ランダムフォレストを用いた特許に関する文書データからの技術適応領域に関する技術抽出

4月18日

富山県立大学　小野田　成晃

研究目的

技術情報が多様化・複雑化
ー＞技術者間での技術傾向に関する情報の共有が難化
そのためにはパテントマップが必要となる

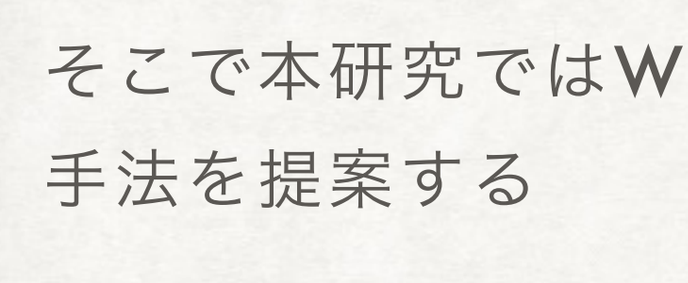
そこで本研究ではWEB上の公開データからパテントマップを作成する手法を提案する

研究概要

研究は以下のプロセスで行われる

J-PlatPat 特許情報プラットフォーム

特許文書データ



スクレイピング

機械学習を用いて知識抽出

パテントマップ作成

パテントマップとは

特許情報の分析ツールの一種

特許の様々な情報に対して２次元マッピングによる可視化を行う

ー＞複雑な技術情報を人が見て知識を抽出できるようにするためのもの

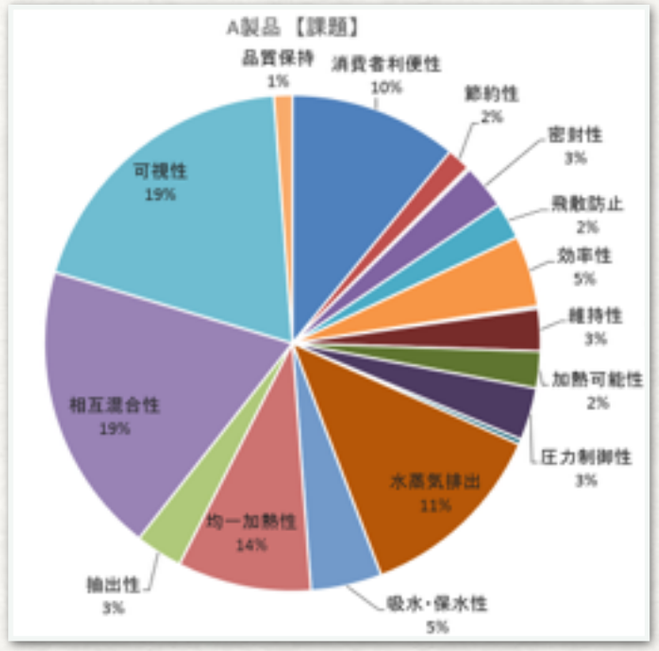
種類は大きく３つに大別される

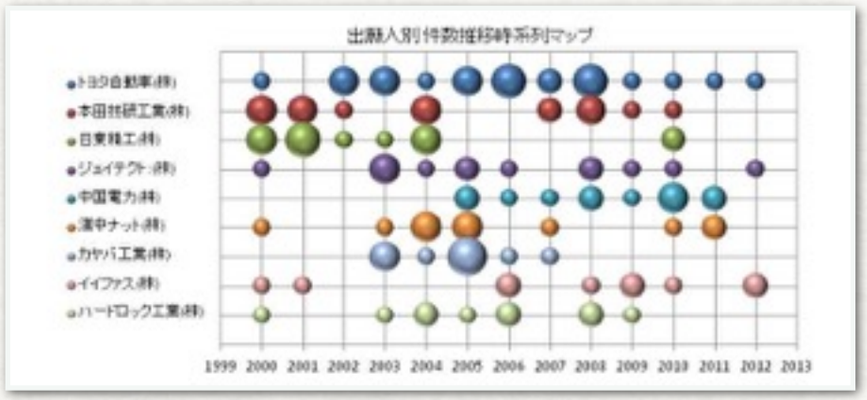
・統計解析型

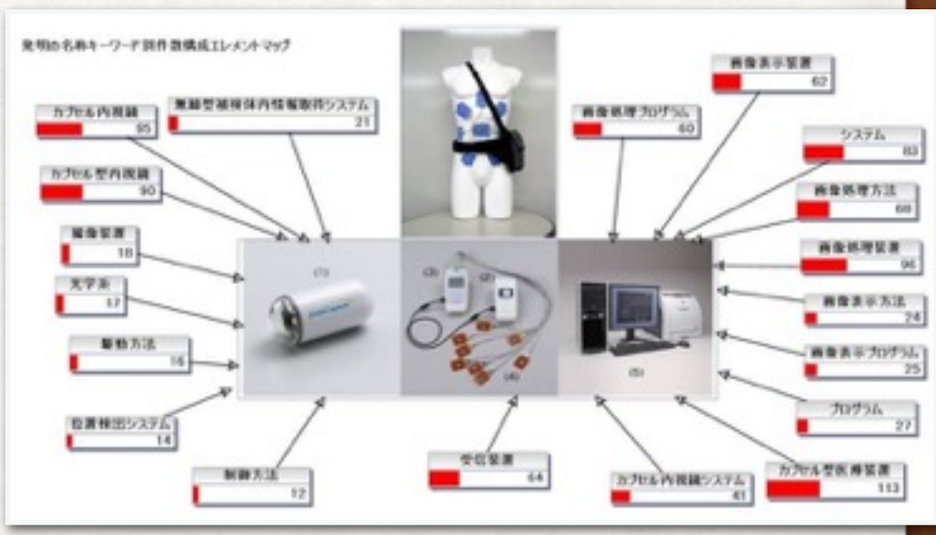
・非統計型

・テキストマイニング型

本研究ではテキストマイニング型を対象とする







*1, HTTP://OUKAJINSUGAWA.HATENADIARY.JP/ENTRY/2013/12/01/091803

