D-021

観光地の口コミに含まれる観光地へのアクセス関連情報の抽出手法
Extraction method of access-related information from reviews of tourist attractions

塩田 雄基[1)]　福本 淳一[1)]　西原 陽子[1)]
Yuki Shioda　Junichi Fukumoto　Yoko Nishihara

1　はじめに

多くの旅行者にとって, 観光地を訪れること自体が魅力であると同時に, 目的地にたどり着くまでの過程も旅の楽しみの一つである. しかし, 現状では, ほとんどの旅行者が公式サイトや経路検索サービスを用いてアクセスを確認しており, そこで得られる情報は最寄り駅やバス停からの所要時間, 交通手段の概要にとどまっている. このため, どのルートを通るとより楽しめるか, 坂道の勾配や途中の土産物屋の有無, 時間帯による混雑度の変動など, 実際の移動体験に関わる詳細は把握しにくいという課題がある . 一方, 「じゃらん net」[2] などの大規模な旅行口コミサイトには, 旅行者の実体験が書き込まれている. しかしながら, 口コミは膨大かつ雑多であり, すべてを読んで有益な情報を探し出す作業は現実的ではない.

そこで本研究では, 旅行口コミサイトから収集された観光地の口コミから観光地へのアクセス関連情報を抽出する手法を提案する.

2　提案手法

提案手法の概要を説明する.

2.1　入力：観光地の口コミ集合

web スクレイピングを用いて, 観光地の口コミサイトから口コミデータを収集する. 収集した口コミは, 入力データとしてそのまま扱うにはノイズが多いため, 主に以下の手順で前処理を行う. まず, テキストに混在する HTML タグや改行コード, 特殊記号などを正規表現で削除し, 複数の空白を一つにまとめるなどのクリーニングを行う. さらに, 全角になっている英数字や記号を半角に統一して処理することで, 表記ゆれによる誤判定を抑制する. 次に日本語テキストは, 同じ文字であっても全角と半角が混在している場合があるため,Unicode 正規化を用いて日本語テキストを全角に統一する. 続いて, 日本語の口コミテキストを機械的に解析するために形態素解析エンジン MeCab を用いて, 文章を単語単位に分割し, 品詞タグ（名詞, 動詞, 形容詞など）を付与する.

2.2　アクセス関連情報を含む部分口コミの抽出

本節では, 口コミからアクセス関連情報を含む部分口コミを抽出する方法について述べる.

2.2.1　開始キーワードと終了キーワードの設定

アクセス関連の情報は、多くの場合「どこから移動して」「どのように辿り着いたか」という順序で記述される. そこで、まず口コミ中でアクセス情報が記述されやすい始点を示す語を「開始キーワード」とし、目的地に着いたことを示唆する語を「終了キーワード」として定義する.

表 1 は、清水寺の場合に設定した具体的な開始キーワードと終了キーワードの一覧である. これらの開始キーワードと終了キーワードを用いて, 口コミからアクセス関連情報を含む部分口コミを抽出する.

2.2.2　部分口コミからの係受けペアの抽出

抽出した部分口コミにはアクセスに直接関連しない表現や, 単に場所を述べるだけの断片が多く含まれる場合がある. そこで本研究では, まず GiNZA を用いて各文から係り受けペアを抽出し, 名詞–動詞および名詞–形容詞の組み合わせを含む文のみを抽出する.

2.3　部分口コミのクラスタリング

抽出後の部分口コミに対して, scikit - learn の TfidfVectorizer を用いてベクトル化を行う. 各部分口コミごとに, テキスト中の単語と抽出された係り受けペアの TF - IDF 値を算出し, アクセス関連情報の重要語（「バス停」「バス」「駅」「タクシー」「車」「徒歩」「坂道」「階段」「混雑」「駐車場」「お店」「お土産」「食べ歩き」「参道」「道中」）に対しては TF - IDF 行列上で重みを 3 倍に増強する. これにより重要語に重みをつけたベクトルが得られ, その結果がクラスタリングにも反映される.

次に, 類似する口コミをクラスタリングする. 観光地に関する口コミデータは, 多様な表現や視点から記述されるため, 同一の情報を異なる言い回しで複数の口コミが存在することが多い. そこで, クラスタリングを用いて類似する口コミをグループ化し, 各クラスタから代表的な口コミを抽出することで, ユーザに提示する口コミ数を最適化し, 情報の質を向上させることを目指す.

2.4　出力：クラスタリング結果

クラスタリング結果は, 各クラスタごとに代表文 1 文を抽出して出力する. 代表文の抽出は, 各部分口コミベクトルとクラスタ中心ベクトルと距離を計算し, 最小距離を示す部分口コミを選択することで行う.

3　評価実験

提案システムの評価を行った.

3.1　実験手順

実験では, 旅行サイト「じゃらん net」に掲載された“クチコミ”欄を対象にデータ収集を行った. 5 つの観光地「清水寺」「伏見稲荷」「平安神宮」「金閣寺」「銀閣寺」を対象として, 階層型クラスタリング（ウォード法）を用いてクラスタリングを行い, 距離閾値を 2.0 から 2.4 まで 0.1 ずつ変化させ, その結果を閾値ごとに人手でアクセス関連情報として有益か判別し, 各観光地のクラスタ数と適合率を求めた.

