トピックモデルを用いた論文分析による企業と研究者のマッチングシステムの提案

大隈　有紗†　　清木　　康††

† 慶應義塾大学総合政策学部　〒 252-0805 神奈川県藤沢市遠藤 5322
†† 慶應義塾大学環境情報学部　〒 252-0805 神奈川県藤沢市遠藤 5322
E-mail: †{s15167ao,kiyoki}@keio.sfc.ac.jp

あらまし　本稿は，電子工学及び情報通信を専門とする研究者を対象とし，対象の研究者の研究内容と類似した研究・事業を行なっている企業を抽出，提示するシステムを提案する. このシステムの機能的特徴は，研究者の論文と，企業の技術速報等の文書を，サーベイ論文を用いて作成したトピックモデルの各トピック群への適合度をベクトル化し，相関量計算を行う所にある. 本研究は，研究者個人や企業が持つ専門性を，不変的な定形の「情報通信」や「電子工学」といった分野へ分類するだけではなく，個々の企業，人物が記述した論文を研究者個人の専門性の特徴として扱うことによって，研究者と企業の出会いの場を創出し，社会の学際的な研究のより一層の促進，研究所の幅広いキャリア取得，企業の専門人材獲得の機会損失の軽減を実現することを目的としている.

キーワード　トピックモデル，テキストデータ，相関量計算，ジョブマッチング

1　は じ め に

主に，世界中の人々がインターネットなどの情報通信技術の発展以降国と国の境界を超えての知的活動や，知識の共有が進み，データベースに保存された世界の論文量は増加傾向にある [1]. しかし，文部科学省による調査では，一つの論文を国の数で除して計算する方法である分数カウント [2]）による集計では，日本の論文数は 10 年前と比較して微減傾向 [1] にあり，日本の科学研究力の低下が問題となっている.

こうした日本の科学研究力の低下を引き起こしている要因の一つに，博士課程への進学者の減少 [3]，すなわち若手研究者の減少があり，減少の一因として，博士号取得後のキャリアパスの不透明さが挙げられる. 科学技術政策研究所が実施した，理系修士学生を対象としたアンケート [4] によると，博士課程進学ではなく就職することを選んだと答えた修士学生のうち，「博士課程に進学すると修了後の就職が心配である.」という質問に約 7.5 割が「そう思う」と回答している.

これらの事から，若手研究者を増やし日本の研究力を高めるためには，博士課程に進学するような，高度な専門性と知見を持つ人材（本稿ではこれを研究者と呼称する.）の就職支援が肝要である事がわかる.

本システムは，既存の枠組みである基本的な学問の分野，分類方式の内側にある細かな分類系統を，学会で発表された解説論文やサーベイ論文の文章群をから抽出されたトピック群として推定し，研究者により執筆された各の研究論文，企業により発表された技報や論文といった文書を生成する専門性を表す特徴量として利用して，研究者-企業間で相互に類似度計算を行う.

これにより，研究者自身の専門的知識をより活かせるような職場の選択ができる様に補助する事，また，副次効果として，企業がより自社にあった専門人材の確保できる様に補助する事を目的としている.

2　本システムで用いる手法

本システムでは，既存のある分野において，その分野を専門とする学会で発表されたサーベイ論文，解説論文を用いて，その研究分野全体の専門性の指向を更に細かく導出し，類似度計算に利用する. 本稿では，この専門性の志向を，文章に潜在する話題（＝潜在トピック）として捉え，潜在意味解析の手法であるトピックモデル [5] の中でも，階層ベイズモデルを用いている LDA（Latent Dirichlet Allocation, 潜在的ディリクレ配分法）[6] を使って導出する.

本節では，LDA における文書生成の考え方と，潜在トピックの導出法について確認する.

LDA において，文書とは，並び順を考慮しない単語の集まりとして考慮される. 単語数 $J(1\sim j)$ 個からなる文書は，LDA では以下の手順で生成される.

