非負値テンソル因子分解を用いた観光行動データからの情報抽出

久保 基[1,2,4]　田中 宏季[1,2,3]　中村 哲[1,2,3]

概要：インバウンド観光を促進するためには，まず観光客の行動や滞在先を分析し，その傾向および要因を理解することが，必要不可欠であると言える．観光客の行動には様々な要因・情報が含まれており，それらを一挙に分析することは容易ではない．そこで，今回我々はスマートフォンアプリから得られた多次元の観光行動ログデータから，日本全国を対象とした外国人観光客の行動パターンを，滞在時間帯/滞在場所/推定居住国，の 3 階テンソルとして構成し，非負値テンソル因子分解 (Nonnegative Tensor Factorization, NTF) を適用することで分析を行った．結果として，3 階テンソルを 5 つのクラスタに分解することで，7 月と 8 月ではアメリカを中心とした訪日外国人観光客のクラスタ及びイタリアを中心とした訪日外国人観光客のクラスタを抽出することができた．

Information Extraction from Tourism Behavior Data using Non-negative Tensor Factorization

MOTOI KUBO[1,2,4]　HIROKI TANAKA[1,2,3]　SATOSHI NAKAMURA[1,2,3]

1. はじめに

インバウンド観光は，外国人が訪日する重要な観光産業の 1 つとして注目を浴びている．訪日外国人観光客の数は年々増加しており，これを受けて日本政府は新たな観光立国推進基本計画を策定するなど，2020 年までには外国人観光客の数を 4000 万人に到達させることを目標として掲げている．日本政府はこの目標を達成するために，様々なプロモーション施策を 3 カ年で設定しており，多様化する個人旅行ニーズに対応することやデジタルマーケティングを活用したビッグデータ分析を行うことが挙げられている．このように今後インバウンド観光を促進する上で，観光情報を予測・理解することは非常に重要であると考えられる．

また，観光庁の調査 [1] によると, 平成 30 年度に日本を訪れた外国人観光客の平均消費額は 153,029 円，その総消費額は 4 兆 5,189 億円であると報告されている．このような大規模な市場に対して適切にマーケティングを行うことでさらなる観光促進が望まれる．

観光客が旅行先で楽しむために，今日様々なサービスが提供されている．それらのサービスは大きく 2 つに大別され，インフラ整備などによるハードウェアサービスとコンテンツなどソフトウェアサービスに分かれる．前者は訪日観光客向けに提供されている無料無線 LAN サービス，後者はスマートフォンアプリなどを用いた観光地やグルメなどの情報提供などが該当する．ソフトウェアサービスを向上させるためには，観光客の嗜好やその行動を把握する必要があるが，そのための調査ではコストがかかることや，訪日観光客ならではのスポットの把握する必要があるという課題が存在する．そのため今回我々は，実際の訪日観光客の観光行動データを分析することにより，その課題を解決する方法を提案する．

既に述べたように，観光客の行動を分析するためにはその要因や情報を多角的に分析する必要があるため，これまで提案されてきている行動予測の手法では難しい．課題としては観光客が観光の際に行動決定する上で、観光データは多種多様な形をとっており ([2])，それらを包括的に考慮することは容易ではないことが考えられる．すなわち，

1　奈良先端科学技術大学院大学 先端科学技術研究科 情報科学領域
2　理化学研究所 革新知能統合研究センター 観光情報解析チーム
3　奈良先端大データ駆動型サイエンス創造センター
4　連絡先:kubo.motoi.kf2@is.naist.jp

それらのデータを包括的に扱えるようなデータ構造かつ分析手法を選ぶ必要がある．

そこで，本研究では様々な種類の観光行動データをテンソルと呼ばれるデータ構造を用いて表現し (図 1)，それに対して非負値テンソル因子分解 (Nonnegative Tensor Factorization, NTF) を適用することにより，有用な行動パターンやその傾向を抽出することを試みる．

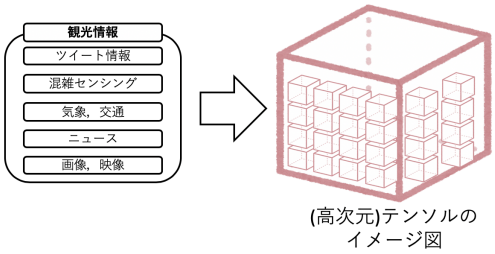

図 1　多様な観光情報とテンソル化

データ分析の手法には，ベクトルや行列の形式で表されているものに適用されていることが多い．しかし，多様な観光行動データを表現するには，ベクトルや行列形式では限界があるため，より高次元の表現が可能なテンソルを用いる必要がある．また，非負値テンソル因子分解は，ベクトルや行列では明らかにできない様々な要因の多項関係を明らかにすることができる手法である．よって今回はそのテンソル形式に適用できる分析手法として非負値テンソル因子分解を用いた．

