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

1st Shigeaki Onoda
Dept. Information Systems Engineering
Toyama Prefectural University
Kurokawa, Imizu, Toyama, Japan
t855005@st.pu-toyama.ac.jp

2nd Mamoru Ota
Dept. Information Systems Engineering
Toyama Prefectural University
Kurokawa, Imizu, Toyama, Japan
ota@pu-toyama.ac.jp

3rd Koji Okuhara
Dept. Information Systems Engineering
Toyama Prefectural University
Kurokawa, Imizu, Toyama, Japan
okuhara@pu-toyama.ac.jp

***Abstract*—The web includes enormous data such as patent. This research find the rule of textual patent data and create new model.**

***Index Terms*—patent, big data, DEA, natural language proccessing**

I. Introduction

ICT 分野の発達により, 民間団体や政府機関のデータをデジタル化することの重要性が増している. 日本の総務省では, 1. 行政の透明性・信頼性の向上, 2. 国民参加, 官民協働の推進, 3. 経済の活性化・行政の効率化のためオープンデータ活用の促進を目指している [1].

その一例として, 災害関連情報を公開する試みが行われている [2]. このように様々な分野で科学技術による効率化が図られているのに対して, 特許分野も例外ではない. 日本の特許庁に提出された特許や実用新案等を誰でも容易に検索・照会可能なサービスの一つとして特許情報プラットフォーム[1]がある. そこでは日本の特許庁に提出された特許や実用新案等が掲載されており,Web サイト上で特許をキーワード検索することで特許利用の効率化を図っている. 特許のような複雑なデータを ICT や機械学習を用いて, 以前よりも効率的に特許の分析・調査が可能になると期待できる.

加えて，人工知能の第三次ブームにより人工知能の多方面の応用研究が盛んになっており，特許調査への応用事例がある．実用化された特許調査ツールには，トーマツデトロイト社が開発した PLSA をもとにした特許調査ツール Doloitte Analytics がある．また日本においても FRONTEO 社と TTDC: Toyota Techno Developement 社が共同で開発した KIBIT が市販されている [3]．しかし, 現状では特許のような非構造データはデジタルデータとして完成しているわけではない.

そこで，本研究では特許データのような文章, 引用数, 特許種, 日時等, 複雑なパラメータを組み合わせからなる非構造化データを統計的・数理計画的な側面からの評価するため, 日本語の特許データに対して DEA(Data Envelopment Analysis) を用いる. 以上の試みにより，特許文書における引用件数等の価値を含んだ単語の重みモデル構築手法を提案する.

特許データの研究に関してはパテントマップ作成を目的とした分析型と特許情報から新しい技術領域や知識を探る推論型がある．パテントマップとは特許調査に用いられる資料の一つで, 特許の情報を分析・整理・抽出などを行い情報の可視化を目的として使われるものである [4]. 分析型の主な研究として, パテントマップを数理的に作成する手法としては，Young らの対象となる技術分野の単語から特許の文書を k-means 法を用いてクラスタリングするものがある [5]．これはクラスタリング後の結果から semantic network を構築し可視化を行う手法である. また大田らのように, 調査対象の特許分野の少数の特許を入力とし，それに対してランダムフォレストを用いることにより特許の集合から目的に合致するサンプルを選択する．そしてそのサンプルの文字データに対して自然言語処理により課題表現を抽出してパテントマップを作成するというものもある [4]. また, 酒井らのように特許の文書の傾向を分析して自然言語処理的なアプローチでの技術課題の抽出手法も提案されている [6]. 例えば「ができる.」等の手がかり表現を含む文節に課題表現があるということが判明している.

しかし，これらの研究は特許の文書以外の情報，例えば特許分類，引用件数や発明者・企業等の情報を考慮していない. そこで推論型の一研究として, 津村らは特許中の出現単語と特許の分類の両方を結合したデータを作成するために，ランダムフォレストによる特許分類の学習を行った．その出力結果に対して MDS を用いたマッピングを行い，特許の分類情報を含んだ特許間の類似度空間を構築した [7]．本研究では津村らのマルチモーダルな単語空間の作成に関連して特許の価値と単語の出現傾向を考慮した研究を行った．価値と単語の頻度を結びつける方法として DEA を用いる．特許分野に関して DEA を用いた研究は Hyonju らの研究がある [8]．Hyonju らは企業と産業領域を対象としてそれらの類似性を発見することを目標としたのに対して本研究では特許の単語情報と引用件数等を素性としたときの価値モデルの構築とそれらから写像される特許文書内での重要単語の発見を目的とする. 津村らとの違いは教師あり学習を用いていないため, 特許分野のドメイン知識がない技術者や経営者が特許分析する際にも利用できる点である.