（1）　トピックの確率分布（トピック分布）が Dirichlet 分布（$Dir(-)$ と表記）から生成されると仮定.　　$\Theta \sim Dir(\alpha)$

（2）　J 個の単語それぞれのトピック n_j を，多項分布（$Multi(-)$ と表記）から生成されると仮定.　　$n_j \sim Multi(\Theta)$

（3）　トピック n_j に対する単語の確率分布に従い，単語 w_j を選択する.　　$p(w_j \mid n_j, \beta)$

β は，単語生成率に関するハイパーパラメータである．

ここで，文書数を i (1〜i)，各文書における単語数を j (1〜j)，全文書に含まれる単語数の合計が M（1〜m）である時，全対象文章群 $\mathbf{D}$ が生成される確率 $P(\mathbf{D}|\alpha,\beta)$ は以下の式で表される．

$$p(\mathbf{D}|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d\mid\alpha)\left(\prod_{j=1}^{J_d}p(n_{dj}\mid\theta_d)p(w_{dj}\mid n_{dj},\beta)\right)d\theta_d$$

α は潜在トピックの混合率に関するハイパーパラメータである．この時，一つ一つの単語（w）の集合である文書が与えられた時のトピックの選択確率と，各単語がどのトピックから生成されたかを示す事後確率分布は以下の式で表される．

$$p(\theta,n\mid\mathbf{w},\alpha,\beta)=\frac{p(p(\theta,n\mathbf{w}\mid\alpha,\beta))}{p(\mathbf{w}\mid\alpha,\beta)}$$

上式での $p(\mathbf{w}|\alpha,\beta)$ は，以下の式で表される．

$$p(\mathbf{w}|\alpha,\beta)=\int p(\theta\mid\alpha)\left(\prod_{j=1}^{J}\sum_{n_j}p(n_j\mid\theta)p(w_j\mid n_j,\beta)\right)d\theta$$

この確率分布 $p(\theta,n\mid\mathbf{w},\alpha,\beta)$ を計算で求めることが困難なため，様々な方法で分布を近似する手法が提案されている．本システムでは変分ベイズ法 [7] を用いて近似を行う．

この時，$p(\theta,n\mid\mathbf{w},\alpha,\beta)$ は次の様な分布で近似される，

$$p(\theta,n\mid\mathbf{w},\alpha,\beta)\simeq q(\theta,n\mid,\gamma,\Phi)=q(\theta\mid\gamma)\prod_{j=1}^{J}q(n_j\mid\Phi_j)$$

$q(\theta\mid\gamma)$ はパラメータ数=トピック数 N の Dirchlet 分布，$q(n\mid\Phi)$ はパラメータ数=単語数 M ×トピック数 N の多項分布である．

この $q(\theta,n\mid,\gamma,\Phi)$ は，既存の文書の生成確率に対する周辺対数尤度 $\log p(\mathbf{w}\mid\alpha,\beta)$ を最大化する α,β を求めるために用いる．この時，Jensen の不等式を用いることによって，下式の様に尤度の下限を定めることによって，解析的に α,β を求めることができる．

$$\log p(\mathbf{w}\mid\alpha,\beta)\geqq\int\sum_{n}q(\theta,n\mid\gamma,\Phi)\log q(\theta,n,\mathbf{w}\mid\alpha,\beta)d\theta-\int\sum_{n}q(\theta,n\mid\gamma,\Phi)\log q(\theta,n\mid\gamma,\Phi)d\theta$$

次節では，本節で述べた手法を用いて作成した本システムの具体的なアプローチについて述べる．

3　本システムのアプローチ

トピックモデルを用いて，類似性のある研究を発見，分析しようとする試みは，共同研究の可能性を見出す事を目的として [8], [9] などの研究で行われている．

本研究のシステムが以上の関連研究と最も異なる点は，LDA を用いて計算を行う際にトピック推定の対象とする文書群に，サーベイ論文や解説論文のみを用いて，類似度計算の対象とする企業文書や，論文は用いない所にある．

これは，対象とする研究分野での既存研究や技術を体系的に解説したサーベイ論文や解説論文が孕んでいる話題（＝潜在トピック）は，その分野での主要技術，活発に行われている研究内容等の意味を持っており，これらのトピックの含有率を用いた類似度計算は，対象とした研究内容の特徴を的確に示すことができると考えたためである．

本節では，本システムの具体的な計算処理の内容について述べる．本システムは，２つの計算工程からなる．

第 1 工程では，前節で述べた LDA を用いて，サーベイ論文，解説論文からなる文書群に含まれる単語群から，統計的に共起する可能性が高いものの集合，すなわち文書に潜在している N 個のトピック（Topic_1〜N）を推定する．

