The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

ニュースサイト閲覧者の興味関心を用いた地域と観光要素の関連性抽出

Extraction of Relevance between Regions and Domestic Enjoyments based on News Site App Users' Interest

川口輝太 *1*2 Kota Kawaguchi　井上裁都 *2 Tatsukuni Inoue　長田誠也 *2 Seiya Osada　山下達雄 *2 Tatsuo Yamashita

*1筑波大学大学院 システム情報工学研究科 Graduate School of Systems and Information Engineering, University of Tsukuba

*2ヤフー株式会社 Yahoo Japan Corporation

We proposed a method to discover domestic enjoyments information that are specific and original to individual domestic area. The proposed method is based on article follow information of news site users and themes given to articles.We extracted the relationship between regional theme and sightseeing theme using co-occurrence information of themes in each article. In addition, We pulled out the relationship between regional theme and sightseeing theme using themes actively selected by the user for personalization. By filtering with similarities between those two types of relevance, accuracy improvement was achieved. Moreover, by focusing on sightseeing theme with low similarity, error candidates for theme linking can be extracted.

1. はじめに

地域活性化のために，我が国の重要な成長分野である観光に注目が集まってきている．日本政府観光局の資料によると，2018 年は年間の訪日外国人数が 3119 万人と過去最高を記録した [1]．産業の観光化を推進するにあたり，観光地として訴求すべき観光コンテンツを正確に把握することは，地域への観光誘致のために重要であり，種々の研究がおこなわれている [2] [3]．

本研究では，実際のユーザの行動データを使って，各地域と観光要素の紐付けを行い，地域ごとの観光要素のクラスタを抽出することを目的とする．分析対象として，ニュースサイト「Yahoo!ニュース」[4] を用いた．Yahoo!ニュースの閲覧者は，興味のある地域や娯楽のテーマをフォローすることで，そのテーマが紐付けられた記事を優先的に閲覧できる．そこで，ニュース閲覧者の興味対象のテーマと実際の記事に付与されているテーマを併用した，観光要素別の地域ランキングと地域別観光要素ランキングの作成手法を提案する．

2. Yahoo!ニュースとテーマ

本研究でのターゲットである，Yahoo!ニュースについて説明する．Yahoo!ニュースは，新聞・通信社が配信する報道をはじめ，個人により執筆された記事や話題の動画，写真などを，ウェブサイトとアプリを通じて，提供するサービスである [4]．Yahoo!ニュースで用いられる「テーマ」とは，サービス側が事前に定義した興味関心の単位である．記事には複数のテーマが自動的に付与される．例えば，青森ねぶた祭りに関する記事は，「祭りと伝統行事」，「青森市」，「青森ねぶた祭り」などのテーマが付与される．これにより，ユーザはテーマ別に記事を閲覧できる．ユーザが上記のテーマを選択することを「フォロー」と呼ぶ．ユーザが各テーマをフォローをすることで，そのテーマが紐付けられた記事を優先的に閲覧できる．図 1 にニュース閲覧者とテーマと記事の関係を示す．

3. 地域テーマと観光テーマの関連性

本章では，「地域テーマ」と「観光テーマ」の関連性について述べる．地域テーマとは各都道府県と市区町村のテーマ，観光テーマとは観光に関するテーマである．この二つの集合の各要素同士のつながりを明らかにすることで，地域ごとの特徴的な観光要素や観光要素ごとの特徴的な地域を抽出できると考える．これらのテーマ間の関連性を取り出すにあたり，ユーザがフォローするテーマ，記事に付与されるテーマの 2 種類のデータセットを用いた．

3.1 ユーザがフォローするテーマ

各ユーザがフォローする複数のテーマに着目した．ユーザがフォローするテーマはそのユーザの嗜好を反映している．そのため，同時にフォローされる複数のテーマはそれぞれに関連性を持つと言える．後述の手法により関連度を計算し，地域と観光要素のつながりを明らかにしていく．以下このテーマをユーザ由来テーマと呼ぶ．

3.2 記事に付与されるテーマ

次に配信される記事に同時に付与される複数のテーマに着目した．ニュース記事には，機械学習や言語処理技術 (エンティティリンキング等 [5]) を用いて複数のテーマを付与している．各記事に付与された地域テーマと観光テーマの共起を調べることにより，それらの関連性が分かる．以下このテーマを記事由来テーマと呼ぶ．

