Selection of Core Words from Textual Patent Data with DEA based on Citation

情報基盤工学講座　M1　小野田成晃

2018/5/16

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DEAを用いて特許価値モデル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

Google Patents

特許文書データ

データ収集

機械学習を用いて知識抽出

価値モデルを生成

機械学習によるパテントマップ作成

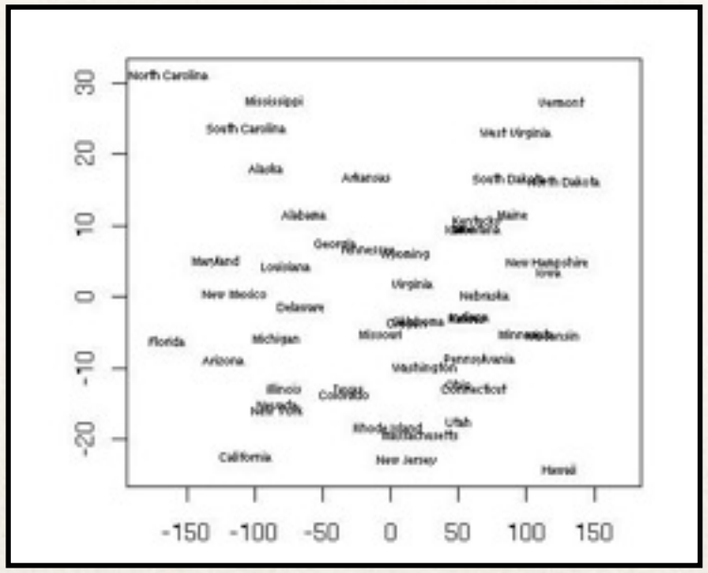
従来：文書間の類似度や決定木によって作成

本研究：文書だけでなく特許価値を考慮

ー＞マルチモーダルな特許価値モデルを生成できる？

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

DMU	効率値	参照集合をあらわすウェイト											
		λA	λB	λC	λD	λE	λF	λG	λH	λI	λJ	λK	λL
A	1	1	0	0	0	0	0	0	0	0	0	0	0
B	0.674	0	0	0	0	0	0	0.404	0	0	0.124	0.054	0
C	0.943	0	0	0	0	0	0	0.889	0	0	0.21	0.113	0
D	0.885	1	0	0	0	0	0	0	0	0	0	0.265	0
E	0.33	0	0	0	0	0	0	0.007	0	0	0.38	0.256	0
F	0.757	0	0	0	0	0	0	0.631	0	0	0	0	0
G	1	0	0	0	0	0	0	1	0	0	0	0	0
H	0.755	0	0	0	0	0	0	0.789	0	0	0.715	0	0
I	0.638	0	0	0	0	0	0	0.276	0	0	0.184	0.368	0
J	1	0	0	0	0	0	0	0	0	0	1	0	0
K	1	0	0	0	0	0	0	0	0	0	0	1	0
L	0.556	0	0	0	0	0	0	0.103	0	0	0.176	0.956	0

10

0でない値を持つ事業体の集合が事業体Eの参照集合

?

データ収集

特許情報プラットフォームは整理されておらず, 時間ごとのアクセス制限を設けていたためビックデータ収集の基盤として不適切

そこで

よく整理されており, マークアップにより構造化されたGoogle Patentを対象としてデータ収集を行った

DEA

- DEAとは各DMUに対して利益が最大となるように重み付け評価をすることで多次元パラメータの効率性を求めるデータ解析手法

$$\left.\begin{aligned} \max_{u,v} \quad & \frac{u^t y_o}{v^t x_o} \\ \text{subject to} \quad & \frac{u^t Y}{v^t X} \leq 1 \\ & u \geq 0 \\ & v \geq 0 \end{aligned}\right\} \qquad (1)$$

$$\left.\begin{aligned} \max_{u} \quad & u^t y_o \\ \text{subject to} \quad & v^t x_o = 1 \\ & -v^t X + u^t Y \leq 0 \\ & u \geq 0 \\ & v \geq 0 \end{aligned}\right\} \qquad (2)$$

