

# 類似性を考慮した訪日外国人旅行者の訪問パターン抽出に関する基礎的研究

古屋 秀樹<sup>1</sup>

<sup>1</sup>正会員 東洋大学教授 国際観光学部国際観光学科 (〒112-8606 東京都文京区白山)  
E-mail: furuya@toyo.jp

本研究は、訪日外国人旅行者の訪問パターンの抽出を都道府県区分よりも細かい地点レベルで行うことを目的としている。論文では、はじめに訪問地の組合せである訪問パターンの抽出に用いる手法として Hierarchical Pachinko Allocation Model(hPAM)を説明するとともに、Latent Dirichlet Allocation Modelとの差異を明らかにする。hPAMは教師データなしの機械学習（セグメンテーション手法）の1つと位置づけられ、各訪問パターンをトピックに区分する確率的導出過程が明示でき、トピック間の関連性を階層構造として考慮できる特色を有する。分析では、観光庁が実施した「訪日外国人消費動向調査（平成27、28年）」データを用いながら、hPAMによって訪問場所の組合せパターンを実際の地点別訪問の有無に基づき分類した。

**Key Words:** *Combination of visiting place, Hierarchical Pachinko Allocation Model*

## 1. はじめに

2017年の訪日外国人旅行者数は約2,869万人を数え、過去最高を記録した。さらなるインバウンド振興のためには、旅行者の行動・意向特性を明らかにしながら、旅行者ニーズへのさらなる対応やプロモーションの強化が必要と言える。

訪日旅行では、その訪日動機・目的を満たすために旅行者は、目的地を定め、一連の訪問を通じて、事前期待を充足し、事後的な評価・満足が形成される。そのため、観光地への訪問の実績データからさかのぼって訪日動機、目的を類推する方法も考えられる。

そこで、本研究は訪日外国人旅行者の訪問パターンの把握を目的とする。この訪問パターンとは、日本における「訪問地の組み合わせ」と設定する。分析では、観光庁が実施した「訪日外国人消費動向調査」データを用いているが、サンプル数が1回の調査につき約4万人を数えるため、訪問パターンも多様な中で、それらを効率的かつ論理的に訪問パターンを集約しなければならない。

そこで、本論文では多数の訪問実績データから訪問パターンをセグメントでき、論理的整合性や具体的なセグメントの導出過程を明示できるHierarchical Pachinko Allocation Model（以下、hPAM）を用いて分析を行い、適用事例があるLatent Dirichlet Allocation（LDA）モデルと比較を行い、その特徴の把握を目的とする。

## 2. 推定手法と本研究の位置づけ

### (1) 訪日外国人消費動向調査<sup>1)</sup>

訪日外国人の訪問パターン把握のために、国籍・地域別の偏りが比較的小さいこと、入国空海港が明確なことから「訪日外国人消費動向調査（観光庁、平成27、28年）」の個票データを用いる。本調査は、18空海港の国際線ターミナル搭乗待合ロビーで出国を待つ訪日外国人旅行者に対して対面形式で行われたものである。

本研究では、その中から、「今回の日本滞在中における訪問地（市区町村より細かな実際の訪問地点を回答）」に関する回答を利用しているが、これは宿泊、日帰りの区分はされていない点に留意する必要がある。

### (2) 分析手法について

旅行者の訪問パターンの分析手法として、ロジットモデルを用いた研究<sup>2)</sup>がみられるが、旅行者の多様性を考慮することが困難と考えられる。

本研究は、訪日外国人旅行者の行動特性把握を第一義的に考え、1つの行動特性である訪問パターンについて実態把握するとともに、適切なセグメント導出を行う。訪問パターンの特徴把握として、特定地点前後の訪問箇所を集計している事例<sup>3)</sup>などがあるが、3地点以上の訪問地の組み合わせの把握が困難と言える。また、バスケット解析、クラスター分析や主成分分析を適用した事例があるが、何故異なるのかという生成過程を明示できてい

ない点が課題といえる。これらを解消した方法として潜在クラス分析があげられ、訪問パターンへの適用事例4, 5, 6がある。しかしながら、潜在クラス分析でも、1旅行の中における旅行者の嗜好を表す「トピック」の混在を考慮できなかったり、全旅行者でトピック構成比率が同一と仮定していること、パラメータの事前確率を想定していないため過学習<sup>7)</sup> (overfitting) の恐れがあるなどの問題点が存在する。これらを改良したものとして、LDAモデル<sup>8,9,10,11)</sup>があり、訪問パターン分析への適用事例として文献12, 13がある。

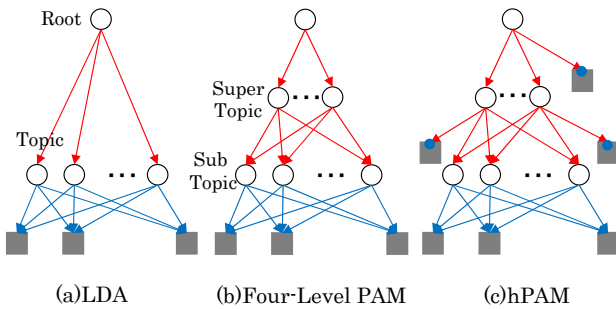


図-1 DAGを用いた各モデルの生成構造  
(○: トピック, ■: 訪問地)

図-1(a)は、LDAモデルの生成構造をDAG(Directed Acyclic Graph, 有向非巡回グラフ)を用いて示したものであるが、これより訪問地は旅行トピックと密接に関連するとともに複数トピックへの包含が考慮可能なこと、トピック間の相関はないと仮定していることがわかる。しかしながら、旅行トピック相互には類似性や関連性が存在することが容易に想定でき、LDAモデルでは考慮できない。そのため、トピック数を多く設定した場合、強制的に「相関のないトピック」に分割するため、説明力の低下も懸念される。

### (3) Hierarchical PAM の概要

トピック間の関連を考慮するために、トピック間の相関係数を明示したCorrelated Topic Models (CTM)<sup>14)</sup>やトピックに階層構造を設定するPAM (Pachinko Allocation Model) の適用がある。例えば、PAMの一例であるFour-Level PAM<sup>15)</sup> (図-1(b)) では、Super TopicとSub Topicと2層のトピックを仮定し、便宜上設けた最上段のRoot TopicからSuper Topic, ならびにSuper Topic からの下層トピック全てに設定されたパスによってトピック間の関連性を考慮している。さらにPAMを拡張したものとしてHierarchical PAM<sup>16)</sup> (図-1(c)). 以下、hPAM) があり、この特徴として全てのトピックが訪問地と関連付いている点があげられる。本研究では、推定の容易性や、比較的説明力が良好であったhPAMを訪問パターン分析に適用する。図-2は、モデル別グラフィカル表現である。