*2, HTTPS://WWW.NIHON-IR.JP/?PAGE_ID=211

*3, HTTP://OUKAJINSUGAWA.HATENADIARY.JP/ENTRY/2014/01/24/063703

機械学習によるパテントマップ作成

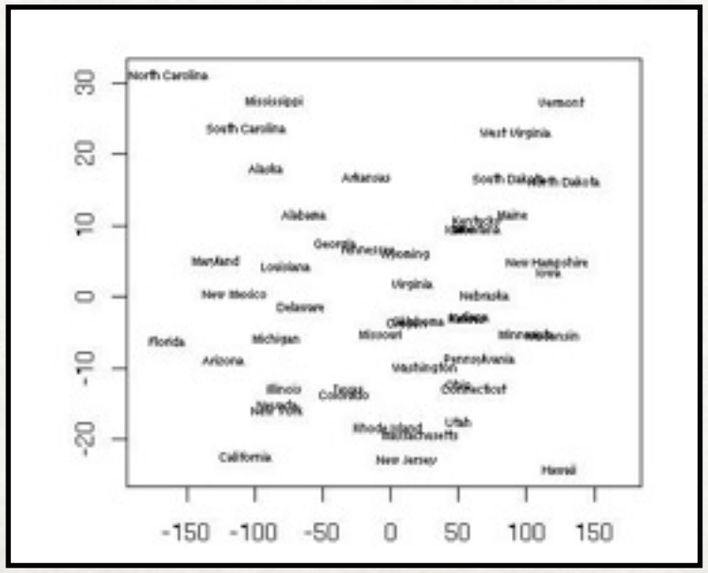
従来：文書間の類似度によって作成

本研究：文書だけでなく分類基準（分野、年代）を考慮

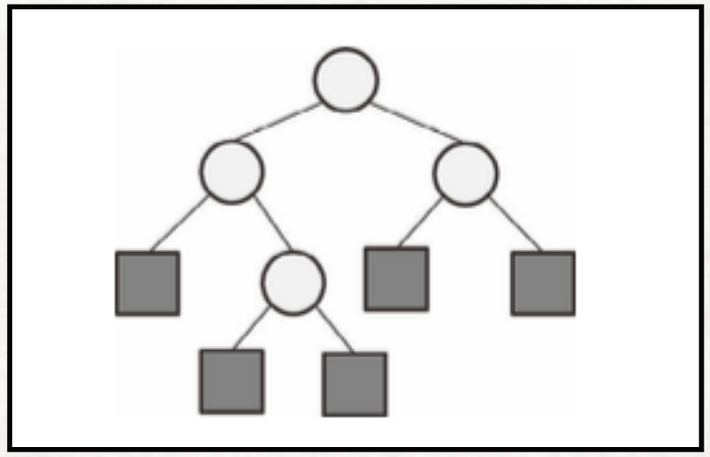
ー＞より意思決定を支援する上で有益な情報を獲得できる

従来

MINAMOTO,0,70,60,90,70,60,80,100,110,80
KASAIRINKAI,70,0,60,80,50,40,20,70,80,60
TONERI,60,60,0,70,50,40,70,80,90,90
HIKARIGAOKA,90,80,70,0,50,60,70,70,80,100
YOYOGI,70,50,50,50,0,40,50,40,40,60
UENO,60,40,40,60,40,0,40,70,80,60
YUMENOSHIMA,80,20,70,70,50,40,0,70,80,60
KOMAZAWA,100,70,80,70,40,70,70,0,40,90
KINUTA,110,80,90,80,40,80,80,40,0,100