3.3 関連性の意味

ここで関連性の意味について議論する必要がある．地域テーマをフォローするユーザは，その地域について興味を持つユーザではあるが，その興味を分類すると少なくとも次の 2 種類がある．

1. 地元民：その地域に住んでいる・係わりの深いユーザ
2. 観光民：その地域に観光等で行きたいユーザ

例えば，地域テーマ「渋谷区」と観光テーマ「スキー」を同時にフォローしているユーザがいるとする．渋谷区にはスキーを行う場所がないので，このユーザは 1 の地元民であると考えられる．そのため「渋谷区」と「スキー」の関連性は低くなるべきである．また，地域テーマ「長野県」と観光テーマ「スキー」を同時にフォローしているユーザの場合，「長野県」で「スキー」はメジャーであるため，このユーザは 2 の観光民の可能性が高い．そのため「長野県」と「スキー」の関連性は高くなるべきである．一方，地域テーマが付与された記事は，基

連絡先: *1s1820785@s.tsukuba.ac.jp, *2{tatinoue, sosada, tayamash} @yahoo-corp.jp

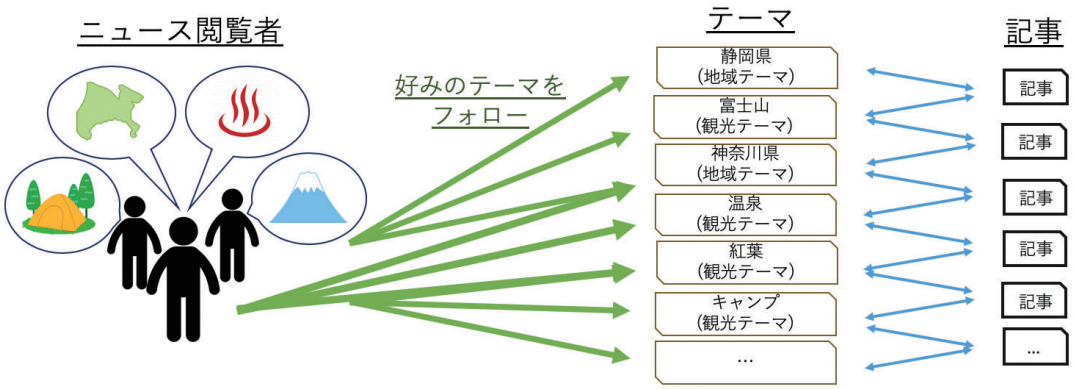

図 1: ニュース閲覧者とテーマと記事の関係

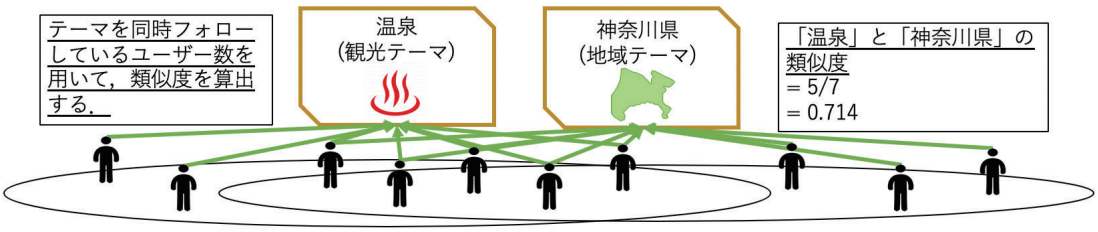

図 2: ユーザのフォローによるテーマ間の類似度の算出方法例

本的にその地域の話題であり，同時に付与された観光テーマがあればそれはその地域での観光要素であると言える．例えば，長野県でのスキー大会のニュース記事に，「長野県」「スキー」の両テーマが付与された場合，これらの関連性は高いと言える．ユーザのフォローするテーマのときとは異なり，記事に付与された地域テーマと観光テーマの関連性は高くなる．

