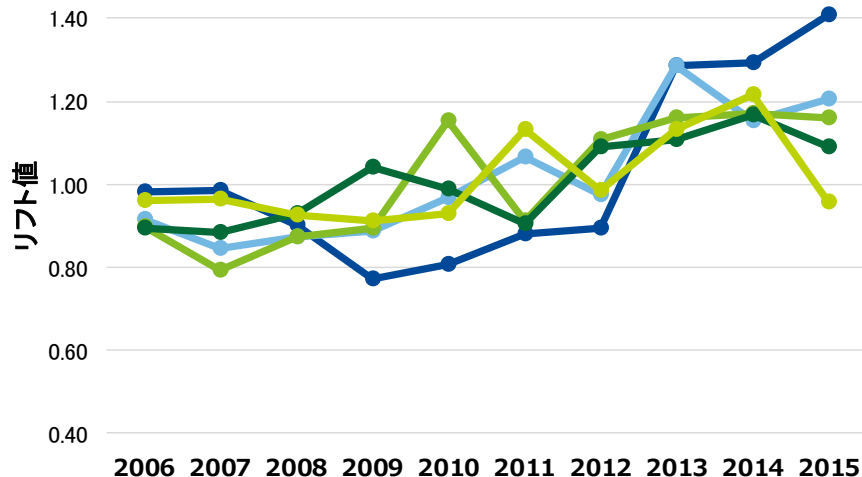
$$P(\text{出願年}Y | \text{トピック}T=1)$$

$$P(\text{出願年}Y)$$

トピックのトレンド

各出願年とのリフト値を集計して全体でのトレンドを見ても、上昇トレンド、下降トレンドを形成するトピックが抽出されていることが分かります

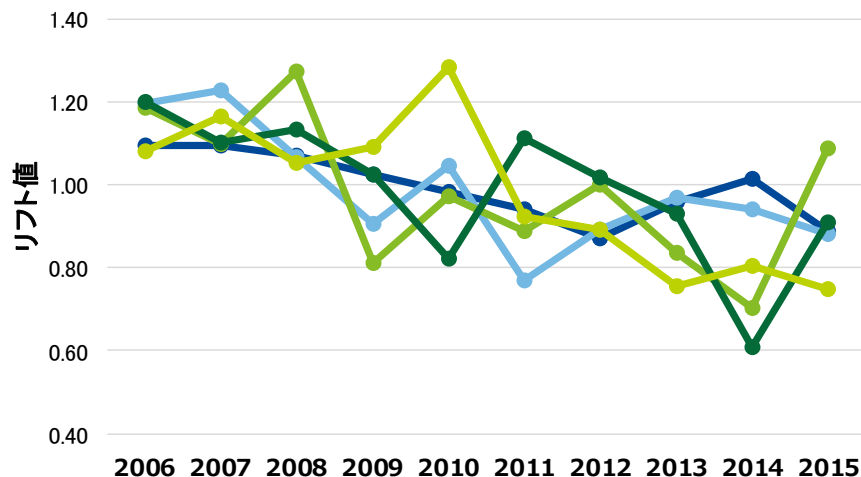
増減率上位5位のトピックのトレンド



2013年前後の増減率

T13.車両用空調の配置	48.5%
T14.配置と形成	30.3%
T04.空気流(吸込と吹出)	23.9%
T12.検出と制御	17.7%
T11.羽の回転	14.9%

増減率下位5位のトピックのトレンド



2013年前後の増減率

T07.空気の燃焼	-5.9%
T09.電解水の生成	-8.8%
T10.塵埃の分離	-17.0%
T06.衣類乾燥	-24.2%
T08.イオン発生	-28.7%

増減率ベスト1位は「車両用空調の配置」に関する技術、ワースト1位は「イオン発生」に関する技術で、それぞれのトピックに該当する特許の要約文の例を確認します

「T13.車両用空調の配置」の特許要約例

発明の名称	出願年
車両用空調装置	2015
課題	
<p>簡単な構成で、除湿暖房時における車両の燃費を向上させることができる車両用空調装置を提供する。</p>	
解決手段	
<p>冷却用熱交換器と、加熱用熱交換器と、冷却用熱交換器から加熱用熱交換器に至るヒータ通過経路と、加熱用熱交換器を迂回するヒータ迂回経路への空調風の流入割合を調整するエアミックスドアと、を備える車両用空調装置である。車両用空調装置は、加熱用熱交換器よりも空調風の流れの上流側で、かつ加熱用熱交換に面する位置に不動に設けられる除湿部材と、除湿部材の車両上方側を覆うように配置され、除湿部材の近傍から車両上方側への空調風の流れを制限する対流制限板と、を備える。</p>	

