令和2年度 機械学習を活用した特許動向分析の調査

令和３年３月

特許庁

特許庁

1 本事業の概要

2 機械学習を活用した特許出願技術動向調査

3 本事業で採用する機械学習モデル

4 再現性の向上の検証

5 再現データの作成

6 新たな調査

7 総括

1.1 本事業の背景

特許庁では毎年、特許出願技術動向調査（以下、「動向調査」と呼ぶ）を実施

【現状】 市場創出に係る技術分野や、国の政策として推進すべき技術分野を中心に、
毎年十数テーマを選定

【意義】 ・先端技術分野等の出願状況や研究開発の方向性の明確化
・企業や大学等における研究開発テーマや技術開発の方向性の決定
・特許庁における審査・審判処理に関する基礎資料　として有用

【課題】 文献を**人手によって読み込んで分類（技術区分）を付与している**ため、コストが高い
・より多くの調査テーマについて調査したいというニーズに対応できていない
・一度動向調査を実施した調査テーマについて動向調査を継続できていない
・調査対象文献数が増加傾向にあり、人手による読み込みでは、文献数に限界がある

【対策】 **機械学習を利用した自動分類付与**の技術が急速に発展

⇒ 過去の動向調査結果を活用して機械学習を行うことにより技術区分を自動付与できれば、
動向調査を**低コストで継続的に実施**できる

1.2 本事業の経緯　1.3 本事業の目的

【経緯】

● 平成30年度に特許庁調査事業「機械学習を活用した特許動向分析の実行可能性調査」を実施
- ・検索結果文献集合から、調査対象でない文献を除外する「**ノイズ排除**」の精度を検証
- ・調査対象文献に対して分類（技術区分）を自動付与する「**技術区分付与**」の精度を検証
- ・3調査テーマを対象として精度を検証

● 令和元年度に特許庁調査事業「機械学習を活用した特許動向分析の調査」を実施
- ・ノイズ排除、技術区分付与の**精度のさらなる向上施策**の有効性を検証
- ・**調査対象テーマを拡張**（4調査テーマ）して精度を検証
- ・**調査対象年「以降」の文献に対してノイズ排除、技術区分付与**を実施し、特許動向を分析

【目的】

上記背景・経緯を踏まえ、機械学習を利用した、動向調査における特許動向分析の再現及び更新の実現性について**さらなる検証**を行い、**動向調査の効率化や新たな手法**について調査する

1.4 本事業の内容

過去に実施された人手による動向調査の結果を学習データとして機械学習を行い、
以下の項目について、さらなる調査・検証を実施

（1）再現性向上の検証（4章）

平成30年度調査事業及び令和元年度調査事業の結果を踏まえ、
ノイズ排除及び技術区分付与の**さらなる精度向上**の可能性について調査・検証を行う

（2）再現データの作成（5章）

平成30年度調査事業の3調査テーマ、令和元年度調査事業の4調査テーマ以外の
新たな4調査テーマを対象として、

- 調査対象年の文献について、ノイズ排除及び技術区分付与の精度を検証する
- 調査対象年「以降」の文献について、ノイズ排除及び技術区分付与を実施する

（3）新たな調査（6章）

上記以外に以下の調査を行う

- 機械学習結果の有効期間の調査
- 動向調査に必要なリソースの調査

2.1 機械学習の位置付け

実際の動向調査における作業プロセス(下図)における、以下の作業に機械学習を適用する

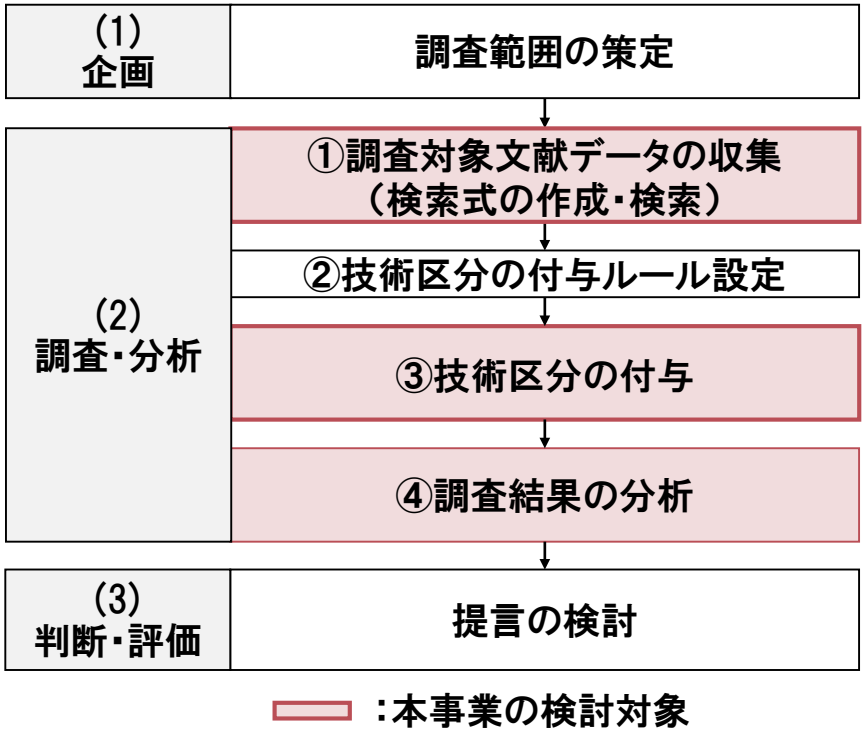
① 調査対象文献データの収集(**ノイズ排除**)