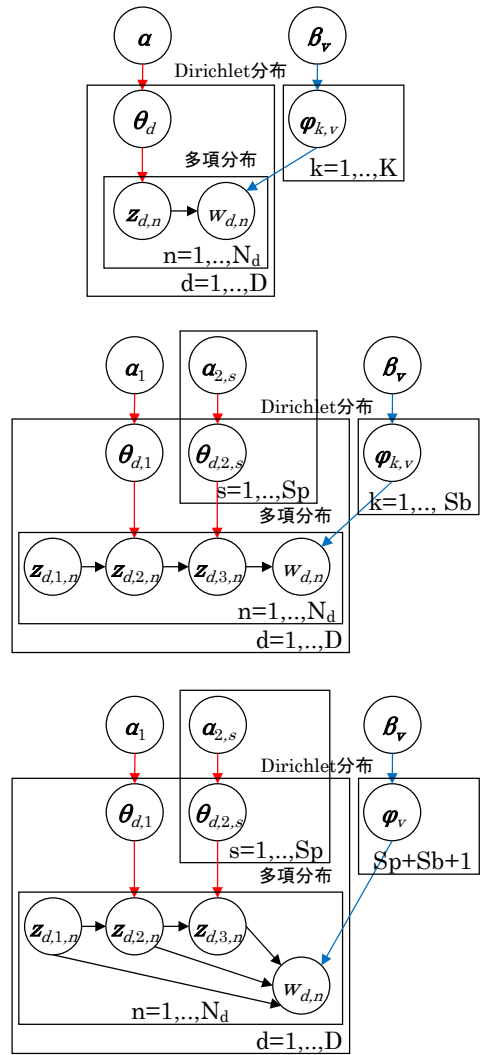


図-2 各モデルのグラフィカル表現  
(上段: LDA, 中段: PAM, 下段: hPAM)

ここで、

$\alpha$ : 旅行トピックを導出する多項分布パラメータを規定するDirichlet分布パラメータ,

$\theta$ : 旅行dのトピックを規定する多項分布パラメータ,

$\beta$ : 訪問地別訪問割合を導出する多項分布パラメータを規定するDirichlet分布パラメータ,

$\phi$ : トピック(もしくはトピックの組合せ)毎の訪問地別訪問割合を規定する多項分布パラメータ,

$z$ : 潜在トピック,

$D$ : 訪日外国人による旅行総数,

$N_d$ : 旅行dにおける訪問地総数.

データと整合するように直接、 $\theta, \phi$ を推定しないのは、事前確率分布を設定しながら、過学習を抑えるためでもある。図-2よりLDAでは、旅行毎のトピック分布パラメータ $\theta$ をサンプリングし、それをもとに単語ごとのトピックをサンプリングする。

LDAに対してPAMでは、トピックが2層のため、旅

行毎のトピック分布パラメータ  $\theta$  のサンプリングが 2 回行われる点が異なる。さらに、hPAM では、ルートトピックからは、それに関連付く訪問地分布と全スーパートピックが経路（経路数： $S_p+1$ ）として、スーパートピックからもそれに関連付く訪問地分布と全サブトピックが経路（同： $S_b+1$ ）となる。それと対応して LDA, PAM では  $\phi$  の次元は、トピック数 ( $K$ )、サブトピック数 ( $S_b$ ) であったが、それに加えて、スーパートピック数 ( $S_p$ ) ならびにルートトピック数 (1) を加えたものとなる。さて、hPAM の生成モデルは下記のように示すことができる。

- (1) 各旅行  $d$  において、スーパートピックパラメータ  $\theta_{d1}$  ならびにスーパートピック毎のサブトピックパラメータ  $\theta_{d2s}$  を確率分布をもとにサンプリングする、
- (2) 各旅行  $d$  におけるそれぞれの訪問地  $n$  に対して、
  - (a) パラメータ  $\theta_{d1}$  からスーパートピック： $z_{d2n}$  をサンプリングする。 $z_{d2n}=0$  の場合、 $\phi_{v0}$  から訪問地をサンプリングする、
  - (b) それ以外の場合、パラメータ  $\theta_{d2s}$  からサブトピック： $z_{d3n}$  をサンプリングする。 $z_{d3n}=0$  の場合、 $\phi_{vz_{d2n}}$  から訪問地をサンプリングする、
  - (c) それ以外の場合、 $\phi_{vz_{d3n}}$  から訪問地をサンプリングする。

以上から、トピックを与えられた旅行： $d$ 、訪問地： $n$  に対するサンプリング分布は(1)式のように示すことができる。

$$P(z_{d,2,n}, z_{d,3,n} | w, z_{d,2 \setminus n}, z_{d,3 \setminus n}, \alpha, \beta) \propto \frac{\alpha_{z_{d,2,n}} + N_d^{z_{d,2,n}}}{\sum_{z'_{d,2,n}} \alpha_{z'_{d,2,n}} + N_d} \times \frac{\alpha_{z_{d,3,n}} + N_d^{z_{d,2,n} z_{d,3,n}}}{\sum_{z'_{d,3,n}} \alpha_{z'_{d,3,n}} + N_d^{z_{d,2,n}}} \times \frac{\beta_n + N_{z_{d,2,n} z_{d,3,n}}^n}{\sum_{n'} \beta_{n'} + N_{z_{d,2,n} z_{d,3,n}}}$$

... (1)式

ここで、 $w$ ：訪問地集合、  
 $N_d$ ：旅行総数、  
 $N_d^{z_{d,2,n}}$ ：トピック  $z_{d,2,n}$  に含まれる旅行数、  
 $\alpha_{z_{d,2,n}}$ ：トピック  $z_{d,2,n}$  に関する Dirichlet 分布パラメータ、  
 $N_{z_{d,2,n} z_{d,3,n}}^n$ ：トピック ( $z_{d,2,n}, z_{d,3,n}$ ) に含まれる目的地  $n$  の出現数。