3.4 ユーザと記事におけるテーマ間関係の差

ユーザがフォローするユーザ由来テーマ，および，記事に付与される記事由来テーマから，それぞれ地域・観光間の関連性を得ることができる．ここで，テーマ間の関連度がユーザ由来と記事由来とでの間でかけ離れていた場合，少なくとも下記の 2 つの可能性がある．

1. 誤った関連
2. ユーザの嗜好と供給されている記事とのミスマッチ

1 は，3.3 節の例で挙げた，ユーザ由来のデータから得られた「渋谷」と「スキー」のような，地域と観光要素の本来の組み合わせでないケースである．2 は，ある地域テーマとある観光テーマの関連性が，ユーザ由来では高く，記事由来では低い場合，ユーザの需要に対してニュース記事が供給不足の可能性がある．これらを実際の分析結果から確認する．

4. 関連性の抽出手法

本章では，地域テーマと観光テーマの関連性を明らかにするための関連度の計算方法について述べる．

4.1 抽出対象テーマの選択

関連性の抽出対象である地域テーマと観光テーマの選択手順について説明する．地域テーマは，47 都道府県＋ 52 市区町村の計 99 テーマを人手で選定した．この 52 市区町村は，観光に強く関連すると考えられる市区町村を主観に基づいて選定した．これらの地域テーマをフォローしているユーザが同時にフォローしているテーマの中から，旅・レジャー・観光に関する，フォロー数が 145 以上の 160 件を「観光テーマ」とした．表 1 に地域テーマと観光テーマの一例を示す．観光テーマの中には，移動手段などの間接的に観光と関係するテーマも含まれている．

4.2 関連度の計算

地域テーマ集合と観光テーマ集合のそれぞれのテーマ同士の関連度を計算する．関連度の計算には，Simpson 係数を採用した (図 2)．集合 X と集合 Y の間の Simpson 係数は，下記の式によって定義される [6]．

$$Simpson(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)}$$

例えば，ユーザ由来テーマを用いる場合，「X = 長野県」「Y = スキー」とすると，$|X|$ は地域テーマ「長野県」をフォローしているユーザ数，$|Y|$ は観光テーマ「スキー」をフォローしているユーザ数，$|X \cap Y|$ は「長野県」と「スキー」を両方フォローしているユーザ数となる．また，記事由来テーマを用いる場合，同じく「X = 長野県」「Y = スキー」とすると，$|X|$ は地域テーマ「長野県」が付与された記事数，$|Y|$ は観光テーマ「スキー」が付与された記事数，$|X \cap Y|$ は「長野県」と「スキー」の両方が付与された記事数となる．

5. 関連性の抽出

関連性の抽出にあたって，2018 年 11 月 8 日時点のユーザのフォローデータを使用した．対象ユーザ数は約 1000 万人であった．ニュース記事は，2018 年 12 月に配信された中から約

表 1: 地域テーマ, 観光テーマの一例

地域テーマ例	観光テーマ例
軽井沢市，日光市，さいたま市，相模原市， 新潟市，倉敷市， 函館市，川越市， 横須賀市，西表島，釧路市，花巻市， 祇園，難波，松山市，久留米市，佐世保市， …	航空業界, 東京ディズニーリゾート, 全日本空輸（ANA）， 日本航空（JAL）, ユニバーサル・スタジオ・ジャパン, 旅行・宿泊（国内）, 温泉, 釣り, 釣り情報, アウトドア, キャンプ場, オリエンタルランド, 登山家, 旅行・宿泊, ...
計 99 テーマ	計 160 テーマ

表 2: ユーザ由来のデータを用いた観光テーマ毎の関連度上位の地域 (括弧内の数字は関連度)

テーマ名/関連度スコア上位 3 位地域	1 位	2 位	3 位
温泉	東京都 (0.196)	箱根町 (0.175)	熱海市 (0.174)
民泊	京都市 (0.236)	東京都 (0.236)	京都府 (0.194)
パワースポット	京都市 (0.134)	京都府 (0.119)	大阪府 (0.104)
グランピング	北海道 (0.196)	兵庫県 (0.196)	京都府 (0.176)
スキューバダイビング	石垣島 (0.167)	東京都 (0.167)	北海道 (0.167)
日本庭園	京都市 (0.438)	大阪市 (0.250)	京都府 (0.250)
シュノーケリング	沖縄県 (0.462)	石垣島 (0.308)	名古屋市 (0.308)
平等院	京都府 (0.500)	京都市 (0.500)	祇園 (0.313)
グリーンツーリズム	北海道 (0.500)	山梨県 (0.500)	宮崎県 (0.500)
宿坊	京都府 (0.500)	山梨県 (0.250)	長野県 (0.250)