本研究では，日本全国を対象にした世界各国から訪れた外国人観光客の観光行動ログデータに対し，非負値テンソル因子分解を適用しいくつかの因子間の関係から行動パターンやその抽出を行った．その結果，国ごとに異なる特徴のあるクラスタを抽出することができた．

本論文の構成を以下に述べる．2 節において，今回分析に用いた訪日外国人観光客の位置・属性情報についてその収集と基礎分析結果について説明する．3 節において，今回分析に用いた手法である非負値テンソル因子分解について述べる．4 節において，観光行動ログデータを 3 階テンソルで構成し，非負値テンソル因子分解を適用することによって得られたクラスタについて述べる．最後に，5 節では本研究のまとめを述べる．

2.　観光行動データ

表 1 は今回分析に使用した観光行動データ*の概要である．各項目について解説する．まず，対象日数については各月でスマートフォンアプリから得られた日時情報から対象となっている日数を示している．対象ログ数は，今回の分析においてインバウンドを対象としているため，国内に居住地を持たず国籍が日本ではないログのみを対象として算出している．

表 1　分析に使用したデータの概要

月/項目	対象日数	対象ログ数
2 月	28	109888
3 月	21	108892
4 月	30	418146
7 月	24	168657
8 月	20	217426
11 月	30	226367

表 2 は各月における訪日外国人観光客の国別ログ数を示している．インバウンド全体に対する割合は 8 月を除いてアメリカが最も高くなっている．日本政府観光局による調査 [10] では，2018 年度の訪日外国人観光客が 3119 万 2000 人に上ると報告しており，その中でも 7 割は東アジアの国々となっている．今回用いるデータでは，東アジアの国々ではなく，アメリカやイタリアといった国のユーザが多くを占めている．その理由として，スマートフォンアプリケーションの展開に地域差が出てしまっていることが挙げられる．今回用いたスマートフォンアプリケーションの配布状況が比較的，東アジアの国々よりも欧米諸国の方がより展開できているため，表 2 で示すような形となっている．

表 3 は各月における訪日外国人観光客が訪れた都道府県別ログ数を示している．各月で上位 3 位までは京都府，大阪府，東京都が独占しており，それ以外の順位で多少誤差で変わる程度であった．

3.　非負値テンソル因子分解 (NTF)

本研究では，1 章でも述べた通り，多次元の観光行動データをテンソルで扱うことでより多くの因子の関係を分析する．そこで，観光情報をテンソルと呼ばれるデータ構造を用いて多種多様な観光データを扱うことで，有用な情報を抽出することを考える．ここでテンソルとは，多次元の配列のことを表すものであり，スカラーは 0 階テンソル，ベクトルは 1 階テンソル，行列は 2 階テンソルということができる．また，テンソルで表現されたパターンの抽出手法の一つに非負値テンソル因子分解 [3] がある．この手法はソーシャルネットワーク分析 [6] や購買行動分析 [5]，観光行動分析 [4] に適応されるなど，多くの分野で適用されている．

3.1　NTF の定義

次項の図 2 に NTF の概念図を示す．

*　データ提供元:(株)Agoop

表 2　各月の訪日外国人観光客の国別ログ数

順位/月	2 月		3 月		4 月		7 月		8 月		11 月	
	国名	ログ数	国名	ログ数	国名	ログ数	国名	ログ数	国名	ログ数	国名	ログ数
1	アメリカ	24455	アメリカ	42588	アメリカ	109400	アメリカ	61455	イタリア	59148	アメリカ	74423
2	中国	9011	フランス	8074	フランス	53657	イタリア	13214	アメリカ	39833	フランス	16120
3	台湾	8837	イギリス	5806	イタリア	39423	スペイン	12803	フランス	33708	イギリス	15036
4	オーストラリア	7548	中国	5342	イギリス	30208	フランス	12658	スペイン	26308	オーストラリア	11948
5	ドイツ	7543	オーストラリア	4992	オーストラリア	29954	オーストラリア	12167	イギリス	7674	ドイツ	10218

