Vol.2020-EIP-87 No.11
2020/2/14

Q&A コミュニティにおける質問文からの観光情報の分析

吉見 憲二[†1]

概要：本研究では，日本最大の Q&A コミュニティである Yahoo!知恵袋の国内観光カテゴリを対象に，質問文から観光情報の分析を行った．「京都」「北海道」「沖縄」を対象とした計量テキスト分析の結果から，観光客の関心や懸念事項が明らかとなった．Yahoo!知恵袋の投稿データは，国立情報学研究所の IDR データセット提供サービスによって提供されているため，位置情報等が含まれるセンシティブなデータに比べて扱いやすく，カテゴリによって投稿者の目的がはっきりしているというメリットがある．

キーワード：Q&A コミュニティ，観光情報，Yahoo!知恵袋，計量テキスト分析

1. はじめに

ソーシャルメディアの普及に伴い，さまざまな分野でソーシャルメディアの投稿データの分析が行われている．特に，観光分野では，観光客の潜在的なニーズの把握のためにソーシャルリスニングに期待される役割は大きい．しかしながら，ソーシャルメディアの投稿はノイズを含むため，精度の高いデータを得るためには多くの工夫が必要である．他方で，Q&A コミュニティでは質問がカテゴリ化されているため，投稿に関わるノイズを低減することができる．本研究では，日本最大の Q&A コミュニティである Yahoo!知恵袋の国内観光カテゴリを対象に，質問文から観光情報の分析を行うことを目的とする．

2. 先行研究

2.1 ソーシャルメディアを用いた観光情報の分析

ソーシャルメディアの普及以前より，ブログ等の Web コミュニティから観光情報を抽出する試みは行われてきた[1]. スマートフォンの普及により，観光情報分析に位置情報や SNS の投稿が利用できるようになると，より高度な分析ができるようになってきている[2]. しかしながら，位置情報データに関しては，その利用者が必ずしも多くはないという問題がある．個別に依頼する場合にも，プライバシーや同意の取得の問題があり，その利用は容易ではない．さらに，SNS の投稿にはノイズが多く含まれるため，精度の高いデータを得るためにはさまざまな工夫が必要となる[3]．近年では，Instagram のように画像や動画を中心としたソーシャルメディアも増えてきているため，対象のメディアに合致した分析手法を考慮しなければならないという問題もある[4].

2.2 Q&A コミュニティの質問文を用いた分析

Q&A コミュニティの質問文を用いた研究として，Yahoo!知恵袋に投稿された質問のうち「赤ちゃん」and「寝かしつけ」を検索ワードとしてヒットした母親の質問 253 件を対象としたものがある．当該の研究では，特徴的なキーワードや投稿者の悩みを質問文から明らかにしている[5]．育児の悩みはソーシャルメディアでも大量に投稿されているが，Q&A コミュニティの質問文を用いることによって，より投稿者の悩みがはっきりした精度の高いデータを取得することに成功している．

2.3 問題意識

前述のように，ソーシャルメディアを用いた観光情報の分析には大きな可能性がある．しかしながら，精度が高いデータを利用するためには様々な困難があり，一般の研究者が分析を行うためには超えるべきハードルが多い．他方で，日本最大の Q&A コミュニティである Yahoo!知恵袋の投稿データは国立情報学研究所の IDR データセット提供サービスによって提供されており，研究者による利用は比較的容易となっている．

本研究では，こうした背景から Yahoo!知恵袋の投稿データを計量テキスト分析の手法を用いて分析し，Q&A コミュニティの質問文から有用な知見を引き出すことを目的とする．

3. 分析内容

3.1 分析に用いるデータ

本研究では，国立情報学研究所が提供する「Yahoo!知恵袋データ(第 3 版)」の 2019 年度提供版における質問文を利用した[6]．Yahoo!知恵袋は日本最大の Q&A コミュニティであり，多くのユーザーから質問

