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

A study of Chinese tourists' travel behavior by using data of travel literature

宋 紫龍*、古屋 秀樹**
SONG Zilong, FURUYA Hideki

本研究は、訪日中国人旅行者の書いた旅行記を用い、そこに記述された旅行情報の抽出方法について検討を行い、記述内容の類似度による旅行記の類型化と共に、それらと個人属性との関連性の明確化を目的とする。このような目的に対し、アンケートデータを用いるケースが多かったが、本論では Web 上の情報を自動的に収集する方法を採用し、旅行のコンテンツが含まれる 1 万編以上の旅行記を集めた。そして、教師データなしの機械学習の 1 つと位置付けられ、データの過学習を抑制できるトピックモデルを用いて分析を行った結果、48 個のトピック（パターン）に分類でき、それを基に訪日中国人旅行者の旅行実態と旅行嗜好を明らかにした。

キーワード：観光行動、類型化、旅行記、クローラ、トピックモデル

１． はじめに

日本の観光産業は、「地方創生」や「経済成長」のためのエンジンということができ、その中でもインバウンドが重要な役割を担う。そこでは、訪日外国人旅行者数に加えて、需要側である旅行者の旅行行動や嗜好を把握し、誘客のための効果的な施策をより詳細に検討することが重要といえる。その中でも訪日中国人旅行者は重要なターゲットと位置付けられ、訪日外国人の 26.5%、訪日外国人旅行消費額の 39.4%を占め（2016 年）、インバウンド旅行市場での大きなマーケットといえる。一方、「爆買い」の沈静化をはじめ、訪日中国人旅行者の旅行行動・嗜好が変化しつつあると言われており、このような中で、訪日中国人旅行者の嗜好や目的地を把握する必要があると考えられる [1]。

そこで、本論文は、Web 上の情報収集による訪日中国人旅行者の嗜好等を把握する分析方法を検討しながら、自然言語処理の方法の 1 つであるトピックモデルを用い、旅行記における訪問地や観光資源の記述の組み合わせ（記述パターン）に基づいた類型化を行うとともに、導かれた類型と個人属性との関連性を明らかにすることを目的とする。これらの分析を通じて、訪日中国人旅行者の嗜好が明らかになるとともに、個人属性ごとの満足度・再訪意向醸成に資する基礎的情報の取得が期待できる。

２． 先行研究と本研究の位置づけ

文献 [2] では、訪日外国人消費動向調査を用いながら、潜在クラスモデルによって外国人旅行者の訪問地を幾つかのパターンに分類した。さらに、文献 [3, 4] では、潜在クラス分析の強い仮定を緩和した一般モデルと位置付けできるトピックモデルを適用している。これらで用いられた分析データは位置情報データであったが、旅行行動の研究は位置情報にとどまらず、旅行の内容、即ちコンテンツも含まれることが望ましい。現時点まで、コンテンツを含める旅行行動に関する研究は十分に行われていないと考えられることから、本研究では旅行者の嗜好、評価等のデータを広範に収集する方法を検討するとともに、トピックモデルを用いて訪日中国人旅行者の類型化を行う。

３． 分析データとデータ収集手法

本研究では、多様な訪日中国人旅行者の嗜好、旅行に対する評価、訪問地を把握するが、その方法としてアンケート調査の利用が考えられる。しかしながら調査時期や被験者が限定的になるため、Web 上の旅行記に着目した。Web データは旅行記の執筆者に偏りを生じるおそれがあるため、より多数の旅行記を集めることによって、偏りを相対的に小さくすることを意図しながら、現在、中国における最大級の観光口コミサイトである「蚂蜂窝（マァファンウォー。以下、蚂蜂窝」」