検索式によって検索された文献(検索式文献)の集合から、調査対象でないノイズ文献を除外し、技術区分の付与対象となる文献(調査対象文献)を収集する

③ **技術区分の付与**、④ 調査結果の分析

調査対象文献に対して分類(技術区分)を自動付与し、付与結果を分析する

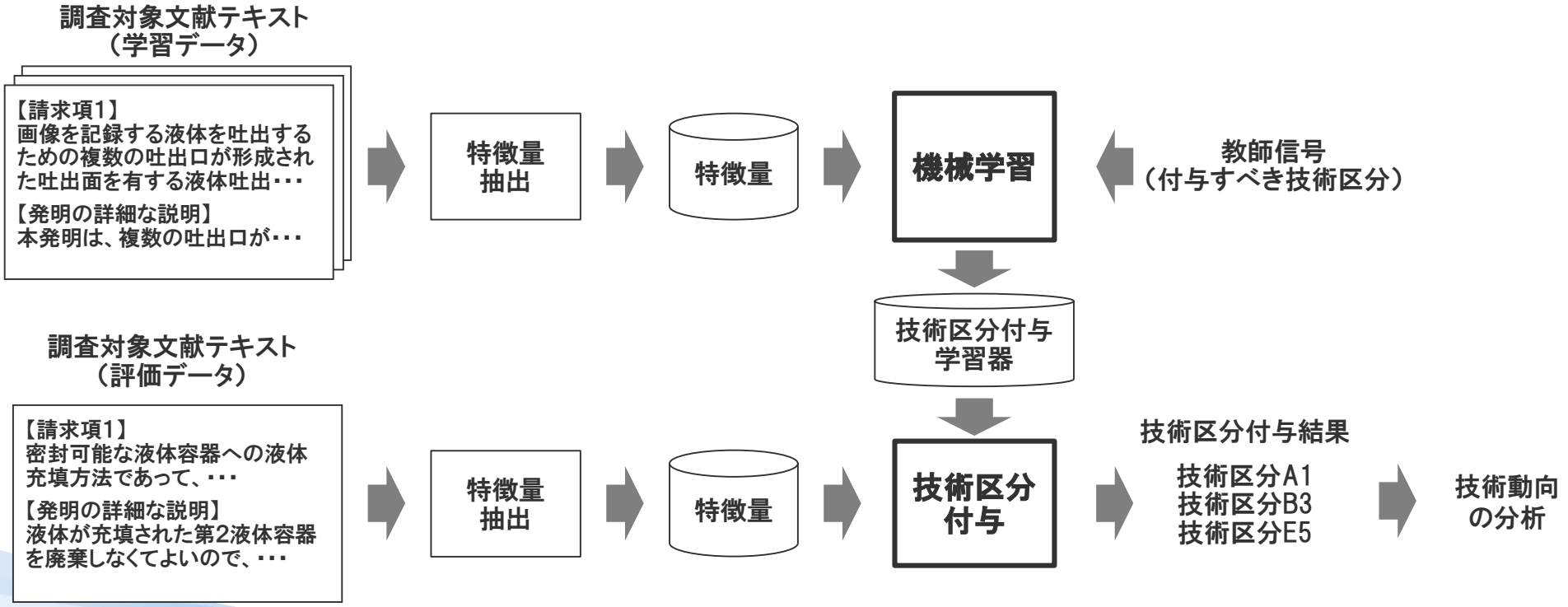
＜動向調査の作業プロセス＞



2.2 機械学習の動向調査への適用イメージ

過去に実施された人手による動向調査の結果を学習データとして機械学習を行い、
動向調査の対象年「以降」の文献に学習結果を適用して、ノイズ排除・技術区分付与を行う

＜技術区分付与に機械学習を適用するイメージ＞



3.1 採用する機械学習モデル

平成30年度調査事業及び令和元年度調査事業で採用した2モデル(SVM, MH-NAM)を流用する

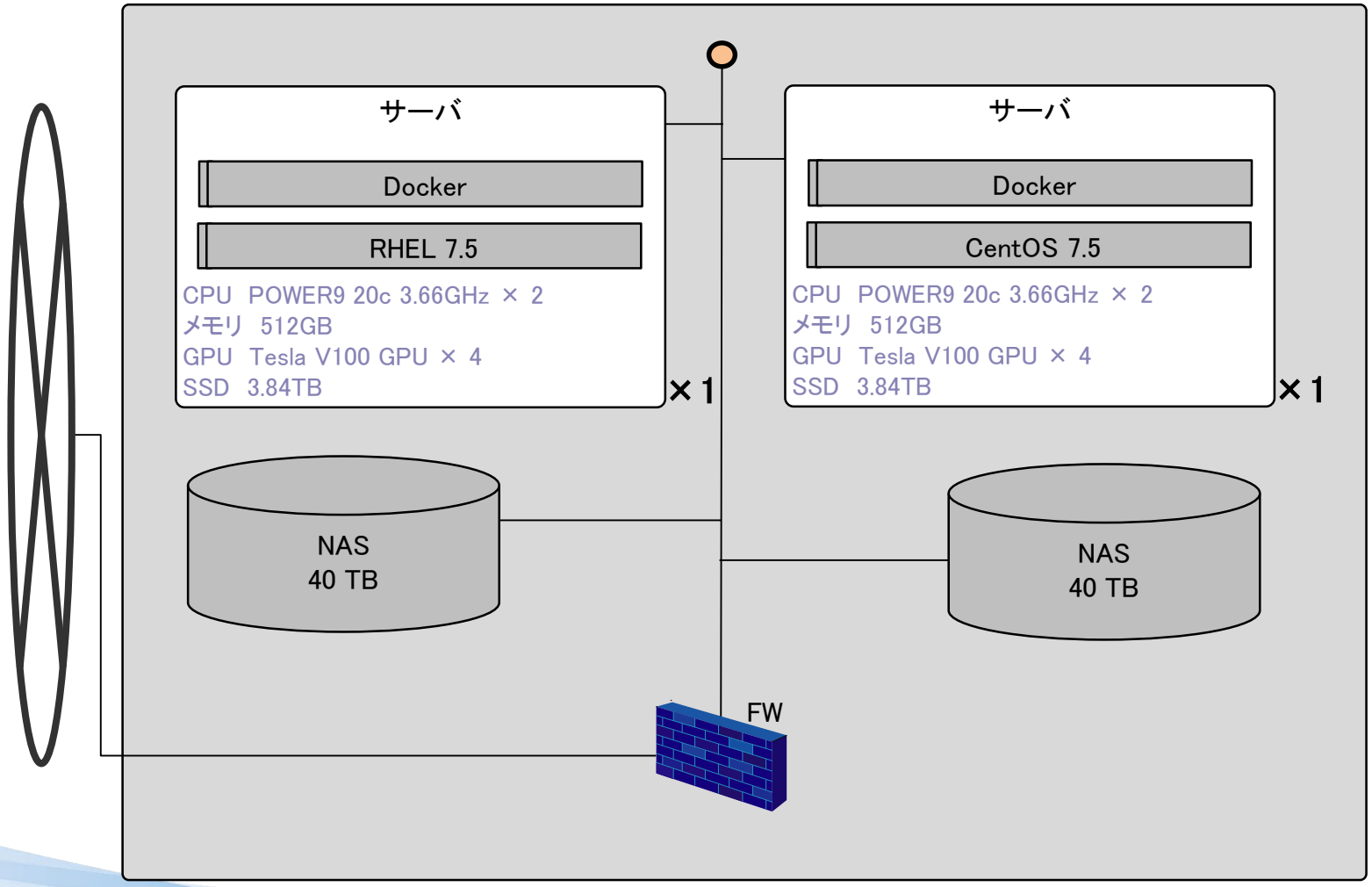
＜採用する機械学習モデル＞

No.	モデル	観点① 適用実績	観点② 学習コスト	観点③ 親和性	モデルの概要
1	サポートベクターマシン(SVM)(ベースライン)	○	○	○	・最も広く適用されているモデル ・特許分類付与への適用実績もあり、高い精度が期待できる ・非ディープラーニングモデルなので学習コストは低い
2	マルチヘッドニューラルアテンションモデル(MH-NAM)	△	△	◎	・特許分類付与への適用実績があるニューラルアテンションモデル(NAM)をベースとしている ・ディープラーニングモデルであるため、学習コストが高い ・文章が長い、表記のゆれがある等のタスクの特性を反映した多くの拡張が可能であり、SVMよりも高い精度が期待できる

3.2 構築環境

【ハードウェア】

検証環境



3.2 構築環境

【主なソフトウェア】

No.	ソフトウェア	概要	環境	特筆事項
1	MeCab	形態素解析器(日本語)	特徴量抽出	日本語テキストからの単語切り出しに使用
2	Word2Vec	テキストをDeep Learningが理解できる数値形式に変換	特徴量抽出	変換したデータを機械学習の入力層に使用することで、学習時間を大幅に短縮可能
3	Scikit-learn	SVMをサポートした機械学習ライブラリ	SVM	精度算出やクロスバリデーション、パラメータ調整等の機能がサポートされている
4	Tensorflow	Deep Learningをサポートした機械学習ライブラリ	MH-NAM	データや処理を可視化可能であり、ニューラルネットワークを構築可能
5	cuDNN	Deep Learning用のCUDAライブラリ	MH-NAM	最適化されたライブラリを使用することで処理性能が向上

4.1　検証の目的

・平成30年度調査事業及び令和元年度調査事業の結果を踏まえ、
　ノイズ排除及び技術区分付与のさらなる精度向上の可能性について調査・検証を行う

・調査実施済の調査テーマの一部を対象として、精度向上が見込める施策について調査を行い、
　ノイズ排除及び技術区分付与の精度向上度合を検証し、施策の有効性を分析する

4.2 検証の内容

平成30年度調査事業及び令和元年度調査事業の結果を踏まえ、
以下の3種類の施策の有効性を検証する

＜精度向上施策の一覧＞

No.	技術課題	検証(施策)の内容	対象モデル
1	ノイズ排除と技術区分付与を連続して機械処理した場合の精度(総合精度)の向上	ノイズ排除精度が100%でないことを前提とした、技術区分付与の機械学習	SVM MH-NAM
2	ノイズ排除の精度低下に起因する技術区分付与の精度低下の防止	ノイズ排除の再現率に基づく閾値設定	SVM MH-NAM
3	複数の原文言語が混在する学習データを使用した場合の機械学習の精度向上	言語ごとの機械学習	SVM

※ 総合精度： ノイズ排除と技術区分付与を「連続して」機械処理した場合の技術区分付与精度
連続して機械処理した場合、ノイズ排除結果にノイズ文献の排除漏れ、非ノイズ文献の誤排除が生じるため、
その分、調査対象文献(ノイズ排除精度100%)に対する技術区分付与精度に比べて精度が低下する傾向にある

4.3 検証の方法 (1) 対象調査テーマ、使用文献数

平成30年度調査事業及び令和元年度調査事業で採用した3調査テーマを対象とする

＜対象調査テーマ＞

No.	調査テーマ	調査実施年度	調査対象年	技術分野	技術区分数※1
1	スマートマニュファクチャリング技術	平成28年度	2005-2014	機械	559区分
2	リチウム二次電池	平成29年度	2009-2015	化学	325区分
3	自動走行システムの運転制御	平成29年度	2010-2014	機械	324区分

※1 上位区分を含めた数。また、技術区分体系の整合性をとるために加えた技術区分を含めた数。

＜使用文献数＞

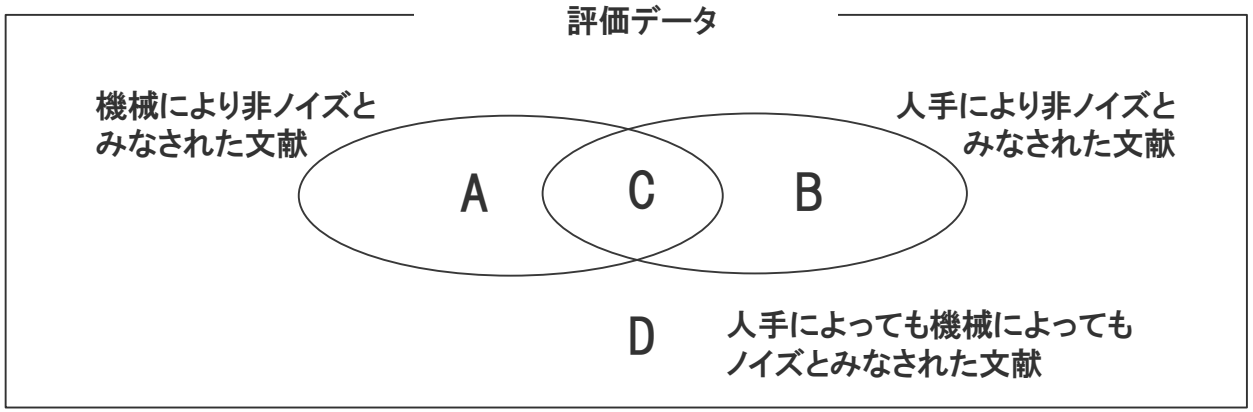
No.	調査テーマ	検索式文献数	調査対象文献数
1	スマートマニュファクチャリング技術	32,949件	11,805件
2	リチウム二次電池	59,033件	52,256件
3	自動走行システムの運転制御	38,472件	8,554件

検索式文献数 ： 調査テーマの調査範囲を規定する検索式によって検索された文献数
調査対象文献数 ： 検索式文献のうち、人手作業によって技術区分が一つ以上付与された文献数

4.3 検証の方法 (2) 精度評価指標

【精度評価指標（ノイズ排除）】

- F値（再現率（漏れのなさ）と適合率（ノイズのなさ）の調和平均）を採用
- ただし、F値は非ノイズ文献の割合が高いほど高くなる
  ⇒平均F値（非ノイズ文献抽出精度（F値1）とノイズ文献排除精度（F値2）の単純平均）も採用



【非ノイズ文献を抽出する精度】
再現率 R1 = C÷(B＋C)
適合率 P1 = C÷(A＋C)
F値1 = (2×R1×P1)÷(R1＋P1)

【ノイズ文献を排除する精度】
再現率 R2 = D÷(A＋D)
適合率 P2 = D÷(B＋D)
F値2 = (2×R2×P2)÷(R2＋P2)

