Analytics Design Lab

株式会社NTTデータ数理システム
ユーザーコンファレンス2018

電気自動車関連の特許文書データからAI技術で導くインサイト

確率的因果意味解析を用いた新たな分析軸の自動抽出とその利用

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11月22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。

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	● 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

代表取締役　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

© 2018 Analytics Design Lab Inc.

昨年の講演のおさらい

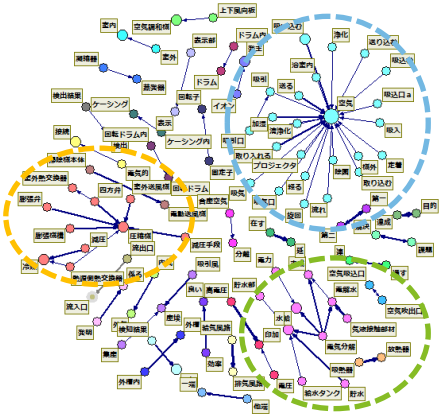
© 2018 Analytics Design Lab Inc.

これまでの特許文書分析と課題

これまでは特許文書から抽出された単語やそれを手動でグルーピングしたカテゴリをベースに、全体の特徴を可視化していましたが、結果が複雑で解釈がしにくいものでした

これまでの特許文書分析のアウトプット例

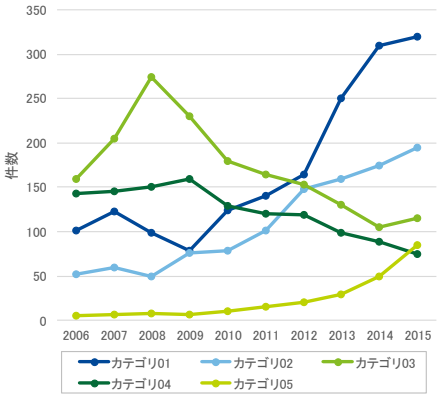
単語共起ネットワークによる全体像把握



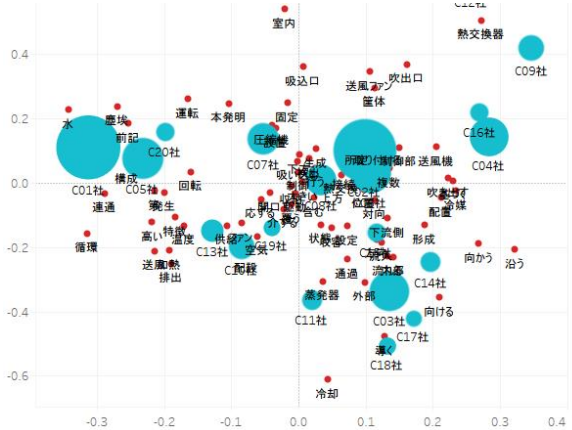
単語のグルーピング（カテゴリの作成）

例）掃除機カテゴリ
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

カテゴリ別トレンド



コレスポンデンス分析による出願人の特徴把握



課題と解決手段のカテゴリのクロス集計による関係把握

解決手段カテゴリ

課題カテゴリ ＼ 解決手段	カテゴリ01	カテゴリ02	カテゴリ03	カテゴリ04	カテゴリ05	カテゴリ06	カテゴリ07	カテゴリ08	カテゴリ09	カテゴリ10
カテゴリ01	206	80	71	184	26	47	11	9	43	1
カテゴリ02	208	76	87	182	23	48	9	15	40	2
カテゴリ03	172	74	53	57	31	35	10	21	20	3
カテゴリ04	176	54	37	59	26	46	29	26	9	5
カテゴリ05	85	39	13	23	14	16	5	0	7	2
カテゴリ06	87	53	31	33	59	37	15	24	28	19
カテゴリ07	79	68	82	28	24	12	6	16	18	15

課題

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

© 2018 Analytics Design Lab Inc.

新たなテキスト分析アプローチ：Nomolytics

※特許登録済（特許第6085888号 ）

テキストマイニング技術に複数のAI技術を応用することで、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する技術を開発しました

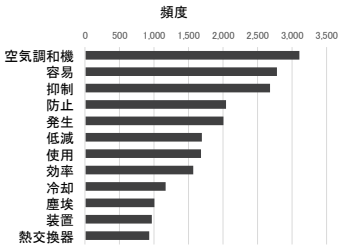
Nomolytics®: Narrative Orchestration Modeling Analytics

テキストマイニング

文章に含まれる単語を抽出し、その出現頻度を集計する

単語抽出

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、膨大な単語を複数のトピックにまとめる

トピック抽出

Visual Mining Studio

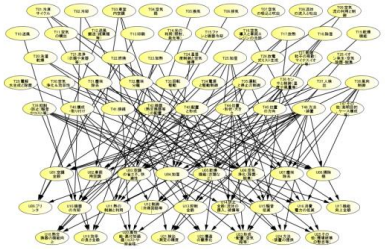
×

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

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

テキスト情報内に潜む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

© 2018 Analytics Design Lab Inc.

Nomolyticsの技術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
- 製品やサービスのニーズをトピックで把握
- 顧客ターゲット別の関心トピックを把握
- 価値観に応じたマーケティングを検討

アンケート
- 自由記述の内容をトピックで把握
- 他の設問と自由記述の関係を把握
- 顧客満足を生む施策を検討

コールセンター履歴
- 問い合わせ内容をトピックで把握
- 各製品の問い合わせ特徴を把握
- 顧客離反を抑制する対策を検討

特許文書
- 特許文書の技術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
- 技術の差別化戦略や提携戦略を検討

営業日報
- 営業活動内容をトピックで把握
- 営業の成約確率を高める活動を把握
- 効果的な営業教育を検討

有価証券報告書
- 各企業の事業内容をトピックで把握
- 各種指標と事業トピックの関係を把握
- 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文の概要をトピックで把握
- 記述トピックに基づく学生の分類
- 学生の絞込みや面接の質問内容を検討

診療記録
- 診療記録、看護記録をトピックで把握
- 聞き取り内容と検査指標の関係を把握
- 定性情報も用いた診療支援を検討

問題発生レポート
- 不具合やヒヤリハットをトピックで把握
- 問題内容と作業環境の関係を把握
- 問題発生を抑制する環境改善を検討

© 2018 Analytics Design Lab Inc.

昨年の講演概要

「風」「空気」に関する特許データ30,039件にNomolyticsを適用し、用途と技術のトピックの抽出、トピックのトレンドや出願人動向の可視化、用途と技術の関係分析を実施しました

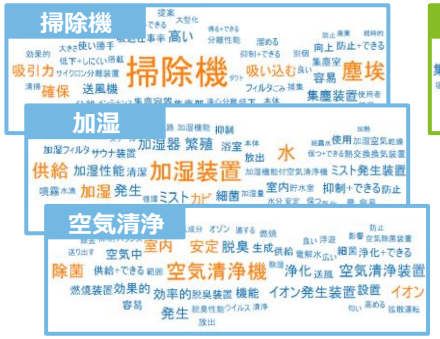
トピックの抽出

テキストマイニングによる単語の抽出

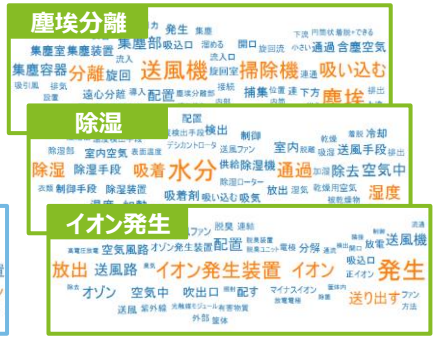
単語	品詞	頻度
空気調和機	名詞	3,106
空気	名詞	2,846
容易な	形容詞	2,790
抑制する	動詞	2,687
塵埃	名詞	1,687
分離する	動詞	1,231
・・・	・・・	…

PLSAの適用

【課題】の要約文から用途トピックを25個抽出



【解決手段】の要約文から技術トピックを47個抽出



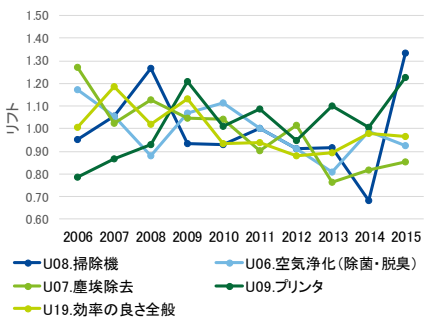
トピックのスコア計算

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	・・・	・・・	…	…	…	…	…	…

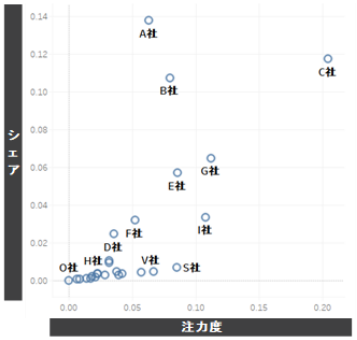
トレンドの可視化

各トピックのスコアを出願年で集計



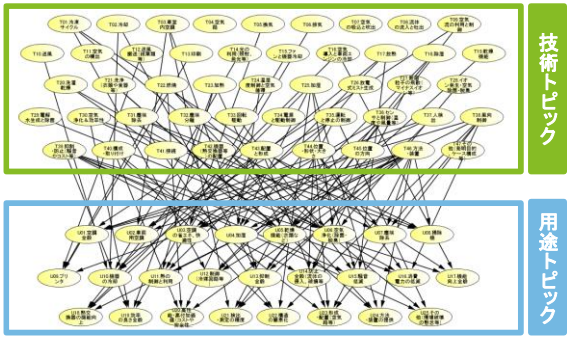
出願人動向の可視化

各トピックのスコアを出願人で集計して各社のポジションを可視化



用途と技術の関係分析

技術トピックに対する用途トピックの関係をベイジアンネットワークでモデル化し技術の新規用途を検討



© 2018 Analytics Design Lab Inc.

今年の講演概要

昨年は全体を表すトピックを抽出してから、属性との関係を分析しましたが、今年はその属性の特徴の要因となるトピックを最初から抽出して、より効果的な特徴の探索を行います

昨年は

データ全体を表現するトピックを抽出

してから

属性との関係を分析

すると

属性の特徴を把握

できた

適用技術 → ***Nomolytics***

今年は

属性の特徴を示す

ような

偏ったトピックを抽出

してから

より顕著な属性との関係を分析

する

適用技術 → ***PCSA(確率的因果意味解析)***

© 2018 Analytics Design Lab Inc.

PCSA：確率的因果意味解析

© 2018 Analytics Design Lab Inc.

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

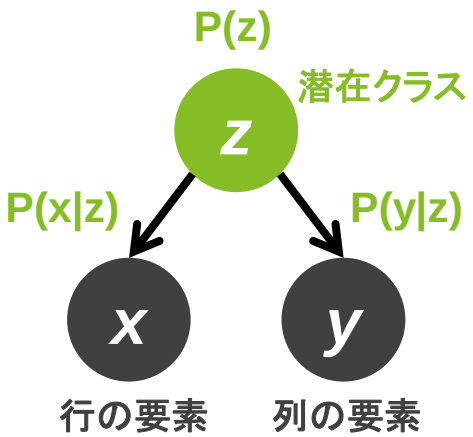
- 行列データの行の要素xと列の要素yの背後にある共通特徴となる潜在クラスzを抽出する手法である
- 元々は文書分類のための手法として開発されている(Hofman, 1999)
- 各文書の出現単語を記録した文書（行）×単語（列）という高次元（列数の多い）共起行列データに適用することで複数の潜在トピックを抽出し、文書（行）×トピック（列）という低次元データに変換して文書を分類する

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

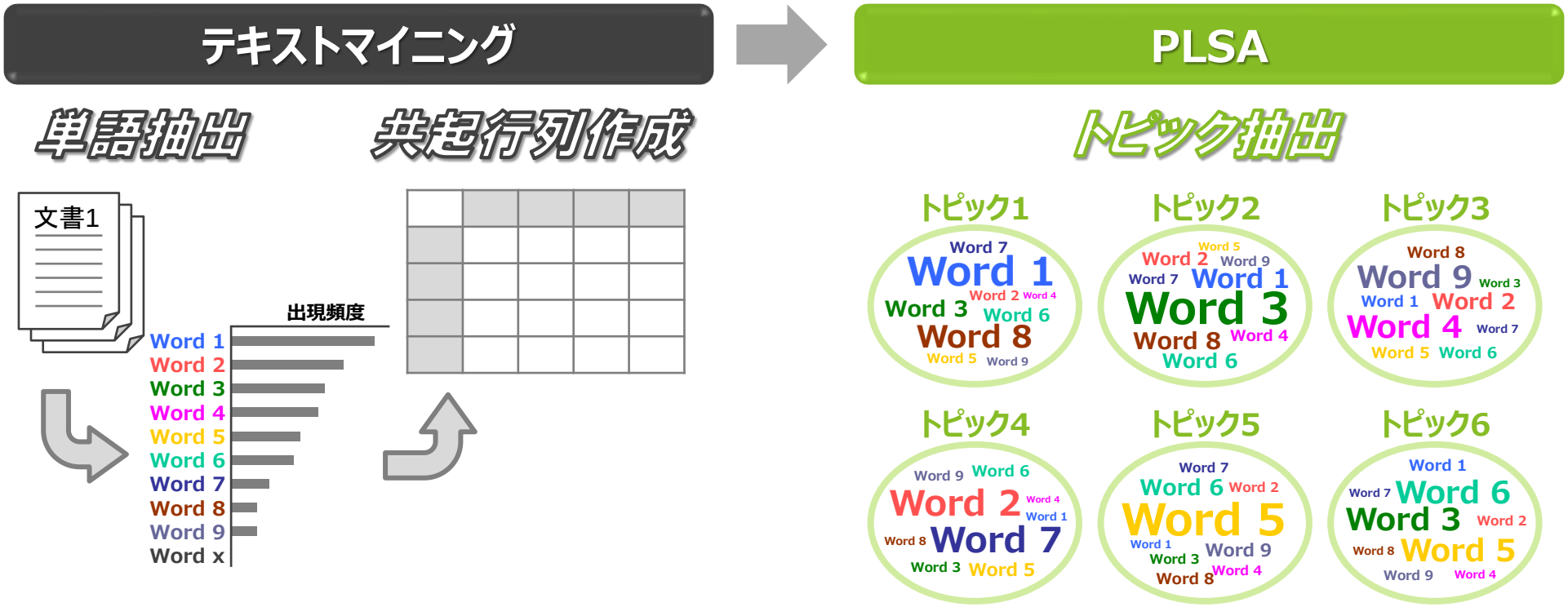
ソフトクラスタリングできる

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

© 2018 Analytics Design Lab Inc.

【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された単語で構成された共起行列にPLSAを適用することで、単語をトピックに集約し、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）

© 2018 Analytics Design Lab Inc.

