所　　属	国際地域学研究科　国際観光学専攻　2年　3820161002番
氏　　名	宋　紫龍
学位の種類	修士（国際観光学）
学位論文題目	トピックモデルを用いた旅行記分析に関する研究 ―訪日中国人旅行者の旅行行動把握に向けて―
論文審査委員	主査　古屋　秀樹　　副査　飯嶋　好彦

論　文　要　旨

The purpose of this thesis is to analyze the travel behavior of Chinese inbound visitors to Japan by travelogue on SNS using Latent Dirichlet Allocation Model. In previous studies, GPS positioning data called “big data” was used for tourists’ behavior analysis. Therefore, there were limitations as to the analysis of the tourist activities, intension and evaluation in details. In this research, we divided 35,552 notes in the travelogue “Mafengwo” into morphemes. The morphemes of place names and keywords concerned with tourist activities and contents were only applied for identification of Chinese tourists’ activities and evaluation by Latent Dirichlet Allocation Model (LDA). LDA is one of the machine learning and natural language processing methods, from which tourist segments were derived from the similarities of the destinations and evaluations. From log-likelihood value which showed the degree of appropriateness, eighty four topics were generated as the most adequate number of topics. By considering the location information and the travel contents’ morpheme in the topics, there are typical travel patterns which are ”in-depth tour” and wide excursion pattern. The composition ratios of these were 70% and 8%. Furthermore, from the view point of travel contents, the top three of the travel contents are found out as ”history culture” (22%), ”wide excursion” (20%) and ”city tourism” (15%). Focusing on topics that related to the above regions, we extracted the visiting patterns in each prefecture by consider with the characteristics of the travel contents' morpheme.

Keywords：Latent Dirichlet Allocation Model, Topic model, travelogues, Travel behavior

キーワード：LDAモデル、トピックモデル、旅行記、旅行行動

1. 本研究の目的：日本のインバウンド観光は急速に拡大しており、その最も大きなターゲットは訪日中国人といえる。日本へのリピーター割合が増加していることから、その変化にともなう中国人旅行者のニーズや旅行行動をより詳細に把握する必要がある。訪問地点やそこでの活動、意向の把握を考えた場合、これまで利用されてきているGPSデータ以外の活用が必要不可欠といえる。そこで、訪問地点ならびに旅行行動・評価が記述された旅行記に着目し、中国で最大規模の旅行記サイト“Mafengwo”の文章データを用いて、類似した訪問地点・旅行行動の抽出を行う。その際に、数多くの形態素と多様な出現の組み合わせから合理的かつ論理的にその類似性に基づくセグメントの抽出手法として、トピックモデルの１つであるLatent Dirichlet Allocation Model（LDA）を採用した。以上から、本研究は、LDAによって抽出したトピック各々について、訪問地の組み合わせ（訪問パターン）や旅行行動を把握するとともに、特徴的な旅行者個人属性との関連性を明らかにして、旅行者と旅行行動との関連性を考察することを目的とする。

2. データ収集及びLDAモデル：SNSサイト“Mafengwo”から自動的にデータを収集できるWebクローラを用いて、2017年4月から2018年6月までに記述された35,552篇の旅行記を収集した。“Mafengwo”は、中国で最大規模の旅行記サイトであり、その記述数も多く時系列の変化も確認することができる。旅行記は文章データであるため形態素に分割を行い、地点名称ならびに旅行行動に関連形態素のみを抽出して（延べ形態素出現数12,585,972個）、LDAモデルによって推定した。LDAモデルは、機械学習・自然言語処理の教師データなしの1手法であり、1つの旅行記は必ずしも1つのトピッ

クに属するとは限らない条件下で、トピック別に形態素と出現頻度のペアの集合を類型化する方法である。トピックモデルの生成過程は以下となる。

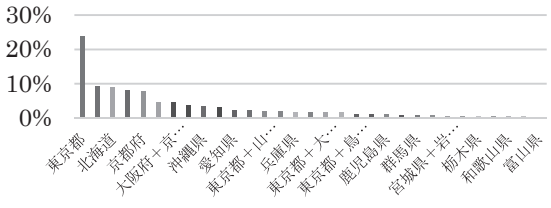
1.For　トピック　k＝1,・・・,K
　(a) 形態素分布を生成$\emptyset_k$～Dirichlet ($\boldsymbol{\beta}$)
2.For　文書　d＝1,・・・,D
　(a) トピック分布を生成$\boldsymbol{\theta}_d$～ Dirichlet ($\boldsymbol{\alpha}$)
　(b) For　形態素 n＝1,・・・,N_d
　　i. トピック生成z_{dn}～Categorical($\boldsymbol{\theta_d}$)
　　ii. 形態素を生成w_{dn}～ Categorical($\emptyset_{\boldsymbol{z_{dn}}}$)