「T08.イオン発生」の特許要約例

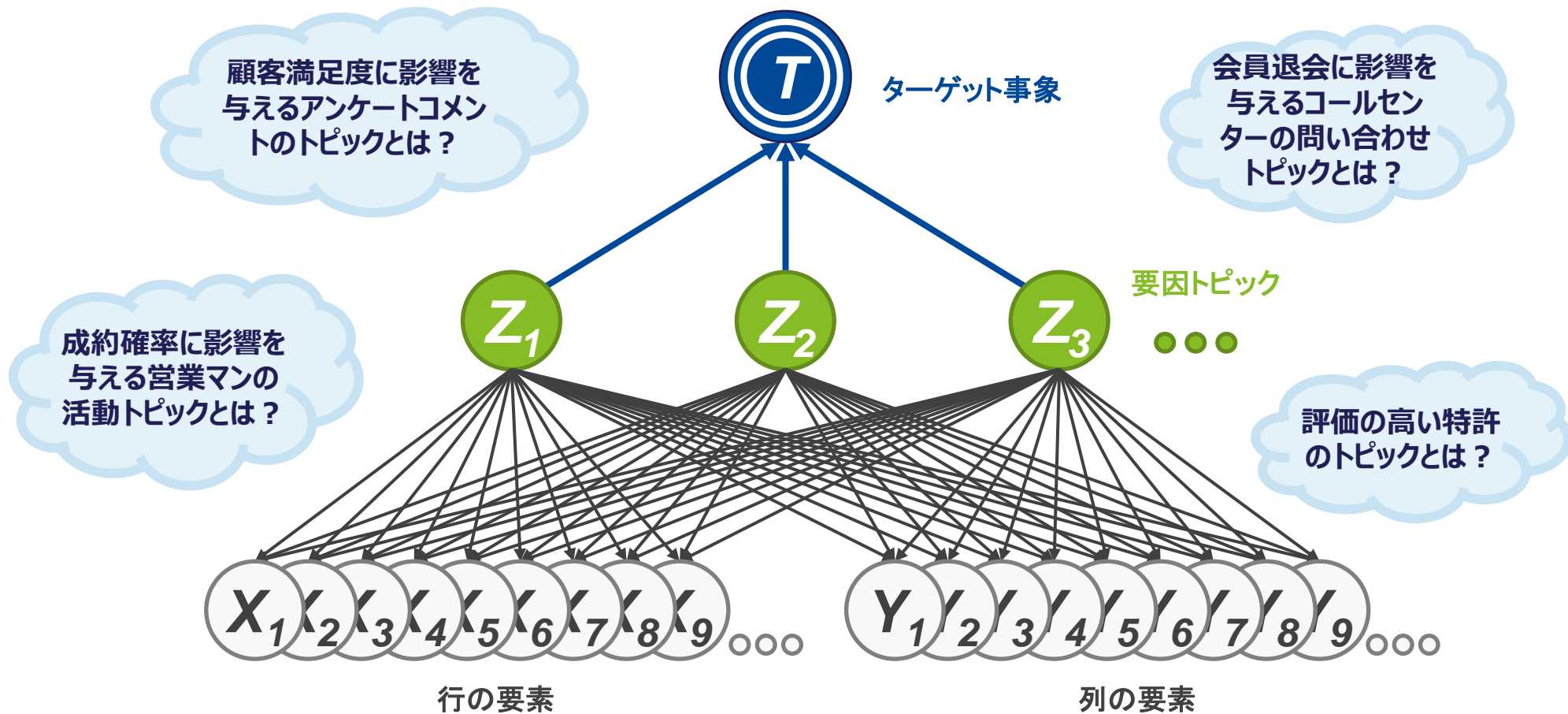
発明の名称	出願年
扇風機	2007
課題	
<p>マイナスイオンを発生することができる扇風機において、寿命の長いイオンを生成する。</p>	
解決手段	
<p>放電電極に水分を供給する貯水タンクおよび吸水管とを備え、放電電極と対向電極との間に高電圧を印加することで放電電極に保持される水分を霧化させる静電霧化装置を具備する。したがって、人体に有益で、かつ消臭や除菌等に効果があるイオンが水に包まれることで、寿命の長いイオンミストとして空中に放出することができ、また水に包まれている分重いイオンミストを、扇風機によって広範囲に拡散することができる</p>	

※対外説明用のため要約文は一部加工している

まとめ

PCSA(確率的因果意味解析)は、ターゲット事象に影響を与える要因トピックを優先的にテキストデータから自動構成でき、効果的なビジネスアクションの検討を支援できます

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



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

