

V 産業日本語研究会ワークショップ講演

平成 30 年 12 月 26 日に開催した産業日本語研究会ワークショップでは、「場」と言葉の関係を研究している言語系学者、および特許文書中の内容の解析と特許情報の引用情報の分析を行っている情報系学者から、2 件の講演をいただいた。以下に当日の講演内容を掲載する。

1. 「場」でわかる日本語～場の言語学への招待～

「場」でわかる日本語 ～場の言語学への招待～

東京学芸大学

岡 智之

場でわかる日本語～場の言語学への招待～

- 質問1: KYの意味を説明してください。
- 質問2: 「場の空気を読む」って外国人に説明できますか？
- 質問3: 「場」ってなんだろう。
- 日本語の理解にとって、「場」の理解は不可欠。

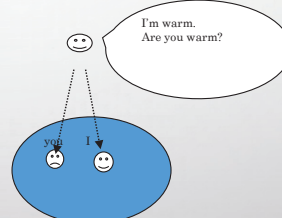
第1章 「場」があるから「主語」はいらない

- 質問： 次のような会話をどう思いますか？
- (1) A: 「今日、私は暖かいですけど、あなたは暖かいですか。桜が満開です。あなたはもう花見に行きましたか？」
- B: 「はい、私も暖かいです。私は昨日、京都に花見に行きました。」
- 日本人の会話では、
- (2) A: 「今日はあったかいね。桜が満開だよ。花見行った？」
- B: 「そうだね。昨日京都に行ってきたよ。」

- 図1 「あったかいね」



- 図2 「I'M WARM」



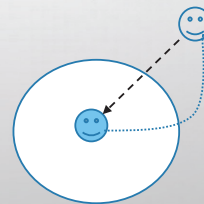
2. 「私はどこですか？」

- 質問： 初めて旅行した町で自分のいる場所がわからなくなった時、歩いている人に聞くとしたら、どう言いますか？
- 「ここはどこですか？」を英語でいうとどうなるでしょうか？

図3



図4



3. 「国境の長いトンネルを抜けると雪国であった」

- 川端康成の小説『雪国』の冒頭の文を、英語で訳すとどうなるでしょうか。
- (3) 国境の長いトンネルを抜けると雪国であった。
- (4) THE TRAIN CAME OUT OF THE LONG TUNNEL INTO THE SNOW COUNTRY.
- (3) の日本語文と、英語の翻訳を絵で描くとどうなるでしょうか。



• 図5 場外在的描写



図6 場内在的描写

4. 「富士山が見える」

- 「富士山が見える」を英語でいうとどうなるでしょうか。
- 「I CAN SEE MT. FUJI」「私は富士山を見ることができる」
- 図7



5. 「が」ってなに？

- 「鳥が飛んでいる」と「富士山が見える」の「が」の共通性は何でしょうか？

- 図8

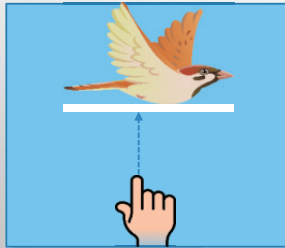
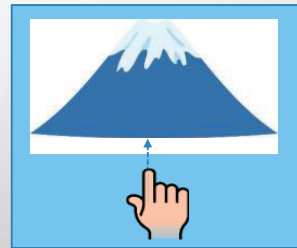


図7



- 「今この場」で一番目立つものを指す

6. 「は」って何？

- 外国人に「は」と「が」の違いを説明できるでしょうか？
- 「太郎は？」の質問にどう答える。
- 「太郎は、コーヒーは好きですが、紅茶は嫌いです。」 対比の「は」とは？

- 図9

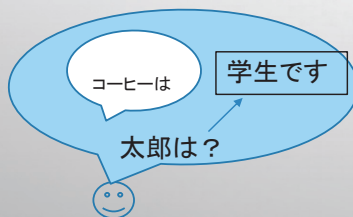


図10 「象は鼻が長い」



「は」は概念的「場」を設定する。

7. 「に」ってなに？

- 助詞「に」の例文を思いつくままにあげてください。
- 「に」をどう説明しますか？

- 図11 学校に行く(着点)

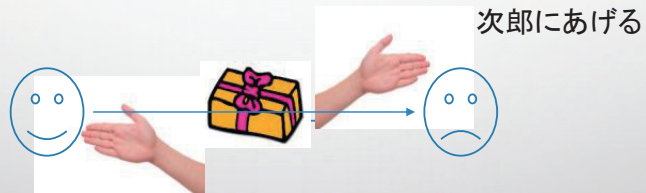


- 図12 机の上に本がある(存在の場所)



- 図13 授受の相手(～にあげる/もらう)

•



- 太郎にもらう
- (5)A. 太郎は次郎にプレゼントをあげた。(あげる相手)
- B. 次郎は太郎にプレゼントをもらった。(もらう相手)

8. 「で」ってなに？

- 「に」と「で」の違いは何でしょうか・

- 図14 グラウンドで走る



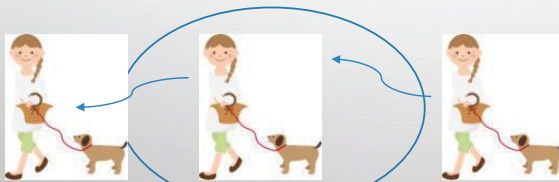
- 図15 グラウンドにいる



- (8) A. 車を踏切の前で止めた。
- B. 車を踏切の前に止めた。

9. 「を」ってなに？

- 「で」と「を」の違いは何でしょうか？
- 「公園を散歩する」「公園で散歩する」
- 図16



- 「山に登る」「山を登る」のちがいは？

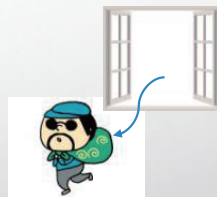
「私は毎日8時に家から出ます」

- 「家を出る」と「家から出る」の違いは何でしょうか？

• 図17



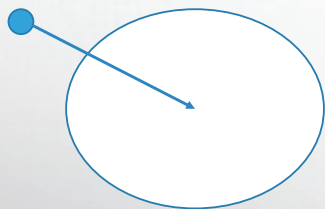
図18



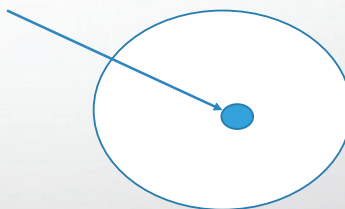
- 今日は家から一歩も出なかった。

格助詞のイメージ 「米洗う 前() 蛍が 二つ三つ」

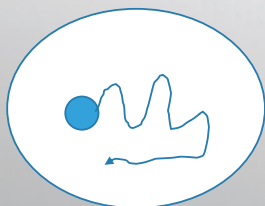
• 「へ」



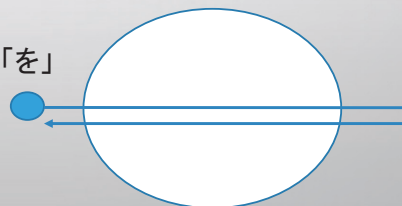
「に」



• 「で」



「を」



第2章 「場」においてコトがナル

1. 「こちらハンバーグ定食になります」

- 「こちらハンバーグ定食になります」は誤用でしょうか？
- では、次の日本語はどうでしょうか？
- 「右手に見えますのは、国会議事堂になります」
- 「トイレは、2階の奥になります」
- 「先生がお座りになった」
- ナルは必ずしも変化を表さない。事物の自然の成立を表す。
- 「柿の木に実がなる」「事がなった」

- どちらを使いますか？
- (6) A. 私たち結婚することにしました。
- B. 私たち結婚することになりました。
- 図19 春になった

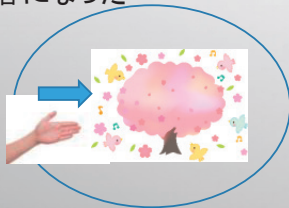
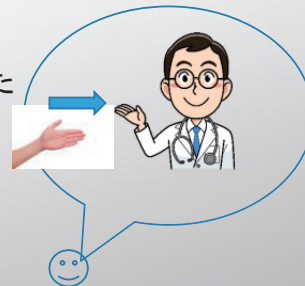


