

観光を対象とした口コミからの耳より情報抽出手法の提案

阪井 奎伍[†] 灘本 明代[†]

甲南大学知能情報学部[†]

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、行ってから後悔したという事態に陥りやすい。そこで、観光地の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、観光の計画をす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い、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。

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、観光地の口コミサイトが普及している。これらの口コミサイトは実際に観光地に行ったことのある人が、その経験に基づく口コミを書いており、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、有名な観光地ほど口コミの量は膨大であり、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「得をした」「参考になった」と感じられる情報を「耳より情報」と呼び、インターネット上の観光地の口コミサイトからこれら耳より情報を抽出する手法の提案を行う。

2. 関連研究

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、旅行の口コミ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[4]は、ユーザにとって有用性の高い口コミを推薦する手法を提案している。しかし商品に対する口コミの評価項目に着目している点で本研究とは異なる。中嶋ら[5]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。

3. 提案手法

以下の手順で耳より情報を抽出する。

- (1) 口コミ情報を文単位で分割
- (2) (1)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耳より情報の候補とする。
- (3) (2)で抽出された耳より情報の候補を形態素解析する。
- (4) (3)の結果から、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耳より情報とする。

3. 1 口コミ情報の分割

同じユーザの口コミの内容には感想等様々な情報と耳より情報の文が混在しているので、本研究では各口コミを文単位に分割する。文章中に「。」「.」「!」「!」「?」「?」「♪」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. 2 旅行を対象とした耳より情報の分類

本研究では、旅行を対象としているため服部ら[3]の耳より情報の分類とそれぞれのキーワードを参考にし、旅行の耳より情報を抽出するための、分類及びそのキーワードを人手によって決定した。表1に分類と各々の耳よりキーワードの一部を示す。

Extracting of Tip information from review of sightseeing
Keigo Sakai[†]

Akiyo Nadamoto[†]

Dept. of Intelligence and Informatics Faculty of Intelligence and Informatics[†]

表1 耳より情報の分類

分類	耳よりキーワード例
提案耳より	おすすめ, 方がいい, 良いと思 い, いいかと, いいのでは, 良 いでしょう, 良くない, すべ き, するべき, いかが, 間違 い, よさげ, どうでしょう
お得耳より	無料, タダ, 安い, 売り切れ, 必須, 必見, ポイント, スム ーズ, 混雑, 役立つ, 見所, 渋 滞, 禁止, 裏道, 近道, 特典, 注意, 警告, ルール, 手軽, セ オリー, コツ, 開始, 解消,
時間情報	春, 夏, 秋, 冬, 朝, 昼, 夕
天気	晴れ, 曇り, 雨

3. 3 耳より情報の抽出

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口コミを耳より情報と定義する. 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, 手順2にて抽出した耳より情報の候補となるすべての文を形態素解析し名詞を抽出する. 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する. ここでクラスタリング手法として, Repeated Bisection[6]を用いる. クラスタ数は実験より 50 とする. ここでクラスタの中心ベクトルとクラスタ内の文各々の類似度を求め, 類似度がある閾値の幅内にある文を耳より情報とする. このとき閾値の幅は, ある値 α と β の間であるとする.

3. 4 実験

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. 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口コミデータである.

得られた口コミデータは「鹿苑寺」が 2075 文「東京ディズニーランド」が 3862 文「東京スカイツリー」が 6025 文「札幌市時計台」が 2216 文「首里城公園」が 3196 文である.

実験は Repeated bisection 法を用いたツールである bayon を用いた. 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2に示す. 表2により α を0.5とし, β を0.6とする.

表2 5つの観光地の結果

	鹿苑寺	ディズニー	スカイツリー	時計台	首里城
0.5	0.413	0.417	0.295	0.348	0.457
0.55	0.488	0.429	0.326	0.467	0.432
0.6	0.395	0.412	0.488	0.341	0.442
0.65	0.341	0.412	0.425	0.395	0.405
0.7	0.268	0.355	0.406	0.317	0.286
0.75	0.256	0.250	0.400	0.231	0.300
0.8	0.314	0.143	0.308	0.188	0.250

4. まとめ

本研究では, 観光地の口コミ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, クラスタリングする. 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する.

ユーザの旅行先の耳より情報を検索し提示することで, 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と思われる. しかし, 抽出した文が耳より情報かの判断を個人で行ったので正解データの正確性が低いという問題がある. また, 耳よりキーワードを含む文数から適切なクラスタ数の推定, 類似度での閾値, これらの適切な値の検討も進めていきたい.

参考文献

- [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/>
- [3]服部 祐基, 灘本 明代, ソーシャルメディアからのタイプ別耳より情報の抽出手法の提案, 第5回データ工学と情報マネジメントに関するフォーラム, 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
- [5]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
- [6] 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.