4000 件を使用した．

表 2 と表 3 にユーザ由来，記事由来における観光テーマ毎の地域ランキングを示す．

表 2 では「温泉」に対して「箱根町」「熱海市」，表 3 でも「大分県」「群馬県」といった温泉地として有名な観光地が抽出できた．また「宿坊」については，表 2/表 3 ともに関連性が高い「京都市」が抽出できた．表 4 と表 5 に地域毎の観光要素ランキングを示す．表 4 では「宮崎県」に対して「海水浴」，「北海道」に対して「トロッコ列車」，表 5 では「京都市」に対して「紅葉」などの妥当な組み合わせが抽出できた．

6. ユーザ由来と記事由来テーマ間の類似度

記事由来テーマ間の関連性は，システムが自動的に記事に付与するテーマに基づくものであり，実際に配信された記事の中身に基づく．一方，ユーザ由来テーマ間の関連性は，ユーザのニュースに対する興味関心に基づくものである．これら記事由来とユーザ由来の観光テーマを比較するための手法を述べる．5. 章で得られた，観光テーマ毎の地域ランキング (表 2, 表 3) データに着目する．各観光テーマは，地域テーマを要素としたベクトルを持つ．各地域テーマは各観光テーマとの関連度を持っている．これをここでは，観光地域ベクトルと呼ぶこととする．$\vec{v_1}$ は，ユーザ由来のデータによる観光地域ベクトル，$\vec{v_2}$ は，記事由来のデータによる観光地域ベクトルである．観光地域ベクトル毎に，ユーザ由来テーマと記事由来テーマ間の類似度を算出した．類似度尺度には，*cos* 類似度を用いた．*cos* 類似度は次のように定義される [6].

$$cos(\vec{v_1}, \vec{v_2}) = \frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}||\vec{v_2}|}$$

表 3: 記事由来のデータを用いた観光テーマ毎の関連度上位の地域 (括弧内の数字は関連度)

テーマ名/関連度スコア上位 3 位地域	1 位	2 位	3 位
クルーズ客船	小樽市 (0.333)	長崎市 (0.222)	鹿児島県 (0.167)
トレッキング	さいたま市 (0.375)	屋久島 (0.250)	岩手県 (0.125)
パワースポット	宮崎市 (0.200)	川越市 (0.167)	宮崎県 (0.125)
宿坊	京都市 (0.667)	祇園 (0.500)	大阪市 (0.333)
キャンプ	宮崎市 (0.200)	宮崎県 (0.125)	金沢市 (0.067)
日本庭園	神戸市 (0.222)	長崎市 (0.111)	さいたま市 (0.071)
釣り	沼津市 (0.500)	相模原市 (0.125)	兵庫県 (0.111)
民泊	新潟県 (0.222)	新潟市 (0.200)	北海道 (0.187)
日本の観光列車	北海道 (0.333)	北九州市 (0.167)	札幌市 (0.067)
スキューバダイビング	大阪市 (0.219)	高知件 (0.167)	富山県 (0.063)
温泉	日光市 (0.857)	大分県 (0.650)	群馬県 (0.622)

v_1 と v_2 の次元数は，地域テーマの数，つまり，99 である．

7. 観光テーマ別地域ランキングの評価

観光テーマを中心に関連する地域テーマが正しく抽出できたか評価を行った．

7.1 評価用データ

評価対象は前述の観光テーマ 160 件のうち，10 記事以上に付与された 71 テーマとした．記事由来とユーザ由来，それぞれにおいて，観光テーマ別地域ランキングの TOP3 にその観光テーマで有名な地域が入っているか否かの 2 値のラベル付けの作業を行った．作業は 2 名で行い，ラベルが一致したテーマのみを評価に用いた．そのうち，ラベルが positive なものを正例とした．記事由来は正例 34 件，負例 16 件，ユーザ由来では正例 37 件，負例 11 件であった．