- 上式のようにv^(t)x_0を1と仮定して線形計画問題に持ち込む手法CCRを用いる

本研究におけるDEA

２つのルールの結果を比較・検証

- ルール１:そのまま単語の頻度分布を入力とする
  - 良い特許における単語の頻度分布が予測可能？
- ルール２：特許i中の単語w_{k}の出現頻度n_{i}の逆数を入力とし,出現が0であれば入力を定数Z=2とおく
  - 単語数が少ない割に仮想出力が大きければ良いDMUという仮説に基づく

DEA入力の概念表

入力

出力

	W1	W2	…	Wk	引用特許	被引用特許
P1	6	0		3	1	0
P2	2	4		0	3	4
.						
.						
Pi	1	2		1	1	1

実験結果

まず, 情報分野G06を対象として分析を行ったが

情報分野は他分野の応用に使われるため, 単語種数が多かった

そのため, 入力値がスパースになり重みが上手くでなかった

そこで, 照明分野H05Bを対象として分析を行った

ルール１，ルール２の結果とそれぞれに次元縮約をかけた計４パタンの結果を示す

	in_減少率	in_AAA	in_AA=	in_ABC	in_AC	in_ACアダプタ	in_AG	in_AIGHT	in_AL	in_ALO	...	in_鶏	in_鶏舎	in_黄土色	in_黄色	in_黒	in_黒色	in_黒鉛	in_齢	out_CBNNUM	out_CNUM
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.250000
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.200000
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.118186	0.254543
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001120	0.142537
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.250000
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.056533	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.333333
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.045455	0.000000
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.250000

DEAの入力に対する重みの表（一部抜粋）

実験結果

各ルールにおいて重みが高く付いた単語

重みの高い上位 10 単語の抜粋（ルール１）

Word.	Weight	Frequency
アントラセン anthracene	5.855055	3
下方 belowdown	2.343041	26
クラック crack	1.785240	10
さ unable to translate	1.721771	117
光 light	1.603364	948
アルミニウム aluminium	1.405620	33
システム system	1.345943	70
エネルギー energy	1.070877	31
アノード anode	0.999145	29
お呼び invitation	0.920917	218

重みの高い上位 10 単語の抜粋（ルール２）

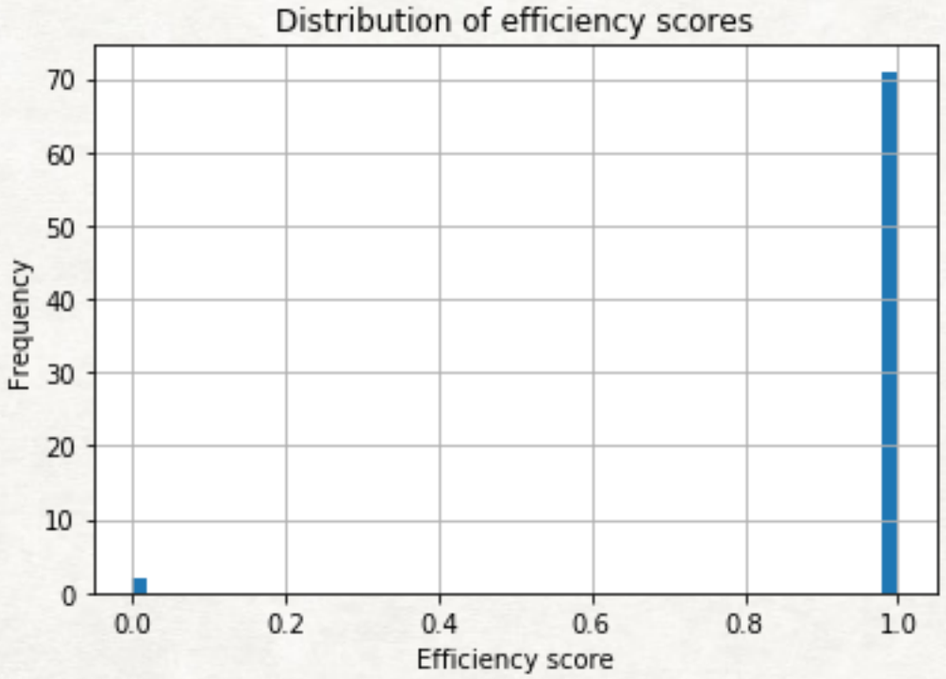
Word.	Weight	Frequency
図 figure	0.656393	2404
発明 invention	0.047136	1545
コイル coil	0.226496	108
実施 enforcement	0.111100	1188
調 style	0.055942	348
所定 stated	0.051680	414
回路 circuit	0.047136	1545
層 layer	0.046728	1694
チューブ tube	0.044257	8
原子 atom	0.040127	21

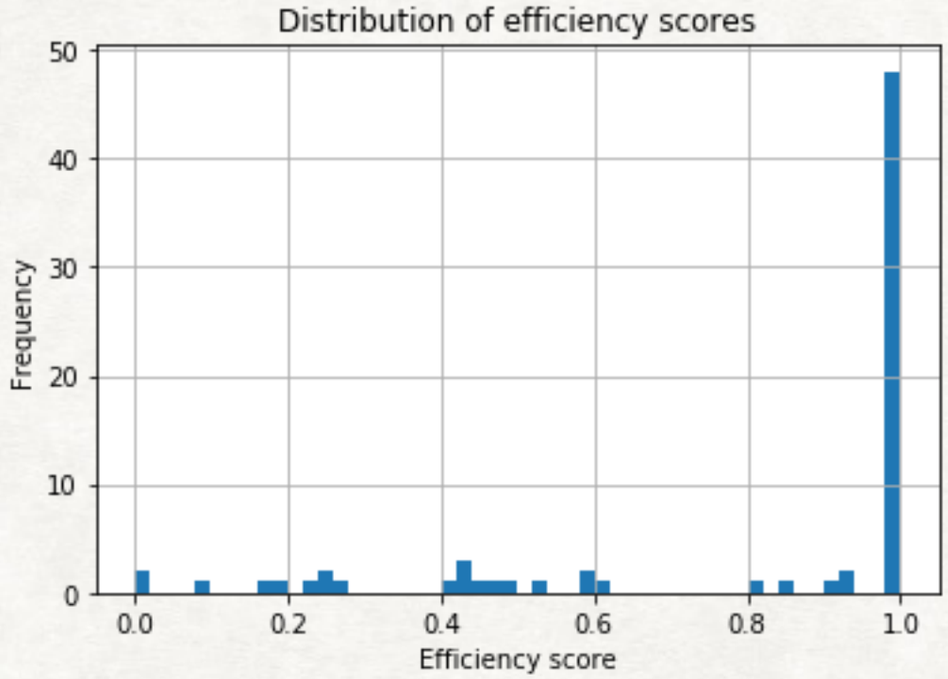
当初の仮説の通り、1は頻度が低いものが重みが付く傾向にあり

2は頻度が高いものが重みが付く傾向にある

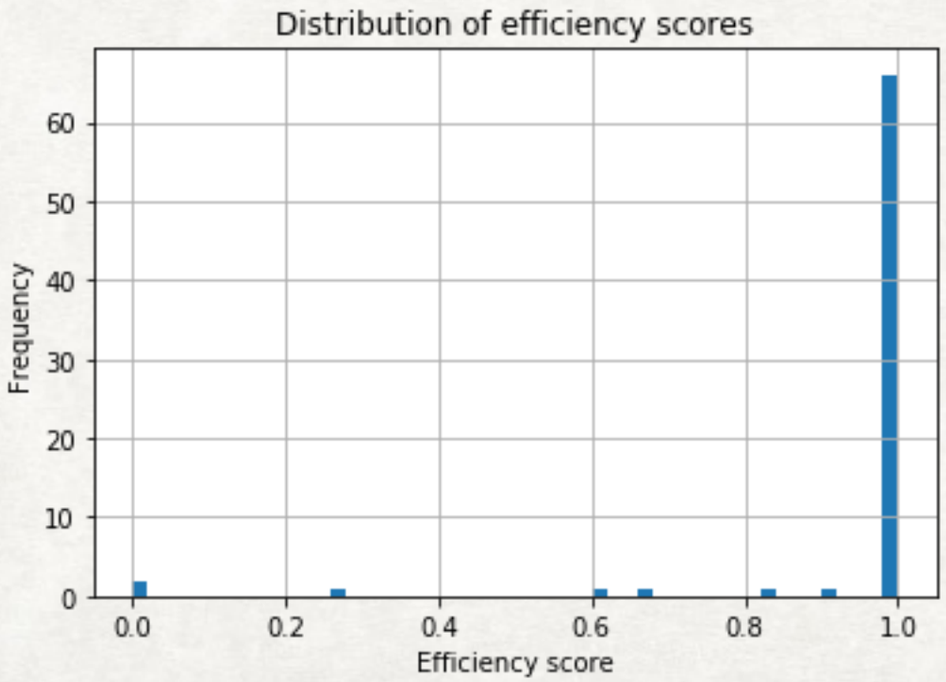
* 単純に頻度だけで順位が決まっているわけではないので引用も考慮されていると考えられる

