平 成 3 1 年 度　卒 業 研 究 論 文

題目　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に関する研究

指導教員　　服部 峻

提出者　　室蘭工業大学 情報電子工学系学科

氏　名　　川村　直輝

学籍番号　　16024047

提出年月日　　令和 2 年 2 月 13 日

目次

図目次

表目次

第 1 章

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．

本研究では，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検討する．また，地域スポットはグルメ，観光スポット，スポーツ観戦，ライブなどの様々なスポットを表す．

第 2 章

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，機械学習を使用し，旅行情報ポータルサイトのレビューを，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．しかし，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，ある特定の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．

第 3 章

提案システム

本章では，複数の旅行サイトからレビューを収集し，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ことにより，様々な旅行スタイルのユーザが目的地や行動で悩んだ際，手助けが出来る旅行支援サイトを自動生成する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

本研究における旅行支援サイト自動生成システムの構成を図 3.1 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を出来る限り網羅的に収集し，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（本研究では『カップル』,『家族』,『友人』,『一人』の 4 種類）に分類を行う．この分類された旅行スタイル別のレビューを用いることで，その旅行スタイルの特徴が反映された地域スポットを抽出でき，ランキング順に示す．このデータを基に図 3.2 のような視覚的に他の旅行スタイルと見比べることで，ユーザが目的地や行動で悩んだ際，直感的に訪れたい地域スポットを選択することが可能な，ユーザの旅行を支援するサイトを自動生成するシステムである．また，昨今では簡単に口コミやレビューを投稿することが可能であり，ある旅行スタイルの新たな旅行レビューが投稿された際，そのレビューをすぐに反映させるため自動生成することにより，常に新鮮な地域スポットを提示する．

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．

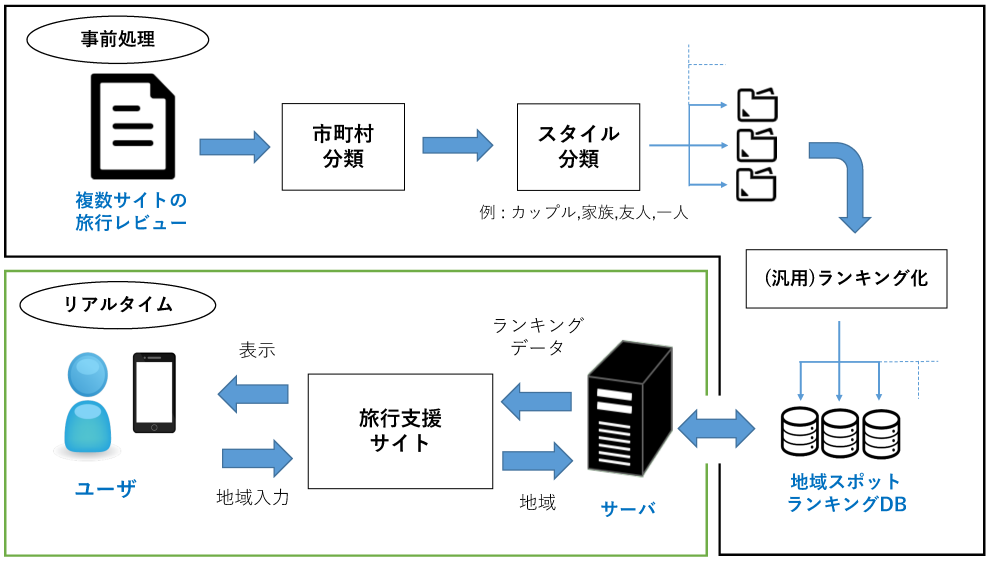

図 3.1　システム構成

3.3　旅行スタイル分類

旅行スタイル分類に関しては，次の 3.4 節における地域スポットのランキングで用いる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（表 3.1）も旅行スタイル別に自動分類する．“じゃらん” や “4travel” 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，本研究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．

表 3.1　“じゃらん”，“4travel” における旅行スタイルのタグの有無に依るレビュー数

	タグ有りレビュー数	タグ無しレビュー数
じゃらん	204,983（90.4%）	21,777　（9.6%）
4travel	45,633（75.2%）	15,009（24.8%）

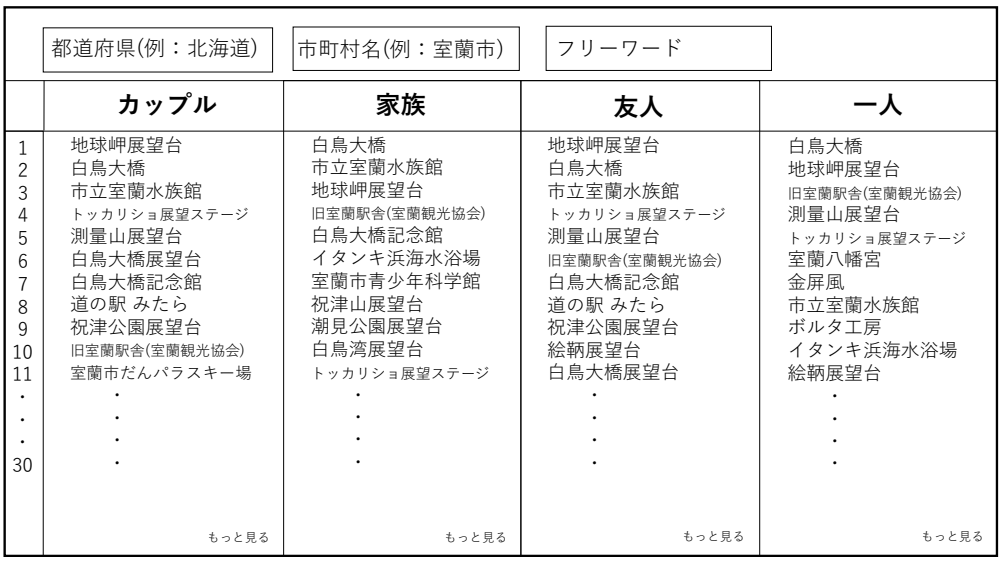

図 3.2　システムイメージ

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 3.2）．旅行スタイル別のレビューを基に TF-IDF 等を用いて，その地域特有の単語を抽出することで，その旅行スタイルに合わせた適切な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，形態素解析を行う際には，形態素解析エンジンである 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．地域スポットのランキングを行うため，抽出した地域スポットにスコアリングを行い，そのスコア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．そのため，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 3.2 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，サーバ上の地域スポットランキングデータベースに問い合わせ，データを読み込み，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．