図20 太郎は医者になった



2. 「風〇窓〇開〇た」？

- 「風で窓が開いた」か「風が窓を開けた」か？
- 図21

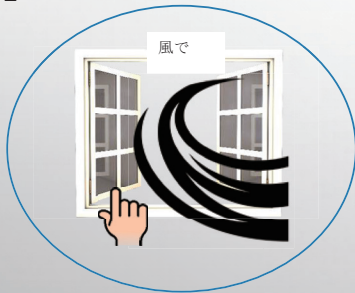
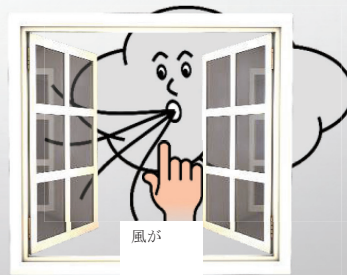


図22



3. 「鈴木さんは中国語を話することができる」

- 「できる」は「いでく」、出て来ること、「出来」(しゅったい)を表す。「コトがデキル」の文法
- 図23 赤ちゃんができた



図24 鈴木さんは中国語を話することができる



4. 「猫にネズミが食べられる」 場における出来文としてのラレル文

- 「られる」の意味は？受け身、可能、尊敬、自発の共通性は？

- 図25 ネズミは猫に食べられる



- 図26 猫はネズミが食べられる



場の言語学のパラダイム

- 西洋近代科学、哲学の原理
- 「主客分離」 「われ思う故に我あり」 主観(主体)が客観(対象)を観察する構図。自他分離。
- 「個物と因果関係」 ニュートン力学 個体が個体にエネルギーを伝えていく。
- 場の言語学のパラダイム
- 「主客非分離」 主体と客体、主観と客観は本来一体のもの。純粹経験。自他非分離。
- 「場における相互作用」 場の量子論。個体の働きは場のはたらきに規定される。
- 日本語文法を場の言語学の観点から捉え直す→日本語教育への応用

2. 「特許のコンテンツ・ネットワークマイニング」

特許のコンテンツ・ネットワークマイニング

国立大学法人長岡技術科学大学
情報・経営システム工学専攻 准教授
野中 尋史

経歴

主な経歴:

2005年3月 修士(工学), 豊橋技科大

2006年4月 NEDOフェロー,
豊橋技科大知財連携CD:知財(パテントマ
ップ作成等)・産学連携

2011年3月 博士(工学),
豊橋技科大(現成蹊大酒井先生, 現東大
坂地先生, 現理研小林先生らと特許コンテ
ンツマイニングの研究)

2015年4月 長岡技術科学大学講師

2018年12月 同大准教授

研究概要

コンテンツマイニング

「特許自動分類・解析」: 特許検索支援

大量の特許
文書



欲しい分野の
特許を自動
検索



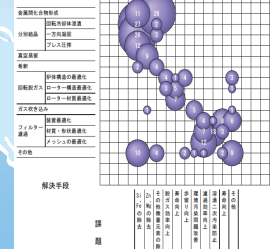
特願2008-9999
〇×システム

本発明は、〇
×システムに
関する。...

意味
解析



パテントマップ化

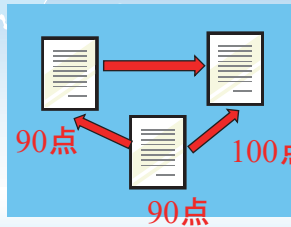


ネットワークマイニング

「特許文書の価値評価」: 知財・投資戦略支援



スコア
リング



株価・財務指標との相関
分析・予測



「特許自動分類・解析」: 特許検索支援

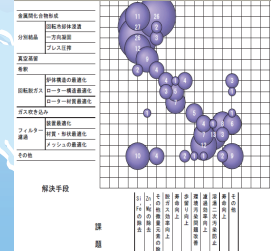
特願2008-9999
〇×システム

本発明は、〇
×システムに
関する。...

意味
解析

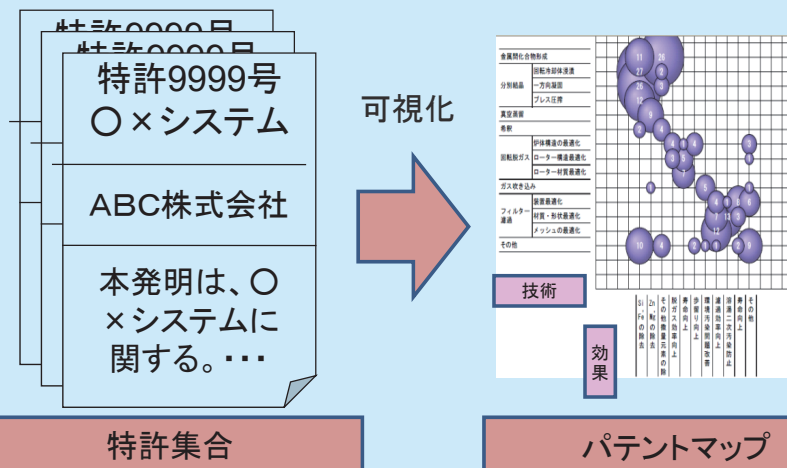


パテントマップ化



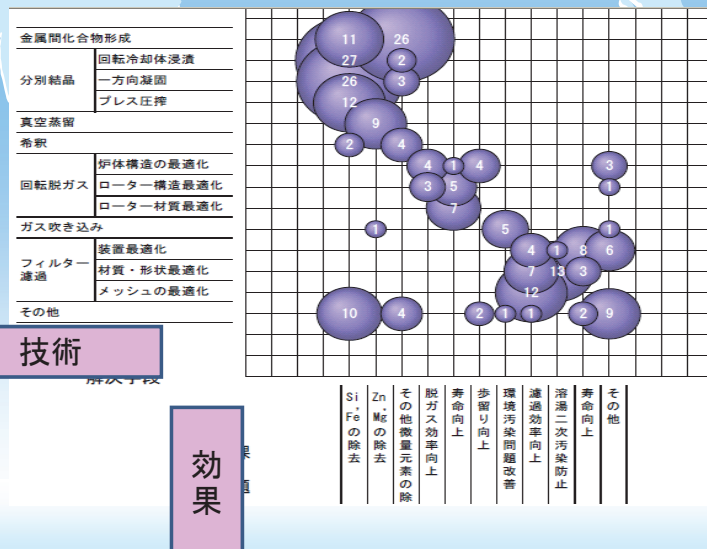
パテントマップとは

- ・特許出願・権利化動向を可視化したもの
 - ・出願戦略策定、研究開発戦略策定※等に使用
- ※研究開発着手時などにおいて、自社や他社の強み弱みについての分析などに使用
- 手作業だと多大な時間必要→情報抽出技術を利用して自動化！



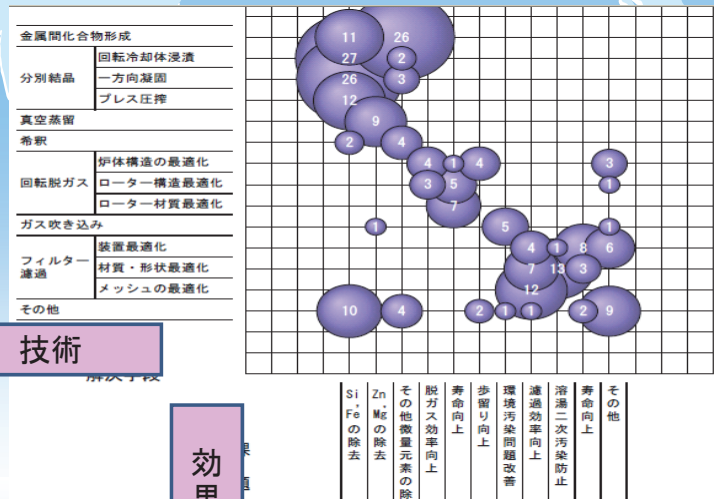
技術-効果型パテントマップ

特許文書がどのような技術で構成され(縦軸)どのような効果を発揮するのか(横軸)出願件数をバブルの大きさとするバブルチャートで表現...どのような(技術-効果)のペアが出願件数大きい(=要チェック分野)か一目で分かる ※特許文書を極限まで要約すると(技術-効果)のペアになること、特許庁の審査官が主に(技術-効果)に着目して審査すること、シーズニーズの分析ができることなどから重要



技術-効果型パテントマップ

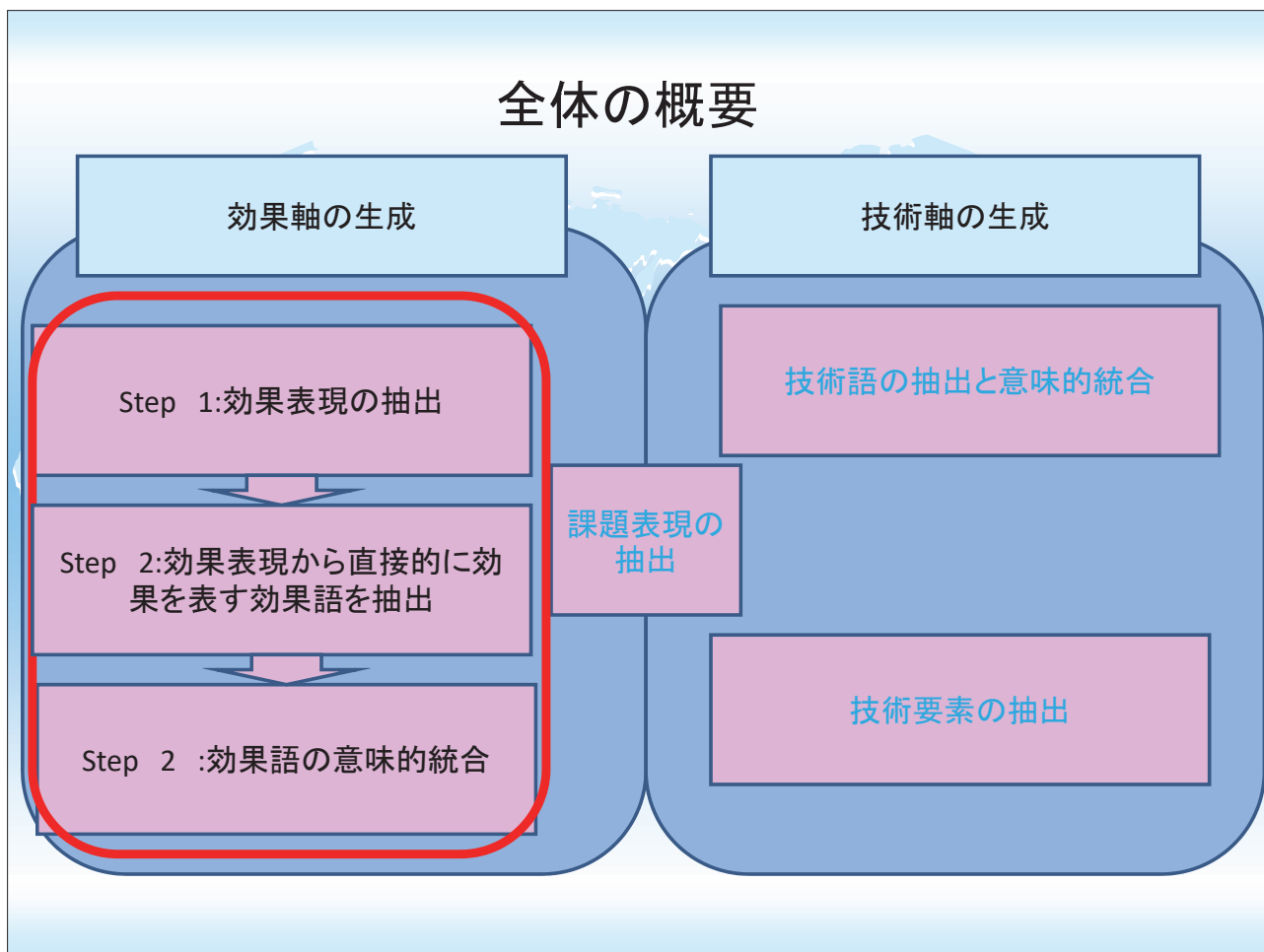
手動で作ると大変な手間→特許文書から自動的に効果と技術に関する語を抜き出そう！



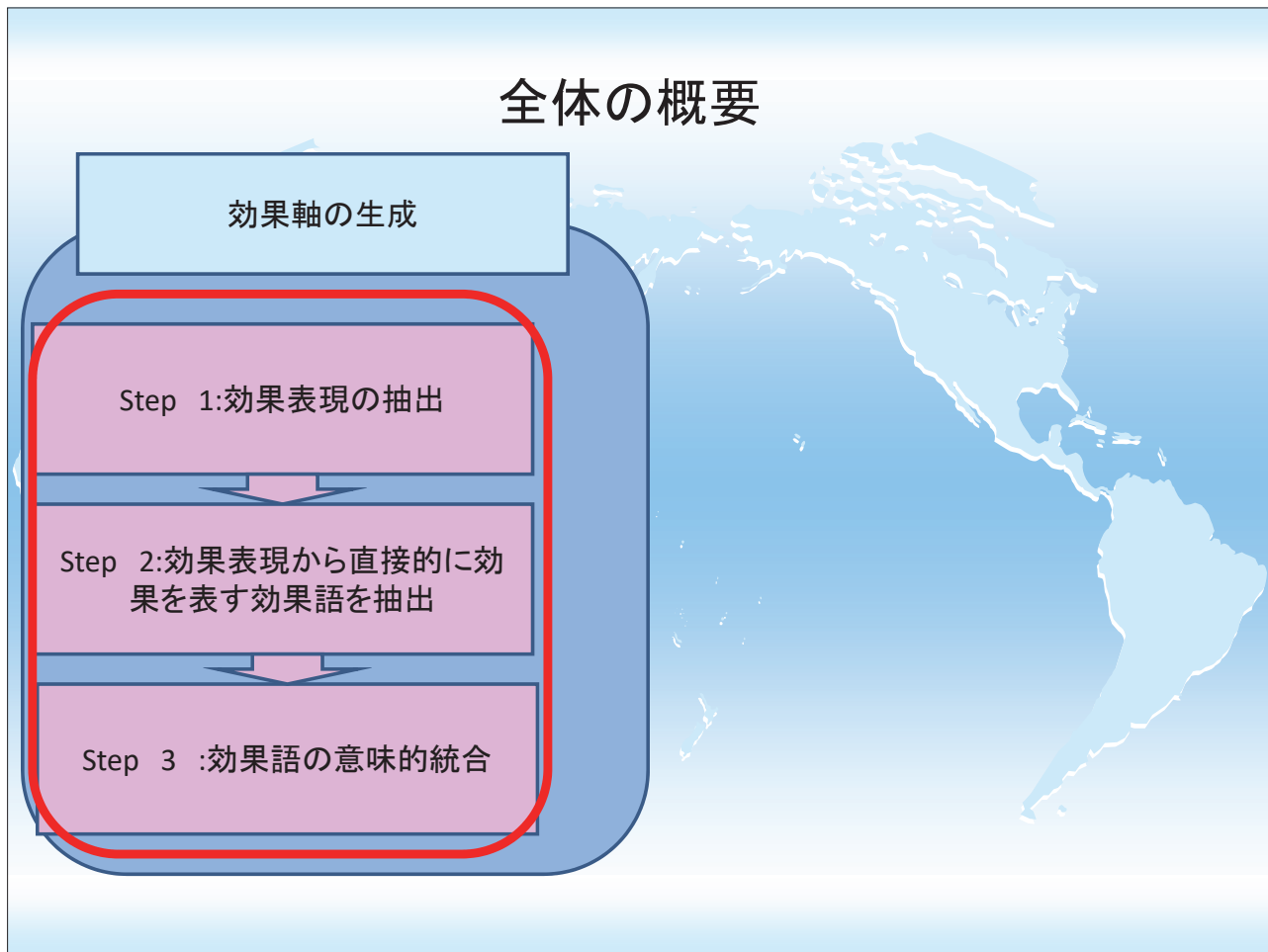
効果軸・技術軸に選択される語・表現の分類

対応する軸	内容・名称	例
効果軸	直接的なユーザの利点に相当する語・表現（効果語・効果表現）	耐久性
効果軸，技術軸	直接的なユーザの利点の根拠となる技術上のメリットに関する表現（課題表現）	「表面にコーティングをすること（により耐久性が向上）」
技術軸	技術の総称を意味する語（技術語）	熱可塑性樹脂
技術軸	技術要素・構成要件	生体情報を計測する計測手段

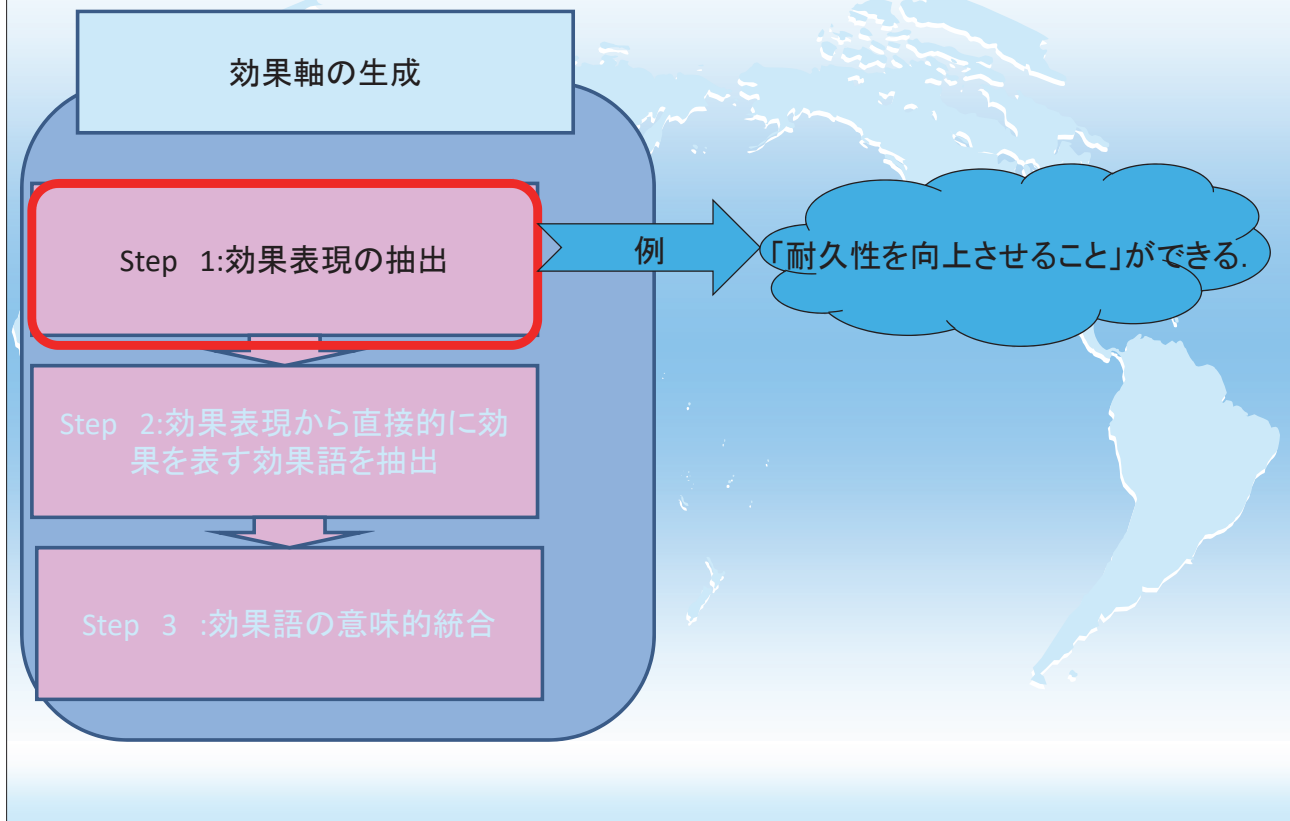
全体の概要



全体の概要



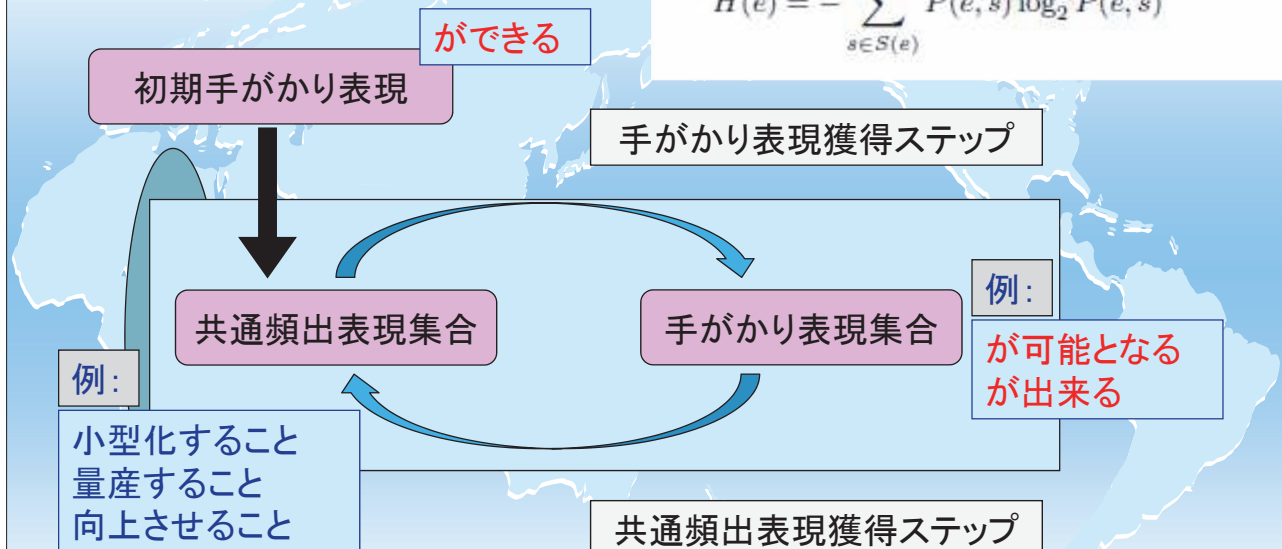
全体の概要



効果表現の抽出

発明の効果タグが付与されている箇所に適用

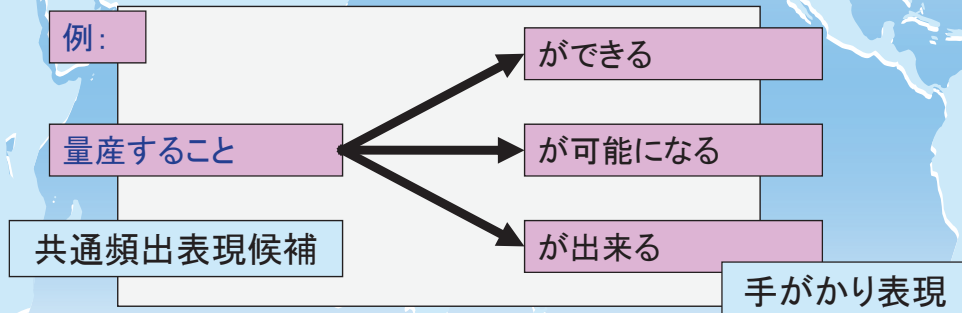
$$H(e) = - \sum_{s \in S(e)} P(e, s) \log_2 P(e, s)$$



※「手がかり表現獲得ステップ」、「共通頻出表現獲得ステップ」を繰り返す

共通頻出表現の選別

- ◆ 共通頻出表現候補から、適切な共通頻出表現を抽出
 - 様々な手がかり表現に係る共通頻出表現は適切

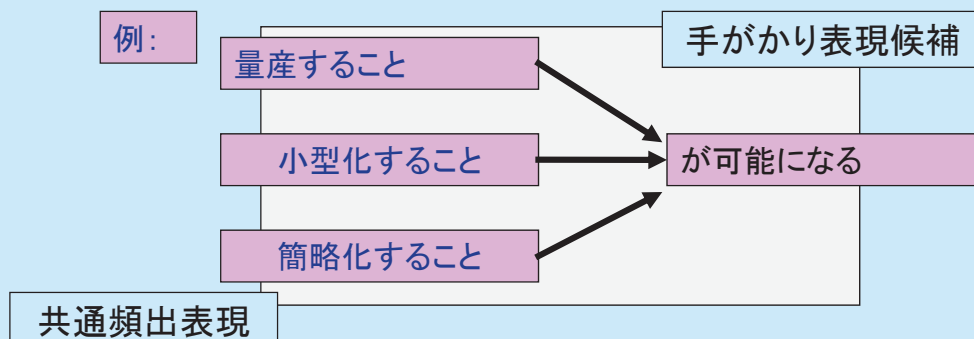


- ◆ 共通頻出表現が手がかり表現に係る確率に基づくエントロピー $H(e)$ を計算
 - ・ エントロピーが高い共通頻出表現候補 → 適切な共通頻出表現

13

手がかり表現の選別

- ◆ 手がかり表現候補から適切な手がかり表現を抽出
 - 様々な共通頻出表現に係る手がかり表現は適切



- ◆ 手がかり表現に対して共通頻出表現に係る確率に基づくエントロピー $H(s)$ を計算
 - ◆ エントロピーが高い手がかり表現候補 → 適切な手がかり表現

14

エントロピー $H(e)$

エントロピー

$$H(e) = - \sum_{s \in S(e)} P(e, s) \log_2 P(e, s)$$

$P(e, s)$: 共通頻出表現 e が手がかり表現 s に係る確率

$S(e)$: 共通頻出表現 e が係る手がかり表現の集合

閾値

$$T_e = \alpha \log_2 |Ns|$$

α : 定数 ($0 < \alpha < 1$)

Ns : 共通頻出表現を取得するのに
使用した手がかり表現の集合

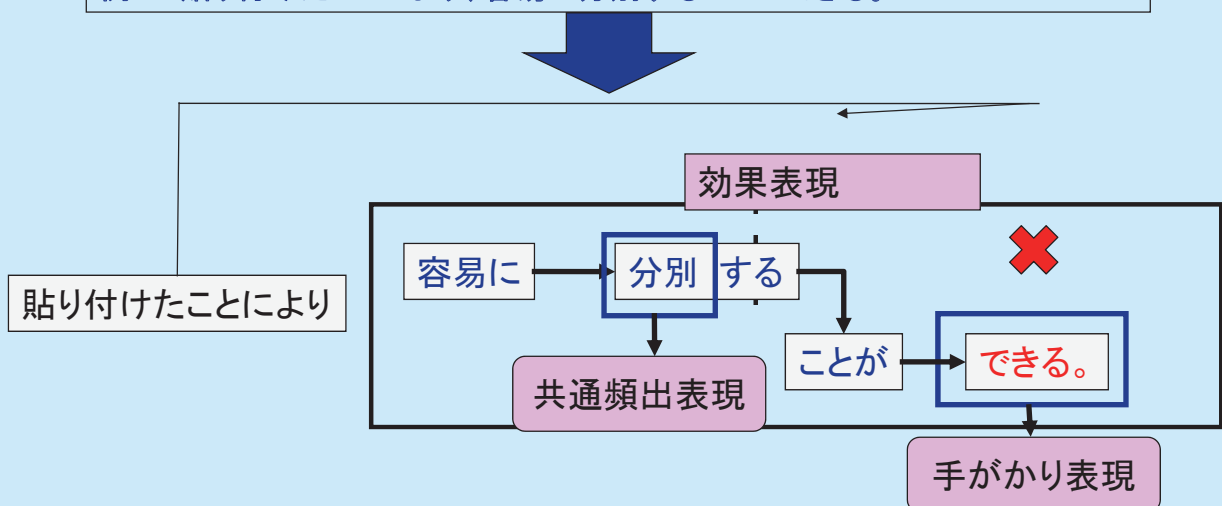
15

効果表現の抽出

Step 1: 抽出した手がかり表現を含む文節を Pclue とする

Step 2: Pclue の直前に出現する文節が Pclue に係る場合、その文節を加えたものを新たに Pclue と定義し、Step 2 を繰り返す。そうでない場合、Pclue を効果表現として出力する

例: ~貼り付けたことにより、容易に分別することができる。



効果表現の抽出

Step 1: 抽出した手がかり表現を含む文節をPclueとする
 Step 2: Pclueの直前に出現する文節がPclueに係る場合、その文節を加えたものを新たにPclueと定義し、Step 2を繰り返す。最終的なPclueを効果表現として出力する

例

F値は？

オリジナル版 0.78

Cross Bootstrap(課題表現と
対に抽出する改良版) 0.89

貼

手がかり表現

全体の概要

効果軸の生成

Step1:効果表現の抽出

Step2:表現から直接的に効果を表す効果語を抽出

Step3:効果語の意味的統合

Step 1-2 効果表現からの効果語抽出

☆ 共通頻出表現、手がかり表現との係受け関係に基づく効果語、技術語に関する抽象的パターンを利用 最終目的:効果語を抽出する

パターンの例

<効果語> <効果語手がかり> <技術語> <共通頻出表現> <手がかり表現>

上記の実例

機械特性 に優れた 熱可塑性樹脂を 提供すること ができる。

効果語 効果語手がかり 技術語

Step 1-2で利用する3つのパターン

☆ 共通頻出表現、手がかり表現との係受け関係に基づく効果語、技術語に関する3つの抽象的パターンを利用

※手がかり=手がかり表現

※共通頻出=共通頻出表現

パターン1

<効果語> <手がかり>

パターン2

<効果語> <共通頻出> <手がかり>

パターン3

<効果語> <効果語手がかり> <技術語> <共通頻出> <手がかり>

パターン1の概要

パターン1の例

<効果語> <手がかり表現>

上記の実例

防音する こと が できる。

使用する こと が できる。

パターン2の概要

パターン2の例

<効果語> <共通頻出表現> <手がかり表現>

上記の実例

機械強度 を 向上すること が できる。

製造費用 を 低減すること が できる。

パターン3の概要

※手がかり＝手がかり表現
※共通頻出＝共通頻出表現
※効果語手がかり：効果語と技術語を接続する手がかり表現

パターン3の例

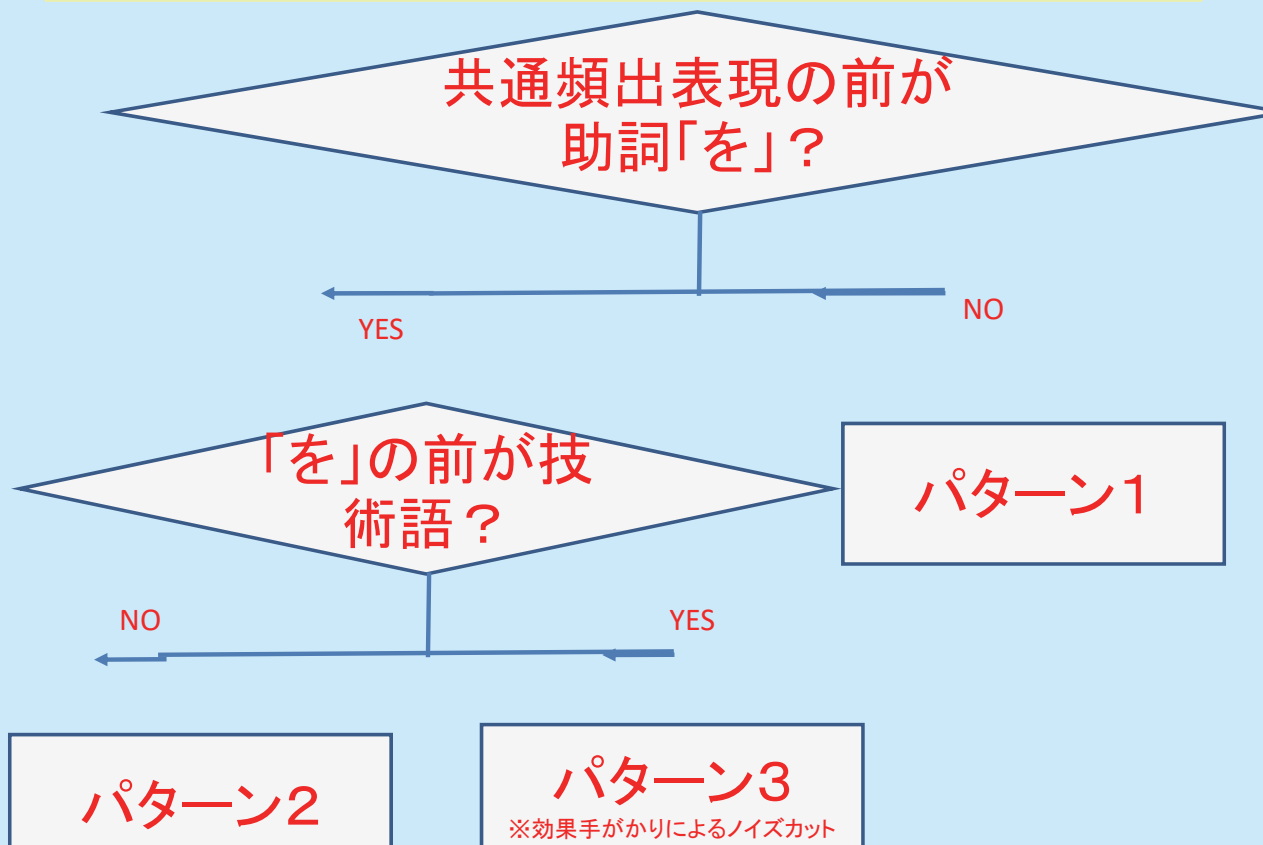
<効果語> <効果語手がかり> <技術語> <共通頻出> <手がかり>

上記の実例

機械特性 に優れた 熱可塑性樹脂を 提供すること ができる。

耐腐食性 を有した 熱可塑性樹脂を 提供すること ができる。

Step 1-2のアルゴリズム概要



全体の概要



n1

全体の概要



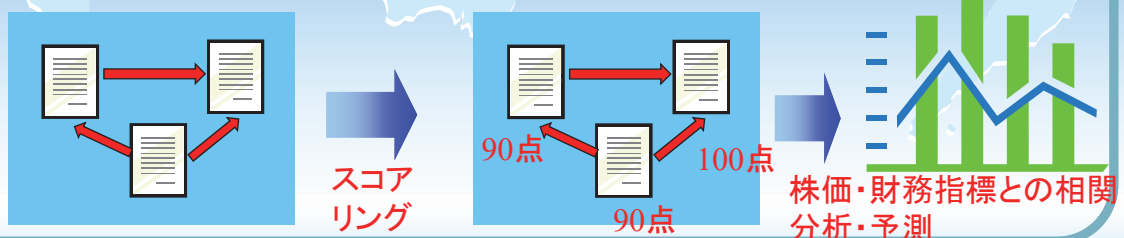
現在取組中

コンテンツマイニングのまとめと 今後

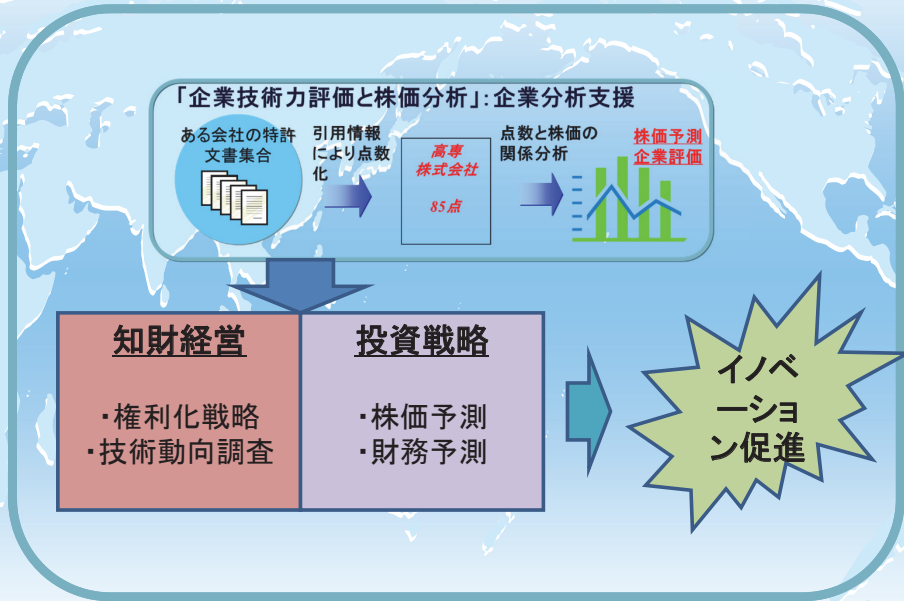
- ・ブートストラップ法ベースの表現抽出を基礎としたパテントマップ生成方法の提案
- ・意味的統合を行う
- ・再現率向上のため機械学習と組み合わせ
- ・他技術分野の評価

研究概要

「特許文書の価値評価」: 知財・投資戦略支援



特許文書の価値評価

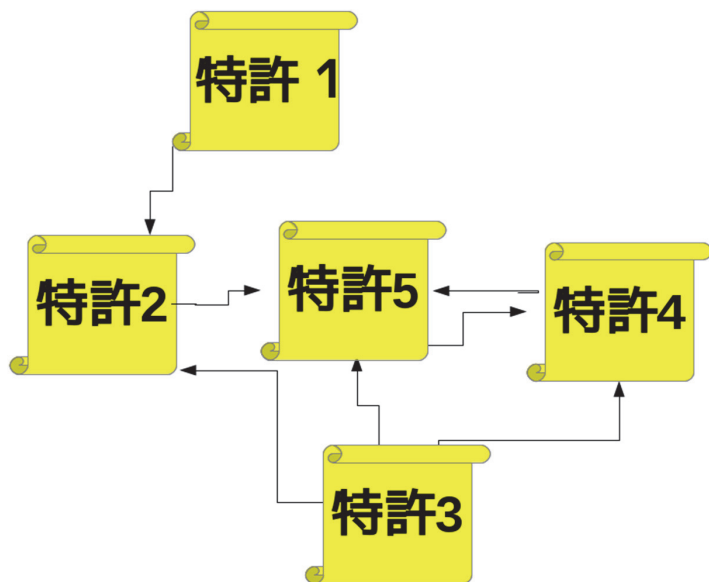


特許のネットワーク

☆出願人引用, 審査官引用, 共同出願, 発明者...



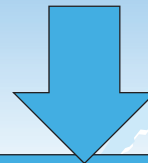
引用ネットワーク



特許文書：過去の特許との
違いについて、過去の特許文
書を引用しながら記載

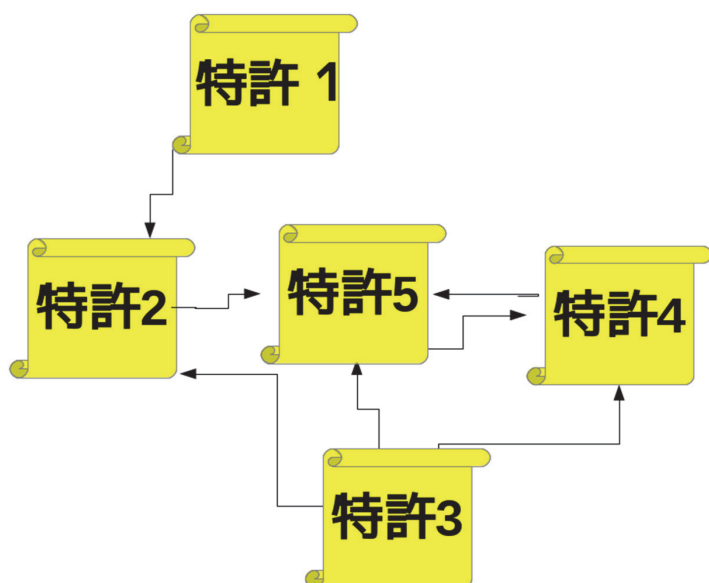


後世の特許から多く引用されて
いるほど重要な技術！



引用の数だけでよい？
引用ネットワークを見よう！

引用ネットワーク



HITS/PageRank
・引用の
連鎖構造あり
(特許2からの重
みは特許3から
の重みは単純な
引用数のだが、
子供を持つ特許
2のほうが大き
いのでは？)

HITSアルゴリズム(3)

- 数式で表すと...

$$\vec{a} = \alpha A^T \vec{h}$$

$$\vec{h} = \beta A \vec{a}$$

$\vec{a}, \vec{h}$: authorityスコア, hubスコア

A : リンク隣接行列

α, β : 収束化項

- 拡張HITS(≡Pagerank)を使用

$$\vec{a}_k = \zeta A^T A \vec{a}_{k-1} + (1 - \zeta) \vec{e}$$

技術経営分野での引用ネットワーク分析利用

Lukach, et al. 2007:

- Pagerank: 引用とPagerankでは性質異なる

佐藤, 岩山 2008:

- 引用, HITS: 重要特許(受賞)特定に引用有効

Nonaka, et al. 2014:

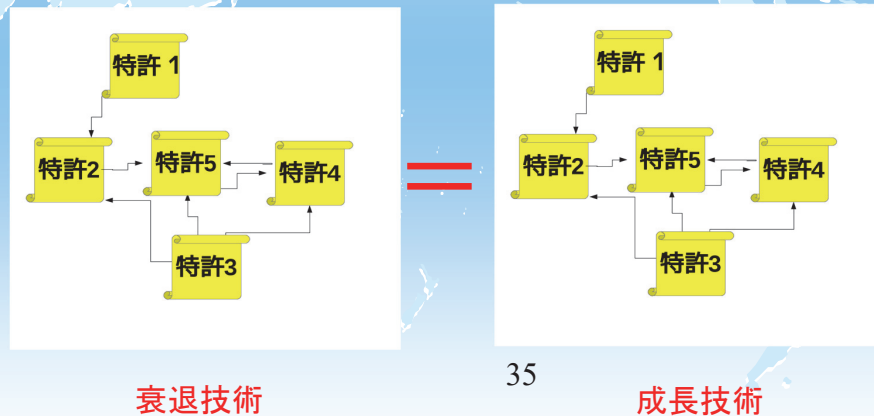
- Expanded HITS: 出願人レベルHITSベーススコアが業種や規模により株価と関連

Bruck 2018:

- PageRank: エマージングテクノロジーの発見にPageRankは有効(ただし少数サンプル)

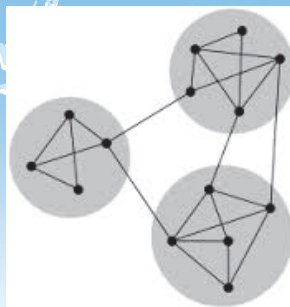
引用リンクを用いたスコアリングの問題

- 衰退技術分野と成長技術分野で同じ引用構造 = 同スコア！
- 衰退技術分野のコア特許 = 成長技術分野のコア特許！



特許引用ネットワークのコミュニティ

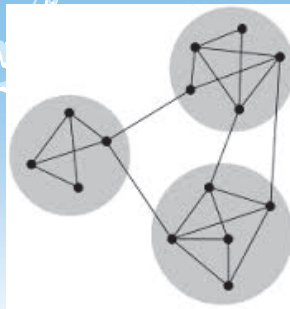
- HITSは1個の特許の価値を推定：個
- コミュニティ毎の特許の価値を推定：群
- コミュニティ特定 = リンクのクラスタリング = 技術分野



36

特許引用ネットワークのコミュニティの特徴

- 「コミュニティ特定 = リンクのクラスタリング = 技術分野」
- 審査官引用だと審査官が引用しやすい塊が抽出
- 出願人引用だと特定観点に基づく塊が抽出
- 塊の「成長性」を見ることができれば成長性に応じた得点



37

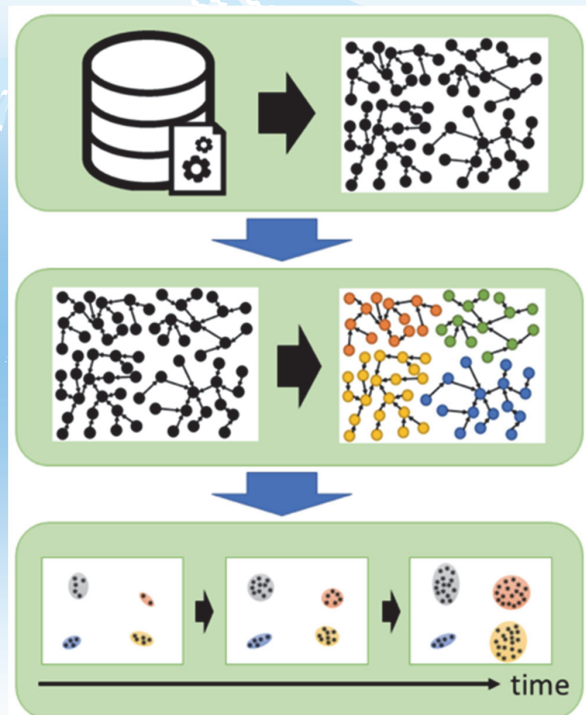
提案手法

1. 引用情報からネットワークを構築

2. Skip-Gramベースのネットワーク埋め込み(Node2vec) & クラスタリングを行う = 塊を見つける！

3. 各コミュニティにおける引用数の時系列予測 = 成長性評価：

- Hawkes process
- ARIMA
- Long Short Term Memory



Overview

1. 引用情報からネットワークを構築



引用データベース

知的財産研究所 (The Institute of Intellectual Property of Japan (IIP)) データベース

※ Akira Goto, Kazuyuki Motohashi (2007)

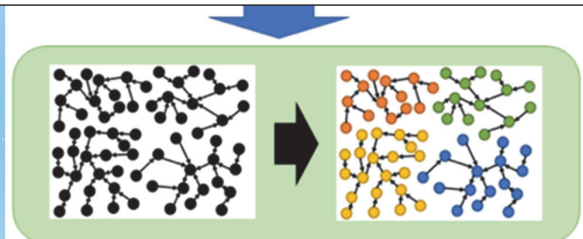
特許数 : 約6,000,000件

期間 : 1960-2017

引用種 : 審査官引用

Overview

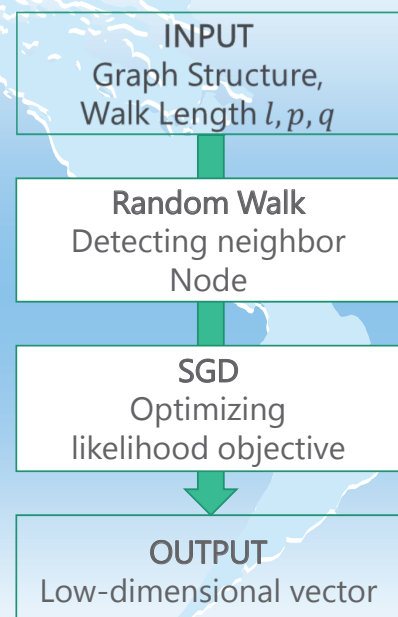
2. Skip-Gramベースのネットワーク埋め込み(Node2vec) & クラスタリングを行う



Node2vec

Skip-Gram (単語):
単語を周辺語彙の類似性に基づき (one-hot) ベクトル化

Skip-Gram (グラフ) = Node2Vec:
グラフ上の節点を隣接ノードの類似性に基づきベクトル化
→ このベクトルをクラスタリングすればコミュニティが抽出！



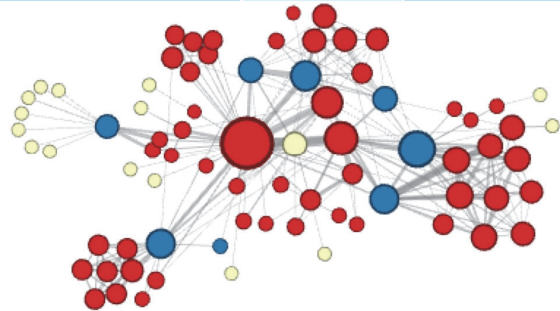
隣接ノードの定義とパラメータ

$$p = 1, q = 0.5$$



DFS-like behavior
Discovers communities

$$p = 1, q = 2$$



BFS-like behavior
Discovers structural roles

Figures source: Aditya Grover and Jure Leskovec, 2016. node2vec. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'16. 855-864.

Experiment of Communities Detection

Target technology field :

Roulette-like ball games (IPC code: A63F5)

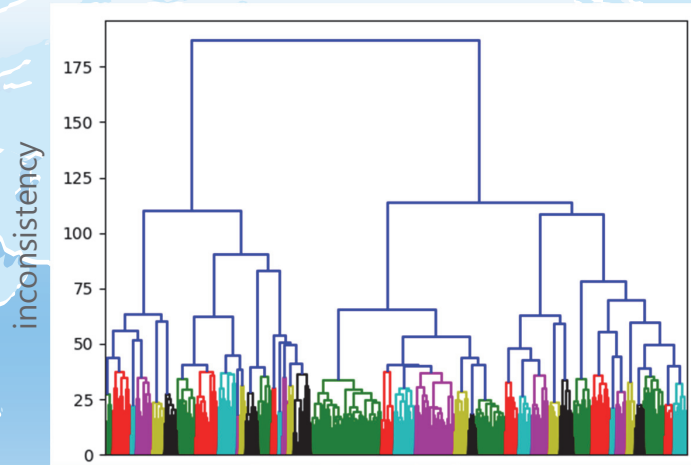
- Number of patent nodes : **29621**
- Number of cited links : **74529**

Node2vec parameters

$$p = 1, \quad q = 0.5, \quad l = 80$$

Apply hierarchical clustering (Ward's method) to
obtained low-dimensional vectors

Results of Communities



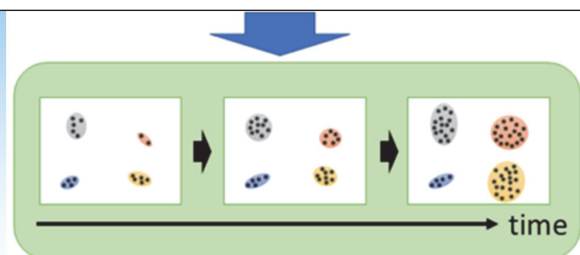
- Threshold value : **20 %**
(of the maximum value of the inconsistency value)

We obtained **39 clusters (正解率0.82)**

Overview

3. 各コミュニティにおける引用数の時系列予測 :

- Hawkes process
- ARIMA
- Long Short Term Memory



Hawkes Process

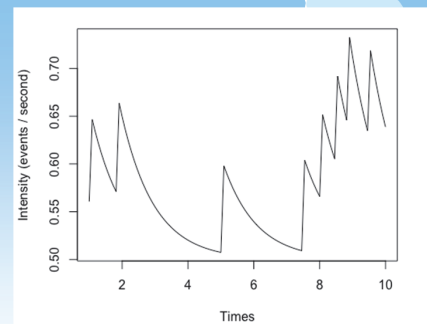
ポアソン過程＋自己励起

$$\lambda_t = \mu + \int_{-\infty}^t \Phi(t-u) dN_u$$

ポアソン過程：希少・ランダム

自己励起：1度おこると連鎖
地震，信用リスク・・・

特許引用も？ [HJ Jang 2017]



ARIMA

Autoregressive Integrated Moving Average
model

伝統的な統計的予測モデル

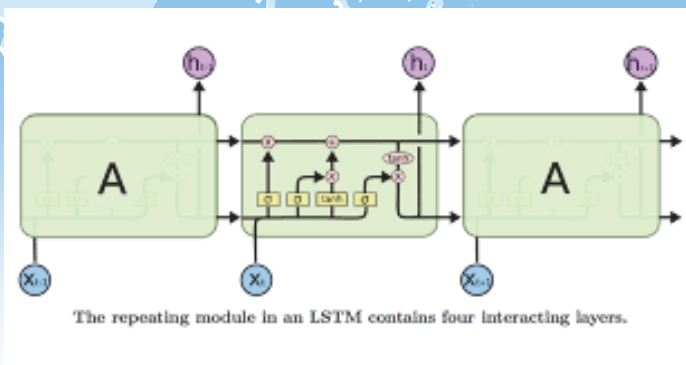
N階階差をとったデータ系列 y に対して過去の
時系列および誤差項の時系列の線形結合から
未来予測

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \\ + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_n \epsilon_{t-n}$$

LSTM

RNNに忘却ゲートを導入したモデル
勾配消失問題を解消し時間方向にディープ

邊土名朝飛2



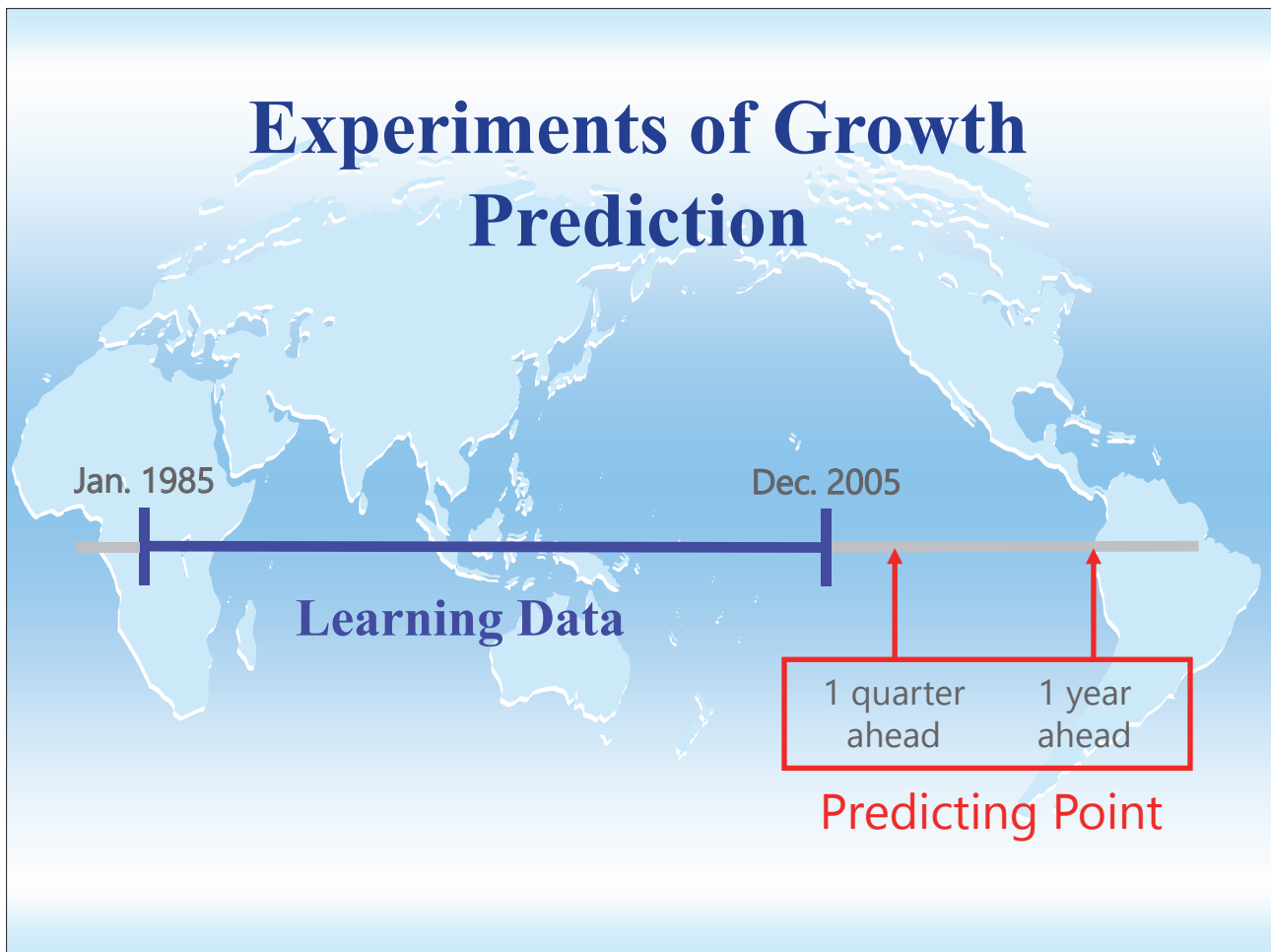
Experiments of Growth Prediction

Jan. 1985

Dec. 2005

Learning Data

Experiments of Growth Prediction



Parameters (Hawkes, ARIMA)

Parameter fitting for Hawkes process and ARIMA

Hawkes process

Fitting Parameters : α , β , μ

Method : Nelder–Mead method

ARIMA

Fitting Parameters : p , 1 , d

Method : Calculate parameters which minimize

Parameters (LSTM)

Cell input / output activation functions : ReLU

Network output activation function : Linear

Number of units : 128

Batch size : 32

Dropout rate : 0.2

Learning rate : 0.001

Optimization algorithm : Adam

MAPE Evaluation Index

$$\frac{1}{N} \sum_{i=1}^K \left| \frac{P_i(t) - R_i(t)}{R_i(t)} \right|$$

$P_i(t)$: Predicted number of patent applications of cluster i at time t

$R_i(t)$: Real number of patent applications of cluster i at time t

Direction Accuracy

$$\frac{TD(t)}{TD(t) + FD(t)}$$

$TD(t)$: Number of clusters that prediction direction is correct at time t

$FD(t)$: Number of clusters that prediction direction is incorrect at time t

予測性能

	Model	MAPE	Direction Accuracy
One quarter	Hawkes	95.22 %	43.59 %
	ARIMA	56.98 %	82.05 %
	LSTM	59.63 %	71.79 %
One year	Hawkes	88.48 %	58.97 %
	ARIMA	71.55 %	82.05 %
	LSTM	82.09 %	64.10 %

Discussion

- ・ARIMA: ベストパフォーマンス
→ 単位根過程を含むデータ (ADF 検定より13コミュニティ) に有効なことが生きた？
- ・LSTM: チューニング不足か？ グラフ畳み込みのほうが良かった？
- ・Hawkes過程はコミュニティレベルの引用ではそもそも希少性満たさない？

ネットワーク分析のまとめと今後

- ・Node2vec & 階層的クラスタリングで検出を行い、その引用成長性予測を行った。予測はARIMAがベストパフォーマンスであった。
- ・グラフ畳み込みの検討
- ・他技術分野への適用
- ・様々な価値評価指標との統合

全体のまとめ

- ・特許のコンテンツマイニングとして表現抽出の
パテントマップ生成への応用、特許のネットワ
ークマイニングとして引用ネットワークコミュニ
ティの引用成長性予測に関する手法の紹介を
行った
- ・教師・評価用データの作成大変。共有化できる
仕組みができれば・・・



ご清聴ありがとうございました！

— 禁無断転載 —

平成 30 年度
産業日本語研究会 報告書
「産業日本語」
Technical Japanese

平成 31 年 3 月

一般財団法人 日本特許情報機構 特許情報研究所

東京都江東区東陽 4 丁目 1 番 7 号

TEL 03-3615-5511