(1)式の初項、第2項、第3項は、それぞれルートトピック、スーパートピック、サブトピックから下層に向けた分岐確率を表現しているものである。なお、実際の推定であるが、尤度最大化または事後尤度最大化による推定も考えられるものの、それらを用いた場合に PAM では解が有限個 (closed-form) とならないため、モーメントマッチングによって近似していた<sup>17)</sup>。しかしながら、

本論文では Gibbs EM 手法<sup>18),19),20)</sup>を用いた。また、導出されたモデルの妥当性評価であるが、LDA では平均分岐数 (perplexity, PPL)<sup>21)</sup>が用いられるケースが比較的多い。しかしながら、パラメータ数が大きく異なり、訪問地がトピックの階層構造によって表現されているケースでは PPL では適当な評価が行えない。そのため、本研究では最終対数尤度、尤度比によって評価を行うものとする。

### 3. 分析データ・基礎的特性把握

#### (1) 分析データの概要

本研究では、「訪日外国人消費動向調査（平成 27 年、28 年）」の個票データを用いる。本調査は、18 空海港の国際線ターミナル搭乗待合ロビーで出国を待つ訪日外国人旅行者に対して、12 言語対応のタブレット端末または紙調査票を用い、外国語を話せる調査員による聞き取り方式によっておこなわれたものである。なお、国籍・地域別の調査枚数は、20 の国籍・地域と「その他の国籍・地域」の計 21 区分について目標回答数を設定し、四半期毎に総数 9,710 サンプルを目標に調査が実施されている。調査項目は、個人属性ならびに訪日旅行に関する訪問地、同行者、旅行支出、土産品の購入実態、旅行情報源、満足度と再訪意向の聞き取りを実施したものであり、サンプル数は 79,739 人 (2015 年：39,783 人、2016 年：39,956 人) である。

#### (2) 個票の拡大処理<sup>22)</sup>

さて、訪日外国人消費動向調査はアンケート調査に協力した被験者データであるため、その属性に偏りが考えられる。本来、抽出台帳に基づいたランダムサンプリングにより標本が決定され、抽出率の逆数である拡大係数を用いて補正する方法が適切といえるが、訪日外国人旅行者には抽出台帳自体がない。そこで、個別の属性などの周辺確率のみ合致させる補正が考えられ、その際には全数把握されているデータの活用が必要となる。例えば、訪日外客数 (JNTO) は国籍・地域別月別に整備されており、出入国管理統計の中では月別の国籍・地域別年齢階層別男女別データがある。本分析では、訪問地の特性把握を重視することとし、それとの関連性が高いと考えられる空海港別国籍・地域別出国者数を用いて、四半期別出国空海港別出国者数に対する国籍・地域別出国空海港別サンプル数の比率の逆数をサンプルに乗じて拡大処理を行った。

なお、訪日外国人旅行者数は、日本永住者等の外国人ならびに乗員上陸数を除く一方、外国人一時上陸客等を加えた入外国籍旅行者のため出入国管理統計と一致しないものの、空海港別の構成比率を考慮することが重要と判断した。また、訪日外国人消費動向調査が出国時の調査のため、出国者数を用いた。

表-1は、四半期別のサンプル数 (2015年、2016年)



ならびに出国者数（出入国管理月報）を示したものである。調査データは、各四半期でほぼ同程度のサンプルが得られている一方、出国者数は増加傾向を示しており、国籍等や出国空海港の構成比率やその偏りをチェックしながら、訪日外国人旅行者の行動特性を把握する必要があると考えられる。

次に、表-2 を用いてサンプル構成比率を確認すると、まず国籍・地域別構成比率（最右欄（列））では、上位 6 カ国において韓国、米国のサンプル比率が実績値より高い一方で、他の地域は低くなる乖離がみられる。一方、出国空海港別構成比率（最下欄（行））では主要 18 空海港を対象としているため実績値合計が 97%に留まることに留意する必要がある。

表-1 四半期別サンプル数ならびに出国者数

調査時期	サンプル数	比率	出国者数	比率
2015.Q1	9,946	12%	4,217,941	10%
2015.Q2	10,045	13%	4,975,296	12%
2015.Q3	9,948	12%	5,112,911	12%
2015.Q4	9,844	12%	5,167,472	12%
2015年計	39,783	50%	19,473,620	46%
2016.Q1	9,945	12%	5,588,907	13%
2016.Q2	10,045	13%	5,681,542	13%
2016.Q3	9,998	13%	5,915,254	14%
2016.Q4	9,968	13%	5,831,450	14%
2016年計	39,956	50%	23,017,153	54%
合計	79,739	100%	42,490,773	100%

表-2 出国空海港・国籍地域別平均立寄り都道府県数（観光庁データ）ならびに構成比率

	新千歳空港	函館空港	仙台空港	新潟空港	羽田空港	成田空港	富士山静岡空港	小松空港	中部空港	関西空港	広島空港	高松空港	福岡空港	鹿児島空港	那覇空港	関門港（下関）	博多港	厳原港	総計	観光庁 (%)	出入国 (%)
中国	1.2	1.3	2.2	2.4	2.3	3.6	4.3	4.3	4.1	3.4	2.0	4.1	2.3	1.3	1.1	2.6			3.2	15	23
韓国	1.0		1.5	1.5	1.5	2.2	2.4	1.8	1.8	2.0	1.5	1.5	1.9	1.1	1.0	2.2	1.8	1.0	1.8	40	23
台湾	1.0	1.2	2.9	2.0	2.0	2.7	2.3	3.5	3.0	2.5	2.8	2.6	2.8	2.1	1.0				2.3	11	18
香港	1.2					1.8	2.6		2.9	2.4	2.9		2.4	1.7	1.0				2.1	3	8
米国	1.5		1.0		2.2	2.8	2.0		2.2	3.5	1.0		2.7		1.0		1.0		2.9	7	5
タイ	1.0				2.1	2.9			2.3	2.8			2.8						2.5	2	4
フィリピン					2.1	2.6			2.2	2.4			1.9						2.4	2	2
オーストラリア	1.0				2.9	3.7			4.0						1.0				3.6	2	2
マレーシア	1.1				2.4	2.7			2.4	2.6			2.2		1.0				2.6	2	2
シンガポール	1.0				2.0	3.0			2.6	2.4			2.7		1.0				2.5	1	2
英国	1.5				2.6	3.4			1.3	3.7			3.3						3.1	2	1
カナダ	1.0				2.7	3.1			1.5	4.0			2.5		1.0				3.1	2	1
フランス					2.8	3.8			4.5	4.1			3.5						3.5	1	1
インドネシア	2.0				2.6	3.2			2.0	3.0			2.0						2.9	2	1
ドイツ					2.8	3.4			2.0	3.4			3.3						3.1	1	1
ベトナム					2.5	3.3			3.1	2.8			2.0						3.0	1	1
インド					2.7	2.4				2.6			1.5						2.5	2	1
イタリア					2.8	3.6			2.0	4.5									3.8	2	1
スペイン					3.1	4.2			6.0	4.2									4.1	1	0
ロシア	1.5					2.8				3.3			4.0						2.7	1	0
その他	1.0				3.3	3.5			2.5	3.5			2.6		3.0	1.0			3.4	1	5
総計	1.1	1.2	2.1	2.0	2.0	3.0	3.5	3.1	2.9	2.6	2.1	2.4	2.0	1.3	1.0	2.2	1.8	1.0	2.4	100	100
観光庁	5	0	0	0	11	34	1	0	4	25	0	0	9	0	5	1	2	1	100		
出入国	5	0	0	0	14	30	0	0	5	26	0	0	7	0	6	0	1	0	97		