具体的には，DEA により各特許を事業所になぞらえて各特許ごとの価値モデルを構築する，DEA の結果により得られた各単語の重みを統計的な分析を行うことにより特許文書における重要な単語を得られる．また，DEA を解くことで各事業所の改善案を得られるが本研究においては特許内の単語の頻度を調整することにより，その特許に不足しているキーワードを数理的に推測する手法を提案する．本手法により単語の情報と特許の価値というマルチモーダルな

[1]https://www.j-platpat.inpit.go.jp/web/all/top/BTmTopPage

テキストマイニングの可能性を探求する．

本稿では，まず本研究で必要な基礎理論である DEA の基本的なモデルについて述べその可視化手法についても説明する. 次に特許データ取得の手順・方法について述べる．さらに 4 章では実験結果ならびに考察を示す．最後に結論で結びとする．

II. Methodology

本章では本研究の基礎となる理論と手法について説明する．

A. DEA

DEA とは Charnes と Cooper が提唱した経営分析手法 [9] であり, 絶対尺度の存在しない企業等の組織評価に用いられる [10]. DEA は, 分析対象となる組織を DMU (Decision Making Unit) に対して産出と投入の比から各 DMU の生産性を測定する. 例えば事業所における従業員数と売上高がわかっていれば, 事業所の効率性を (売上高 / 従業員数) として計算できる. しかし，実社会の場合は 2 変数だけということは少なく，来客数，店舗数，在庫数等あらゆる変数が複雑に組み合わさっている. DEA はそのような多変量で絶対尺度が存在しない場合のユニットの効率性を求める場合に効果を発揮する. また効率的な DMU と非効率な DMU を比較することにより, どのリソースを調整すれば，より効率的になるか数理的に推測できる k 番目の DMU の効率性 θ_k を求める分数計画問題は次式で定義される.

$$\begin{aligned} \text{max}: \quad & \theta_k = \frac{\sum_{n=1}^{N} v_{kn} y_{kn}}{\sum_{m=1}^{M} u_{km} x_{km}} \\ \text{subject to}: \quad & \frac{\sum_{n=1}^{N} v_{kn} y_{sn}}{\sum_{m=1}^{M} u_{km} x_{sm}} \leq 1 \ (s = 1, 2, 3, ..., K) \\ & u_{km} \geq 0 \ (m = 1, 2, 3, .., M) \\ & v_{kn} \geq 0 \ (n = 1, 2, 3, .., N) \end{aligned} \tag{1}$$

ここで K は DMU の数で $k \in K$ となり, x, y は入力, 出力を表し, M, N はそれぞれ入力・出力変数の数で $m \in M, n \in N$ となる. u, v はそれぞれ指標に対する重みとし s は評価対象の DMU の番号とする. 式 (1) の目的関数の分母が仮想入力を表し, 分子が仮想出力を表す. この式では k の効率性は仮想出力/仮想入力の式で, k にとって最大の効率値が得られるように重み付けを行う. つまり, 自分に対して最も有利に評価するのが DEA の肝である．また式 (1) の DEA モデルは入力となる投入が少ないほど良いというスキームに基づいている．

式 (1) の分数計画問題を線形計画法のアルゴリズムで解くために, CCR モデルが実用的と考えれられている b10. 本研究で用いた CCR モデルについて次項で説明する.

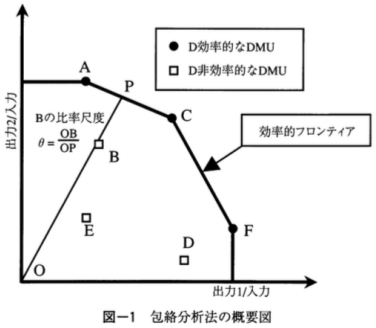
B. CCR Model

式 (1) の分数計画問題を線形計画問題として解くために分母または分子を 1 にする DEA の基礎モデルを CCR モデルと呼ぶ. 分母・分子どちらを 1 にするかは入力と出力のどちらを基準とするかで変わる. 仮想入力を 1 として式 (2) のように定式化したものを入力指向モデルといい. 逆に仮想出力を 1 としたものを出力指向モデルと言い式 (3) のように定式化できる．入力指向モデルでは, 効率値の最大化は 仮想出力の最大化に等しくなるので良い出力を得られる DMU を評価できる. 反対に出力指向モデルでは効率性の最大化は入力の最小化に等しいため，入力に関しての改善案を提示できる．どちらのモデルが合っているかは入力・出力のどちらを改善したいかで判断することとなる．