7.2 評価結果

観光テーマ別地域ランキングの記事由来での評価 (article theme) と，ユーザ由来での評価 (follow theme) を図 3 に示す．6. 章で得られた cos 類似度に下限値を設けて，下限値以上での判定結果における再現率，適合率をプロットした．ユーザ由来と記事由来の類似度でフィルタリングしないときの精度はグラフ上，$Recall = 1$ に対応する $Precision$ であり，この時のユーザ由来と記事由来の $Precision$ は，それぞれ 0.771, 0.680 であった．ユーザの嗜好を直接反映していると考えられるユーザ由来の精度が，両手法の類似度でフィルタリングすることで，向上している．記事由来でもフィルタリングによる精度の向上が見られるが，ユーザ由来の精度には及ばないことがわかる．

8. ユーザ由来と記事由来テーマの差異分析

7.2 節で述べたように記事由来とユーザ由来の間で類似度が低いテーマをフィルタリングすると適合率が向上する．そこで，類似度が低い観光テーマについて，個別のテーマ毎に 3.4

表 4: ユーザ由来のデータを用いた地域テーマ毎の関連度上位の観光テーマ (括弧内の数字は関連度)

テーマ名/関連度上位 3 件	1 位	2 位	3 位
北海道	トロッコ列車 (0.556)	グリーンツーリズム (0.500)	隠れ宿 (0.304)
静岡県	富士山 (0.272)	カヌー (0.200)	フライフィッシング (0.188)
京都府	宿坊 (0.500)	日本庭園 (0.250)	グリーンツーリズム (0.250)
宮城県	海水浴 (0.304)	カヌー (0.200)	シュノーケリング (0.154)
大阪市	観光競争力 (0.281)	海水浴 (0.217)	日本庭園 (0.125)
長野県	しなの鉄道 (0.667)	グリーンツーリズム (0.250)	宿坊 (0.250)
山梨県	グリーンツーリズム (0.500)	宿坊 (0.250)	トロッコ列車 (0.222)
金沢市	隠れ宿 (0.087)	リアル脱出ゲーム (0.083)	鮎釣り (0.071)
宮崎市	キャンプ (0.118)	鮎釣り (0.071)	フライフィッシング (0.063)
青森市	青森ねぶた祭り (0.353)	鮎釣り (0.071)	温泉 (0.051)
佐世保市	カヌー (0.200)	フライフィッシング (0.063)	海水浴 (0.043)
沼津市	フライフィッシング (0.125)	釣り大会 (0.105)	ワカサギ釣り (0.091)
熱海市	温泉 (0.174)	隠れ宿 (0.087)	秘湯 (0.087)
箱根町	温泉 (0.175)	紅葉 (0.100)	秘湯 (0.075)
日光市	温泉 (0.083)	日本の観光列車 (0.083)	リアル脱出ゲーム (0.083)

表 5: 記事由来のデータを用いた地域テーマ毎の関連度上位の観光テーマ (括弧内の数字は関連度)

テーマ名/類似度上位 3 件	1 位	2 位	3 位
大分県	温泉 (0.650)	土産 (0.100)	温泉宿 (0.100)
金沢市	土産 (0.267)	紅葉 (0.267)	はとバス (0.167)
滋賀県	紅葉 (0.316)	鮎釣り (0.286)	土産 (0.105)
群馬県	温泉 (0.622)	秘湯 (0.333)	温泉宿 (0.077)
京都市	宿坊 (0.667)	紅葉 (0.250)	グリーンツーリズム (0.25)
北海道	ワカサギ釣り (1.000)	鉄道旅行 (0.600)	日本の観光列車 (0.333)
兵庫県	バーベキュー (0.333)	温泉 (0.270)	渓流釣り (0.222)
山梨県	富士急ハイランド (0.667)	富士山 (0.490)	渓流釣り (0.222)
大阪市	天王寺動物園 (0.9)	ユニバーサル・スタジオ・ジャパン (0.608)	宿坊 (0.333)

