川崎医療福祉学会誌　Vol. 29　No. 1　2019　127 − 137

原　著

トピックモデルによるDPCデータの分析と病院機能についての考察

田　中　昌　昭*1

要　　約

医療機能の分化・連携を推進するには，客観的なデータに基づいて病院機能を把握し，それを医療政策の意思決定に活かす必要がある．そこで，本研究では自然言語処理分野で培われた手法を用いて公表DPC（Diagnosis Procedure Combination）データから病院機能を分析する手法を考案し，既存の病院機能分類との関係性を調べた．文書を病院，単語を疾患，そして単語の出現頻度を疾患の出現頻度に対応させて病院に隠されたトピックを推定するトピックモデルを構築した．得られたトピックを基本的な医療機能とみなして，その含有率パターンで病院機能を定義した．各病院の基本的な医療機能（トピック）を特徴量として階層クラスター分析を行った結果，クラスターと既存の病院機能の間に対応関係が認められた．特に，特定機能病院を高精度に推定することができた（適合率72/79=0.91，再現率72/84=0.86）．考案した手法は，病院機能を複数の基本的な医療機能（トピック）の組み合わせとして捉える点に特徴があり，それによって病院の機能分化の現状を定量的に測定できる可能性が示唆された．しかしながら，病院機能は取り扱う疾患の数だけで定義できるものではない．地理的な条件に加え，病院が保有する人的･物的資源なども考慮して，より精度を上げる必要がある．

1．緒言

近年の少子高齢化による人口構造の変化，それに伴う疾病構造の変化によって医療提供体制に抜本的な改革が求められている．たとえば2014年度から導入が始まった病床機能報告制度[1)]はそのための施策の一つで，医療機能の分化や連携を促進し，地域に必要な医療機能の明確化と強化を目的としている．

一口に病院機能と言っても，地域医療構想[2)]における高度急性期機能・急性期機能・回復期機能・慢性期機能といった患者のステージに着目した機能，病床数・診療科数・医療機器の台数・職員数といった物的・人的資源に着目した機能，DPC病院・特定機能病院･地域医療支援病院･在宅療養支援病院･在宅療養後方支援病院・三次救急医療施設といった医療制度に基づく機能など様々である．しかし，病院をこれらの機能に分類する目的は，限られた医療資源を地域で最適化し，住民に必要な医療を提供するとともに無駄を省き医療費を抑えることにある．そのため，病院機能を客観的に高精度で評価することが重要となる．

あらためて病院機能とは何かと考えた場合，それを客観的に評価する指標は必ずしも確立されているとは言えない．そこで，本研究では医療に関する公表データに自然言語処理分野で培われた技術を応用することにより，病院の機能を分類する新しい手法を提案し，その利用可能性を確かめることを目的とする．

2．トピックモデル

トピックモデル[3)]は，文書の分類や検索を目的として自然言語処理分野で開発された言語モデルである．トピックモデルでは，文書には隠れたトピックがあって，単語はあらかじめトピックごとに決められた確率で出現すると考える．例えば政治についての論説には「議会」や「内閣」といった単語がスポーツ記事に比べて高い確率で出現するであろう．この「政治」や「スポーツ」がトピックである．しかし，「政治とスポーツ」といった複数のトピックを持つ

*1 川崎医療福祉大学　医療福祉マネジメント学部　医療情報学科
（連絡先）田中昌昭　〒701-0193　倉敷市松島288　川崎医療福祉大学
E-mail : mtanaka@mw.kawasaki-m.ac.jp

文書も存在する．そこで，文書はいくつかのトピックからなり，単語はそれらのトピックに応じて決まるある確率で発生し，その結果として複数のトピックが混ざり合った文書が作られると考える．このようなトピックモデルを Latent Dirichlet Allocation (LDA)[4]と呼び，自然言語処理のみならず，問診データの解析[5]，最も医療資源を投入した疾患の推定[6]，入院患者に発行するオーダパターンの予測[7]，ICU 退院後の死亡率（post-discharge ICU mortality）の正確な予測[8]，医薬品の副作用の予測[9]など，様々な課題に適用されている．

LDA をグラフィカルモデルで表したのが図1である．図で$\boldsymbol{\theta_d} = (\theta_{d,1}, \cdots, \theta_{d,K})$は，文書$d$がトピック$k \in \{1, \cdots, K\}$を含む確率$\theta_{d,k}$を要素とする確率ベクトル（文書のトピック構成比率）である．ここで，Kはトピック数で，定義より$\sum_{k=1}^{K} \theta_{d,k} = 1$である．$\boldsymbol{\theta_d}$を含む矩形の右下にある$D$は文書数で，矩形は$d = 1 \sim D$の繰り返しを表す．また，$\boldsymbol{\phi_k} = (\phi_{k,1}, \cdots, \phi_{k,|V|})$は，トピック$k$から単語$v$が生成される確率$\phi_{k,v}$を要素とする確率ベクトル（単語の出現分布）である．ここで，Vは語彙集合で$|V|$は語彙数である．定義より$\sum_{v \in V} \phi_{k,v} = 1$である．$z_{d,i}$は，文書$d$の$i$番目の単語のトピックで，$w_{d,i}$はそのトピックから確率$\phi_{z_{d,i},w_{d,i}}$で発生した単語である．$N_d$は文書$d$に含まれる単語数である．$z_{d,i}$および$w_{d,i}$は離散値なので，それぞれ$\boldsymbol{\theta}_d$および$\boldsymbol{\phi}_k$をパラメタとする多項分布から生成されるものとする．さらに$\boldsymbol{\theta}_d$および$\boldsymbol{\phi}_k$は確率ベクトルなので，それぞれ$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$および$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{|V|})$をパラメタとするディリクレ分布から生成されると仮定する．図で網掛けをしたディリクレ分布のパラメタ$\boldsymbol{\alpha}, \boldsymbol{\beta}$と観測されたデータ$\boldsymbol{w}$から確率分布$\boldsymbol{\theta}_d$と$\boldsymbol{\phi}_k$を求める．その方法として変分ベイズ法やギブスサンプリングがあるが，本研究ではギブスサンプリングを用いて$\boldsymbol{\theta}_d$と$\boldsymbol{\phi}_k$を推測した．