第 4 章

提案手法：旅行スタイル分類

本研究では，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．「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のではないか」という仮説を立て，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．本研究においては，著者が手動で連想語集合を準備した．例としては図 4.1 に，旅行スタイル『家族』の連想語集合 $W_{s=\text{family}}$ の一部を示し，詳細は 5.1.1 項に記述している．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の単語の出現頻度（Count）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 種類の方式を提案する．目的の 1 つは，著者が手動で準備した連想語集合以上の幅広い単語を用いてカウントするために，拡張性を加味した方式である．もう 1 つは，旅行スタイルに対する相応しさの度合いを連想語集合の各単語それぞれについて計算し，より分類精度を高めるために，指示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．

拡張性（Scalability）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] というライブラリを用いて，連想語集合の単語を増加させることにより，あるレビューの中で連想語がパターンマッチする可能性を高める．

指示性（Indicatability）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，連想語の重みを求める．

従って，ベースの手法とこれら 2 種類の方式を組み合わせ，表 4.1 に示した，4 つの手法（**C**，**CS**，**CI**，**CSI**）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．

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 4.1　レビュー r と連想語集合『家族』の例

表 4.1　算出方式

手法名	**概要**
C	単語の頻度のみで算出する手法
CS	単語の頻度と拡張性を加味して算出する手法
CI	単語の頻度と指示性を加味して算出する手法
CSI	単語の頻度と拡張性，指示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，出現頻度が最も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 $\mathrm{score}_r(s)$ 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，$\mathrm{wc}_r(w)$ 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．

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

4.2　手法 C に連想語集合の拡張性を加味した手法（CS）

4.1 節のベース手法で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，手法 **CS** では，再現率を向上させるため，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から，類似単語及び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，著者が設定した，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，新たに追加される単語は，既に連想語集合内に存在する場合，類似度がより高い類似度を用いる．その追加された類似度を加味して，4.1 節で求めた単語出現頻度，及び，本節で新たに求めた類似度を用いて $\mathrm{score}_r(s)$ を算出し，各旅行スタイルに分類する．以下に示す式中の $\mathrm{scl}_s(w)$ 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．

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

4.3　手法 C に指示性を加味した手法（CI）

4.1 節のベース手法では，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 $\mathrm{score}_r(s)$ を算出した．この時，旅行スタイル s の連想語集合の中には，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 だけでなく，そうではない単語も含まれてしまっていると考えられた．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 $\mathrm{idct}_{r,s}(w)$ も加味した推定モデルを検討する．

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

4.3.1　地域による指示性の変化

まず，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．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 $s = \mathrm{family}$ 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 $W_{s=\mathrm{family}}$ 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 $s = \mathrm{family}$ 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 $W_{s=\mathrm{family}}$ 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 に対する旅行スタイル s への指示度 $\text{idct}_{r,s}(w)$ の算出方式を議論していく．具体的には，TF-IDF 法に倣い，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 $\text{idct}_{r,s}(w)$ 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（4.3.2 項）と idf（4.3.3 項）の 2 つの尺度の積から算出する．

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

4.3.2　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 4.2）．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数を表している．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）出現頻度を示している．

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

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

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

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

表 4.2　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

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

指示度 $\mathrm{idct}_{r,s}(w)$ における因子 $\mathrm{idf}_{r,s}(w)$ 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，本研究では 7 種類のレビュー集合を定義した（表 4.3，図 4.2）．

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

表 4.3　指示性の因子 $\mathrm{idf}_{r,s}(w)$ に用いる 7 種類の文書集合 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

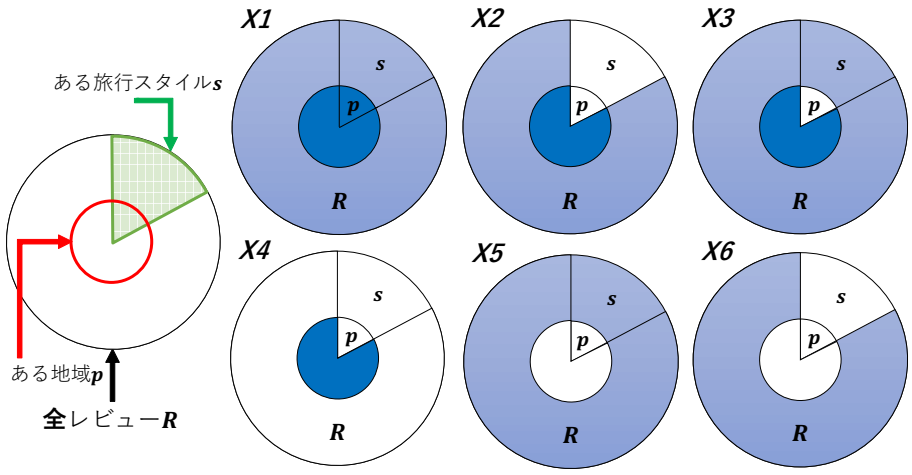

図 4.2　$\mathrm{idf}_{r,s}(w)$ に用いる文書集合 X のイメージ図

4.4　拡張性と指示性の双方を加味した手法（CSI）

この手法では，ベース及び 2 つの手法を考慮し，$\mathrm{score}_r(s)$ を算出する．拡張性，指示性は異なるコンセプトによって設計されているが，これら 2 つの手法を組み合わせることにより各算出方式ではカバーしきれない部分を補うことが出来ると期待できる．特に，拡張性は旅行スタイルに相応しくない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下を補うために指示性と組み合わせ，指示性は単語の少なさから存在する再現率の低下を補うため拡張性と組み合わせる．

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

第 5 章

評価実験

本章では，4 章で述べたパターンマッチによる 4 種類の提案手法を用いて，収集した旅行レビューを旅行スタイルへ分類し，その分類されたレビューに関して評価実験を行う．その結果を基に，旅行スタイル分類の精度を比較し，考察を行う．