SHINOZAKI,80,60,90,100,60,60,60,90,100,0



本研究



研究のながれ

1.特許データから要約、特許請求の範囲でマイニング

2.形態素解析により文書データから名詞を抽出

3.単語の頻度と任意の特許の行列を作成

4.NMFにより行列をトピックに因子分解（縮約）

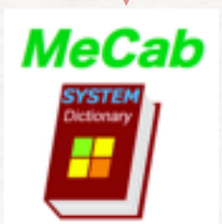
5.縮約とクラスを入力としてランダムフォレスト適用

6.予測が一致した木の数を類似度としてMDS適用

7.マッピングされた特許情報にカーネル密度推定法を用いて密集領域を特定

->パテントマップのできあがり

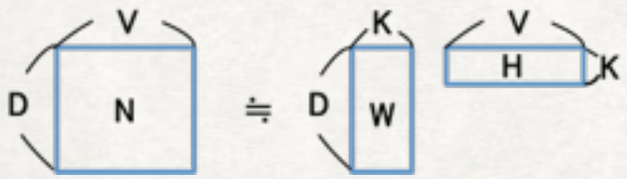
1



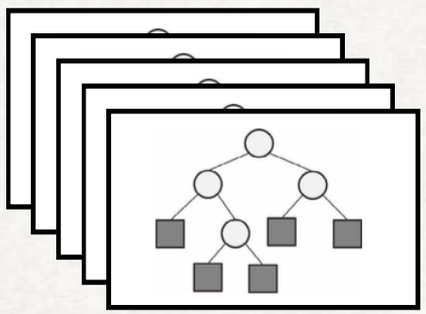
2 3

PATENT A,0,70,60,90,70,60,80,100,110,80
PATENT B,70,0,60,80,50,40,20,70,80,60
PATENT C,60,60,0,70,50,40,70,80,90,90
PATENT D,90,80,70,0,50,60,70,70,80,100
PATENT E,70,50,50,50,0,40,50,40,40,60
PATENT F,60,40,40,60,40,0,40,70,80,60
PATENT F,80,20,70,70,50,40,0,70,80,60