平均F値 = (F値1＋F値2)÷2

4.3 検証の方法 (2) 精度評価指標

【精度評価指標（技術区分付与）】

(1) **F値**

F値の算出方法として、**マイクロ平均F値**、及び、**マクロ平均F値**を採用

- マイクロ平均F値

  複数の技術区分を一つの技術区分に統合した場合のF値
  個々の技術区分に対する文献ごとの結果をマージ(加算)してから、F値を算出

- マクロ平均F値

  各技術区分に対して算出したF値の単純平均

(2) **乖離度**（必要に応じて採用）

乖離度の算出方法として、**トータル差**、**単純平均**、**2乗平均**を採用

- トータル差：「技術区分ごとの人手付与文献数の総和」と「技術区分ごとの機械付与文献数の総和」の差の絶対値を、「技術区分ごとの人手付与文献数の総和」で除算した値
- 単純平均 ： 技術区分別の付与文献数の差の絶対値の単純平均
- 2乗平均 ： 技術区分別の付与文献数の差の2乗の単純平均

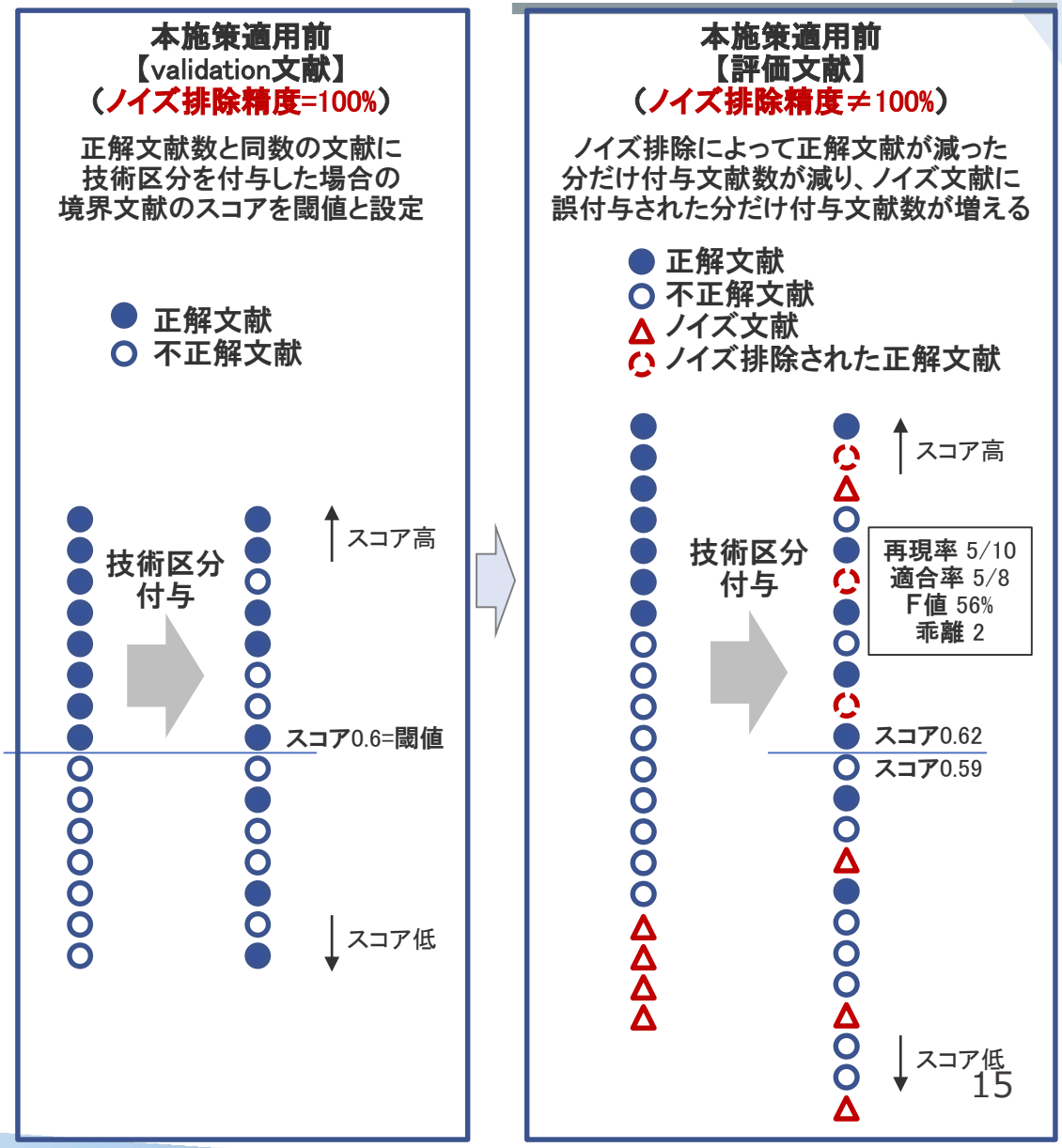
4.4 再現性の向上施策（1） ノイズ排除精度が100%でないことを前提とした技術区分付与の機械学習

【概要】

技術区分付与において、**閾値の設定に使用する学習文献（validation文献）の選定方法を変更**することにより、技術区分付与精度を向上できるかを検証

＜本施策適用前（右図）＞

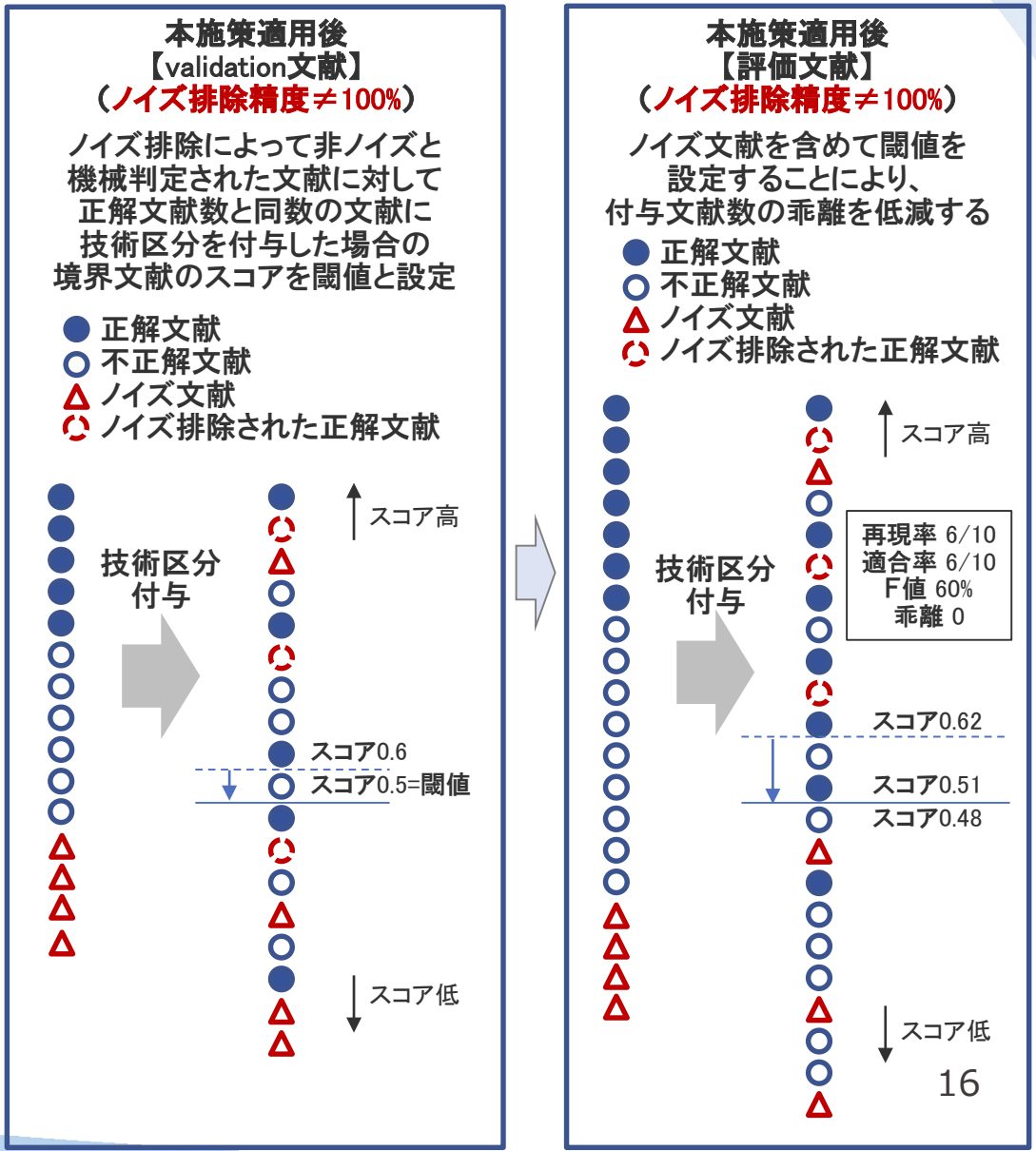
ノイズ排除精度=100%の文献を技術区分付与のvalidation文献として使用



4.4 再現性の向上施策（1） ノイズ排除精度が100%でないことを前提とした技術区分付与の機械学習

＜本施策適用後（右図）＞

validation文献に対してノイズ排除した結果、
非ノイズと機械判定された文献を
技術区分付与のvalidation文献として使用



4.4 再現性の向上施策（1） ノイズ排除精度が100%でないことを前提とした技術区分付与の機械学習

【検証対象調査テーマ】 スマートマニュファクチャリング技術、自動走行システムの運転制御

【有効性検証結果】 精度（F値）の改善は見られないが、**一部の調査テーマにおいて、乖離度が改善**

調査テーマ	施策適用	技術区分付与の総合精度							
		SVM				MH-NAM			
		マイク平均F値	差	マクロ平均F値	差	マイクロ平均F値	差	マクロ平均F値	差
スマートマニュファクチャリング技術	適用前	26.54%	-	8.53%	-	27.68%	-1.78%	7.23%	-0.29%
	適用後	25.97%	0.57%	8.29%	0.24%	25.90%		6.94%	
自動走行システムの運転制御	適用前	56.74%	-	26.59%	0.09%	54.76%	-0.55%	20.20%	-0.22%
	適用後	56.31%	0.43%	26.68%		54.21%		19.98%	