※観光庁：訪日外国人消費動向調査（2カ年分）の構成比率  
 ※出入国：出入国管理統計月報（2カ年分）の構成比率  
 ※空白セル：サンプル数0を示す。

その中で、成田空港が若干高く、他の空海港はほぼ実績値と近似している。以上から、国籍・地域と出国空海港の構成比率において、十分な整合性をもっているとは言い難く、サンプルから導出される値が偏りを持つ恐れがある。

これらを念頭に、都道府県別の訪問を考えて、表-2 中央部に出国空海港別国籍・地域別平均立寄り都道府県数を示した。出国空海港や国籍・地域によって平均値にばらつきがみられるため、サンプル構成比率の適切な処理が必要と考えられる。これらを考慮するために、四半期別にコントロールトータル（空海港別出国者数）に対する国籍・地域別出国空海港別サンプル数の比率の逆数をサンプルに乗じて拡大処理を行った。なお、サンプルの欠損が考えられるものより細かい出国空海港別の特徴を捉えること、ならびに JNTO が重点市場と設定していることから、18 空海港ならびにその他を含む 21 国籍・地域区分を設定した（合計 378セル）。なお、サンプルが存在しないセルは、その拡大係数を設定することができないため、拡大自体を行わないものとした。

(3) 拡大後データの基礎的特性把握

次に、国籍・地域別訪日旅行者の基本的特性の把握を行った。表-3 は、拡大後サンプル数・リピータ比率を示したものである。2016年では、上位 4 カ国が占める割合が 71%と高いこと、2 回以上の訪問であるリピーター比率は全体で 60%となっているものの、中国で低かったり、欧米諸国で減少傾向を示す特徴が見られる。

表-3 国籍・地域別サンプル数・リピータ比率

No.	国籍・地域	① 15年 出国者数	② 16年 出国者数	③ 16年 構成比率	④ 増加比率 ((②-①)/①)	⑤ 15年 リピータ比率	⑥ 16年 リピータ比率	⑦ 増分 (⑥-⑤)
1	中国	429	498	23%	16%	37%	42%	5%
2	韓国	387	497	23%	28%	67%	66%	0%
3	台湾	346	390	18%	13%	79%	81%	2%
4	香港	144	168	8%	17%	82%	81%	-1%
5	米国	96	122	6%	26%	53%	56%	3%
6	タイ	81	91	4%	13%	65%	64%	0%
7	フィリピン	37	45	2%	23%	51%	55%	4%
8	オーストラリア	33	41	2%	23%	47%	46%	-1%
9	シンガポール	29	35	2%	19%	70%	71%	1%
10	マレーシア	25	33	1%	29%	55%	48%	-7%
11	英国	22	27	1%	19%	44%	44%	0%
12	インドネシア	19	25	1%	31%	50%	48%	-2%
13	フランス	21	25	1%	18%	46%	43%	-3%
14	カナダ	21	24	1%	13%	50%	47%	-3%
15	ベトナム	15	19	1%	29%	48%	44%	-4%
16	ドイツ	16	18	1%	12%	58%	52%	-6%
17	イタリア	9	11	1%	17%	42%	37%	-5%
18	インド	9	10	0%	18%	52%	54%	2%
19	スペイン	7	8	0%	16%	29%	28%	0%
20	ロシア	5	5	0%	1%	58%	56%	-2%
21	その他	84	96	4%	15%	45%	46%	0%
	総計	1,834	2,187	100%	19%	59%	60%	1%

※出国者数単位：万人

表4 は、観光・レジャー目的、性・年齢階層別構成比率である。東アジア諸国を中心に観光・レジャー目的が7~8割を占める一方、インド、米国が低いこと、同様に性別別構成比率では近隣諸国で女性が多い一方、その他では男性の比率が多いこと、40歳代までが多くを占めることがわかる。

表-4 国籍・地域別観光・レジャー目的、性年齢階層構成比率

No.	国籍・地域	15年 観光 等目的	16年 観光 等目的	男性 29歳 以下	男性 30代	男性 40代	男性 50代	男性 60歳 以上	女性 29歳 以下	女性 30代	女性 40代	女性 50代	女性 60歳 以上
1	中国	73	73	14	16	18	4	2	21	18	10	5	2
2	韓国	74	78	19	15	10	6	3	25	11	6	4	2
3	台湾	83	83	12	15	19	5	3	17	18	12	6	3
4	香港	87	89	12	13	12	6	2	18	14	13	8	2
5	米国	34	43	15	16	15	15	7	9	7	5	5	8
6	タイ	68	72	13	16	10	5	1	21	19	9	5	2
7	フィリピン	47	50	14	14	10	5	2	19	15	11	7	3
8	オーストラリア	72	78	20	13	11	8	7	17	8	7	5	5
9	シンガポール	63	69	15	15	13	8	5	15	10	10	6	3
10	マレーシア	63	67	15	15	12	10	3	13	17	7	5	2
11	英国	44	56	20	16	14	11	7	12	6	5	3	4
12	フランス	57	61	21	21	13	7	3	13	8	6	4	2
13	カナダ	55	60	17	15	13	11	8	13	6	4	6	7
14	インドネシア	50	52	17	17	13	6	3	19	13	8	3	2
15	ドイツ	38	40	16	21	20	13	4	10	7	4	2	3
16	ベトナム	29	33	22	23	17	4	3	16	15	3	4	3
17	イタリア	48	56	15	25	17	10	3	10	12	5	3	2
18	インド	15	13	23	32	18	7	4	8	4	2	1	1
19	スペイン	68	70	14	30	13	6	3	10	15	4	3	1
20	ロシア	32	42	18	19	12	6	2	19	12	6	4	3
21	その他	46	51	19	21	17	7	5	13	7	7	3	1
	総計	70	72	15	16	10	6	3	19	14	9	5	3