【開発手法】 PCSAによる要因トピックの抽出

※特許出願中（特願2018-090885）

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、そのターゲット事象に影響を与えるトピックを優先して抽出します

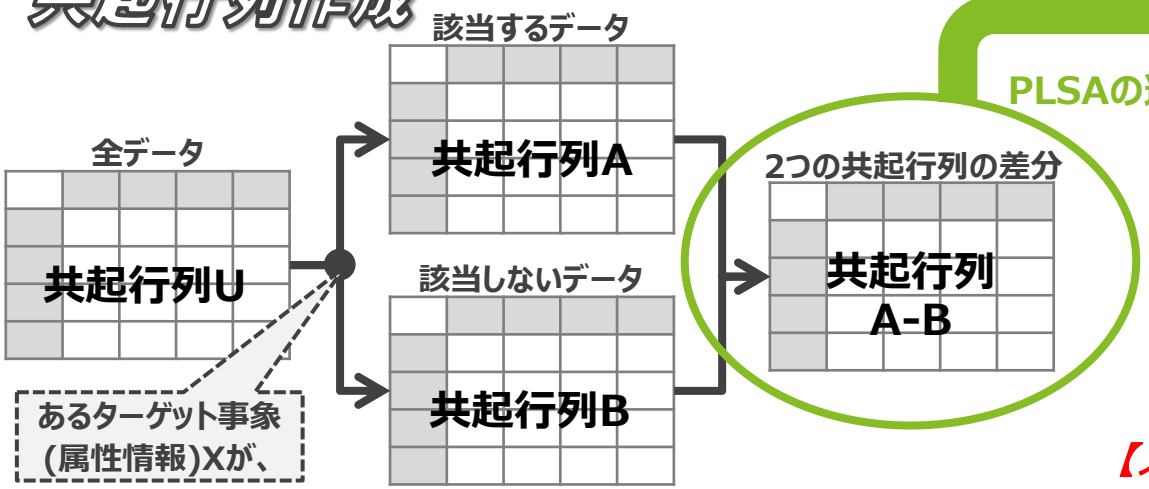
確率的因果意味解析（*PCSA: Probabilistic Causal Semantic Analysis*）

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

共起行列作成



トピック抽出

ターゲット事象Xの該当有無に影響を与える潜在トピックを優先的にテキスト情報から抽出できる

【人工知能学会 2018年度全国大会優秀賞 受賞】

© 2018 Analytics Design Lab Inc.

Nomolyticsでは教師なし学習(PLSA)を終了してから、その結果を使った教師あり学習をしていますが、PCSAでは教師の情報を考慮したような教師なし学習を実行しています

Nomolytics

テキスト
マイニング

教師なし学習と教師あり学習をそれぞれ独立させて順番に実行している

PLSA
確率的潜在
意味解析

教師なし学習

ベイジアン
ネットワーク

教師あり学習

PCSA(確率的因果意味解析)

テキスト
マイニング

教師あり情報を差し込んだ教師なし学習を実行しているようなイメージ

差分
共起行列

教師あり

PLSA
確率的潜在
意味解析

教師なし学習

© 2018 Analytics Design Lab Inc.

PCSAの適用事例：
電気自動車関連の特許文書データの分析

※本事例は外部発表用に分析した事例であり、実際にサービスとして提供した分析の事例ではありません

© 2018 Analytics Design Lab Inc.

「車」「電気」を含む10年分の特許データ26,419件の要約文を対象に、全体のトピックをPLSAで、出願年・パテントスコア・出願人の属性で特徴を示すトピックをPCSAで抽出します

分析データの抽出条件

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

分析対象

- トピック抽出対象：要約文のテキスト情報
- 使用する属性情報
  - 出願年：2007年～2016年
  - 出願人：出願件数上位26社 （個人は除く）
    ※グループ会社を統一するなどの名寄せ済み
  - PIスコア
    請求項の広さ、被引用回数に比例する指標
    （Patent Integrationの独自評価指標）

トピックの抽出対象

①PLSAで抽出する全体の集約トピック

U	全体	26,419件

▶ データ全体を表すトピックの分類を把握する

②PCSAで抽出する属性別の特徴トピック

A	出願年	2014年以後	6,405件（24.2%）
		2013年以前	20,014件（75.8%）

▶ 最近の出願特許の特徴を把握する

B	PIスコア	5点以上	7,558件（28.6%）
		5点未満	18,861件（71.4%）

▶ スコアの高い有用技術の特徴を把握する

C	出願人	上位5社	6,617件（25.0%）
		5社以外	19,802件（75.0%）

▶ 出願量の多い企業とそうでない企業の棲み分けを把握する

© 2018 Analytics Design Lab Inc.

全体のトピックの抽出

該当件数：26,419件

© 2018 Analytics Design Lab Inc.

テキストマイニングで要約文から名詞と係り受け表現を抽出し、名詞×係り受けで構成される共起行列を作成します

テキストマイニングの実行 Text Mining Studio → 共起行列の作成

要約文に含まれる「名詞」と「係り受け表現」を抽出する

単語	頻度
構成	4,997
制御	4,360
配置	3,895
モータ	3,486
形成	3,459
供給	3,309
検出	3,215
電気自動車	3,181
バッテリ	2,985
電力	2,828
駆動	2,496
方法	2,286
制御装置	2,092
…	…

名詞：3,020語

係り受け表現	頻度
電力-供給	1,208
否-判定	517
モータ-駆動	460
バッテリ-充電	440
効率-良い	419
供給-電力	332
電気自動車-提供	285
充電-行う	273
モータ-供給	270
並列-接続	267
バッテリ-接続	252
モータ-備える	242
簡易-構成	228
…	…

係り受け：2,128表現

- 係り受け表現は「名詞×動詞（サ変名詞含む）・形容詞・形容動詞」の係り受けペアを抽出
- 頻度20件以上を対象に抽出
- ノイズとなるような語はあらかじめ除外して抽出

名詞×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

	電力-供給	否-判定	モータ-駆動	バッテリ-充電	効率-良い	供給-電力	電気自動車-提供	充電-行う	モータ-供給	並列-接続
構成	118	33	36	33	24	32	10	25	30	46
制御	268	73	108	85	12	115	2	41	74	40
配置	69	2	29	8	6	15	5	9	18	12
モータ	239	61	494	58	31	54	33	19	280	27
形成	31	4	20	8	1	7	0	5	4	12
供給	1,350	43	53	85	10	368	7	43	280	43
検出	134	99	56	36	4	44	8	19	33	27
電気自動車	193	73	79	129	59	53	289	114	54	13
バッテリ	337	87	60	529	44	81	20	72	106	70
電力	1,350	50	75	127	35	368	20	57	189	36

名詞：3,020語×係り受け2,128表現

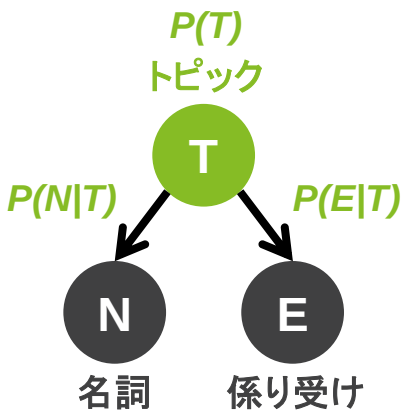
© 2018 Analytics Design Lab Inc.

作成した共起行列にPLSAを適用してトピックを抽出し、それぞれのトピックに所属する名詞と係り受けの所属確率からそのトピックの意味を解釈します

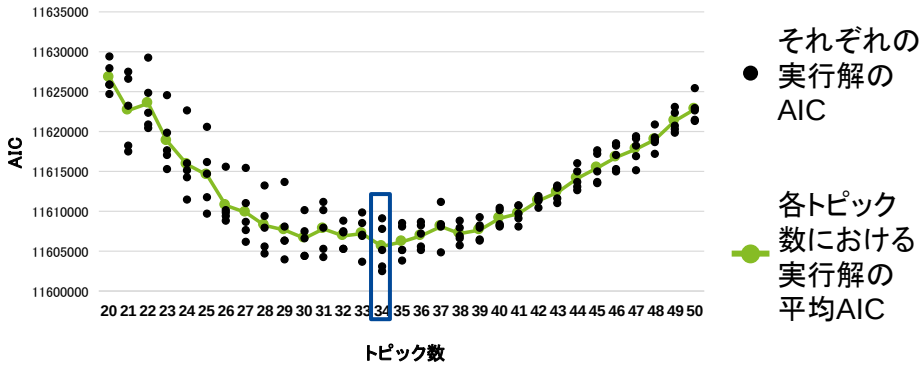
PLSAの実行

Visual Mining Studio

共起行列にPLSAを適用し、AICを評価基準とした最適なトピックを抽出する



- PLSAはVisual Mining Studioの二項ソフトクラスタリングを適用する
- PLSAはトピック数を指定する必要があり、初期値により解が異なる特性があるため、トピック数を20から50まで1刻みで変化させ、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



トピックの解釈

各トピックに所属する「名詞」と「係り受け」の所属確率からそのトピックの意味を解釈する

- PLSAでは以下の3つの確率がアウトプットとなる
  - P(T)・・・トピックの存在確率
  - P(N|T)・・・トピックにおける名詞の所属確率
  - P(E|T)・・・トピックにおける係り受けの所属確率
- P(N|T)とP(E|T)からトピックの意味を解釈する

（例）Tu13の解釈：電気自動車の蓄電池充電

P(N|U)	名詞
12.6%	充電
8.9%	電気自動車
6.5%	蓄電装置
3.0%	バッテリ
2.0%	充電システム
2.0%	蓄電池
1.9%	電力
1.7%	制御
1.5%	充電スタンド
1.5%	放電
1.3%	外部電源
1.2%	充電+できる
1.0%	充電ケーブル
0.8%	検出
0.7%	情報
…	・・・

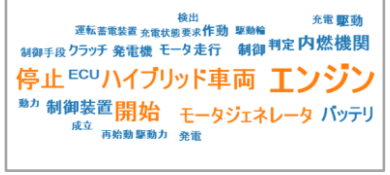
P(E|U)	係り受け
5.1%	バッテリ-充電
4.0%	充電-行う
3.9%	電気自動車-充電
1.9%	蓄電池-充電
1.6%	蓄電装置-充電
1.6%	電力-供給
1.3%	充電-開始
1.2%	電気自動車-接続
1.2%	充電-蓄電装置
1.1%	用いる-充電
1.1%	充電-制御
1.0%	供給-電力
0.9%	蓄電装置-備える
0.9%	電力-充電
0.8%	蓄電装置-提供
…	・・・

© 2018 Analytics Design Lab Inc.

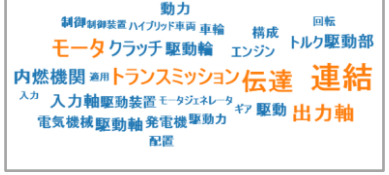
全体のトピック34個の一覧①

全体を集約するトピックでは、エンジン、動力伝達、モータ、ブレーキ、電力変換、二次電池、充電、情報通信、異常検出、筐体、構成、接続、小型化、安全性など34個抽出されました

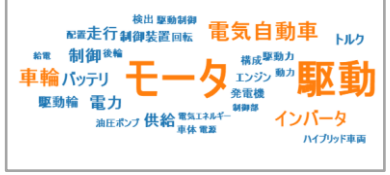
Tu01.エンジンの始動と停止



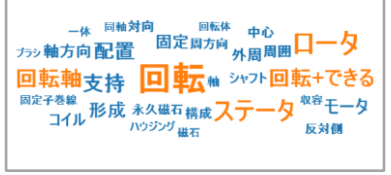
Tu02.動力の伝達



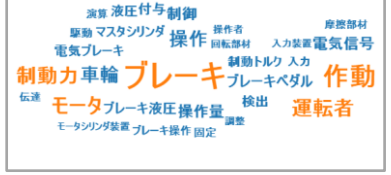
Tu03.モータ駆動



Tu04.ロータ・ステータなど回転部品の構成

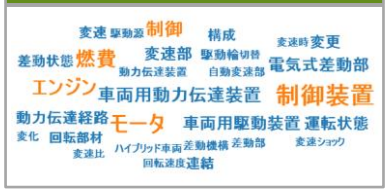


Tu05.ブレーキ装置

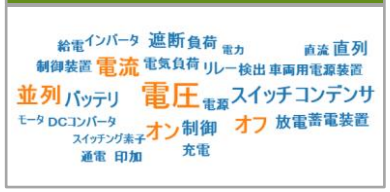


Tu06.動作制御

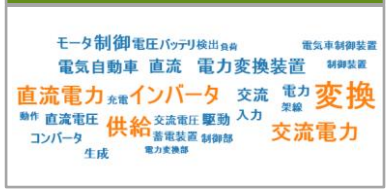
Tu07.動力伝達の制御



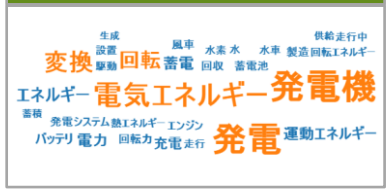
Tu08.スイッチの切り替え



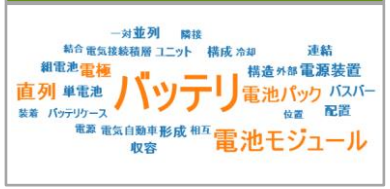
Tu09.交流・直流の変換



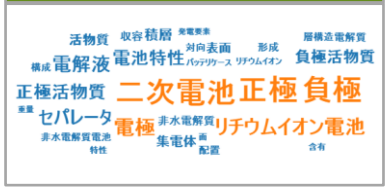
Tu10.エネルギーの変換



Tu11.電池モジュールの提供

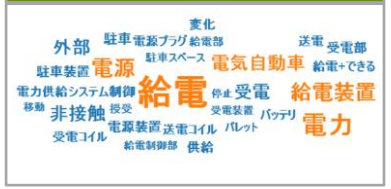


Tu12.二次電池の構成



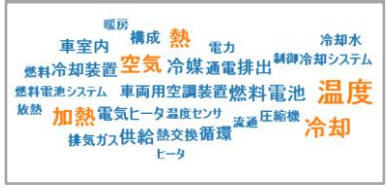
Tu13.電気自動車の蓄電池充電

Tu14.非接触受電など給電装置

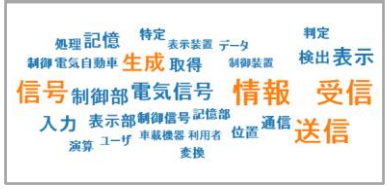


Tu15.外部への電力供給

Tu16.空調などの冷却・加熱



Tu17.情報通信



Tu18.演算・推定

Tu19.機器の異常検出

Tu20.操作スイッチ

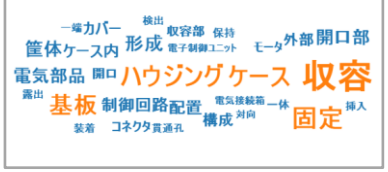
※文字の大きさはトピックに対する関係の強さを表現している
（上位5つの単語を赤色で表示している）

© 2018 Analytics Design Lab Inc.

