観光地のレビューからの耳より情報抽出手法

阪井 奎伍† 灘本 明代††

† 甲南大学大学院自然科学研究科知能情報学専攻 †† 甲南大学知能情報学部

a) *m1524003@s.konan-u.ac.jp* b) *nadamoto@konan-u.ac.jp*

概要 今日トリップアドバイザーやフォートラベルのような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には実際に観光地に行ったことのある人がその経験に基づくレビューを書いており，観光地の公式サイトにはない様々なお得な情報が多数記載されている．しかしながら，有用な観光地ほどレビューの量は膨大であり，その中から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」と感じる情報を耳より情報と呼び，この耳より情報を抽出する手法を提案する．我々の提案する耳より情報は，ユーザにとって有用であり，そして人々があまり知らない，ある程度レアな情報であると考え，これらを考慮して観光地のレビューから耳より情報の抽出手法を提案する．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有用な情報とし抽出する．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ルを構成する文との類似度からある程度レアな情報を抽出し，そのある程度レアな情報を耳より情報として，ユーザに提示する．

キーワード 観光，レビュー，情報抽出

1 はじめに

現在インターネット上には様々な観光に関する情報が多数存在する．人々は旅行に行く前に，観光地に関する情報をインターネットから取得し，計画を立てる事を行う場合が多数ある．この時，初めて訪れる観光地ではわからないことも多く，行ってから「〜しておけば」と後悔したという事態に陥りやすい．そこで，観光地の公式サイトにあるような基本的な情報だけでなく，経験に基づくお得と思われる情報を取得することは，観光の計画を立てる際に有用であると考えられる．しかしながら，検索エンジンを用いてこのような情報を集めようとしても公式サイトにあるような基本的な情報が多く，経験に基づくお得と思われる情報を上手く取得できない．また，Q & A サイトでは質問に対する回答の為，得られる情報がピンポイントになってしまうことやユーザが思いついた内容以外の有用な情報を得ることが困難であること，検索結果が多いため全ての質問を閲覧するには時間がかかってしまうことがある．

一方，トリップアドバイザー[1] やフォートラベル[2] 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．これらのレビューサイトは実際に観光地に行ったことのある人が，その経験に基づくレビューを書いており，観光地の公式サイトにある情報ではない，経験に基づく様々なお得と思われる情報が多数記載されている．そして，これらお得な情報の中にはあまり知られていない有用な情報が多数存在している．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，その中から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「知って得をした」,「参考になった」と感じられる情報を「耳より情報」と呼ぶ．我々の提案する耳より情報は，ユーザにとって有用であり，そして人々があまり知らないある程度レアな情報であると考え，これらを考慮して観光地のレビューから耳より情報の抽出手法を提案する．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有用な情報とし抽出する．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ルを構成する文との類似度からある程度レアな情報を抽出し，それを耳より情報とする．

2 関連研究

レビューに関する研究は種々ある．服部ら [1] 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．耳より情報を含む書き込み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，タイプ別のキーワードを提案している．本研究ではこの 4 つのタイプと各々のタイプにおけるキーワードを参考にし，観光地のレビューを対象としたタイプと耳よりキーワードを提案する．河中ら [2] は，ユーザにとって有用性の高いレビューを推薦する手法を提案している．しかし商品に対するレビューの評価項目に着目している点で本研究とは異なる．小林ら [3] はレビューには意見が生まれた理由が製品と特長を表していると考え,「条件」,「理由」,「態度」,「対象」の要素を定義し，レビューからこれらの要素を抽出する手法を提案している．小林ら [3] の定義した 4 つの要素の

[1] トリップアドバイザー http://www.tripadvisor.jp/

[2] フォートラベル http://4travel.jp/

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する点がある．しかし本研究では耳よりキーワードで抽出した有用な情報からさらに類似度を用いる点で異なる．