$$\begin{aligned} \text{max}: \quad & y_k = \textstyle\sum_{n=1}^{N} v_{kn} y_{kn} \\ \text{subject to}: \quad & \textstyle\sum_{m=1}^{M} u_{km} x_{sm} - \sum_{n=1}^{N} v_{kn} y_{sn} \geq 0 \\ & \textstyle\sum_{m=1}^{M} u_{km} x_{km} = 1 \ (s = 1, 2, 3, ..., K) \\ & u_{km} \geq 0 \ (m = 1, 2, 3, ..., M) \\ & v_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \end{aligned} \tag{2}$$

$$\begin{aligned} \text{min}: \quad & x_k = \textstyle\sum_{m=1}^{M} u_k x_{km} \\ \text{subject to}: \quad & \textstyle\sum_{m=1}^{M} u_{km} x_{sm} - \sum_{n=1}^{N} v_{kn} y_{sn} \geq 0 \\ & \textstyle\sum_{n=1}^{N} v_{kn} y_{kn} = 1 \ (s = 1, 2, 3, ..., K) \\ & u_{km} \geq 0 \ (m = 1, 2, 3, ..., M) \\ & v_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \end{aligned} \tag{3}$$

図 1 にに DEA の概念図を示す. 図中の効率的フロンティアとは DEA により示された効率的かどうかの境界であり, ここに近ければ近いほど DMU は効率的であるといえ, 重なっている DMU は効率値が最大の 1 を取るものとなる.



図−1 包絡分析法の概要図

Fig. 1 DEA の概念図

DEA を適用することにより，効率値が 1 の DMU は効率的, それ以外は非効率と分別される. 効率的と非効率な DMU が導かれる. そして非効率な DMU_k の存在はそれを非効率たらしめる優位な DMU から成り立っている. これに対して優位な DMU の集合を参照集合という. 例えば, 図 1 中の B の参照集合は A と C である. つまり A と C の存在により B が非効率と判断されているのである. 非効率な DMU_k をより効率に改善するにはこの参照集合に入力を近づける必要がある．

そしてある DMU_k に対しての参照集合中の DMU の参照度合を導くには上述の CCR モデルの双対問題を解くことで判明する．具体的には入力指向の場合は式 (4)，出力指向の場合は式 (5) のようになる．

入力指向モデルの双対問題

$$\begin{aligned} \text{min}: \quad & \theta_k \\ \text{subject to}: \quad & \lambda_s \geq 0 \ (s = 1, 2, 3, .., K) \\ & \theta_k x_{km} - \sum_{s=1}^{K} \lambda_s x_{sm} \geq 0 \ (m = 1, 2, 3, ..., M) \\ & \sum_{s=1}^{K} \lambda_s y_{sn} - y_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \end{aligned} \tag{4}$$

出力指向モデルの双対問題

$$\begin{aligned} \text{min}: \quad & \theta_k \\ \text{subject to}: \quad & \lambda_s \geq 0 \ (s = 1, 2, 3, .., K) \\ & \sum_{s=1}^{K} \lambda_s y_{sn} - \theta_k y_{kn} \geq 0 \ (n = 1, 2, 3, ..., N) \\ & x_{km} - \sum_{s=1}^{K} \lambda_s x_{sm} \geq 0 \ (m = 1, 2, 3, ..., M) \end{aligned} \tag{5}$$

ここで λ_s は対象となる k 番目の DMU に対しての参照となる s 番目の DMU の参照度合である. 全く参照していない場合は 0 を取る. また,λ は次に述べる可視化で用いる.

C. DEA の視覚化

例えば 2 入力 1 出力のように対象となる問題の変数が少なければ縦軸を入力 2/出力，横軸を入力 1/出力として二次元グラフをマッピングし，その関係性を視覚的に観察することができる．一方で, 多入力多出力の視覚化には．コレスポンデンス分析が利用できる [11].

