社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IN2019-67 (2020-01)

e旅行: 旅行スタイル別レビュー分析に基づく旅行支援サイト自動生成

川村　直輝†　荒澤　孔明†　服部　峻††

†,††室蘭工業大学 ウェブ知能時空間研究室　〒 050–8585 北海道室蘭市水元町 27–1
E-mail: †{16024047,18096001}@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

あらまし　多くの人々は旅行や観光のために，旅行レビューサイトや旅行情報誌を利用しており，旅行レビューサイトでは友人や家族，カップルなどの旅行スタイル別に観光スポットやグルメを閲覧することが出来る機能も使われている．ユーザの中には，旅行を計画している人や移動中にどこかへ立ち寄る際にこのような機能を利用する人々がいる．しかし，実際に旅行する際に，例えば友人を旅行スタイルの条件に加え，観光スポットや飲食店を検索すると，検索結果のレビュー数が少ない場所が存在し，参考になりにくい場合が多く，従来サイトの旅行スタイル別の検索では未だ十分ではない．そこで本稿では，旅行支援サイトを自動生成するために，複数の旅行レビューサイトから地域ごとのレビューを出来る限り網羅的に収集し，旅行スタイル別の分類を行った上で，旅行スタイル別のレビューに絞って分析する．この分析結果より各地域特有の地域スポットの抽出及びランキングを行うことで，旅行支援サイトを自動生成するシステムを開発することを目的としている．

キーワード　旅行支援，文章分類，情報抽出，Web マイニング，機械学習

e-Travel: Automatical Travel Support Site Generation based on Review Analysis per Travel Style

Naoki KAWAMURA†, Komei ARASAWA†, and Shun HATTORI††

†,†† Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan
E-mail: †{16024047,18096001}@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

Abstract　For traveling and sightseeing, many people often use travel review sites and travel information magazines. Some travel review sites allow users to browse sightseeing spots and food of a target place depending on travel styles such as “alone,” “friend,” “family,” and “couple.” Some users try to utilize such a function while planning their travel and/or stopping over at somewhere (not a planed place) in their travel. However, the function of search per travel style of the existing travel review sites is not enough useful when a traveler searches for sightseeing spots and restaurants by querying a target place and adding “friend” as a travel style condition, because the search results for some places are not helpful for her/him due to lack of reviews. Therefore, this paper aims at developing an automatical travel support site generation system that extracts place-specific spots and offers users with the ranking of them per travel style, by exhaustively collecting place-specific reviews from many travel review sites and analyzing them classified per travel style.

Key words　Travel Support, Text Classification, Information Extraction, Web Mining, Machine Learning

1. ま え が き

旅行や観光の計画を立てる際には，様々な要素が存在する．その要素には「誰と」,「何のために」,「どこへ」,「いつ」，などが挙げられる．どの要素も旅行形態に少なからず影響はあるが，特に『誰と』は家族旅行や友達と観戦，一人で出張など様々な旅行の目的により変動し易いと考えられる．また，多くの人々は旅行や観光の計画を立てる際，旅行レビューサイトや旅行情報誌を利用している．旅行情報誌は包括的に情報を掲載しているが，旅行レビューサイトでは検索機能を使用することで友人や家族，カップルなどの「誰と」行く旅行であるかに応じた旅行スタイル別に観光スポットやグルメを閲覧することが出来る．また，ユーザの中には，事前に旅行計画を立てている時の他，移動中に立ち寄る際にもこのような機能を利用する人々

もいる．しかし，実際に旅行する際に，例えば『友人』を旅行スタイルの条件に加え，その条件に即した観光スポットや飲食店を検索すると，そのレビュー数が少ないため参考にしにくい場合が多く，従来の旅行スタイル別の検索では不十分である．

本稿では，これらの問題を解決するため旅行スタイル別に支援を行えるサイトを自動生成するために，複数の旅行レビューサイトから地域ごとのレビューを出来る限り網羅的に収集し，旅行スタイル別に分類を行う手法について提案する．さらに，旅行スタイル別のレビューに絞って分析し，この分析結果から旅行スタイル別の各地域特有の地域スポットの抽出を行い，これらを用いてランキング順にし，ユーザに提示する手法について提案する．また，地域スポットとしてグルメ，観光スポット，観戦などのメジャーなスポットだけでなく，潜在的なスポットも抽出可能にする手法についても検討する．

2. 関 連 研 究

本研究ではレビュー分類に文章分類の手法を利用するが，文章分類に関する研究は多く行われている．西川ら [1] は，機械学習を用いて，旅行情報ポータルサイトのレビューを用いて，ホテルなどの利用者の状況・状態に応じて 1 人の利用か，複数人の利用かの極性判定を行い分類している．しかし，この研究では目標の 1 つである旅行スタイルに合わせた分類まで達しておらず，旅行スタイル別のランキングを作成できるとは言い難い．また，滝川ら [2] は，十分な学習データを用意できない状態で，特定分野に対する専門性のある短い文章を推定する手法を提案していた．学習データが無いため機械学習を用いず，単語重要度を求め，単語に重みを付与することで分野の推定を試みている．しかし，この研究ではある特定の分野かどうかを調べるために，ある特定分野以外の文書を一般的な文書としている．そのため，旅行レビューでスタイル別に分類したレビューをある特定の分野と考えた際に，旅行ではどのようなレビューであっても，いずれかの旅行スタイルに分類されてしまうので，単語重要度を求めるための旅行スタイルとして分類されない一般的な文書が存在しない．

そこで，本研究では，文章を 2 種類へ極性判定するのではなく複数種類へ分類を行い，他の一般的な文書を用いずに短いレビュー文章を分類するため，TF-IDF 法を参考に単語重要度を付与し，旅行スタイル各々の重要語に基づいて分類する手法を試みている．

3. 提案システム

本章では，旅行スタイル別の旅行サイトからレビューを出来る限り網羅的に収集し，そのレビューを旅行スタイル分類し，旅行スタイル別に分類したレビューを基に各地域の地域スポットをランキング順に表示するための，旅行支援サイト自動生成システムについて詳細を述べる．

3.1 旅行支援サイト自動生成システム全体の概要

本稿における旅行支援サイト自動生成システムの構成を図 1 に示す．この旅行支援サイトではユーザが旅行で訪れたい地域を入力すると，ユーザに旅行スタイル別の地域スポットのランキングを提示する．その際に，事前処理として複数サイトの旅行レビューの収集を行い，これらを市町村ごとの地域に分類する．さらに，旅行スタイル別（本稿では『カップル』,『家族』,『友人』,『一人』の 4 種類）に分類を行う．分類された旅行スタイル別のレビューから旅行スタイル別の地域スポットをランキング順に示し，このデータを基に図 2 のような旅行支援サイトを自動生成するシステムである．

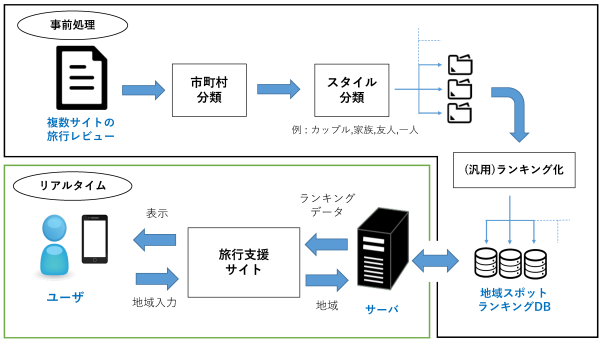

図 1　システム構成

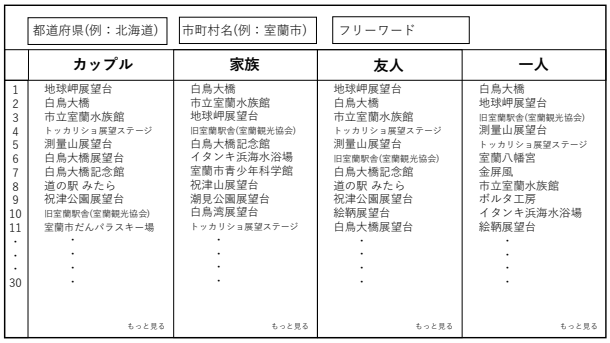

図 2　e 旅行：システムイメージ

3.2 市町村分類

市町村分類に関しては，各地域スポットのレビューが書き込まれている “じゃらん [3]” や “4travel [4]” などの旅行レビューサイトにて，北海道の各地域スポットのレビューを，約 180 件の市町村に分類し，収集する．

3.3 旅行スタイル分類

旅行スタイル分類に関しては，旅行スタイル別のレビュー数を増加させるために，旅行レビューサイトにて投稿されているレビューから旅行スタイルが判明していないレビューも旅行スタイル別に自動分類する．“じゃらん” の旅行レビューサイトでは既に旅行スタイルのタグが付与されているレビューが数多く存在している．レビューに付与されているタグは『カップル・夫婦』,『家族』,『友達同士』,『一人』,『その他』の 5 種類あり，このタグが付与されているレビューは 3.1 節で定義した 4 種類の旅行スタイル別へそのまま分類する．そして，タグが付与されていないレビューはレビューの内容により，それぞれの旅行スタイルにパターンベースの手法で分類する．

しかし，旅行スタイル別に分類する際，『その他』のタグが付与されている様に，本稿における旅行スタイル分類ではレビューによって，どの旅行スタイルにも属さない可能性がある．そこで，タグが付与されていないが『その他』に分類される様

レビュー r	連想語集合
水族館で**子供**が遊べる遊具がありました。ペンギンを見ることができ**娘**が大喜びでした。	家族，家族連れ，**子供**，子ども，子どもたち，子ども達，**娘**，父，父親，祖母，大人，ベビーカー

図 3　レビュー r と連想語集合『家族』の例

なレビューを，旅行スタイル別のレビューとして，誤分類してしまうケースが増加する可能性があり，旅行スタイルに基づいたその地域特有のスポットを抽出する際，精度に影響を及ぼす可能性があるため，どの旅行スタイルにも属さないレビューやそれに近しいレビューを出来る限り分類しないようにする．

3.4　地域スポットのランキング

ランキングに関しては，各市町村に存在する地域スポット及び潜在的なスポットを旅行スタイル別に分類されたレビューから抽出し，ランキング化する（図 2）．旅行スタイル別のレビューを基に単語を抽出することで，その旅行スタイルに影響を受けたスポットを抽出できると仮説を立てた．地域スポットなどは固有の単語であると考えられるので，形態素解析を行うため，MeCab とシステム辞書には mecab-ipadic-NEologd を使用し，『固有名詞』を抽出する．抽出した地域スポットのランキングを行うため，単語に重要度を付与し，その重要度に基づいてランキングを作成し，データベースを構築する予定である．

また，レビューから固有名詞を抽出する際，抽出された単語にノイズが含まれる可能性がある．特に「○○は “室蘭市” の中でも」など地域の名前はよく出現し易い単語である．地域名などはストップワード等の手法で除去する必要がある．

3.5　旅行支援サイト

旅行支援サイトに関しては，システムイメージのインターフェースを図 2 に示している．ユーザには都道府県や市町村を入力してもらい，入力された内容に合わせ，地域スポットランキングデータベースをサーバから読み込み，各旅行スタイルのランキングをユーザに提示する．「フリーワード」は入力することで，地域スポットの種類（観光，グルメ，観戦など）や季節を考慮し，地域スポットをリランキングすることが出来る．

4.　提 案 手 法

本稿では，旅行スタイルに基づき，あるレビューを分類する手法を議論する．著者らは「旅行スタイル s にて旅行した際のレビューには，その旅行スタイル s を表す単語が多く含まれる」という仮説から，パターンマッチに基づく手法を提案する．

旅行スタイルに基づき，あるレビューを分類するために，各旅行スタイル s の連想語集合 W_s を用意する．例として図 3 に，連想語集合 $W_{s=\text{family}}$ を示す．そして，あるレビュー r の中に，旅行スタイル s の連想語 $w \in W_s$ がどの程度含まれているかに着目し，旅行スタイル別にレビューを分類する．

この手法を基に，目的を持たせた 2 つの方式を提案する．目的の 1 つは，より相応しい度合いを求めるために，指示性を加味した方式である．もう 1 つは，連想語集合以上の幅広い単語を求めるために，拡張性を加味した方式である．以下にそれぞれの方式について記述していく．

指示性を加味した方式では，TF-IDF 法を参考に，単語の重要度を求める．単語出現頻度 TF と文書出現頻度 DF また，場合によっては逆文書頻度 IDF の値を求め，その値を利用し，どの旅行スタイル s が相応しいのか，連想語の重みを表す．

拡張性を加味した方式では，ある単語との類似単語，及び，その単語間の類似度を取得できる Word2Vec [5] という手法を用いて，連想語集合の単語を増加させることでパターンマッチする可能性を高める．

従って，ベースの手法とこれら 2 つの手法を組み合わせ，表 1 に示した，4 つの手法（**C**，**CI**，**CS**，**CIS**）を利用し，レビュー r を投稿したユーザが，その時ある旅行スタイル $s \in \{\text{couple}, \text{family}, \text{friend}, \text{alone}\}$ で旅行していたかを表す度合い，$\text{score}_r(s)$ を算出し，最もスコアが大きい旅行スタイルに分類する．以降より，$\text{score}_r(s)$ の算出方式を論述していく．

表 1　算 出 方 式

手法名	概要
C	単語の頻度のみで算出する手法
CI	単語の頻度と指示性を加味して算出する手法
CS	単語の頻度と拡張性を加味して算出する手法
CIS	単語の頻度と指示性，拡張性を加味して算出する手法