調査テーマ	施策適用	技術区分付与の乖離度											
		SVM						MH-NAM					
		トータル差	差	単純平均	差	2乗平均	差	トータル差	差	単純平均	差	2乗平均	差
スマートマニュファクチャリング技術	適用前	35.4%	2.4%	56.4	3.4	23561	3634	28.94%	6.7%	55.6	12.4	19742	12111
	適用後	37.9%		59.8		27195		39.44%		68.1		31853	
自動走行システムの運転制御	適用前	3.9%	-2.7%	9.8	-1.2	296	-97	3.43%	-0.4%	11.1	-0.01	519	-19
	適用後	1.2%		8.6		199		3.05%		11.1		500	

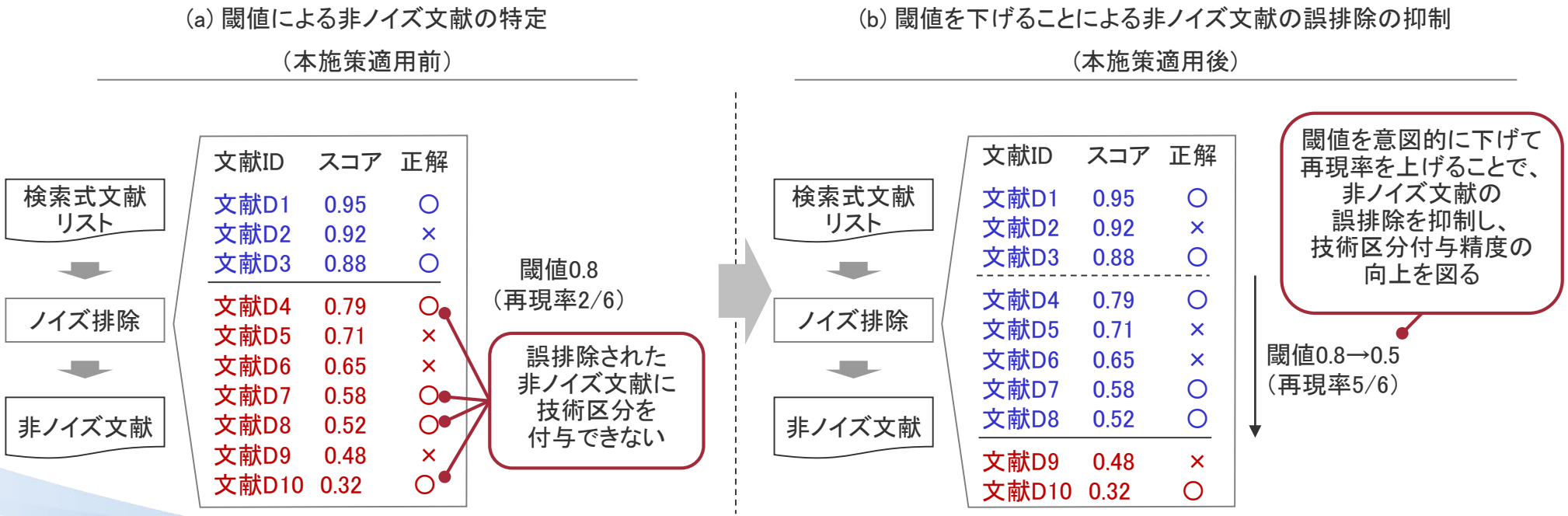
乖離度が改善

※ 差の値が黒字は改善、赤字は悪化を示す。乖離度は値が低いほど良い。

4.5 再現性の向上施策（2） ノイズ排除の再現率に基づく閾値設定

【概要】 ノイズ排除において、ノイズ文献を判別する**閾値を「意図的に」下げて再現率を高く設定**し、非ノイズ文献の誤排除を抑制することにより、技術区分付与精度を改善できるかを検証

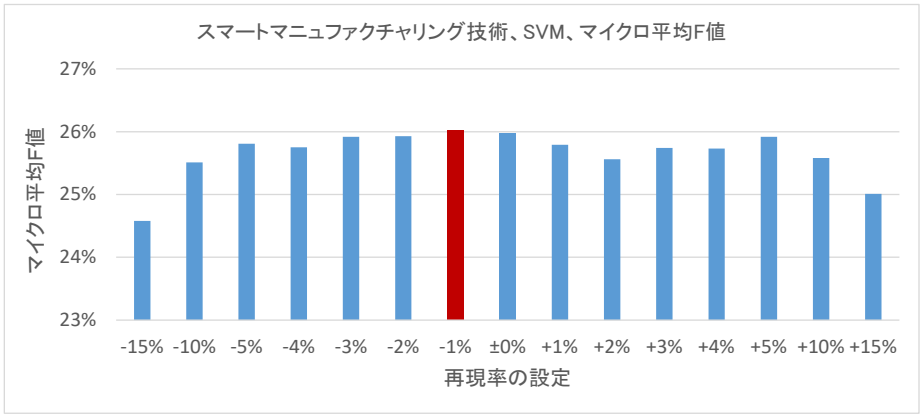
＜本施策適用前＞ validation文献において、再現率=適合率となるノイズ排除スコアを閾値と設定
＜本施策適用後＞ validation文献において、**再現率を -15%から+15% まで変動**させた時のノイズ排除スコアを閾値と設定

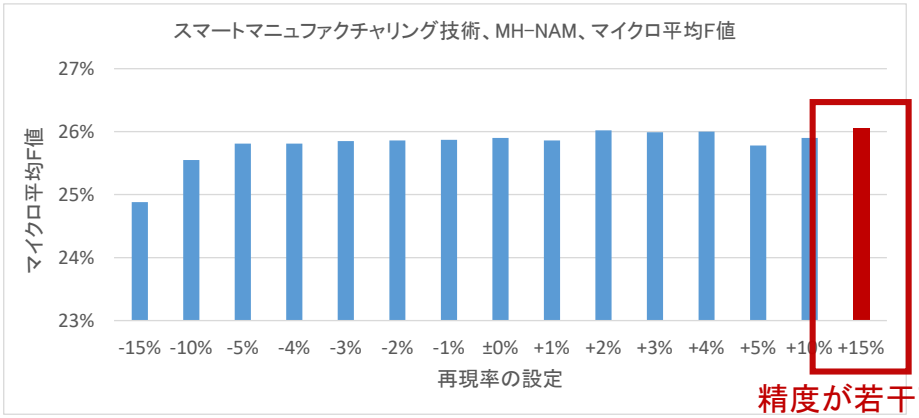


4.5 再現性の向上施策（2） ノイズ排除の再現率に基づく閾値設定

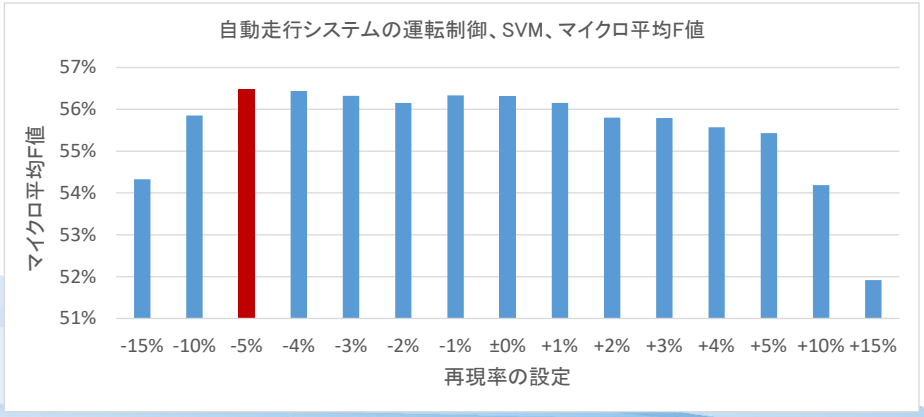
【検証対象調査テーマ】 スマートマニュファクチャリング技術、自動走行システムの運転制御