旅行のレビューに関する研究も種々ある．中嶋ら [4] は，旅行ブログからまず名所を抽出し，さらに書き手の体験情報，評価表現，書き手がその場を訪れることで得られる状態情報，名所の由来や歴史などの記述箇所の 4 つを名所付随情報として抽出する手法を提案している．体験情報は耳より情報に類似しているが，ブログ記事から抽出している点で異なる．藤井ら [5] 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．本研究ではレビューを対象としている点，キーワードを分類している点で異なる．松本ら [6] は観光地のクチコミから特徴語を抽出し観光地検索を支援するシステムを提案している．本研究では目的の観光地は決まっている点で異なる．安藤ら [7] 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．本研究では旅行のレビューにおいて各々の文との類似度を用いて耳より情報を抽出している点で異なる．

3　耳より情報抽出手法

耳より情報を抽出する手順を以下に示す．

(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レビューを検索する．

(2) システムは，検索結果のレビューを取得しこれらレビューを文単位で分割する．

(3) (2) の中で我々の提案する耳よりキーワードが含まれている文すべてを有用な情報とする．

(4) (3) で抽出された有用な情報において，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) 各クラスタ内でクラスタの中心ベクトルとそ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にある文をみんながあまり知らないある程度レアな情報とし，これを耳より情報とする．

ここで，一つのレビューの中には個人の感想や経験談といった様々な情報と耳より情報の文が混在しているため，これらを分割する必要がある．実際にレビューを見てみると文単位でこれらの文が分割可能である事が分かる．そこで，本研究では各レビューを文単位に分割し，耳より情報をこの文単位に抽出することを行う．文の分割は，文中に「。」,「.」,「！」,「!」,「？」,「?」,「♪」または改行が一つ以上の場合にそこが文の終わりと考え分割する．

3.1　有用な情報の抽出

服部ら [1] 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．耳より情報を含む書き込み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，タイプ別のキーワードを提案している．本研究では，この 4 つのタイプと各々のタイプ別のキーワードを参考にし，旅行のレビューを対象としたタイプと各々のキーワードを人手により決定した．タイプは「お得耳より」,「提案耳より」,「時間情報」,「天気情報」の 4 つであり，表 1 にそれぞれのタイプとその耳よりキーワードを示す．

表 1 耳よりキーワード

タイプ	耳よりキーワード例
お得耳より	スムーズ，便利，必須，穴場，役立つ，がらがら，十分，定番，見ごろ，最善，対策，緊急，画期的，人気，好都合，有名，損，警告，注意，混雑，残念
提案耳より	おすすめ，良かった，方がいい，良いと思，するべき，できる，できない，それより，なので，だから，行くべき，すぐに，実際は
時間情報	朝，昼，夕方，夜
天気情報	晴れ，曇り，雨 ，寒い，暑い

東京スカイツリーのレビューを例として，それぞれのタイプの説明をする．お得耳よりは「展望台に登る時は、双眼鏡を持って行くと便利」という文の“便利”といった特に意識していなくても文をぱっと見たときに「お得だ」と感じるような単語だけでなく,「注意書きにもあったのですが、４５０展望台ではトイレが狭くとても並ぶ」という文の“注意”といったネガティブな単語でも失敗や後悔の回避に繋がる単語も対象とする．提案耳よりは「当日予約をすると相当並ぶのでクレジットカードを片手にネットで事前予約することをお勧めします」という文の“お勧め”といった東京スカイツリーに行った体験・経験をしていない相手に対して自身の体験・経験を基に根拠を持って提案するような言葉である．また，この提案耳よりには自身の体験・経験を基にした意見な

のでポジティブな意見だけでなく,「土曜日の 18 時半ぐらいに着いたんですがちょうど 18 時 30 分〜19 時の整理券を配っていて中は行列が…これなら平日の昼間に来た方がいいのかなと」のようにネガティブな意見であっても有用だと考えられる．時間情報は時期的なものを表す単語，天気情報は天気や気候を表す単語である．これら時間情報や天気情報は時間や季節，天気によって耳より情報が変わると考え，旅行レビューサイトからの耳より情報の抽出に必要である．本論文では耳より情報を抽出するはじめの一歩として，提案耳よりとお得耳よりについての抽出を行う．そこで文に提案耳よりとお得耳よりの耳よりキーワードが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 ある程度レアな情報の抽出