全体のトピック34個の一覧②

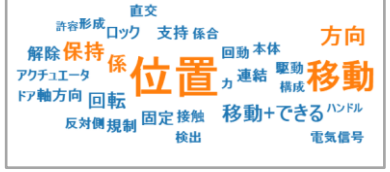
全体を集約するトピックでは、エンジン、動力伝達、モータ、ブレーキ、電力変換、二次電池、充電、情報通信、異常検出、筐体、構成、接続、小型化、安全性など34個抽出されました

Tu21.筐体



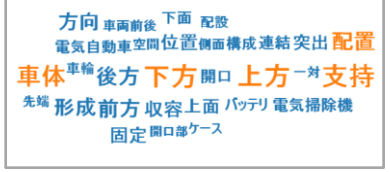
Tu22.表面の形成

Tu23.位置とその移動



Tu24.配置・位置・方向

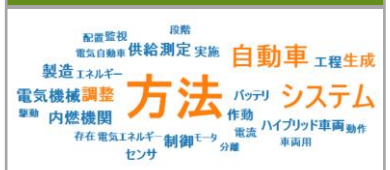
Tu25.構成の方位



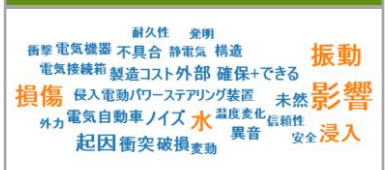
Tu26.構成

Tu27.接続

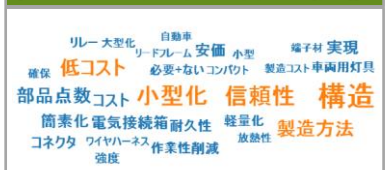
Tu28.方法の提供



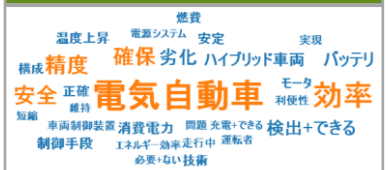
Tu29.損傷や浸水など不具合の防止



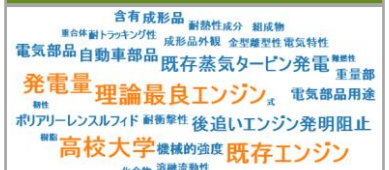
Tu30.小型化・簡素化・低コスト化など付加価値



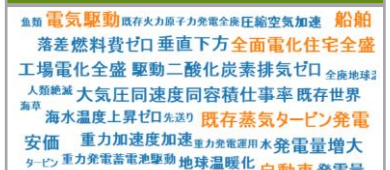
Tu31.効率性・安全性の向上



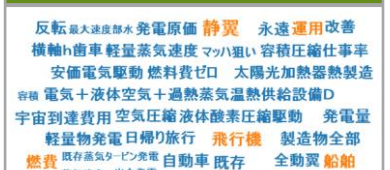
Tu32.既存エンジンへの警鐘・樹脂組成物の提供



Tu33.重力発電の活用による地球温暖化防止



Tu34.タービン発電の出力向上・燃費低減



Tu32,Tu33,Tu34は特定の出願人による重複した要約内容の特許から抽出された

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

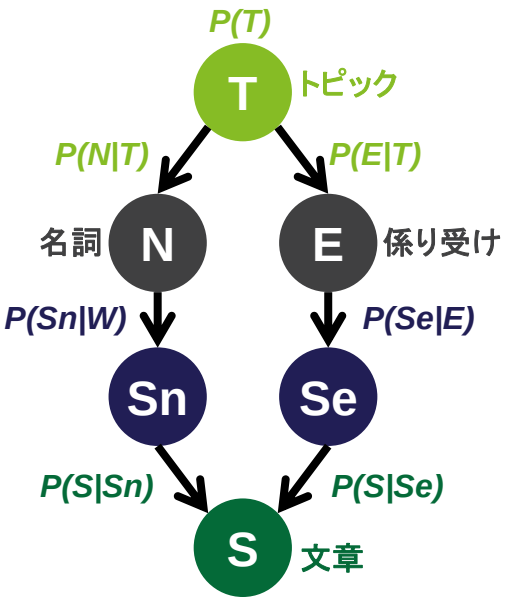
© 2018 Analytics Design Lab Inc.

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



文章を名詞で定義される文章Sn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

名詞N_iで定義される文章Sn_h

$$Sn_h = \{N_1, N_2, \cdots, N_i\}$$

トピックT_kを条件とした文章Sn_hの出現確率

$$P(Sn_h|T_k) = \sum_i P(Sn_h|N_i)P(N_i|T_k)$$

名詞N_iが出現する中で文章Sn_hが出現する確率
（N_iの出現文章数の逆数）

$$P(Sn_h|N_i) = 1/n(N_i)$$

係り受けE_jで定義される文章Se_h

$$Se_h = \{E_1, E_2, \cdots, E_J\}$$

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
※$P(S_h|Sn_h)$と$P(S_h|Se_h)$はともに1/2とする

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

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	・・・	T34
1	1	3.1	0.9	2.0		3.5
1	2	1.4	0.2	5.5		8.4
2	1	0.8	5.8	1.3		2.9
2	2	1.2	3.2	1.7		4.0
2	3	0.6	1.8	2.6		9.6
・・・						

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	・・・	T34
1	3.1	0.9	5.5		8.4
2	1.2	5.8	2.6		9.6
・・・					

③閾値を設定してフラグに変換する

※閾値は5に設定した

特許ID	T01	T02	T03	・・・	T34
1	0	0	1		1
2	0	1	0		1
・・・					

© 2018 Analytics Design Lab Inc.

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

トピックのスコア(フラグ情報)を紐づけた特許データ

特許ID	出願番号	要約	出願年	出願人	PIスコア	トピック Tu01	トピック Tu02	・・・	トピック Tu34
1	特願2007-XXXX	【課題】電気式変速操作装・・・	2007	A社	0	1	1		0
2	特願2009-XXXX	【課題】従来の電気自動車・・・	2009	B社	1	0	1		1
3	特願2012-XXXX	エンジンのための方法及び・・・	2012	C社	13	0	1		1
4	特願2013-XXXX	【課題】駐車場に設置された・・・	2013	D社	7	1	0		0
・・・	・・・		・・・	・・・		・・・	・・・		・・・
26,419	特願2016-XXXX	充電ステーションが電気エネ・・・	2016	X社	0	1	0		1

トピック×属性の様々な集計・分析が可能に

© 2018 Analytics Design Lab Inc.

PCSAの適用：
出願年の特徴トピックの抽出

2014年以後	2013年以前
該当件数：6,405件（24.2%）	該当件数：20,014件（75.8%）

分析のポイント

最近ホットなトピックとは？
もうトレンドが終わっているトピックとは？

© 2018 Analytics Design Lab Inc.

特徴トピックの抽出手順

①2014年以後データ、②2013年以前データで作成した2つの共起行列の差分にPLSAを適用することで、2014年前後の出願に特徴のあるトピックを優先的に抽出します

2つの共起行列の作成 → 差分の共起行列の作成 → PLSAの適用

①出願年が2014年以後のデータ

6,405件（文章数：10,723件）

	電力-供給	否-判定	モータ-駆動	バッテリ-充電	効率-良い
構成	21	3	9	8	6
制御	60	24	28	19	3
配置	13	1	6	0	2
モータ	47	19	138	12	5
形成	7	2	7	0	0

②出願年が2013年以前のデータ

20,014件（文章数：33,113件）

	電力-供給	否-判定	モータ-駆動	バッテリ-充電	効率-良い
構成	97	30	27	25	18
制御	208	49	80	66	9
配置	56	1	23	8	4
モータ	192	42	356	46	26
形成	24	2	13	8	1

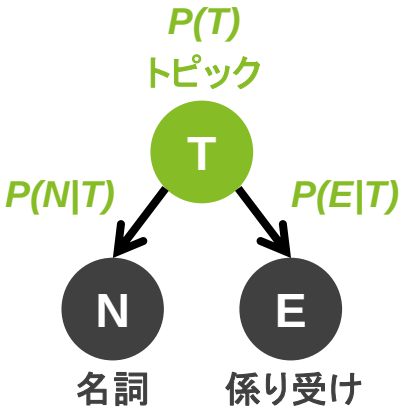
①2014年以後の共起行列と
②2013年以前の共起行列の
差の絶対値を計算した共起行列を作成する

| ①－重み調整済みの② |

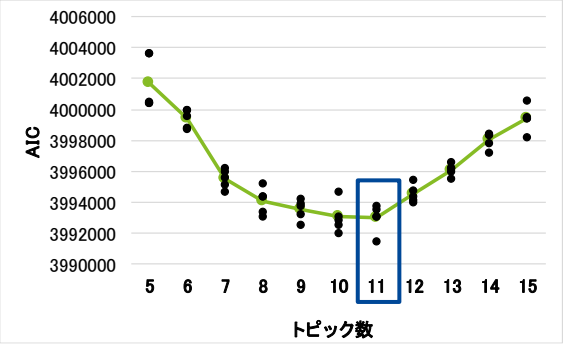
	電力-供給	否-判定	モータ-駆動	バッテリ-充電	効率-良い
構成	10.4	6.7	0.3	0.1	0.2
制御	7.4	8.1	2.1	2.4	0.1
配置	5.1	0.7	1.4	2.6	0.7
モータ	15.2	5.4	22.7	2.9	3.4
形成	0.8	1.4	2.8	2.6	0.3

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、②2013年以前の共起行列の頻度を2つの文章数の比率（10,723/33,113）で重み調整してから、①2014年以後の共起行列との差の絶対値を計算する

差分の共起行列にPLSAを適用する



AICを評価基準とした最適なトピックを抽出する

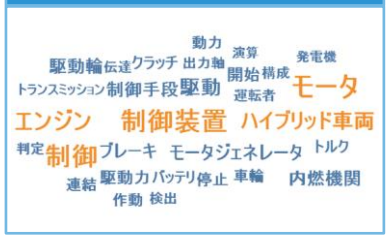


© 2018 Analytics Design Lab Inc.

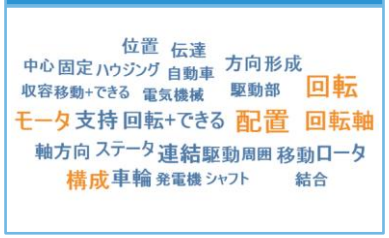
出願年の特徴トピック11個の一覧

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

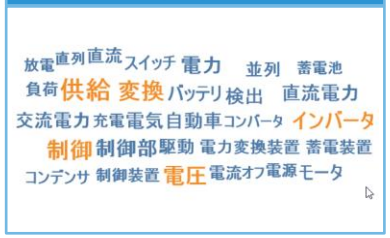
Ta01.エンジン駆動・動力伝達の制御



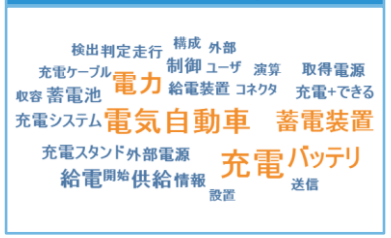
Ta02.モータの回転構成



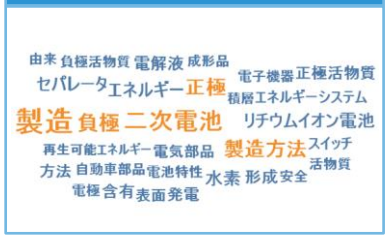
Ta03.交流・直流の変換



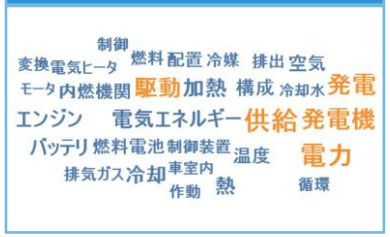
Ta04.電気自動車への充電、給電装置



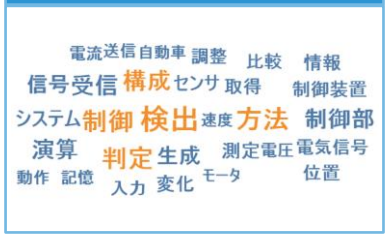
Ta05.二次電池の製造方法



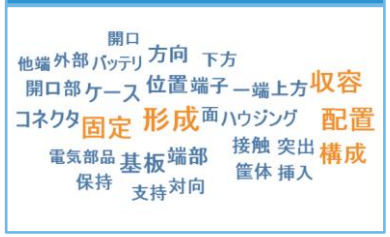
Ta06.空調などの冷却・加熱



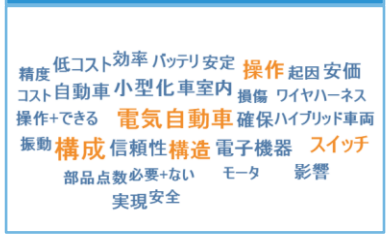
Ta07.情報通信、検出判定システム



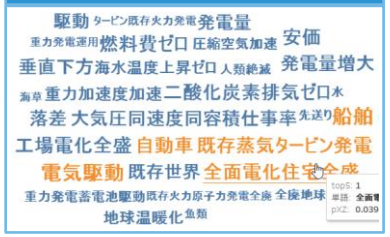
Ta08.形成・配置



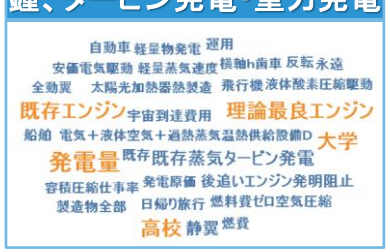
Ta09.小型化・低コスト化・簡素化・操作性向上



Ta10.重力発電の活用による地球温暖化防止



Ta11.既存エンジンへの警鐘、タービン発電・重力発電



Ta10,Ta11は特定の出願人による重複した要約内容の特許から抽出された

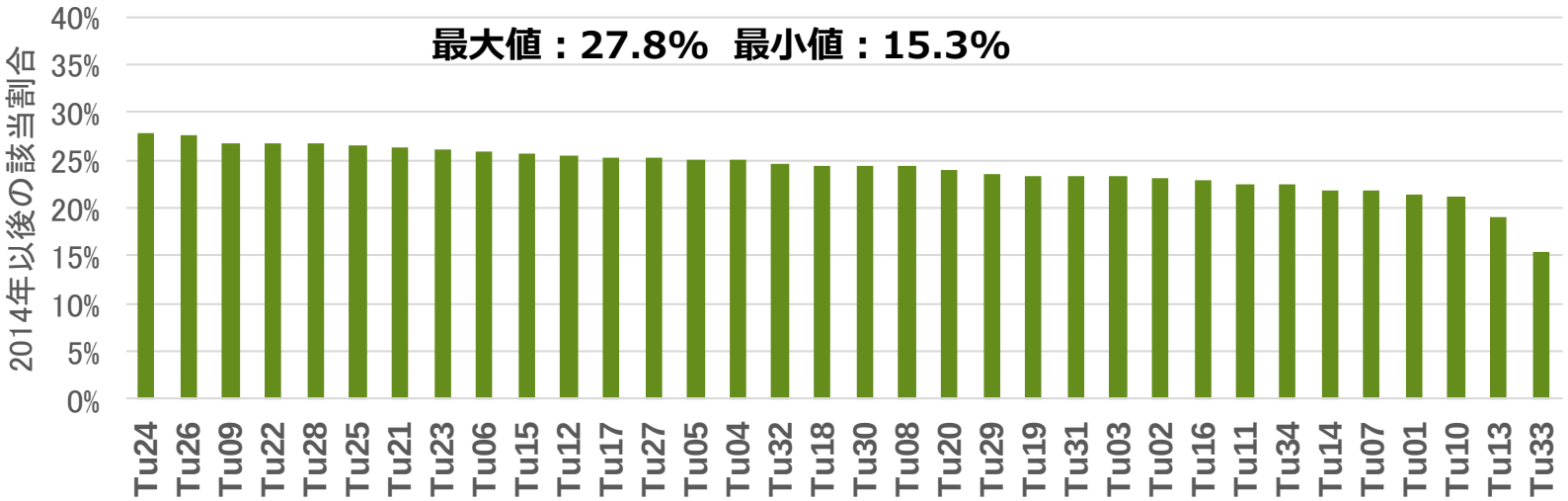
※文字の大きさはトピックに対する関係の強さを表現している（上位5つの単語を赤色で表示している）

© 2018 Analytics Design Lab Inc.

