3W-04

観光情報可視化システムのための Web 上における観光特徴抽出手法の提案

渡邉 小百合 †　　吉野 孝 †
† 和歌山大学

1 はじめに

観光庁による旅行・観光消費動向調査 [1] では，2010 年から 2013 年にかけて日本人の旅行平均回数が増加している．近年ドラマやアニメの舞台への聖地巡礼等の新しい形態の旅行だけでなく，外国人旅行者の増加など，新しいニーズが出てきていることから，観光地側も観光客のニーズや問題点を知り，観光地の発展に繋げていく必要がある．しかし，観光庁の調査 [1] では，東京都で宿泊旅行をした人数が約 33000 人であるのに対し，和歌山で宿泊旅行をした人数は約 4600 人という結果が出ている．これより，メジャーな観光地と比べマイナーな観光地は情報が少ないため，問題点の発見が困難であると考えられる．

我々は，観光業者に有益な情報を提供する，観光情報可視化システムの開発を目指している．そのためには，マイナーな観光地であっても，観光地間の類似性等から改善のための示唆を得ることが出来る環境の構築が必要である．そこで，本稿では，観光情報可視化システムを構築するための，Web 上における観光特徴の抽出手法を提案する．

2 関連研究

Web 上から観光特徴を抽出する研究として，遠藤らの地域サイトとブログ内で多く出現する名詞から，観光情報を融合して提供するシステムがある [2]．この研究では，抽出する観光特徴を名詞のみに限定している．本研究では，名詞だけではなく，「のどかな」「神秘的な」といった観光地の雰囲気や抽象的な表現も抽出する点が異なる．

また，松本らは 4 種類の特徴量算出手法を用いてクチコミから特徴語を抽出する実験を行っている [3]．この実験では，特徴語の抽出に計算的定義を用いているが，本研究では特徴の抽出手法として文書解析を行うトピックモデルの LDA を用いる点で異なる．

以上の研究は，観光特徴の抽出を行い，その結果を来訪者に対しての情報として使用している．本研究は，抽出した観光特徴を，観光業者に対して有益な情報として利用することを想定している．

3 観光特徴の抽出

本研究では，Web 上から観光特徴を抽出するために，トピックモデルの一種である LDA(Latent Dirichlet Allocation, 以下 LDA)[4] を用いる．

以下に観光特徴抽出の手順を示す．

(1) 形態素解析を用いた単語の抽出

表 1: 観光特徴のカテゴリと単語例

カテゴリ	例	単語数
見る・遊ぶ	観光　体験　ウォーキング	71
祭り・イベント	コンテスト　祭り　花火大会	15
自然・文化	世界遺産　歴史　海	88
食べる・泊まる	温泉　グルメ　旅館	31
お土産・特産品	特産物　名物　伝統工芸品	15
印象・雰囲気	厳かな　美しい　開放感	36

形態素解析システム JUMAN[*1] を用いて，Web ページのテキストデータを解析し，その中から観光特徴になりうると思われる単語を取り出す．Web ページは，47 都道府県魅力度ランキング 2015[*2] における 1 位から 10 位までの 10 都道府県の観光サイトを参考とし，合計 354 単語を取得した．

(2) クラウドソーシングを用いた観光特徴の分類
クラウドソーシングを用いて，遠藤らが定義したもの [2] に，「印象・雰囲気」を追加した 6 つのカテゴリに，取得した観光特徴を分類する．「印象・雰囲気」を追加した理由としては，観光地が持つ雰囲気と与える印象によって，来訪者の振る舞いや抱える問題点が異なるのではないかと考えたためである．クラウドソーシングには，Yahoo!クラウドソーシング [*3] を用い，10 人で分類を行った．回答結果において，最も多い回答が過半数以上だったものをその単語のカテゴリとし，最も多い回答が過半数以下のものは観光特徴として曖昧であるとし，取り除いた．その結果，取得した観光特徴の約 72%がカテゴリ付けされた．表 1 に，カテゴリとカテゴリに含まれる観光特徴の例と単語数を示す．

(3) LDA モデルの作成
1 つのカテゴリを 1 文書とした，計 6 文書と，Python ライブラリ Gensim[*4] を用いて LDA モデルを作成する．作成した LDA モデルに，形態素解析で分かち書きしたテキストデータを入力することで，モデルの元となった文書と入力したテキストデータの類似度を出力する．

4 評価実験

4.1 実験概要

実験は，カテゴリ付けした単語で作成した LDA モデルを用いて，観光地間の類似度を検証する．メジャーな観光地とマイナーな観光地の類似点を発見することを目的としているため，本実験では，トリップアドバイザーによる日本観光ランキング 10 選 [*5] において，各

Proposal of tourism feature extraction methods on the web for the tourist information visualization system
Sayuri Watanabe† Takashi Yoshino†
†Wakayama University

[*1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN
[*2] http://tiiki.jp/news/05_research/survey2015/2858.html
[*3] http://crowdsourcing.yahoo.co.jp/
[*4] https://radimrehurek.com/gensim/index.html
[*5] https://www.tripadvisor.jp/Attractions-g294232-Activities-Japan.html

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.