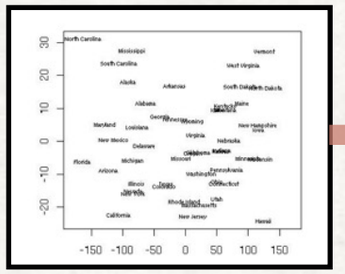
4



5

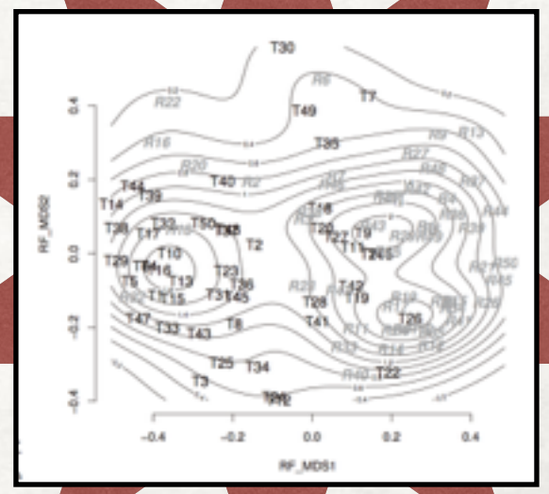


6



7

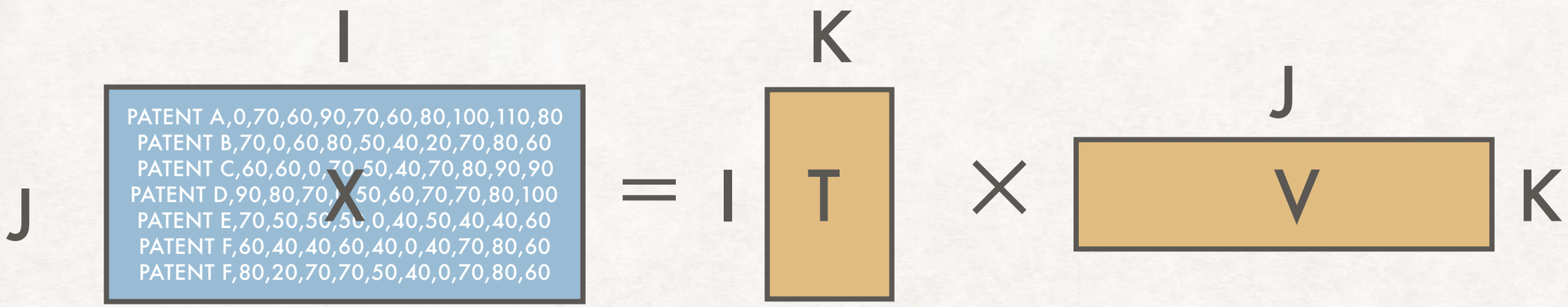
COMPLETE!



非負値因子分解NMF(NON-NEGATIVE MATRIX FACTORIZATION)