第 2 工程では，研究者 (Resarcher) が執筆した文章と，企業 (Company) が発表した文章が，N 個の各潜在トピック (Topic N) をどれ程に割合で含んでいるかという各トピック含有率 $\theta_N(\sum_{k=1}^{n}\theta_n=1)$ を，それぞれ，研究者の専門性特良ベクトル Resarcher_vec（$\vec{R}=\theta_r1,\theta_r2\cdots\theta_rn$），企業の専門性特徴ベクトル Company_vec（$\vec{C}=\theta_c1,\theta_c2\cdots\theta_cn$）として扱い，対象の研究者と企業の類似度をコサイン類似度を用いて計算する．

第 1 工程，第 2 工程の処理内容の詳細を以下に記述する．

第 1 工程は，図 1 の様に 3 つの処理から構成されている．各処理の内容を以下に記す．

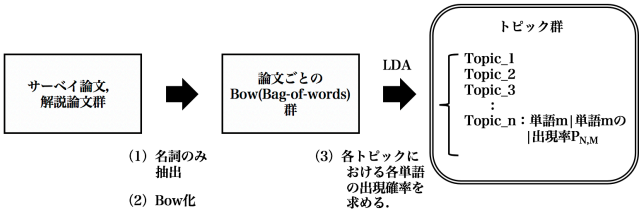

図 1 ：第 1 工程

（1） サーベイ論文，解説論文からなる文書群（以下解説論文群）に対して形態素解析ソフトである Mecab [10] を用いて，各論文それぞれに含まれる名詞のみを抽出する．

（2） 各論文ごとに，抽出された M 個の単語（名詞）の出現頻度を並べたベクトルである Bow（Bag-of-words）を求める．

（3） （2）で求めた BoW 群を用いて，N 個の各トピックにおける M 個の単語それぞれの出現確率 $P_{N,M}$（$P_{1,1}\cdots P_{N,M}$）

を求める．

（1）では，より多くの専門用語を抽出するために，形態素解析を行うために使う辞書に，専門用語（キーワード）自動抽出システム [11] をによって作成したユーザー辞書と，mecab-ipadic-NEologd [12] をシステム辞書として用い，URL や日付などのストップワードの除外を行った．

専門用語（キーワード）自動抽出システムは，与えられた文書を形態素解析し，その結果を複合語に組み立てて，重要度順に返すシステムである．このシステム用いて，解説論文群の各論文から抽出された複合語を，日付や大学名などの団体名などの専門用語と見なせないものを取り除き，その中から重要度の上位 10 位まで抽出して，形態素解析を行うための辞書データに，名詞として登録した．

また，各文書での出現率が 1 回である単語は，文書を生成する上でのトピック含有率決定への貢献度が低いと判断し，除外している．

（2）では，各論文ごとの各単語の出現率を BoW 群として求め，この BOW 群から，（3）において N 個の各トピックにおける M 個の単語それぞれの出現確率 $\{P_{1,1} \cdots P_{K,M}\}$ を，前節で述べた LDA を用いて求める．以上が本システム第 1 工程の内容である．

続いて，本システムの第 2 工程について解説する．第 2 工程は，図 2 の様に，2 つの計算処理から構成されている．各処理の内容を以下に記す．

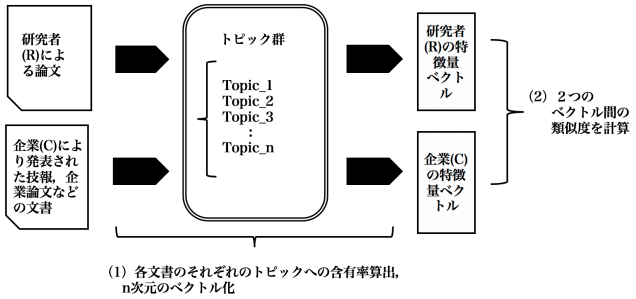

図 2：第 2 工程

（1） 研究者（R），企業（C）それぞれの発表文書（論文，技報）に対して，第 1 工程で求めたトピックの含有率 θ_{Rn}, θ_{Cn} を求める．

（2） （1）で求めたそれぞれの含有率 θ_{Rn}, θ_{Cn} を，専門性特徴量ベクトル Resarcher_vec（$\vec{R}$），Company_vec（$\vec{C}$）に変換し，ベクトル間のコサイン類似度を求める