表 3　各月の訪日外国人観光客が訪れた都道府県別ログ数

順位/月	2 月		3 月		4 月		7 月		8 月		11 月	
	都道府県名	ログ数	都道府県名	ログ数	都道府県名	ログ数	都道府県名	ログ数	都道府県名	ログ数	都道府県名	ログ数
1	京都	48532	京都	50476	京都	219678	京都	88228	京都	116294	京都	109606
2	大阪	16966	大阪	15429	大阪	43033	大阪	19914	大阪	20148	大阪	29739
3	東京	8395	東京	8797	東京	28506	東京	11544	東京	14464	東京	16018
4	静岡	5955	静岡	5068	静岡	16863	静岡	7202	奈良	8706	愛知	9856
5	愛知	5418	滋賀	4543	滋賀	15623	愛知	6415	広島	7666	静岡	9493

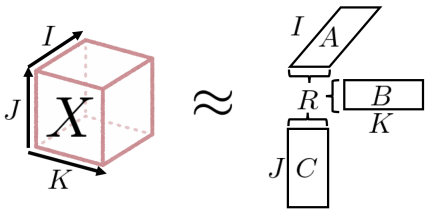

図 2　テンソル因子分解の概念図

図 2 に示す様に，$I \times J \times K$ の 3 階テンソルを考える．今回は例として 3 階テンソルを対象としているが，実際には 3 階以上のテンソルを扱うことも可能である．ここで，3 階テンソルは $X = [x_{ijk}] \in \mathbb{R}_+^{I \times J \times K}$ と書き表すことができる．ここで，I, J, K は各モード (=因子の軸) の要素数，すなわち値ラベルの種類の数を示している．この値は通常どういったモードで，どのような分析を行うかによって設定するパラメータである．また，$\mathbb{R}_+$ は非負の実数値である．NTF では，与えられたテンソルを次頁の式 (1)

$$\boldsymbol{X} \approx \boldsymbol{A} \otimes \boldsymbol{B} \otimes \boldsymbol{C} \tag{1}$$

のように 3 つの因子行列の内積で近似するように因子分解する．

本研究では，テンソル分解の代表的な手法である CANDECOMP/PARAFRAC decomposition [8, 9] を用いて，3 階テンソル $\boldsymbol{X}$ をランク数 R とした 3 つの行列 $\boldsymbol{A} = [a_{ir}] \in \mathbb{R}_+^{I \times R}$，$\boldsymbol{B} = [b_{ir}] \in \mathbb{R}_+^{J \times R}$，$\boldsymbol{C} = [c_{ir}] \in \mathbb{R}_+^{K \times R}$ の内積として分解を行う．ここでランク数 R は基底の数のことを指し，いくつのクラスタに分解するのかを決定するパラメータである．また，クラスタとは因子間の関係において同じような傾向であるものをまとめたものであり，一つの傾向を表す集合としている．

$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ はそれぞれ各要素の因子行列であり，それらを用いて以下の式 (2) のように

$$\hat{\boldsymbol{X}} = [\hat{x}_{ijk}] = \boldsymbol{A} \otimes \boldsymbol{B} \otimes \boldsymbol{C} \tag{2}$$

として得られるテンソル $\hat{\boldsymbol{X}}$ に対し，

$$D = \sum_{i}^{I} \sum_{j}^{J} \sum_{k}^{K} \mathcal{D}(x_{ijk}, \hat{x}_{ijk}) \tag{3}$$

を最小化するように，$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ を計算する．ここで，$\mathcal{D}(x_{ijk}, \hat{x}_{ijk})$ は距離関数を表しており，基本的にこの距離関数には KL ダイバージェンスやユークリッド距離が用いられる．本研究では，ログなどの数値の取りうる値が大きいデータに対して有効である一般化 KL ダイバージェンスを用いた．NTF はこれらの因子行列 $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ の各要素の非負性を保ちながら，D を最小にする $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ を求める手法である．

4. 分析結果

4.1 NTF による因子分解結果

ここでは，2 節で述べた位置情報，及び属性情報に対して NTF を適用した結果について説明する．入力として用いたデータ及びそのパラメータは以下に示す通りである．

入力テンソル　{ 滞在していた時間/滞在した場所 (都道府県単位)/推定居住国 } で構成される 3 階テンソル，テンソル内の要素は「ある時間にどの国に住んでいる人がどの都道府県にどれだけ滞在しているか」を表す．

表 4　図 3 及び図 4 における推定居住国 ID の対応表

7 月		8 月	
ラベル ID	都道府県名	ラベル ID	都道府県名
26	京都	26	京都
27	大阪	27	大阪
13	東京	13	東京
22	静岡	29	奈良
23	愛知	34	広島