4.1　単語の頻度のみで算出する手法（C）

この手法では，レビュー r 中の単語と旅行スタイル s の連想語集合の単語を比較し，出現頻度の一番高い旅行スタイルに分類する．但し，頻度が同じ場合どちらにも分類を行わない．あるレビュー r の中に，旅行スタイル s の連想語集合内の単語が幾つ含まれているかに着目して $\text{score}_r(s)$ を算出し，各旅行スタイルに分類する．以下に示す式中の W_s は旅行スタイル s の連想語集合を示しており，$\text{wc}_r(w)$ は，旅行スタイル s の連想語集合の集合 W_s に含まれる単語 w が，レビュー r 内で何回出現したかを示している．

$$\text{score}_r(s) = \sum_{w \in W_s} \text{wc}_r(w)$$

4.2　手法 C に指示性を加味した手法（CI）

前節の手法では，あるレビュー r の中に，旅行スタイル s の連想語集合の単語が幾つ含まれているかに着目して $\text{score}_r(s)$ を算出した．この時，旅行スタイル s の連想語集合の中に，連想語集合内の単語がレビュー r 内に存在した際，そのレビュー r の旅行スタイルが s である確度を高くするような単語 w とそうでない単語があると考えられた．この考えに基づき本節では，単語 w の，レビュー r に対する旅行スタイル s への指示度 $\text{idct}_{r,s}(w)$ も加味した推定モデルを検討する．

$$\text{score}_r(s) = \sum_{w \in W_s} \text{wc}_r(w) \cdot \text{idct}_{r,s}(w)$$

4.2.1 指示性の高い単語抽出

次に，旅行スタイル s で旅したユーザ群によって投稿されたレビュー集合（以降，旅行スタイル s のレビュー集合）の重要語を推定する手法を検討した．著者らは，単語 w が，旅行スタイル s のレビュー集合の重要語であるか否かは，そのレビュアーが訪れた場所にも依存するという点に着目した．

例えば，地域 p = 北海道留寿都村 の主な観光目的地は「ルスツリゾート」である．すなわち，地域 p = 留寿都村 に旅行スタイル s = family で訪れたユーザ群が投稿するレビュー集合の特徴語としては，連想語集合 $W_{s=\text{family}}$ の中でも，特に「子供」等の単語が現れ易くなるであろう．また，地域 p = 北海道登別市 の主な観光目的地は「登別温泉」である．すなわち，地域 p = 登別 に旅行スタイル s = family で訪れたユーザ群が投稿するレビュー集合の特徴語としては，連想語集合 $W_{s=\text{family}}$ の中でも，特に「祖父母」等の単語が現れ易くなるであろう．

他方，地域 p = 北海道札幌市 の場合，その観光目的地は分散する．すなわち，地域 p = 札幌 に旅行スタイル s = family で訪れたユーザ群が投稿するレビュー集合の特徴語としては，連想語集合 $W_{s=\text{family}}$ 内の各単語 w が一様に現れるであろう．

従って次項より，レビュー r のレビュアーが訪れた地域 p に依存する場合とそうでない場合の 2 つを考慮し，単語 w の，レビュー r に対するスタイル s への指示度 $\text{idct}_{r,s}(w)$ の算出方式を議論する．具体的には，TF-IDF 法に倣い，単語 w の，レビュー r に対するスタイル s への指示度 $\text{idct}_{r,s}(w)$ を，旅行スタイル s のレビュー集合における単語 w の特徴量とみなし，tf（4.2.2）と idf（4.2.3）の 2 つの尺度の積から算出する．

$$\text{idct}_{r,s}(w) = \text{tf}_{r,s}(w) \cdot \text{idf}_{r,s}(w)$$

4.2.2 指示度 $\text{idct}_{r,s}(w)$ における因子 $\text{tf}_{r,s}(w)$ について

指示度 $\text{idct}_{r,s}(w)$ における因子 $\text{tf}_{r,s}(w)$ としては，次式の通り 4 種類の算出方式を定義した（表 2）．但し，式中の $|R_s|$ と $|R_{s,p}|$ は，それぞれ旅行スタイル s のレビュー集合と地域 p の旅行スタイル s のレビュー集合の要素を表している．また，$\text{dc}_{R_s}(w)$ は，旅行スタイル s のレビュー集合 R_s 内の単語 w の文書（レビュー）出現回数を示している．

$$\text{TF}_s(w) = \frac{\sum_{r \in R_s} \text{wc}_r(w)}{\sum_{\forall w'} \sum_{r \in R_s} \text{wc}_r(w')} \in [0,1]$$

$$\text{TF}_{s,p}(w) = \frac{\sum_{r \in R_{s,p}} \text{wc}_r(w)}{\sum_{\forall w'} \sum_{r \in R_{s,p}} \text{wc}_r(w')} \in [0,1]$$