＊ 東洋大学大学院国際地域学研究科国際観光学専攻 ＊＊東洋大学国際観光学科

を分析対象とした。「蚂蜂窝」では、それぞれの旅行者が記述した訪問先、旅行行動やその評価とともに、旅行記の題目、居住地、性別、出発時間、同行人物、滞在日数、旅行費用も収集できる。さらに、2013 年から、「蚂蜂窝」は中国旅游研究院に認められ、連携し、「全球自由行報告」をはじめ、毎年に観光に関する報告書を発行している。そのため、「蚂蜂窝」からのデータの真実性を保証できるだけではなく、説得力も高いと想定できる。以上から、作者の個人属性を含めた情報が豊富的に記録されているため、本研究にとって適当な分析データと判断した。

そして、保存されている Web 上のサーバから旅行記を効率的に獲得するために、クローラを使用した。クローラとは、ウェブ上のデータを周期的に取得し、自動的にデータベース化するプログラムであり、これをプログラミング言語（Python）で構築したことによって、訪日旅行記を 2017 年 4 月から 7 月中旬にかけて合計 10,606 編（訪日旅行の出発日時：2015 年 1 月～2017 年 7 月）収集し、これらを用いてトピックモデル分析を行った。

４． トピックモデル[5)、6)]

トピックモデルは自然言語処理方法で、教師データなしの機械学習の 1 つである。旅行記内での形態素（言語において、それ以上分解したら意味をなさなくなるところまで分割された最小単位）の出現の組み合わせから、尤度に基づき類似した旅行記を非排他的かつ明示された導出過程に基づきセグメント（トピックの割り当て）するものである。1 つの旅行記は必ずしも単一のトピックに帰属するとは限らない点を考慮しながら、トピック別に形態素と出現頻度とのペアの集合をモデル化する方法である。

さて、トピックモデル分析を行う前に、各旅行記を形態素に分割する必要がある。本研究は 10,606 編の旅行記の中で位置情報と旅行コンテンツの内容を表せる地名、観光スポット名及び名詞を対象として、形態素に分割した。その結果、地名・観光スポット名称 2,495 個ならびに名詞 96,386 個を取得し、これをトピックモデルに導入した。トピックモデルの生成過程は以下である。

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θ_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($\emptyset_{\boldsymbol{z_{dn}}}$)

トピック分布θ_dと形態素分布集合$\emptyset$が与えられた際の旅行記 $\boldsymbol{W}_d$の生起確率は、(1)式で示される。

$$\boldsymbol{p}(\boldsymbol{w_d}|\boldsymbol{\theta_d}, \Phi) = \prod_{\boldsymbol{n=1}}^{\boldsymbol{N_d}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{p}(\boldsymbol{z_{dn}} = \boldsymbol{k}|\boldsymbol{\theta_d})\, \boldsymbol{p}\,(\boldsymbol{w_{dn}}|\emptyset_{\boldsymbol{k}})$$

$$= \prod_{\boldsymbol{n=1}}^{\boldsymbol{N_d}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\theta_{dk}} \emptyset_{\boldsymbol{kw_{dn}}} \qquad \text{... (1)式}$$

上記によって、教師なしデータである旅行記から、出現の仕方が類似した形態素の組み合わせ（形態素パターン）を、あらかじめ設定したトピック数のもとで抽出することができる。これより、あるトピックのもとで、旅行記の中で頻繁に使用される形態素パターンが明らかとなり、その形態素から旅行者の嗜好する目的地、観光資源を類推することが可能となる。

さて、このトピック数は分析者が任意に設定した中で、モデルの説明力にもとづいて決定するのが一般的である。これを判断するために、分岐数または選択肢数を表し、確率の逆数で定義されているパープレキシティ（perplexity）を用いる。パープレキシティが小さいほど高い精度で予測できるモデルであることを示す。

本研究では、トピック数を 1~50 まで設定して各々20 回推定を行い、PPL の平均値を算出した。その結果、PPL は 48 トピックの時に最小となったため、以下の分析では 48 トピックの結果について考察を行う。