3. 分析結果及び考察：LDA分析では、推定のフィッティングを示す対数尤度によって最適なトピック数が決定されるが、本分析では84トピックとなった（表－1)。

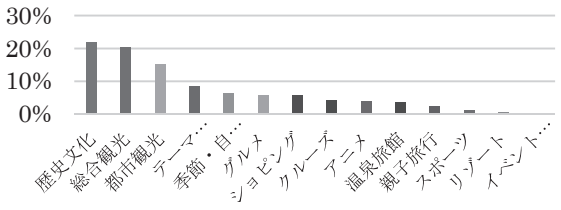
表－1　「トピック結果」及び「トピック集約」の例

トピック構成比率	形態素（15個）							
トピック1	民宿	行李	地铁	飞机	价格	日元	地铁站	排队
	3%	3%	3%	2%	1%	1%	1%	1%
10.21%	人民币	便利店	房东	现金	入境	工作人员	航班	
	1%	1%	1%	1%	1%	1%	1%	
トピック24	樱花	樱花季	樱花树	新宿御苑	目黑川	花期	开花	花瓣
	30%	4%	4%	2%	2%	1%	1%	1%
1.19%	粉色	造币	上野公园	季节	赏花	千鸟渊	野餐	
	1%	1%	1%	1%	1%	1%	1%	
トピック41	小豆岛	直岛	栗林公园	高松市	濑户内海	艺术	作品	丰岛
	6%	5%	4%	3%	3%	2%	2%	2%
0.57%	琴平	港口	金刀比罗宫	天使	小岛	橄榄	酱油	
	2%	2%	1%	1%	1%	1%	1%	

集約結果（例）	
訪問地	－
旅行内容	通常内容
訪問地	東京都
旅行内容	季節・自然風景
訪問地	香川県
旅行内容	総合観光



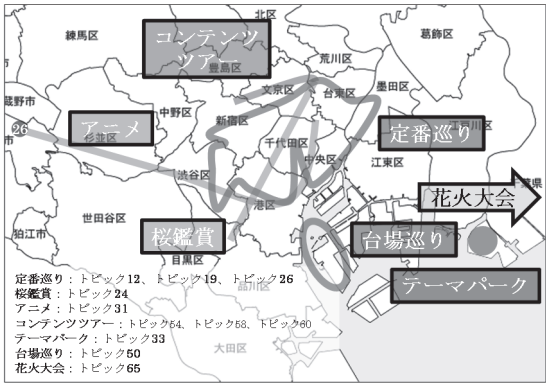
図－1　集約した訪問地割合



図－2　集約した旅行内容割合

上位トピックでは通常の旅行内容またはゴールデンルートでの伝統的な旅行に関連するトピックが抽出されたのに対して、下位トピックほどトピックの構成比率が減少し、地方でのニッチな旅行が多くなる傾向を示した。次に、84個のトピックは15の位置情報または旅行内容の情報を持つ形態素から構成されたため、「位置情報」と「旅行内容」に着目し、いくつかの大項目に集約することを行った。表－1は集約結果の一例である。

次に、「位置情報」を表す内容を「訪問地点」として都道府県ごとに集約した。訪問地の集約項目の割合（図－1）を考察すると、訪日中国人旅行者は、①訪問地割合の70%以上が１つの地域で数多くのスポットを周遊する「深度游」形態で旅行していること、②8%以上が地方を訪問し、広域ルートで旅行行動を行っていることが分かった。また、旅行内容を表す名詞の意味を考慮し、「旅行・観光消費動向調査（観光庁)」を参考にしながら、旅行内容の集約項目を名付け、15区分に集約した。旅行内容の集約項目の割合（図－2）を考察すると、旅行内容の上位3位とそれぞれの構成比率は、「歴史文化」: 22%、「総合観光」: 20%、「都市観光」: 15%となった。比較的構成比率の高かった「ショッピング」を旅行内容とするトピックの割合は6%となっており、「爆買い」の沈静化の現象であると考えられる。



図－3　東京都における旅行形態

表－2　東京都における旅行形態の特徴

旅行形態	居住地		
	北京市	上海市	東北
定番巡り	1.30%	0.91%	0.82%
桜鑑賞	-0.66%	-0.21%	-0.55%
アニメ	0.36%	-0.40%	-0.05%
コンテンツツアー	0.16%	-0.09%	2.20%
テーマパーク	0.75%	-0.29%	2.34%
台場巡り	0.16%	0.27%	-0.41%
花火大会	-0.27%	1.00%	-0.92%

旅行形態	居住地			
	華北	華中	華南	西部
定番巡り	-0.41%	-1.29%	-0.70%	-1.85%
桜鑑賞	-2.84%	-0.85%	-0.45%	0.29%
アニメ	-1.58%	-0.36%	0.20%	0.17%
コンテンツツアー	-2.06%	-0.77%	-0.75%	-0.17%
テーマパーク	-2.21%	0.26%	-1.46%	-0.35%
台場巡り	0.12%	-0.12%	-0.08%	-0.75%
花火大会	-1.72%	0.11%	-1.30%	-1.19%

更に、各深度游地域及び広域地域に関連するトピックに着目し、旅行形態の抽出を通じ、訪日中国人旅行者の旅行行動及び特徴の把握を行った。ここで、東京都を例として紹介する。図－3に示すように東京における7つ旅行形態を抽出した。位置情報と旅行内容の出現の組み合わせから、訪問パターン（本研究では「旅行形態」）を名付け、抽出した。例えば、「桜鑑賞」の旅行形態を抽出する際に、トピック24に注目したところ（表－1）、まず、位置情報を表す「新宿御苑（2.40%）」、「目黒川（1.70%）」、「上野公園（0.90%）」、「千鳥ヶ淵（0.80%）」から東京都への訪問が、「桜（38.80%）」といった旅行内容から、季節・自然風景のような桜鑑賞の主題を表すトピックと類推した。これより、上野公園→千鳥ヶ淵→目黒川→新宿御苑を立寄り場所とする「桜鑑賞」の旅行形態とした。以上の方法を用いながら、東京都における「定番巡り（関連トピックの構成比率：4.69%）」、「桜鑑賞（1.19%）」、「アニメ（0.83%）」、「コンテンツツアー（0.78%）」、「テーマパーク（0.78%）」、「台場巡り（0.38%）」、「花火大会（0.19%）」の7つ旅行形態を名付け、抽出した。また、トピックモデルの分析結果とする「各旅行記のトピック別構成比率」を使い、「特定属性（個人属性または旅程情報）に所属する各旅行記の特定旅行形態を主題とするトピック別構成比率」の平均値（A_1）と「特定旅行形態を主題とする各旅行記のトピック別構成比率」の平均値（A_2）の差（$A_1 - A_2$）を算出する。例えば、表－2の中の値は算出した平均値の差であり、カラースケールは濃ければ濃いほど、ある居住地に所属する全ての旅行記が特定旅行形態に対する内容の記述は平均より多い傾向があると判断することから、西部に住む人は桜鑑賞を行う傾向がある特徴を把握できた。以上の方法を用いながら、他の深度游地域及び広域地域の旅行形態の抽出及び特徴把握を行った。

4. 結論と今後の課題：旅行記に対するLDA分析によって、位置情報と旅行内容を考慮しながら、トピックに対する集約及び旅行形態の抽出を通じ、訪日中国人旅行者の旅行行動及び特徴を把握できた。今後はトピックモデルの派生モデルの適用を検討しながら、分析方法を精緻化し、データを増やして分析を行うことを課題とする。

参考文献

1) David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003), pp.993-1022, 2003
2) 古屋秀樹、岡本直久、野津直樹：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、2017
3) 宋紫龍、古屋秀樹：訪日中国人旅行者の旅行記を用いた旅行情報抽出方法の基礎的分析、日本観光研究学会全国大会学術論文集、32、pp.109-112、2017
4) 岩田具治：トピックモデル(機械学習プロフェッショナルシリーズ)、講談社、2015