（1）では，研究者によって書かれた論文と，企業によって発表された技報あるいは企業論文に内在する専門性の特徴量を，第 1 工程によって作成した潜在トピック群（1〜N）の含有率ベクトル $\{\theta_{R1} \cdots \theta_{Rn}\}$（$\sum_{k=1}^{n} \theta_{Rn} = 1$），$\{\theta_{C1} \cdots \theta_{Cn}\}$（$\sum_{k=1}^{n} \theta_{Cn} = 1$）として，それぞれ求める．

（2）では，（1）で求めた含有率 $\{\theta_{R1} \cdots \theta_{Rn}\}$ と $\{\theta_{C1} \cdots \theta_{Cn}\}$ を，それぞれ研究者の専門性特徴量ベクトル Resarcher_vec（$\vec{R} = \theta'_{R1}, \theta'_{R2} \cdots \theta'_{Rn}$），企業の専門性特徴量ベクトル Company_vec（$\vec{C} = \theta'_{C1}, \theta'_{C2} \cdots \theta'_{Cn}$）に変換する．この変換の際，導出された含有率ベクトルのうち，含有率 θ_n が 10^{-4} を下回る場合は -1 として扱う．これは，類似度計算を行う際，関係性が極めて低いトピックへの関与をより強く否定し，類似度の計算の精度を高めるための処理である．

最後に，導出した特徴量ベクトルを用いて，対象とする研究者と企業の類似度を下式の様に表されるコサイン類似度を用いて計算する．

$$cos(\vec{R}, \vec{C}) = \frac{\vec{R} \cdot \vec{C}}{|\vec{R}| \cdot |\vec{C}|} = \frac{\vec{R}}{|\vec{R}|} \cdot \frac{\vec{C}}{|\vec{C}|} = \frac{\sum_{i=1}^{|V|} R_i C_i}{\sqrt{\sum_{i=1}^{|V|} R_i^2} \cdot \sqrt{\sum_{i=1}^{|V|} C_i^2}}$$

以上が，本システム第 2 工程の内容である．

本稿では，本システムの実装は，Python [13] を用いて行なっている．その際，第 1 工程（1）で述べた専門用語（キーワード）自動抽出システム [11] の使用には，同システムの公式サイトで配布されている Python モジュール（termextract）を用いた．また，第 1 工程（2），（3）の処理には，genism [14] という，トピックモデル計算用の python ソフトウェアを用いた．

4 実　　験

今回は，電子工学及び情報通信の分野を対象として，本システムの精度を確かめる実験を行なった．本節での実験は，以下の手順で行われる．

（1） 今回の実験対象分野である電子工学及び情報通信の分野の解説論文群に，前節で第 1 工程として述べた処理を施し，N 個の潜在トピックの内容を推定するモデルを作成．

（2） 電子工学及び情報通信の分野の技術に関するキーワードを定める．

（3） （2）で選出したキーワード関する内容の，企業から発表された文書（企業文書）を，1 キーワードにつき 1 文書選択する．

（4） （3）で選択した文書と研究内容が近しい論文を，CiNII を用いて検索し，選択する．

（5） （3），（4）で選択した企業論文，研究者論文に対して，前節第 2 工程で述べた処理を施し，類似度計算を行う．

まず始めに，手順（1）を行う．本節での実験では，トピック抽出のための解説論文群に，解説論文群には EIC（電子情報通信学会）情報・システムソサイエティの『電子情報通信学会論文誌「情報・システム：D」(和文論文誌 D)』と，電気学会の論文誌『電気学会論文誌. C（電子・情報・システム部門誌)』で発表されている，サーベイ論文，解説論文そして，技術解説のために書かれた招待論文から，無料公開されているものを合計 340 文書を用いて，総数 $N = 8$ の潜在トピックのモデルを作成した．