表 5　図 3 及び図 4 における推定居住国 ID の対応表

7 月		8 月	
ラベル ID	国名	ラベル ID	国名
45	アメリカ	18	イタリア
18	イタリア	41	アメリカ
39	スペイン	11	フランス
11	フランス	35	スペイン
0	オーストラリア	40	イギリス
12	ドイツ	36	スイス
40	スイス	12	ドイツ
4	中央アフリカ共和国	37	台湾
44	イギリス	5	中国
41	台湾	4	中央アフリカ共和国

滞在していた時間は 24 時間幅で設定しており，滞在した場所は北海道から沖縄県まで日本全国を対象としている．また，推定居住国はそのログデータのユーザが住んでいる国を推定し，そのデータとなっている．

クラスタ数　5 クラスタ数は対象とするデータに存在する傾向をいくつに分類するかを決めるパラメータである．これは，大きすぎても小さすぎても抽出できる傾向などが異なってしまう．事前にパラメータの数を決定する必要あり，今回は計算時間の関係でクラスタ数はある一定時間内に分解できる最大の数とした．

今回は特に分解結果が顕著であった 7 月及び 8 月の因子分解結果について報告する．図はクラスタごとに色分けされており，滞在時間帯/滞在場所/推定居住地の関係をそれぞれ表している．それぞれのグラフにおいて，縦軸は，各因子のそれぞれの項目に対する出現頻度を表しており，この値が大きいほどそのクラスタ内での傾向が強く出ていると考えられる．また，横軸はラベルデータと表している．例えば，滞在時間帯のグラフの横軸は 24 時間幅であるため，0〜24 となっている．滞在場所については北海道から沖縄県まで，1 から 47 の都道府県コードが割り当てられている．表 4 に上位 5 位の都道府県のラベル ID を示す．

推定居住国は，出現順に番号が振られており，月によって対象ユーザの推定居住国が異なるため，ログ数の上位 10 カ国の国ラベル ID を表 5 に示す．

月ごとに登場する国が異なっているため，同じ国でも異なるラベル ID となっているので，こちらの表を参照しながら分析結果を説明する．

4.1.1　7 月の訪日外国人観光客の観光行動データにおける因子分解結果

図 3 に 7 月における訪日外国人観光客の観光行動データに NTF を適用した因子分解結果を示す．

7 月における訪日外国人観光客の観光行動では，推定居住国を中心に見るとクラスタ 1-4 とクラスタ 5 に分かれている．前者はアメリカを中心とする諸国を居住国とするユーザのクラスタになっており，クラスタ 5 は推定居住国がコロンビアを中心としたクラスタになっており，ユーザで異なるクラスタのグループが抽出されていることがわかる．

また，クラスタ 1-4 の中では滞在場所がクラスタごとによって異なっており，クラスタ 1 は主に京都に滞在したクラスタ，クラスタ 2 は京都や大阪を中心とした関西に滞在したクラスタであることが読み取れる．クラスタ 3 については，関西圏だけでなく東京を中心とした首都圏にも滞在したユーザの多いクラスタであることがわかる．特にクラスタ 1 では，各推定居住国のピークがいくつか見られるため，京都観光するユーザの行動時間帯などその傾向が現れていると言える．

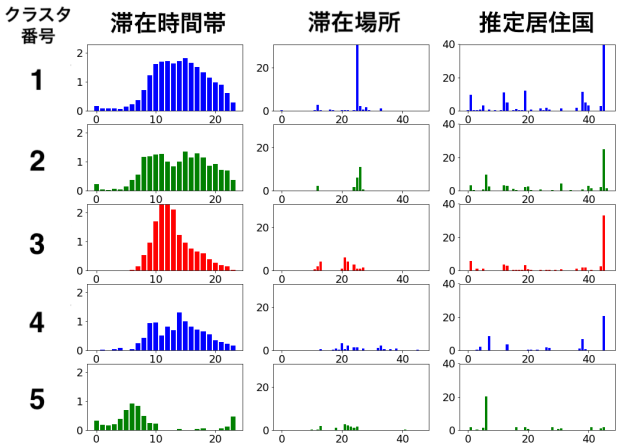

図 3　NTF を用いた因子分解結果 (7 月)

4.1.2　8 月の訪日外国人観光客の観光行動データにおける因子分解結果

次項の図 4 に 8 月における訪日外国人観光客の観光行動データに NTF を適用した因子分解結果を示す．

8 月における訪日外国人観光客の観光行動では，まずそれぞれのクラスタで推定居住国のピークが異なっている．クラスタ 1 ではフランス・イタリア・アメリカ・スペインがピークとなっており，この国々のユーザが京都を中心に観光したということがわかる．クラスタ 2 及びクラスタ 3 では，アメリカと中国及び中央アフリカ共和国がピークとなっており，これらの国々のユーザが関西圏を中心に観光