５． 分析結果と考察

10,606 編の旅行記を 48 トピックに類型化することができたが、上位 7 トピックならびに出現頻度上位 15 個の形態素を表－1 に示す。表－1 に示した各トピックの構成比率は、全文書におけるトピックの構成比率を示し、各形態素の下に示す数値は、各トピックにおける形態素の構成比率（各トピックが表している主題を判断するための参考値）である。

表－１　上位７トピックにおける頻出形態素

トピック	各トピック構成比率	単語（キーワード）				
トピック1	12.0% 通常の旅行コンテンツ	感覚	場所	時間	写真	民泊
		2.7%	2.6%	1.7%	1.1%	1.0%
		時間	旅程	荷物	親友	友達
		1.0%	0.9%	0.9%	0.8%	0.8%
		行列	味	天気	地図	計画
		0.7%	0.6%	0.6%	0.6%	0.5%
トピック2	10.8% 通常の旅行コンテンツ	円	空港	時間	交通	地下鉄
		3.3%	2.5%	2.4%	2.2%	2.1%
		価格	旅程	観光地	場所	アドバイス
		1.4%	1.3%	1.2%	1.2%	1.0%
		荷物	時間	航空券	民泊	宿泊
		1.0%	0.9%	0.8%	0.8%	0.8%
トピック3	8.8% 通常の旅行コンテンツ	ホテル	空港	時間	荷物	客室
		18.0%	3.8%	2.3%	1.9%	1.7%
		旅程	時間	飛行機	場所	朝食
		1.5%	1.4%	1.1%	1.1%	1.1%
		感覚	晩餐	価格	航空券	駅
		0.8%	0.8%	0.7%	0.7%	0.6%
トピック4	6.7% 購買行動	価格	ドラッグストア	空港	買い物	味
		1.4%	1.2%	1.1%	1.1%	0.9%
		百貨	百貨店	感覚	店	円
		0.8%	0.8%	0.8%	0.6%	0.6%
		服	元	場所	時間	ラーメン
		0.6%	0.6%	0.6%	0.6%	0.5%
トピック5	4.4% 歴史文化	神社	世界	観光者	文化	建物
		1.6%	1.4%	1.1%	1.0%	1.0%
		歴史	場所	街並	映画	伝統
		0.8%	0.7%	0.7%	0.7%	0.6%
		国家	体験	物語	鳥居	シーン
		0.6%	0.6%	0.5%	0.5%	0.5%
トピック6	3.9% グルメ	寿司	味	ラーメン	鰻	牛肉
		3.2%	2.6%	2.5%	2.0%	1.7%
		レストラン	定食	本店	グルメ	アドレス
		1.7%	1.6%	1.6%	1.5%	1.3%
		食感	河豚	円	味	食材
		1.2%	1.1%	1.0%	1.0%	0.9%
トピック7	3.9% ゴールデンルートの歴史文化	清水寺	稲荷伏見大社	金閣寺	二条城	八坂神社
		7.5%	5.7%	4.5%	3.8%	2.6%
		春日大社	神社	奈良公園	東大寺	花見小路
		2.0%	1.9%	1.6%	1.6%	1.5%
		稲荷	鴨川	鳥居	三年坂二年坂	建物
		1.4%	1.3%	1.2%	1.2%	1.0%

上位３位のトピックをみると、「ホテル」、「空港」、「時間」、「交通」といった形態素から構成され、「通常の（一般的な）旅行コンテンツ」であるといえる。次に、トピック 4～7 をみると、トピック 4：「価格」、「ドラッグストア」、「買物」から構成される「購買行動」トピック、トピック 5：「神社」、「文化」、「建物」から構成される「歴史文化」トピック、トピック 6：「寿司」、「味」、「ラーメン」から構成される「グルメ」トピック、トピック 7：京都と奈良における観光地名から構成される「ゴールデンルートの歴史文化」トピックと考えられる。

図－１　京都府における異なる旅行形式

また、上位トピック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が、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。

さらに、京都府に関連するトピックに着目し（トピック 7、13、36 が該当）、それぞれの頻出地点を示したものが図－１である。3 つぞれぞれの特徴を考えると、トピック 7：京都市の東側における古都巡り、トピック 13：嵐山から東福寺までの地域で紅葉鑑賞、トピック 36：宇治市で抹茶体験に分類でき、旅行記から京都府における訪問地点の差異を抽出できた。

次に、トピックモデルでは合計 48 トピックが抽出されたが、各旅行記がどのトピックに所属しているか帰属確率も同時に推定できる。そこで、トピックを集約しながら、旅行記トピックと各旅行記執筆者および旅行行程とのクロス集計を行った。トピック集約では、各トピックで頻出の位置情報（訪問地、観光スポット名）と旅行コンテンツに着目し、それらをいくつかの大項目に集約した。表－２は、その概要を示しているが、訪問地は「1.ゴールデンルート」、それに含まれる「2.特定都市」に加えて、3.北海道、4.九州、5.沖縄、6.その他地方（地方部）の 6 区分に集約した。また、コンテンツは、観光の実態と志向等を参考にしながら主要な旅行行動を 13 区分設定し、通常の観光コンテンツ、それ以外複数にわたるコンテンツの 15 項目に集約した。

表－２ トピックの主題項目

訪問地	コンテンツ
1.ゴールデンルート 2.特定都市 3.北海道 4.九州 5.沖縄 6.その他地方（地方部）	1. 通常の旅行コンテンツ 2. 購買行動 3. 歴史文化 4. グルメ 5. 大都市遊覧 6. 季節風景 7. テーマパーク 8. 親子旅行 9. 温泉旅館 10. アニメ 11. クルーズ 12. リゾート 13. イベント参加 14. スポーツ観戦 15. 総合的な観光

表－３　個人属性とのクロス集計結果

項目		性別		住所						
		男性	女性	北京	上海	東北	華北	華中	華南	西部
訪問場所	ゴールデンルート	-0.2%	0.2%	-0.4%	-2.0%	1.1%	0.8%	-0.8%	-0.4%	1.7%
	特定都市	-0.1%	0.1%	1.3%	0.5%	-2.0%	-0.4%	0.1%	1.2%	-0.6%
	北海道	-0.1%	0.1%	0.2%	0.0%	-0.4%	-0.1%	-0.2%	0.6%	-0.1%
	沖縄	0.0%	0.0%	0.1%	0.1%	-0.1%	0.3%	-0.1%	-0.2%	-0.2%
	九州	0.0%	0.0%	0.1%	0.5%	0.0%	-0.2%	-0.1%	0.0%	-0.3%
	その他地方（地方部）	-0.2%	0.2%	0.2%	2.6%	0.7%	-0.8%	-0.9%	-1.0%	-0.8%
コンテンツ	歴史文化	-0.8%	0.8%	0.3%	-0.8%	-0.1%	-0.5%	-1.1%	0.4%	1.8%
	総合的観光	-0.6%	0.6%	0.0%	1.4%	1.5%	0.2%	-1.5%	-0.6%	-1.0%
	大都市遊覧	0.1%	-0.1%	0.7%	-0.1%	0.1%	-0.9%	-0.4%	0.3%	0.2%
	購買行動	0.4%	-0.4%	-0.4%	-1.3%	1.5%	2.1%	-0.2%	-1.0%	-0.7%
	グルメ	0.1%	-0.1%	0.6%	1.0%	-0.4%	-0.3%	0.0%	0.1%	-1.0%
	季節風景	0.0%	0.0%	0.0%	0.1%	-0.8%	-0.4%	-0.4%	1.1%	0.4%
	テーマパーク	0.0%	0.0%	0.7%	-0.3%	0.0%	0.7%	0.3%	-0.9%	-0.5%
	アニメ	0.0%	0.0%	0.3%	0.0%	0.2%	-0.2%	-0.4%	0.0%	0.2%
	親子旅行	0.2%	-0.2%	0.8%	-0.1%	1.1%	0.3%	-0.2%	-1.0%	-0.7%
	クルーズ	0.0%	0.0%	-0.6%	0.4%	0.5%	0.5%	0.7%	-0.7%	-0.8%
	温泉旅館	0.0%	0.0%	0.0%	0.4%	-0.6%	-0.5%	0.0%	0.7%	-0.1%
	リゾート	0.0%	0.0%	0.1%	0.1%	-0.1%	0.3%	-0.1%	-0.2%	-0.2%
	イベント参加	0.0%	0.0%	0.0%	0.1%	-0.1%	-0.1%	0.0%	-0.1%	0.1%
	スポーツ、コンサート	0.0%	0.0%	0.1%	0.0%	-0.1%	-0.1%	0.0%	0.1%	0.0%