節であげた項目を検証した．

- 誤った関連

例 1 :「東京都」「温泉」のような明らかな誤り

「東京都」と「温泉」テーマの間では，ユーザ由来テーマでは，関連度が高かったが，記事由来では，関連度は低かった (表 2，表 3)．つまり,「東京都」をフォローしている人が同時に「温泉」テーマをフォローしていると考えられるので 3.3 節の 1 の地元民による「温泉」テーマのフォローであると考えられる．

例 2 :「大阪府」「スキューバダイビング」のような記事テーマ付与の誤り候補

「大阪府」と「スキューバダイビング」のテーマの間では，ユーザー由来の関連度が低く，記事由来の関連度が高かった (表 2，表 3)．実際に,「大阪府」と「スキューバダイビング」に同時に紐付く記事を見てみると，大阪府の水族館でコスプレをしたスキューバダイバーが水槽内に現れたという内容であった．ここでは，スキューバダイビングのテーマに紐付くのは，誤りであると考えられる．記事に対するテーマの紐付けの誤り候補を取り出すことができた．

- ユーザの嗜好と供給されている記事とのミスマッチ

ユーザ由来と記事由来間の類似度の低いテーマ「日本庭園」に着目する.「日本庭園」と「京都市」は，ユーザ由来のテーマでは関連が高いが，記事由来のテーマでは，関連が小さい (表 2，表 3)．しかし，例 1 のような誤った地域は確認されず，配信側の記事テーマとユーザーの嗜好の間に差がある可能性がある．

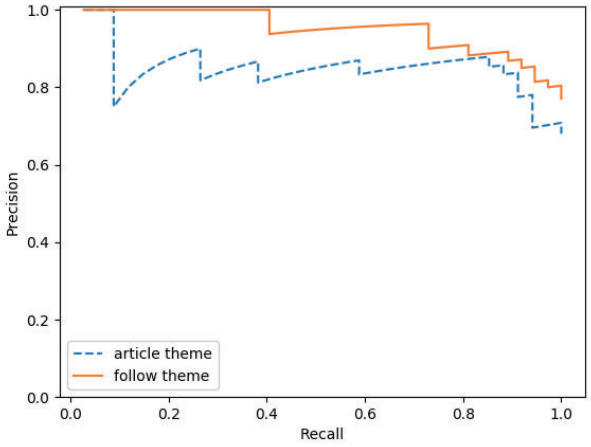

図 3: 観光テーマ別地域ランキングの人手評価結果

9. おわりに

本研究では，ニュースサイトにおいて,「ユーザによって能動的に選択されたテーマ」および「記事に内容に基づき自動的に付与されたテーマ」のそれぞれを用いて地域と観光要素の紐付けを行い，観光要素別地域ランキングと地域別観光要素ランキングを作成した．そして，両者の類似度を利用することにより精度の向上を確認できた．また，類似度が低いテーマを分析し，テーマ紐付けの誤りパターンの知見を得た．さらに，ユーザと記事配信の需要供給の間にミスマッチが存在すると思われるケースも発見した，今後は分析を深めて，より正確な地域・観光要素間の関連性の抽出を実現し，地域の意外な観光テーマの発見など，地域活性化に貢献していきたい．

参考文献

[1] 日本政府観光局 (JNTO). 国籍/月別 訪日外客数. https://www.jnto.go.jp/jpn/statistics/since2003_visitor_arrivals.pdf.

[2] 王怡青, 土井俊弥, 井上祐輔, 宇津呂武仁. ご当地グルメを題材とするクイズ・コンテンツの作成. 第 8 回 DEIM フォーラム論文集, 2016.

[3] 土井俊弥, 王怡青, 井上祐輔, 宇津呂武仁. 観光情報 PR のための旅ゲー風アプリの提案およびご当地グルメ版の作成. 言語処理学会第 22 回年次大会論文集, pp. 59–62, March 2016.

[4] Yahoo!ニュース. https://news.yahoo.co.jp/.

[5] 井上裁都, 末永圭吾, 長田誠也, 立石健二. Entity linking を用いたニュース記事に対する市区町村単位の地域情報の付与. 言語処理学会第 22 回年次大会論文集, pp. 45–48, March 2016.

[6] 相澤彰子. 大規模テキストコーパスを用いた語の類似度計算に関する考察. 情報処理学会論文誌, Vol. 49, No. 3, pp. 1426–1436, 2008.