行列を基底に基づいた行列の積として近似的に表現する次元縮約手法

BOWはスケールが大きいので当手法で縮約を行う



K:基底数, I:データ次元数, J:データサンプル数

ここでX≒TVとなるように近似させる

一般的にT,Vはそれぞれデータの重み、特徴量を表している

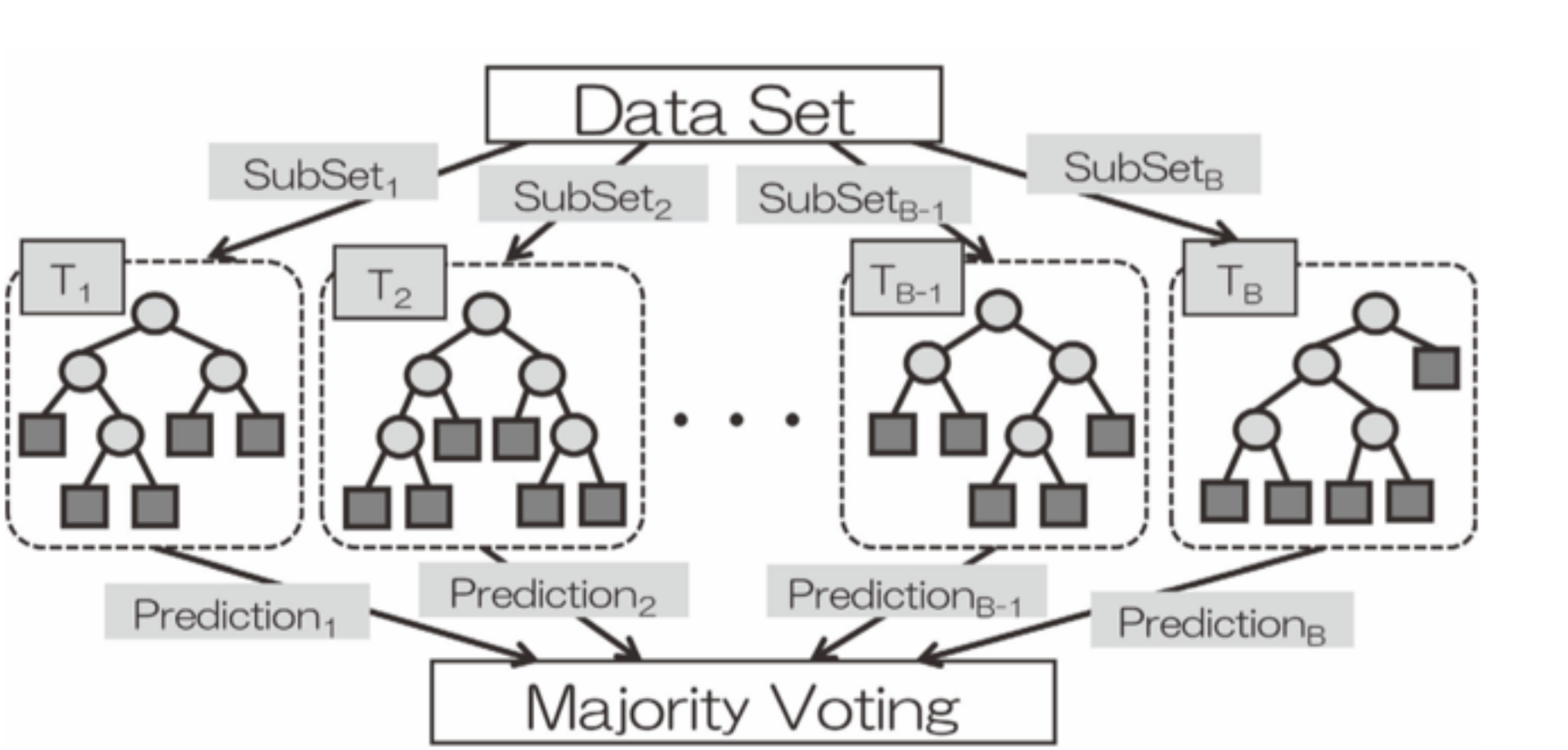
本研究ではVをトピック、Tを重みとする

そして|X − TV|の局所最小値を取るような行列に変換

ランダムフォレストRF(決定木)

技術分野を分類対象のクラスとして

NMFによる低次元密行列を元に学習を行う



ここで各決定木が識別結果を出力して正解だった木の個数を特許間の類似度として出力

ー＞これにより単語の頻度だけでなく、クラスを考慮したマッピング可能

多次元尺度構成法MDS(MULTI-DIMENSIONAL SCALING)

比例尺度や間隔尺度等の定量的数値データをマッピングする手法

本研究ではRFからの類似度行列に対して適用

0,70,60,90,70,60,80,100,110,80
70,0,60,80,50,40,20,70,80,60
60,60,0,70,50,40,70,80,90,90
90,80,70,0,50,60,70,70,80,100
70,50,50,50,0,40,50,40,40,60
60,40,40,60,40,0,40,70,80,60
80,20,70,70,50,40,0,70,80,60

$$z_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \sum_{i=1}^{n} \frac{d_{ij}^2}{n} - \sum_{j=1}^{n} \frac{d_{ij}^2}{n} + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{d_{ij}^2}{n^2} \right)$$