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，“4travel” から 2019 年 10 月 11 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，“じゃらん” が『カップル』72,196 件，『家族』63,579 件，『友人』34,445 件，『一人』34,763 件，合計 204,983 件であり，“4travel” が『カップル』13,202 件，『家族』8,965 件，『友人』5,934 件，『一人』17,532 件，合計 45,633 件であった．そして，“じゃらん” のみのレビュー集合，“4travel” のみのレビュー集合，“じゃらん” と “4travel” の双方を使用するレビュー集合の 3 種類を用いる．これら 3 種類のレビュー集合は TF-IDF 法に倣った指示度 $\mathrm{idct}_{r,s}(w)$ の各出現頻度（TF，DF，IDF）を求める際にそれぞれ使用する．また，各レビュー集合をこれまで提案してきた 4 手法（**C**，**CS**，**CI**，**CSI**）で旅行スタイル別に分類し，レビュー投稿者（旅行者）が付与した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．また，手法 **CI** と手法 **CSI** は指示性の因子 $\mathrm{idf}_{r,s}(w)$ を加味しない $X0$ とし，手法 **CI**′ と手法 **CSI**′ は指示性の因子 $\mathrm{idf}_{r,s}(w)$ を加味するものとする．

5.1.1 連想語集合の作成

本研究では，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，著者が手動で，その旅行スタイルに近しいと考えることが出来る連想語を，旅行スタイル別に選び，その選んだ単語で連想語集合を作成する．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（表 5.1）．

表 5.1　各旅行スタイルの連想語集合と単語数

旅行スタイル	連想語集合内の連想語	単語数
カップル	カップル，恋人，夫婦，彼氏，彼女，彼，夫，妻，女房，二人，2 人，旦那，旦那さん，家内，デート，デートスポット，デートコース，初デート	18
家族	家族，家族連れ，子供，子ども，こども，子どもたち，子供たち，子供達，子ども達，子連れ，娘，息子，母，父，孫，お父さん，父親，義父，お母さん，母親，義母，父母，祖父，祖母，祖父母，大人，ベビーカー	27
友人	友人，友達，友だち，ともだち，仲間，同僚，仲間たち，仲間達，知人，ママ友，女友達，同級生，3 人，4 人，5 人，6 人，7 人，8 人	18
一人	一人，1 人，ひとり，ひとり旅，一人旅，個人，個人的，個人旅行，小生，出張，仕事，通勤，通学，バイク，ツーリング	15

5.1.2　Word2Vec のモデル準備

旅行スタイル別の連想語集合を拡張する際，旅行スタイルを表す単語と取得する候補の単語の類似度を高めるため，本研究では，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・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，“じゃらん” の旅行スタイル別のレビュー集合をテストデータとしてそのまま用いた．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 章で提案した 4 種類の手法で分類した結果を基に，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．“じゃらん” のみのレビュー集合の結果は図 5.1 から図 5.3，“4travel” のみのレビュー集合の結果は図 5.4 から図 5.6，“じゃらん” と “4travel” の双方を使用したレビュー集合の結果は図 5.7 から図 5.9 である．

これらの図より，精度が最も良かった再現率と F 値の 2 種類の図はどのレビュー集合においても折れ線グラフの形が似ている．すなわち，再現率が上がると F 値も同様に上がっている．一方で，精度が最も良かった適合率だけの図 5.1 や図 5.4，図 5.7 を見ると，適合率が高いだけでは再現率や F 値は極僅かしか上がっていない．

5.2.1　拡張性の有用性に関する考察

まず，手法 **C** と手法 **CS** を見比べる．拡張性を加味した手法 **CS** は手法 **C** と比べ，再現率が向上していることが図 5.2，図 5.5，図 5.8 より確認できる．そのため，Word2Vec を用いることで連想語を増やすことができ，パターンのマッチ数の増加が確認することが出来た．また，再現率が向上する一方，手法 **C** と比べ，著しく適合率が下がっていることが図 5.2，図 5.8 より分かる．

表 5.2 と表 5.3 は “じゃらん” のみのレビューで分類した，手法 **C** と手法 **CS** による，システムが分類したレビューの旅行スタイル別での適合率を表している．表のヒットレビュー数とは，正解セットとしたテストレビュー集合が，分類されたレビューとヒット（マッチ）したレビューの総数である．特に，旅行スタイル『カップル』と旅行スタイル『一人』の分類レビュー数が増加しているが，正しい旅行スタイルには分類されていない．旅行スタイル『カップル』や旅行スタイル『一人』に多くレビューが分類されてしまう原因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．ここでのノイジーな単語とは 2 種類存在する．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，一人称である「私」や感情を表す「好き」，季節を表す「夏」，地域を表す「北海道」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」，旅行スタイル『カップル』の単語「恋人」などが挙げられる．実際に追加されてしまった単語には「場所」，「夜」，「スポット」，「。」などが含まれており，これらはいずれの旅行スタイルでも出現頻度が高い単語であるため，誤分類に大きな影響を与えたと考えられる．

しかし，“4travel” のみのレビュー集合を使用する際は手法 **C** と手法 **CS** を比べると，僅かに手法 **CS** の適合率が上回っていることが図 5.5 より分かる．原因としては 2 つ考えられ，1 つ目は手法 **C** の適合率が低いことから，設定した連想語が相応しくないという点であり，2 つ目は，全体の適合率の算出には全旅行スタイルのレビュー数の合算値を用いるため，分類レビュー数の多い旅行スタイルから結果が大きく受ける点である．後者について詳しく確認するため，次は “4travel” のみのレビューで分類した手法 **C** と手法 **CS** における，システムが分類したレビューの旅行スタイル別での適合率を表した，表 5.4 と表 5.5 で見比べていく．旅行スタイル『一人』へ分類されたレビューが 9 割以上を占めているため，『一人』の分類精度に大きく左右されている．加えて，旅行スタイル『カップル』は分類レビュー数が増加しているが，他の旅行スタイルへの分類レビュー数が減少しており，これらは旅行スタイル『一人』に多く分類されていることが分かる．旅行スタイル『一人』に追加されているノイジーな単語は「の」，「好き」，「北海道旅行」，「家族」などが含まれており，特に「の」という単語はいずれのレビューにおいても頻繁に使用される単語であるため，大きな影響があったと考えられる．