算出されたトピックごとに，出現頻度の高い上位 20 単語を，表 1 にまとめる．

Topic_1		Topic_2	
0.009	DSS	0.009	特徴量
0.007	周波数	0.006	ペトリネット
0.005	並列処理	0.005	識別器
0.005	ASIC	0.004	モデル化
0.005	DSP	0.003	音声認識
0.004	GaN	0.003	状態空間表現
0.003	DNA	0.003	HMM
0.003	ループ	0.002	深層学習
0.003	時系列	0.002	RTL
0.003	学習オートマトン	0.002	ブランチ
0.003	コンパイラ	0.002	検出法
0.003	GPS	0.002	カルマンフィルタ
0.002	並列性	0.002	XML
0.002	染色体	0.002	モータ
0.002	タンパク質	0.002	イミュニティ
0.002	オートマトン	0.002	サブシステム
0.002	CPS	0.002	周波数
0.002	通信網	0.002	DNN
0.002	スケジューリング	0.002	エージェント
0.002	ゲートアレイ	0.002	オブジェクト

Topic_3		Topic_4	
0.006	モジュール	0.006	ボルツマンマシン
0.006	畳み込み	0.005	観測データ
0.004	スケジューリング	0.005	RBM
0.004	制御器	0.005	深層学習
0.004	SAM	0.005	学習者
0.003	アーキテクチャ	0.004	アレイ
0.003	アート	0.004	トラヒック
0.003	半導体	0.004	サンプリング
0.003	ファジィ制御	0.004	偏波変動
0.003	チャネル数	0.003	CNN
0.003	LPV	0.003	CMOS
0.002	ドナー	0.003	コマンド
0.002	MOSFET	0.003	ユニット
0.002	アプリケーション	0.003	ナノ粒子
0.002	レイヤ	0.003	プーリング
0.002	不等式	0.003	半導体レーザ
0.002	ストーン	0.002	教師なし学習
0.002	CNN	0.002	OPGW
0.002	ゲイン	0.002	GaN
0.002	カーリング	0.002	微細化

表 1：各トピックごとの出現頻度上位 10 単語（1/2）

Topic_5		Topic_6	
0.005	エキスパートシステム	0.006	計算モデル
0.005	機械翻訳	0.006	ICタグ
0.004	GaN	0.005	電磁波
0.004	双極子	0.004	LOM
0.003	排出量	0.004	相互作用
0.003	自然言語処理	0.004	超小型
0.002	レート	0.004	制御系
0.002	深層学習	0.003	SAW
0.002	航空機	0.003	トップダウン
0.002	WDM	0.003	マップ
0.002	TEM	0.003	ゲイン
0.002	通信路	0.003	PID
0.002	半導体	0.003	MAS
0.002	キャリア	0.003	力覚情報
0.002	イオン	0.003	FRIT
0.002	最適化問題	0.003	周波数
0.002	リモートセンシング	0.002	ボトムアップ
0.002	特徴量	0.002	MPC
0.002	エラー	0.002	電力システム
0.002	シミュレータ	0.002	プラント

Topic_7		Topic8	
0.006	ハードウェア	0.015	歩行者
0.005	組込みシステム	0.01	PID
0.004	マシンビジョン	0.009	CNN
0.004	PLC	0.007	周波数
0.004	YAG	0.006	特徴量
0.004	CALS	0.005	データセット
0.003	システムLSI	0.005	GDSS
0.003	ファイバ	0.004	ニューラルネットワーク
0.003	STC	0.003	インタフェース
0.003	パルス	0.003	エレベータ
0.002	レーザー	0.003	ニューロン
0.002	センシング	0.003	コンテンツ
0.002	レーザアブレーション	0.003	動画像
0.002	ビジョン	0.003	制御技術
0.002	EDA	0.002	内部表現
0.002	補償器	0.002	検出法
0.002	マイクロ	0.002	ユニット
0.002	目視検査	0.002	パーソナル化
0.002	構造物	0.002	FRIT
0.002	消費電力	0.002	テスト

表 1：各トピックごとの出現頻度上位 10 単語（2/2）
（票左列：出現確率　表右列：該当トピックで，
左列同行の数値の出現確率を持つ単語）

表 1 において，各単語の左横の数字がトピック内でのその単語の出現確率を表している．

これらのトピック間の関係を，主成分分析 [15] の手法を用いて，二次元に落とし込み可視化した結果が図 3 である．この可視化には，python のライブラリである pyLDAvis [16] を用いた．

図 3 において，より近い距離にあるトピック同士は，同じ文書内に占める含有率が高い，つまり，共起性が高い事を示している．

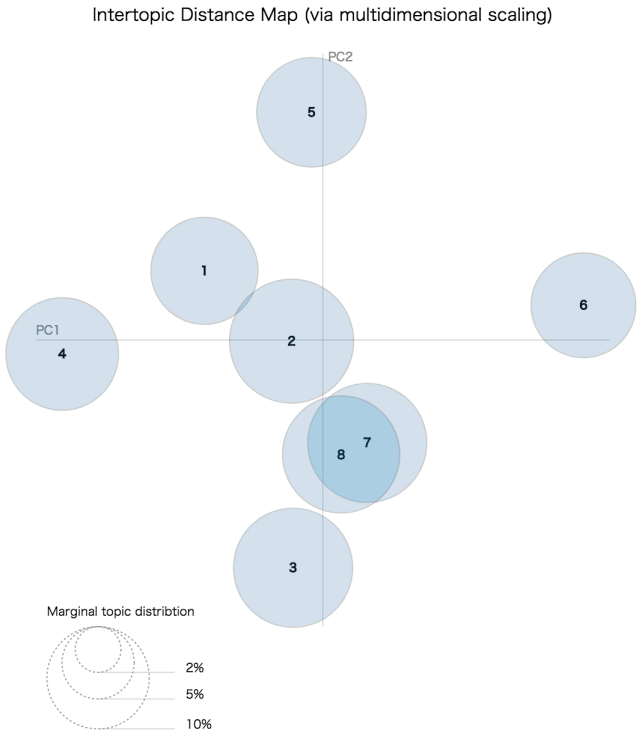

図 3：トピック 1〜N 間の距離関係

続いて，手順（2），（3）を行う．

本節での実験では，本システムの研究者と企業間の類似度の算出精度を検証するために，研究内容的に近しいと判断できる文書を用いて実験を行なった．電子工学及び情報通信の分野から，今回選んだキーワードは「IoT」，「AI」，「セキュリティ」の 3 つである．そして，先ほど選んどキーワードに関連する企業文書を，企業による発表文書から 1 文書のみ選択し，実験対象とした．

今回実験対象とした会社は，IT 系企業から技術報告，企業論文等を無料公開している企業を，合計 3 社を選択した．それぞれ，A 社，B 社，C 社とする．選択文書は，それぞれの会社から 1 文書ずつ，A 社より IoT に関する文書を企業文書 A，B 社より AI に関する文書を企業文書 B，C 社よりセキュリティに関する文書を企業文書 C として選択した．

表 2 は，選択した文書内で出現する単語の出現回数をカウントし，その内出現回数の上位 20 単語を，多い順で並べたものである．

企業文書A

単語数	単語
71	アイテム
24	棚
18	ポイント
17	重量
15	識別
14	競技
14	未知
14	商品
12	ディープラーニング
11	収納
10	ボックス
10	バウンディ
9	特徴量
9	既知
8	マッチング
8	StowTask
8	SVM
8	45分
7	認識システム
7	ラウンド

企業文書B

単語数	単語
43	顧客
38	AI
28	回答
28	コンタクトセンター
25	接点
24	問い合わせ
22	CHORDSHIP
19	FAQ
17	対話
17	チャットボット
15	高度化
13	オペレーター
12	マーケティング
11	ハイブリッド
10	機械学習
10	企業
10	チャット
9	体験
8	蓄積
8	ソリューション

企業文書C

単語数	単語
34	セキュリティー
33	制御システム
16	ホワイトリスト
11	制御機器
9	監視
7	アプリケーション
6	長期間
6	運用開始
6	情報システム
6	安全
6	制御プログラム
6	ハッシュ値
6	OODAループ
5	forKernel
5	Linux
4	負荷
4	接続
4	悪意
4	制御技術
4	再起動

表 2：企業文書 A，B，C における出現回数上位 20 単語
（票左列：出現回数　表右列：該当文書で，
左列同行の数値の出現回数を持つ単語）

研究者文書A

単語数	単語
27	適応制御
27	自律
24	エリア
16	人員
13	倉庫
12	最適
12	コンテナ
11	出庫
10	現場
10	注文
9	移動
9	完了
7	物流
7	リアルタイム
7	システム構成
6	算出
6	指示
6	作業人員
6	人員配置
6	フロー

研究者文書B

単語数	単語
64	人狼
25	対戦
21	処刑
20	プレイヤ
19	表出
17	知能
16	議論
16	役職
15	エージェント
14	擬人化
13	初日
12	人工知能
11	協力
10	1人
9	村人
9	占い師
9	ゲーム
9	AI
8	印象
7	調査

研究者文書C

単語数	単語
34	セキュリティー
33	制御システム
16	ホワイトリスト
11	制御機器
9	監視
7	アプリケーション
6	長期間
6	運用開始
6	情報システム
6	安全
6	制御プログラム
6	ハッシュ値
6	OODAループ
5	forKernel
5	Linux
4	負荷
4	接続
4	悪意
4	制御技術
4	再起動

表 3：研究者文書 A，B，C における出現回数上位 20 単語
（票左列：出現回数　表右列：該当文書で，
左列同行の数値の出現回数を持つ単語）

次に，手順（4）を行う．これらの企業文書 A，B，C と類似した研究内容の論文を，企業文書 1 つにつき 1 つ選出した．この選出作業には，国立情報学研究所が運営している，学術論文や図書・雑誌等の学術情報データベースである CiNII [17] を用いた．CiNII [17] での検索結果から，3 つの論文をそれぞれ研究者文書 A，B，C として選出した．

企業文書と研究者文書は，それぞれ末尾のアルファベット（A，B，C）が同じ組み合わせが，研究内容が近しい文書同士の組み合わせである．また，表 3 は，選択した研究者文書内で出現する単語の出現回数をカウントし，その内出現回数の上位 20 単語を，多い順で並べたものである．

最後に，手順（5）を行う．まず，前節第 2 工程（1）で述べた処理である，それぞれの文書における，各トピック（1〜N，今回は $N = 8$）の含有率の算出を行う．算出の結果，含有率の数値が 10^{-4} を下回る場合は，-1 に置き換える．表 4 は，企業文書 A，B，C における各トピックの含有率，表 5 は，研究者文書 A，B，C における各トピックの含有率である．表 4，表 5 共に，含有率の数値が 10^{-4} 以下のものは -1 への置き換え処理を行なってある．

トピック ＼ 企業文書	企業文書A	企業文書B	企業文書C
Topic_1	-1	0.03035484	0.26743364
Topic_2	0.33545041	0.20605122	0.20274591
Topic_3	0.03120301	0.11132185	0.24311352
Topic_4	0.46877396	0.40180436	-1
Topic_5	-1	0.03688988	-1
Topic_6	0.03916926	-1	-1
Topic_7	0.07091584	0.05957781	0.16612029
Topic_8	0.05359893	0.15334064	0.11811555

表 4：企業文書 A，B，C における各トピック含有率

トピック ＼ 研究者文書	研究者文書A	研究者文書B	研究者文書C
Topic_1	0.06389018	0.00133045	0.01066206
Topic_2	0.06411263	0.27168465	0.25885421
Topic_3	0.20217288	0.56346774	-1
Topic_4	0.257447	0.00133175	-1
Topic_5	-1	0.06580558	0.06481946
Topic_6	0.04569629	0.00133172	0.08279105
Topic_7	0.36492994	0.01440274	0.12534484
Topic_8	-1	0.08064536	0.45618179

表 5：研究者文書 A，B，C における各トピック含有率

手順（5）の最終工程として，表 4，表 5，で算出された含有率を，それぞれの文書ごとに，特徴量ベクトルに置き換えて，コサイン類似度による類似度計算を行う．企業文書 A，B，C は，それぞれ企業特徴量ベクトル Company_A，Company_B，Company_C へ，研究者文書は，それぞれ研究者特徴量ベクトル Resarcher_A，Resarcher_B，Resarcher_C へと変換を行なった．この変換の後，企業-研究者間の特徴量ベクトルの類似度の計算を行った結果を表 6 にまとめる．

研究者特徴量ベクトル ＼ 企業特徴量ベクトル	Company_A	Company_B	Company_C
Resarcher_A	0.4610602	-0.0436501	0.2672162
Resarcher_B	0.0490311	0.1887184	0.1194975
Resarcher_C	-0.1946576	-0.2730946	0.2718335