式A

d_{ij}: 2点i, jの距離

式Aから求めた行列Zの固有値ベクトルをマッピングする座標値とする

距離（非類似度）や類似度を求める手法は様々ある
論文では割愛されていたがユークリッド距離、ミンコフスキー距離、ピアソン相関係数等が有名

実験結果

表1　分析におけるパラメータ設定

パラメータ	値
森のサイズ:B	50
各木で選択されるの変数の数	7
NMF の基底数:K	50
NMF の更新回数	100

実験は「ロボット」、「テキストマイニング」（以下R,Tと呼ぶ）を対象として計50件分析

・比較対象として決定木の代わりにコサイン類似度によるマップを表示

表２はNMF適用の特徴量（トピック）とその内訳であるが、

例

表2　NMF の学習結果において重み係数が大きい単語の一例

変数	対応する単語とその重み									
V31	制御	移動	脚	駆動	接地	障害	位置	動作	衝突	歩行
	49.28	49.04	35.96	26.74	22.15	14.30	12.87	11.70	11.44	10.89
	設定	指令	軌道	車輪	旋回	関節	追随	計測	スイング	車体
	9.98	8.88	8.69	8.30	8.30	7.39	6.96	6.94	6.72	6.23
V20	光	クリーニング	集	位置	ミラー	装置	EUV	搬送	発生	回転
	86.79	40.27	70.96	69.19	52.38	43.16	31.96	26.01	18.49	13.89
	反応	クリーニングチャンバー	ガス	寿命	物質	汚染	交換	付着	対応	反射
	8.06	10.43	6.92	4.43	4.22	4.11	3.61	3.44	3.25	2.80
V46	情報	コミュニケーション	タイム	ライン	抽出	ログ	ライフ	商品	提供	関連
	142.97	36.73	28.29	27.80	23.93	22.97	22.97	21.35	19.89	19.19
	収集	データベース	検出	ユーザ	個人	端末	要素	蓄積	取得	DB
	17.91	11.61	16.65	16.19	13.29	12.53	10.69	7.53	6.34	5.41
V40	請求	コンテンツ	結果	情報	要因	クエリ	デバイス	ユーザ	記載	フォーマット
	58.47	28.72	48.53	45.99	45.84	43.30	28.93	28.23	27.97	21.08
	統計	レイティング	選択	アイテム	広告	スコア	販売	システム	ランキング	モジュール
	27.85	9.85	17.60	15.32	14.56	12.48	9.85	9.55	8.65	8.39
V37	情報	グルーピング	文書	データ	ユーザ	検索	オブジェクト	端末	登録	サーバ
	96.86	9.99	41.03	28.97	20.27	11.76	9.96	8.96	8.37	6.87
	管理	プレゼンテーション	ファイル	プログラム	送信	保存	グラフ	サマリー	企画	取得
	6.84	4.41	4.40	3.47	3.31	3.09	2.94	2.92	2.66	2.55