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ノイジーな単語については，分類への影響力を小さくする必要があり，且つ，それぞれに適した連想語を設定する必要がある．

表 5.2　“じゃらん” における手法 **C** による旅行スタイル別分類の内訳

正解＼システム	Couple	Family	Friend	Alone	分類レビュー総数
カップル （ヒットレビュー数）	**0.672** (2,225)	0.154 (511)	0.104 (343)	0.071 (234)	3,313
家族 （ヒットレビュー数）	0.152 (3,007)	**0.711** (14,012)	0.086 (1,700)	0.050 (992)	19,711
友人 （ヒットレビュー数）	0.121 (283)	0.125 (292)	**0.671** (1,573)	0.084 (196)	2,344
一人 （ヒットレビュー数）	0.257 (908)	0.163 (576)	0.136 (480)	**0.443** (1,564)	3,528

表 5.3　“じゃらん” における手法 **CS** による旅行スタイル別分類の内訳

正解＼システム	Couple	Family	Friend	Alone	分類レビュー総数
カップル （ヒットレビュー数）	**0.391** (8,167)	0.287 (5,997)	0.177 (3,701)	0.144 (3,007)	20,872
家族 （ヒットレビュー数）	0.190 (3,781)	**0.633** (12,617)	0.117 (2,333)	0.060 (1,204)	19,935
友人 （ヒットレビュー数）	0.219 (390)	0.196 (348)	**0.429** (763)	0.157 (279)	1,780
一人 （ヒットレビュー数）	0.370 (57,500)	0.275 (42,782)	0.169 (26,222)	**0.187** (29,107)	155,611

表 5.4　“4travel” における手法 **C** による旅行スタイル別分類の内訳

正解＼システム	Couple	Family	Friend	Alone	分類レビュー総数
カップル	**0.371**	0.174	0.127	0.332	558
（ヒットレビュー数）	(207)	(97)	(69)	(185)	
家族	0.221	**0.360**	0.118	0.302	3,184
（ヒットレビュー数）	(704)	(1,145)	(375)	(960)	
友人	0.237	0.191	**0.324**	0.247	518
（ヒットレビュー数）	(283)	(292)	(1,573)	(196)	
一人	0.269	0.153	0.101	**0.477**	1,441
（ヒットレビュー数）	(388)	(220)	(145)	(688)	

表 5.5　“4travel” における手法 **CS** による旅行スタイル別分類の内訳

正解＼システム	Couple	Family	Friend	Alone	分類レビュー総数
カップル	**0.349**	0.216	0.140	0.295	1,391
（ヒットレビュー数）	(486)	(300)	(195)	(410)	
家族	0.236	**0.407**	0.116	0.240	1,430
（ヒットレビュー数）	(338)	(583)	(166)	(343)	
友人	0.221	0.216	**0.314**	0.247	194
（ヒットレビュー数）	(43)	(42)	(61)	(48)	
一人	0.289	0.189	0.129	**0.393**	42,228
（ヒットレビュー数）	(12,210)	(7,960)	(5,459)	(16,599)	

5.2.2　指示性の有用性に関する考察

まず，手法 **C** と手法 **CI** を見比べる．指示性を加味した手法 **CI** は手法 **C** と比べ，僅かに適合率が向上していることが図 5.1，図 5.4，図 5.7 より分かる．また，“じゃらん”，“4travel”，“じゃらん” と “4travel” の各レビュー集合における手法 **C** と手法 **CI** における因子 $\mathrm{tf}_{r,s}(w)$ を考慮した適合率と分類レビュー数及び，ヒットレビュー数を表 5.6 から表 5.8 に示す．表 5.6 と表 5.7 を見ると手法 **C** と比べ，因子 $\mathrm{tf}_{r,s}(w)$ は旅行スタイルと地域のどちらも考慮した $\mathrm{TF}_{s,p}$ と $\mathrm{DF}_{s,p}$ の方が，適合率の向上を確認できる．

しかし，旅行スタイルだけを考慮した TF_s と DF_s は指示性を加えることで曖昧なレビューを分類可能に出来たが，いずれのレビュー集合においても適合率が手法 **C** よりも下がった．この原因として，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存在すると，「家族連れ」が『家族』の連想語であり，「彼女」が『カップル』の連想語であるため，TF_s と DF_s はどちらの旅行スタイルが正しいのか指示性を適切に与えることが出来ず，正しく分類を行えなかったためである．これは，連想語集合内の単語の出現頻度の差が挙げられる．“じゃらん” のレビュー集合で確認していくと，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 つの単語が出現していた場合，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．

一方，手法 **C** では単語の出現頻度のみで分類するので，先程の例の場合，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．同様のレビューが “じゃらん” のレビューでは 2,192 件，“4travel” のレビューでは 382 件存在し，その曖昧なレビュー群は手法 **C** では考慮していないため，適合率が TF_s と DF_s よりも上回ったと考えられる．

また，旅行スタイルと地域に着目する $\mathrm{TF}_{s,p}$ や $\mathrm{DF}_{s,p}$ は手法 **C** と比べ，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．他方，$\mathrm{TF}_{s,p}$ と $\mathrm{DF}_{s,p}$ の場合は，ある地域 p である旅行スタイル s の連想語またはレビューの出現頻度と指示性を与える範囲を絞っているため，ある地域 p である旅行スタイル s の連想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ことが出来る．よって，適合率が僅かに上回ったと考えられる．

“じゃらん” と “4travel” のレビューを使用した場合においては表 5.8 を見ると，$\mathrm{TF}_{s,p}$ や $\mathrm{DF}_{s,p}$ は手法 **C** と比べ，どちらも適合率が下回った．原因として，上記で述べているように，もし，ある地域 p で，ある旅行スタイル s の連想語がたった 1 つ存在した場合，その連想語は指示性が与えられるため，他の旅行スタイルにおいて，ある旅行スタイル s の連想語が出現し