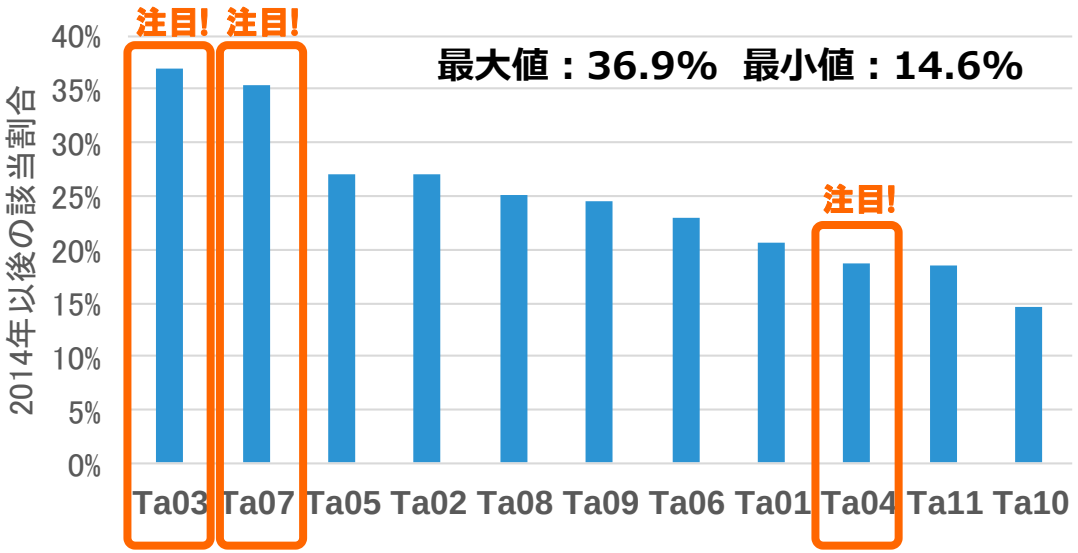
全体のトピックとの属性分布の比較

トピックごとに2014年以後データの該当割合を比較すると、全体ではおおむね25%前後ですが、PCSAのトピックでは高いものから低いものまで(特に高いものが)抽出されています

全体のトピック



PCSA適用の特徴トピック



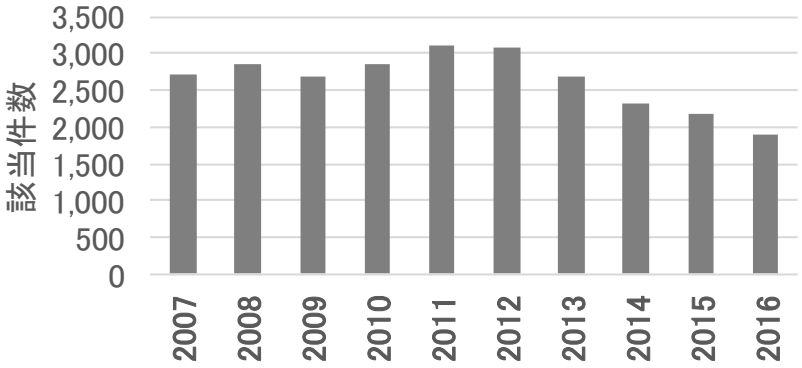
- それぞれ2014年以後のデータの該当割合の高い順にトピックを並べている
- 全体のトピックではおおむね25%前後のトピックが抽出されているが、PCSAのトピックでは割合が高いものから低いものまで抽出されており、最大値も高い
- 元々の2014年以後データの割合は24.2%

© 2018 Analytics Design Lab Inc.

2014年以後の該当割合が低いTa04は2011年が出願のピークとなっており、2014年以後の該当割合が高いTa03、Ta07は近年出願が多くなっています

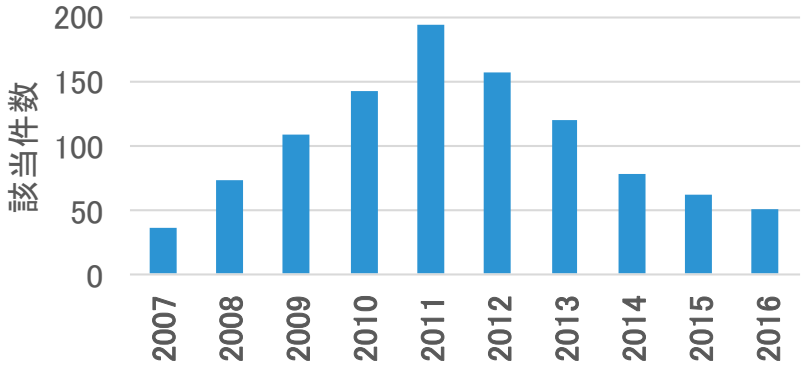
全体の出願件数の推移

全体では緩やかに減少傾向にある



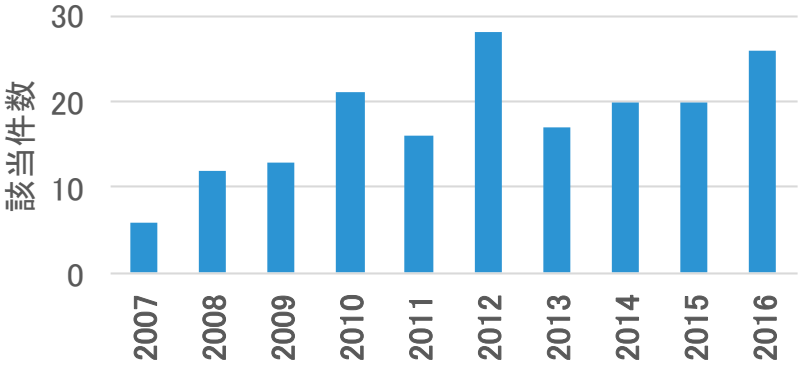
「Ta04.電気自動車への充電、給電装置」の推移

Ta04は2014年以後の該当割合ワースト3位



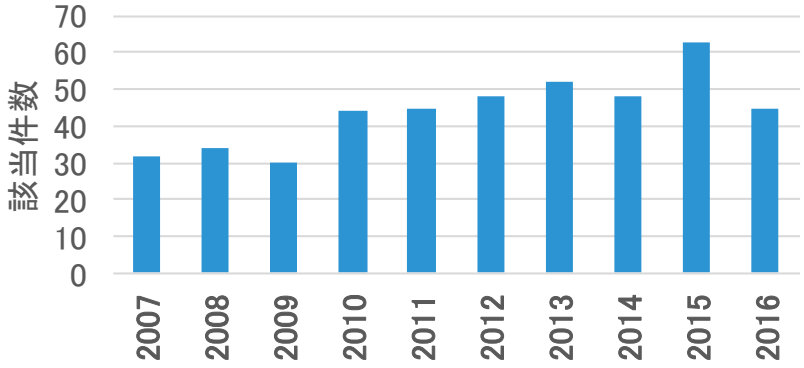
「Ta03.交流・直流の変換」の推移

Ta03は2014年以後の該当割合1位



「Ta07.情報通信、検出判定システム」の推移

Ta07は2014年以後の該当割合2位



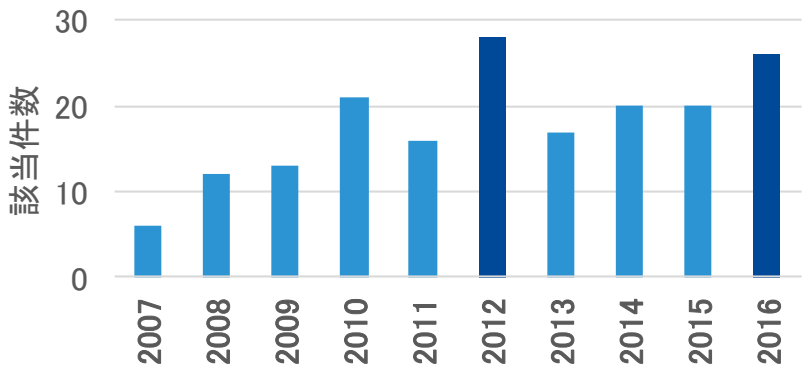
© 2018 Analytics Design Lab Inc.

特徴が目立つ特許群のキーワード

該当件数の多いTa03の2012年は電圧変換、Ta03の2016年は電気自動車のバッテリの電力変換、Ta07の2015年2016年は光通信に関するキーワードが特徴的です

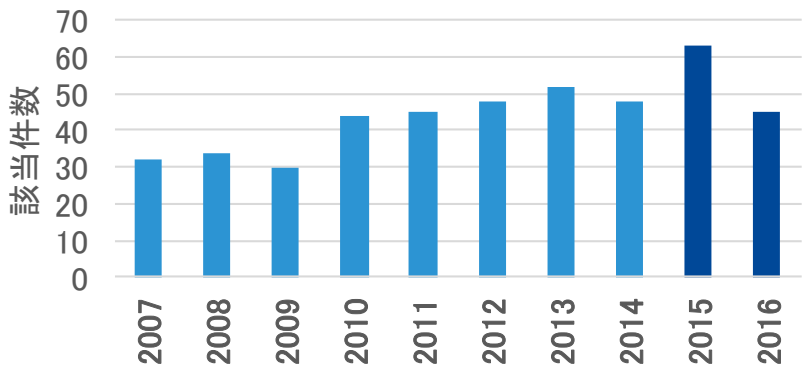
Ta03の該当が多い年の特許のキーワード

Ta03.交流・直流の変換



Ta07の該当が多い年の特許のキーワード

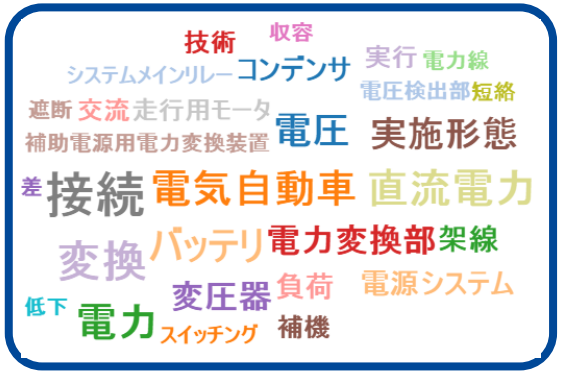
Ta07.情報通信、検出判定システム



Ta03×2012年の要約キーワード

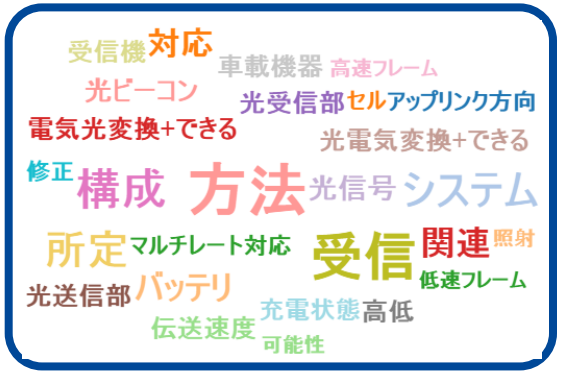
電圧の変換に関するキーワードが多い

Ta03×2016年の要約キーワード



電気自動車やバッテリの電力変換に関するキーワードが多い

Ta07×2015年2016年の要約キーワード



光信号の送受信のシステムに関するキーワードが多い

© 2018 Analytics Design Lab Inc.

PCSAの適用：
PIスコアの特徴トピックの抽出

５点以上	５点未満
該当件数：7,558件（28.6%）	該当件数：18,861件（71.4%）

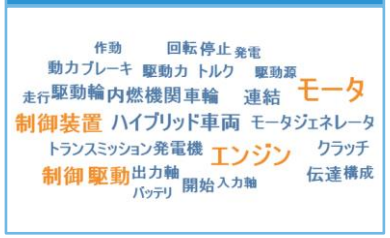
分析のポイント　スコアの高い有用な技術とは？

© 2018 Analytics Design Lab Inc.

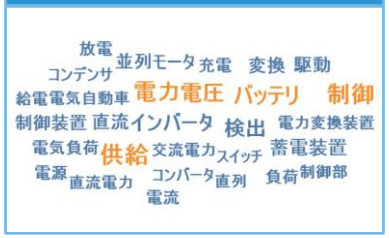
PIスコアの特徴トピック10個の一覧

PIスコアで特徴を示すトピックは、エンジン駆動、電力変換、充電、二次電池の製造方法、冷却・加熱、検出判定、小型化・低コスト化など10個抽出されました

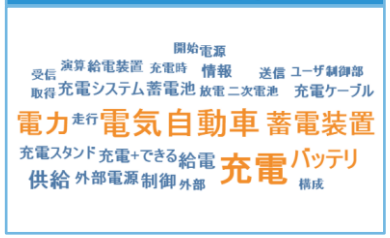
Tb01.エンジン駆動・動力伝達の制御



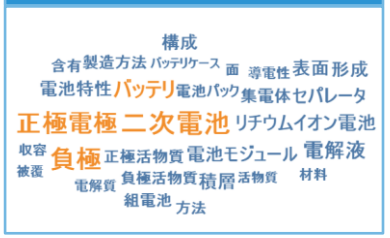
Tb02.交流・直流の変換



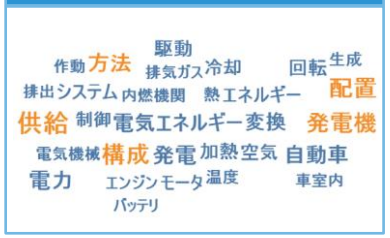
Tb03.電気自動車への充電、給電装置



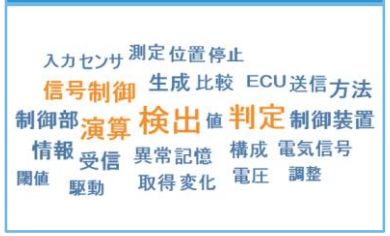
Tb04.二次電池の製造方法



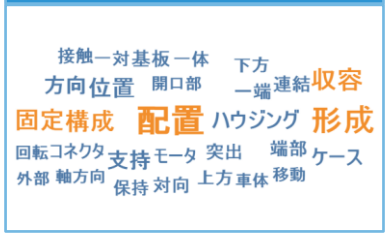
Tb05.空調などの冷却・加熱、ガス処理



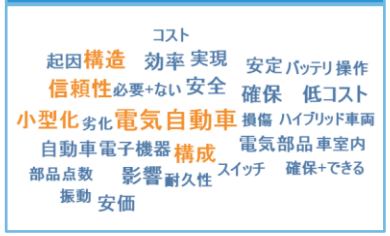
Tb06.状態の検出と判定



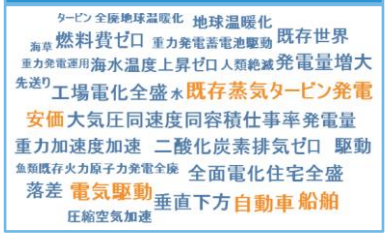
Tb07.形成・配置



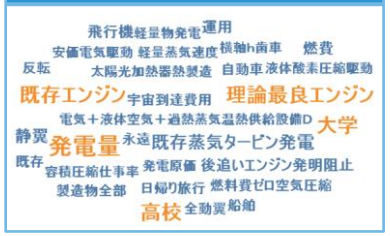
Tb08.小型化・低コスト化・簡素化・安全性向上



Tb09.重力発電の活用による地球温暖化防止



Tb10.既存エンジンへの警鐘・タービン発電の提案



Tb09,Tb10は特定の出願人による重複した要約内容の特許から抽出された

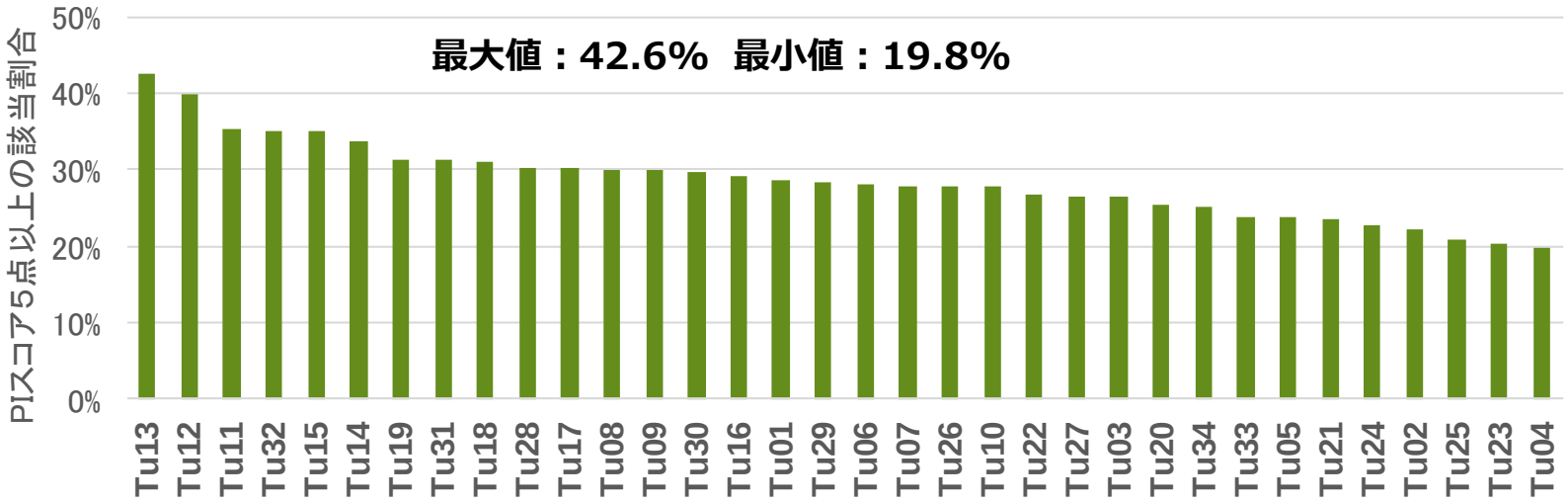
※文字の大きさはトピックに対する関係の強さを表現している （上位5つの単語を赤色で表示している）

© 2018 Analytics Design Lab Inc.