†1 佛教大学 Bukkyo University

と回答が寄せられている．当該データにおける質問数は約 264 万件, 回答数は約 611 万件となっている．ただし，収録データは収録期間（2014 年 4 月 1 日－2017 年 3 月 31 日）に投稿され解決した質問の 10%がランダムサンプリングされたものである．

Q&A コミュニティにおける質問文は，ソーシャルメディアの投稿に比べて，カテゴリ化されている点や投稿者の問題意識がはっきり表れている点で精度が高いデータとなっていることが期待できる．分析に当たっては, 提供されたデータから「>地域, 旅行, お出かけ>国内>観光地, 行楽地」カテゴリの投稿のみを抽出した. 対象となった投稿は 16,557 件となった.

3.2 分析対象

分析対象のエリアを設定するために，質問文に含まれる都道府県の言及数（1 投稿当たりの重複を含まない）を表 1 にまとめている．ここでは京都府や東京都が上位に来ている．この結果を各年の都道府県別延べ宿泊者数（表 2）と比較すると，3 つの都道府県を除いて一致が見られた. なお, 本来であれば日帰りの観光客数も含まれる「観光入込客統計」を用いることが望ましいが, 大阪府が未導入のため「宿泊旅行統計調査」の都道府県別延べ宿泊者数の数字を用いている．

本研究では，特に観光目的での来訪が多いと考えられる京都府, 北海道, 沖縄県を対象に計量テキスト分析の手法を用いて，Q&A コミュニティから得られる観光情報の分析を行った．

表 1　都道府県別言及数

都道府県	2014	2015	2016	2017	合計
京都府	576	562	479	81	1698
東京都	525	457	404	80	1466
大阪府	432	418	318	69	1237
沖縄県	283	278	252	46	859
北海道	233	265	215	40	753
広島県	119	142	111	21	393
福岡県	139	132	103	19	393
長野県	86	87	92	18	283
長崎県	75	100	74	18	267
愛知県	96	74	79	16	265

表 2　都道府県別延べ宿泊者数

順位	2014	2015	2016	2017
1 位	**東京都**	**東京都**	**東京都**	**東京都**
2 位	**北海道**	**北海道**	**北海道**	**北海道**
3 位	**大阪府**	**大阪府**	**大阪府**	**大阪府**
4 位	千葉県	千葉県	千葉県	千葉県
5 位	静岡県	静岡県	静岡県	**沖縄県**
6 位	**沖縄県**	**沖縄県**	**沖縄県**	静岡県
7 位	**京都府**	**長野県**	神奈川県	神奈川県
8 位	**長野県**	神奈川県	**長野県**	**京都府**
9 位	神奈川県	**京都府**	**京都府**	**長野県**
10 位	**福岡県**	**愛知県**	**愛知県**	**愛知県**

出典：「宿泊旅行統計調査」（国土交通省）

3.3 分析方法

分析に当たってはフリーのテキストマイニングソフトウェアである KH Coder（https://khcoder.net/）を使用した．形態素解析は付属の ChaSen（茶筌）を用いているが，一部観光施設名等で辞書に登録されていないものについては強制抽出の対象とした．

分析の手順として，まず経年的な全体の傾向を対応分析（コレスポンデンス分析）で可視化した. 次に, 共起ネットワーク分析を用いて抽出した語句の単語間の共起関係について示した．共起ネットワーク分析の描写に当たっては Jaccard 係数 0.1 以上の共起関係を基準とした. 最後に, 対象とした地名における共起頻度が高い観光用語について，頻出上位 10 語を抽出した．

4. 分析結果

4.1 対応分析

対象となる 16,557 件の質問投稿に関して，対応分析（コレスポンデンス分析）を行った．結果は図の通りである．横軸を示す成分 1 は 55.03%，縦軸を示す成分 2 は 32.21%の寄与率となった．成分 1 は経年的な変化，成分 2 は 2015 年の傾向をそれぞれ示していると解釈できる．