てしまうと，誤分類されてしまったと考えられる．

指示性の中でも特に，旅行スタイルと地域のどちらも着目することにより，確かに単語へ重みを付与することが出来たと考えることが出来る．

表 5.6　“じゃらん” の手法 **C** と手法 **CI** における適合率とレビュー数の比較

手法	$\mathrm{tf}_{r,s}(w)$	適合率	分類レビュー数	ヒットレビュー数
手法 **C**	–	0.671	28,896	19,374
手法 **CI**	TF_s	0.652	31,088	20,282
	$\mathrm{TF}_{s,p}$	**0.673**	30,258	20,361
	DF_s	0.652	31,088	20,271
	$\mathrm{DF}_{s,p}$	0.672	30,258	20,347

表 5.7　“4travel” の手法 **C** と手法 **CI** における適合率とレビュー数の比較

手法	$\mathrm{tf}_{r,s}(w)$	適合率	分類レビュー数	ヒットレビュー数
手法 **C**	–	0.387	5,701	2,208
手法 **CI**	TF_s	0.386	6,083	2,347
	$\mathrm{TF}_{s,p}$	**0.428**	5,658	2,420
	DF_s	0.386	6,083	2,347
	$\mathrm{DF}_{s,p}$	**0.428**	5,658	2,420

表 5.8　“じゃらん” と “4travel” の手法 **C** と手法 **CI** における適合率とレビュー数の比較

手法	$\mathrm{tf}_{r,s}(w)$	適合率	分類レビュー数	ヒットレビュー数
手法 **C**	–	0.624	34,597	21,582
手法 **CI**	TF_s	0.609	37,173	22,629
	$\mathrm{TF}_{s,p}$	**0.619**	36,774	22,776
	DF_s	0.609	37,173	22,628
	$\mathrm{DF}_{s,p}$	**0.619**	36,774	22,769

5.2.3　拡張性と指示性をハイブリッドした手法の有用性に関する考察

まず，図 5.3 や図 5.6，図 5.9 より，手法 **CS** や手法 **CI** と手法 **CSI** を見比べると，F 値が向上していることが分かる．F 値が向上しているということは，拡張性を加味することにより誤分類が数多く起きてしまう弊害及び，指示性におけるパターンマッチ数の少なさをある程度は補い合うことが出来たと考えられるが，どの図を見ても手法 **CSI** では適合率が大きく下がっている．これは，連想語が増加すると，単語出現頻度 TF と文書出現頻度の DF だけでは，それぞれの旅行スタイルの特徴的である単語に確実な重みを付与できていないためである．

しかし，指示性の因子 $\mathrm{idf}_{r,s}(w)$ を加味した手法 **CSI**′ を見ると適合率及び，再現率が向上している．表 5.9 は “じゃらん”，“4travel”，“じゃらん” と “4travel” の各レビュー集合において指示性の因子 $\mathrm{idf}_{r,s}(w)$ を加味した際のレビュー集合で最も適合率が良かった数値を表しており，これより手法 **CI**′ と手法 **CSI**′ は僅かではあるが，様々な指示性の因子 $\mathrm{idf}_{r,s}(w)$ を加味することで適合率が手法 **CI**（$X0$）と手法 **CSI**（$X0$）よりも上回っていることが分かる．また，“4travel” の手法 **CSI**′ の際の集合 $X2$ の場合を除き，大きな差は見られなかった．理由として考えられるのは，連想語集合の単語の変化が無いためである．特に，手法 **CI** では拡張性を加味していないので確実に変化していないと言える．表 5.10 には指示性の因子 $\mathrm{idf}_{r,s}(w)$ のレビュー集合に関する学習モデルの変化を表しており，手法 **CSI** の場合も，指示性の因子 $\mathrm{idf}_{r,s}(w)$ を加味した際の学習モデルと違いが見られず，拡張性を用いて取得する単語が殆ど同じである．“4travel” の手法 **CSI**′ の際の集合 $X2$（以降，4tra-**CSI**′-$X2$ と表す）の場合では，表 5.10 より学習モデルが唯一変化していることが確認でき，連想語集合が変わっているため，大きな変化が存在した．変化とは，分類レビュー数の増加はなかったが適合率が向上していることである．4tra-**CSI**′-$X2$ の分類レビュー数を見ると 5,150 件であり，手法 **C** や手法 **CI** と比べても少ないが，ヒットレビュー数は 2,951 件と手法 **C** や手法 **CI** と比べて数多く正しく分類できていたと言える．

以上より，拡張性，指示性どちらも加味することで連想語だけでは不足していた，再現率と適合率の向上及び，欠点であった誤分類の弊害やパターンのマッチ数の低下を補うことが可能である．

表 5.9　指示性の因子 $\mathrm{idf}_{r,s}(w)$ に関する適合率の比較

	じゃらん		4travel		じゃらん＋ 4travel	
集合	**CI′**	**CSI′**	**CI′**	**CSI′**	**CI′**	**CSI′**
$X0$	0.673	0.613	0.427	0.443	0.619	0.570
$X1$	0.680	0.622	0.426	0.449	0.618	0.578
$X2$	0.681	0.623	0.431	**0.573**	0.626	0.577
$X3$	0.680	0.622	0.431	0.448	0.625	0.578
$X4$	**0.682**	**0.624**	0.430	0.447	**0.627**	**0.579**
$X5$	0.680	0.622	0.431	0.448	0.625	0.578
$X6$	0.680	0.622	**0.432**	0.448	0.625	0.578

表 5.10　指示性の因子 $\mathrm{idf}_{r,s}(w)$ に関する手法 **CSI′** の学習モデルの変化

	じゃらん		4travel		じゃらん＋ 4travel	
集合	次元数	学習回数	次元数	学習回数	次元数	学習回数
$X0$	200	30	200	20	200	30
$X1$	200	30	200	20	200	30
$X2$	200	30	**150**	**30**	200	30
$X3$	200	30	200	20	200	30
$X4$	200	30	200	20	200	30
$X5$	200	30	200	20	200	30
$X6$	200	30	200	20	200	30