※単位：%

表-5 国籍・地域別個人手配率・平均訪問都道府県数

No.	国籍・地域	①15年 個人 手配率	②16年 個人 手配率	③ 変化量 (②-①)	④15年 訪問箇 所数	⑤16年 訪問箇 所数	⑥ 変化量 (⑤-④)
1	中国	42%	51%	9%	3.27	3.12	0.15
2	韓国	76%	83%	8%	1.81	1.73	0.09
3	台湾	44%	53%	9%	2.25	2.25	0.01
4	香港	59%	67%	8%	2.06	2.17	0.11
5	米国	93%	90%	-3%	2.86	2.86	0.01
6	タイ	69%	71%	2%	2.51	2.54	0.04
7	フィリピン	88%	90%	2%	2.46	2.35	0.11
8	オーストラリア	87%	87%	0%	3.55	3.71	0.16
9	シンガポール	85%	85%	0%	2.47	2.59	0.12
10	マレーシア	81%	73%	-8%	2.73	2.40	0.33
11	英国	88%	83%	-5%	3.12	3.11	0.02
12	フランス	90%	90%	0%	3.50	3.54	0.04
13	カナダ	88%	89%	1%	3.03	3.11	0.09
14	インドネシア	72%	79%	7%	3.07	2.82	0.26
15	ドイツ	90%	87%	-3%	3.00	3.22	0.22
16	ベトナム	72%	74%	2%	3.11	2.84	0.27
17	イタリア	86%	83%	-3%	3.59	3.93	0.33
18	インド	95%	94%	-1%	2.40	2.51	0.10
19	スペイン	89%	87%	-2%	4.09	4.11	0.03
20	ロシア	81%	87%	6%	2.70	2.70	0.00
21	その他	89%	91%	2%	3.32	3.51	0.20
	総計	63%	70%	7%	2.46	2.41	0.05

表5 から、「往復航空（船舶）券や宿泊等を個別に手配した」個別手配率(全サンプル)は平均 70%である。これ以外は、団体ツアー参加や往復航空（船舶）券と宿泊等がセットになった個人旅行向けパッケージ商品の購入であるが、アジア諸国で低いものの、16年には個人手配旅行が増加していることがわかる。

さて、「訪日外国人消費動向調査」では最大9箇所の問地を回答できる。それをもとにした都道府県別訪問率を上位6カ国・地域ごとに表6に示す。中国人では千葉、東京やゴールデンルートに含まれる山梨、長野、愛知が減少するのに対して、京都、奈良している。リピート率が最も増加しているため、来訪者属性による影響も考えられる。さらに、その他の傾向として京阪地域での増加、台湾、香港で沖縄が、韓国で九州が選好されるなど旅行者属性と訪問地との関連性を類推することができる。

表-6 都道府県別訪問率(2016年、単位：%)

	中国		韓国		台湾		香港		米国		タイ		全体	
	訪問率	増減	訪問率	増減	訪問率	増減	訪問率	増減	訪問率	増減	訪問率	増減	訪問率	増減
北海道	7	1	6	1	11	1	9	1	5	1	12	0	8	0
青森県	0	0	0	0	1	0	0	0	1	0	0	0	0	0
岩手県	0	0	0	0	1	0	0	0	0	0	0	0	0	0
宮城県	1	0	0	0	2	1	0	0	1	0	1	0	1	0
秋田県	0	0	0	0	1	0	1	0	0	0	1	0	0	0
山形県	0	0	0	0	1	0	0	0	0	0	0	0	0	0
福島県	0	0	0	0	0	0	0	0	0	0	0	0	0	0
茨城県	1	0	0	0	1	0	1	0	1	0	3	1	1	0
栃木県	1	0	0	0	2	1	1	0	1	2	7	1	2	0
群馬県	0	0	0	0	1	0	1	1	1	0	2	1	1	0
埼玉県	1	0	1	0	2	1	1	1	1	0	2	1	1	0
千葉県	42	6	16	0	30	2	28	1	75	1	50	0	38	3
東京都	64	4	24	2	34	2	34	1	75	3	55	1	49	2
神奈川県	16	4	3	1	6	1	4	1	14	3	10	0	9	2
新潟県	1	0	0	0	1	0	0	0	1	0	0	0	1	0
富山県	0	0	0	0	3	1	2	0	0	0	1	0	1	0
石川県	1	0	0	0	4	0	2	0	2	0	1	1	2	0
福井県	0	0	0	0	0	0	1	0	0	0	0	0	0	0
山梨県	13	2	0	3	0	1	1	2	0	9	2	5	1	0
長野県	1	0	0	0	6	0	3	0	2	0	3	1	3	0
岐阜県	2	0	0	0	4	1	4	0	2	0	5	1	3	0
静岡県	10	2	1	0	2	1	3	0	3	0	6	0	5	1
愛知県	20	3	3	1	6	0	8	0	5	0	8	0	9	1
三重県	1	1	0	0	1	0	1	0	0	0	1	0	1	0
滋賀県	0	0	0	0	1	0	0	0	0	0	1	1	1	0
京都府	39	2	19	3	20	2	22	3	25	2	17	0	27	2
大阪府	55	2	35	3	31	1	36	1	23	1	32	3	38	0
兵庫県	5	1	6	1	8	1	8	0	3	1	9	2	6	1
奈良県	10	2	4	0	8	2	7	2	3	0	6	1	7	1
和歌山県	1	0	0	0	1	0	3	0	1	0	1	0	1	0
鳥取県	0	0	0	0	1	0	1	0	0	0	0	0	0	0
島根県	0	0	0	0	0	0	0	0	0	0	0	0	0	0
岡山県	0	0	0	0	1	0	2	1	1	0	0	0	1	0
広島県	1	0	1	0	2	0	2	1	7	1	2	0	3	1
山口県	0	0	2	1	1	0	1	1	1	0	1	0	1	0
徳島県	0	0	0	0	0	0	0	1	0	0	0	0	0	0
香川県	1	0	0	0	1	0	0	1	1	0	1	0	1	0
愛媛県	0	0	0	0	1	0	0	0	0	0	0	0	0	0
高知県	0	0	0	0	0	0	0	0	0	0	0	0	0	0
福岡県	4	1	22	2	7	0	9	1	3	1	6	2	9	1
佐賀県	0	0	1	0	1	1	2	1	0	0	1	1	1	0
長崎県	1	0	4	1	2	0	2	0	2	0	2	0	2	0
熊本県	1	1	2	3	3	2	2	2	0	0	1	2	1	2
大分県	1	1	10	2	3	1	3	1	0	0	2	2	3	1
宮崎県	0	0	0	0	0	0	1	1	0	0	0	0	0	0
鹿児島県	0	0	1	0	1	0	2	1	1	0	0	0	1	0
沖縄県	5	0	9	1	11	1	11	1	4	0	0	0	7	0