実験結果

・上図が本研究におけるパテントマップ

・下図がMDSとコサイン類似度を用いたパテントマップ

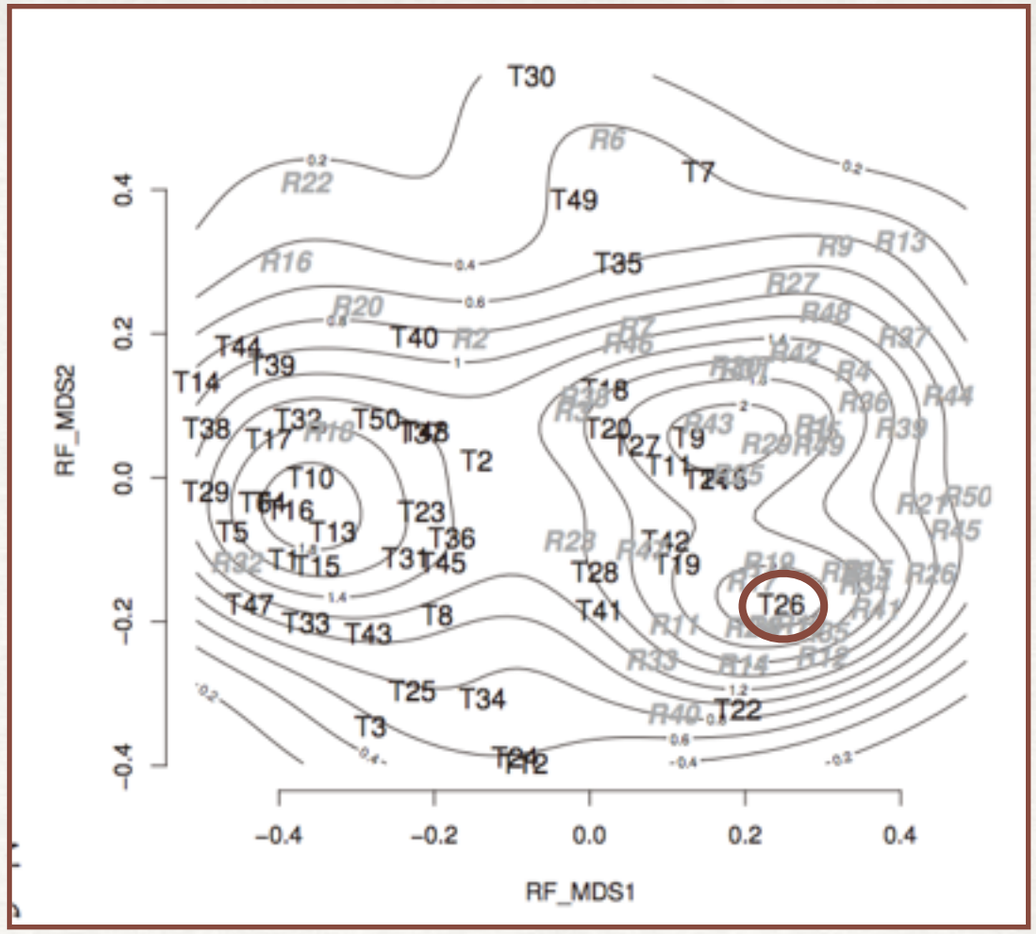
上図は右にR、左にTが密集している

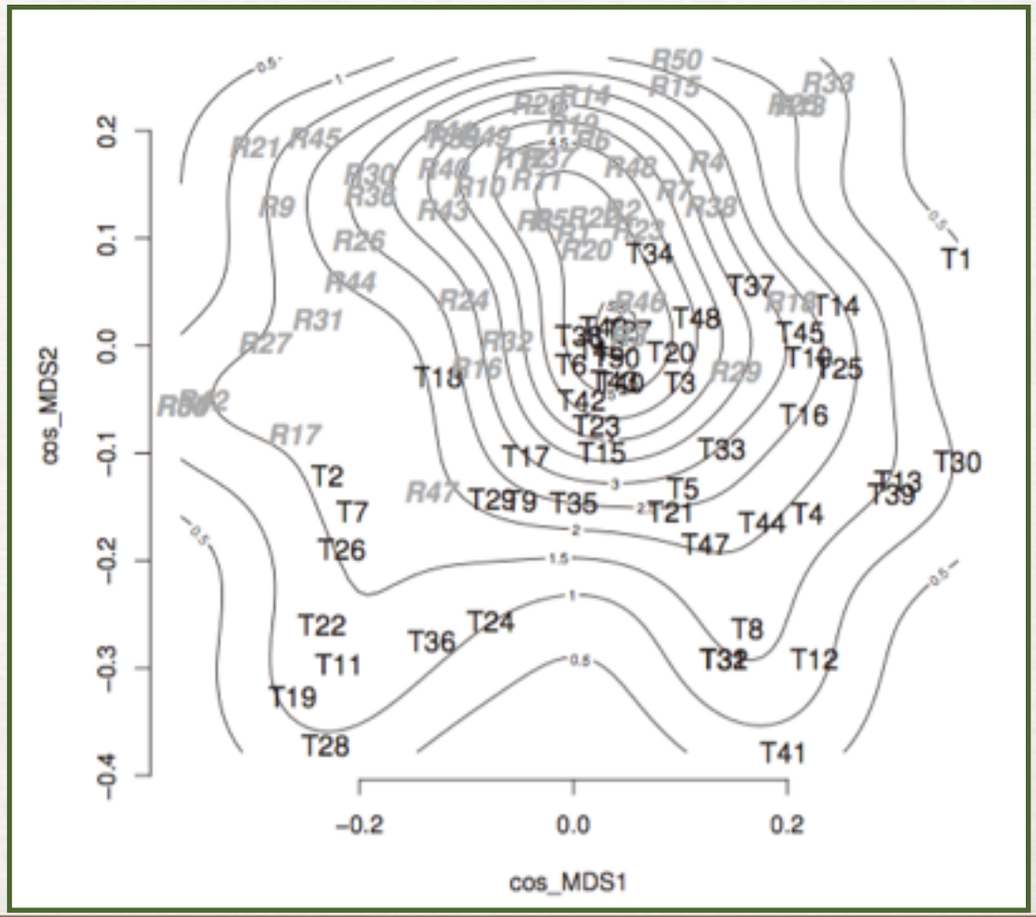
・Rはさらに上と下に領域が別れている

=>Rは機械や制御等の多様分野からなることを表している

・T26はT分野であるがR分野への応用が期待できる有益な個体であると解釈できる

・下図は領域は分類できているが、密度が一箇所に集中しており異分野間の適用領域が把握が困難





まとめ

・本研究により特許に関するデータの視覚的な理解支援を実現
・将来的には機械学習や計算機統計学をベースにした技術情報分析の期待は高まる可能性

課題

・ネットワーク分析、係り受け解析等の別種のマップ化
・文書以外の情報とのデータ統合

ー＞この研究のクラスを「技術分野」から「特許の価値」におきかえれば特許評価モデルも作成可能？