表 1　清水寺の場合の開始キーワードと終了キーワード

開始 キーワード	清水道, 京都駅, 清水五条駅, 駅から, バス停, 祇園四条駅, 最寄りのバス停, 歩いて, 徒歩で, タクシーで, 車で, バスで, 電車で, 自転車で, 清水寺まで, 行くまで, 清水寺への, 参道, 道中
終了 キーワード	清水寺, 到着, 着きました, 着きます, 到着した, 到着する, 舞台, 境内, 行けます, 行きました, 辿り着きました, 着いた, 着く, おすすめ

1)　立命館大学

表 2　距離閾値 2.0 から 2.4 までの適合率

距離閾値	2.0	2.1	2.2	2.3	2.4
クラスタ数	117	96	81	65	53
清水寺	0.51	0.53	0.69	0.70	0.73
クラスタ数	34	27	17	12	11
伏見稲荷	0.56	0.52	0.76	0.75	0.73
クラスタ数	11	8	4	2	2
平安神宮	0.55	0.625	0.75	0.50	0.50
クラスタ数	5	3	3	2	2
金閣寺	0.40	0.66	0.66	1.00	1.00
クラスタ数	13	10	6	4	3
銀閣寺	0.46	0.50	0.50	0.75	0.66

3.2　実験結果

各観光地におけるクラスタ数と適合率をまとめた結果を表 2 に示す.

表 2 より, 距離閾値が 2.2 のとき, 5 つの観光地のうち 2 つの観光地で適合率が最も高く, 閾値を増やしても適合率の上昇が少ないことが示された.

3.3　考察

表 2 において, クラスタ数が減少する一方, 適合率が上昇した. これは, 距離閾値が大きくなるほど類似度の高い部分口コミがより大きなまとまりとして合併され, 同じテーマを共有する口コミが集まりやすくなるためだと考えられる. しかし, 閾値を過度に大きくするとアクセス関連情報が少ない観光地ではクラスタ数が極端に少なくなり, 異なるテーマの口コミが同一クラスタに混在することで代表文の一貫性が低下する恐れがある. 本研究の実験では距離閾値 2.2 が, 5 つの観光地で最適なバランスを示す閾値であると考えられる.

抽出されたアクセス関連情報を含む口コミの例を表 3, 表 4 に示す.

正しく抽出された口コミは「坂道の有無」「道中の土産物屋などの状況」「バス停や地下鉄駅からの所要時間」といった具体的なアクセス情報を豊富に含む. 一方, 誤って抽出された例は「観光地の景観や施設内の様子に関する表現」「アクセスとは直接関係のないチケット情報」「文脈上主要なアクセス情報が無い断片的表現」であった. また, 伏見稲荷のように観光地内が広く, どこまでをアクセス関連情報として有益と判断するか難しい例もあった.

4　おわりに

本研究では, 口コミデータを対象に観光地までのアクセス関連情報を抽出する手法を提案した. 実験の結果, 提案手法により一定程度, アクセス関連情報を含む口コミを抽出可能なことが示された.

参考文献

[1] “音羽山清水寺” https://www.kiyomizudera.or.jp/ （2025 年 4 月 23 日アクセス確認）
[2] “じゃらん net,” https://www.jalan.net/ （2025 年 4 月 23 日アクセス確認）
[3] 馬場優大, 藤生慎, 森崎裕磨: 旅行情報サイトに投稿された口コミデータを用いた観光地の改善点抽出システムの提案, *AI・データサイエンス論文集*, vol.4, no.3, pp.942–951, (2023)
[4] 阪井奎伍, 瀧本明代: 観光地のレビューからの耳寄り情報抽出手法, 知能と情報, vol.6, no.0, pp.49–54, (2015)
[5] 上原尚, 嶋田和孝, 遠藤勉: Web 上に混在する観光情報を活用した観光地推薦システム, 信学技報, NLC2012-35, Dec, (2012)
[6] Shogo Isoda, Masato Hidaka, Yuki Matsuda, Hirohiko Suwa, Keiichi Yasumoto: Timeliness-aware On-site Planning Method for Tour Navigation, *Smart Cities 2020*,3(4), 1383-1404, (2020)

表 3　正しく抽出されたアクセス関連情報の口コミ例

観光地	**例 1**	**例 2**
清水寺	道中にたくさんお土産屋さんがあります	駅からバスで行くとアクセスがいいです
伏見稲荷	道中に景色の良い場所や神聖なスポットが随所にありますので、余裕のある方は山頂コースで巡って参拝されると良いかと思います	道中の道のりは砂利、階段はもちろんのこと、山道のようにでこぼこしていました
平安神宮	バス停も近く、地下鉄からは徒歩 10 分、アクセスは良し	岡崎公園から大鳥居、応天門に向かって歩くと平安京の大内裏に入って行く様な雰囲気が感じられ、壮大なスケール感がとても好きです
金閣寺	道中もなかなかの見応え	バス停からすぐ金閣寺
銀閣寺	ですが、歩く道中で様々なお店があるので楽しいです	バス停があったりしますが、かなり離れていることもありますのでご注意下さい

表 4　誤って抽出されたアクセス関連情報の口コミ例

観光地	**例 1**	**例 2**
清水寺	京都の中でもっとも京都を感じさせる場所、それが清水寺	歩いて三重塔などを見ながら人の流れに従って進み、本堂・清水の舞台に到着
伏見稲荷	参道の朱色の鳥居は圧巻です	稲荷大社の先も奥社行きました
平安神宮	テラス席もあるので観光に疲れたらコーヒーを飲みながら平安神宮	岡崎公園からの雰囲気やスケール感は感じられるが、アクセス手段の具体的情報がない
金閣寺	バスで, 地下鉄・バス共通二日券もおすすめ	（該当する誤抽出例なし）
銀閣寺	バスで 10 分から 15 分ぐらいで到着します私は銀閣寺	金閣寺道, 後は、これも金閣寺と同じなのが、拝観料を払ってくれる点です