テーマをもつ観光地グループの自動生成

中村 みなみ † 乙武 北斗 ‡ 吉村 賢治 ‡
†福岡大学大学院工学研究科 ‡福岡大学工学部
td152007@cis.fukuoka-u.ac.jp , {ototake, yosimura}@fukuoka-u.ac.jp.

1 はじめに

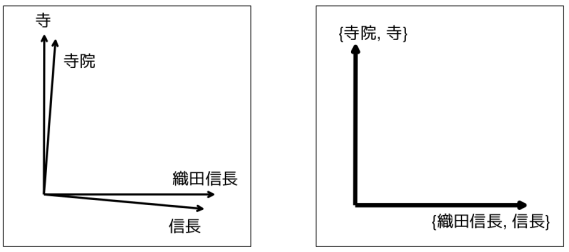
地方自治体はそれぞれの地域で多くの観光資源をもっており，その観光資源を活かすために，テーマをもつ観光地の周遊ルートを作成したいという要望をもっている．また，近隣の複数の地方自治体が協力して周遊ルートを作成したいという要望もある．

本稿におけるテーマとは，ルートに含まれる観光地に関連している人や出来事，歴史的背景等である．例えば，京都府の「南蛮寺跡」「本能寺跡」「相国寺」は「織田信長」にゆかりがある地となっており [1]，この 3 つを結ぶ観光ルートのテーマは「織田信長」となる．

このようなテーマをもつルートの作成は，現状人手で行われており，観光地や観光地周辺の歴史的背景等の専門知識が必要であること，複数人で決定する作業には費用や時間のコストが多くかかること等の問題がある．

それぞれのユーザーに適した観光情報の自動抽出であれば，石野らによる旅行ブログを利用した観光情報の自動抽出 [2] や，小作らの新聞記事を利用した観光地のイベント情報の抽出 [3] 等がある．また，キーワードを用いて予め人間が用意しているグループに分類する手法も提案されている [4]．これらの先行研究においては，観光地の情報収集や，予め人間がテーマをもつグループを用意しておく必要があるためその部分で人手による作業コストが発生する．

このような問題を解決するために，本稿では複数の観光地に関連するテーマを発見し，テーマをもつ観光地のグループを自動生成する手法を提案する．人手を介さないこと，自動分類された各グループはテーマをもつことを目的とする．これにより，観光地について専門的な知識がなくとも，テーマのある観光ルートを生成することができ，時間や費用といったコスト削減にもつながる．また，京都府のような多くの観光資源をもつ地域だけでなく，地方の観光地でも利用することができ，観光情報の発信の材料にもなる，人手では気付かなかったような観光地同士の繋がりを知ることができる，地方での観光資源を活かすことができる等の利点が考えられる．



(a) 単語素性軸の例　(b) 類似単語集合素性軸の例

図 1 文書ベクトルの素性軸

2 提案手法

本章では，観光地及び案内文を入力として，その観光地のテーマを出力する 2 つの手法について提案する．1 つは観光地の類似性を利用した手法で，もう 1 つは関連性を利用した手法である．以下その詳細を述べる．

2.1 類似性を利用した手法

同一の人物や出来事が出現する案内文をもつ観光地を巡るルートはテーマをもつルートであると考える．その案内文から観光のテーマとなる語，つまり共通の人物や出来事については TF-IDF を用いて抽出ができ，文書ベクトルの余弦類似度をテーマ性のある観光地の分類の第一次近似として利用する．また，文書ベクトルを構築するにあたり，素性空間の軸の直交性が重要となる．そこで TF-IDF を用いて特徴語として抽出された単語に対し，word2vec[5] を使用し，類似した単語の集合を 1 つの軸とする．図 1 に例を示す．また，この手法では案内文をクラスタリングをした際にクラスタの中心となったベクトルがテーマとなる．今回，ハードクラスタリング，ソフトクラスタリングには bayon[*1] を使用した．以下に手順を示す．

[*1] Mizuki Fujisawa. bayon
https://github.com/fujimizu/bayon

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

手順 1： 文書ベクトルの構築

(i) 案内文から TF-IDF で抽出した単語について word2vec を用いて 200 次元のベクトル表現を求める．

(ii) ベクトル化された単語をハードクラスタリングする．この結果で得られた各クラスタを文書ベクトルの素性とする．

手順 2： 案内文のクラスタリング

(i) 手順 1-(ii) で次元削減された文書ベクトル空間を用いて各案内文のハードクラスタリングを行い，クラスタの中心となるベクトルを求める．

(ii) 手順 2-(i) で求めたクラスタの中心ベクトルとの余弦類似度を用いて案内文のソフトクラスタリングを行い，その中心をクラスタのテーマとする．

2.2 関連性を利用した手法

観光地の案内文，もしくは観光地の名称を Web 検索し，その結果から取得されたページに含まれている単語を観光地と共起関係のある単語とし，その共起頻度を観光地とその単語の関連度とみなす．また，この単語がテーマ候補となる．以下に手順を示す．

手順 1： Web 検索を行う．観光地名，あるいは観光地案内文から抽出した観光地に存在している観光資源と観光地がある都道府県を検索ワードとし，Web 検索を行う．本研究では BingSearchAPI[*2]を使用した．取得した記事数は 1 回の検索に対し，100 件である．

手順 2： 検索結果から記事内に出てくる固有名詞のみを抽出する．この抽出された固有名詞が各観光地を巡るルートのテーマとなりうる単語である．

手順 3： 記事から抽出された固有名詞と観光地の共起頻度をとる．人名は goo ラボ固有表現抽出 API[*3]において予め抽出し，表記の統一を図る．例えば，”坂本竜馬”，”坂本龍馬”, ”竜馬”, ”龍馬”という表現は全て”坂本竜馬”に統一する．

[*2] Bing Search API. https://azure.microsoft.com/ja-jp/services/cognitive-services/search/

[*3] goo ラボ固有表現抽出 API https://labs.goo.ne.jp/api/jp/named-entity-extraction/

表 1 実験に用いた正解データ

テーマ	観光地
夏目漱石	下鴨神社, 高野川, 赤山禅院
後醍醐天皇	大覚寺, 天龍寺
坂本竜馬	岬神社, 酢屋, 寺田屋 , 京都霊山護国神社
紫式部	廬山寺, 誠心院, 東北院の軒端の梅
小堀遠州	大徳寺, 二条城, 金地院, 御香宮神社, 伏見奉行所跡
小野小町	下御霊神社, 補陀洛寺, 随心院
織田信長	南蛮寺跡, 本能寺跡, 旧二条城跡 相国寺, 大徳寺, 妙心寺
千利休	大徳寺, 北野天満宮, 一条戻橋, 晴明神社
川端康成	清滝, 八坂神社御旅所, 円山公園
足利義満	石清水八幡宮, 若宮八幡宮, 御所八幡宮, 花の御所, 相国寺, 金閣寺, 等持院

手順 4： 共起頻度の閾値によって固有名詞と観光地のフィルタリングし，固有名詞で観光地をグループ化する．その結果，作成されたグループの固有名詞をテーマとする．1 つのテーマに対し，観光地が 2 つから 7 つまでのものが観光ルートに適するものとする．

3 評価実験

3.1 実験対象

実験に使用したデータは以下の通りである．

(1) word2vec の学習に用いたデータ
2016 年 6 月時点の Wikipedia の全記事約 1,200,000 記事 (2.5GB)

(2) 分類対象のデータ
京都観光 Navi[1] から収集した観光ルート 10 件に含まれる観光地 37 件．この 10 件の観光ルートを正解データとする．それぞれのテーマには観光地が 2 つから 7 つ存在する．その内容を表 1 に示す．

(3) 形態素解析器とそこで使用した辞書
MeCab
mecab-ipadic-NEologd

3.2 実験方法

類似性を利用した手法，関連性を利用した手法についてそれぞれ分類対象データに対してテーマと観光ルートを求める．

類似性を利用した手法については，クラスタリング

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

結果からシステムの出現クラスタ数，正解数，及び，システム提案分類の中から実際に第一著者が観光地についてWeb検索をし，テーマが1度でも出現すれば妥当性の高いクラスタであると判断する．妥当性の判断は正解とは異なるクラスタに分類されたテーマに対して行う．

関連性については観光地名と都道府県で検索した結果と，観光地に存在する観光資源と都道府県で検索した結果に対して，2-2の手順2から4の処理を行い，その結果得られるグループに対して，共起頻度の閾値との関係も求める．関連性の手法の結果の正否の判断としては，適合率，再現率のF値を用いている．
適合率，再現率の定義を以下の式(1)，(2)に示す．

$$\text{適合率} = \frac{a_1}{\text{正解データの個数}} \tag{1}$$

$$\text{再現率} = \frac{a_1}{a_2} \tag{2}$$

a_1とは，システム出力のテーマ内にある観光地の中で実際の正解データの観光地と一致している数，a_2はシステム出力のテーマ内にある全ての観光地の個数である．例えば，「坂本竜馬」というテーマでは，正解データは，「岬神社」「酢屋」「寺田屋」「京都霊山護国神社」であり正解データの個数は4，それに対し，システムが出力した結果が，「岬神社」「酢屋」「寺田屋」「京都霊山護国神社」「下鴨神社」「円山公園」であった場合，a_1の値は4となり，適合率は4/4= 1となる．また，システムが出力した結果は6つの観光地であるので，a_2の値は6である．よって再現率は4/6となる．この適合率，再現率によりF値を求め，F値が0.3以上のものを正解のテーマと判断する．

また，テーマの妥当性の判断としては，共起頻度の閾値25と30のものでそれぞれの観光地とテーマの共起頻度の合計値が高い10個のテーマを第一著者が実際にWeb検索をし，検証する．検証方法としては式(3)に示す検証値の値が0.5以上のものが妥当性の高いテーマであると判断する．

$$\text{検証値} = \frac{\text{テーマと関連があった観光地数}}{\text{そのテーマで提案している観光地数}} \tag{3}$$

例えば，「足利義満」というテーマに対して観光地が「金閣寺」「天龍寺」「相国寺」「大徳寺」「花の御所」の5つの観光地が出力された場合，テーマで提案している観光地数は5となる．またWeb検索により，足利義満と関連があるものは「金閣寺」「相国寺」「大徳寺」「花の御所」の4つであると判断された場合，テーマと関連があった観光地数は4となる．この場合の検証値は4/5である．

表2　類似性を利用した手法の結果

出力クラスタ数	正解数	妥当性の高いクラスタ数
17	2	8

表3　類似性を利用した手法の出力例 (一部抜粋)

出力観光地	出力テーマ	正解テーマ
京都霊山護国神社，酢屋	坂本竜馬，明治維新	坂本竜馬
廬山寺，誠心院	紫式部，阿弥陀如来	紫式部

表4　関連性を利用した手法-観光地名で検索

閾値	出力テーマ数	正解数
10	236	6
15	135	6
20	77	5
25	55	5
30	34	3
35	20	2
40	13	2

表5　関連性を利用した手法-観光資源で検索

閾値	出力テーマ数	正解数
10	1156	4
15	821	6
20	598	5
25	471	5
30	380	5
35	307	4
40	260	4

3.3　実験結果

類似性を利用した手法の精度を表2に，実際のシステムの出力例を表3に示す．関連性を利用した手法について，観光地のみで検索した結果を表4に，観光地内の観光資源をキーワードとして検索した結果を表5に示す．また，関連性を利用した手法については共起頻度の合計値が高い10個のテーマのうち，妥当性の高い，つまり検証値が0.5以上のものがいくつあったのかを表6に示す．

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.

表 6　関連性を利用した手法-10 件中妥当性の高いもの

検索キーワード	閾値	検証値 0.5 以上の数
観光地名	25	7
観光地名	30	8
観光資源	25	6
観光資源	30	7

4　考察

類似性を利用した手法よりも関連性を利用した手法の方が提案数が多く，また，妥当性の高い提案が多い．類似性を使った手法が悪くなる原因として考えられるものは，文書の類似性で求められるものは”似ているもの”であり，共通している人物や出来事といった今回のテーマを集めることは困難であったためと考えられる．

関連性を利用した手法においては，観光地のみの検索に比べ，観光資源で検索したものの提案観光地数が多くなっている．これは観光地に含まれている観光資源の数に応じて取得される記事が観光地毎に 100 から 500 件とバラバラになってしまったため，不必要なものが増加してしまったためではないかと考えられる．また，観光資源で検索した場合，不適切なテーマが出力される傾向がある．例えば，「鹿苑寺」というテーマで「金閣寺」「大徳寺」「花の御所」が出たが，これらは「足利家」が共通しているものであり，「鹿苑寺」はテーマとしてふさわしくない．2 種類の検索での正解データの数や妥当性が高いと判断されたテーマの数はほぼ同じである．抽出した観光地の案内文もインターネット内にあるものを利用したため，観光資源での検索はあまり効果がなかったように感じる．

この結果から，テーマをもつ観光地グループを導き出すには，観光地名で検索をし，その結果，共起関係のあるものがテーマとなり，100 件の記事に対しての共起頻度 25 から 30 が妥当な結果が出ることがわかった．結果の精度を上げるにはいくつかの方法が考えられる．

1. 人名統一の充実

   今回人名変換を手動で作成したが，歴史上の人物で中学校の教科書に出てくるような有名な人物のみであり，全ての人物の変換表とはなっていない．今後，この変換表の種類を豊富にしていくことで異なるテーマに分かれている人物を同一テーマにすることができると考えられる．また，人名だけでなく，観光地名についても変換表を作成することで表記揺れの問題が解決するのではないかと考えられる．(例：「金閣寺」と「鹿苑寺」)

2. ストップワードリストの作成

   今回手動でテーマにはならないホームページ特有の単語 (例： ブックマーク) と入力データの観光地名をテーマとして出力しないためにストップワードとした．今回は，京都の全ての観光地一覧を取得することができなかったため，ストップワードリストには周辺の観光地名を入れることができていなく，ホームページ特有の単語がほとんどとなっている．観光地名はテーマにはならないため，「清水寺」のような今回の正解データには入っていないが，周辺にある観光地もテーマからはずすストップワードにするべきである．

3. テーマの選定

   今回テーマとなるものは形態素解析で固有名詞と判断されたものとしている．使用した辞書の mecab-ipadic-NEologd は多くの普通名詞を固有名詞表現とすることが多い．そのため，不要なテーマを出力することが多い．1 つの形態素解析器だけを使用するのではなく，複数のものを使用してテーマを選定する工夫が今後必要となると考えられる．

5　おわりに

本稿ではテーマをもつ観光地グループの自動生成を行う 2 つの手法を提案し，それぞれ評価実験を行った．評価実験の結果，テーマ毎の分類に関しては関連性を利用した手法が有効であることが分かった．4 で上げた方法等の工夫をしてテーマの抽出の精度を上げることが今後の課題である．

参考文献

[1] 京都観光 Navi(織田信長ゆかりの地をたどる) https://kanko.city.kyoto.lg.jp/travelroute.php?InforKindCode=7&ManageCode=5000023

[2] 石野 亜耶, 難波 英嗣, 竹澤 寿幸. 旅行ブログエントリからの観光情報の自動抽出. 知能と情報, Vol. 22, No. 6 ,pp. 667-679, 2010.

[3] 小作 浩美, 内山 将夫, 井佐原 均, 河野 恭之, 木戸出 正継. 新聞記事コーパスでの単語出現特徴を利用した観光イベント情報の検索支援. 人工知能学会論文誌, Vol. 19, No. 4, pp.225-233, 2004.

[4] 岡田 真, 小山 雅史, 獅々堀 正幹, 青江 順一. キーワード抽出を用いた文書自動分類手法. 情報処理学会第 55 回全国大会, pp.210-211, 1997.

[5] T.Mikolov et al.,“Distributed representations of words and phrases and their compositionality,”In Proceedings of NIPS, 2013.

Copyright(C) 2017 The Association for Natural Language Processing.
All Rights Reserved.