公知でないレアな情報はユーザが得をしたと感じる耳より情報ではないかと考え，このレアな情報を抽出することを行う．この時，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．そこで本論文では，他の耳より情報の候補と類似する話題について述べながら，他の候補と少し異なる事を述べているレビューをある程度レアな情報とする．さらに，ある程度レアな情報とは話題ごとにそれぞれ異なると考え話題ごとに分類し，ある程度レアな情報を抽出する．つまりは，話題ごと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 節にて抽出した耳より情報の候補となる有用な情報のすべての文を juman[8] により形態素解析し名詞を抽出する．ここで単に名詞のみを抽出すると「期間限定」などの 2 つの名詞が連続して 1 つの単語になっているものが「期間」と「限定」に分かれてしまうため，名詞が連続している場合に限り名詞連結を行い 1 つの名詞として扱う．そして，耳よりキーワード以外の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 [9]Repeated Bisection[10] を用いてクラスタリングを行う．

3.3 Repeated Bisection

本研究では Reipeated Bisection 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[11] や CLUTO[12] 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，より類似度の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をそれぞれ比較し直す．

(6) (5) を移動できる要素が無くなるまで繰り返し行う．

3.4 ある程度レアな情報の決定

Repeated Bisection を用いてクラスタリングし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，そのトピックに近いほど種々の人が発言している公知の情報であると言える．逆に中心から離れた文はそのクラスタ内であまり発言されていない情報であり，そのクラスタのトピックと関係ない情報である場合が多い．そこで，我々はある程度レアな情報は，トピックと関係があるが，あまり話題になっていない情報をもつと考え，中心からある程度距離が離れているが，そんなに離れていない文をある程度レアな情報であるとする．つまりは，図 1 に示すように，中心からある程度離れた距離，α と β の間にある文がある程度レアな情報であるとする．このように求めたある程度レアな情報を耳より情報とする．

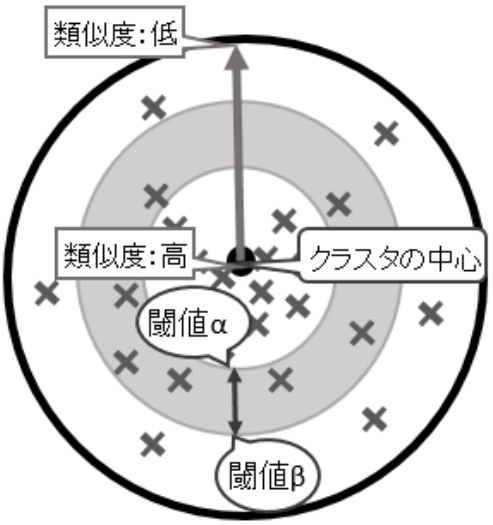

図 1 1 クラスタ内の概要

例えば，東京スカイツリーの景色という話題の中では，閾値 α 以上なら「景色がよかった」などその話題の中で多くの人がレビューしている情報が集まっている公知な情報であり，閾値 β 以下ならほとんどレビューされて

表 3 閾値 0.5〜0.8 の適合率

閾値	ディズニー	スカイツリー	金閣寺	首里城公園	札幌市時計台
0.50	0.417	0.295	0.413	0.457	0.348
0.55	0.429	0.326	0.488	0.432	0.467
0.60	0.412	0.488	0.395	0.442	0.341
0.65	0.412	0.425	0.341	0.405	0.395
0.70	0.355	0.406	0.268	0.286	0.317
0.75	0.250	0.400	0.256	0.300	0.231
0.80	0.143	0.308	0.314	0.250	0.188