Correspondence analysis とは列と行の対応指標を含んだ集計表 (主にクロス集計表が用いられる) をもとに分析する記述的手法である [12]．これは行の要素と列の要素の相関係数が最大となるように行と列の要素をソートして，表形式のデータを二次元空間にマッピングする．この結果からカテゴリーごとの構造や類似性を観察することができる [13].

コレスポンデンス分析のため豊澄らは，DEA を用いて参照集合 R_k と効率値 θ_k を求め, この結果をもとに作成された表 1 のクロス集計表を用いた．表の行には効率的な DMU_{n1}, 列には非効率な DMU_{n2} を用いる．ここで $n1 + n2 = n$ で n は分析対象となる DMU の数である．$lambda$ は式 (4), (5) の双対問題の参照集合の要素である. 本研究では, 豊澄らの二次元配置手法により, DEA により得られた結果の視覚化を行なった.

TABLE I
SUBSTITUTION TABLE TO CORRESPONDENCE ANALYSIS

		Efficient DMU		
Inefficient DMU		DMU_7	...	DMU_{n1}
	DMU_1	$\lambda_{1,7}$	...	$\lambda_{1,n1}$
	DMU_2	$\lambda_{2,7}$	...	$\lambda_{2,n1}$
	.	.	...	.
	DMU_{n2}	$\lambda_{n2,7}$	...	$\lambda_{n2,n1}$

III. COLLECT PATENT DATA FROM OPEN DATA

A. Collecting Method

前述のオープンデータは人手で少数の特許事例を調べるのには必要充分であるが，ビックデータとして特許全体の分析を行いたい場合は整理されているとはいい難い．例えば, データの保存形式が PDF 形式の場合や, 被引用特許の件数が掲載されていない等の問題点がある.

そこで, 今回特許の定量的な分析をするためのリソースとして, 日本語ドメインの Patent - Google[2]がある．Patent - Google は Google 検索オプションの一つで, 世界各国の特許データが html 形式で公開されている. これは PDF などの非構造データに比べてデータ整理・収集しやすい利点がある. またこの検索プラットフォームの他に Google Patent[3]がある. こちらは独自のドメインを持っており検索インタフェースと検索結果に多少の違いがある．現在の Google の特許検索の状況を整理するため図 2 を付す．Google Patent と Patent - Google はいずれも特許記事自体は patents.google.com ドメインで公開されている．そのため 2 つの特許の文書に本質的差異はない.

本研究では, 検索オプションが豊富で通常の Google 検索エンジンと同様に使える Patent - Google を情報収集のプラットフォームとして利用した.

まず, 必要なデータとして,1. 特許 ID,2. 発明者,3. タイトル,4. 承認日,5. 引用特許数,6. 被引用特許数,7. 本文に含まれる特許内の単語とその頻度とした. また, 本研究では津村らと同様, 特許に含まれている単語を分析対象としているため, 文章から抽出する素性は名詞のみとした [7]. なお収集特許は実験後の単語を詳細に分析するため，著者の母語である日本語で提出されたものを対象とした.

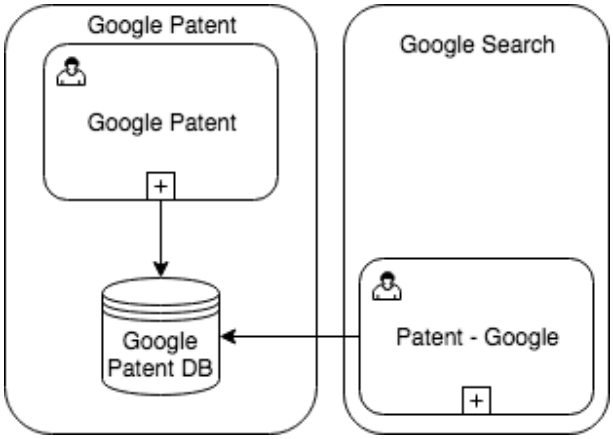


Fig. 2 Google における特許検索のプラットフォーム

B. DataBase

前項で収集したデータを蓄積・分析するためにデータベースを構築した. 収集データは 7 種類あるが, そのうち単語に関しては各特許に対して抽出できる種類数が異なるため, スケーラビリティに富む NoSQL である mongoDB を用いた. また, 収集した全特許に含まれる全単語種を分析に用いるため, 別途全単語辞書を構築した.

