D-37

観光情報マイニングのための地域情報の獲得

井上 俊右† 嶋田 和孝‡ 遠藤 勉‡
†九州工業大学情報工学部 ‡九州工業大学大学院情報工学府

1 はじめに

現在，さまざまな地域の基幹産業の一つとして観光業が注目されている．その環境下で，Web に存在する観光情報がいま重要な役割を持っている [1]．本研究では，Web 上の観光情報を収集し，評判分析を行い，可視化する観光情報分析システムの構築を目指している．対象都市は著者らの所属する大学のある，福岡県飯塚市近辺とする．本論文ではシステムの前処理となる，観光情報の抽出の流れと実際に抽出を行った場合の問題点について言及する．情報抽出の対象はリアルタイム性に優れ，外出先での投稿が頻繁に行われる Twitter とする．

2 提案手法

基本的な処理の流れは，(1) 基本クエリの獲得，(2) クエリ関連語の獲得，(3) クエリの生成と検索，(4) フィルタリングとなる．以降，その詳細について述べる．

基本クエリ及びその関連語の獲得 抽出の対象となる基本的なクエリを，市などが用意した観光情報ポータルサイトから取得する．基本クエリとしては，サイト上に記載された観光地名，地域のイベント，飲食店名の名称とする．例えば「嘉穂劇場」や「飯塚観光こども山笠」などである．しかし，基本クエリだけでは，観光に来た人が必ずその地名を言及するとは限らず，またサイトに未掲載の観光情報を取り逃すという問題がある．この問題に対処するために，クエリの関連語を取得する．ポータルサイトからたどれるリンクにある説明文を対象とし，文中の各単語の重要度を算出する．重要度の高い単語から，人間がクエリを選出する．単語への分割には，形態素解析器 MeCab を利用する．重要度の算出には Okapi-BM25[2] を使用する．

クエリの生成と検索 得られたクエリをもとに，Twitter の API を利用して tweet(投稿文) の検索を行う．基本クエリや，その関連語との組み合わせなどで検索を行うが，施設の正式名称がそのまま記述されているとは限らない．例えば「旧伊藤伝右衛門邸」などの長い名称においては,「旧伊藤邸」や「伝衛門邸」などと省略されて記述されている場合がある．そこで基本クエリから考えられる省略語を人手で生成し，クエリとして加えていく．

フィルタリング クエリを基に検索を行った結果，例えば「飯塚」の場合，対象都市としての「飯塚」と人名としての「飯塚」が含まれていることがある．その判別を行うために，いくつかのルールを設定することにより簡易なフィルタリングを行う．例えば，MeCab を用いた単語の分析結果に含まれる「名詞-人名」や「名詞-地名」などの情報を利用する．また手動で簡易な接尾語のルール (〜さんや〜市など) を作成し，それらとのマッチング処理により分別を行う．

3 結果と考察

提案手法を実際の Twitter データに対して適応した．しかし，対象都市のような小規模な観光地では tweet の投稿数が少なく，十分な量の情報が得られたとは言い難い．より多くの tweet を取得するために，さらなるクエリの生成や，Twitter 以外の情報を利用していかなければならない．

検索後，フィルタリングによってノイズの除去を図ったが MeCab では誤解析が多くみられた．その多くは地名を人名と誤認識しているものであり，精度の向上に対して再現率の低下が著しいものとなった．

また，地名と判断されたものであっても，それが対象地域のものなのかという問題がある．例えば飯塚市の「八木山」の場合，仙台市にも「八木山」という地名があり，これらの判別も必要となる．この問題に対しては，その土地特有に現れる共起語，例えば飯塚市なら「バイパス」，仙台市なら「ベニーランド」による判定ができる．また，対象 tweet の前後の発言を調べることで，その土地の関連語句 (「飯塚市」など) を得ることができ，見分けることが可能であった．さらに，発言者自身の情報を利用する (福岡市在住など) と，居住区の近くの観光を行っているとの類推も可能ではあるが，その情報に依存しすぎるのは危険である．なぜなら観光を目的としている以上，あらゆる場所から訪れる観光客を取得することが必要になるためである．

対象地名を呟いていたとしても，そこに観光情報が含まれていない場合がある．発言者が通勤などにおいて対象地を日常的に通過しているだけ，という状況も含まれてしまうためである．そのため，発言の「観光性」についても判断できるような枠組みを考えなければならない．考えられる手法としては，Twitter の投稿時間に着目することである．早朝や深夜ならば生活の一環としての発言が多く，昼間になるにつれて観光としての発言が増すような傾向があった．さらに休日であることや，イベントの当日などの日時によっても判断していく．

4 おわりに

本論文では，Twitter を対象とした観光地情報の取得方法について提案した．小規模な観光地情報の抽出はその情報の少なさにより困難であった．今後はクエリの拡充や情報源の拡大などでより再現率を高めていく．また，略語の生成において人手でのコストを低減するために自動化の仕組みを模索中である．

参考文献

[1] 斉藤. Web における観光情報の提供と分析. 人工知能学会誌, 26(3):234-239, 2011.

[2] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC 1994), 1994.