表－３は、個人属性と旅行嗜好とのクロス集計結果である。トピックモデルの結果とする「各旅行記のトピック別構成比率」を使い、「特定属性に所属する各旅行記の特定項目を主題とするトピック別構成比率」の平均値：P_1と「特定項目を主題とする各旅行記のトピック別構成比率」の平均値：P_2を算出でき、表－３の数値は（P_1 －P_2）を示す。カラースケールは濃ければ濃いほど、所属している項目の嗜好が強いといえる。

表－３より、性別では、男性は「大都市遊覧、購買行動」の嗜好があり、女性が特定都市で「歴史文化や総合的観光」の嗜好があることがわかる。また、居住地では、北京居住者は「特定都市＋大都市遊覧、親子旅行」の嗜好があり、上海居住者は「地方部＋総合的観光、グルメ」の嗜好があることが分かった。一方、中国の地方部をみると、東北と華北居住者は「ゴールデンルート＋購買行動」の嗜好が、西部居住者は、「購買行動」を行わず、「歴史文化」の嗜好が、華南居住者は「特定都市や北海道＋季節風景、温泉旅行」の嗜好があることが分かった。また、中国の沿岸部居住者はクルーズに参加する割合が高いことが明らかになった。

他のクロス集計結果として、春と夏の来訪者は、「特定都市＋季節風景」の嗜好があり、冬季来訪者は「北海道＋総合的観光」の嗜好があること、一人旅の旅行者は「地方部＋歴史文化、アニメ」を、家族の方は「ゴールデンルートへ伝統的な旅行行動」ならびに「クルーズ」への嗜好が高いことが明らかとなった。以上から、個人属性及び旅程情報別の訪日中国人旅行者の旅行嗜好を明らかにできたと考えることができる。

６．　まとめと今後の課題

本研究はクローラによる自動的な Web ページ上の旅行記の収集を行い、10,606 編の旅行記の中のおよそ 10 万種類・280 万個の旅行情報に関する形態素を抽出し、教師データなしの機械学習の 1 つと位置付けられ、データの過学習を抑制できるトピックモデルを用いて分析を行った。その結果、48 個のトピック（パターン）に分類でき、それをもとに訪日中国人旅行者の旅行実態と旅行嗜好を明らかにできた。

今後の課題として、さらにデータ量を増やした分析を行うこと、トピック数の縮約方法の検討、他言語を用いた分析を行うとともに、比較検証を行い旅行者の国籍・地域による差異の把握が考えられる。

謝辞：本研究の分析にあたり、王旭氏（東京工科大学）に協力を頂いた。ここに深謝の意を表します。

【参考文献】

1) 日比野直彦・森地茂・島田貴子（2011）：居住地域別訪日中国人旅行者の日本国内における観光行動 －インバウンド戦略検討のため基礎的分析－、『交通学研究』第 54 巻、p55-64
2) 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、Vol. 72、No. 5、p. I_571-I_583
3) 古屋秀樹（2016B）：トピックモデルによる訪日外国人旅行者の訪問パターンの基礎分析、第 53 回土木計画学研究発表会講演集，No.53
4) 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
5) 佐藤一誠（2015）：トピックモデルによる統計的潜在意味解析、コロナ社
6) 岩田具治（2015）：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