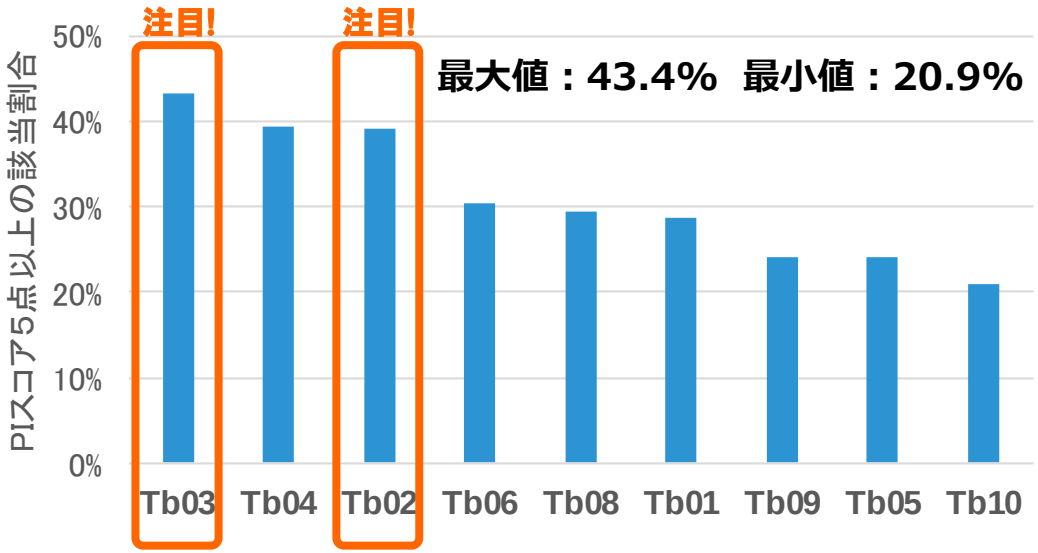
全体のトピックとの属性分布の比較

トピックごとにPIスコア5点以上のデータの該当割合を比較すると、最大値・最小値はほぼ変わりませんが、全体と比べてPCSAでは平均的な該当割合のトピックの抽出は限定されます

全体のトピック



PCSA適用の特徴トピック

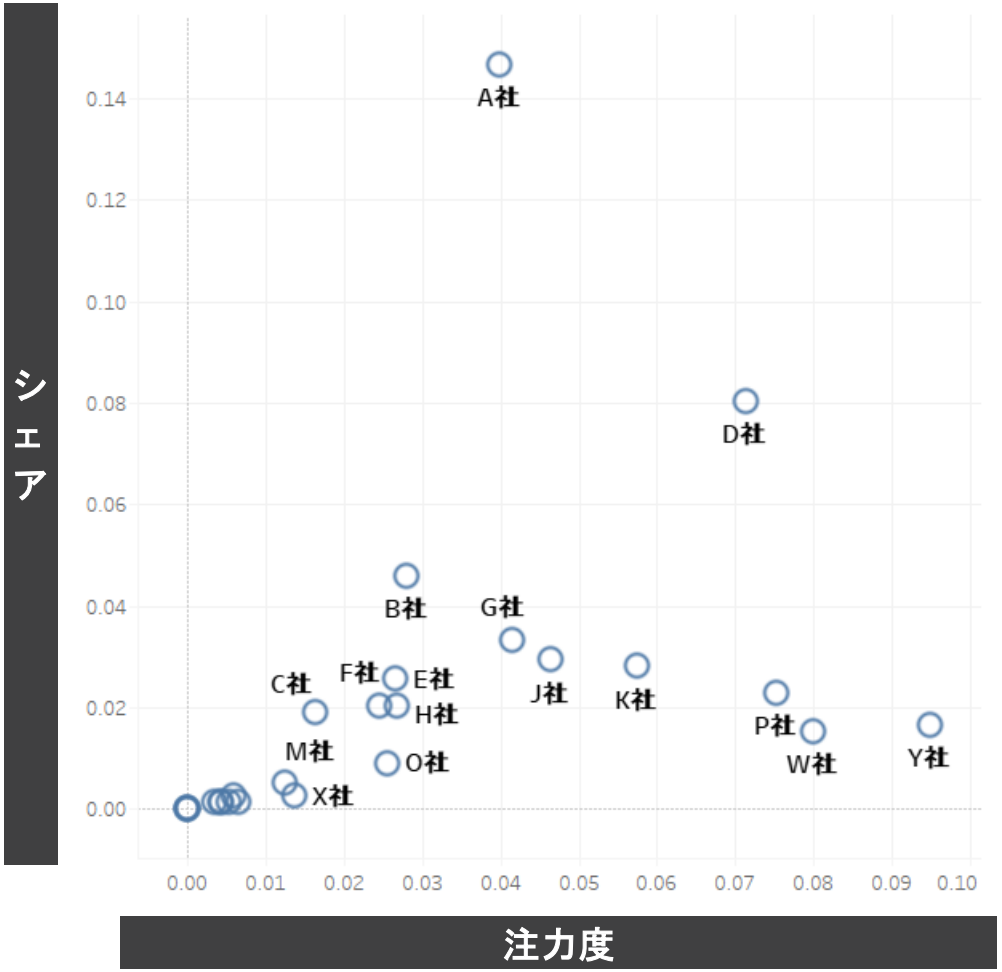


- それぞれPIスコア5点以上のデータの該当割合の高い順にトピックを並べている
- 全体のトピックもPCSAのトピックも割合の最大値・最小値はほぼ変わらないが、全体では平均的な割合のトピックが多く抽出され、PCSAではそれが限定的である
- Tb07は該当データがなかった
- 元々のPIスコア5点以上データの割合は28.6%

© 2018 Analytics Design Lab Inc.

電気自動車への充電に関するトピック(Tb03)はPIスコアが比較的高い傾向にありますが、メインプレーヤーは２社で、シェアがより高い１社と、注力度が高い１社となります

注力度とシェアを軸とした出願人のポジショニング



結果の解釈

- 最もシェアの高いのはA社で、2番手は少しギャップを開けてD社であるが、このトピックの領域ではこの2社がメインプレーヤーといえる
- D社はA社にシェアで劣るものの、注力度が高く、独自性の強い技術が開発されている可能性がある

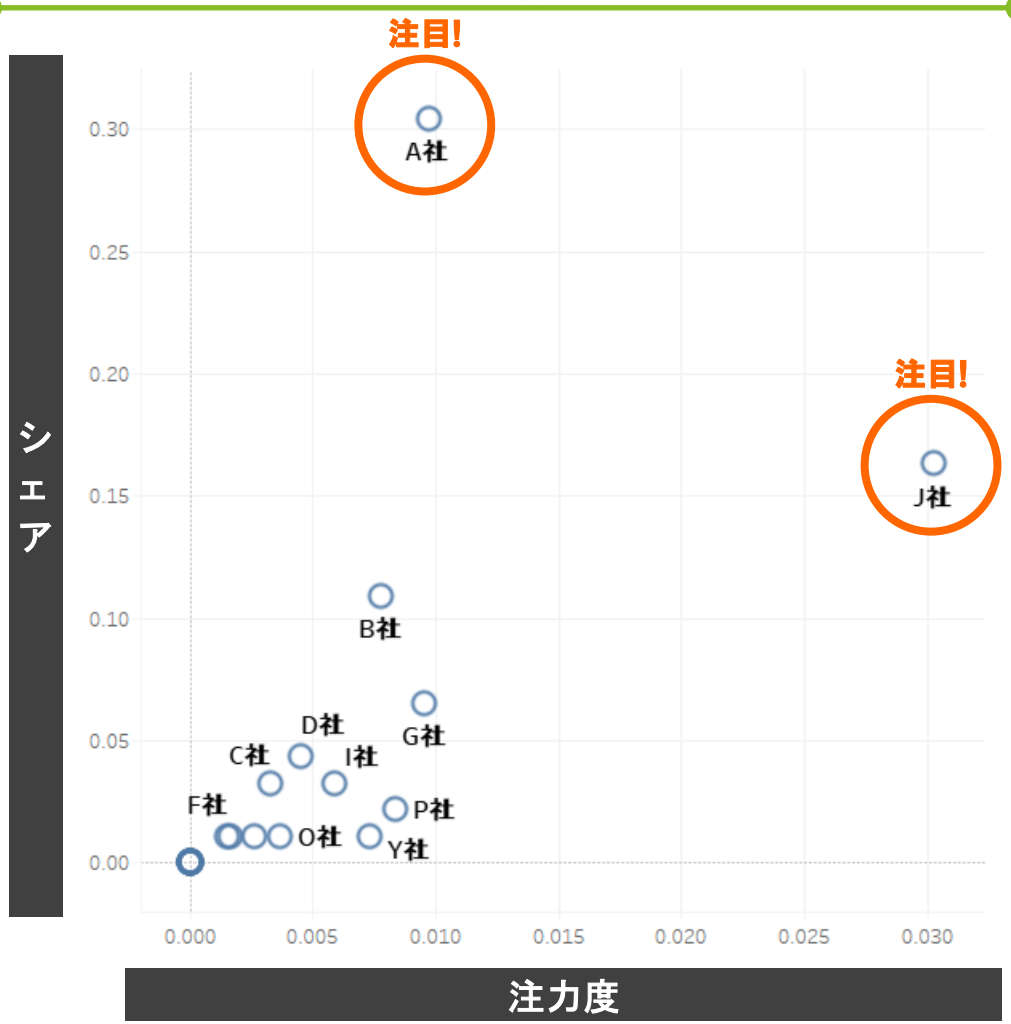
注力度とシェア

- **注力度**： $P(\text{トピック}T = 1 \mid \text{出願人}X = 1)$
  - 出願人Xの出願特許の中で、どれくらいの割合がそのトピックTに該当するものか、つまり出願人がどれくらいそのトピックに注力しているのかを示している
- **シェア**： $P(\text{出願人}X = 1 \mid \text{トピック}T = 1)$
  - トピックTが該当する特許の中で、どれくらいの割合がその出願人Xの出願によるものか、つまりトピックのなかでどれくらいその出願人が占めているのかを示している

© 2018 Analytics Design Lab Inc.

交流・直流の電力変換に関するトピック(Tb02)はPIスコアが比較的高い傾向にありますが、メインプレーヤーはおおよそ３社で、特に１社の注力度の高さが際立ちます

注力度とシェアを軸とした出願人のポジショニング



結果の解釈

- 最もシェアの高いのはA社で、2番手はギャップを開けてJ社、3番手はさらに少しギャップを開けてB社であるが、このトピックの領域ではこの3社がメインプレーヤーといえる
- 注力度ではJ社がとても高く、他社を大きく突き放しており、独自性の強い技術が開発されている可能性がある

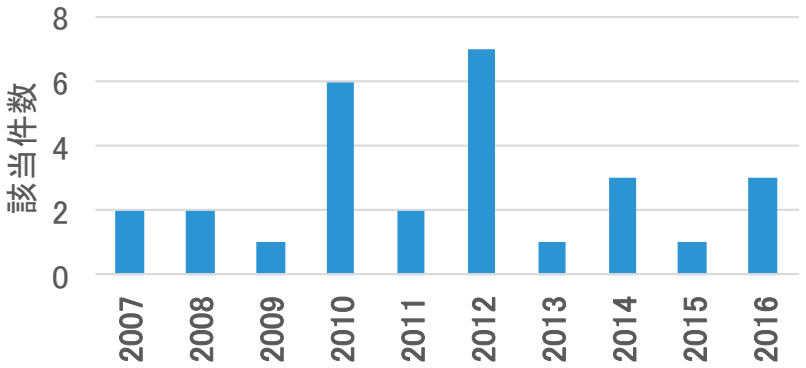
注力度とシェア （再掲）

- **注力度**： $P(\text{トピック}T = 1 \mid \text{出願人}X = 1)$
  - 出願人Xの出願特許の中で、どれくらいの割合がそのトピックTに該当するものか、つまり出願人がどれくらいそのトピックに注力しているのかを示している
- **シェア**： $P(\text{出願人}X = 1 \mid \text{トピック}T = 1)$
  - トピックTが該当する特許の中で、どれくらいの割合がその出願人Xの出願によるものか、つまりトピックのなかでどれくらいその出願人が占めているのかを示している

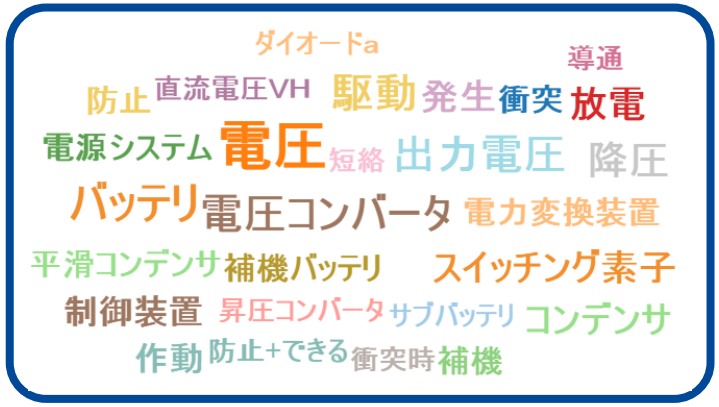
© 2018 Analytics Design Lab Inc.

シェアが最も高いA社は電圧変換に関するものが多いですが最近は出願が少なく、注力度の最も高いJ社は直近で出願が増えており鉄道の電力変換に関するものが多いです

A社（該当28件）の出願推移と要約キーワード

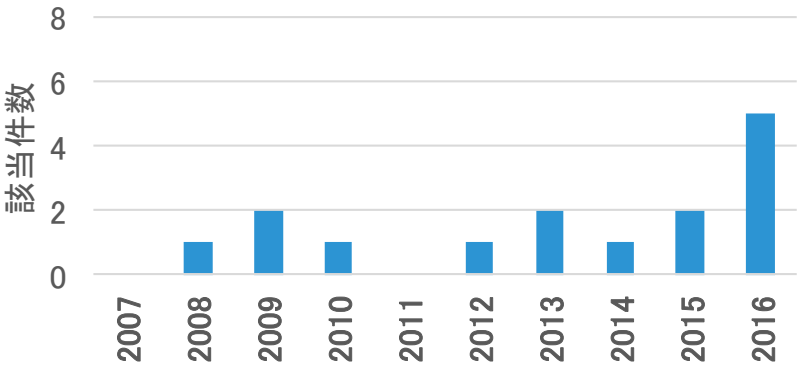


Tb02×A社の要約キーワード



電圧の変換に関するキーワードが多い

J社（該当15件）の出願推移と要約キーワード



Tb02×J社の要約キーワード

架線など鉄道車両の電力変換に関するキーワードが多い

© 2018 Analytics Design Lab Inc.