表 2 クラスタ数 30〜70 の適合率

クラスタ数	30	40	50	60	70
データ数	187	217	243	261	268
ディズニー	0.422	0.47	0.465	0.418	0.392
データ数	128	139	152	145	136
スカイツリー	0.414	0.412	0.418	0.322	0.408
データ数	192	204	209	217	226
金閣寺	0.267	0.32	0.35	0.33	0.271
データ数	307	354	371	475	395
首里城公園	0.417	0.431	0.45	0.406	0.367
データ数	90	97	100	91	107
札幌市時計台	0.305	0.309	0.368	0.366	0.338

いない情報といったあまりにもレアな情報や，もしくは「天気が悪い」など話題と関係ない情報であり，閾値 α 〜β の幅内にあるレビューは「○○の場所からならよく見えた」など公知でない話題と関係があるようなある程度レアな情報となり，耳より情報とする．

4 パラメータの決定の実験

3.1 節より抽出した有用な情報を話題毎に分けるためクラスタリングを行う際の最適なクラスタ数，3.2 節である程度レア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．クラスタリングには Repeated bisection 法を用いている bayon を用いた．さらに，提案手法の有用性を示すユーザ実験を行った．

4.1 クラスタ数の決定

ある程度レアな情報とは話題毎にそれぞれ異なっていると考えられるため本研究では 1 つの観光地に対して抽出された有用な情報からある程度レアな情報を抽出するために適した話題の数を求めるために実験を行った．データはフォートラベルから任意の 5 つの単語各々をクエリとして得られたレビューデータを用いる．レビューデータは東京ディズニーランド 3862 文，東京スカイツリー 6025 文，金閣寺 2075 文，首里城公園 3196 文，札幌市時計台 2216 文を用いた．実験はクラスタ数を 30〜70 間 10 単位で変化させた．その結果を人手で耳より情報か耳より情報ではないかを判定し適合率を求めた．結果を表 2 に示す．表 2 よりクラスタ数が 50 のときに 5 つの観光地のうち 4 つにおいて適合率が最も高くなったのでクラスタ数を 50 とする．

4.2 閾値 α，β の決定

耳より情報を求める類似度の閾値の幅を決定する値 α と β を求めるために実験を行った．実験は 4.1 で用いたデータ，クラスタリング手法も 4.1 と同様である．そして閾値を 0.5〜0.8 間 0.05 単位で変化させた．その結果を人手で耳より情報か耳より情報でないかを判定し適合率を求めた．結果を表 3 に示す．表 3 より閾値 0.55 のとき 5 つの観光地のうち 3 つにおいて適合率が最も高くなったので閾値 0.55 を中心とするために α を 0.5 とし，β を 0.6 とする．

5 評価実験

3 章で提案した観光地のレビューサイトから耳より情報の抽出手法について，その有用性を測るために，ユーザ評価実験を行った．

5.1 実験条件

被験者は，20 代，男女 8 名の被験者とし，実験データは提案手法により抽出された東京ディズニーランド 360 文，東京スカイツリー 242 文，金閣寺 99 文，首里城公園 209 文，札幌市時計台 151 文を用いた．はじめに，被験者に 5 つの観光地それぞれに行く予定を立てるために情報収集をするという状況を想定した上で実験を行った．実験方法は，提案手法により抽出された文が「参考になった」「知って得をした」と感じる耳より情報であるかそうではなかったかを 2 択で判断し，適合率を求めた．本研究では時間情報を考慮していないため「夏は相当暑いので日焼け対策を」といった季節限定など，ある期間に行くなら耳より情報となるが期間外では耳より情報とならない文に関してはその期間に行くと仮定し，耳