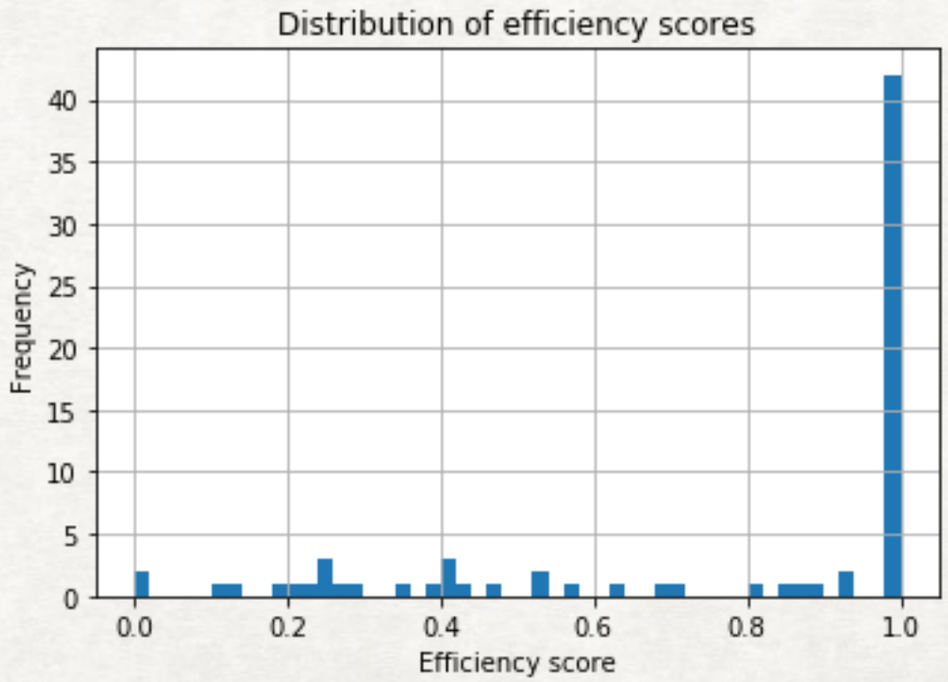
実験結果





ルール1の結果（左：非圧縮, 右：圧縮後）