PCSAの適用：
出願人の特徴トピックの抽出

上位5社	5社以外
該当件数：6,617件（25.0%）	該当件数：19,802件（71.4%）

分析のポイント

出願件数の多い企業で牛耳られている技術とは？
出願件数の多い企業とそうでない企業で棲み分けができている技術とは？

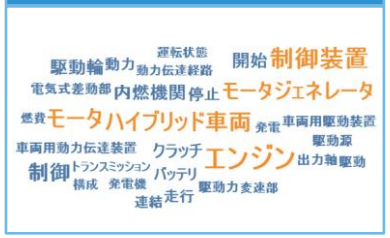
※出願件数の多い上位5社とは、本資料ではA社、B社、C社、D社、E社に該当する

© 2018 Analytics Design Lab Inc.

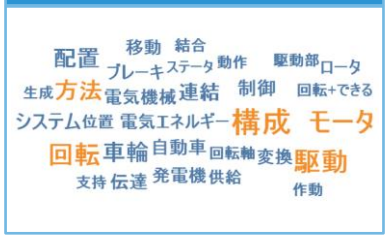
出願人の特徴トピック13個の一覧

出願人(上位5社か否か)で特徴を示すトピックは、エンジン駆動、モータ駆動、電力変換、充電、二次電池、温度制御、検出判定、操作性、小型化・低コスト化など13個抽出されました

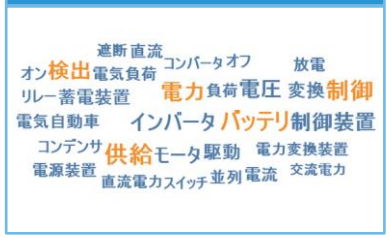
Tc01.エンジン駆動・動力伝達の制御



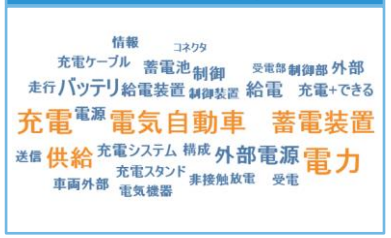
Tc02.モータ駆動、ブレーキシステム



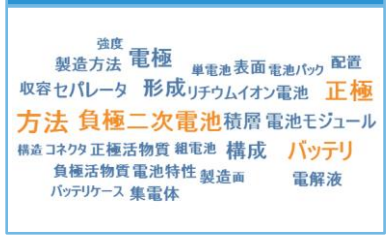
Tc03.交流・直流の変換



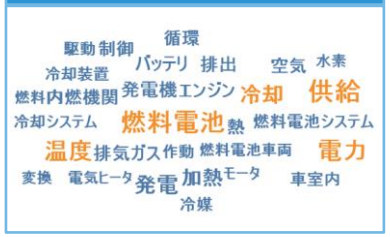
Tc04.電気自動車への充電、給電装置



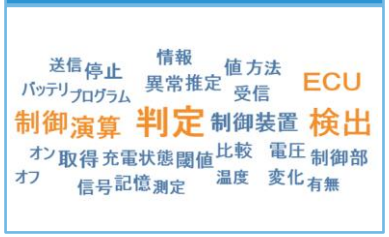
Tc05.二次電池の構成・製造方法



Tc06.空気や燃料電池などの温度制御

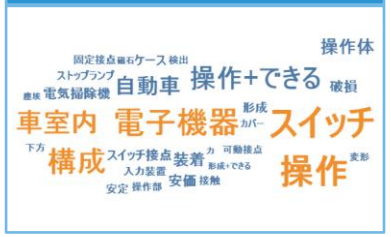


Tc07.状態の検出と判定

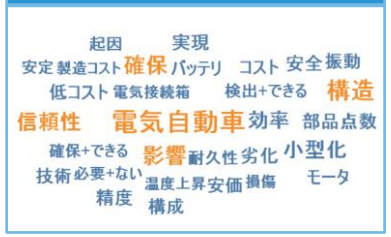


Tc08.形成・配置

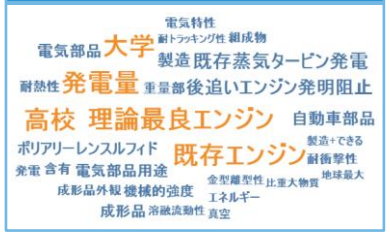
Tc09.スイッチの操作性



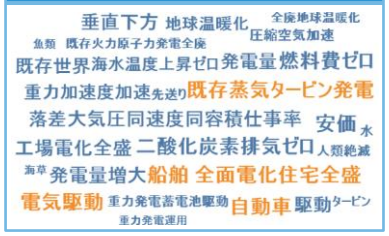
Tc10.小型化・低コスト化・簡素化・信頼性向上



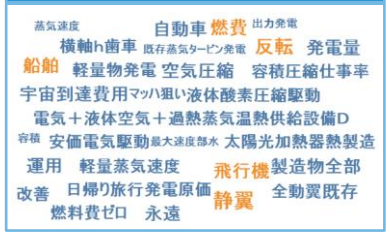
Tc11.既存エンジンへの警鐘・樹脂組成物の提供



Tc12.重力発電の活用による地球温暖化防止



Tc13.タービン発電の出力向上・燃費低減



Tc11,Tc12,Tc13は特定の出願人による重複した要約内容の特許から抽出された

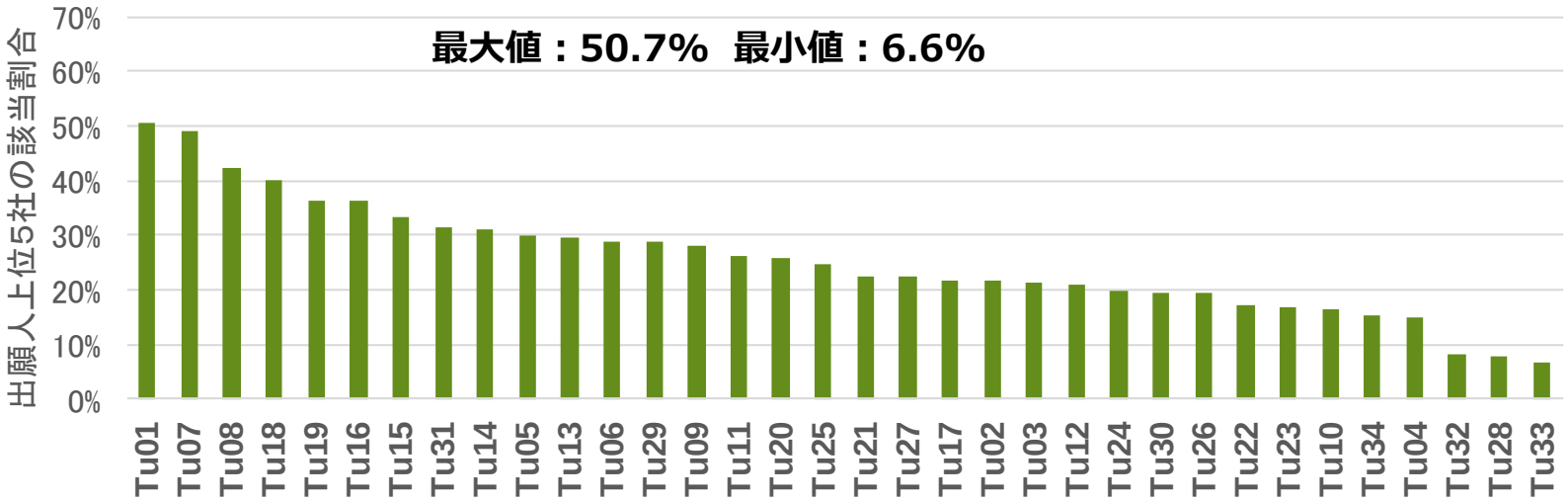
※文字の大きさはトピックに対する関係の強さを表現している （上位5つの単語を赤色で表示している）

© 2018 Analytics Design Lab Inc.

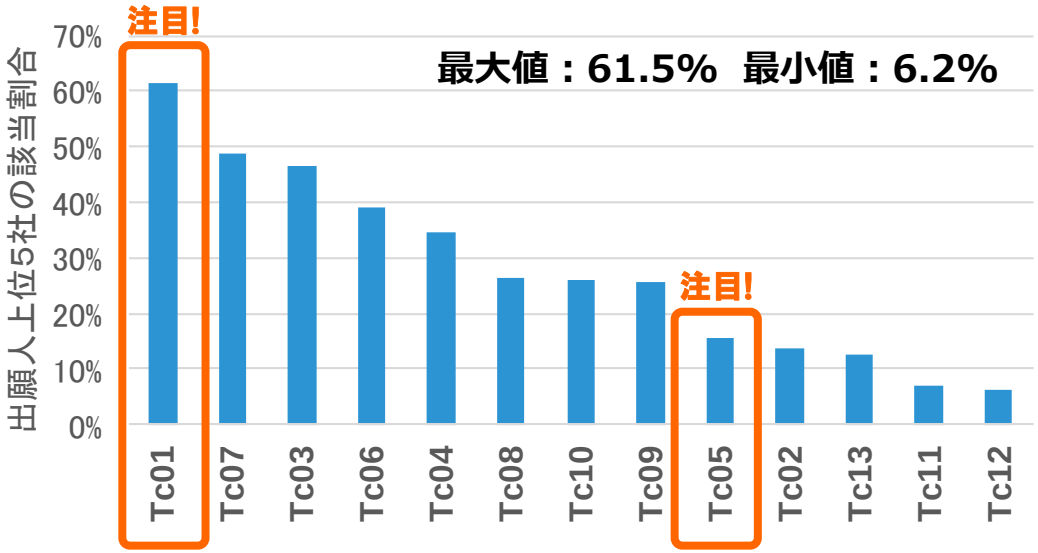
全体のトピックとの属性分布の比較

トピックごとに出願人上位５社データの該当割合を比較すると、どちらも高いものから低いものまで抽出されていますが、PCSAはその高低差が大きく、数も限定的に抽出されています

全体のトピック



PCSA適用の特徴トピック

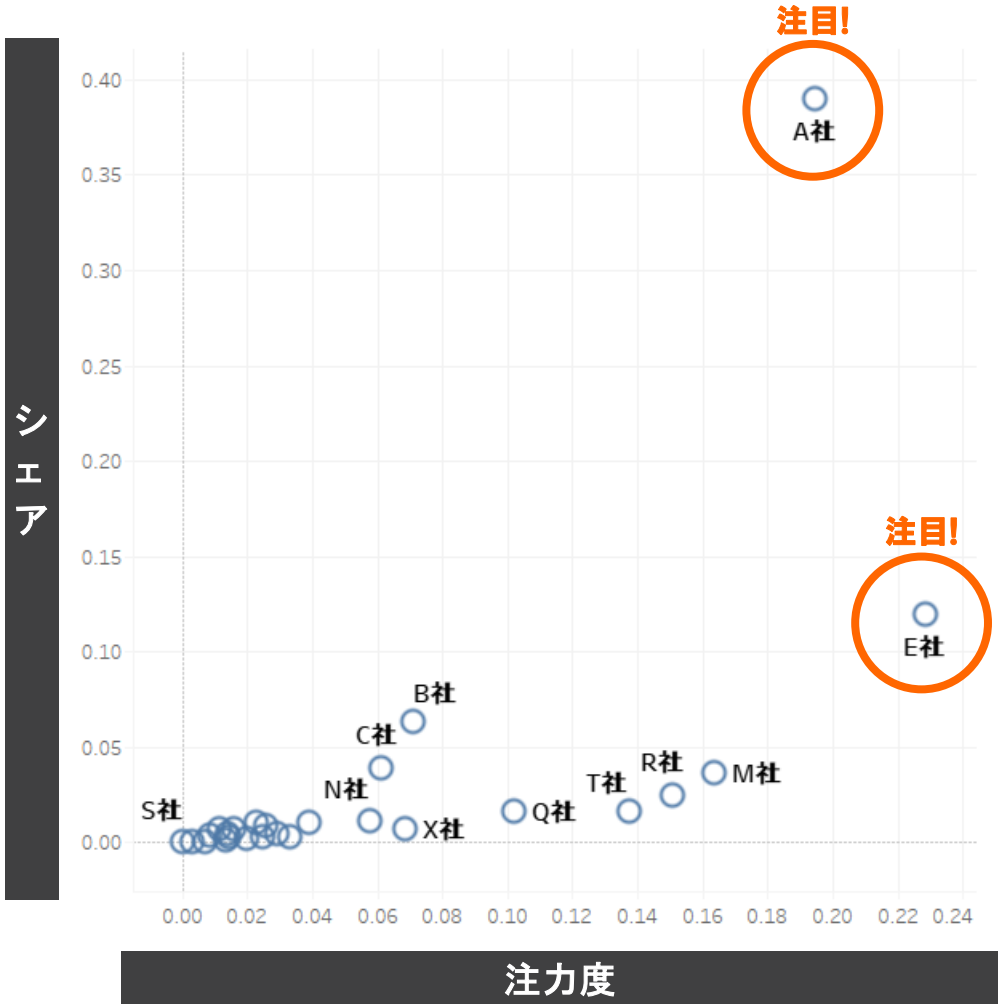


- それぞれ出願人上位5社のデータの該当割合の高い順にトピックを並べている
- 全体のトピックもPCSAのトピックも割合の高いものから低いものまで抽出されているが、PCSAではその高低差が大きく、最大値も高く、そうしたトピックが全体よりも限定的に抽出されている
- 元々の上位5社データの割合は25.0%

© 2018 Analytics Design Lab Inc.

エンジン駆動や動力伝達の制御に関するトピック(Tc01)は上位５社の該当割合が高い傾向にありますが、実際にメインプレーヤーはその５社のうちの２社で、どちらも注力度が高いです

注力度とシェアを軸とした出願人のポジショニング



結果の解釈

- 最もシェアの高いのはA社で他社を大きく突き放しており、2番手は大きくギャップを開けてE社であるが、E社も比較的高い水準のシェアを保有している
- 注力度も最も高いのはほぼ同じ水準でA社とE社であり、このトピックの領域ではこの2社がメインプレーヤーといえる

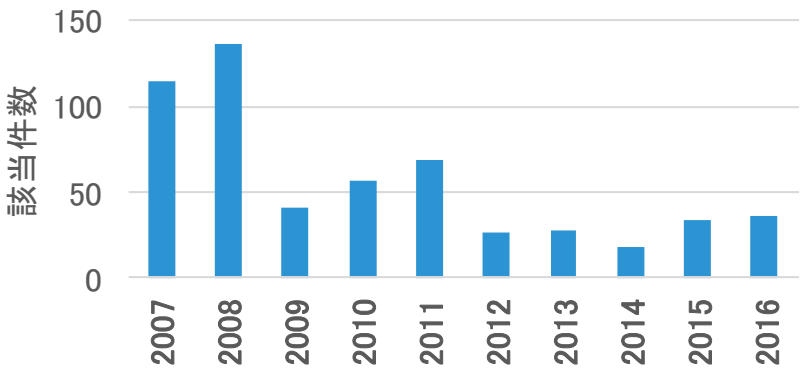
注力度とシェア （再掲）

- **注力度**： $P(\text{トピック}T = 1 \mid \text{出願人}X = 1)$
  - 出願人Xの出願特許の中で、どれくらいの割合がそのトピックTに該当するものか、つまり出願人がどれくらいそのトピックに注力しているのかを示している
- **シェア**： $P(\text{出願人}X = 1 \mid \text{トピック}T = 1)$
  - トピックTが該当する特許の中で、どれくらいの割合がその出願人Xの出願によるものか、つまりトピックのなかでどれくらいその出願人が占めているのかを示している

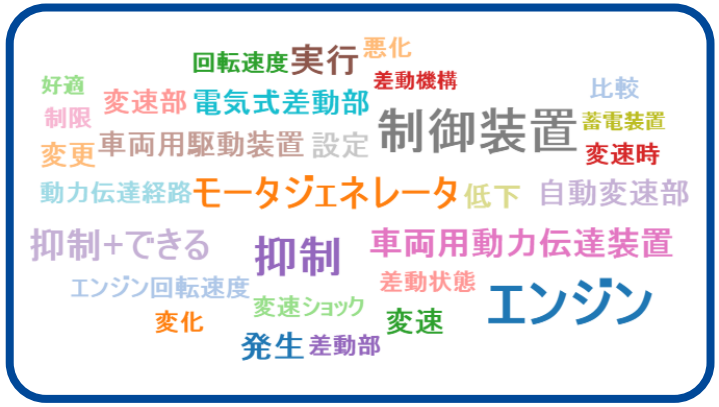
© 2018 Analytics Design Lab Inc.