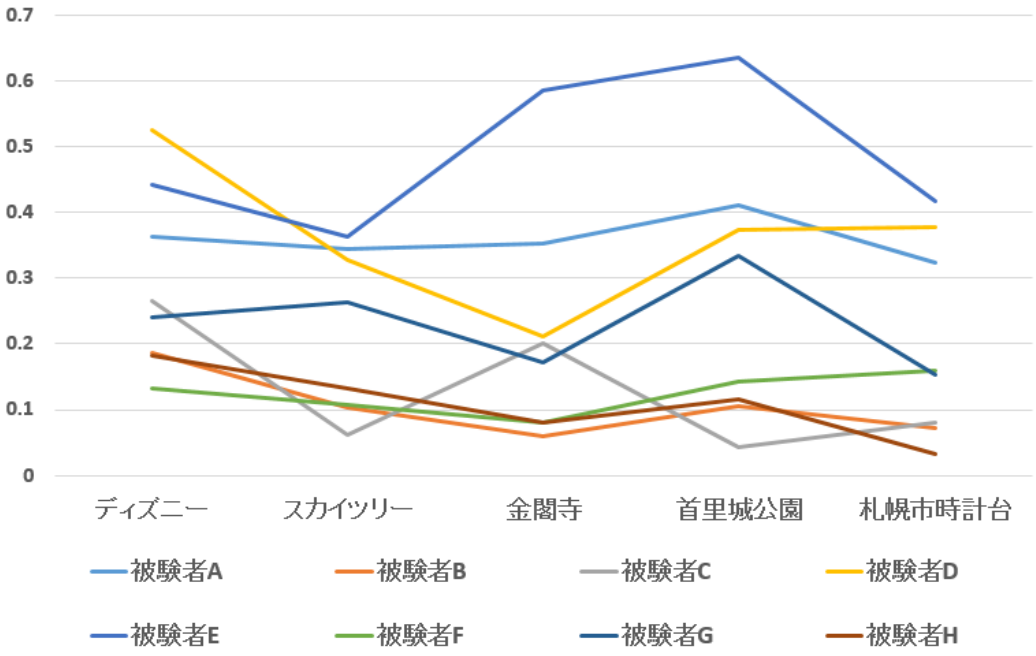

図 2 ユーザ別適合率

より情報であるかそうでないかの判断を行った．また，クラスタリングによって話題毎に分けることができているかを評価するためにそれぞれのクラスタの話題とクラスタ内の文の内容が一致しているかしていないかについても 2 択で判断した．

5.2 耳より情報抽出の結果と考察

観光地毎の平均適合率を表 4，全体の結果を図 2 に示す．表 4 から全体的に適合率が 30%以下と低い結果となったことが分かる．しかし，図 2 のように被験者を個別に見てみると被験者 A は全ての観光地において 30%を超え，被験者 E は 60%～40%と高い結果となる場合もある．このことから被験者によって適合率に大きな差が出ると考えられる．

表 4 耳より情報平均適合率

観光地	適合率
東京ディズニーランド	0.292%
東京スカイツリー	0.213%
金閣寺	0.218%
首里城公園	0.270%
札幌市時計台	0.202%
平均	0.239%

原因としては，被験者が旅行自体に興味があるかどうかや各々の観光地について興味があるかどうかを考慮してない点や被験者によって耳より情報と感じる文が異なっていた．この例として，観光地が東京スカイツリーの場合に耳より情報として抽出された「都心の交通量の減る休日や空気が澄む冬の晴れ，工場が止まるお正月なら綺麗に見える確率が高いと思いました」という文がある．この文の場合，被験者 8 名のうち 4 名が耳より情報と判断したが残りの 4 名は耳より情報ではないと判断する結果となった．また，1 つの話題の中で耳より情報と判断された文が 1 つもない場合があった．これは，「下」や「とき」，「場合」など話題と言えない単語が話題になってしまったことが原因として考えられる．これらの単語をストップワードとする必要がある．