したことが読み取れる．クラスタ 4 ではアメリカのピークが最大で，関西圏だけでなく東海地方や首都圏にも観光したクラスタとなっている．クラスタ 5 は，主にイタリアがピークとなっており，東海地方から関西圏にかけて観光したクラスタとなっている．滞在時間帯を見ても，クラスタ 1 と同様の割合でユーザが滞在していることがわかる．8 月の観光行動分析では，ヨーロッパ圏のユーザが増えていることがわかり，イタリアのユーザは関西圏を観光したクラスタと関東圏や東海地方など広く観光したクラスタが抽出された．

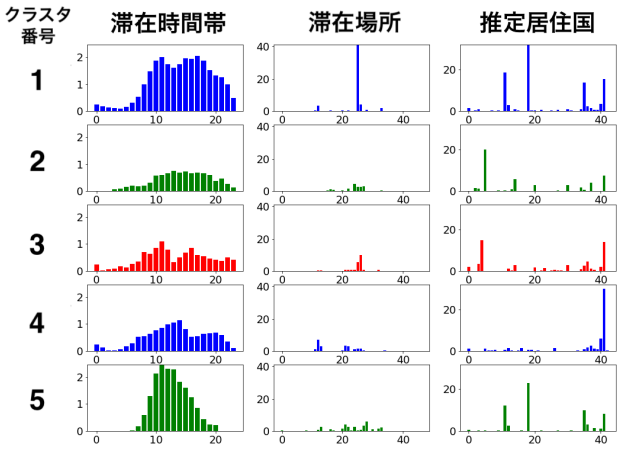

図 4　NTF を用いた因子分解結果 (8 月)

5.　まとめ

本研究において我々は，訪日外国人観光客のさらなる増加を見込んだインバウンド観光促進をするため，観光行動ログデータの分析を行った．今回我々は，訪日外国人観光客に配布されたスマートフォンアプリケーションより得られた，位置情報や時間情報及びユーザの属性情報が含まれる観光行動ログデータに対し，非負値テンソル因子分解を適用することによって，主にアメリカやイタリアなどを中心とした居住国のユーザのクラスタ及びその傾向を明らかにし，時間帯や滞在場所及び居住国ごとに異なる様々な行動パターンを抽出した．

今後の課題としては，今回の分析で扱えていない因子を含めた上で，どのような因子選びをすることでより詳しい情報を抽出することができるかを考えることで，より精度の高い分析が可能になると考える．また，今回は滞在地の対象が広くなったため，地域別に分けてユーザごとの観光ルートから非負値テンソル因子分解を適用することによって，対象ユーザがどのようなルートで観光したかを明らかにすることで，観光推薦などに応用することができると考えられる．

参考文献

[1] J. T. Agency. Consumption trend survey for foreigners visiting japan. Technical report, Ministry of Land, Infrastructure, Transport and Tourism, 2018.

[2] 観光庁:2018 2020 年度 訪日プロモーション全体方針, http://www.mlit.go.jp/common/001227859.pdf [Online] (2018).

[3] Shashua A., and Tamir H. Non-negative tensor factorization with applications to statistics adn computer vision. Proceedings of the 22nd International Conference on Machine Learning (ICML2005), ACM, 2005.

[4] 熊谷雄介, 今井良太, 松林達史, 佐藤吉秀, 堀岡力. 非負値複合テンソル因子分解を用いた訪日外国人観光客の回遊行動分析. IEICE, IEICE 信学技報, 2015.

[5] 松林達史, 幸島匡宏, 林亜紀, 澤田宏. 非負値テンソル因子分解を用いた購買行動におけるブランド選択分析. 人工知能学会論文誌 30 巻 6 号, pp713-720, 2015.

[6] Koji Hashimoto, Toni Iwasaki, Tetsuo Furukawa. Tensor Decomposition using Self-Organizing Map and Missing Data Estimation. IEICE, IEICE Technical Report, 2013.

[7] Koh Takeuchi, Hisashi Kashima, Naonori Ueda. Autoregressive Tensor Factorization for Spatio-temporal Predictions. JSAI, JSAI 大会論文集, 2018.

[8] Kolda, Tamara G., And Brett W. Bader. Tensor decompositions and applications. SIAM review 51.3 (2009): 455-500.

[9] Cichocki, Andrzej, et al. Nonnegative matrix and tensor factorizations:applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons (2009).

[10] 日本政府観光局 (JNTO):訪日外客統計 2018 年度推計値, https://www.jnto.go.jp/jpn/statistics/data_info_listing/pdf/190116_monthly.pdf [Online] (2018).