圧倒的シェアのA社は差動や変速に関するものが多いですが出願が多いのは10年前で、シェア２位のE社はハイブリッド車やクラッチに関するものが多いですが最近は出願が減っています

A社（該当560件）の出願推移と要約キーワード

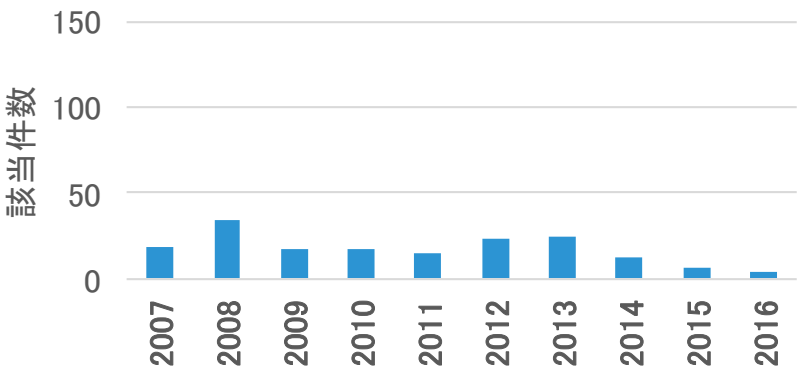


Tc01×A社の要約キーワード

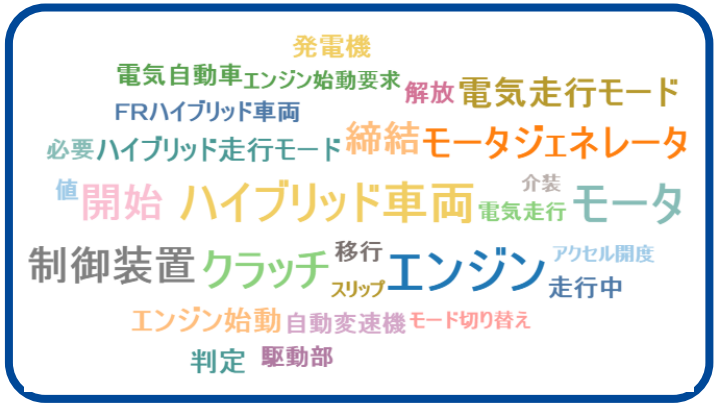


差動や回転速度・変速に関する動力制御のキーワードが多い

E社（該当172件）の出願推移と要約キーワード



Tc01×E社の要約キーワード

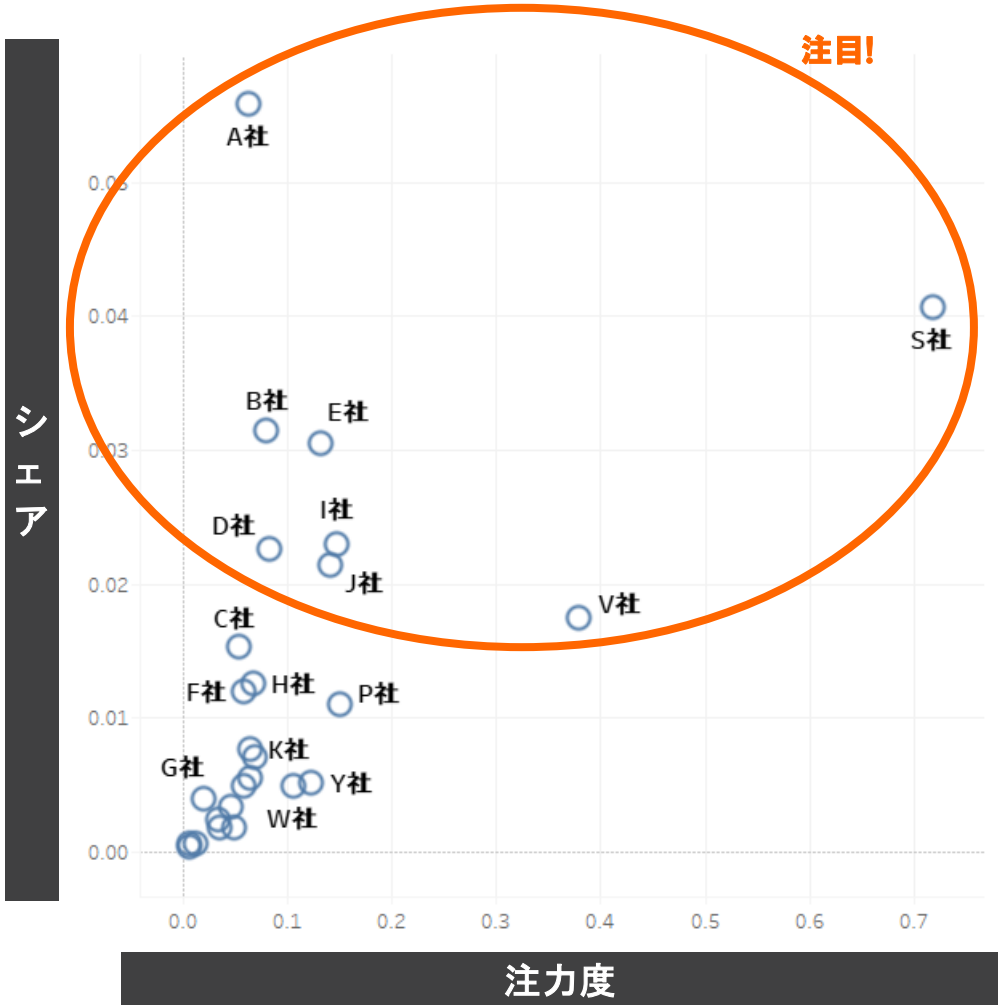


ハイブリッド車両やクラッチに関する動力制御のキーワードが多い

© 2018 Analytics Design Lab Inc.

二次電池に関するトピック(Tc05)は上位５社の該当割合が低い傾向にありますが、実際にプレーヤーが多く、５社以外の出願人も高シェア・高注力度の水準に位置しています

注力度とシェアを軸とした出願人のポジショニング



結果の解釈

- 最もシェアの高いのはA社で、次いでS社だが、他にも比較的シェアを保有している出願人も複数あり、プレーヤーが多い領域といえる
- 注力度ではS社が最も高く、他社を突き放しており、次いでV社の注力度も高い水準にある
- 巨大企業だけが占有するのではなく、各社が独自の技術を保有して棲み分けがされている可能性がある

注力度とシェア

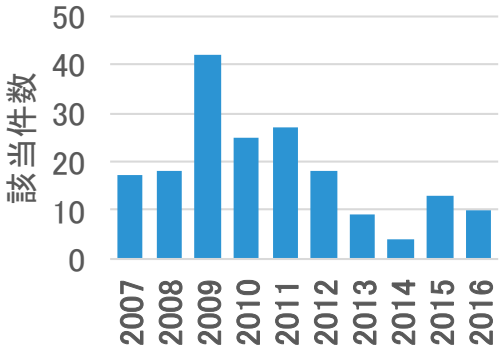
- **注力度**： $P(\text{トピック}T = 1 \mid \text{出願人}X = 1)$
  - 出願人Xの出願特許の中で、どれくらいの割合がそのトピックTに該当するものか、つまり出願人がどれくらいそのトピックに注力しているのかを示している
- **シェア**： $P(\text{出願人}X = 1 \mid \text{トピック}T = 1)$
  - トピックTが該当する特許の中で、どれくらいの割合がその出願人Xの出願によるものか、つまりトピックのなかでどれくらいその出願人が占めているのかを示している

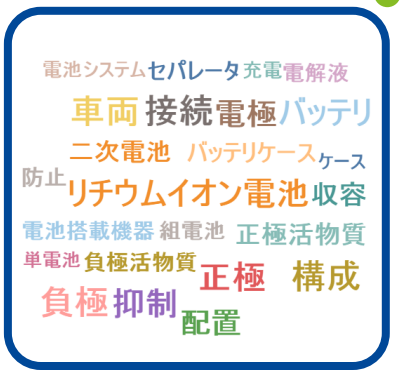
© 2018 Analytics Design Lab Inc.

PCSAトピックにおける各出願人の出願推移とキーワード:「Tc05.二次電池の構成・製造方法」①

シェア1位のA社は2009年が出願ピークでリチウムイオン電池に関するものが多く、シェア2位注力度1位のS社は年10~20件の安定した出願で電池モジュールに関するものが多いです

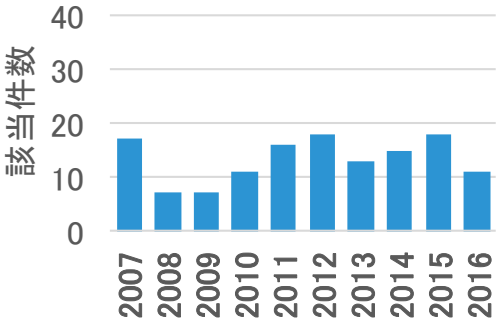
A社（該当183件）の出願推移と要約キーワード

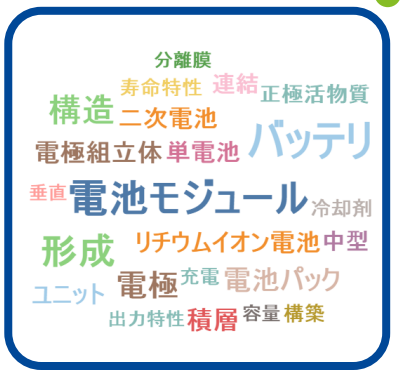




リチウムイオン電池や電極、接続・構成、収容ケースなど二次電池の全般的なキーワードが多い

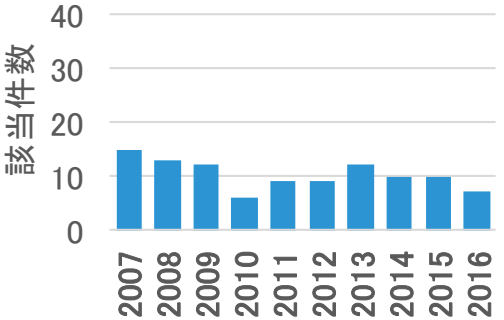
S社（該当133件）の出願推移と要約キーワード

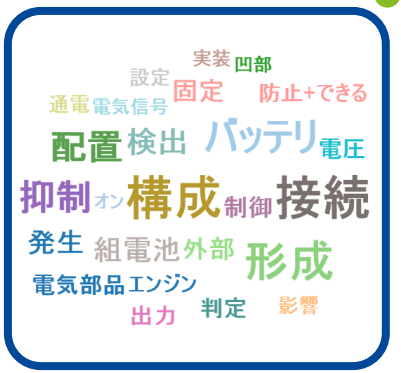




電池モジュールや電池パック、積層された電極組立体などに関するキーワードが多い

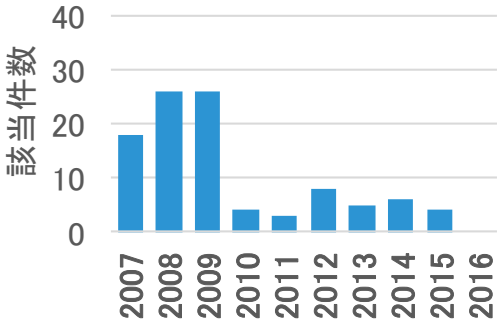
B社（該当103件）の出願推移と要約キーワード

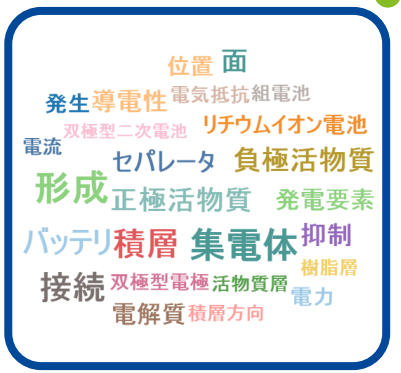




構成・形成・接続や組電池などに関するキーワードが多い

E社（該当100件）の出願推移と要約キーワード



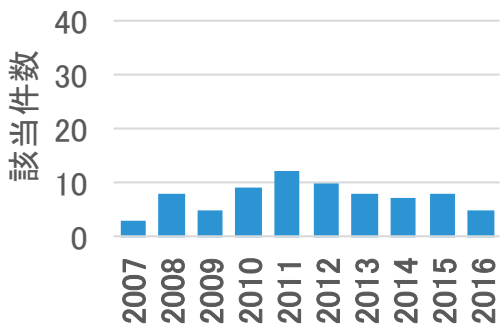


集電体、電極活物質、積層、双極型電池などに関するキーワードが多い

© 2018 Analytics Design Lab Inc.

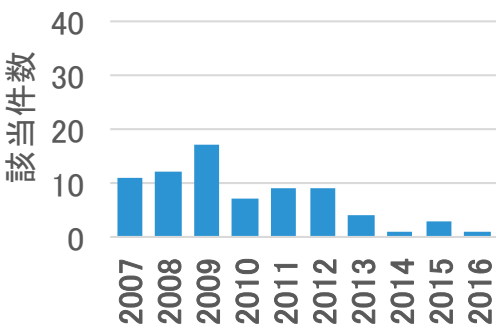
J社はシェア・注力度ともに高い水準ではないですが、直近で出願が急増しており、非水電解質や活物質に関するものが多く、注力度2位のV社は金属材料に関するものが多いです

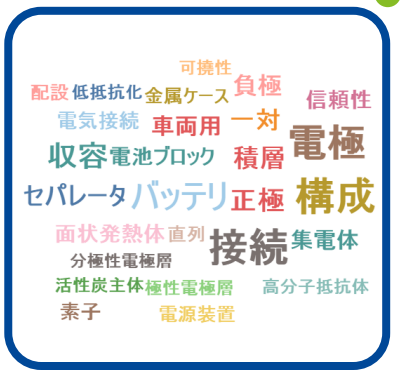
I社（該当75件）の出願推移と要約キーワード



ケース、ハウジングなど筐体に関するキーワードが多い

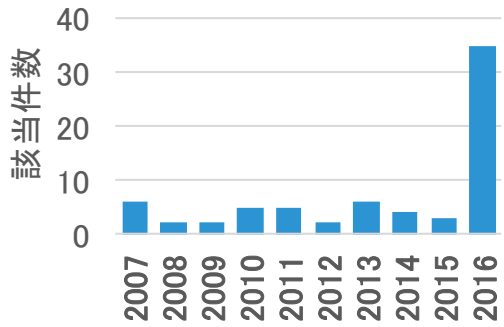
D社（該当74件）の出願推移と要約キーワード

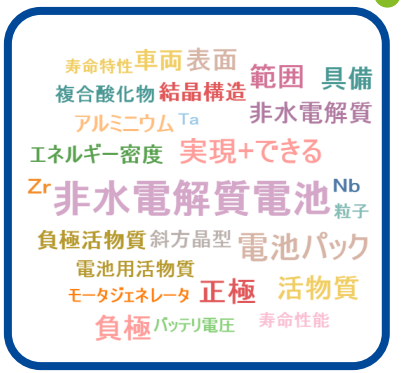




電極、セパレータ、構成などに関するキーワードが多い

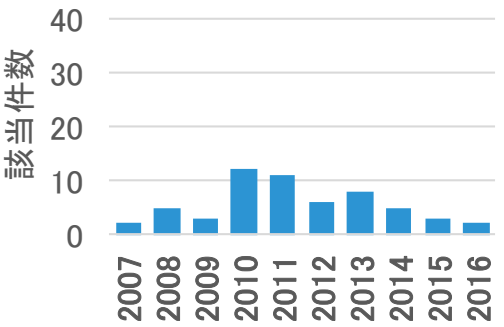
J社（該当70件）の出願推移と要約キーワード

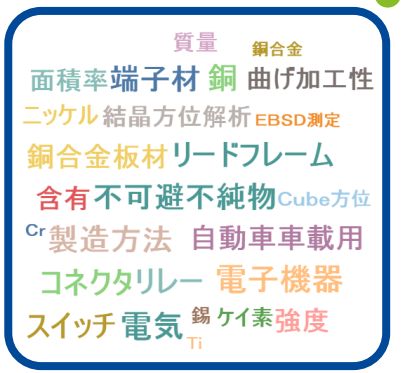




非水電解質、活物質に関するキーワードが多い

V社（該当57件）の出願推移と要約キーワード





電極の金属箔でも用いられる銅など金属材料に関するキーワードが多い

© 2018 Analytics Design Lab Inc.