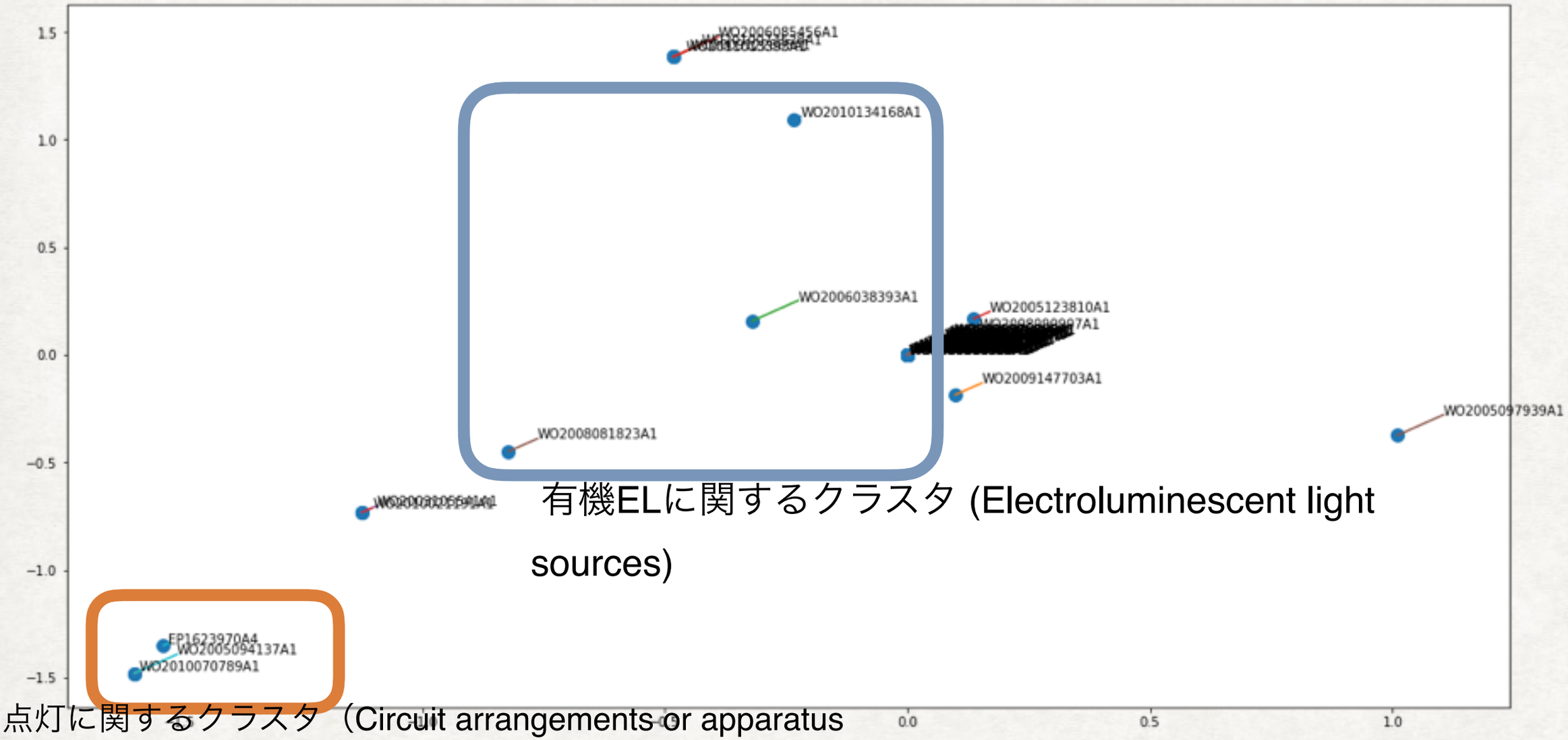




ルール2の結果（左：非圧縮, 右：圧縮後）

実験結果

コレスポンデンス分析による可視化(結果はルール2に対してのもの)