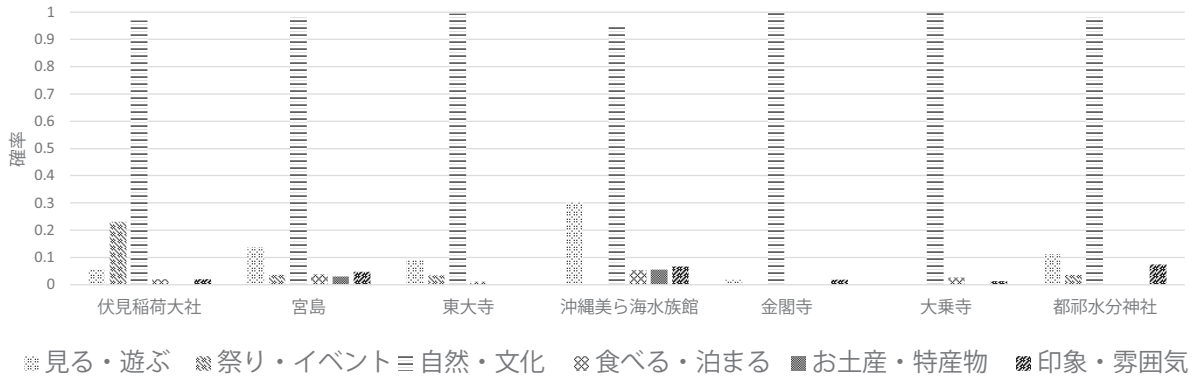

図 1: 観光サイトに含まれるカテゴリの確率

地域上位 3 位以内の観光地をメジャーな観光地，ワースト 3 位以内の観光地をマイナーな観光地とする．メジャーな観光地は伏見稲荷大社 [*1]，宮島 [*2]，東大寺 [*3]，沖縄美ら海水族館 [*4]，金閣寺 [*5] とし，マイナーな観光地は大乗寺 [*6]，都祁水分神社 [*7] とした．

JUMAN によって分かち書きした各 Web サイトのテキストデータを LDA モデルに入力し，テキストデータに含まれているカテゴリの確率とトピックの確率の出力結果から類似度を測る．

4.2 実験結果と考察

図 1 に実験結果を示す．縦軸は，観光地のテキストデータにカテゴリの単語が含まれている確率を表す．以下の 2 つの観点から考察する．

(1) メジャーな観光地とマイナーな観光地の類似度

大乗寺 金閣寺と，自然・文化以外のカテゴリの確率が低い点で類似していることがわかる．これは，どちらの観光サイトも，主に観光地そのものや歴史についてを取り上げているためだと考えられる．

都祁水分神社 東大寺と，見る・遊ぶ，祭り・イベント，自然・文化，お土産・特産物カテゴリの確率が類似している．しかし，観光サイトを比べると，お土産の項目がない点や展示物の説明がある点では類似していたが，東大寺の方が行事に関する説明が多かった．これは，行事の名前と説明に「祭り」や「コンテスト」といったカテゴリの単語が含まれているものが少なかったことが原因であると考えられる．

(2) メジャー同士，マイナー同士の類似度

メジャーな観光地 ほとんどの観光地において，6 つ中 5 つ以上のカテゴリが含まれている．これは，観光地のサイトにおいて，サイト内の項目が多いため，観光地に関する情報量が多いことが考えられる．

マイナーな観光地 含まれていないカテゴリが 2 つ以上ある．これは，観光サイトのトップページにあるメニュー項目以外の個別ページがなく，情報量が少ないことが考えられる．これより，マイナーな観光地には情報が少ないことがわかるので，メジャーな観光地との類似点を発見する必要がある．

全ての観光地において，自然・文化カテゴリの値が他のカテゴリと比べて非常に高くなっている．これは，観光地に自然や歴史に関係するものが多いという理由もあるが，LDA モデルに使用しているカテゴリ内の単語数の差が問題として挙げられる．表 1 に示すように，自然・文化カテゴリの単語数は，本実験で最も値が低いお土産・特産物カテゴリの単語数と，約 6 倍の差がある．これより，全カテゴリの単語数を調整する必要があると考えられる．また，この手法を観光情報可視化システムに組み込むにあたって，類似度や確率といった数値の表現方法も検討する必要がある．

5 おわりに

本研究では，観光業者に有益な情報を提供する，観光情報可視化システムの開発を目指している．本稿では，クラウドソーシングによってカテゴリ付けした観光特徴でトピックモデルの一種である LDA モデルを作成し，観光地間の類似度を測る実験を行った．その結果，含まれているカテゴリの確率やバランスから，観光地間の類似点，相違点が推測できることがわかった．

問題点として，カテゴリに含まれる単語数に差があることが挙げられる．今後は，カテゴリの単語数の調整や，類似度の可視化方法について検討していく．

参考文献

[1] 国土交通省 観光庁：「旅行・観光産業の経済効果に関する調査研究」(2013 年版), http://www.mlit.go.jp/common/001091028.pdf (参照 2016-1-5).

[2] 遠藤雅樹, 中村信也, 奥秋清次, 大野成義：地域サイト及びブログからの観光情報抽出と融合の提案, 情報処理学会研究報告, Vol.2012-DBS-155 No.6, pp.1-6 (2012).

[3] 松本敦志, 杉本徹：クチコミから抽出した特徴語を利用する観光地検索支援, 情報処理学会第 75 回全国大会, pp.307-308 (2013).

[4] David M.Blei,Andrew Y.Ng,Michael I.Jordan：Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp.993-1022 (2003).

[*1] http://inari.jp/
[*2] http://www.miyajima-wch.jp/index.html
[*3] http://www.todaiji.or.jp/index.html
[*4] http://oki-churaumi.jp/index.html
[*5] http://www.shokoku-ji.jp/k_about.html
[*6] http://www.daijyoji.or.jp/main/index.html
[*7] http://www.tsugemikumari.or.jp/

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.