J社が2016年に出願した特許には、自動車への搭載を想定した寿命性能、放電性能、低温性能などに優れた非水電解質電池の技術があり、一部は登録済みとなっています

Tc05×J社×2016年の該当特許35件から抜粋した課題要約文

発明の名称	登録	PIスコア
電池パック及び自動車	済み	2

優れた急速充放電性能及び優れた繰り返し充放電特性を示すことができる活物質を含む電池パック及び自動車を提供する

発明の名称	登録	PIスコア
組電池、電池パック、及び車両	済み	1

高いエネルギー密度及び高い電池電圧を示すことができ且つ寿命特性に優れた非水電解質電池を実現することができる電池用活物質を提供する

発明の名称	登録	PIスコア
組電池、電池パック及び自動車	済み	1

高いエネルギー密度、高い電池電圧、及び優れた寿命特性を示すことができ、電圧管理を容易に行うことができる非水電解質電池を実現する

発明の名称	登録	PIスコア
バイポーラ電池、電池パック及び車	済み	0

放電性能、低温性能、サイクル寿命性能及び高温貯蔵性能に優れた電気化学セルと、これを用いた電池パック及び車と、電気化学セルの製造方法とを提供する

発明の名称	登録	PIスコア
固体電解質、リチウム電池、電池パック、及び車両	ー	27

構成元素の化学的安定性を有しながら低コスト化が可能で、室温よりも低温でのリチウムイオン導電性が高い固体電解質、この固体電解質を含み、低温での放電レート性能が優れるリチウム電池、このリチウム電池を含む電池パック、及びこの電池パックが搭載された車両を提供する

発明の名称	登録	PIスコア
電極、非水電解質電池、電池パック、及び車両	ー	10

高温においても優れた寿命性能を示す非水電解質電池を実現できる電極、高温においても優れた寿命性能を示す非水電解質電池、この非水電解質電池を含む電池パック、及びこの非水電解質電池が搭載された車両を提供する

※登録の確認は2018年9月7日時点

© 2018 Analytics Design Lab Inc.

まとめ

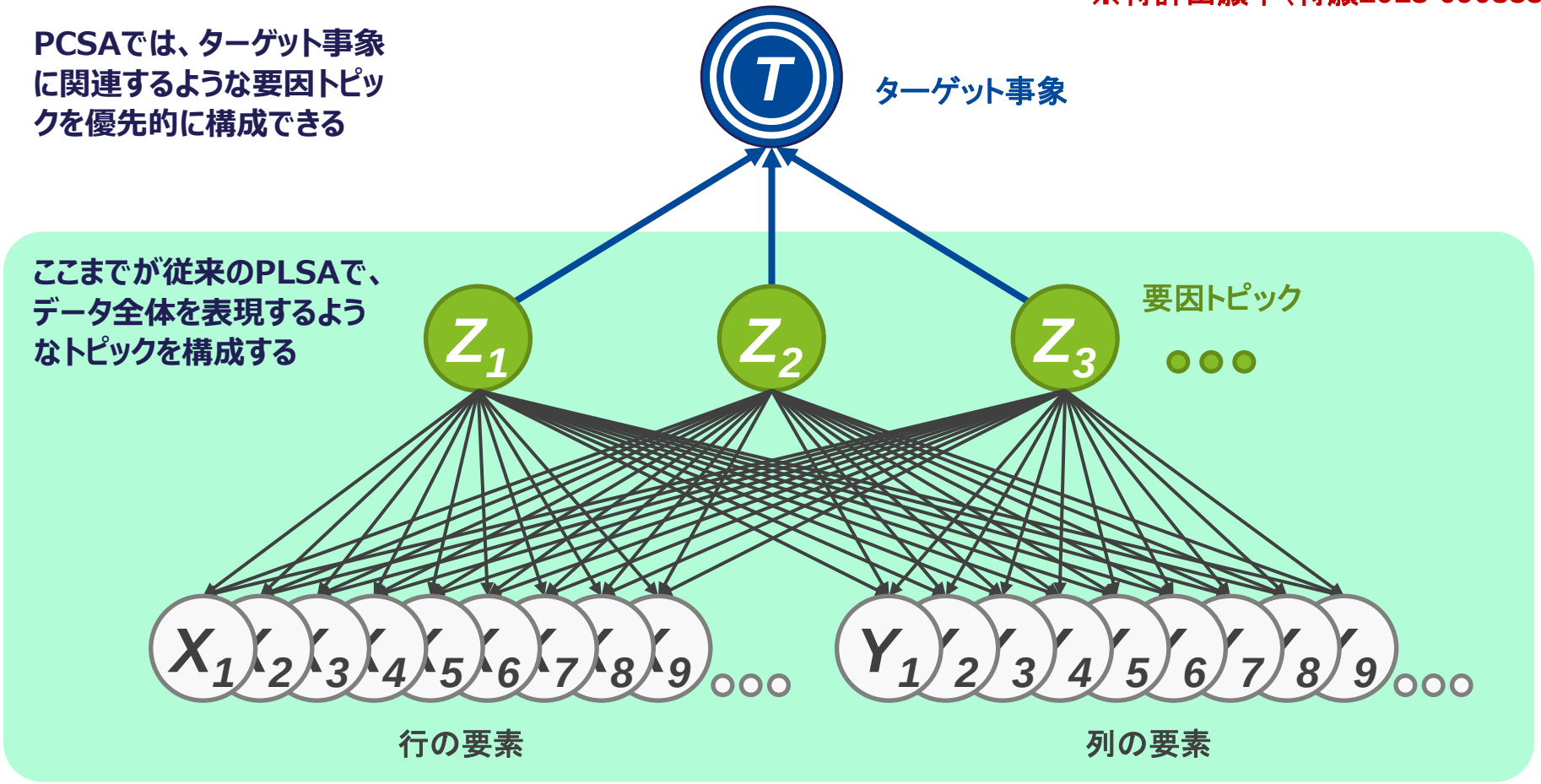
© 2018 Analytics Design Lab Inc.

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、よりインサイトに近づく探索軸を獲得することができます

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

※特許出願中（特願2018-090885）

PCSAでは、ターゲット事象に関連するような要因トピックを優先的に構成できる



© 2018 Analytics Design Lab Inc.

PCSAで特許文書を分析することで、確認したい対象に合わせた特徴要因をテキスト情報から自動構成できるため、それを分析軸に探索することで効果的なインサイト獲得が期待できます

Step1 ターゲット事象の設定

出願年 最近の出願特許の特徴とは？

PIスコア スコアの高い有用技術の特徴とは？

出願人 出願件数の多い企業で牛耳られている技術の特徴、逆に棲み分けができている技術の特徴とは？

Step2 差分共起行列の作成とPLSAの適用による特徴トピックの抽出

テキストマイニングによる単語の抽出

単語	頻度
構成	4,997
制御	4,360
配置	3,895
モータ	3,486
…	…

係り受け	頻度
電力-供給	1,208
否-判定	517
モータ-駆動	460
バッテリ-充電	440
…	…

ターゲット事象の該当有無別に共起行列を2つ作成

共起行列U

	電力-供給	否-判定	モータ-駆動	バッテリ-充電
構成	118	33	36	33
制御	268	73	108	85
配置	69	2	29	8
モータ	239	61	494	58

共起行列A

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

共起行列B

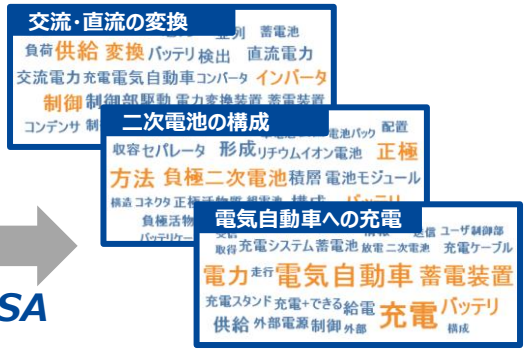
	電力-供給	否-判定	モータ-駆動	バッテリ-充電
構成	21	3	9	8
制御	60	24	28	19
配置	13	1	6	0
モータ	47	19	138	12

共起行列A-B

	電力-供給	否-判定	モータ-駆動	バッテリ-充電
構成	10.4	6.7	0.3	0.1
制御	7.7	[illegible]	[illegible]	2.4
配置	5.1	[illegible]	[illegible]	2.6
モータ	15.2	5.4	22.7	2.9

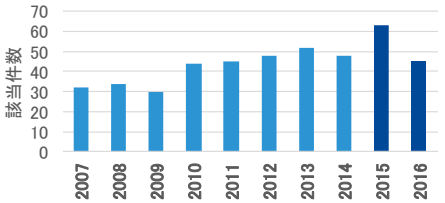
2つの共起行列の差分共起行列にPLSAを適用してトピックを抽出

PLSA

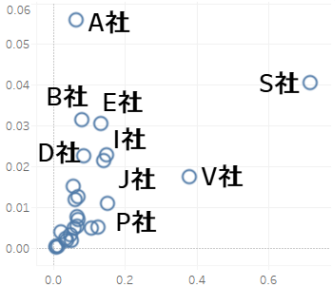


Step3 特徴トピックを軸とした探索的分析

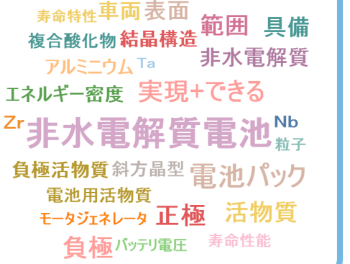
出願件数のトレンドを可視化



出願人のポジショニングを可視化



要約の特徴キーワードを可視化



© 2018 Analytics Design Lab Inc.

PCSAの適用対象例

PCSAの技術は様々な業務のテキストデータに適用でき、効果的なビジネスアクションの検討を支援します

口コミ
評価得点に影響を与える
口コミのトピックとは？

アンケート
顧客満足度に影響を与え
る自由記述トピックとは？

コールセンター履歴
解約・退会に影響を与える
問い合わせトピックとは？

特許文書
近年出願が伸びている
技術トピックとは？

営業日報
契約成立に影響を与える
営業活動トピックとは？

有価証券報告書
業績指標に影響を与える
定性的なトピックとは？

エントリーシート
選考通過に影響を与える
自己PRトピックとは？

診療記録
検査指標に影響を与える
定性的なトピックとは？

問題発生レポート
問題の発生に影響を与え
る報告記録トピックとは？

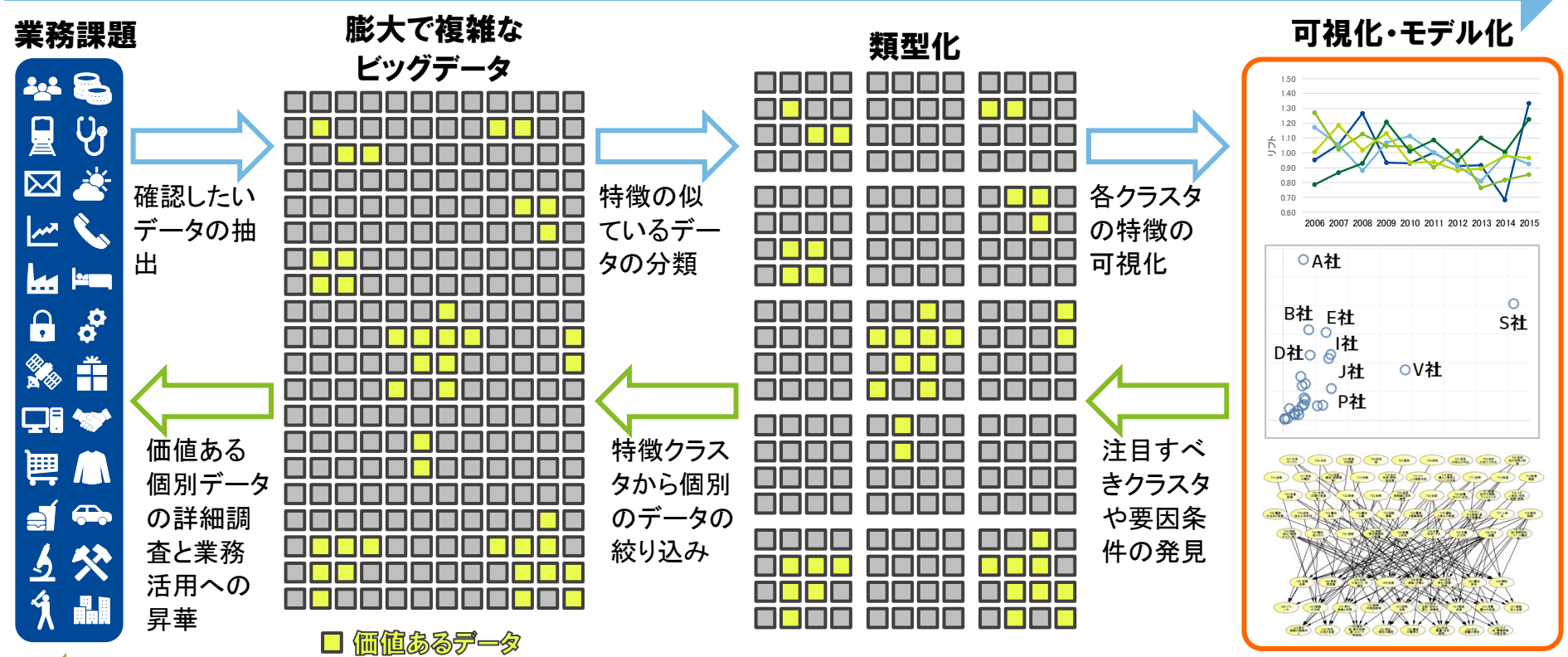
© 2018 Analytics Design Lab Inc.

ビッグデータからインサイト獲得のためのダブルアプローチ

ビッグデータはそのままでは複雑で理解不能なので、まず理解できる形に抽象化して特徴を発見しますが、今度は抽象度が高くて業務に活用できないので、その特徴の個別のデータに着目して再度具体化します

特徴を発見しやすくするために抽象化する（量的分析に軍配を上げ平均的な存在の間にある普遍的な特徴を得る）

マクロ探索アプローチ



ミクロ探索アプローチ

価値ある個別のデータを発見するために具体化する（質的分析に軍配を上げ平均の中の個別の特性を確認する）

© 2018 Analytics Design Lab Inc.

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

analytics.office@analyticsdlab.co.jp

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

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

株式会社アナリティクスデザインラボ

Analytics Design Lab

© 2018 Analytics Design Lab Inc.