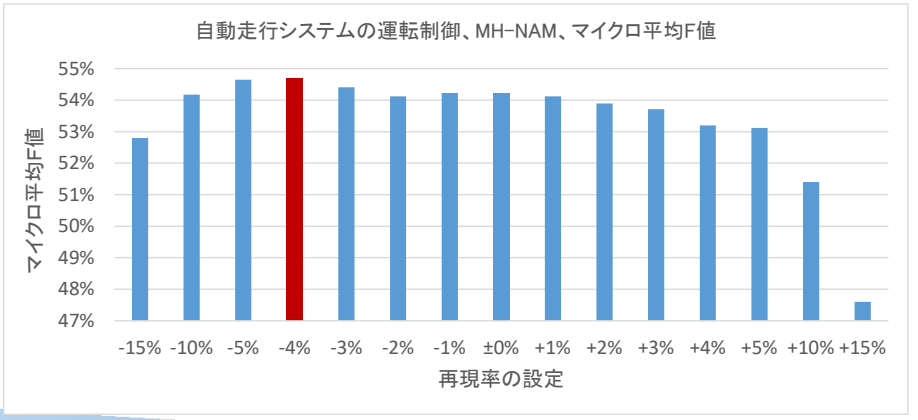
【有効性検証結果】 一部の調査テーマにおいて精度が改善したが、効果は小さい





精度が若干改善



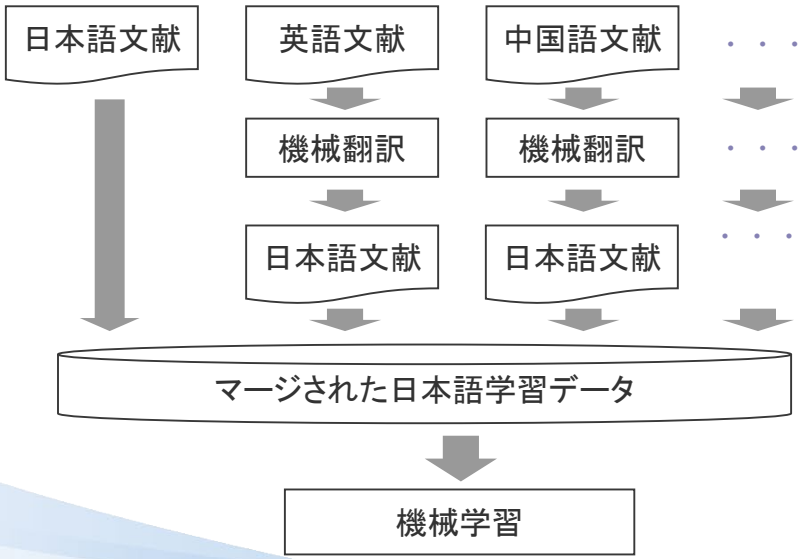


4.6 再現性の向上施策 (3)　言語ごとの機械学習

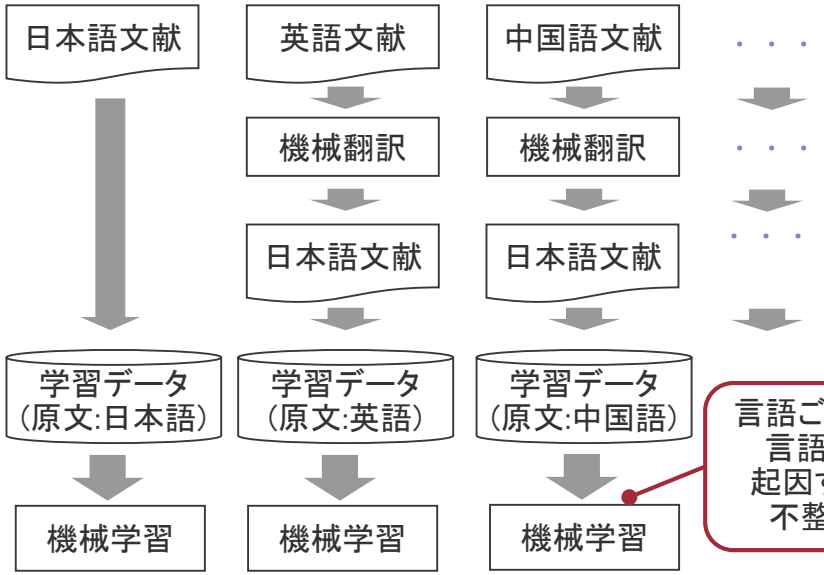
【概要】 文献を言語ごとに分け、**言語(出願国)ごとに学習・推測**した場合の精度を比較

＜本施策適用前＞　学習文献を日本語に機械翻訳した結果をすべてマージして学習
＜本施策適用後＞　学習文献を**言語(出願国)ごとに分けて個別に学習**

(a) 機械翻訳した日本語文献をマージして学習
(本施策適用前)



(b) 原文の言語ごとに分けて学習
(本施策適用後)



4.6 再現性の向上施策 (3)　言語ごとの機械学習

【検証対象調査テーマ】 リチウム二次電池、自動走行システムの運転制御

【有効性検証結果】 **一部の調査テーマの一部の言語で精度が改善**
言語別に学習する場合、学習データ量が減少することが精度低下原因の一つ

【リチウム二次電池】

言語	評価文献数	付与精度(総合精度)(マイクロ平均F値)		
		全言語で学習	言語別に学習	差
全体	19,067件	54.20%	52.45%	-1.75%
CN	5,051件	62.51%	61.92%	-0.59%
DE	417件	48.93%	41.92%	-7.01%
EP	65件	46.53%	28.43%	-18.10%
JP	8,339件	50.49%	50.03%	-0.46%
KR	2,491件	53.03%	47.62%	-5.41%
US	2,538件	49.26%	46.67%	-2.59%
WO	166件	48.35%	35.23%	-13.12%

【自動走行システムの運転制御】

言語	評価文献数	付与精度(総合精度)(マイクロ平均F値)			
		全言語で学習	言語別に学習	差	
全体	7,696件	56.74%	49.35%	-7.39%	
CN	1,097件	40.81%	43.67%	2.86%	精度改善
DE	893件	49.91%	22.34%	-27.57%	
EP	54件	37.66%	5.17%	-32.49%	
JP	2,758件	63.91%	64.45%	0.54%	精度改善
KR	856件	42.56%	42.03%	-0.53%	
US	1,812件	50.88%	48.92%	-1.96%	
WO	226件	51.72%	36.66%	-15.06%	

4.7 検証のまとめ

施策(1)から施策(3)を検証したが、一部の調査テーマにおいて精度改善が見られたが、顕著な効果は見られなかった

＜各施策の有効性＞

No.	技術課題	検証(施策)の内容	効果
1	ノイズ排除と技術区分付与を連続して機械処理した場合の精度(総合精度)の向上	ノイズ排除精度が100%でないことを前提とした、技術区分付与の機械学習	一部の調査テーマで有効性を確認
2	ノイズ排除の精度低下に起因する技術区分付与の精度低下の防止	ノイズ排除の再現率に基づく閾値設定	一部の調査テーマで有効性を確認
3	複数の原文言語が混在する学習データを使用した場合の機械学習の精度向上	言語ごとの機械学習	一部の調査テーマの一部の言語で有効性を確認

5.1 検証の目的　　5.2 検証の内容

【目的】

平成30年度調査事業の3調査テーマ、及び、令和元年度の4調査テーマ以外の
新たな4調査テーマを対象として、調査対象の特許文献における
ノイズ排除及び技術動向の再現（技術区分付与）の実現可能性を検証する

【内容】

以下の検証項目について検証を行う。

（1）調査対象年の特許文献におけるノイズ排除及び技術動向の再現の精度検証

新たな4調査テーマを対象として、調査対象年の特許文献におけるノイズ排除及び
技術区分付与の精度を検証する

（2）調査対象外の特許文献に対するノイズ排除及び技術動向の再現の実施

調査対象外の特許文献（調査対象年以降の特許文献、動向調査の時点でDBに未登録
だった調査対象年の特許文献）に対してノイズ排除及び技術区分付与を実施し、
その結果を分析する

5.3 調査対象年の特許文献の精度検証（１）検証の方法

新たな4調査テーマを対象として、調査対象年の特許文献におけるノイズ排除及び技術区分付与の精度を検証する。

＜対象調査テーマ＞

No.	調査テーマ	調査実施年度	調査対象年	技術分野	技術区分数※
1	リチウム二次電池	H29年度	2009-2015	化学	325区分
2	ドローン	H30年度	2007-2016	機械	474区分
3	三次元計測	H30年度	2006-2016	一般	600区分
4	情報セキュリティ技術	H27年度	2009-2013	電気・電子	423区分

※　上位区分を含めた数。また、技術区分体系の整合性をとるために加えた技術区分を含めた数。

＜使用文献数＞

No.	調査テーマ	検索式文献数	調査対象文献数
1	リチウム二次電池	58,250件	51,964件
2	ドローン	27,190件	21,608件
3	三次元計測	46,651件	35,427件
4	情報セキュリティ技術	45,664件	14,397件

5.3 調査対象年の特許文献の精度検証 （２）検証の結果

＜ノイズ排除精度＞ 平均F値 71%～83% MH-NAMの方が精度高い

調査テーマ	評価文献数	モデル	非ノイズ文献抽出のF値	ノイズ文献排除のF値	平均F値	差
リチウム二次電池	11,650件	SVM	95.39%	60.95%	78.17%	1.52%
		MH-NAM	95.63%	63.74%	79.69%	
ドローン	5,438件	SVM	93.01%	72.99%	83.00%	-0.52%
		MH-NAM	92.72%	72.24%	82.48%	
三次元計測	9,331件	SVM	85.94%	55.47%	70.71%	0.54%
		MH-NAM	86.17%	56.33%	71.25%	
情報セキュリティ技術	9,136件	SVM	67.82%	85.29%	76.56%	3.94%
		MH-NAM	73.26%	87.73%	80.50%	

非ノイズ文献抽出のF値 ： 非ノイズ文献を正しく抽出できたかを示す精度値
ノイズ文献排除のF値 ： ノイズ文献を正しく排除できたかを示す精度値

5.3 調査対象年の特許文献の精度検証 （２)検証の結果

＜技術区分付与精度＞ マイクロ平均 41%～56% マクロ平均 15%～38% SVMの方が精度高い

調査テーマ	評価文献数	モデル	総合精度							
			マイクロ平均				マクロ平均			
			再現率	適合率	F値	差	再現率	適合率	F値	差
リチウム二次電池	11,650件	SVM	55.77%	55.51%	55.64%	0.76%	37.83%	37.88%	37.55%	-2.88%
		MH-NAM	56.44%	56.36%	56.40%		34.61%	34.96%	34.67%	
ドローン	5,438件	SVM	45.92%	46.15%	46.03%	-3.12%	28.31%	30.34%	28.67%	-6.36%
		MH-NAM	42.72%	43.10%	42.91%		22.29%	23.25%	22.31%	
三次元計測	9,331件	SVM	44.56%	45.18%	44.87%	-3.58%	20.84%	21.97%	20.89%	-4.25%
		MH-NAM	41.24%	41.35%	41.29%		16.95%	17.08%	16.64%	
情報セキュリティ技術	2,881件	SVM	41.70%	39.96%	40.81%	0.10%	22.17%	21.79%	20.98%	-5.92%
		MH-NAM	42.13%	39.76%	40.91%		15.96%	15.30%	15.06%	

※ 総合精度： ノイズ排除と技術区分付与を連続して機械処理した場合の技術区分付与精度

5.3　調査対象年の特許文献の精度検証（３）結果の考察

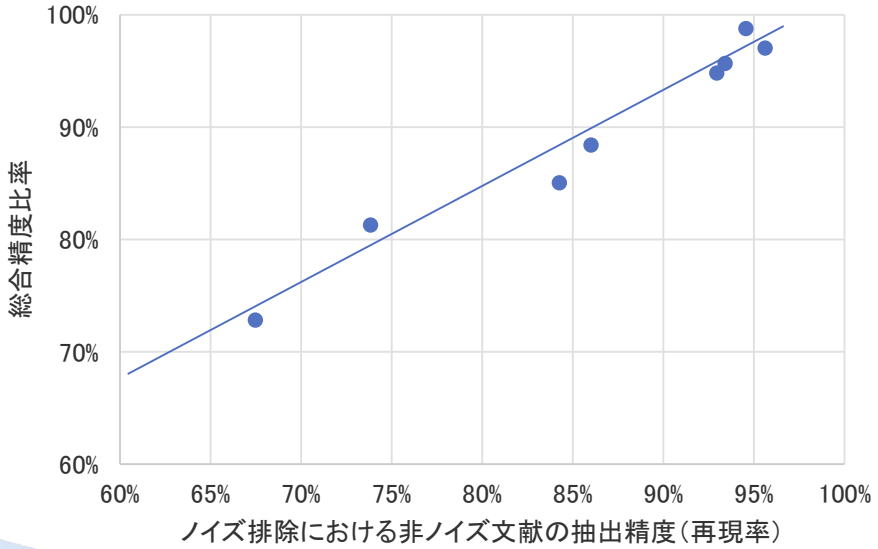
（a）モデル間の精度比較

全般的に、ノイズ排除及び付与文献数の多い技術区分では、MH-NAMの方が精度が高いが
付与文献数の少ない技術区分では、SVMの方が精度が高い

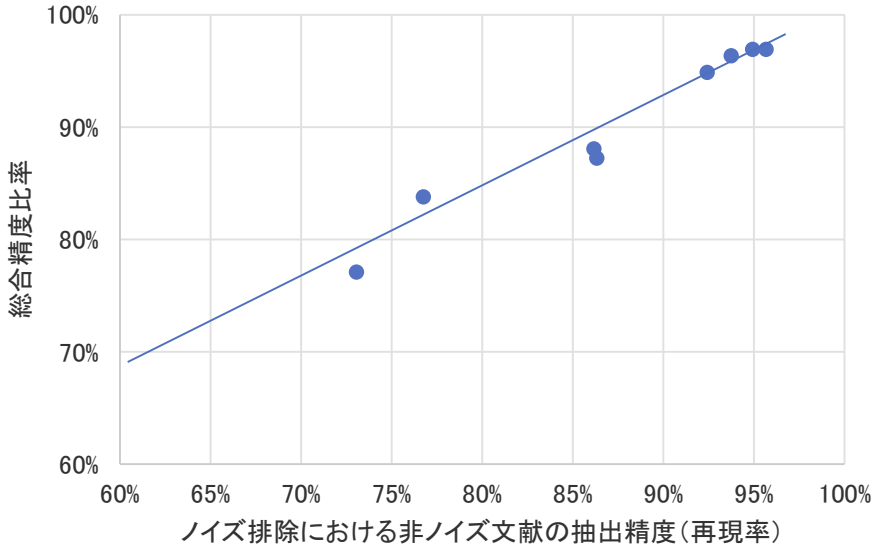
（b）総合精度の比較

ノイズ排除の再現率が低い調査テーマほど、総合精度が大きく低下する
（ノイズ排除で誤排除されてしまった非ノイズ文献には技術区分を付与できないため）

非ノイズ文献の抽出精度（再現率）と
総合精度比率の相関（SVM）



非ノイズ文献の抽出精度（再現率）と
総合精度比率の相関（MH-NAM）



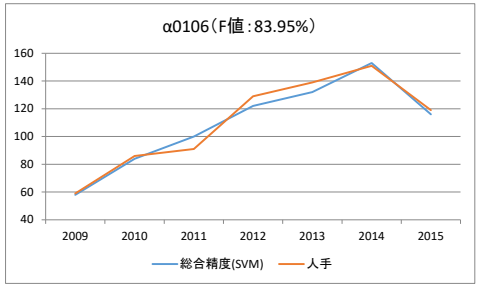
※ 総合精度比率 =「ノイズ排除精度100%の場合の技術区分付与精度」に対する、
「ノイズ排除と技術区分付与を連続して行った場合の技術区分付与精度（総合精度）」の割合