[2]https://www.google.co.jp/?tbm=pts
[3]https://patents.google.com

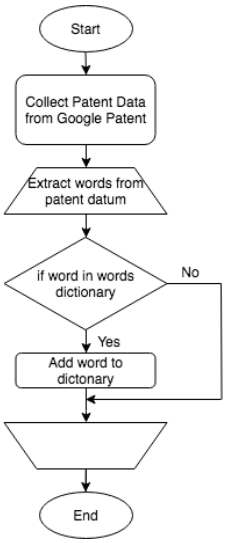


Fig. 3 単語辞書作成の手順

次に収集した特許の全単語辞書を生成する手順を図 3 に示す．辞書には単語とその合計頻度が入っている．まず，前述 Google Patent を対象として各特許記事から上述の 7 個の情報を抽出・保存する．そこから特許データベースを作成する．次にそのデータベースに対して再帰的に単語データを検索しそのデータを辞書に加える．そして, すべての含有単語を探索した後, 図 3 のループを終了して辞書を生成が完了する. なおすでに単語が登録されていた場合は回数のみ加算する．

IV. シミュレーション結果ならびに考察

A. *DEA* 手法の適用

DEA 手法の特許データに対しての適用としては,Hyeonju らの研究がある [8].Hyeonju らは特許の単語データから新技術の適用領域を探るために用いていたが, 本研究では入力を単語の頻度行列に絞ることで, 価値のある特許の算出とともに, 価値のある特許に含まれる単語の予測も目的の一つとして行う. これにより, 特許というドメインでのコアな単語を浮き彫りにできると考えられる. 次に具体的な定式化を示す.

本研究では DEA における仮想入力項目を特許に含まれる単語とその頻度とし, 仮想出力を引用件数, 被引用件数とする. これは後藤らの研究から特許の価値で重要な素性として被引用特許数と引用特許数が挙げられたためである [14]. つまり基本的に引用数が多いほど価値のある特許と判断できる．通常 DEA では仮想入力は小さく, 仮想出力が大きいほど効率的であるとするが, 本研究のもとでは仮想入力が単語のため, このルールが適用可能かどうかは実証されていない. そのため 2 つのルールを設け DEA を適用した.

- **ルール 1** そのまま単語の頻度分布を入力とする
- **ルール 2** 特許 i 中の単語 w_i 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 とおく

ルール 1 は含まれている単語数が少ない割に仮想出力が大きければ良いという仮設に基いて行い, 価値の高い特許モデルを生成することを目的とする. さらに良い特許での単語分布を予測するために用いる. またルール 2 は Luhn の文書の重要度の決め方に基づいて, 特許中の単語の出現が多ければ重要という仮説を用いる [15].

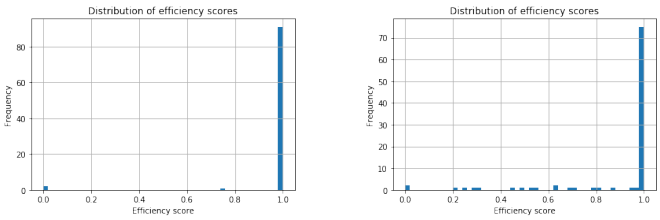
具体的には投入である単語が多いの対して, より大きな出力であれば効率的と評価するモデルである．これは DEA の投入が少なくて出力が大きいほど効率的という元のモデルと逆になるが，ルール 1 の結果との比較のため用いた．そしてこれらはそれぞれ異なる仮設に基づくのでそれぞれの出力結果を比較・分析することにより双方の結果を検証する.

また, 入力/出力の概念表を表 2 に示す．この表では入力は各特許 k における単語 w_i の頻度 f_{ki} としその特許に存在しない単語は上述のルールに従いルール 1 なら 0 をルール 2 なら Z を与える．なお本研究では $Z = 2$ と設定した. なお入力は 4 章で述べた全単語辞書と比較して作成する．本研究では DEA の CCR のモデルを用いた．特許データは情報 G06 分野と照明分野 H05B に対して分析を行い．前述の 2 ルールに対してそのまま入力指向での DEA を用いた場合と，入力データに対して次元縮約を行ったものに DEA を適応した計 4 パターンの実験を行った．出力指向モデルに関しては式 3 で示される通り，入力 x と重み v の積の総和が最小値となるような解を求めるので，目的関数の次元数が出現する単語分だけになり, 導く変数の数が膨大がになるため，本研究では入力指向を採用した.