経年的な変化としては,「海」という単語が一番左に来ており，年々登場数が少なくなってきていることが示唆されている. これは昨今言われている「海離れ」という現象とも合致する．他には,「車」という単語がやや左に配置され,「レンタカー」や「バス」という単語が相対的に右に位置している．

成分 2 は 2015 年の傾向として,「札幌」が登場傾向の多い単語となっているが，その背景要因についてははっきりとしない．

Vol.2020-EIP-87 No.11
2020/2/14

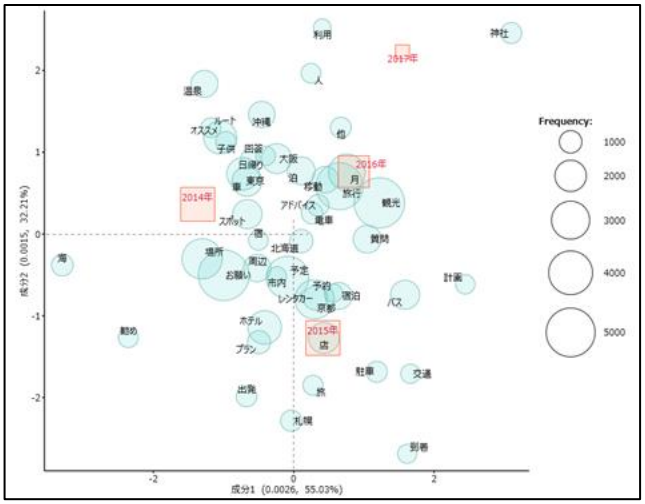

図 1　対応分析（コレスポンデンス分析）の結果

4.2 京都の観光情報

「京都」を含む関連語の共起ネットワーク分析の結果は図 2 の通りとなった．「金閣寺」「清水寺」「伏見稲荷」「嵐山」といった定番の観光地が他の多くの単語と共起しており，さまざまな質問文に用いられていることが読み取れる．同時に，「バス」「時間」「効率」といった単語も他の多くの単語と共起しているため，時間の制約がある中で定番の観光地をバスで巡る旅行者像が想起される．

他にも，「貴船神社」「下鴨神社」「八坂神社」といったパワースポットとして近年話題になっている神社や「銀閣寺」「二条城」といった世界遺産の共起関係が見られている．さらに，「奈良」は京都市内の主要な観光スポットとは共起関係がなく，「修学旅行」との共起関係が見られたことも興味深い点であった．

Jaccard 係数上位の観光用語は表 3 に示している．共起ネットワークと同様に主要な観光地の名称が登場しているほか，「バス」「大阪」「修学旅行」といった単語が「京都」と高い共起関係を見せている．

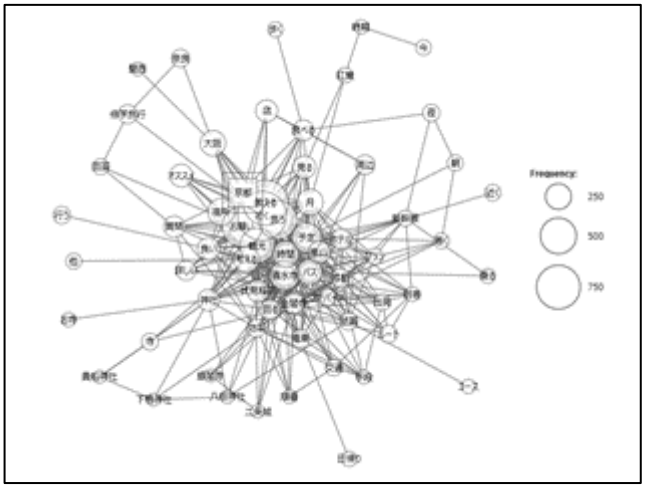

図 2　「京都」の共起ネットワーク

表 3　Jaccard 係数上位の観光用語（京都）

	観光用語	Jaccard 係数
1	清水寺	0.1853
2	嵐山	0.107
3	伏見稲荷	0.1046
4	観光	0.0985
5	金閣寺	0.0942
6	バス	0.0916
7	大阪	0.0892
8	神社	0.0831
9	祇園	0.0667
10	修学旅行	0.0667