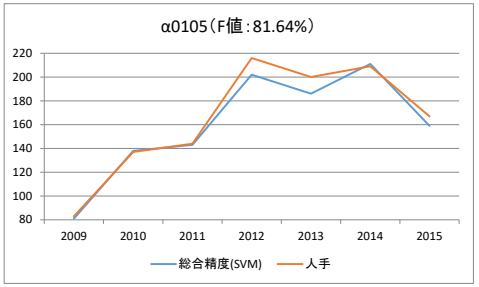
5.3 調査対象年の特許文献の精度検証 （３）結果の考察

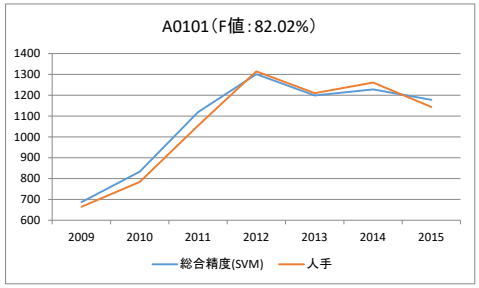
（c）技術動向の再現の実現性

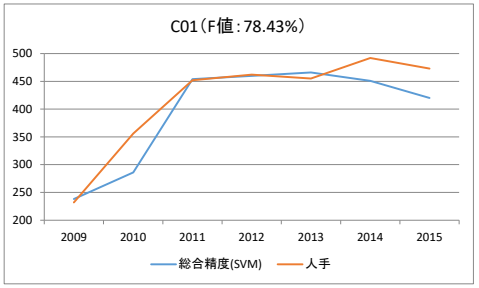
F値の高い技術区分においては、技術動向を再現できているものが多い

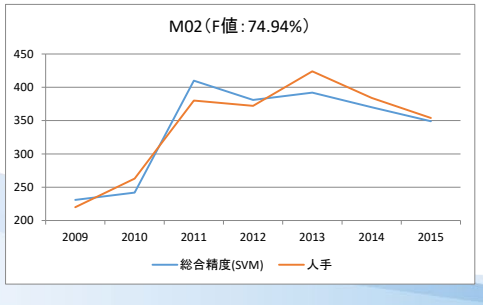
【リチウム二次電池】

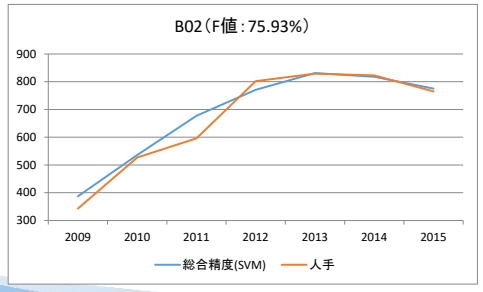




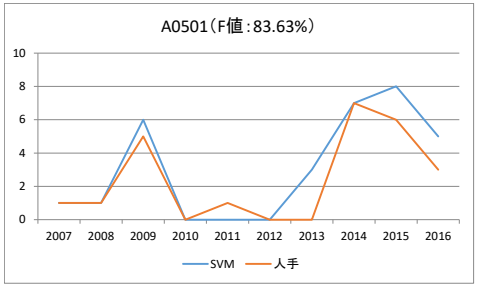


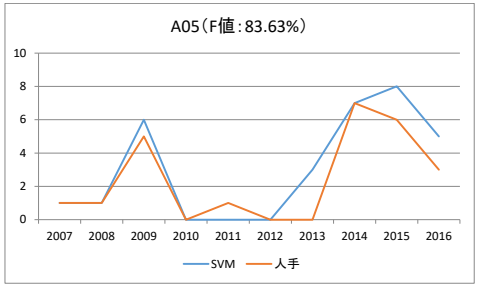


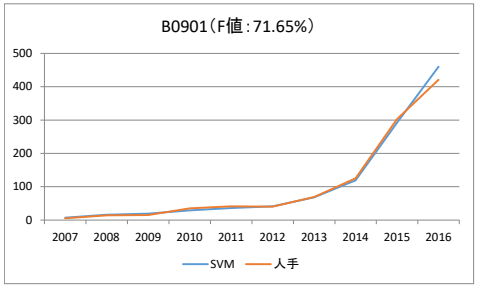


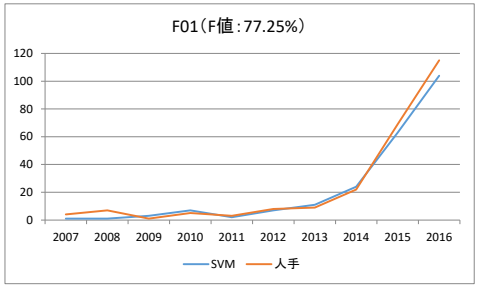


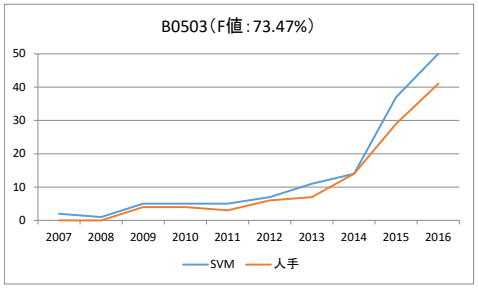
【ドローン】

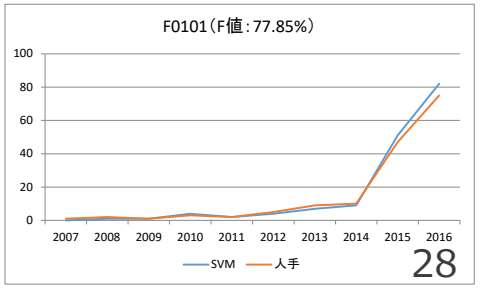










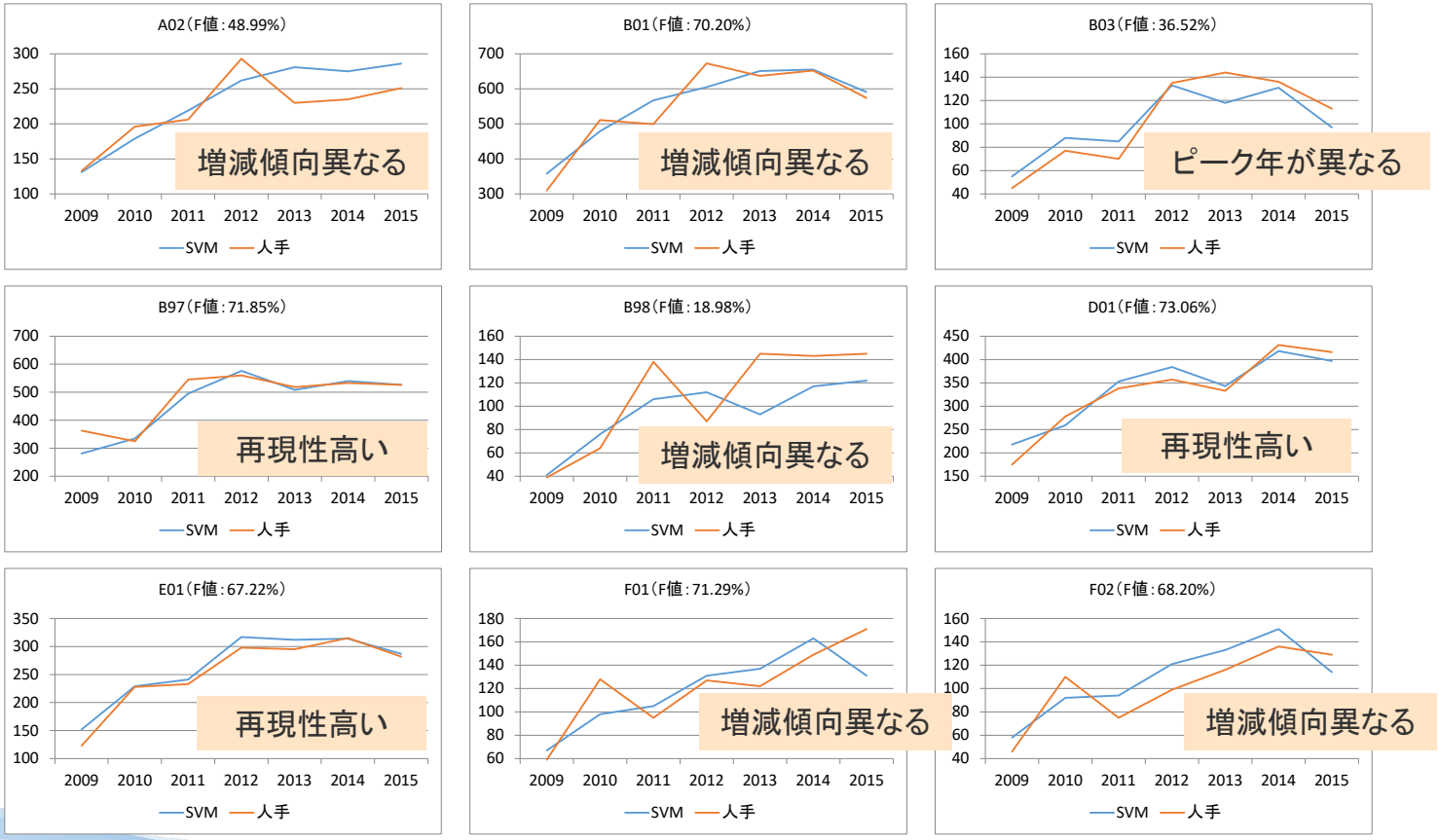


5.3 調査対象年の特許文献の精度検証（３）結果の考察

（c）技術動向の再現の実現性

技術区分の中区分で見ると、再現性が高い区分、増減傾向の異なる区分、ピーク年が異なる区分等が存在する

【リチウム二次電池】



5.4 調査対象外の特許文献の精度検証 （1）検証の方法

以下の2種類の特許文献に対して、ノイズ排除及び技術区分付与を実施し、その結果を分析する。

（1）動向調査の時点でDBに未登録であった調査対象年範囲内の特許文献
（2）調査対象年以降の特許文献

No.	調査テーマ	動向調査当時にデータベース登録済の検索式文献数	（1）動向調査当時にデータベース未登録の検索式文献数	（2）調査対象年以降の検索式文献数
1	リチウム二次電池	58,250件 (2009年-2015年)	1,994件 (2009年-20015年)	44,554件 (2016年-2018年)
2	ドローン	27,190件 (2007年-2016年)	1,966件 (2007年-2016年)	27,876件 (2017年-2018年)
3	三次元計測	46,651件 (2006年-2016年)	2,005件 (2006年-2016年)	16,226件 (2017年-2018年)
4	情報セキュリティ技術	45,664件 (2009年-2013年)	11,802件 (2009年-2013年)	108,016件 (2014年-2018年)

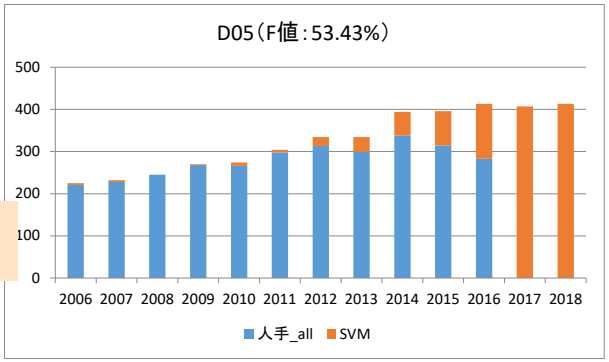
5.4 調査対象外の特許文献の精度検証 (2) 検証の結果

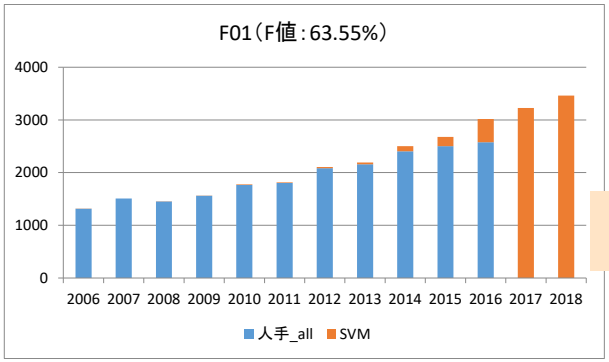
・調査対象年において、動向調査を実施した時点での人手による付与文献数(下図青グラフ)と、DB未登録の文献に対する機械による技術区分付与結果を加算した場合の付与文献数(下図青＋赤グラフ)の増減傾向が異なる場合がある

・調査対象年以降の文献に対する技術区分付与結果を追記することにより、調査対象年以降の技術動向を把握できる (ただし、調査対象年以降のDB未登録の文献が含まれていないことに留意)

【三次元計測】

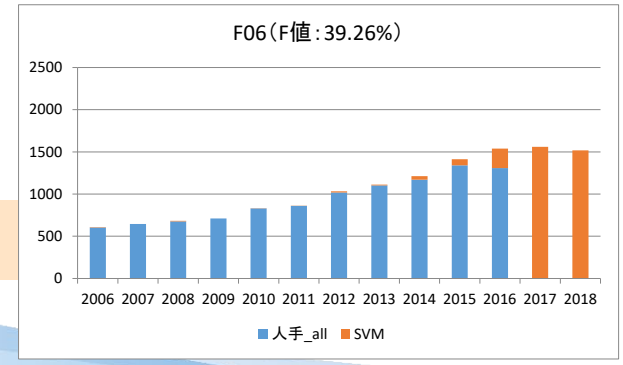
青　　:増加から減少に
青+赤:増加から横ばいに

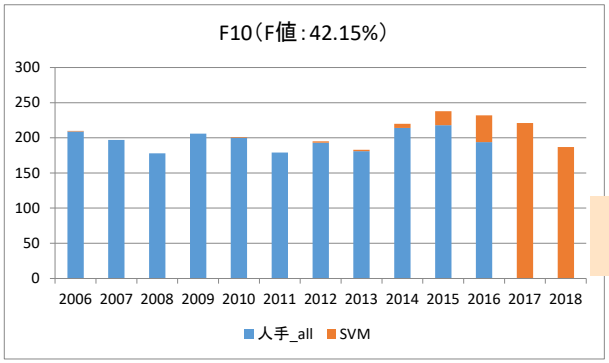




青　　:増加から横ばいに
青+赤:増加が続く

調査対象年以降
頭打ちに





調査対象年以降
減少に転ずる

5.5　検証のまとめ

・新たな4調査テーマにおけるノイズ排除精度

非ノイズ文献の抽出精度　　：　68%～96%
ノイズ文献の排除精度　　：　55%～88%
平均F値　　：　71%～83%

MH-NAMの方が精度高い

・新たな4調査テーマにおける技術区分付与精度

技術区分付与総合精度　　：　マイクロ平均41%～56%　マクロ平均15%～38%

SVMの方が精度高い

・非ノイズ文献の抽出精度（再現率）が低いほど、総合精度の低下度合が大きい

・F値の高い技術区分や中区分の一部において、技術動向が再現できている

・DBに未登録の文献、及び、調査対象年以降の文献を加味すると、技術動向のトレンドが変化する場合がある

6.1 調査の目的　6.2 調査の内容

【目的】

前章までで報告した調査内容以外で、本事業に資する調査を行う

【内容】

新たな調査として、以下の調査を実施

（1）機械学習結果の有効期間の調査
（2）動向調査に必要なリソースの調査

6.3 調査（1） 機械学習結果の有効期間の調査

【概要】

機械学習を調査対象年以降の文献に適用した場合の「**出願年ごとの付与精度の変化**」を計測することにより、**どのくらいの期間、機械学習による精度を維持できる**かを調査