TABLE II
DEA MODEL OF THIS STUDY

	inputs			outputs	
No.	w_1	...	w_i	citation	cited by
$Patent_1$	f_{11}	...	f_{1i}	c_1	cb_1
$Patent_2$	f_{21}	...	f_{2i}	c_2	cb_2
.	.	...	.	.	.
$Patent_k$	f_{k1}	...	f_{ki}	c_k	cb_k

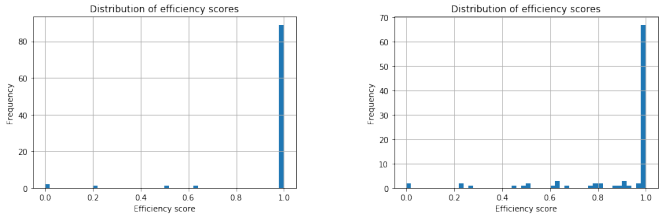
B. 実験結果



A: uncompressed　　B: compressed

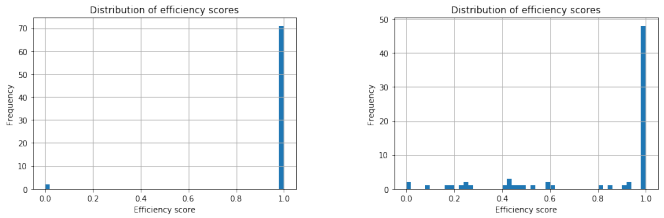
Fig. 4 Distribution of Efficiency Score on Rule1 in G06 field

まず情報分野 G06 に対して DEA を採用した. その結果をルール 1, ルール 2 それぞれ図 4，図 5 に示す. この分野は 94 記事中の単語の数が 10873 単語あり，単語自体も「走行中」や「化学構造」等他分野に渡るため，図 4, 5 の A の通り DEA の入力次元数が大きくなりほとんどの効率値が同じ値で 1 になった．原因としては入力行列がスパースになりどの特許の単語の出現も差異がないと見られたためだと考えられる．加えてループコイルとみられる単語「ルプコイル」等の誤字脱字が多く含まれていた．またこのデータに対し



A: uncompressed B: compressed

Fig. 5 Distribution of Efficiency Score on Rule2 in G06 field

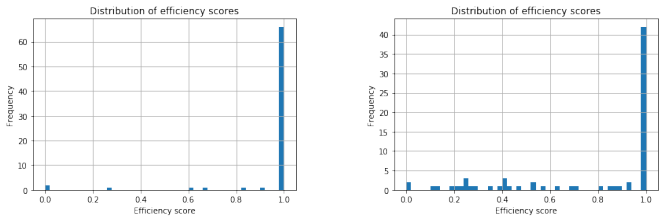


A: uncompressed B: compressed

Fig. 6 Distribution of Efficiency Score on Rule1 in H05B field

て，次元縮約をかけた場合は図 4, 5 の B ように効率値の分散が大きくなった．そこで次元数を減らせば効率値，重みの分散が大きくなると考え，単語の種類数や分野の幅が情報分野に比べ小さい照明分野 H05B の特許 74 記事，6252 単語に対して分析を行った．その結果を図 6, 図 7 に示す, それぞれルール 1, ルール 2 の効率値の分布を表している．依然としてほとんどの DMU が効率的と判断されているので，各 DMU の違いが浮き彫りにならなかった．そこで，次元縮約を行った結果は図 5 の左図に示すとおり効率値の差異がより明確になった．また, ルール 1 とルール 2 の効率値の分布は大きな違いが見られなかった.