表 6：類似度計算の結果

5 考　　察

前節表 6 より，今回の実験では，名前の末尾が同じアルファベットの特徴量ベクトル（Company_A ならば Resarcher_A）が，企業軸（列），研究者軸（行）の双方で最も高い類似度を示した．このため，実験の目的であった「類似度計算の精度の検証」においては，近しい研究内容の企業，研究者に対する類似度を最も高く算出できたという点で，一定の信頼性を持てる様な精度の計算ができるという検証結果が出た．

また，表 6 の第 3 列の Company_C の軸での Resarcher_A にに対する類似度は，末尾が同じアルファベットのもの（Resarcher_C）に対して，大幅な差が見られない様な数値が出ている事がわかる．

この結果より，第 2 節で述べた LDA の文書生成の仕組みを考慮すると，本システムにおいては，Company_C は，Resarcher_A とも似た様なトピックを背景に持っている，すなわち，近しい技術を用いた研究を行っている部分があると判断されていることが分かる．

6 総論及び，今後の展望

本稿では，電子工学及び情報工学における，トピックモデルを用いた論文分析による企業と研究者のマッチングシステムを提案した．

実験の結果により，本システムは一定の信頼性のおける精度の計算を行えることが分かったため，論文等の文書を用いたジョブマッチングシステムの新たな可能性を見出すことができた．

今後の課題としては，使用文書の事前処理が挙げられる．本稿において実験に使用した論文群は，全て元 PDF ファイルからテキストデータへの抽出を行なってから運用しているが，それぞれのフォーマットの違いによっては，長い単語が途中で切れて，意味のない文字列として検出されてしまう事があった．そのため，そういった意味を持たない文字列はストップワードを除外する段階で一緒に削除せざるを得ず，結果として正確なモデル作成が困難になっていた．

使用文書の事前処理は，モデルの精度，類似度計算の精度を高める上で非常に肝要に部分であるので，今後は，専門用語の抽出をより正確に行えるような事前処理を行える様にする事で，本システムの精度をより洗練させる必要がある．

また，今回は電子情報通信学会，電気学会で発表されたサーベイ論文，解説論文類のみをシステム構築に用いたため，今後は電子工学及び情報工学の分野の他の学会で発表された論文も，システムに組み込んでいくことで，より，電子工学及び情報工学の分野の専門性の抽出精度を高めていきたい．

文　　献

[1] 文部科学省　科学技術政策研究所　『日本の科学研究力の現状と課題』(NISTEP ブックレット;001 ver. 5)2018.12 http://hdl.handle.net/11035/2456

[2] 国立研究開発法人　科学技術振興機構　『世界の論文 — 科学技術情報プラットフォーム』2019.1.2 https://jipsti.jst.go.jp/foresight/dataranking/treatises/count/

[3] 文部科学省　『文部科学統計要覧』（平成 30 年度版）2018.4

[4] 科学技術政策研究所　第 1 調査研究グループ　『日本の理工系修士学生の進路決定に関する意識調査』(p.22　図 7)2009.3 http://hdl.handle.net/11035/895

[5] David M. Blei “Probabilistic Topic Models” April 2012, Communications of the ACM, Vol. 55 No. 4, Pages 77-84

[6] David M. Blei, Andrew Y. Ng and Michael I. Jordan “Latent Dirichlet Allocation” 2003

[7] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation”, in Journal of Machine Learning Research, 2003, p.1107-1135

[8] 荒木 将貴，桂井麻里衣，大向 一輝，武田 英明『研究成果データベースを用いた異分野の共同研究者の推薦』2016

[9] 小野 龍太郎，冨浦 洋一，田中 省作，上瀧 恵里子『オーサートピックモデルを用いた論文分析による潜在的研究グループの発掘に関する研究』2014.3

[10] MeCab 0.996 http://taku910.github.io/mecab/

[11] 専門用語（キーワード）自動抽出システム

[12] mecab-ipadic-NEologd https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md

[13] Python 3.5.0 https://www.python.org/

[14] genism https://radimrehurek.com/gensim/

[15] Karl Pearso, “On lines and planes of closest fit to systems of points in space. ”, Philosophical Magazine 2, 1901 ,p.559-572.

[16] pyLDAvis

https://github.com/bmabey/pyLDAvis
[17] CiNii
https://ci.nii.ac.jp/
[18] 奥村 学『トピックモデルによる統計的潜在意味解析』初版第 2 刷，株式会社コロナ社，2015