※学習モデル共通点：min-count は 1，取得類似単語数は 5

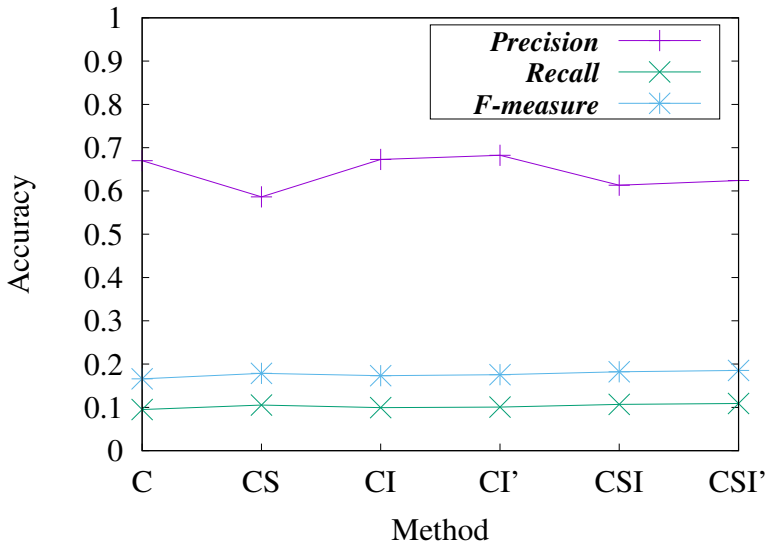

図 5.1　“じゃらん” のレビューのみの各手法における最も適合率が高かった際の評価

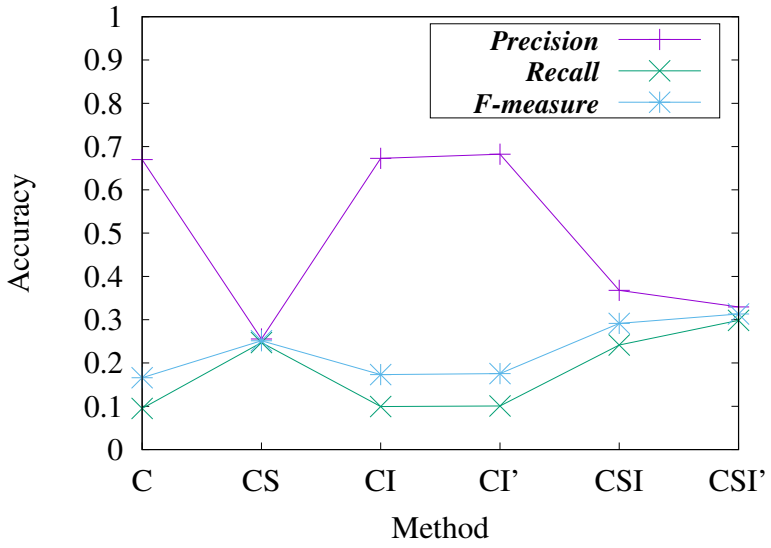

図 5.2　“じゃらん” のレビューのみの各手法における最も再現率が高かった際の評価

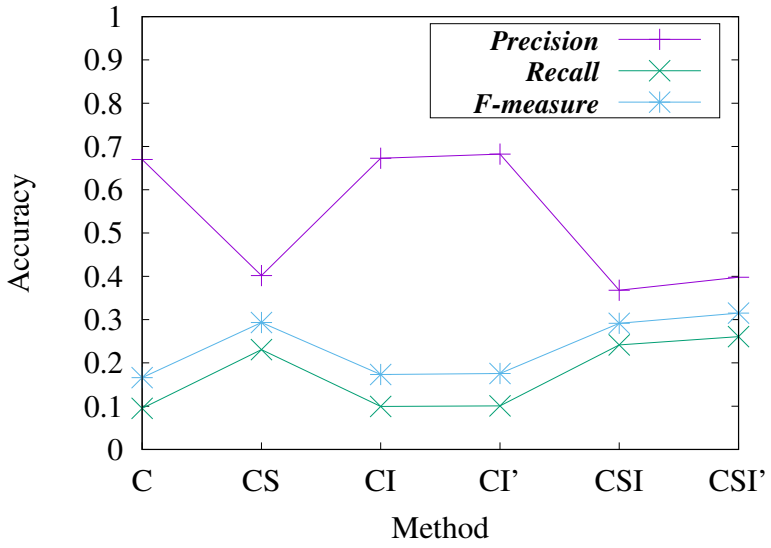

図 5.3　“じゃらん” のレビューのみの各手法における最も F 値が高かった際の評価

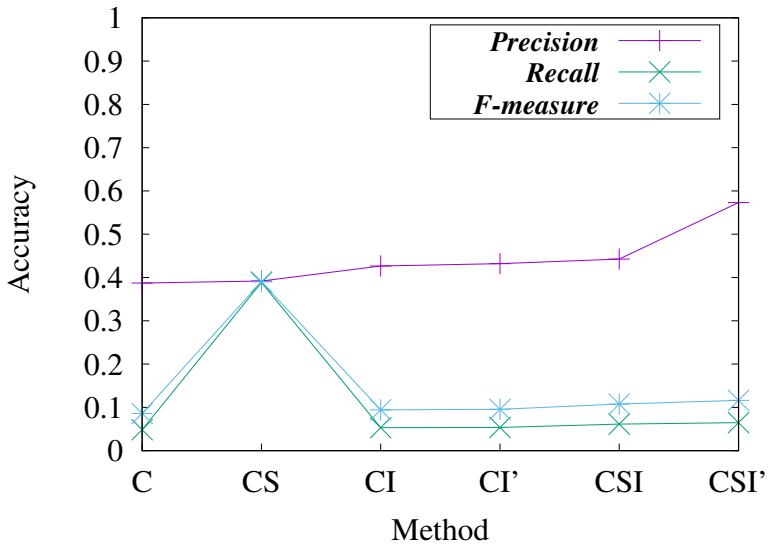

図 5.4　“4travel” のレビューのみの各手法における最も適合率が高かった際の評価

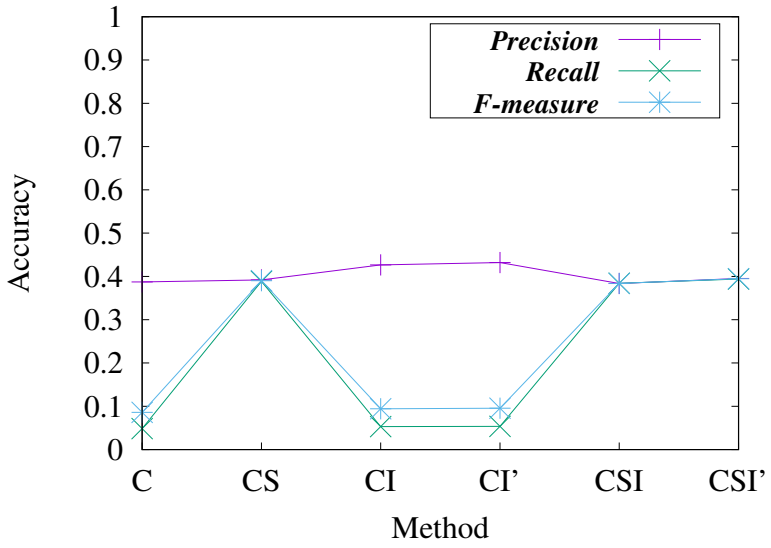

図 5.5　“4travel” のレビューのみの各手法における最も再現率が高かった際の評価

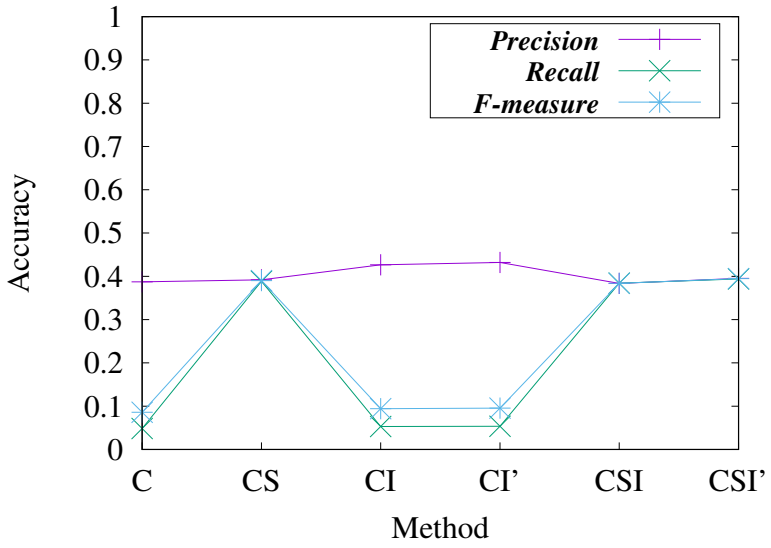

図 5.6　“4travel” のレビューのみの各手法における最も F 値が高かった際の評価

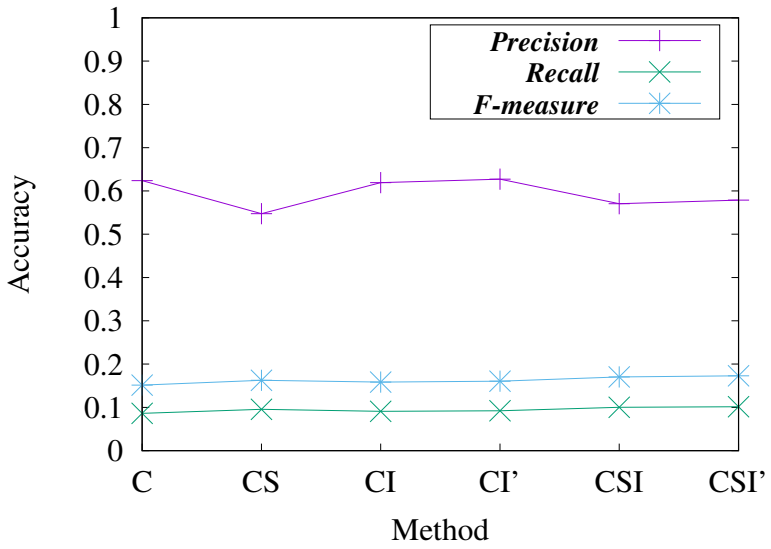

図 5.7　“じゃらん” と “4travel” のレビューの各手法における最も適合率が高かった際の評価

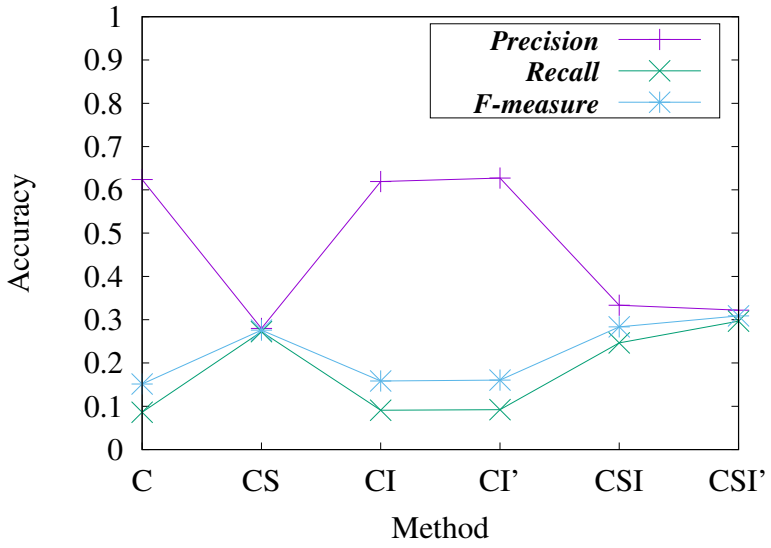

図 5.8　“じゃらん” と “4travel” のレビューの各手法における最も再現率が高かった際の評価

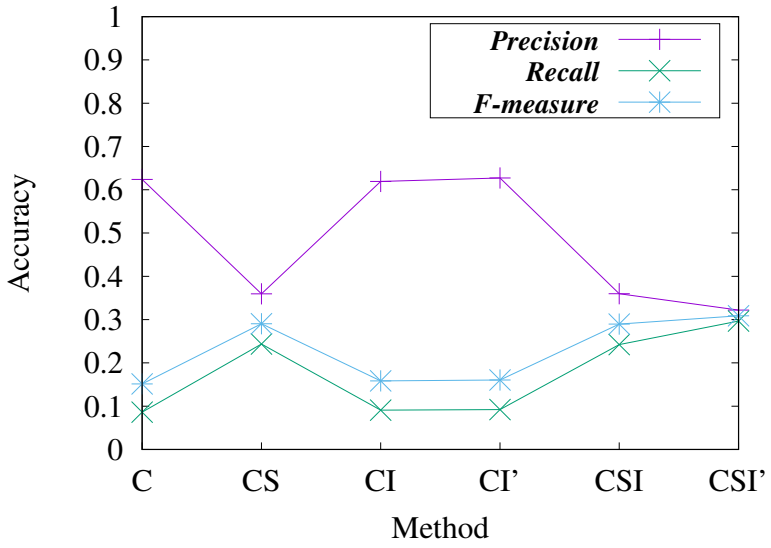

図 5.9　“じゃらん” と “4travel” のレビューの各手法における最も F 値が高かった際の評価

第 6 章

まとめと今後の課題

本研究では，複数サイトの地域スポットのレビューを出来る限り網羅的に収集し，それらのレビューを 4 種類の旅行スタイルへパターンマッチの手法を用いて自動で分類する手法について提案した．また，旅行支援サイトの自動生成についても検討した．