4.3 北海道の観光情報

続いて「北海道」を含む関連語の共起ネットワーク分析の結果を図 3 に示している．「北海道」の共起ネットワークでは，「札幌」「小樽」「富良野」「旭川」「美瑛」といった地名が中心に登場していた．メジャーな観光地に比べて共起している単語の数は少ないが，「帯広」「釧路」「知床」「洞爺湖」「登別」といった地名も挙がっている．施設名としては，「新千歳空港」と「旭山動物園」が登場している．

「レンタカー」「移動」という単語も多くの単語と共起関係を有しており，レンタカーを用いてさまざまな地域を巡る旅行者像が想起される．他には，「美味しい」「海鮮」「店」という関係から食に関する興味もうかがえた．

Jaccard 係数上位の観光用語は表 4 に示している．共起ネットワークと同様に主要な地域の名称が登場しており，その中でも「札幌」は際立って高い共起関係を見せている．「レンタカー」が上位に登場するのも北海道の傾向であった．

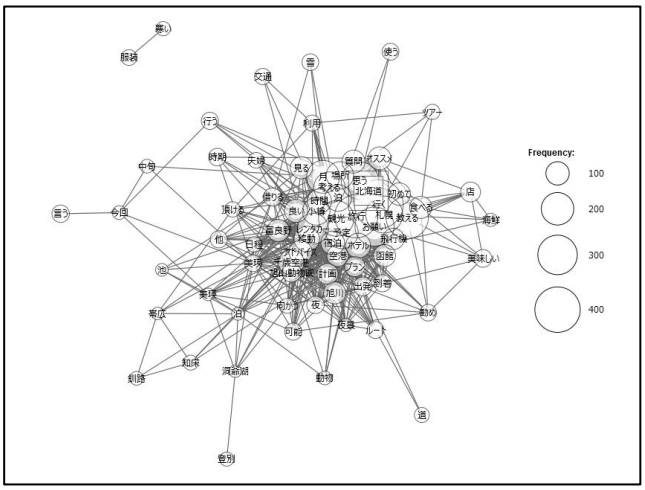

図 3　「北海道」の共起ネットワーク

表 4　Jaccard 係数上位の観光用語（北海道）

	観光用語	共起	Jaccard 係数
1	札幌	261 (0.347)	0.2605
2	小樽	142 (0.189)	0.1782
3	千歳空港	119 (0.158)	0.1522
4	レンタカー	178 (0.236)	0.1234
5	富良野	89 (0.118)	0.1137
6	旭川	89 (0.118)	0.1134
7	函館	88 (0.117)	0.1046
8	旅行	360 (0.478)	0.0912
9	旭山動物園	66 (0.088)	0.0853
10	飛行機	83 (0.110)	0.0801

4.4 沖縄の観光情報

最後に，「沖縄」を含む関連語の共起ネットワーク分析の結果を図 4 に示している．「沖縄」の共起ネットワークでは,「美ら海水族館」「国際通り」「首里城」「青の洞窟」という観光名所に加えて,「本島」「離島」「石垣島」「古宇利島」「恩納村」「海」「ビーチ」「リゾート」といった沖縄ならではの用語が多く見られた．「シュノーケリング」や「ダイビング」といったアクティビティが登場するのも沖縄の特徴となっている．

Jaccard 係数上位の観光用語は表 5 に示している．北海道と同様にレンタカーが上位に登場しており，移動に関する関心が高いことが読み取れる．加えて，他地域と比べて「ホテル」が上位に来ていることも特徴的であった．