また，各 DMU の各単語の重みの算術平均を算出して重みの高い単語上位 10 個を抜粋したものを表 3, 表 4 に示す，なおルール 1,2 ともに, 単語数が DMU 数に比して多いためほとんど 0 のスパースであった．そして，表 4 に示すとおり「発明」,「図」等の一般的な単語が上位に来ていることからルール 1 の仮説の通り投入が少ない割に良い出力がでている単語が高い重みづけされている．特に「アントラセン」は本特許データ中，3 回しか出現しておらず照明分野において良い特許に含まれる単語（以降，E-word と呼ぶ）を予測できている．



A: uncompressed B: compressed

Fig. 7 Distribution of Efficiency Score on Rule2 in H05B field

TABLE III
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 (unknown)	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

TABLE IV
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

同様にルール 2 についての重みの高い単語も表 5 に示す．こちらはルール 1 と反して「発明」,「図」,「コイル」等の特許文書内でより一般的な単語が含まれていることがわかる．これはルール 2 投入が多いわりに大きな出力であるという仮説と合致している．実際，これらは照明分野で高頻度の単語の関係と部分一致する．

以上の結果を見るとルール 1 で高い重みであった「アントラセン」や「アルミニウム」はルール 2 では重みの平均は 0 となった．逆にルール 2 で高かった「発明」,「図」はルール 1 ではそれぞれ重み平均が 0.020 と 0 という上位単語の重みに比べ低い値であった．双方のモデルともにある程度の重みが付けされる単語もいくつか観察できたが上位に入るような特異なものは見受けられなかった．

今回の入力単語に対して形態素解析をかけ，名詞を抽出したが,「いずれ」,「それら」のようなノイズと思われる単語も含まれていた．よってルール 1 では収集した特許分野での E-word を，ルール 2 ではその分野での一般的な単語（以降 G-word と呼ぶ）を抽出できたことが伺える．

C. 可視化の結果

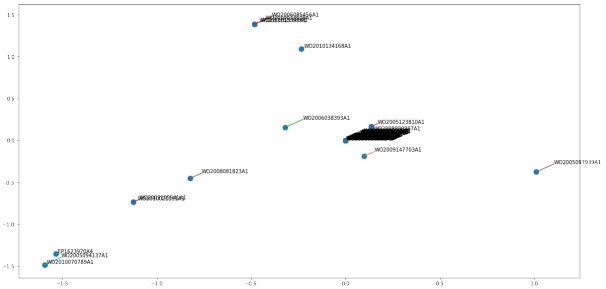


Fig7 Visualized Efficient DMU on Rule2

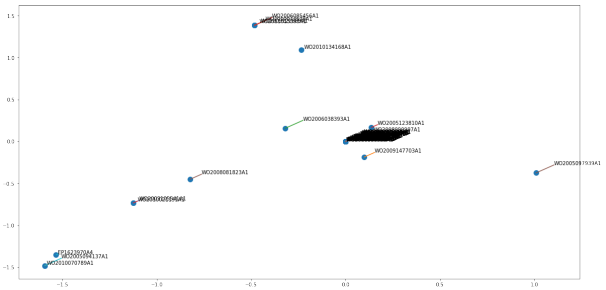


Fig8 Visualized Inefficient DMU on Rule2

前節で分析した DEA の結果を 1 章で述べた豊澄らの手法で可視化を行った結果を図 7，図 8 に示す．この通り DEA の分析結果から各特許の位置関係をマッピングしており，例えば図 7 の左下のクラスタである WO2005094137A1 や WO2010070789A1 は H05B41(Circuit arrangements or apparatus for igniting or operating discharge lamps) という同じ H05B の下位分類に属しており，中心のクラスタである WO2008081823A1 は WO2010134168A1 は有機 EL に関しての特許でどちらも同じ H05B33(Electroluminescent light sources) という下位分類である．

このようにある程度特許間の傾向をマッピングできている．また，縦軸と横軸の値が両方とも 0.0 をとる特許群はその DMU の参照集合が 0 であったため上手く相関取れなかったものである．これは元のデータがスパースであることに起因していると考えられる．

V. 結論

本研究は特許のデータから文書の傾向と引用件数等の評価という非構造なマルチモーダルデータから特許文書内の単語の重み付けを行った．これにより，特許の専門家が良い特許かどうか判断してラベル付けを行う教師あり学習とは異なり，特許の知識がない技術者でも特許内単語の価値を含んだ単語の重み付け手法を提案した．この評価指標は特許のみならず，文書と価値を含む変数がある論文等のデータにも適応可能である．この手法で導かれた単語の重みを利用して自然言語処理やデータマイニングにおけるマルチモーダルな機械学習に応用できると考えられる．また，単語の重み手法としても Luhn が用いた手法や TF-IDF のような頻度と位置だけから重要度を判断する手法よりも複合的な情報を含んだ重み付けが行えているため，評価値が絡む分野で適応可能である．さらに，それに対しての可視化を行うことで技術者や経営者の意思決定の支援に利用できると思われる．