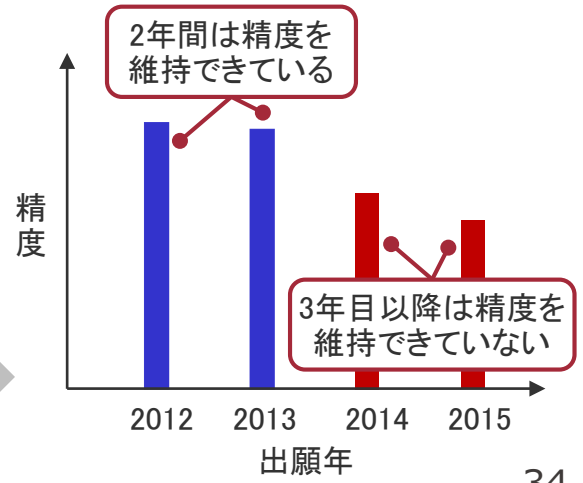
【方法】

調査対象年範囲内の検索式文献を、時系列（出願年）に沿って大きく2グループに分け、出願年が古い方のグループで機械学習を行い、出願年が新しい方のグループの付与精度を出願年ごとに算出し、付与精度の時系列的推移（低下度合）を分析

ヒト幹細胞関連技術 （調査対象年: 2007～2015、検索式文献: 26,733件）

<学習データと評価データの割当>

<精度の年次推移>



6.3 調査（1） 機械学習結果の有効期間の調査

【対象調査テーマ】 ヒト幹細胞関連技術 （調査対象年 2007年～2015年）

【結果】 技術区分付与精度は、時系列的に低下する傾向にある
時系列的な精度低下を伴った機械処理による動向調査結果の扱いについて考慮が必要

※赤数値は精度悪化を示す

※ベース： 2007年～2011年の検索式文献の一部を評価文献として時系列横断的に抽出して精度検証したもの

調査テーマ	出願年	マイクロ平均				マクロ平均			
		再現率	適合率	F値	ベースと比較した精度差	再現率	適合率	F値	ベースと比較した精度差
ヒト幹細胞関連技術	ベース	49.34%	49.03%	49.18%	-	20.62%	20.41%	19.81%	-
	2012年	47.77%	49.25%	48.50%	-0.68%	19.80%	22.04%	20.16%	+0.35%
	2013年	47.56%	48.49%	48.02%	-1.16%	20.02%	21.21%	19.83%	+0.02%
	2014年	44.40%	46.79%	45.57%	-3.61%	18.56%	19.82%	18.52%	-1.29%
	2015年	43.90%	48.02%	45.87%	-3.31%	18.49%	21.10%	18.68%	-1.13%

調査テーマ	出願年	相関係数	乖離度（トータル差）	乖離度（単純平均）	乖離度（2乗平均）
ヒト幹細胞関連技術	ベース	0.992	0.65%	4.20	49.73
	2012年	0.990	3.00%	4.93	71.06
	2013年	0.994	1.92%	6.12	114.28
	2014年	0.993	5.11%	6.73	142.46
	2015年	0.980	8.58%	8.35	222.37

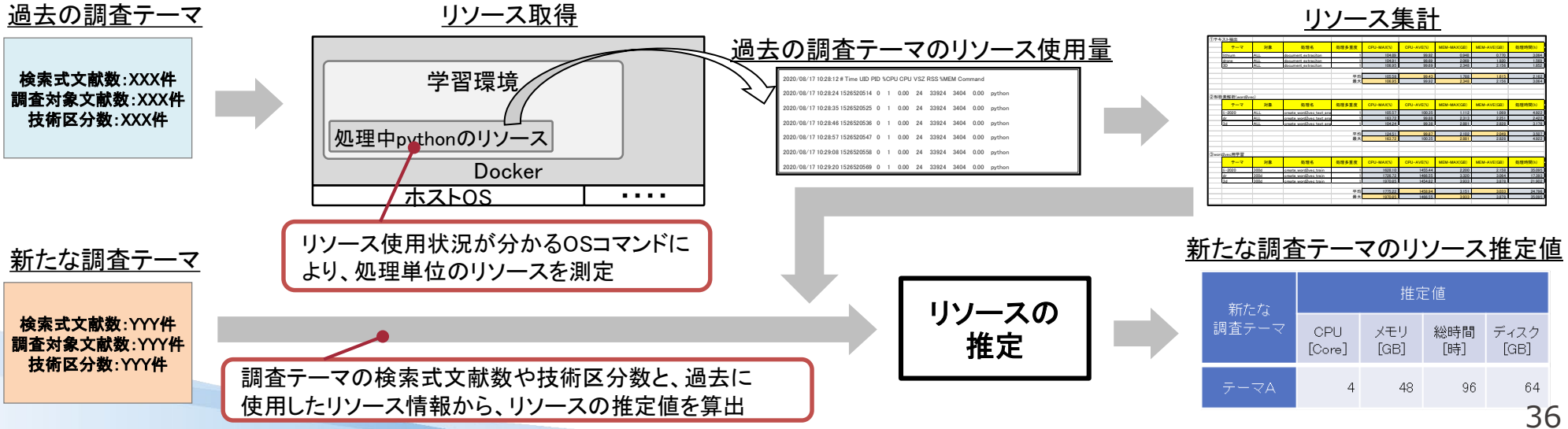
6.4 調査（2） 動向調査に必要なリソースの調査

【概要】

過去の調査テーマに対して機械処理による機械学習等の処理単位ごとの「**リソース情報**」を計測することにより、**新たな調査テーマにおける動向調査に必要なリソース量**を推定する

【方法】

過去の調査テーマの処理単位ごとのリソース情報の結果を踏まえ、新たな調査テーマの検索式文献数を基準に、調査テーマの検索式文献数から各リソース比を算出し、調査テーマごとにリソース比を乗じたリソース情報からリソース量を推定することで、新たな調査テーマに必要なリソース量を算出



新たな調査テーマ	推定値			
	CPU [Core]	メモリ [GB]	総時間 [時]	ディスク [GB]
テーマA	4	48	96	64

6.4 調査 (2) 動向調査に必要なリソースの調査

【過去の調査テーマで使用されたリソース情報の収集結果】

収集した過去の調査テーマ： リチウム二次電池、ドローン、三次元計測 の3テーマ

(a) 特徴量抽出（テキスト抽出、形態素解析、特徴量解析等）処理で使用したリソース情報について

- 処理の総時間及びディスク使用量は、検索式文献数に比例して増加する傾向
- CPU使用率及びメモリ使用量については、検索式文献数との間に相関が見られない

(b) 機械学習処理で使用したリソース情報について

- SVM及びMH-NAMによるノイズ排除の機械学習処理において、
  処理の総時間は、検索式文献数が多い調査テーマで増加する傾向
- SVMによる技術区分付与の機械学習処理において、
  処理の総時間は、調査対象文献数に比例して増加する傾向
  ディスク使用量は、技術区分数が多い調査テーマで増加する傾向
- MH-NAMによる技術区分付与の機械学習処理において、
  処理の総時間は、調査対象文献数と技術区分グループ数の積に比例して増加する傾向
  ディスク使用量は、技術区分グループ数が多い調査テーマで増加する傾向

6.4 調査（2） 動向調査に必要なリソースの調査

【新たな調査テーマにおいて必要となるリソース量の推定結果】

新たな調査テーマ「高効率火力発電・発電用ガスタービン」において、検索式文献数及び技術区分数を使用することにより、動向調査に必要なリソース量の推定結果を算出可能

ただし、動向調査に必要なリソースの推定は、1つの調査テーマの不規則なデータに影響を受けてしまうため、サンプリング対象となる調査テーマを増やすことが必要

＜機械学習モデル SVM＞

新たな調査テーマ	推定値				推定値（5年先までの5年分）			
	CPU [Core]	メモリ [GB]	総時間 [時]	ディスク [GB]	CPU [Core]	メモリ [GB]	総時間 [時]	ディスク [GB]
高効率火力発電・発電用ガスタービン	4	48	96	64	4	54	91	63

＜機械学習モデル MH-NAM＞

新たな調査テーマ	推定値				推定値（5年先までの5年分）			
	CPU [Core]	メモリ [GB]	総時間 [時]	ディスク [GB]	CPU [Core]	メモリ [GB]	総時間 [時]	ディスク [GB]
高効率火力発電・発電用ガスタービン	20	260	1335	210	16	286	32	49

7.1 今後の課題

【今後の課題】

本調査において、ノイズ排除は、業務への適用が十分に期待できた。
また、技術区分付与においては、F値が高い技術区分の一部に関して、動向調査における人手で技術区分付与した場合の再現ができた。

今後、業務への適用を考慮し、ノイズ排除及び技術区分付与において解決すべき主な課題を以下に示す。

(1)付与文献数が少ない技術区分の付与精度が低い。そのため、事前学習や転移学習等の適用により、付与文献数が少ない技術区分の精度を向上させる必要がある。

(2)人手付与文献数と機械付与文献数に乖離がある技術区分が存在する。そのため、この乖離を小さくするための機械処理方式を検討する必要がある。

(3)継続的な動向調査を高精度に行うためには、学習器の信頼性(いつまで精度を維持できるか)を担保する必要がある。そのため、学習器の信頼性を継続的に評価する方式を検討する必要がある。