提案手法に関して，TF-IDF 法に倣い旅行レビューの分類を行い，評価実験によって，拡張性と指示性を加味することで，それぞれの欠点を補い合えたことにより，どちらか一方の時よりも F 値の向上が確認できた．拡張性においては，Word2Vec を用いることで旅行レビューに対応した単語を取得することが出来，再現率が向上した．旅行スタイル別のレビューを用いて Word2Vec の学習モデルを作成すると旅行レビューに対応した単語は取得できているが，各旅行スタイルに適した単語だけを取得することは出来なかったため，学習させる文書を考慮する必要がある．また，取得した単語が旅行スタイルに適切ではないと明確に判断できるようなノイジーな単語が出現する可能性がある．指示性においては，単語出現頻度 TF や文書出現頻度 DF を用いて検討したが，レビュー内で同じ単語が使用される可能性が低いため，大きな違いは存在しなかった．しかし，ある地域に着目し頻度を求めることで，適合率が向上した．ある地域に特定することにより，連想語へ指示性を付与する範囲を絞ることが出来，フィルターが自然と掛かっていると考えられる．逆文書頻度である IDF はどの文書集合が最良であると断言できないが，指示性の因子 $\mathrm{idf}_{r,s}(w)$ を加味した場合においては指示性の因子 $\mathrm{tf}_{r,s}(w)$ が旅行スタイルと地域のどちらも考慮した頻度の方がより良い結果となったため，新たな尺度として加味することは効果がある．

本研究の現状の社会的貢献は，旅行スタイルが判明していない旅行レビューを特定の旅行スタイルに適したレビューとして分類を行えたことである．今後の社会的貢献としては，本研究で使用した旅行レビューサイトだけではなく，Twitter や Instagram などの SNS の情報や Google マップやトリップアドバイザーなどのレビューなどを旅行スタイルへ分類可能にすることで，より旅行スタイルに適した地域スポットの抽出を行うことが出来，且つ，「多くの人には知られていないが知る人ぞ知る」といった，潜在的なスポットの抽出が可能であると考えられる．また，それらの地域スポットをランキングで表示する際に，抽出した地域スポットへ詳細なデータ（例えば，グルメや観光，旅行スタイルに適している度合いなど）を与えることで，簡単なワードや検索フィルターを掛けることにより，様々なユーザが目的地や行動で悩ん

だ際，直感的に選択できるようなランキングを提示し旅行を支援することである．