$$\text{DF}_s(w) = \frac{\text{dc}_{R_s}(w)}{|R_s|} \in [0,1]$$

$$\text{DF}_{s,p}(w) = \frac{\text{dc}_{R_{s,p}}(w)}{|R_{s,p}|} \in [0,1]$$

4.2.3 指示度 $\text{idct}_{r,s}(w)$ における因子 $\text{idf}_{r,s}(w)$ について

指示度 $\text{idct}_{r,s}(w)$ における因子 $\text{idf}_{r,s}(w)$ は次式の通り算出する．但し，X は任意のレビュー集合を示しており，本稿では 7 種類のレビュー集合を定義した（表 3，図 4）．

$$\text{idf}_{r,s}(w) = \frac{|X|}{\text{dc}_X(w) + 1}$$

表 2　指示性の因子 $\text{tf}_{r,s}(w)$ の 4 種類の定義

	文書集合
TF_s	旅行スタイル s の単語出現頻度
$\text{TF}_{s,p}$	地域 p で旅行スタイル s の単語出現頻度
DF_s	旅行スタイル s のレビューにおける文書出現頻度
$\text{DF}_{s,p}$	地域 p で旅行スタイル s のレビューにおける文書出現頻度

表 3　指示性の因子 $\text{idf}_{r,s}(w)$ に用いる 7 種類の文書 X

X	文書集合
$X0$	使用しない
$X1$	全レビュー集合 R
$X2$	R における R_s の差集合 $R \backslash R_s$
$X3$	R における $R_{s,p}$ の差集合 $R \backslash R_{s,p}$
$X4$	R_p における $R_{s,p}$ の差集合 $R_p \backslash R_{s,p}$
$X5$	R における R_p の差集合 $R \backslash R_p$
$X6$	R における R_s と R_p の和集合との差集合 $R \backslash (R_s \cup R_p)$

R_s：旅行スタイル s のレビュー集合

R_p：地域 p のレビュー集合

$R_{s,p}$：地域 p で旅行スタイル s のレビュー集合

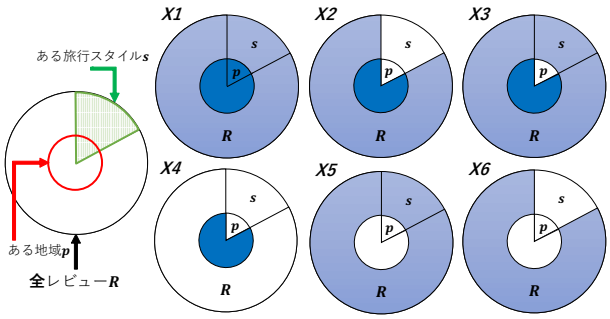

図 4　$\text{idf}_{r,s}(w)$ に用いる文書集合 X のイメージ図

4.3 手法 C に連想語集合の拡張性を加味した手法（CS）

手法 **C** の連想語集合の単語だけでは分類されるレビューに限りがあり，再現率を上げるため，連想語集合の単語を増やすこと（拡張性）を目的とする．Word2Vec を用いて連想語集合の単語より，類似単語及びその単語間の類似度を取得して，連想語集合に追加し，新しい連想語集合を作成する．但し，初めから連想語集合内に存在する単語は類似度を 1.0 とし，追加される単語が既出の類似単語であった場合，類似度はより高い類似度を優先する．その追加された類似度を加味して，手法 **C** で求めた単語出現頻度及び類似度を用いて $\text{score}_r(s)$ を算出し，各旅行スタイルに分類する．以下に示す式中の $\text{sim}_s(w)$ は旅行スタイル s における Word2Vec の手法を用いて作成された新たな連想語集合の単語 w とその類似度である．

$$\text{score}_r(s) = \sum_{w \in W_s} \text{wc}_r(w) \cdot \text{sim}_s(w)$$

4.4 指示性と拡張性の双方を加味した手法（CIS）

この手法では，ベース及び 2 つの手法を考慮し，$\text{score}_r(s)$ を算出する．指示性，拡張性は共に違うことを目的としているが，合わせることにより各算出方式で足りていなかった部分を補うことが出来ると予想できる．特に，指示性は単語の少なさから存在する再現率の低さを補うため拡張性と組み合わせ，拡張性

表 4　旅行スタイル分類精度結果

	Precision			Recall			F-measure		
手法	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F
C	0.670	0.095	0.166	0.670	0.095	0.166	0.670	0.095	0.166
CI	0.673	0.099	0.173	0.673	0.099	0.173	0.673	0.099	0.173
CI′	**0.682**	0.101	0.176	0.682	0.101	0.176	0.682	0.101	0.176
CS	0.586	0.105	0.179	0.256	0.247	0.251	0.402	0.231	0.293
CIS	0.613	0.107	0.182	0.368	0.241	0.292	0.368	0.241	0.292
CIS′	0.624	0.109	0.185	0.330	**0.298**	0.313	0.398	0.261	**0.315**

は様々な単語が増えることで起きる誤分類の弊害による適合率の低さを補うために指示性と組み合わせる．

$$\mathrm{score}_r(s) = \sum_{w \in W_s} \mathrm{wc}_r(w) \cdot \mathrm{idct}_{r,s}(w) \cdot \mathrm{sim}_s(w)$$

5. 評価実験

5.1 実験概要

まず，予め正解の旅行スタイルが付与されているテストレビューを用意する．これらのレビューは“じゃらん”から 2019 年 10 月 8 日に収集した．その時の旅行スタイル別のレビュー総数は，『カップル』72,196 件，『家族』63,579 件，『友人』34,445 件，『一人』34,763 件であり，合計 204,983 件であった．そして，これまで提案してきた 4 手法（**C**，**CI**，**CS**，**CIS**）で，テストレビュー集合を旅行スタイルで分類し，正解の旅行スタイルと比較することで，精確に分類できているかを評価する．但し，その評価尺度には，適合率，再現率，F 値を用いる．

5.1.1 連想語集合の作成

本稿では，旅行スタイルとして『カップル』,『家族』,『友人』,『一人』の 4 種類を定義し，各旅行スタイルの連想語集合に関しては著者らが手動で作成した．但し，その時の旅行スタイル別の連想語数は，『カップル』18 個，『家族』27 個，『友人』18 個，『一人』15 個である．

5.1.2 Word2Vec のモデル準備

旅行スタイル別の単語 1 つ 1 つの類似度を高めるため，本稿では，旅行スタイル別に異なった Word2Vec モデルを用いた．モデルの学習には，“じゃらん”の旅行スタイル別のレビューから取得した名詞・動詞・形容詞・句読点・句点を用いた．その際の単語の取得には 3.4 節で用いた MeCab を使用した．但し，学習させる文書としては，上記で述べたテストデータをそのまま用いた．また学習時のパラメータには，Skip-gram モデルを用いて，その window には 10 を与えた．さらに，次元数は 100，125，150，175，200 の 5 種類，学習回数は 20，25，30 の 3 種類，`min_count` は 1，5，10 の 3 種類で変化させたモデルもバリエーションとして作成した．

5.2 実験結果

4 章の提案手法で分類した結果を基に，表 4 に各々の手法の中で一番良かった適合率，再現率，F 値の結果を示す．

5.2.1 適合率に関する考察

はじめに，手法 **CI** と手法 **CIS** は $\mathrm{idf}_{r,s}(w)$ を加味しない $X0$ とし，手法 **CI′** と手法 **CIS′** は $\mathrm{idf}_{r,s}(w)$ を加味するものとする．$\mathrm{idf}_{r,s}(w)$ を考慮すると，最も適合率が高いのは指示性を加味した，手法 **CI′** であった．その時の，$\mathrm{tf}_{r,s}(w)$ と $\mathrm{idf}_{r,s}(w)$ はそれぞれ，$\mathrm{DF}_{s,p}$ と $X4$ であった．

表 5　手法 **C** と手法 **CI** の旅行スタイル別適合率

手法	カップル	家族	友人	一人	分類レビュー数
C	0.671	**0.711**	0.671	0.443	28,896
TF_s	0.665	0.689	0.642	0.433	31,088
$\mathrm{TF}_{s,p}$	**0.672**	0.700	0.690	**0.488**	30,258
DF_s	0.664	0.689	0.642	0.433	31,088
$\mathrm{DF}_{s,p}$	0.670	0.699	**0.693**	0.487	30,258

表 6　手法 **CS** と手法 **CIS** の平均適合率

	CS	**CIS**			
$\mathrm{tf}_{r,s}(w)$	-	TF_s	$\mathrm{TF}_{s,p}$	DF_s	$\mathrm{DF}_{s,p}$
`ave_Precision`	0.434	0.424	**0.463**	0.424	0.462

表 7　指示度の因子 $\mathrm{idf}_{r,s}(w)$ に関する精度結果

	CI′			**CIS′**Pre			**CIS′**F		
X	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F
$X0$	0.673	0.099	0.173	0.613	0.107	0.182	0.368	0.241	0.292
$X1$	0.680	0.100	0.175	0.622	0.109	0.185	0.386	0.253	0.306
$X2$	0.681	0.101	0.175	0.623	0.109	0.185	0.398	**0.261**	**0.315**
$X3$	0.680	0.100	0.175	0.622	0.109	0.185	0.387	0.254	0.307
$X4$	**0.682**	0.101	0.176	0.624	0.109	0.185	0.391	0.256	0.310
$X5$	0.680	0.100	0.175	0.622	0.109	0.185	0.419	0.241	0.306
$X6$	0.680	0.100	0.175	0.622	0.109	0.185	0.391	0.256	0.310

表 5 は手法 **C** と手法 **CI** の旅行スタイル別適合率と分類レビュー数である．但し，分類レビュー数とは『分類なし』ではなく，4 種類の旅行スタイルいずれかに分類されたレビューである．表 5 より，手法 **C** と手法 **CI** を見比べると，値はほぼ変化していないが，分類レビュー数が極僅かに変化している．例えば，「彼女と遊園地に行きました。家族連れが多かったです。」と書かれたレビューがあるとする．手法 **C** では頻度のみで分類するので，「家族連れ」が『家族』の単語で，「彼女」が『カップル』の単語である．この場合，スコアがどちらも 1 となり，同点であるので分類されない．同様に，スコアが同点となるようなレビューが 2,192 件存在した．一方，手法 **CI** では旅行スタイルごとに連想語集合の単語へ指示性を与えることで，分類できなかったレビューを分類可能とした．

また，TF_s と DF_s では手法 **C** と比較して，旅行スタイル別での適合率が下がっている． 適合率が下がった原因として考えられるのは，レビュー r 内の単語頻度が同じであったレビューがそれほど多くはないことと，連想語集合内の単語の出現頻度の差があることが挙げられる．後者の差とは，旅行スタイル『家族』のレビューの中で書かれた連想語集合における「子供」の単語数は 5,921 個で，旅行スタイル『友人』のレビューの中で書かれた連想語集合の「友達」の単語数は 873 個と，各連想語集合の単語により出現頻度に差が存在するということである．もし，『友人』との旅行レビュー内にこの 2 つの単語が書かれていた場合，本来ならば『友人』へ分類されるべきであるが，計算上，『家族』へと誤分類されてしまうことが考えられる．

しかしながら，ある旅行スタイルのある地域に着目する $\mathrm{TF}_{s,p}$ と $\mathrm{DF}_{s,p}$ は手法 **C** と比べて，適合率が僅かに上回った．他の

手法の場合は，正しい旅行スタイルにて，ある地域 p で旅行スタイル別の連想語集合内の単語が出現しなかったとしても，それぞれの単語に頻度や重みが存在する．そのため，ある旅行スタイル s の連想語集合の単語が，ある旅行スタイル s 以外に出現してしまうと，違う旅行スタイルであるが単語の頻度や指示性が反映されて誤分類してしまう．一方，手法 $\mathrm{TF}_{s,p}$ と手法 $\mathrm{DF}_{s,p}$ の場合は，ある旅行スタイル s のある地域 p の単語またはレビューの出現頻度であるので，ある地域 p で旅行スタイル s の連想語集合の単語が存在しなかった際，頻度や指示性が存在しないため反映されず誤分類を防ぐ．よって，適合率が僅かに上回ったと考えられる．

表 6 は表 5 の計算式の各々に拡張性を加味した手法 **CS** と手法 **CIS** の平均適合率を示しており，90 個の学習モデルから適合率を求め，その平均適合率を算出した．表 6 より，やはり TF_s と DF_s の 2 つは手法 **CS** よりも適合率がやや下回っているが，一方，ある旅行スタイル s のある地域 p に着目する $\mathrm{TF}_{s,p}$ と $\mathrm{DF}_{s,p}$ の 2 つは手法 **CS** よりも適合率が上回っている．

表 4 より，手法 **CI′** と手法 **CIS′** は僅かではあるが，適合率が手法 **CI** と手法 **CIS** よりも上回っていることがわかる．また表 7 は，**CIS′**Pre が手法 **CIS′** の最大適合率を表し，**CIS′**F は最大 F 値を表している．表 7 より，様々な $\mathrm{idf}_{r,s}(w)$ を加味することで，適合率は僅かに向上するが，連想語集合の単語が同じ場合，$\mathrm{idf}_{r,s}(w)$ による大きな変化は殆ど見えない．同じ単語であると断定する理由は，手法 **CI′** では連想語集合は変化しないことと，手法 **CIS′**Pre では使用されている学習モデルが次元数 200，学習回数 30，`min_count` = 1，取得類似単語数 5 件と全ての $\mathrm{idf}_{r,s}(w)$ のパターンにて同じであったためである．連想語集合に変化があった場合，表 7 の手法 **CIS′**F の列を見ると，適合率が上下している．しかし，$\mathrm{idf}_{r,s}(w)$ を加味しない場合の手法 **CIS** と比べると，いずれかの $\mathrm{idf}_{r,s}(w)$ であったとしても，$\mathrm{idf}_{r,s}(w)$ を加味した場合の方が適合率は高いと言える．

以上より，適合率の向上には，指示性を加味する際，ある地域 p を考慮することは効果的であり，$\mathrm{idf}_{r,s}(w)$ も効果的である．

5.2.2 再現率に関する考察

まず，表 4 により，手法 **CS**，手法 **CIS** と手法 **C**，手法 **CI** を見比べると手法の目的の 1 つである，再現率の向上が確認できた．すなわち，Word2Vec の手法を用いることで連想語集合を，少なからず旅行スタイルに関する類似単語で増やすことが出来ていると言える．しかし，再現率が上がる一方で，適合率が著しく下がっていることが問題である．理由として考えられるのは，Word2Vec で類似単語を取得する際に，ノイジーな類似単語が追加されるためである．この場で述べているノイジーな単語とは 2 つある．1 つ目は，ある旅行スタイルだけに関係している単語ではなく，いずれの旅行スタイルにおいても出現する単語が追加される場合であり，例えば，代名詞である「私」や感情を表す「好き」，季節を表す「夏」などが挙げられる．2 つ目は，ある旅行スタイルではなく，その他の旅行スタイルの連想語集合内の単語が追加される場合であり，例えば，旅行スタイル『家族』の単語「家族」や「子供」などが挙げられる．適合率を著しく下げないようにするため，これらのノイジーな単語を，含まないようにする必要性がある．もしくは，含むとしても追加された単語を考慮しないようにする必要がある．

しかし，適合率が下がってしまったが，再現率の向上に拡張性を加味することは効果的であると言える．

5.2.3 F 値と学習モデルに関する考察

表 4 より，最も F 値が高いのは指示性，拡張性の両方を加味した手法 **CIS′** となった．しかし，手法 **CIS** と手法 **CS** の結果を見ると大差ない，もしくは F 値が低い結果となっている．本来望む結果の予想としては，手法 **CIS** が手法 **CS** よりも F 値が高い状態であった．この原因として考えられるのは，類似単語や類似度を求める際の学習モデルが違うので，連想語集合が微かに変化し，分類されるレビュー数に変化が生じているためである．特に，分類されたレビュー数は手法 **CS** が 117,697 件，手法 **CIS** が 134,527 件と約 2 万件弱違う．モデルに関しては，次元数 200，`min_count` = 10，取得類似単語上位 10 件と一緒であるが，学習回数は手法 **CS** が 25 回，手法 **CIS** は 30 回であり，モデルの違いが見受けられた．

どのモデルを使用するかにより，適合率と再現率が変わるため，最適なモデルを探す必要がある．

6. まとめと今後の研究課題

本稿では，ある地域に関する旅行レビューを旅行スタイル別へ分類する手法について提案した．評価実験によって，適合率に関しては，ある旅行スタイルだけではなく，ある地域も考慮して単語の重みを求めて，分類することが効果的であることがわかった．加えて，様々な $\mathrm{idf}_{r,s}(w)$ を用いることで分類精度が向上することがわかった．再現率に関しては，類似単語で旅行スタイル別の連想語集合を拡張することが効果的であることがわかった．しかし，ノイジーな単語を取得する危険も存在するため，ノイジーな単語を除去，もしくは含まないように工夫する必要がある．F 値に関しては，頻度，指示度，類似度を組み合わせることで向上することがわかった．

本稿の結果では，再現率が向上すると適合率が低下していたが，今後は再現率を向上させつつ，適合率を保持する必要がある．そのために，連想語集合を手動で作成していたので，自動化し，より旅行スタイルに適した連想語集合を作成することや，旅行スタイル独自の単語と単語の繋がりがあると考えられるのでレビューの係り受け解析を行い，新たなパターンを作成するなど課題が残されている．

文　献

[1] 西川 崇哉, 岡田 真, 橋本 喜代太, “レビュー文章の自動分類におけるテキストの前処理手法の検証,” 言語処理学会第 18 回年次大会発表論文集, pp.517–520 (2012).

[2] 滝川 真弘, 山名 早人, “特定分野における単語重要度計算手法の提案と短い文章における著者の専門性推定への適応,” 情報処理学会研究報告「自然言語処理」, Vol.2017-NL-233, No.15, pp.1–6 (2017).

[3] じゃらん, https://www.jalan.net/kankou/ (2019).

[4] 4travel, https://4travel.jp (2019).

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Advances in Neural Information Processing Systems 26, pp.3111–3119 (2013).