また，前述の 2 ルールに関しては抽出したい単語や，どのような問題に適応するかでどちらのルールを使うか分ける必要があることが判明した．例えば専門用語の抜き出しに利用するのであれば E-word が適しており，特許らしい文書の自動作成に応用する場合は G-word が必要となるであろう．

今後の課題としては,「いずれ」,「それら」のようなノイズと思われる単語の存在が浮き彫りになった．本研究では名詞の抽出の際にストップワードの除去も行ったがいくつか漏れがあることが明らかになった．なので，ストップワード除去をより入念に行いデータの素性として抽出する単語をより具体的に選定することが必要となった．可視化についてはマッピングに対して意味付けの検証を行うため，今後特許に詳しい専門家の精査が必要であるだろう．

そして，入力次元数をそのまま単語の種類数とするとスパースな行列となり効率値にうまく差異があらわれなかったので，NMF のような入力単語の復元が可能な次元縮約手法を用いて次元数を減らしつつ個々の単語の重みも分析できるような試みを行う必要がある．また，取得する特許の数を増やし，入力次元数と DMU 数の比を減らした場合の振る舞いの検証も行っていきたい．さらに，これらの問題を解決することで，分野を絞らず特許という文書自体に関しての分析も行っていきたい．

REFERENCES

[1] Ministry of Public Management, ”Boost Open Data Strategy”, http://www.soumu.go.jp /menu_seisaku/ictseisaku/ictriyou/opendata/, Accessed: May 5, 2018.

[2] NTT Communications, ”Report of Deploying Information Infrastructure of Hazzard Data”, ”No. 1.0, 2013.

[3] Tsutomu Kiriyaam, Toshiyuki Ando, ”Patent Information and AI: Outline”, *Information Science and Technology*, Vol. 67, No. 7, pp. 340-349, 2017.

[4] Takahisa Ota, ”Patent Research using Machine Learning Methods”, *Information Science and Technology*, Vol. 67, No. 7, 2017.

[5] Young Gil Kim, Jong Hwan Suh, Sang Chan Park, ”Visualization of patent analysis for emerging technology”, *Expert Systems with Applications*, vol. 34 pp. 1804 UTF20131812, 2008.

[6] Hiroyuki Sakai, Hirohumi Nonaka, Shigeru Masuyama, ”Extraction of Information on the Technical Effect from a Patent Document”, *Journal of the Japan Society for Artificial Intelligence*, Vol. 24, No. 6, *I*, 2009.

[7] Takumi Tsumura, Fumiaki Saito, Shohei Ishizu, ”Knowledge Extraction from Textual Patent Data Using a Random Forest”, *J Jpn Ind Manage Assoc*, vol. 68, No. 3, pp. 161-170, 2017.

[8] Hyonju Seol, Sungjoo Lee, Chulhyun Kim, ”Identifying new business areas using patent informatin: A DEA and text mining approach”, *Expert Systems with Applications*, vol. 38, pp. 2933-2941, 2011.

[9] Charnes, A., Cooper, W.W., Rhodes, E., ”Measuring the Efficiency of Decision Making Units”, *European Journal of Operational Research*, Vol. 2, , pp. 429-444, 1978.

[10] Toshiyuki Sueyoshi, ”DEA: Business Analysis method”, Asakura, 2001.

[11] Kotaro Toyozumi, Yusuke Nishiuchi, Shingo Aoki, Hiroshi Tsuji, ”Correspondense Analysis based Position Visualization for DEA”, SCI’06, No. 50, Session ID: 2W1-2, 2006.

[12] K.Yamashita, Hang.Li, ”Mining Open A nswer in Questionnaire Data”, *IEEE Intelligent Systems*, Sept./Oct., pp.58-63, 2002.

[13] ”How To Analyze Simple Two-Way and Multi-Way Table, Correspondence Analysis”, URL:http://www.statsoft.com/Textbook/Correspondence-Analysis, Accessed Jun 1 2018.

[14] Akira Goto, Kiminori Genba, Jun Suzuki, Schumpeter Tamada, ”Classification Index of Important Patent”, *RIETI Discussion Paper Series*, 06-J-018, 2018.

[15] Luhn, H. P, “The Automatic Creation of Literature Abstracts”, *IBM Journal of Research and Development*, vol. 2, No. 2, pp. 159-165, 1958.