次に提案手法により抽出された耳より情報を観光地毎に良い例，悪い例を挙げる．良い例としては，東京スカイツリーで「無料で写真をとってくれるサービスもあるので，カメラをお忘れなく」という耳より情報が得られた．この耳より情報は被験者 8 名のうち 7 名が耳より情報と感じた文である．この文は，「無料」と「サービス」という耳よりキーワードにより有用な情報として抽出されたと考えられる．一方，悪い例では，東京スカイツリーで「夏休みの混雑前の天気の良い日をセレクトしたつもりですが，ピーカンとまではいかず」というレビューを書いた人の体験経験だけで有用な情報がないものや「ちょうどライトアップされる瞬間を撮影できました」といった感想だけの情報が抽出されてしまった．これらは耳より情報ではないレビューでも「混雑」や「撮影」といった耳よりキーワードが含まれていることが原因の一つと考えられる．

5.3 話題と文の一致の結果と考察

観光地毎の平均適合率を表5に示す．表5より5つの観光地のうち3つにおいて適合率が0.9を超える結果となったのでクラスタリングによりほぼ正確に話題毎に分けることができたとわかった．話題と一致してない文ではクラスタ内の類似度がより低い文に多かった．公知な情報と話題と関係ない情報を省くことによって，ある程度レアな情報の抽出を行うために閾値を0.5〜0.6の幅としたがこの幅では関係ない情報が含まれてしまうことが分かった．例としては，話題が「入場料」で文が「富士山も綺麗に見ることが出来ました」という全く関係ないと分かるものが多い．しかし，話題が「入場券」で文が「インターネットから予約でき日時指定もできる」という文には入場券が含まれていないが文の内容から入場券の話題と分かる文もある．この文は被験者8名のうち2名しか話題と一致していないと判断しているので文の内容よりは文に話題の単語が含まれている方が話題と文は一致していると考えられる．

表5 話題一致平均適合率

観光地	適合率
東京ディズニーランド	0.855%
東京スカイツリー	0.841%
金閣寺	0.927%
首里城公園	0.928%
札幌市時計台	0.914%
平均	0.893%

6 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提示した．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，次の行動を起こす際の参考になると思われる．

今後の課題として評価実験によりユーザの興味に合わせてパーソナライズする必要があることがわかった.「おすすめ」といった耳よりキーワードそのものが有用な情報として抽出されて文もあったため極端に短い文は有用な情報から省くことも考えている．また，レビューなのですべての文が体験・経験を基に書かれていると考えていたが，実際は1人のユーザのレビュー全体で見ると体験・経験の内容を書いていても文単位に分割してしまうことで経験情報が含まれてない文が出てきてしまったため，経験マイニングをする必要があることが分かった．ある程度レアな情報の抽出では耳より情報が全く無い話題や話題そのものが何なのか理解しにくい話題があったためこれらも改善する必要がある．また，ある程度レアな情報の抽出には類似度のみを用いているが，単語の出現頻度なども視野に入れて適合率の向上を図りたい．有名な観光地とそうでない観光地にではレビューの件数に大きな差があるため抽出される有用な情報の量も変化してしまう．評価実験では予備実験と同様の観光地を対象としたが今後はユーザが興味のある観光地を対象としていく考えなので，ユーザが指定した観光地の有用な情報の量に応じて最適なクラスタ数を算出できるようにすることや類似度での閾値 α，β，これらの適切な値の検討も進めていきたい．

参考文献

[1] Hattori,Y. and Nadamoto,A.: Tip Information from Social Media based on Topic Detection, International Journal of Web Information Systems, Vol. 9, No. 1, pp.83-94, 2013.

[2] 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

[3] 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 Forun 2009 C6-4，2009

[4] 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

[5] 藤田一輝，石井亜耶，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

[6] 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

[7] 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

[8] 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.kyoto-u.ac.jp/index.php?JUMAN

[9] 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

[10] 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.

[11] Bayon - a simple and fast clustering tool - Google Project Hosting http://code.google.com/p/Bayon/

[12] CLUTO - Software for Clustering High-Dimensional Datasets http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview