地方都市を対象とした観光情報の分析に向けて

嶋田 和孝　前田 裕　井上 俊右　遠藤 勉
九州工業大学 情報工学部 知能情報工学科
{shimada, h_maeda, s_inoue, endo}@pluto.ai.kyutech.ac.jp

概要: 近年，観光は様々な地域において基幹産業の 1 つになっている．そのような環境下で，Web に存在する多くの観光情報の分析は重要な役割を持ってきている．本発表では，地方都市を対象とした観光情報の抽出および分析ツールの構築に向けた議論をする．まず， Web の意見情報と既存の観光情報（各都市で用意されたのポータルサイトなどの情報）との対応付けについて議論する．ここでは，分析に有用な情報の獲得方法について考察する．地方都市の場合，必ずしも十分な情報を得られるとは限らないため，様々なアプローチが必要となると考えられる．次に，観光情報の分析手法の一つとして，情報の P/N 判定（肯定的意見か否定的意見かの判断）を題材に，その手法を考察し，実験を行う．Web には様々な特有の表現があり，その問題に対処した手法について提案する．
Keywords: 観光情報の抽出，評判分析，地方都市の観光情報．

1　はじめに

近年，観光は様々な地域において基幹産業の 1 つになっている．観光の活性化は，地域の活性化に繋がると考えられる．そのような環境下で，Web に存在する多くの観光情報の分析は重要な役割を持ってきている [10, 7]．一方で，Web 上には膨大な量の情報があり，またその内容は玉石混淆であるため，それらすべてを人間がチェックするのは現実的ではない．そこで，本研究では，Web から観光に関する様々な情報を抽出し，その内容について自動で分析を行い，整理して提示することで，旅行者の動向や要望を可視化し，人間がさらに詳細な分析するための基盤技術とインタフェースの構築を目指す．図 1 に最終的なシステムのイメージを示す．提案システムを利用することで，観光による地域振興に有用な情報をスムーズに整理および理解できるようになると考えられる．

本論文では，その観光情報分析システムの構築のための基礎技術について提案する．具体的には，対象となる都市に関する観光情報の抽出と評判分析について検討，実装および評価する．ここで，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．観光情報抽出の部分では，基本的な処理の流れと実際に抽出処理を行った場合の問題点について議論する．評判分析では，文の PN 分類を行う．PN 分類とは，その文が肯定的な意見 (P) なのか，もしくは否定的な意見 (N) なのかを分類することである．自動的に獲得された訓練データに基づく機械学習を適用し，その有効性を検証する．なお，本論文では，対象データとして，Twitter[1] の情報を扱う．これは，Twitter のようなマイクロブログでは，ユーザの状況がリアルタイムにログとして残される傾向があり，観光的なイベントや事象に対しての有用な情報が多く存在すると考えられるためである．

2　観光情報の抽出

本節では，観光情報の抽出について述べる．

2.1　手法

基本的な処理の流れは，以下のようになる．

1. 基本クエリの獲得
2. ポータルサイトからの関連語の獲得
3. クエリの生成と検索
4. フィルタリング

以降では，各処理の詳細について述べる．

2.1.1　基本クエリと関連語の獲得

検索の対象となる基本的な情報としては，市などが用意した観光情報のポータルサイトなどを利用する．基本クエリとは，ポータルサイト上に記述された観光施設や飲食店，お祭りなどのイベントの名称を指す．図 2 は飯塚市観光ポータル[2]から得られる観光地情報の例で

[1] http://twitter.com/
[2] http://portal.kankou-iizuka.jp/

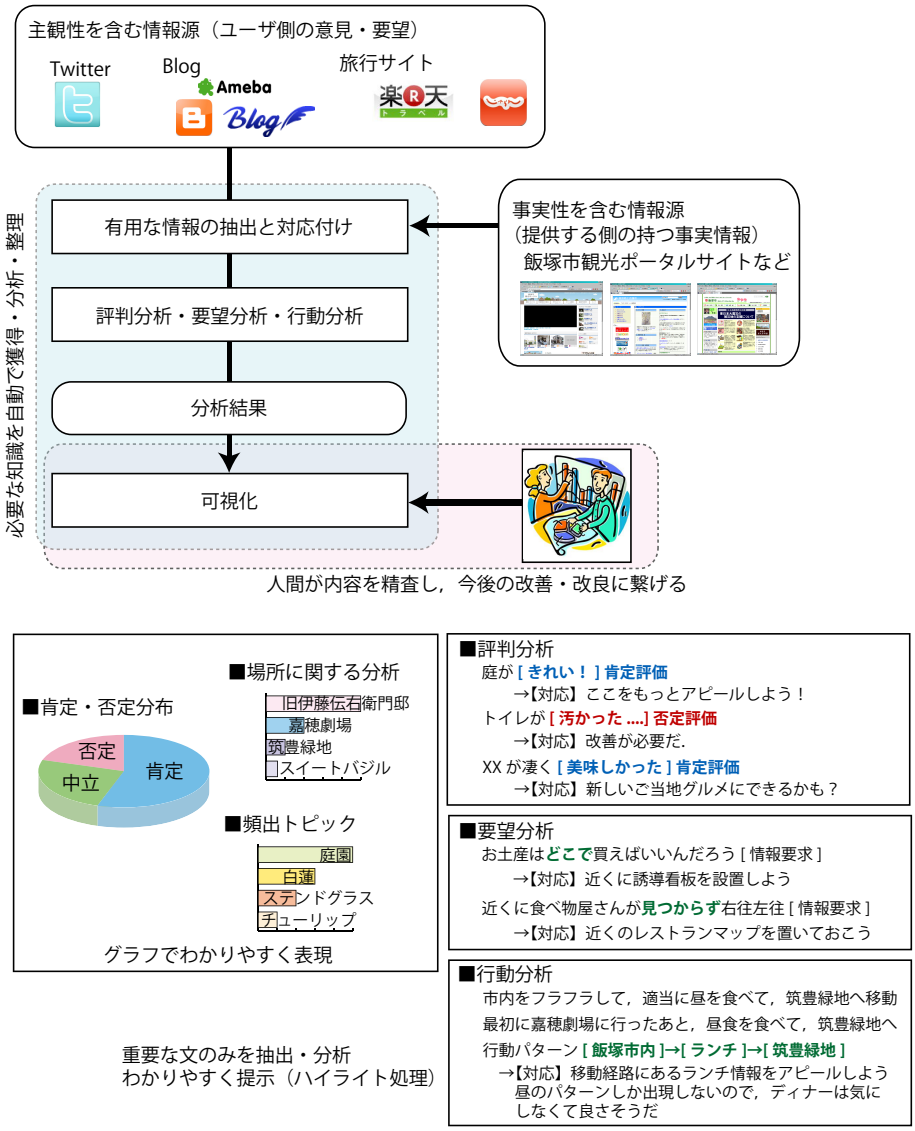

図 1: 最終的なシステムのイメージ図

ある．ここで,「嘉穂劇場」や「旧伊藤伝右衛門邸」などが基本クエリとなる．現段階で，基本クエリは 200 個程度存在する．

しかしながら，観光に訪れた人が必ず施設名などを言及するとは限らないという問題や見出しになっていない関連施設に関するクエリの不在など，基本クエリだけでは，十分ではない．そこで，Web サイト上の情報を利用して，関連語を獲得する．ポータルサイトから辿れるリンク先の説明文などを対象とし，各単語の重要度を算出する．文から単語への分割には，形態素解析器 MeCab[3]を利用する．重要度の算出には，文書検索などで用いられる Okapi-BM25[3] を利用する．具体的には，ある文書 D 中の単語 q_i の重要度は以下の式で計算される．

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_i\cdot(1-b+b\cdot\frac{|D|}{avgdl})} \quad (1)$$

$$IDF(q_i) = log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \quad (2)$$

ここで，$f(q_i,D)$ は，文書 D における単語 q_i の出現頻度を表し，$n(q_i)$ は文書集合の内に q_i を含む文書数，$|D|$ は文書 D の長さ，$avgdl$ は文書集合における文書の平均長，n は全文書数である．b および k は重み付けの定数である．図 3 は「旧伊藤伝右衛門邸」に関する「館内の紹介」であり，このような文から，図中に

[3]http://mecab.sourceforge.net/

図 2: ポータルサイト上の情報

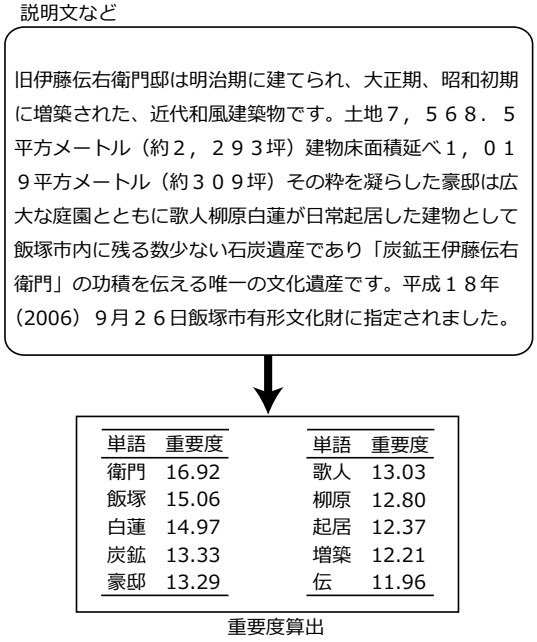

図 3: 説明文と単語の重要度

示されるような重要度付きの単語群が得られる．この例では，この施設に関連の高い「白蓮」や「炭鉱」,「歌人」などの単語が上位に位置している．この単語群から，クエリとしてふさわしいと思われる単語を人手で選択する．

2.1.2 検索とフィルタリング

得られたクエリ候補をもとに，Twitter の API を利用して tweet（投稿文）の検索を行う．検索のクエリとしては，基本クエリ単体や関連語との組み合わせを利用する．しかしながら，正式名称がそのまま tweet 中に記述されているとは限らない．たとえば，飯塚市で最も知られた観光施設の一つである「旧伊藤伝右衛門邸」に関していえば，この名称そのものをユーザが入力している可能性はかなり低い．そこで，考えられる省略形を生成し，それらもクエリとして扱う．この例の場合,「旧伊藤邸」や「衛門邸」などが新たにクエリとして加わる．

次に，検索された結果に対しての処理を行う．得られた tweet すべてが，対象都市（この場合「飯塚市」）に関連しているとは限らない．例えば,「飯塚」という地名の場合，人名としての「飯塚」も検索結果に含まれることが多々ある．そこで，いくつかのルールなどを基に簡易のフィルタリングを行う．具体的には，まず，MeCab の解析結果を利用する．MeCab は，解析結果の品詞層に「固有名詞-人名」や「固有名詞-地域」などの情報を含んでいる．この結果を利用して，人名など，不要な要素を削除する．同様に，手作業で簡易の接尾ルール（～さんや～選手，～市など）を作成し，それらとのマッチング処理を行い，結果を選別する．

2.2 結果と考察

2.1 節で述べた手法を実装し，実際に Twitter のデータに適用した．しかしながら，各クエリで得られる情報は，数が少なく，十分な量とはいえない状態であった．我々が対象としているような地方都市の場合，大型観光地の比較して，必ずしも十分な情報を得られるとは限らない．情報の抽出には,（1）さらにクエリを追加し，再現率の向上を図る,（2）Twitter 以外の情報源からも収集を図る，などの対応が必要である．

クエリの生成では，基本的には手動で略語（「旧伊藤伝右衛門邸」に対する「旧伊藤邸」）を生成した．手動による手法では，クエリ数が増えた場合のコストが大きい．略語については，自動推定する手法などが提案されており [9]，今後の課題の一つである．また，Twitter に関していえば，スマートフォンなどからの入力も多くあり，それらの入力デバイスの特性も含めた略語パターンの推定も興味深いタスクである．

本手法では，簡単なフィルタリングによって，検索結果に含まれるノイズの低減を図ったが，十分ではない．フィルタリングで対応すべき問題は，人名・地名の判別だけでなく，同名異地域の問題も存在する．例えば，対象都市としている飯塚には「八木山」という地名

が存在するが，実際に検索すると仙台にある「八木山」に関する tweet も得られる．これはともに地名であり，提案したようなフィルタリングでは対応できない．これらを分別するために調査したところ，発信者の持つ文脈（例えばプロフィール欄に記載された居住地など）やその単語が表れる前後の tweet の中身（周辺文脈に「福岡」や隣接した地域の名前が出てくるなど）を解析すれば，ある程度の分別が可能であることを確認した．一方で，例えば居住地情報に固執すれば，遠くからの観光客の投稿を見逃す可能性も出てくる．様々な文脈情報を考慮し，より精度の高いフィルタリング処理を行うことも重要な課題の一つである．

フィルタリングして得られた結果は，対象都市に関する観光地の情報を含む文章であるが，それが必ずしも観光情報を意味するわけではない．例えば，日常的に通勤の際に近くをただ通っているだけという場合もある．すなわち，得られた結果の「観光性判断」も大きな課題の一つであろう．ここで，Twitter のリアルタイム性[4]を考えると投稿時間が大きな役割を持つ可能性がある．具体的には，朝早くや夜遅くに観光をしている可能性は低く，それらの時間に投稿された tweet は，「観光地名」を含んでいても日常の tweet である，などの解釈が可能になる．また，リアルタイムではなく，あとで感想を書くことが多い blog などと比較すると，イベントなどの場合はその投稿された日時も重要な判別要素になる．具体的には,「山笠（北部九州でみられるお祭り）」というキーワードの場合，博多を含む様々な地域の投稿がマッチする．この場合，距離が隣接しており，先の述べた投稿者もしくは投稿そのものがもつ文脈情報では切り分けが難しい．しかし，開催される時期にはズレがあり，投稿日時は有効な分別判断の材料となるだろう．Twitter 特有の要素としては,「なう」や「だん」などの特徴的な表現があり，これらの表現と観光地名などの組み合わせは観光性判断に有効であることを予備実験的に確認しており，今後観光情報抽出処理に組み込んでいく予定である．

3 評判分析

本論文では，観光情報の分析の一つとして，評判分析を行う．評判分析は，近年注目されている言語処理タスクの一つである [5]．本論文では，特に文章の内容が肯定的な意見（Positive）か否定的な意見（Negative）かの 2 値に分類するタスク（以降，PN 分類と呼ぶ）を対象とする．

[4] システムレスポンスという意味でのリアルタイム性ではなく，投稿者が実際にその場やその動作をしているときに投稿しているという意味でのリアルタイム性を指す．

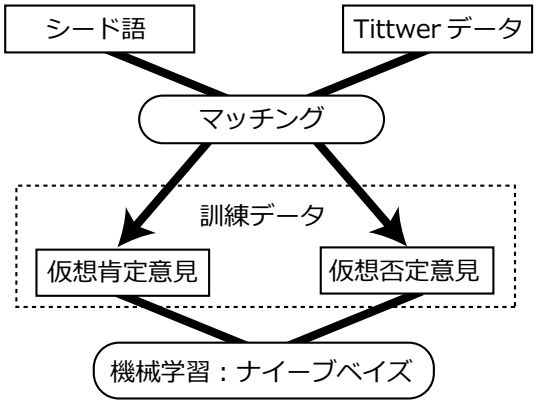

図 4: PN 分類の処理の流れ

3.1 手法

テキストの PN 分類については，様々なアプローチがある．その一つは，機械学習を用いた手法である．Pang ら [2] は映画のレビュー記事を対象とした PN 分類において，いくつかの機械学習に基づく手法の有効性および比較を行っている．機械学習を用いれば，高い精度が見込めるが，一方で高い精度を得るためには十分な訓練データが必要となる．しかしながら一般に，訓練データとなるタグ付きコーパス作成のための作業は高コストである．

本論文では，人手によるタグ付きコーパスなしで，高精度な PN 分類を行う枠組みを導入する．Turney[4] は，語やフレーズの極性値を自動的に算出するために，“excellent” や “poor” といった肯定および否定に関して強い極性をよく表す語との共起度を利用する手法を提案した．本手法もこれに倣い，シード語（種語）から自動的に訓練データを作成する．図 4 に処理の流れを示す．

まず，大量の tweet から，シードを含む tweet を抽出する．ここでは，肯定側 (P) のシードとして「♪」を利用し，否定側 (N) のシードとして「orz」を用いる．「良い」や「悪い」などの単語を用いなかった理由は，Twitter に投稿される文章の特性として，その表記が口語体に近く，また顔文字を代表とする様々な記号のかたまりが頻出する傾向があるためである．このシードと共起した単語群を肯定の表現もしくは否定の表現と仮定し，機械学習に適用する．

機械学習の手法は様々あるが，本論文では，ナイーブベイズ分類器を利用する．ナイーブベイス分類器は，ベイズの定理に基づく確率モデルによる分類器である．

$$P(c|d) = \frac{P(c)P(d|c)}{p(d)} \tag{3}$$

この右辺が最大となるクラスを最終結果とする．ここで，分母の $P(d)$ はクラスに依存しないため，実際には

次のような式で求められる．

$$\hat{c} = argmax_c P(c) \prod_{i=1}^{n} P(x_i|c) \quad (4)$$

ここで，c がクラス（すなわち P もしくは N）であり，x_i は文を構成する単語を表す．ナイーブベイズ分類器は，単語が互いに独立であるというナイーブな仮定に基づく手法であるが，比較的高い精度が得られることが知られている．

3.2 実験・考察

前節で述べた手法を実装し，評価した．評価データとして，人手で 116 tweets をラベル付けした．そのうち，P の数は 64 であり，N の数は 52 であった．

本手法はシードを用いて訓練データを獲得する必要がある．そのためのタグなしデータとして，約 600 万 tweets を利用した．すなわち，この 600 万 tweets に対して,「♪」と「orz」がマッチしたものがそれぞれ P および N の訓練事例である．

実験では，既存の PN 辞書を利用したスコアリング手法を比較対象とした．辞書としては，小林らによって作成された日本語評価極性辞書（用言編）[6] と東山らによる日本語評価極性辞書（名詞編）[11]，鍛治らによる評価表現辞書 [1] を用いた．それぞれの辞書中のエントリには実数値が付与されており（P であれば＋側，N であれば－側の値），この値を基に，評価データ中の各文とのマッチングをとり，そのスコアが＋になるか－になるかで，最終的な PN 判定を行った[5]．

実験結果を表 1 に示す．表中の精度は以下の式で計算される．

$$\text{精度} = \frac{\text{正しく } PN \text{ 判定された事例数}}{\text{全事例数}}$$

表中の NB はナイーブベイズ分類器，Sc はスコアリングを意味する．実験結果から分かるように，提案手法の精度が辞書ベースの手法を上回った．これは特に，評価データ中で，口語体が多く，また特有の記号表現が多用されているような事例で，辞書ベースの手法では正しく PN 判定できなかったためである．このことから，実際に投稿された tweet を訓練データとすることの有効性が確認された．

実験結果は，P および N にそれぞれ 1 つずつのシードを利用した場合の結果である．訓練データはシードに依存するため，精度をさらに向上させるためには，より有効なシードを探し出すことが不可欠であり，今後，実験的に調査を進める予定である．今回は学習器として，ナイーブベイズ分類器を利用した．しかし，より精度のよい様々な学習器が機械学習の分野で提案されており，それらの適用も今後の課題の一つである．

[5]実際には，辞書によってスコアのレンジが異なるので，それについては実験的に正規化している．

表 1: PN 分類の精度.

手法	提案手法+NB	辞書+Sc
精度	**0.82**	0.76

4 おわりに

本論文では，Web から観光に関する様々な情報を抽出し，整理して可視化することで，できるだけ容易に観光情報の分析が可能なインタフェースの構築を目指し，その第一歩として，Twitter を対象とした観光情報の抽出と PN 分類の手法について，提案・実装・評価した．観光情報の抽出に関しては，既存のポータルサイトの情報を利用し，関連語を推定することでクエリを拡張する手法について述べた．提案手法を実装し，抽出処理を行ったが，得られる情報の量は十分ではなかった．地方都市の観光地を対象とした場合，大型観光地の比較して，観光情報の抽出は困難なタスクであり，クエリの拡充や様々な情報源からの情報抽出が必要となる．一方で，徳久ら [8] は，類似した観光地の情報を利用することで，観光開発の支援を行う手法を提案しており，この知見は，我々が対象とする地方都市の観光情報分析にも有効であると考えられ，今後検討が必要である．

評判分析では，文章単位の PN 分類を行った．分類器としてはナイーブベイズ分類器を利用した．一般に高精度な分類器作成のためには十分な量の訓練データが必要だが，その作成は一般に高コストである．この問題を解決するために，本手法では 2 つのシード語を利用し，自動的に訓練データを獲得する枠組みを採用し，実装・評価した．既存の辞書ベースのスコアリング手法と比較して，精度が 6%程度向上し，本手法の有効性が確認された．より高精度な PN 分類を実現するためには，適切なシードの拡張や様々な学習器での実験などが必要になる．また，今回の評価データは 100 文程度で十分でない．より正確な有効性検証のために，大規模な評価データでの実験も今後の課題の一つである．

本論文では，対象都市の観光情報抽出と PN 分類についてのみ実装，評価したが，最終的な目標は，それらだけでなく，要望などの様々な情報を獲得し，分析することであり，これらのタスクは今後の課題である．さらに，分析システムとしては，可視化手法についても十分な検討が必要である．

参考文献

[1] N. Kaji and M. Kitsuregawa. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-CoNLL2007)*, 2007.

[2] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, 2002.

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

[4] P. D. Turney. Thumbs up? or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, 2002.

[5] 乾, 奥村. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, 13(3):201–242, 2006.

[6] 小林, 乾, 松本, 立石, 福島. 意見抽出のための評価表現の収集. 自然言語処理, 12(2):203–222, 2005.

[7] 斎藤. 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.

[8] 徳久, 奥村, 村田. 観光開発支援のためのブログ記事からの評判分析. 観光と情報, 7(1):85–98, 2011.

[9] 村山, 奥村. Noisy-channel model を用いた略語自動推定. 言語処理学会第 12 回年次大会, pp. 763–766, 2006.

[10] 川村, 鈴木, 山本, 松原. 観光情報学. 情報処理, 51(6):642–648, 2010.

[11] 東山, 乾, 松本. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, pp. 584–587, 2008.