Vol.2015-DBS-162 No.13
2015/11/26

観光を対象としたレビューからの耳より情報抽出

阪井 奎伍[1,a)]　灘本 明代[2,b)]

概要：今日トリップアドバイザーやフォートラベルに代表されるように観光地のレビューサイトが普及している．これらレビューサイトは観光地での経験に基づくレビューが多く書かれており，観光地などの公式サイトにはない様々なお得と思われる情報が多数記載されている．しかしながら，有名な観光地ほどレビューの量は膨大であり，その中には不要な情報も多数記載されているためユーザにとって有用な情報を見つけることは困難である．そこで，本研究ではユーザが観光地のレビューを読んだときに「参考になった」「知って得をした」と感じる文を「耳より情報」と呼び，この耳より情報の抽出手法を提案する．具体的には，我々の提案する耳よりキーワードを含むレビューを文単位でクラスタリングし，そのクラスタの中心ベクトルを構成する文との類似度がある程度低い文を耳より情報として抽出し，ユーザに提示する．

1. はじめに

現在，インターネット上には観光に関する様々な情報が多数存在している．人々は旅行に行く前にこれらインターネット上の観光に関する様々な情報を収集し旅行の計画を立てることが多数ある．この時，旅行先が慣れていない土地や初めて訪れる土地であれば，事前に十分に情報を収集しておくことでより快適な旅行に繋がると考えられる．しかしながら，検索エンジンを用いて観光地の情報を収集しようとしても，観光地の様々な公式サイトでは基本的な情報が多く，それ以上の情報を見つけることは困難である．その結果，実際に観光地に訪れてから「〜しておけば」と後悔する場合が多数生じている．また，Q&A サイトでは質問に対してピンポイントな回答になるため，あらかじめユーザが知りたいと考えている内容以外の情報を収集することは困難である．また，検索結果が多いため 1 つの観光地に対する質問と回答を全て閲覧するには時間がかかってしまうことがある．

一方で，フォートラベル [1] やトリップアドバイザー [2] に代表されるように，様々な観光地のレビューサイトが普及してきている．これらのサービスでは実際に観光地に訪れたことのある人が，その観光地に訪れた経験を基にしたレビューを書いている．この経験に基づくレビューは観光地の公式サイトにある情報ではないお得な情報が多数記載されている場合が多いため，旅行前の観光地の情報収集においてとても役立つと考えられる．しかしながら，これらのレビューサイトでは有名な観光地ほどレビューの量が膨大であり，ユーザがレビュー全てを把握することは困難である．また，膨大なレビューの中には個人の感想等の不要な情報も多数あり，その中からユーザにとってお得と思われる情報を見つけることは困難である．そこで，我々はこのお得と思われる情報を自動でレビューサイトから収集できたら便利であると考え，このユーザがお得と感じる情報を自動で抽出し提示する手法の提案を行う．本研究では，このユーザがお得と思われる情報を「耳より情報」と呼ぶ．我々の提案する耳より情報は，「帰ろうとしたとき，休憩所があり，無料でお茶が飲めました」等の金銭や時間に関する情報，「自販機などはないから飲み水は下から用意して持って行ったほうが良い」等のあると便利と考えられる情報，「京都で車の運転はかなり気を使いますね，特に有名な場所は観光客が多いのでベビーカーや車椅子を使わない場合は公共の交通機関の利用を強く推奨します」等の交通手段のお得と感じる情報，「ほうが良い」や「推奨します」といった相手にアドバイスをしている情報を耳より情報の候補である有用な情報とし，その中で人々にあまり知られていない意外な情報やユーザが興味のある情報，未知の情報であると考える．本論文では，耳より情報を抽出するはじめの一歩として有用な情報であり，かつ意外な情報である情報を耳より情報として抽出する．具体的には，観光地のレビューから我々の提案する耳よりキーワードを含む文を文単位で抽出する．そして，抽出された有用な情報をキーワード毎にクラスタリングし，そのクラスタの中心ベクト