4. 推定結果

拡大したサンプル約 4249 万人のデータを用いて、hPAM を用いたトピックの抽出を行った。SP 数, SB 数は任意に設定しながら、最終対数尤度、尤度比に基づき適当なトピック数の設定を行う必要があるため、表-6, 表-7 のように試行を行った。表-6 から、SP=2, SB=5 のケースが最も最終対数尤度が小さくなったが、これはトピック数が増加すると相対的に対数初期尤度が大きくなるためとも考えられる。そこで、モデルによる説明力の改善度合いを示す尤度比 (=1-最終尤度/初期尤度) を導出した (表-7)。その結果、SP=10, SB=20 のケースで最も尤度比が高くなった。

これらを踏まえて、適当なトピック数を決定する必要があるが、既存研究では最終尤度によって決定している事例が多かった。一方、表-6 (SP=2, SB=5) や表-7 (SP=10, SB=20) では、各トピックの解釈が容易でなく、同一トピック内に複数のトピックが混在していると考えられた。そこで、今回は、hPAM の特徴を把握することを意図して、既存研究からトピック数を (SP=20, SB=40) と設定して解釈を進めた。

表-6 モデルの最終対数尤度 (縦項目: SB 数)

	LDA モデル	hPAM SP:2	hPAM SP:3	hPAM SP:4	hPAM SP:5	hPAM SP:10	hPAM SP:15	hPAM SP:20	hPAM SP:25
5	-4.64	-4.25	-4.58	-4.74					
10	-4.72	-4.37	-4.64	-4.68	-4.80				
20	-5.10	-4.54	-4.73	-4.87	-5.01	-5.26			
30	-5.32	-4.59	-4.78	-4.82	-4.98	-5.35	-5.50		
40	-5.47	-4.72	-4.83	-4.93	-5.19	-5.36	-5.60	-5.67	
50	-5.60	-4.77	-4.87	-5.03	-5.11	-5.52	-5.67	-5.73	-5.81
60	-5.72	-4.80	-4.97	-5.05	-5.23	-5.60	-5.71	-5.83	
70	-5.82	-4.79	-4.92	-5.23	-5.22	-5.50	-5.65		
80	-5.86	-4.89	-4.96	-5.08	-5.40	-5.65	-5.79		
90	-5.94	-4.84	-5.06	-5.16	-5.47	-5.65			
100	-6.00	-4.94	-5.07	-5.18	-5.38				

表-7 モデルの尤度比 (縦項目: SB 数)

	LDA モデル	hPAM SP:2	hPAM SP:3	hPAM SP:4	hPAM SP:5	hPAM SP:10	hPAM SP:15	hPAM SP:20	hPAM SP:25
5	0.27	0.12	0.15	0.19					
10	0.33	0.17	0.18	0.20	0.24				
20	0.34	0.16	0.20	0.21	0.24	0.27			
30	0.35	0.16	0.20	0.22	0.24	0.25	0.23		
40	0.35	0.13	0.18	0.20	0.24	0.24	0.24	0.24	
50	0.35	0.13	0.19	0.20	0.23	0.23	0.22	0.22	0.24
60	0.35	0.12	0.16	0.20	0.22	0.21	0.21	0.23	
70	0.35	0.15	0.18	0.18	0.23	0.22	0.23		
80	0.36	0.12	0.19	0.22	0.21	0.21	0.21		
90	0.35	0.12	0.17	0.19	0.20	0.22			
100	0.35	0.12	0.18	0.21	0.21				

表-8 スーパートピック:北部九州のサブトピック分布ならびに主要訪問地

クラス	割合	訪問地1	訪問地2	訪問地3	訪問地4	訪問地5	訪問地6
SP0*那覇	0%	—	—	—	—	—	—
SB28福岡市	32%	福岡市	別府	由布院	阿蘇	熊本市	太宰府
SB36博多	28%	博多	天神	福岡市	由布院	太宰府	キャナルシティ
SB4福岡市	28%	福岡市	長崎市	熊本市	由布院	博多	別府
SB22福岡市	9%	福岡市	下関	小倉	広島市	博多	門司

次に、表-8 に SP トピック:北部九州の SB トピック分布ならびに主要訪問地を示す。個々のトピックには、それと関連が深い訪問地が導かれるため、それを用いてトピックの名称を付与した。これを用いて、Root ならびに SP トピック下の分岐率を示したものが表-9 である。尤度の最大化を通じて、類似した SB トピックが SP トピックの下に分岐する形式となっており、例えば SP1: 北海道に強く関係するものとして、SB12 (訪問地: 札幌ほか), SB33 (訪問地: 札幌ほか) に加え、SP1\*がある。トピック名称と同一で \* がついているものは、hPAM 特有のトピックに直接関連づいた訪問地である。また、いくつかの SB トピックが複数の SP トピックで出現しており、図-1 の DAG で示したような SB トピックに複数の SP トピックからのパスの存在が確認できる。逆に、対応関係が複数あるために、それを明瞭に示すことが困難であるといえる。そこで、構造を表現するためにネットワーク解析ソフトウェアを利用した。