本研究の現状の技術的貢献は，TF-IDF 法を用いて文章分類を行っており，要素を文書のみとせず 2 つの要素（本研究では旅行スタイルと地域）を考慮することで，各単語へより適切な特徴量を付与できることが確認できた．今後の技術的貢献としては，TF-IDF 法の TF ではなく IDF の有効性をより詳しく解析することによって，2 つの要素を考慮することの重要性が更に明確になると期待される．また，文書数が少ない場合において，明確な区分が可能であるならば，要素を 2 つだけではなく更に多くの要素として考慮することで，それぞれにより相応しい特徴量を単語へ付与できるのではないかと考えられる．

今後の課題として，再現率向上に伴い適合率が著しく低下してしまったため，旅行スタイルごとの連想語集合拡張による適合率の低下を防ぐ必要がある．現在は連想語集合を手動で作成しているため，本研究で定義した 4 種類の旅行スタイル以外の新たな旅行スタイルを追加する際などを考慮し，連想語集合の作成を自動化する必要性がある．そのため，TF-IDF 法や Okapi BM25 などを用いて単語の重要度を測り，相応のパラメータやいずれかのフィルターを掛けることで，各旅行スタイルに適した単語を連想語集合として抽出できると考えられる．また，Word2Vec だけではノイズとなる単語が多く取得できる可能性が存在するため，単語と単語の繋がりを解析，もしくは，文章の意味を解析することにより Word2Vec 以外の拡張性を加味した手法を試みる必要性があると考える．その他に，本研究では主に地域を「北海道」の市町村とし，旅行スタイルの分類を行ったが，これを更に「日本」の「北海道」の市町村と地域の範囲を拡げて旅行スタイルの分類を行うと，「北海道」という地域の特色が加味されたレビューの分類が可能であると期待できる．最後に，旅行スタイル別のレビューを基に地域スポットのランキングを行うことが出来ていないため，地域スポット抽出方法を考える必要性がある．

謝辞

本研究に際して，様々なご指導を頂きました服部峻助教に厚く御礼申し上げます．また，日常の議論を通じて多くの知識や示唆を頂いた服部研究室の皆様にも深く感謝の意を表します．また，実験に使用した Java と Python のライブラリ製作者の皆様にも感謝致します．そして，本研究で用いた，フリーのオープンソースソフトウェアを提供している皆様に感謝致します．

参考文献

[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).