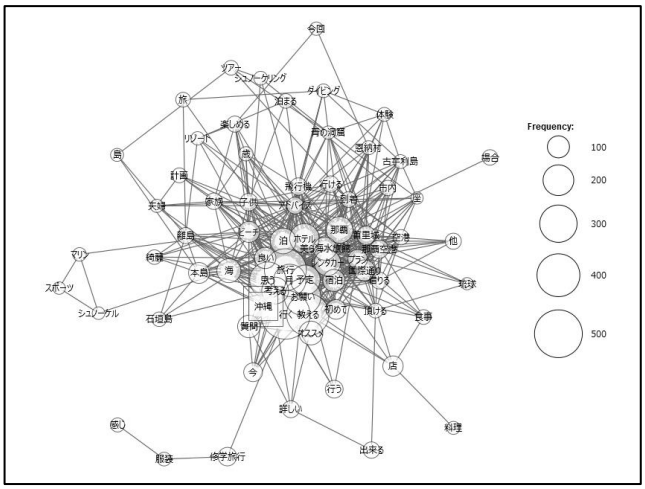

図 4　「沖縄」の共起ネットワーク

表 5　Jaccard 係数上位の観光用語（沖縄）

	観光用語	共起	Jaccard 係数
1	那覇	148 (0.172)	0.1689
2	美ら海水族館	112 (0.130)	0.1296
3	国際通り	90 (0.105)	0.1039
4	本島	89 (0.104)	0.103
5	ビーチ	90 (0.105)	0.0951
6	旅行	382 (0.445)	0.0947
7	海	119 (0.139)	0.0922
8	レンタカー	144 (0.168)	0.091
9	ホテル	182 (0.212)	0.0847
10	首里城	69 (0.080)	0.0797

5. おわりに

本研究では，国立情報学研究所が提供する「Yahoo! 知恵袋データ(第 3 版)」の 2019 年度提供版データを利用し，国内旅行カテゴリにおける質問文から観光情報の分析を行った．「京都」「北海道」「沖縄」を対象とした計量テキスト分析の結果から，それぞれの地域に特有の観光客の関心や懸念事項を明らかにすることができた．これらの結果は一般的な観光地に対する理解から大きく外れるものではないが，Q&A コミュニティの質問文から有用な知見を引き出させることを示した点で意義のあるものとなっている．

今後は，より直観に反するような意外な発見をするために，引き続きデータを分析していきたい．

謝辞

本研究は ROIS-DS-JOINT(課題番号:00032,研究代表者:小舘亮之)の助成を受けた.また,国立情報学研究所の IDR データセット提供サービスにより,ヤフー株式会社から提供の「Yahoo!知恵袋データ(第 3 版)」を利用した.

参考文献

[1] 斎藤一（2011）「Web における観光情報提供と分析」『人工知能学会誌』26 巻 3 号，pp.234-239.
[2] 相原健郎（2017）「ビッグデータを用いた観光動態把握とその活用：動体データで訪日外客の動きをとらえる」『情報管理』vol.59, No.11，pp.743-754.
[3] 渡邉小百合，吉野孝（2018）「観光地名なしツイートからの観光地に関する感想の抽出手法」『情報処理学会論文誌』Vol.59, No.1，pp.43-51.
[4] 田尾彩美，小舘亮之（2017）「ソーシャルメディアが観光に与える影響 : Instagram への投稿事例分析を中心として」『信学技報』vol. 116, no. 488， pp. 117-122.
[5] 佐々木裕子，高橋眞理（2015）「インターネットの Q&A コミュニティサイトにみる 0～4 ヵ月児の母親の育児における寝かしつけの悩み　テキストマイニングによる分析」『医療看護研究』11 巻 2 号，pp.28-35.
[6] 国立情報学研究所：“情報学研究データリポジトリ” (2019/1/10 閲覧), https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html