表-9 スーパートピック, サブトピックの構成比率

スーパートピック	名称	割合	割合	サブトピック1	サブトピック2	サブトピック3
SP1	北海道	7%	7%	SP1* 札幌	SB12 札幌	SB33 札幌
SP7	関東以北	3%	3%	SB5 東京(都)	SB9 東京	SP7* 東京
SP9	都区内1	6%	25%	SB2 新宿	SB1 新宿	SB24 新宿
SP14	都区内2	6%		SP14* 浅草	SB2 新宿	SB1 新宿
SP18	都区内3	3%		SB0 新宿	SB24 新宿	SB23 成田
SP19	都区内4	7%		SP19* 新宿	SB2 新宿	SB1 新宿
SP8	東京近郊	4%	SP8* 軽井沢	SB9 東京	SB8 新宿	
SP2	京阪奈神1	4%	18%	SB30 大阪市	SB38 大阪市	SB7 大阪市
SP3	京阪奈神2	6%		SP3* 京都(府)	SB35 大阪市	SB7 大阪市
SP5	京阪奈神3	4%		SP5* 広島市	SB30 大阪市	SB37 大阪市
SP11	京阪奈神4	4%	SB38 大阪市	SB7 大阪市	SB37 大阪市	
SP12	GR1	3%	28%	SP12* 新宿	SB23 成田	SB13 広島市
SP13	GR,昇龍道1	2%		SP13* 沖縄(県)	SB31 東京	SB19 東京(都)
SP15	GR,昇龍道2	6%	SB17 東京	SB26 大阪(府)	SB15 名古屋	
SP16	GR2	9%	SP16* 大阪(府)	SB26 大阪(府)	SB17 東京	
SP17	GR3	5%	SB31 東京	SB19 東京(都)	SB18 東京(都)	
SP6	GR4	5%	SB31 東京	SB18 東京(都)	SB19 東京(都)	
SP0	北部九州	7%	9%	SB28 福岡市	SB36 博多	SB4 福岡市
SP10	九州	2%		SP10* 福岡(県)	SB28 福岡市	SB4 福岡市
SP4	沖縄	7%	7%	SP4* 那覇	SB14 那覇	SB11 那覇
合計		97%	97%	—	—	—

※赤: 構成比率 ≥ 40%, 橙: ≥ 30%, 青: ≥ 20%



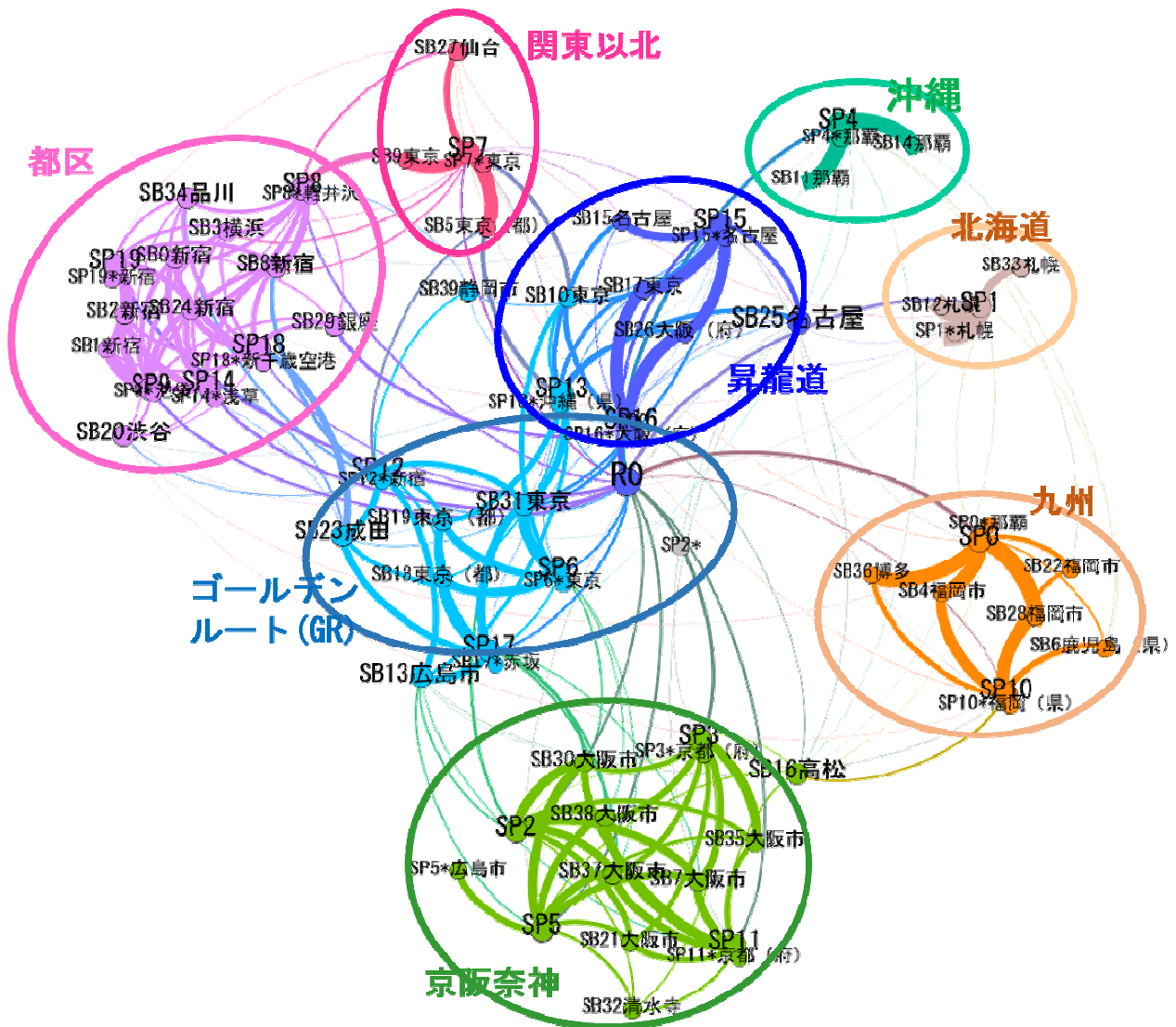


図-1 SP:20–SB:40のトピック関連性のDAG表現

Root から SP トピック，SP トピックから SB トピック への分岐確率をデータとして用いて，ソフトウェアは Gephi を利用した．グラフ可視化アルゴリズムに Force Atlas を用いて，各種パラメータは視認性が向上するように設定した．図中の線の太さが確率を表している一方，プロットの場合は，視認性が改善されるように任意に設定している．また，図中の色は，現在高速かつ精度が高い手法として知られている Louvain 法に基づいたモジュラリティ<sup>23, 24</sup>を実行して，上記データからコミュニティを検出している．

作成したグラフから，訪問パターンとして大きく 8 つに区分できること，その中の SP トピック，SB トピックの包含関係を視覚的に把握することができた．

## 5. まとめ

本研究は，訪日外国人旅行者の訪問パターンの特性把握を目的として，分析に用いた hPAM の概念整理や潜在クラスモデル，LDA モデルとの比較を行った．次に，「訪日外国人消費動向調査（平成 27，28 年）」データを用いて hPAM の推定を行った結果，1 地方における訪問パターンとして，都内・東京近郊が 25%と最も高く，京阪奈神が 18%，九州ならびに北海道が 9%，沖縄が 7%と続いていることがわかった．一方，複数地方の周遊として東京ー京阪を訪問するゴールデングルートが 20%，中部地方を南北に貫く昇龍道が 8%を占めることがわかった．

なお，今回の分析では，抽出された訪問パターン間の関連性を中心に着目したが，各パターンを占めるサンプルの特性把握を行う必要がある．旅行者の国籍・地域による特徴や訪日経験などの個人属性との関連を明らかに

して、リピーターの増加が見込まれる中で、訪問パターンの構成率の変化や、あらたなパターンが出現していないか、注視することも必要と考えられる。

今後の課題として、最新データへの適用、適切なトピック数導出のための指標の検討、より効果的なプロモーションの検討、効果的な観光振興策への落とし込みが考えられる。

#### 【参考文献】

- 1) 観光庁(2012,2013) (訪日外国人消費動向調査), <http://www.mlit.go.jp/kankocho/siryou/toukei/syouthityousa.html> (2016.2.6 閲覧)
- 2) 森地茂, 兵藤哲朗, 岡本直久 (1992), 時間軸を考慮した観光周遊行動に関する研究, 土木計画学研究・論文集, Vol.110, pp.63-70,
- 3) 国土交通省総合政策局 (2018), 地域のモビリティ確保の知恵袋 2017～訪日外国人旅行者の地方誘客を支える交通施策～, [http://www.mlit.go.jp/sogoseisaku/soukou/sogoseisaku\\_soukou\\_tk\\_000042.html](http://www.mlit.go.jp/sogoseisaku/soukou/sogoseisaku_soukou_tk_000042.html) (Viewed on 2018.4.30)
- 4) 劉瑜娟, 古屋秀樹 (2015) : 潜在クラス分析を用いた訪日外客の訪問パターンに関する基礎的分析, 第 52 回土木計画学研究発表会講演集(CD-ROM), No.52
- 5) 古屋秀樹, 劉瑜娟 (2015) : 訪日外客の 47 都道府県の訪問パターン分析, 日本観光研究学会第 30 年全国大会研究発表論文集 (CD-ROM)
- 6) 古屋秀樹, 劉瑜娟 (2016) : 潜在クラス分析を用いた訪日外国人旅行者の訪問パターン分析, 土木学会論文集 D3・特集号 (土木計画学研究・論文集), Vol.33
- 7) 佐藤一誠: トピックモデルによる統計的潜在意味解析, コロナ社, 2015
- 8) 岩田具治: トピックモデル, 講談社, 2015
- 9) Graham Neubig : 奈良先端科学技術大学院大学 HP (NLP Programming Tutorial 7-トピックモデル), <http://www.phontron.com/slides/nlp-programming-ja-07-topic.pdf>, 2016.2.16 閲覧
- 10) 石黒勝彦 (2013) : トピックモデルの応用: 相関・構造をもつトピックモデル, 統計推理研究所 H24 年度公開講座資料
- 11) 塚井誠人 (2018) : ベイジアンアプローチを使ってみる第 4 回「トピックモデル」, 交通工学, Vol.53, No.1, pp.46-53
- 12) 古屋秀樹 (2016) : 訪日外国人旅行者の地区レベル訪問パターンの基礎分析, 第 53 回土木計画学研究・講演集(CD-ROM), No.53
- 13) 古屋秀樹, 岡本直久, 野津直樹 (2017), (GPS ログデータを用いた訪日外国人旅行者の訪問パターンの分析手法の開発, 運輸政策研究, Vol.76
- 14) Blei and Lafferty(2007), A Correlated Topic Model of Sci-ence, The Annals of Applied Statistics, Vol. 1(1), pp. 17-35
- 15) Wei Li and Andrew McCallum(2006), Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, in Proc. ICML, <https://people.cs.umass.edu/~mccallum/papers/pam-icml06.pdf>(viewed on 2018.4.30)
- 16) David Mimno, Wei Li and Andrew (2007), Mixtures of Hierarchical Topics with Pachinko Allocation, in Proc. ICML, <https://people.cs.umass.edu/~mccallum/papers/pam08jmlrs.pdf> (Viewed on 2018.4.30)
- 17) 前掲 14
- 18) Hanna M. Wallach: Topic Modeling: Beyond Bag-of-Words. ICML 2006, <http://people.ee.duke.edu/~lcarin/icml2006.pdf> (Viewed on 2018.4.30)
- 19) 前掲 15
- 20) Machine Learning for Language Toolkit, <http://mallet.cs.umass.edu/>(viewed on 2018.4.30)
- 21) 前掲 7,8
- 22) 古屋秀樹: 出国空港別構成比率を考慮した訪日外国人旅行者の都道府県別訪問率の推定, 土木計画学研究・講演集, Vol.56, 2017年 11月
- 23) 飯田恭弘他 (2014), 大規模グラフ構造データからのコミュニティ抽出と重要度計算—高速化への取組みと応用—, 人工知能 29 巻 5 号, pp.472-479
- 24) 尾崎直人他 (2015), 大規模ネットワークにおけるコミュニティ抽出手法の改良, 人工知能学会全国大会論文集, No. 29, pp.1-4

(2018. 7. 31 受付)