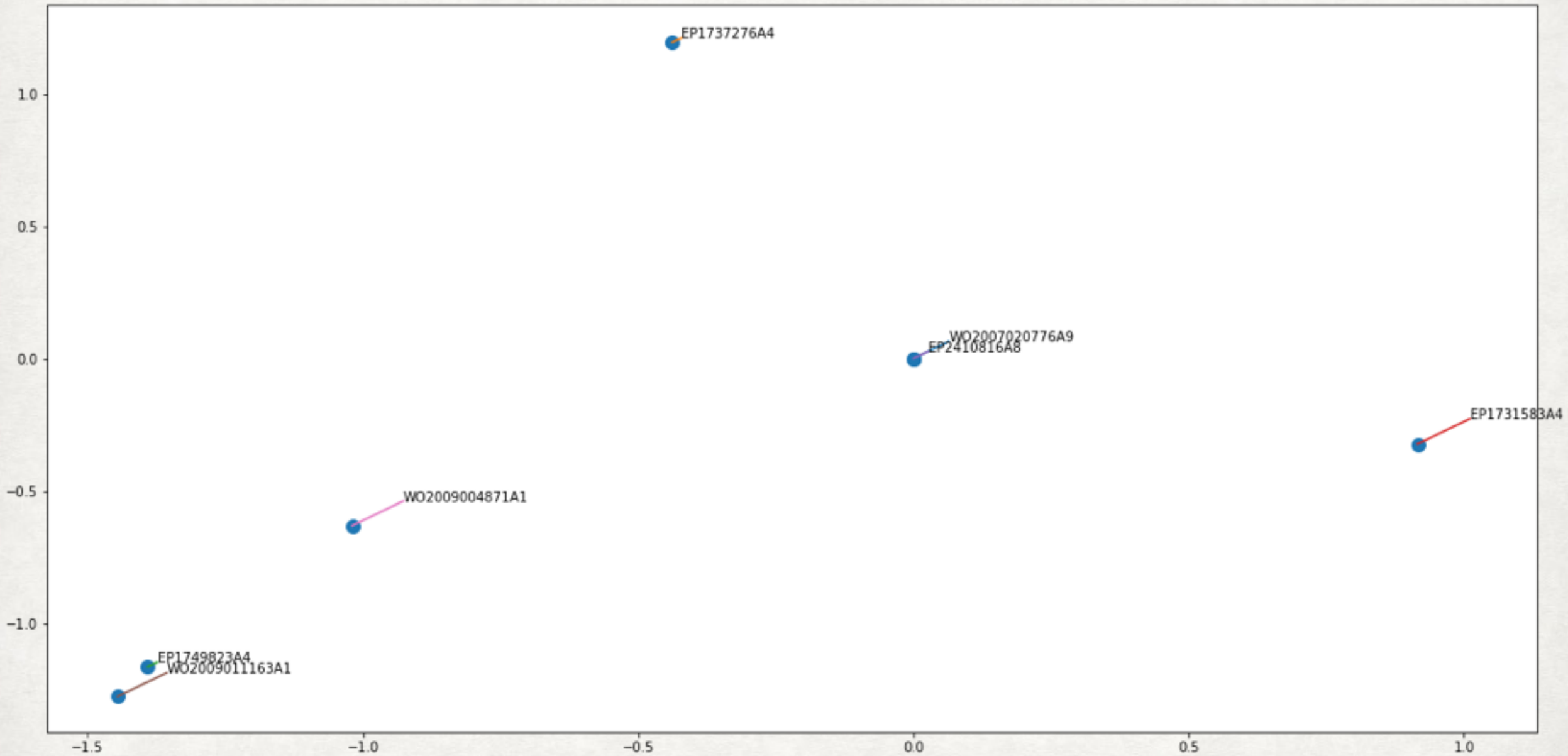
有機ELに関するクラスタ (Electroluminescent light sources)

点灯に関するクラスタ（Circuit arrangements or apparatus for igniting or operating discharge lamps）

効率的なDMU66個の位置関係

実験結果

コレスポンデンス分析による可視化(結果はルール2に対してのもの)



非効率的なDMU7個の位置関係

おわりに

- 可視化した結果から同じ効率値が同じ値の中での関係性が感覚的に把握できる

- また特許の価値を含んだ数値データができるので今後の新しい知財創出の助けとなる

- 入力がスパースな問題を縮約手法等を用いて改善する

- 単語のステミング問題が深刻なため解決策を思案中

- =>対応は富山の国際会議以降になる