[1] 甲南大学　自然科学研究科
Konan University, Kobe, Hyogo 658-0072, Japan
[2] 甲南大学　知能情報学部
Konan University, Kobe, Hyogo 658-0072, Japan
[a)] m1524003@s.konan-u.ac.jp
[b)] nadamoto@konan-u.ac.jp

ルを構成する文との類似度がある程度低い情報を耳より情報として抽出する．

以下本論文では 2 章でレビューや旅行に関する関連研究を，3 章で耳より情報抽出手法について述べる．そして第 4 章では実験の結果と考察ついて述べ，5 章でまとめと今後の課題について述べる．

2. 関連研究

レビューに関する研究は種々ある．服部ら [3] は，mixi のコミュニティからイベントを対象とした耳より情報を抽出する手法を提案している．耳より情報を抽出する際のキーワードやキャッチフレーズを「提案・推薦」,「抑止抑制」,「現状状況説明」,「可能・不可能」の 4 つのタイプに分類し，タイプ別にキーワードを提案している．本研究では被験者実験より観光地のレビューを対象に耳より情報を人手で抽出し，観光地を対象とした耳より情報を抽出する際の耳よりキーワードを提案する．

河中ら [4] は，通販サイトのレビューに対してユーザにとって有用性の高いレビューのランク付け手法を検討している．ユーザにとって有用な情報を抽出する点は類似しているが，対象が通販であり，評価項目毎にランク付けしている点で異なる．

小林ら [5] は，ユーザの意見が生まれた理由が商品の重要な特徴であると考え,「条件」,「理由」,「態度」,「対象」の要素を定義し，Web 上の商品レビューから理由の書かれている部分の抽出手法を提案している．小林らの定義した 4 つの要素のうち「だと思う」といった「理由」や「なので」といった「態度」は本研究で提案する耳よりキーワードと類似しているが，本研究では耳よりキーワードにより抽出した情報にクラスタリングを行い，類似度を用いている点で異なる．

佐々木ら [6] は，商品レビューを対象に有用なレビューの特徴を分析し，機械学習で用いる素性の選択を行い，素性を適切に組み合わせることでレビューの有用性の判定を行っている．レビューの有用性を考慮している点は類似しているが，対象が商品であることや機械学習を用いている点で異なる．

石川ら [7] は，旅行先の宿のレビューにおいて，レビューを分類しユーザが必要とする情報を見つけやすくするし，さらにユーザの重視する評価項目によるランキング手法を提案している．ユーザにとって必要な情報を提示する点は類似しているが，評価点を用いて抽出している点が異なる．また，旅行に関する研究も種々ある．

遠藤ら [8] は，web 上に存在する膨大な観光情報に着目し，特定地域の有用な観光情報の自動抽出・融合を行うための特定地域に限定せず低コストに対象地域の有用な観光キーワードを自動取得手法を提案している．観光地のキーワードを用いて有用な情報を抽出する点は類似しているが，観光地の情報をレビューに限定している点やキーワードの取得手法については異なる．

藤井ら [9] は旅行ブログを対象に「買う」,「食べる」,「体験する」,「泊まる」,「見る」,「その他」の 6 種類のタイプへ自動的に分類し，タイプごとに地図上に提示する手法を提案している．観光情報を抽出する点は類似しているが，観光情報を分類分けし提示する手法に対し，本研究では有用な情報を抽出する点で異なる．

中嶋ら [10] は，旅行ブログから旅行先の地の名所を抽出し，各所付随情報として体験情報，評価表現，状態情報，名所の由来や歴史の 4 つを抽出する手法を提案している．体験情報は耳より情報と類似しているが，ブログの記事から抽出している点で異なる．