以上がトピックモデルの概要であるが，このモデルにおいて文書を病院，単語を疾患，そして単語の出現頻度をその疾患の出現頻度に置き換え，トピックを基本的な医療機能に見立てて病院が担っている機能を分析しようというのが本研究のアイデアである．これは，取り扱う疾患によって病院の機能を捉えるのは理にかなっていると考えられるからである．

3．材料と分析方法

3．1　材料

本研究では病院の診療実績を厚生労働省が公開している平成28年度の DPC データ[10]から入手した．DPC とは診断群分類（Diagnostic Procedure Combination）のことで，診断と処置（手術，検査等）を組み合わせたものである．これが医療費の支払い制度と結び付いて DPC 制度が誕生した．DPC 制度に参加する病院は毎年行われる DPC 調査で患者の診療データを提出し，それを集計したものが公開されている．DPC データの公開サイトから「施設概要表」と参考資料2の「(8) 疾患別手術別集計 MDC01～ MDC18」をダウンロードして，病院ごと

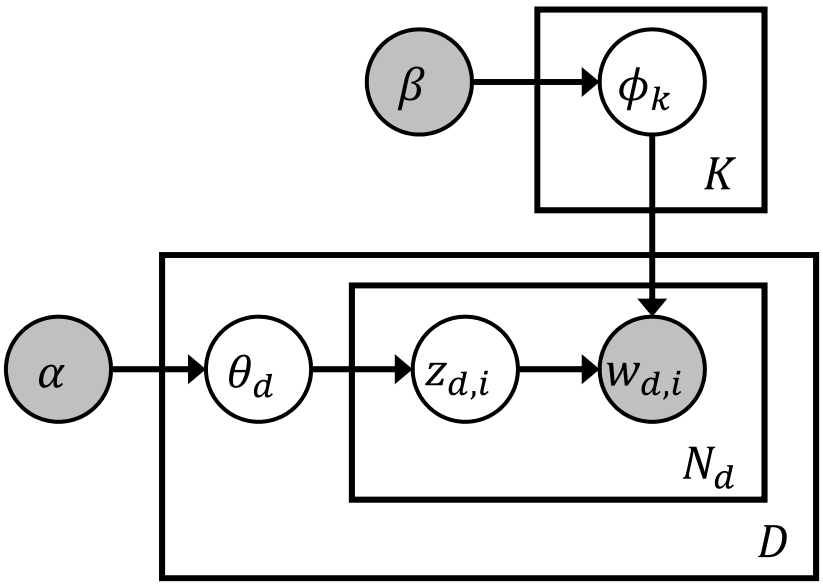

図1　LDA のグラフィカルモデル

に疾患の出現頻度を抽出した．なお，ここで疾患の出現頻度とは DPC コードの先頭6桁で表される疾患コードと DPC コードの9～10桁目の手術を組み合わせた8桁のコードで識別される「疾患・手術」の件数であるとした．平成28年度の DPC データには DPC 対象病院が1,666施設，DPC 準備病院が276施設，そして出来高算定病院施設が1,559施設あるが，今回は DPC 対象病院1,666施設を分析の対象とした．

また，本研究の提案手法で得られた病院機能と既存の機能分類の関連性を調べるために特定機能病院，地域医療支援病院，DPC 群の種類，総合入院体制加算，在宅療養支援病院，在宅療養後方支援病院，三次救急医療施設などの病院属性を記録した平成28年度病床機能報告公表データを厚生労働省の病床機能報告サイト[1)]からダウンロードして DPC データと紐づけた．

３．２　分析方法

まず，行を病院，列を疾患，要素をその件数とする病院疾患行列を作成して R の topicmodels パッケージ[11)]の LDA 関数を使って LDA モデルを構築した．その際，トピック数は40とし，前述したようにモデルの推定にはギブスサンプリングを用いた．LDA 関数のパラメタにはトピック数：k=40以外に稼働検査期間：burnin=500，サンプリング回数：iter=1000を与え，それ以外はデフォルト値を使った．

得られた LDA モデルからトピックkにおける疾患の出現分布$\boldsymbol{\phi_k}$を取り出して，次式で与えられる確率を要素とする主要診断群（MDC, Major Diagnostic Category）の出現分布$\hat{\boldsymbol{\phi}}_{\boldsymbol{k}}$を計算した．

$$\hat{\phi}_{k,j} = \sum_{v \in \mathrm{MDC}_j} \phi_{k,v}$$

ここで，vは疾患を表し，$j \in \{1,\cdots,18\}$は MDC コード[12)]である．これによってトピックを MDC の構成比率で表すことができる．これは，病院機能を疾患の分布で解釈するよりも，粒度の粗い MDC の分布で阻視化した方が直感的に理解しやすいと考えたからである．

次に，LDA モデルから病院$d \in \{1,\cdots,D\}$（Dは病院数）のトピック分布$\boldsymbol{\theta_d}$を取り出して，それをデータポイントに用いてウォード法による階層クラスター分析を行った．得られた樹形図（デンドログラム）をクラスター数が20になるようカットオフして各々の病院に1から20までのクラスター番号を割り当てた．また，各病院に対して次式で与えられるハーフィンダール・ハーシュマン・インデックス（HHI, Herfindahl-Hirschman Index）[13)]を計算した．

$$H_d = \sum_{k=1}^{K} \theta_{d,k}^2$$

HHI は市場の寡占を表す指標で，独占状態では1，均等なシェア（すべてのkに対して$\theta_{d,k}$=$1/K$）では$1/K$という値をとる．したがってH_dは病院dにおける機能（トピック）の寡占を表す指標と解釈でき，その逆数$F_d = 1/H_d$はその病院に含まれるおおよその基本的な医療機能の数（以後，これを機能指数と呼ぶ）とみなすことができる．

３．３　評価方法

トピックモデルは教師なし学習であるため，得られた結果の評価が難しい．幸い，本研究で扱うデータには教師情報が含まれている．まず，公表 DPC データの「施設概要表」には「DPC 算定病床数」，「DPC 算定病床の入院基本料」などの項目がある．前者からは病院の規模，後者からは看護配置がわかる．そして，病床機能報告公表データには特定機能病院かどうか，地域医療支援病院かどうか[†1)]，DPC 群の種類（Ⅰ群，Ⅱ群，Ⅲ群）[†2)]，総合入院体制加算（加算1，加算2，加算3），在宅療養支援病院かどうか，在宅療養後方支援病院かどうか，そして三次救急医療施設かどうかなどの情報がある．これらの機能分類は，病院がその地域に対して実際に果たしている役割というよりも病院に対して期待される役割あるいはその病院が自ら主張する役割という意味合いが強いが，それらの間には何らかの関連があることは否定できない．というよりも，関連がなければならない．そこで，これらの情報を提案手法で得られた病院機能の検証情報として利用する．具体的には，DPC 算定病床数と機能指数F_dの関係，クラスター別の特定機能病院数，地域医療支援病院数，DPC 群の各群（Ⅰ，Ⅱ，Ⅲ）の病院数，総合入院体制加算の各加算（加算1，加算2，加算3）の病院数，在宅療養支援病院数，在宅療養後方支援病院数，そして三次救急医療施設数を求める．こうして何らかの関係性が示されれば提案手法は病院機能の一側面を捉えているとみなすことができる．

４．結果

４．１　分析対象データの抽出結果

1,666施設の DPC 対象病院のうち，すべての疾患の出現頻度が10件未満の病院が1施設あったので分析対象から除外した．また，年度内に合併した病院が2施設あったので合併後のデータにマージした．その結果，分析対象として1,664施設の病院データが得られた（D =1664）．また，出現頻度が10件以

表1 病院機能の MDC（主要診断群）分布

	MDC01	MDC02	MDC03	MDC04	MDC05	MDC06	MDC07	MDC08	MDC09	MDC10	MDC11	MDC12	MDC13	MDC14	MDC15	MDC16	MDC18
TOPIC01	15.9%	0.0%	0.2%	3.9%	0.5%	1.3%	28.8%	1.9%	0.0%	45.6%	0.3%	0.0%	0.2%	0.0%	0.0%	0.1%	1.4%
TOPIC02	0.4%	0.0%	0.2%	1.6%	0.0%	7.4%	0.2%	0.1%	0.6%	1.3%	2.9%	84.5%	0.5%	0.1%	0.0%	0.0%	0.2%
TOPIC03	0.0%	0.1%	0.0%	2.7%	1.6%	1.8%	1.0%	0.0%	0.0%	4.2%	68.9%	0.0%	0.0%	0.4%	0.0%	0.0%	19.2%
TOPIC04	0.1%	0.0%	0.0%	0.0%	98.7%	0.4%	0.3%	0.0%	0.0%	0.2%	0.3%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%
TOPIC05	0.1%	0.0%	0.0%	0.4%	0.0%	4.8%	0.2%	0.5%	0.0%	0.0%	91.9%	1.6%	0.1%	0.0%	0.0%	0.0%	0.3%
TOPIC06	0.0%	0.1%	0.0%	0.0%	0.0%	3.3%	0.9%	0.0%	93.6%	0.5%	0.0%	1.3%	0.0%	0.0%	0.0%	0.0%	0.1%
TOPIC07	0.0%	0.0%	0.0%	5.8%	0.7%	87.9%	0.0%	0.1%	0.0%	2.0%	2.1%	0.0%	0.7%	0.0%	0.0%	0.0%	0.5%
TOPIC08	3.0%	0.0%	1.8%	0.9%	0.0%	3.4%	0.1%	0.5%	0.0%	0.8%	2.6%	21.3%	0.0%	67.9%	0.0%	0.5%	0.2%
TOPIC09	90.3%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	9.1%	0.0%
TOPIC10	0.3%	41.7%	0.1%	2.5%	0.8%	8.9%	33.7%	1.0%	0.0%	3.2%	0.9%	1.0%	0.2%	0.0%	0.0%	5.4%	0.3%
TOPIC11	0.0%	0.1%	0.4%	0.0%	0.0%	1.2%	0.1%	0.0%	0.0%	0.4%	0.3%	94.9%	0.1%	2.2%	0.0%	0.0%	0.0%
TOPIC12	0.1%	80.7%	1.7%	0.1%	0.1%	1.7%	0.4%	0.5%	0.0%	9.9%	2.2%	0.1%	0.0%	0.0%	0.0%	0.5%	2.1%
TOPIC13	0.0%	0.0%	0.0%	0.7%	0.0%	99.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TOPIC14	0.0%	0.9%	0.0%	0.1%	0.2%	96.4%	0.3%	0.8%	0.0%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.9%
TOPIC15	17.7%	0.0%	0.0%	10.5%	9.9%	26.4%	1.5%	1.5%	2.2%	2.2%	9.7%	1.7%	0.8%	0.0%	0.0%	14.6%	1.4%
TOPIC16	6.8%	6.1%	5.4%	8.4%	3.9%	22.1%	11.1%	4.3%	0.7%	6.8%	10.3%	4.1%	3.9%	2.4%	0.0%	1.6%	2.2%
TOPIC17	2.3%	0.0%	7.3%	63.4%	7.5%	4.3%	0.5%	2.6%	0.0%	2.8%	9.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TOPIC18	3.4%	0.0%	4.6%	42.6%	10.3%	16.9%	1.6%	2.3%	0.0%	2.6%	11.2%	0.0%	0.7%	0.0%	0.0%	3.7%	0.0%
TOPIC19	0.0%	98.6%	0.0%	0.0%	0.0%	0.3%	0.0%	0.4%	0.0%	0.5%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%
TOPIC20	0.0%	0.0%	0.0%	0.0%	0.2%	98.9%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.3%	0.0%
TOPIC21	0.0%	0.0%	68.1%	18.1%	5.1%	1.9%	0.1%	0.9%	0.0%	2.5%	5.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TOPIC22	2.1%	0.0%	2.2%	6.7%	5.4%	13.4%	0.3%	1.3%	0.0%	1.5%	61.7%	0.1%	0.1%	0.0%	0.0%	4.9%	0.3%
TOPIC23	0.0%	0.0%	0.2%	3.0%	0.0%	7.7%	1.6%	0.2%	0.0%	1.1%	2.8%	0.4%	82.3%	0.0%	0.0%	0.0%	0.4%
TOPIC24	70.6%	0.0%	8.1%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	20.3%	0.0%
TOPIC25	0.0%	0.0%	0.0%	13.8%	70.5%	2.9%	5.5%	2.1%	0.0%	1.1%	2.4%	0.0%	0.1%	0.0%	0.0%	0.6%	0.8%
TOPIC26	0.0%	0.0%	0.0%	1.2%	96.6%	0.8%	0.0%	0.0%	0.0%	0.5%	0.0%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%
TOPIC27	0.1%	0.0%	0.5%	0.0%	0.0%	0.2%	22.0%	0.5%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	76.5%	0.0%
TOPIC28	8.9%	5.9%	2.5%	5.7%	1.0%	11.3%	3.1%	3.0%	0.0%	3.0%	5.7%	1.6%	2.9%	40.2%	1.2%	2.0%	2.1%
TOPIC29	1.6%	0.2%	9.7%	16.0%	0.0%	42.0%	4.2%	0.8%	6.1%	1.4%	10.0%	3.5%	2.8%	0.0%	0.0%	0.0%	1.6%
TOPIC30	82.7%	0.0%	2.6%	3.2%	0.0%	1.6%	4.9%	0.1%	0.0%	1.2%	1.3%	0.1%	0.0%	0.0%	0.0%	1.4%	1.0%
TOPIC31	0.0%	0.0%	0.0%	0.0%	99.4%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%
TOPIC32	0.0%	0.0%	0.5%	0.0%	0.0%	15.8%	0.4%	0.5%	0.1%	0.0%	18.7%	63.7%	0.3%	0.0%	0.0%	0.0%	0.0%
TOPIC33	0.0%	0.0%	0.0%	0.7%	0.2%	95.9%	0.0%	0.0%	0.0%	0.7%	0.8%	0.0%	0.4%	0.0%	0.0%	0.0%	1.3%
TOPIC34	0.0%	0.0%	0.0%	95.3%	0.1%	3.0%	1.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TOPIC35	0.0%	0.0%	0.0%	6.2%	5.5%	82.6%	0.0%	0.0%	0.3%	1.4%	1.9%	0.0%	1.1%	0.0%	0.0%	0.8%	0.1%
TOPIC36	0.1%	0.0%	0.0%	1.6%	94.2%	0.8%	0.2%	0.0%	0.0%	0.2%	0.4%	0.0%	0.1%	0.4%	0.0%	0.0%	2.0%
TOPIC37	0.1%	0.0%	0.1%	0.4%	0.4%	0.2%	85.2%	0.1%	0.2%	0.2%	0.1%	0.0%	0.2%	0.2%	0.0%	5.5%	7.4%
TOPIC38	1.5%	5.0%	32.6%	1.3%	0.3%	15.5%	4.2%	13.7%	0.6%	7.5%	6.2%	4.1%	0.0%	0.5%	0.0%	3.5%	3.5%
TOPIC39	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	4.4%	1.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	93.6%	0.6%
TOPIC40	0.4%	0.0%	0.0%	0.0%	0.1%	0.1%	95.5%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	3.6%	0.0%

上ある疾患の数は801であった（$|V|=801$）．分析対象病院のうち，特定機能病院は84施設，地域医療支援病院の数は525施設，DPC Ⅰ群，Ⅱ群，Ⅲ群はそれぞれ82，151，1424施設であった．総合入院体制加算は加算1，2，3がそれぞれ34，52，246施設で，在宅療養支援病院，在宅療養後方支援病院，そして三次救急医療施設はそれぞれ121，219，282施設であった．

4．2 病院機能の MDC 分布

病院機能を構成する MDC の分布を表1に示す．行はトピック TOPIC01～ TOPIC40で，列は主要診断群 MDC01～ MDC18（ただしデータが存在しない MDC17の精神疾患は除く），そして各セルはトピックに含まれる MDC の構成比率である．データバーを描いて構成比率の大小を視覚的に捉えられるようにしている．この表から多くのトピックが1～数個の MDC から構成されていることがわかる．たとえば，TOPIC13の99.2% は MDC06（消化器系疾患，肝臓・胆道・膵臓疾患），TOPIC31の99.4% は MDC05（循環器系疾患）といった具合である．一方，TOPIC15や TOPIC16のように，突出した MDC はなく，多くの MDC が混ざり合ったトピックもある．これらのトピックに含まれる疾患を構成比率の大きいものから10個ずつ挙げたのが表2である（ただし TOPIC13には該当する疾患が8個しかなかった）．「手術」は電子点数表[12]に記載された手術区分（DPC コードの9～10桁目）で，01～06は定義テーブルに定義された手術あり，97はその他の手術あり，そして99は手術なしを意味している．トピック番号の右にある括弧内の数字は当該トピックに含まれる疾患の数である．ただし，構成比率が閾値（$1/|V|=1/801$）を超えないものはカウントしていない．また，その横のパーセンテージは当該トピックに含まれる疾患のうち，手術ありの割合である．

4．3 病院のクラスター分析

図2は，各病院のトピック構成比率$\theta_{d,k}$をヒートマップで描いたものである．縦軸は病院dで，水平の白い線によって病院が属するクラスターを区切り，クラスター内では上から DPC 算定病床数の昇順に並べてある．横軸はトピックkである．ヒートマップは構成比率$\theta_{d,k}$が大きいものほど暗い色で表示している．クラスターによってトピック構成パターンが異なっており，それがクラスターを特徴づけている．数個の突出したトピックを持つクラスターもあれば，似たような構成比率を持ついくつかのトピックから構成されるクラスターもある．

図3は，クラスターごとに DPC 算定病床数と機能指数の平均値を求め，両者の相関関係をバブルチャートにしたものである．横軸は平均 DPC 算定病床数で，その値が大きいほど病院の規模が大きいことを表し，縦軸は平均機能指数で，その値が大きいほど病院が多くの機能を持っていることを表す．

表2　トピックに含まれる疾患の例

TOPIC13(8)93.6%	手術	割合	MDC
小腸大腸の良性疾患（良性腫瘍を含む.）	01	92.6%	MDC06
食道，胃，十二指腸，他腸の炎症（その他良性疾患）	99	1.9%	MDC06
穿孔または膿瘍を伴わない憩室性疾患	99	1.6%	MDC06
ヘルニアの記載のない腸閉塞	99	1.5%	MDC06
胃の悪性腫瘍	04	0.8%	MDC06
肺炎等	99	0.7%	MDC04
虚血性腸炎	99	0.6%	MDC06
胃の良性腫瘍	02	0.2%	MDC06
TOPIC15(151)46.2%	手術	割合	MDC
脳梗塞	99	6.9%	MDC01
てんかん	99	3.3%	MDC01
誤嚥性肺炎	99	3.2%	MDC04
非外傷性頭蓋内血腫（非外傷性硬膜下血腫以外）	99	3.2%	MDC01
急性心筋梗塞（続発性合併症を含む.），再発性心筋梗塞	97	2.7%	MDC05
頭蓋・頭蓋内損傷	99	2.6%	MDC16
心不全	99	2.4%	MDC05
腎臓または尿路の感染症	99	2.3%	MDC11
股関節大腿近位骨折	01	2.3%	MDC16
胆管（肝内外）結石，胆管炎	03	2.2%	MDC06
TOPIC16(254)52.7%	手術	割合	MDC
肺の悪性腫瘍	99	3.7%	MDC04
全身性臓器障害を伴う自己免疫性疾患	99	3.5%	MDC07
慢性腎炎症候群・慢性間質性腎炎・慢性腎不全	99	2.4%	MDC11
肺の悪性腫瘍	97	2.2%	MDC04
肝・肝内胆管の悪性腫瘍（続発性を含む.）	97	2.0%	MDC06
副腎皮質機能亢進症，非機能性副腎皮質腫瘍	99	1.8%	MDC10
緑内障	97	1.7%	MDC02
前立腺の悪性腫瘍	99	1.5%	MDC11
慢性化膿性中耳炎・中耳真珠腫	01	1.4%	MDC03
皮膚の悪性腫瘍（黒色腫以外）	01	1.3%	MDC08
TOPIC31(15)26.6%	手術	割合	MDC
狭心症、慢性虚血性心疾患	99	63.2%	MDC05
狭心症、慢性虚血性心疾患	02	18.7%	MDC05
頻脈性不整脈	99	4.4%	MDC05
急性心筋梗塞（続発性合併症を含む.），再発性心筋梗塞	97	3.9%	MDC05
徐脈性不整脈	97	3.7%	MDC05
弁膜症（連合弁膜症を含む.）	99	2.1%	MDC05
閉塞性動脈疾患	99	1.5%	MDC05
徐脈性不整脈	99	0.7%	MDC05
高血圧性疾患	99	0.5%	MDC05
心筋症（拡張型心筋症を含む.）	99	0.5%	MDC05

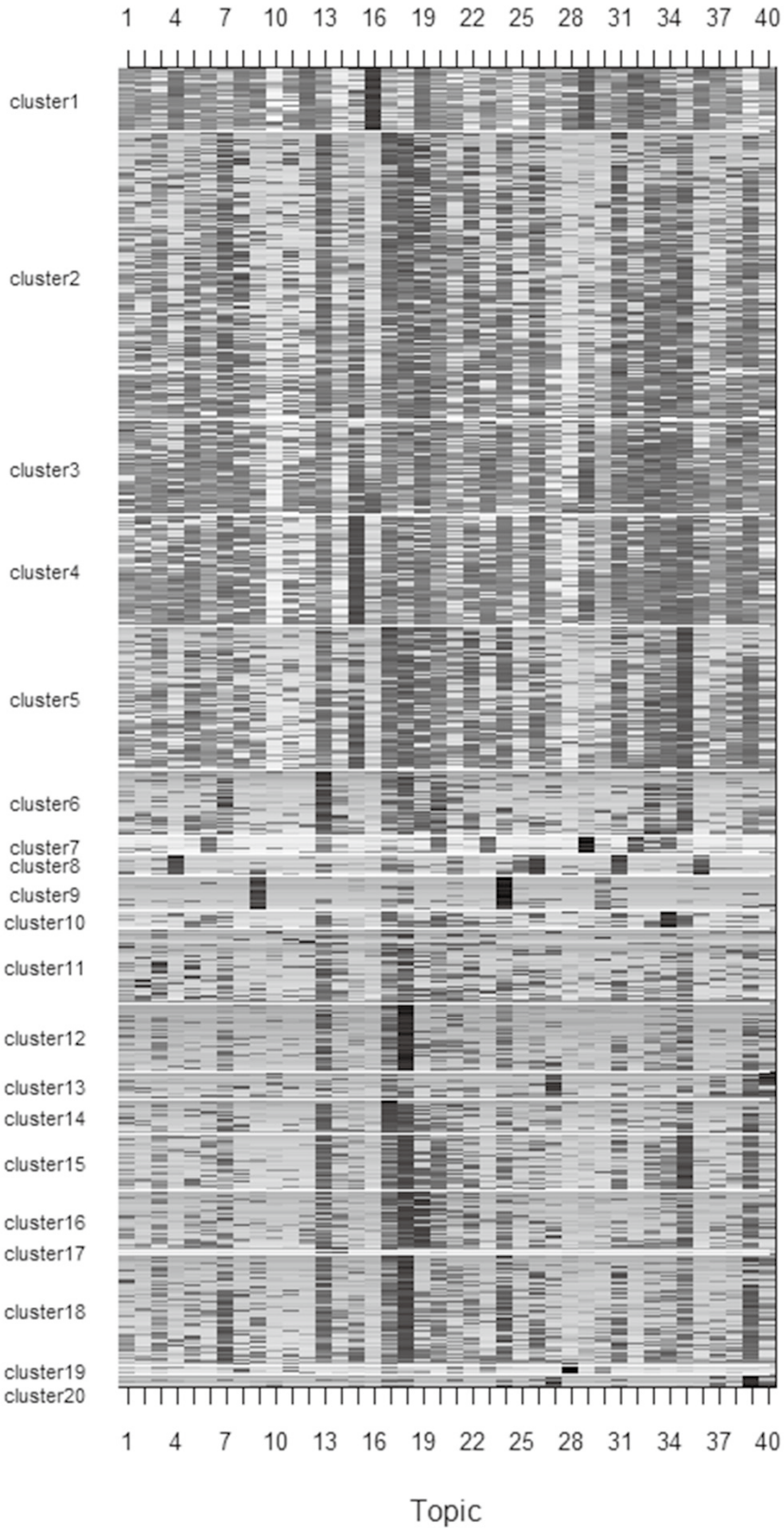

図2　トピックヒートマップ

バブルに付けた数字はクラスター番号で，バブルのサイズはそのクラスターに属する病院数を表している．図から病院規模と機能数の間には正の相関が見られる．それを確認するためにスピアマンの順位相関係数を求めたところ0.66（$p = 0.002$）となり，有意な相関が認められた．

表3は，クラスターごとに病院数，特定機能病院数，地域医療支援病院数，DPC Ⅰ，Ⅱ，Ⅲ群の病院数，総合入院体制加算1，2，3の病院数，在宅療養支援病院数，在宅療養後方支援病院数，そして三次救急医療施設数を集計したものである．各項目において，平均値＋標準偏差以上の値を太字にしている．病院数が最も多いのはクラスター2で，DPC Ⅲ群の病院数，総合入院体制加算3の病院数，地域医療支援病院数，在宅療養後方支援病院数でもトップである．図3からクラスター2の病院群は平均 DPC 算定病床数が約300床，平均機能指数が約14で，多くの機能を有している中規模病院であることがわかる．病院数79のクラスター1は，その大半を占める72病院が特定機能病院で，DPC Ⅰ群に占める病院数もトップである．図3からクラスター1の病院群は平均 DPC 算定病床数が約800床，平均機能指数が約10で，比較的多くの機能を有している大規模病院であることがわかる．しかしながら，図2のヒートマップを見るとクラスター1と2ではトピック構成の様相が大きく異なる．クラスター1では TOPIC16が大きな比重を占めている．一方，クラスター2には際立って大きな比重を占めるトピックが見られない．

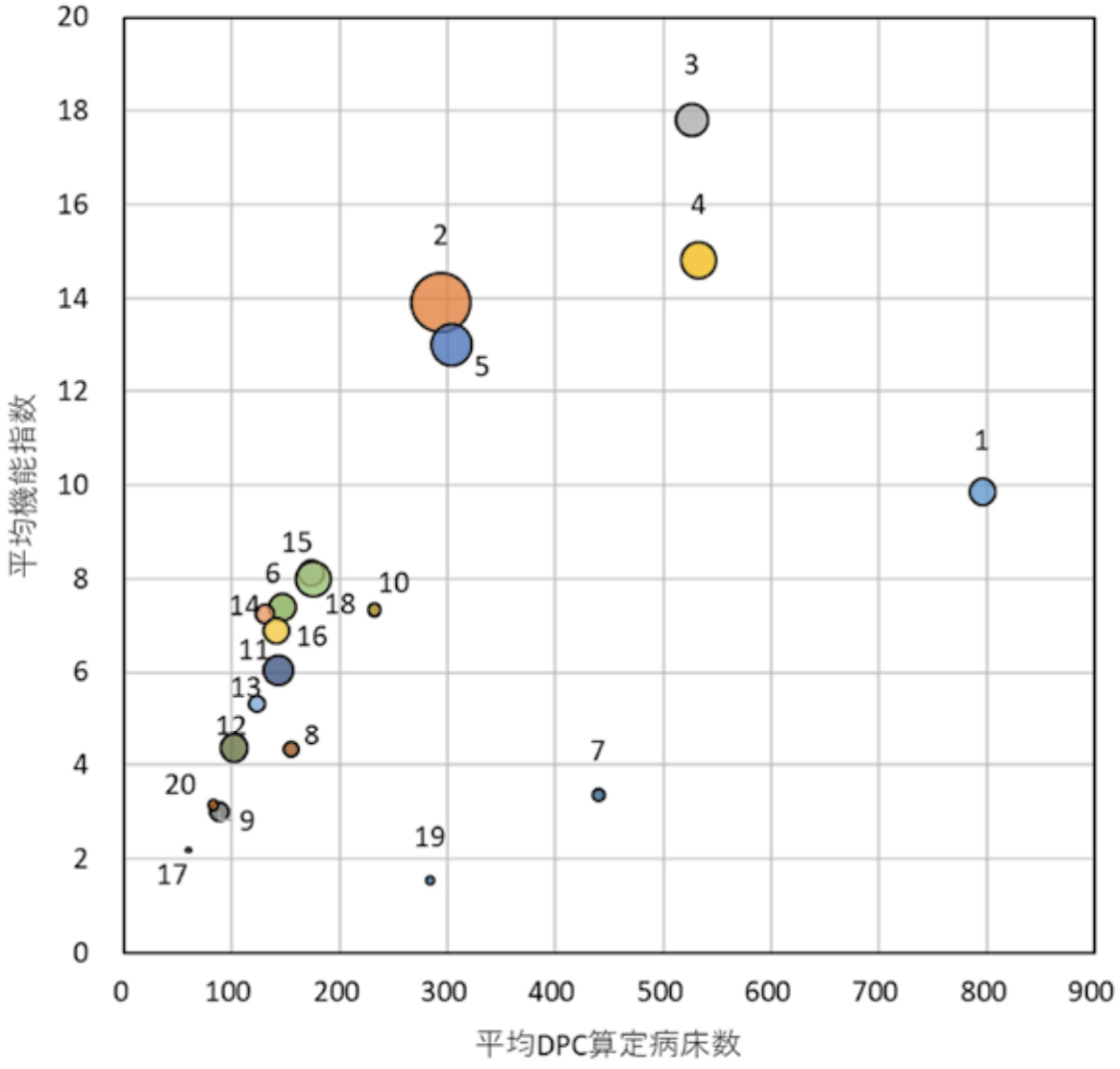

図3　クラスターごとの平均 DPC 算定病院数と平均機能指数の相関

表3 クラスターごとの病院機能の集計

クラスター	病院数	特定機能病院	DPC群の種類			総合入院体制加算			地域医療支援	在宅療養支援	在宅療養後方支援	三次救急医療施設
			Ⅰ群	Ⅱ群	Ⅲ群	加算1	加算2	加算3				
1	79	**72**	**74**	5	0	0	1	1	5	0	1	**59**
2	**365**	0	2	18	**342**	3	6	**81**	**154**	7	**64**	36
3	118	6	6	**48**	64	**13**	**15**	**43**	**78**	1	9	**56**
4	142	1	0	**57**	85	**15**	**24**	**61**	**107**	0	8	**104**
5	**180**	0	0	9	**170**	1	4	**39**	**85**	3	**38**	17
6	84	0	0	1	83	1	1	1	16	**16**	10	0
7	22	4	0	6	16	0	1	2	1	0	0	1
8	28	1	0	1	26	0	0	0	6	0	0	2
9	44	0	0	0	44	0	0	0	0	1	1	0
10	23	0	0	0	23	0	0	0	10	0	4	1
11	93	0	0	1	92	0	0	4	8	**18**	9	1
12	88	0	0	0	87	0	0	0	4	**25**	7	0
13	34	0	0	1	32	0	0	1	3	4	4	0
14	44	0	0	0	44	0	0	2	2	12	5	1
15	73	0	0	0	73	0	0	1	9	1	16	0
16	74	0	0	0	74	0	0	1	7	**14**	13	0
17	6	0	0	0	6	0	0	0	0	3	0	0
18	139	0	0	4	135	1	0	9	23	**16**	28	4
19	12	0	0	0	12	0	0	0	7	0	1	0
20	16	0	0	0	16	0	0	0	0	0	1	0
合計	1664	84	82	151	1424	34	52	246	525	121	219	282
平均	83.2	4.2	4.1	7.6	71.2	1.7	2.6	12.3	26.3	6.1	11.0	14.1
標準偏差	82.2	16.0	16.5	16.1	77.6	4.3	6.2	23.8	43.5	7.8	15.9	28.2

5．考察

トピックモデルは教師なし学習であり，病院が取り扱う疾患の出現頻度以外の情報は何も与えていないにもかかわらず，実際に用いられている制度上の病院機能を的確に推定していた．たとえば，表3に示すようにトピックモデルに基づくクラスター分析によって特定機能病院を高精度（適合率72/79=0.91, 再現率72/84=0.86）で識別している（クラスター1）．また，大学病院本院なみの高度な医療を提供するDPC Ⅱ群（現DPC特定病院群）を2つのクラスターに集約している（クラスター3と4）．さらに在宅療養後方支援病院（クラスター2，5）や在宅療養支援病院（クラスター6，11，12，16），そして三次救急医療施設（クラスター1，3，4）なども識別している．また，病院が持つ機能は病院規模が大きいほど多くなると考えられるが，図3の結果はそれを裏付けている．これらの結果から，トピックモデルは何らかの観点で病院機能をトピックとして抽出し，それに基づいて病院機能を分類できる可能性が示された．

本提案手法の利点は病院機能を抽出する際の柔軟性にある．これはトピックモデルの特徴であるソフトクラスタリングに由来している．ソフトクラスタリングとは，データが複数のクラスターに属することを許すクラスタリング手法のことである．トピックモデルでは，文書が複数のトピックを持つという仮定を置くことによりそれを実現している．実際の病院を考えても，高度な医療を提供しながら地域の診療所や病院と連携して地域医療を支援する病院もある．単一の観点からだけで病院機能を見るのは不自然である．

もう一つの利点として，現実とのギャップを分析するツールとしての利用価値がある．実際に使われている病院機能分類の多くは機能分化を誘導するために政策的に作り上げられたものである．そのため，必ずしも現状に即していないケースがある．それを是正するには現状との乖離を把握する必要があるが，本提案手法はその材料を提供できる．たとえば表3においてクラスター2の多くの病院はDPC Ⅲ群（現DPC標準病院群）であるが，Ⅰ群やⅡ群が若干混ざっている．なぜ，それらの病院がクラスター2に分類されたのかを精査することにより，あるべ

き姿との乖離を究明して医療政策や病院経営に活かせるかもしれない.

一方，課題としてはモデル選択の問題がある．トピックモデルに限らず，一般的にクラスタリング手法はクラスター数Kを外部から与える必要がある．LDA の場合，クラスター数はトピック数と等価であるが，できあがるモデルはトピック数によって異なるため，事前にKを決めなければならない．これをモデル選択という．LDA で最適なトピック数Kを決定する方法として Perplexity を利用する方法や尤度を計算する方法がある[3]．Perplexity は単語の平均分岐数を表しており，トピック内に現れる単語の均質性を示す指標になっている．Perplexity は低いほど良いモデルとされており，この値が最小となるトピック数を採用する．一方，尤度はデータの当てはまりの良さを示す指標である．この値が大きいほどモデルがデータに適合していることになるので，尤度が最大になるトピック数を採用する．しかしながら，予備実験で Perplexity（と尤度）のトピック数依存性を調べたところ，K〜200くらいまで Perplexity は急激に減少（尤度は急激に増加）するが，その後はKの増加に伴って緩やかに減少し続け（尤度は増加し続け）極致に達する気配が見られない．当然のことながらトピック数は疾患数$|V| = 801$を超えることはない．そして，$K = |V|$は1トピック1疾患という究極の状態を表し，理論的には誤りではないが，トピックとしての意味がない．予備実験では疾患数のトピック平均がトピック数Kの増加に伴って単調に減少するという結果が得られている．これはモデルがデータに過剰適合している可能性を示唆している．そこで，1トピック当たりの疾患数が20程度になるように本研究では$K = 40$としてモデルを構築した．しかし，これでは恣意的で妥当性に欠けるので，最適なトピック数の決定については今後の課題として残された.

最後にトピックの解釈について考察する．トピックの解釈については，表2に示すように各トピックに含まれる疾患を構成比率の高いものから並べて人間がそれを見て解釈を行うことになる．一般にトピックモデルは得られたトピックの解釈が難しいとされているが，本研究の場合は MDC が解釈の助けとなる．表2の例で言えば TOPIC13に含まれるほとんどの疾患が MDC06なので，消化器系疾患を扱う診療機能と解釈できる．しかも手術ありが93.6% とかなりの割合を占めているので，そういったスタッフや設備の整った医療機能を提供しているものと考えられる．同様に TOPIC31は循環器系疾患を扱う診療機能と解釈できる．ただし，手術ありが26.6% と比較的低いので内科的なアプローチを主に提供する機能と考えられる．一方，TOPIC15や TOPIC16はそれぞれ151，254と多くの疾患を含み，しかも突出して構成比率の高いものはなく，MDC も多岐にわたっているため，一見して解釈に戸惑う．しかし，図2を見ると TOPIC16はクラスター1においてひと際目立つトピックであり，クラスター1の多くが特定機能病院であることを考えると，広範な領域で高度な医療を提供する機能と解釈できる．TOPIC15についても同様で，クラスター4に顕著に現れることから，総合的な体制を充実させた機能あるいは重篤な疾患に対する高度な救急医療を提供する機能を表しているものと考えられる．このようにトピックは MDC そのものではなく，MDC の組み合わせによって診療機能を表現している．多くの病院は単一のトピックだけからなるのではなく，複数のトピックから構成され，その構成比率がその病院を特徴づけ，病院の機能を表している．このような病院機能の捉え方に本研究の新規性がある.

6．結語

本研究では自然言語処理分野で文書の分類や検索に用いられるトピックモデルを利用して公表 DPC データから病院機能を推定する手法を提案した．具体的には文書を病院，単語を疾患（厳密には疾患＋手術），そして単語の出現頻度をその疾患の出現頻度に対応させてモデルを構築し，トピックを抽出した．得られたトピックを用いて病院を分類したところ，特定機能病院や地域医療支援病院など制度上の病院機能分類との整合性が見られたので，病院が提供する診療機能をトピックによって推定できる可能性を示すことができた.

一方，今回の試みはトピックモデルの入力データとして疾患の出現頻度のみを利用しているため，包括的な病院機能を把握する上では限定的と言わざるを得ない．例えば地域医療支援病院は，紹介患者に対する医療の提供や救急医療の提供等，地域で必要とされる様々な取り組みを通じて，かかりつけ医等を支援する医療機関と位置付けられているので，取り扱う疾患の数だけでは機能を抽出できない．地域の実情に合った病院の機能を評価するには地理的な条件や医療スタッフなど人的資源，医療機器などの物的資源，そして紹介率・逆紹介率など地域医療連携の指標も加味する必要があろう．これらについては今後の課題としたい.

注

†1）地域医療支援病院であるかどうかという属性は平成28年度病床機能報告公表データにはなかったので，これのみ平成29年度病床機能報告公表データのものを用いた．

†2）2018年度から，Ⅰ群は「大学病院本院群」，Ⅱ群は「DPC 特定病院群」，そしてⅢ群は「DPC 標準病院群」に変更になっている．

文 献

1）厚生労働省：病床機能報告．
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000055891.html, 2018.（2019.2.19確認）

2）厚生労働省：地域医療構想．
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000080850.html, [2016].（2019.2.19確認）

3）佐藤一誠，奥村学監修：トピックモデルによる統計的潜在意味解析．コロナ社，東京，2015.

4）Blei DM, Ng AY and Jordan MI：Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022, 2003.

5）畠山豊，宮野伊知郎，片岡浩巳，中島典昭，渡部輝明，奥原義保：問診データに対する潜在トピックモデルに基づく健診データ解析．医療情報学，**33**(5)，2013.

6）Hatakeyama Y, Ogawa T, Ikeda H and Haseyama M：A most resource-consuming disease estimation method from electronic claim data based on labeled LDA. *IEICE Transactions on Information and Systems*, **E99.D**(3), 763-768, 2016.

7）Chen JH, Goldstein MK, Asch SM, Mackey L and Altman RB：Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *Journal of the American Medical Informatics Association*, **24**(3)，472-480, 2017.

8）Luo YF and Rumshisky A：Interpretable topic features for post-ICU mortality prediction. *American Medical Informatics Association（AMIA）2016 Annual Symposium Proceedings*, 827-836, 2016.

9）Xiao C, Zhang P, Chaowalitwongse WA, Hu J and Wang F：Adverse drug reaction prediction with symbolic Latent Dirichlet Allocation. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence（AAAI-17）*，1590-1596, 2017.

10）厚生労働省：平成28年度 DPC 導入の影響評価に係る調査「退院患者調査」の結果報告について．
https://www.mhlw.go.jp/stf/shingi2/0000196043.html, 2018.（2019.2.21確認）

11）Grünand B and Hornik K：Topic models.
https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf, 2018.（2019.2.21確認）

12）厚生労働省：診断群分類（DPC）電子点数表について．
https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000039920.html, 2017.（2019.2.21確認）

13）Laine CR：The Herfindahl-Hirschman index: A concentration measure taking the consumer's point of view. *The Antitrust Bulletin*, **40**(2)，423-432, 1995.

（令和元年6月7日受理）

Analysis of DPC Data by Topic Model and Consideration on Function of Hospitals

Masaaki TANAKA

（Accepted Jun. 7, 2019）

Key words : topic model, DPC (Diagnostic Procedure Combination), function of hospital, LDA (Latent Dirichlet Allocation), NLP (Natural Language Processing)

Abstract

In order to promote the differentiation and cooperation of medical functions, it is necessary to grasp the hospital function based on objective data and make use of it in decision making of medical policy. Therefore, in this research, the author devised a method to analyze hospital functions from published DPC (Diagnosis Procedure Combination) data using the method cultivated in the natural language processing field and examined the relationship with the existing hospital function classification. The author applied a topic model that estimates topics hidden in hospitals by associating hospitals with documents, words as diseases, and word frequencies as disease frequencies. The author considered the topic obtained as a basic medical function and defined the hospital function by its composition ratio. As a result of hierarchical cluster analysis using the basic medical function (topic) of each hospital as a feature, correspondence was found between the cluster and the existing hospital function. In particular, the author was able to estimate Special Functioning Hospitals with high accuracy (precision 72/79 = 0.91, recall 72/84 = 0.86). The devised method is characterized by grasping the hospital function as a combination of a plurality of basic medical functions (topics), suggesting the possibility of quantitatively measuring the current state of functional differentiation of the hospital. However, the hospital function can not be identified only by the number of diseases actually treated. It is necessary to further improve the accuracy considering geographical conditions as well as the human and material resources possessed by the hospital.

Correspondence to : Masaaki TANAKA

Department of Health Informatics
Faculty of Health and Welfare Services Administration
Kawasaki University of Medical Welfare
Kurashiki, 701-0193, Japan
E-mail：mtanaka@mw.kawasaki-m.ac.jp

（Kawasaki Medical Welfare Journal Vol.29, No.1, 2019 127－137）