松本ら [11] は，観光地のレビューから特徴語を抽出しユーザに提示することでレビューを見ずに観光地の特徴を把握し旅行の際に訪れる観光地検索を支援する手法を提案している．レビューの特徴を抽出する点は耳よりキーワードの抽出と類似しているが，特徴をそのままユーザに提示することに対し耳より情報の抽出に用いている点で異なる．

安藤ら [12] は，楽天トラベルのレビューにおいて「良くも悪くもユーザの心をぐっと掴むような極端なレビュー」を集め，読み手の心に響く表現を「インパクトのある表現」と定義し，読み手の心を動かすような情報について分析している．読み手の心を動かすような情報は耳より情報となる場合があると考え，耳より情報を抽出する際のキーワードを作成する際の参考とした．

3. 耳より情報抽出手法

我々の提案する耳より情報抽出の手順を以下に示す．

(1) ユーザは耳より情報を取得したい観光地をクエリとして入力する．
(2) システムは，そのクエリを用いた検索結果のレビューを取得しこれらのレビューを文単位で分割する．
(3) (2) の中で耳よりキーワードを含む文を有用な情報として抽出する．
(4) (3) で抽出された有用な情報を，耳よりキーワード以外の名詞を用いてクラスタリングを行う．
(5) 各クラスタの中心ベクトルと各々のクラスタを構成する文の類似度を求める．
(6) 類似度がある閾値の幅内である文を有用な情報かつ人々にあまり知られていない意外な情報とし，これを耳より情報とする．

ここで，観光地の各々のレビューの中には個人の感想や意見，経験談，そして耳より情報と様々な情報が混在しているため，これらを分割する必要がある．そこでこれらのレビューは実際に見てみると文単位で分割可能であることが分かる．そこで本研究では取得したレビューを全て文単位で分割し，文単位で耳より情報の抽出を行う．文の分割

表 1　耳よりキーワード

耳よりキーワード例
おすすめ，推奨，のがいい，方がいい，適して，出来ます，知る人ぞ知る，避けたほうが，もってのほか，あきらめ，困りました，あらかじめ，改善，大切，許容範囲，疑問，詐欺，だけでなく，可能，時間余裕，時間半程，昼頃，%off，期間限定，長蛇の列，待たずに，空いてい，行列覚悟，無制限，先着順，整理券，値段，無料，参加費，割引，履物，必需品，水分補給，レンタル，使用可能，便利，自販機，途中に，地図，賑わう，駐車サービス，ロードマップ，アクセス，トイレ，車椅子，ベビーカー，足腰，接客

には，「。」，「.」，「！」，「!」，「♪」，「☆」，「★」，「(笑)」，「w」，タブまたは改行が 1 つ以上の場合にそこが文の終わりと考え分割する．またレビューの中には「小雨が降っていたので，足元も悪く，すべって躓いている方もいました」という文とこれに続く「なので歩きやすい靴で行くのをお勧めします」という文のように 2 文以上で耳より情報となる場合も考えられるが本論文ではこのような場合は考慮していない．

3.1　有用な情報の抽出

実際の観光地のレビューにある耳より情報には，耳より情報の要因となるキーワードが含まれていると考えられる．例えば伏見稲荷大社のレビューの場合，「帰ろうとしたとき，休憩所があり，無料でお茶が飲めました」という文がある．この場合，「無料」という語が金銭に関する有用な情報に繋がると考えられる．また，「自販機などはないから飲み水は下から用意して持って行ったほうが良い」という文の場合，「用意」という語が持ち物としてあると便利になる有用な情報に繋がり，「ほうが良い」という語が相手に対してアドバイスしている有用な情報に繋がると考えられる．

このように耳より情報の要因となるキーワードを耳よりキーワードと呼び，これらを抽出するために予備実験を行った．具体的には，まず対象とする観光地を決定するためにフォートラベルにある観光地の中から過去 3 年以内に 100 件以上のレビューがある観光地を無作為に 100 件取得した．そして，その観光地に対しての興味の有無と過去 3 年以内での訪問経験の有無について被験者 15 名の意見を聞き，100 件の観光地の中から興味有り訪問経験有り，興味有り訪問経験無しが各 3 名以上となる「伏見稲荷大社」，「京都水族館」，「水木しげるロード」，「嵐山」，「横浜中華街」の 5 件を対象の観光地と決定した．次にこの 5 件各々の観光地に対して興味有り訪問経験有り，興味有り訪問経験無しとなった被験者各 3 名の計 6 名が観光地のレビューの文を読み，耳より情報であると感じた文とそうでない文と 2 択で判断した．6 名の内，興味有り訪問経験有り，興味有り訪問経験無しで各 2 名以上となる文を耳より情報の教師データとし，この耳より情報の中から耳よりキーワードを人手で抽出した．**表 1** に耳よりキーワードの例を示す．表 1 の耳よりキーワードが観光地のレビューの文に含まれていれば有用な情報として抽出する．この有用な情報を耳より情報の候補とする．

3.2　意外な情報

公知ではないレアな情報はユーザにとって意外な情報である場合がありそれが耳より情報になると考え，3.1 節で抽出した有用な情報からさらに意外な情報を抽出することを行う．この時，公知な情報やあまりにもレアな情報は旅行者にとって有用でない場合がある．そこで本研究では，他の有用な情報と類似する話題を述べながら他の有用な情報とは少し異なることを述べているレビューをある程度レアな情報と考え，これを意外な情報とする．また有用な情報の中には様々な話題が存在すると考え，有用な情報を話題毎に分類し意外な情報を抽出する．つまりは，話題毎にクラスタリングを行い，そのクラスタ内の中心ベクトルからある程度離れている範囲にある文を耳より情報とする．具体的には，3.1 節で抽出した耳より情報の候補となる有用な情報全ての文を対象に juman[13] により形態素解析を行い名詞を抽出する．この時，単に名詞だけを抽出すると「期間限定」のような 2 つの名詞が連続して 1 つの名詞となる語が「期間」と「限定」に分かれてしまうため，名詞が連続して出現した場合に限り連続している名詞同士を連結し，1 つの名詞として抽出する．また，有用な情報抽出に用いた耳よりキーワードには名詞の語も含まれているため，耳よりキーワードを除く名詞を対象としてクラスタリングを行う．クラスタリングの手法は種々あるが，単文にある程度適していると考えられる [14]Repeated Bisection[15] を用いてクラスタリングを行う．

3.3　Repeated Bisection

本研究では ReipeatedBisection を用いてクラスタリングを行うが，Repeated Bisection について簡単に述べる．Repeated Bisection はクラスタリングツール bayon[16] や CLUTO[17] で使用されているクラスタリング手法であり，K-means 法を k = 2 で n - 1 回繰り返して n 個のクラスタを得る．すべてのデータを 1 つのクラスタに格納し，以下の手順を繰り返し，クラスタを 2 分割していき，クラスタリングを行う．

(1) 全クラスタ中から最もまとまりの悪いクラスタを 1 つ選択する．
(2) クラスタの中からランダムに 2 つの要素を選択しそれぞれを格納したクラスタを作成する．
(3) 元のクラスタ内の全ての要素に対し，ランダムに選択した要素との類似度を比較する．
(4) 類似度を比較した結果，より類似殿高いクラスタに要素を格納する．
(5) クラスタ間で要素の移動を行い，クラスタ内で類似度

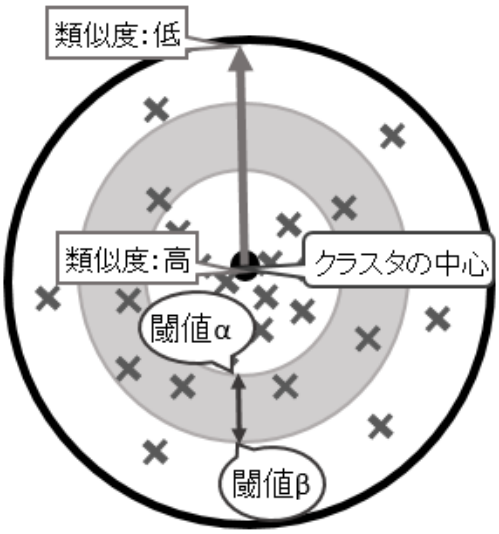

図 1 1 クラスタ内の概要

をそれぞれ比較し直す．

(6) (5) を移動できる要素がなくなるまで繰り返し行う．

3.4 意外な情報の抽出

Repeated Bisection を用いてクラスタリングを行った結果の各クラスタは中心ベクトルがそのクラスタを代表するトピックであるため，このトピックの中心に近い程多くの人々がレビューしている公知な情報であると考えられる．また，トピックの中心から遠い程人々がレビューしていない情報やトピックとは関係のない情報になる場合が多い．そこで，我々は人々があまりレビューを書いていないが，トピックと関係のある内容の情報は，そのトピック内では意外な情報であると考え，中心からある程度離れているが離れすぎていない範囲にある情報を意外な情報とする．つまりは，**図 1** に示すようにクラスタの中心からある程度離れた距離，閾値 α と β の間にある文が意外な情報であるとする．このように求めた文は有用な情報かつ意外な情報であるので耳より情報とする．

伏見稲荷大社の休憩というトピックのクラスタを例にすると，閾値 α 以上なら「休憩できます」といった多くの人がレビューしている内容が集まっている公知な情報であり，閾値 β 以下なら「途中で休憩しながらでも結構な距離なので半端な気持ちでは辛いです」といった個人の感想や「結構歩くので疲れます」といったトピックとあまり関係ない情報であり，閾値 $\alpha \sim \beta$ の幅内にあるレビューは「帰ろうとしたとき，休憩所があり，無料でお茶が飲めました」といった公知ではなくトピックとも関係のある情報となり，耳より情報となる．

4. 実験

3.1 節で提案した耳よりキーワード有用性を測り，3.4 節で意外な情報を抽出するための閾値 α，β を求めるために実験を行った．また，訪れたことのある観光地かそうでないかで耳より情報に違いがあるのかについても同時に調べた．それぞれについて以下に結果と考察を述べる．

表 2 有用な情報抽出結果

観光地	有用な情報	適合率	再現率
伏見	249	0.24	0.61
水族館	109	0.06	0.64
水木しげる	69	0.41	0.68
嵐山	80	0.23	0.68
中華街	227	0.2	0.62

4.1 実験条件

以下の条件で実験を行った．

データセット：

伏見稲荷大社：2337 文
京都水族館：818 文
水木しげるロード：741 文
嵐山：636 文
横浜中華街：2202 文

被験者： 20 代 6 名

実験方法：

各々の観光地に対して興味有り訪問経験有り，興味有り訪問経験無しとなる各 3 名の被験者 6 名が観光地のレビューの文を読み，耳より情報であると感じた文とそうでない文と 2 択で判断した．6 名の内，興味有り訪問経験有り，興味有り訪問経験無しで各 2 名以上となる文を耳より情報の正解データとする．

正解データ：

伏見稲荷大社：96 文
京都水族館：11 文
水木しげるロード：28 文
嵐山：28 文
横浜中華街：74 文

4.2 耳よりキーワードの有用性

3.1 節で提案した耳よりキーワードを用いて実験データから有用な情報を抽出した際の適合率と再現率を**表 2** に示す．適合率は全体的に悪かった．これは耳よりキーワードを決定する際，耳より情報の要因となると考えられるキーワードだが耳より情報よりもそうでない文に多く含まれているキーワードは省いてしまったため適合率が落ちてしまったと考えられる．しかしながら，レビューの総数から有用な情報として 1/10 前後の文数に絞り，いずれの観光地も再現率は 0.6 以上なので耳よりキーワードは改善の余地はあるが有用であると考えられる．

4.3 閾値 α，β の決定

3.4 節で提案した意外な情報抽出においての中心ベクト

ルと文の類似度の閾値α，βを決定するために，耳よりキーワードを用いて抽出した有用な情報を対象にクラスタリングを行い，閾値を0〜1間0.05単位で適合率と再現率を求めた．話題毎に分ける際のクラスタリングにはRepeated Bisection法を用いているbayonを用いた．適合率の結果を**表 3**，再現率の結果を**表 4**に示す．表3より類似度が0もしくは1付近のときにも耳より情報が含まれていることが分かる．類似度0付近は紅葉の話題で「レンタサイクル屋で地図がもらえます」といったクラスタリングの結果分けられたトピックとまったく関係ないトピックについての内容だが耳より情報というパターンであった．これは中心ベクトルは1つのトピックに偏っているものが望ましいが，有用な情報の中に類似するトピックがなかったために複数のトピックを元に中心ベクトルを構成してしまい，クラスタ内での類似度が下がってしまう結果となってしまった．このことより，複数のトピックをもつクラスタについては本手法では対応できないことが分かった．類似度1付近は1クラスタ内の文が少なく，すべての文が1になってしまい類似度が意味をなさなくなってしまっていた．これは，クラスタリングを行う際に必要以上のクラスタに分かれてしまったためと考えられるので有用な情報の文数によって最適なクラスタ数を推定する必要があることが分かった．一方で，類似度[0.55,0.6)の場合の適合率は各々の観光地で最も高いということはないが，嵐山以外の4つの観光地において有用な情報の適合率と比べ増加している．表4より耳より情報の類似度に大きく集中した範囲がなく，分散している為全体的にかなり低い結果となった．これらを踏まえて，現段階では意外な情報を抽出する際の閾値α，βをそれぞれ0.55，0.6とする．

4.4 提案手法の有用性

対象の観光地の訪問経験の有無により提案手法の有用性が変わるのかを図った．具体的には，正解データを観光地の訪問経験の有る人，ない人それぞれに作成し，提案手法による結果の適合率と再現率を示す．訪問経験の有る人3名のうち2名が耳より情報といった文を訪問経験の有る人の正解データとし，同様に訪問経験のない人3名のうち2名が耳より情報といった文を訪問経験のない人の正解データとした．**表 5**に結果を示す．再現率は全体的に低い結果となったが，訪問経験のある場合はない場合より「伏見稲荷大社」,「水木しげるロード」,「嵐山」,「横浜中華街」の4つの観光地で適合率が高くなった．これは，訪問経験のない人よりある人の方が耳より情報と判断した文が多かったためと考えられる．適合率の低かった「京都水族館」については，元のレビュー自体に耳より情報が少なく不要な情報が多く含まれているからと考えられる．また，実際に耳より情報と判断した文をみると，訪問経験がある場合にはその観光地の歴史や設備の情報，観光にかかる時間の目安となる情報，レビューを書いた人がどんなコースを回ったかといった情報を耳より情報と判断する傾向があった．一方訪問経験のない場合には交通手段や渋滞の情報や駅からの距離の情報といった観光地へ行くまでの情報を耳より情

表 3 閾値 0〜1 の適合率

類似度	伏見	水族館	水木しげる	嵐山	中華街
1	0.25	0.5	0	0	0
[0.95,1)	0.50	0	0	0	0.13
[0.9,0.95)	0.29	0	0	0.67	0.18
[0.85,0.9)	0.11	0	0	0.14	0.44
[0.8,0.85)	0.20	0	0.75	0.2	0.1
[0.75,0.8)	0.27	0.09	0.5	0.5	0
[0.7,0.75)	0.08	0.09	0.43	0.33	0.13
[0.65,0.7)	0.23	0	0.5	0.33	0.15
[0.6,0.65)	0.28	0	0.38	0	0.21
[0.55,0.6)	0.29	0.11	0.62	0.22	0.31
[0.5,0.55)	0.20	0	0.83	0.11	0.33
[0.45,0.5)	0.50	0	0	0	0
[0.4,0.45)	0	0.17	0	0.33	0.18
[0.35,0.4)	0	0	0	0	0.17
[0.3,0.35)	0	0	0	0	0.33
[0.25,0.3)	0	0	0	0	0
[0.2,0.25)	0	0	0	0	0
[0.15,0.2)	0	0	0	0	0
[0.1,0.15)	0	0	0	0	0
[0.05,0.1)	0	0	0	0.67	0
(0,0.05)	0	0	0	0	0
0	0.27	0	0.4	0.22	0.08

表 4 閾値 0〜1 の再現率

類似度	伏見	水族館	水木しげる	嵐山	中華街
1	0.01	0.091	0	0	0
[0.95,1)	0.042	0	0	0	0.014
[0.9,0.95)	0.042	0	0	0.071	0.027
[0.85,0.9)	0.021	0	0	0.036	0.054
[0.8,0.85)	0.031	0	0.073	0.036	0.014
[0.75,0.8)	0.031	0.091	0.049	0.071	0
[0.7,0.75)	0.021	0.091	0.073	0.107	0.041
[0.65,0.7)	0.052	0	0.049	0.071	0.054
[0.6,0.65)	0.031	0	0.073	0	0.041
[0.55,0.6)	0.083	0.182	0.195	0.071	0.149
[0.5,0.55)	0.094	0	0.122	0.071	0.149
[0.45,0.5)	0.021	0	0	0	0
[0.4,0.45)	0.01	0.091	0	0.071	0.027
[0.35,0.4)	0.021	0	0	0	0.014
[0.3,0.35)	0	0	0	0	0.014
[0.25,0.3)	0.01	0	0	0	0
[0.2,0.25)	0	0	0	0	0
[0.15,0.2)	0	0	0	0	0
[0.1,0.15)	0	0.091	0	0	0
[0.05,0.1)	0	0	0	0.071	0
(0,0.05)	0	0	0	0	0
0	0.094	0	0.049	0.071	0.027

表 5　提案手法の結果

観光地	訪問	適合率	再現率
伏見	無し	0.128	0.063
	有り	0.426	0.208
水族館	無し	0.190	0.364
	有り	0.095	0.182
水木しげる	無し	0.462	0.146
	有り	0.615	0.195
嵐山	無し	0.167	0.071
	有り	0.556	0.179
中華街	無し	0.250	0.122
	有り	0.556	0.270

報と判断する傾向があった．これは，実際に訪問したことがあれば観光地までのアクセスに関しては特に必要ないからであると考えられる．逆に訪問経験がなければ，迷わず着けるのか，車と電車ではどちらがいいのか等が判断しづらく，このようなアクセスに関する情報を求めているからだと考えられる．これらのことから訪問経験の有無が耳より情報かどうかの判断に影響を与えていることが分かった．

5. まとめと今後の課題

本研究では，観光地のレビューサイトから耳より情報を含む文を抽出する手法を提案した．具体的には，まずレビューサイトからレビューを取得し，文単位に分割する．そしてそれらの文から耳より情報のキーワードを含む文を抽出し，クラスタリングを行った．次に，クラスタの中心ベクトルとの類似度が閾値 α と β の間にある文すべてを耳より情報とした．我々の提案する手法を用いることで，ユーザは効率的に新たな知識を得て，旅行の計画を立てる際の参考になると思われる．

今後の課題として，意外な情報抽出の際のクラスタリングにおいて極端に文数の少ないクラスタが存在し，これが原因で耳より情報であっても類似度が 0 や 1 になってしまうため，クラスタ内の文数を考慮して最適なクラスタ数を推定する必要がある．また，中心ベクトルが複数のトピックで構成されている場合，そのクラスタ内の文の話題が様々なものになってしまうので複数のトピックで構成されたクラスタについて類似度ではない別の手法を考えなくてはならない．4.3 節では訪問経験の有無が耳より情報かどうかの判断に影響を与えることが分かったので，訪問経験の有無によって優先すべき情報とそうでない情報があるため，耳より情報にも優先度があり，それらを考慮する必要がある．耳よりキーワードを用いて抽出した有用な情報の中には，「おすすめです」といった内容のない情報も含まれてしまっているため，耳よりキーワードだけでなく経験情報であるかを判定する必要がある．本論文では 5 つの観光地のレビューから作成した耳より情報の正解データを作成し，それを元に耳よりキーワードと閾値 α，β を決定したため他の観光地を対象にした場合に同じ結果が得られるかわからないので，それらも確認していきたい．そして，本研究では有用な情報かつ人々にあまり知られていない意外な情報やユーザが興味のある情報，未知の情報が耳より情報と考えているので，ユーザの興味や知らないであろう未知の情報を考慮した耳より情報抽出手法を進めていきたい．

謝辞　本研究の一部は JSPS 研究費 26330347 及び，私学助成金（大学間連携研究補助金）の助成によるものです．ここに記して謝意を表します．

参考文献

[1] フォートラベル http://4travel.jp/
[2] トリップアドバイザー http://www.tripadvisor.jp/
[3] Yuki Hattori and Akiyo Nadamoto “Tip Information from Social Media based on Topic Detection” International Journal of Web Information Systems, Vol. 9, No. 1, pp. 83-94, 2013.
[4] 河中 照平, 井上 潮, “閲覧者にとって有用性の高い Web ユーザレビューランク付け手法の提案”, DEIM Forum 2014 B5-5, 2014.
[5] 小林 大祐, 井上 潮,“Web 上のレビュー情報からユーザが重要視する製品の特徴うを抽出する手法の提案”, DEIM Forum 2009 C6-4, 2009.
[6] 佐々木 優衣, 関 洋平, “商品レビューを対象とした有用性の定義と判別”, DEIM Forum B5-1, 2014.
[7] 石川 瑛子, 中山 伸一, 真栄城 哲也, “旅行のクチコミサイトにおける利用者の必要性に応じた情報表示”, 第 72 回全国大会講演論文集, pp.865-866, 2010.
[8] 遠藤 雅樹, 横山 昌平, 大野 成義, 石川 博, “特定地域に限定しない観光キーワードの自動抽出”, DEIM Forum 2014 E9-2, 2014.
[9] 藤井 一輝, 石野 亜耶, 藤原 泰士, 前田 剛, 難波 英嗣, 竹澤 寿幸, “多言語旅行ブログエントリを用いた観光情報提示システム”, DEIM Forum2014 P4-1, 2014.
[10] 中嶋 勇人, 太田 学, “旅行ブログからの名所とその付随情報の抽出”, DEIM Forum 2013 B8-4, 2013.
[11] 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”, 第 75 回全国大会講演論文集, pp.307-308, 2013.
[12] 安藤 まや, 石崎 俊, “インパクトの視点に基づく WEB 上のユーザレビューの分析”, 言語処理学会第 18 回年次大会, pp.731-734, 2012.
[13] 京都大学 大学院情報学研究科 知能情報学専攻知能メディア講座言語メディア分野:日本語構文解析システム JUMAN : http://nlp.ist.i.kyotou.ac.jp/index.php?JUMA
[14] 花井俊介, 灘本明代, “酷似レシピ抽出のためのクラスタリング手法の提案”, DEIM Forum 2014 F8-6, 2014.
[15] Ying Zhao and George Karypis. Comparison of agglomerative and partitional document clustering algorithms. Technical report, Department of Computer Science,University of Minnesota, Minneapolis, MN 55455, 2002.
[16] Bayon - a simple and fast clustering tool - Google Project Hosting http://code.google.com/p/Bayon/
[17] CLUTO - Software for Clustering High-Dimensional Datasets http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview