

# ホテル宿泊者レビューからの評価表現抽出とレーティング比較

2006MI006 浅井 宣光

指導教員 河野 浩之

## 1 はじめに

近年さまざまなホテル予約サイトが作られている一方ホテル予約サイトが多すぎてどこを使えばいいのかわからないという問題点がある。ホテルを選ぶ際にやはり重要となるのはホテルの評判、価格、レビューである。本研究ではそれぞれのレビューの評価、実際にその評価がどれだけ信頼できるかというものを比較する。

## 2 ホテルサイトの現状と関連技術

最近のホテル検索サイトでは様々な評価を用いてレビューを作っているサイトが増えている。例えば日本のホテル専門サイト「一休」(<http://www.ikyuu.com>)では施設からの返信があったりする。特に最近多いのはさまざまな国のホテルを扱っているサイトが増えてきており、世界的に有名で特にアジアに強いといわれる「agoda」(<http://www.agoda.com>)、世界くまなく提携ホテルを抱えている「Hotel Club」(<http://www.hotelclub.com>)などがある。

## 3 評価表現抽出における先行研究

藤村ら [2] は語の頻出度に応じてのスコア化しスコアを計算している。電子掲示板の評判情報を文単位で肯定、否定に分類する。彼らの研究では「価格.com」のノート PC に関する掲示板、2003 年度の書き込みの内、肯定的な評判 935 件、否定的な評判 551 件を用いている。この評判を基に肯定(否定)的な評判には肯定(否定)的な概念を持った語が多く含まれているはずであると仮定し、肯定的な評判と否定的な評判の差をとっている。なお一般的な語はどちらの文章にも同様に出現するはずなのでその影響は打ち消され、肯定的な評価表現においては正の値をもって、否定的な評価表現について

は負の値をもって抽出される。なお表 2 からわかるように決定木を生成するためのアルゴリズム C4.5、境界に最も近いサンプルとの距離が最大となる SVM(Support Vector Machine) よりも藤村らの手法は (1) と手法 (2)、どちらに置いても正確性が高いことがわかる。

$$score(w_i) = \frac{P_{pos}(w_i) - P_{neg}(w_i)}{P_{pos}(w_i) + P_{neg}(w_i) + k} \quad (1)$$

$$(-1 \leq score(w_i) \leq 1)$$

ここで  $P_{pos}(w_i)$  は肯定的な評判で属性  $w_i$  が出現する確率である。ここで  $P_{neg}(w_i)$  は否定的な評判で属性  $w_i$  が出現する確率である。また  $k$  は、例えば  $P_{neg}(w_i)$  が 0 であった際に起こる 1/1 の問題を解決するために分母に加えた実数である。 $score(w_i)$  はこれらを基にした数字を式に当てはめて計算した結果とする。各文章に含まれる属性のスコアの総和が 0 より大きければ、肯定的な評判とし、0 より小さければ否定的な評判であるというように分類する。

## 4 評価抽出の構成技術

### 4.1 形態素解析

TreeTagger という英文形態素解析ツールを用いる。例えば、「部屋がきれいで新しい」という文章を解析すると「部屋」名詞、「が」格助詞(一般)「きれい」形容動詞、「で」助動詞、「新しい」形容詞というように言語で意味を持つ最小単位の語に分割し、それぞれの品詞を判別する作業を行っていく。

### 4.2 構成手順

まず評判データはコピーアンドペーストによりテキスト形式で収集する。

次に形態素解析ツール, TreeTagger により文書を単語レベルまで分割する。ホテルのレビューでは TreeTagger の辞書に乗っていないような海外の地名などの未知語などが頻出するので (1) の属性を使用する。

表 1 ホテル予約サイトの比較

サイト名	レビュー		言語	特徴	登録ホテル件数
	有無	項目数			
agoda		6	日本語, 英語	全て項目が採点されている	11 万 2 千
一休		7	日本語	施設からの返信がある	数万
tripadvisor		6	日本語, 英語	日本語のレビューが少ない	3 千万以上 (レビュー)
AsiaRooms.com		6	英語	評価をしてない項目がある	数千
HotelClub		5	日本語, 英語	評価が何を点数付けしているのか不明	6 万
AsiaTravel.com		6	日本語, 英語	お勧め項目があるがレビューが少ない	5 万

表 2 P/N 分類精度表 参照:電子掲示板からの  
評価表現および評判情報抽出

分類手法	(1) 形容詞, 形容動詞		(2), (1)+ 名 詞, 未知語	
	肯定	否定	肯定	否定
藤村らの手法	83.8	71.3	86.2	72.9
C4.5	79.1	60.5	78.0	60.3
SVM	79.6	71.8	80.4	73

しかし英語には形容動詞というものがないので形容詞のみを使用する．解析した結果より形容詞を抜き出しテキストに書き込む．

次に単語ごとに否定(肯定)的な文には否定(肯定)的な概念を持った語が多く含まれているはずであるというのを念頭におき、藤村らの手法である数式(1)を使い、否定的な評判と肯定的な評判の差を取り評価表現辞書を作成していきこれらのプログラムを Active Perl(5.12.1)により作成し Mysql(14.12) に格納する．

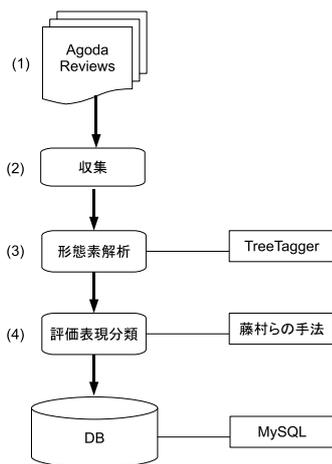


図 1 実装の構成手順

## 5 性能評価

実際にタイ、バンコクにあるホテル宿泊者レビューの中よりランダムに 5 件選びアンケートを取った．アンケートは参考にゲストがつけたスコア，ゲストが書いたレビューを書きスコアを 10 段階評価採点をしてもらい、実際に評価表現辞書を使用し、スコアリングをした．評価表現辞書を作成する際に肯定，否定文という分け目になるポイントを平均点である 7.9，レビューが最も多かった部分よりも小さい直近の分岐点である 7.5 の 2 つの分岐点を用いてスコアリングを行った結果下の表 3 のようにアンケートの平均点に 7.5 を分岐点とした場合にアンケートに近いスコアを得られた．そこでホテルの

レーティングと文書にはやはり関係がないのだと思っホテル 2.5 星から 5 件ずつ、計 30 件を収集しスコアリングをした結果ゲスト評価が 8 よりも低い場合に評価表現辞書でスコアリングをした場合に 81.25% の確率でゲスト評価よりも高い評価，ゲスト評価が 8 以上の場合には 35.71% の確率でゲスト評価よりも高いスコアが算出されている．よって宿泊者のスコアにはスコアが高かったり低かったりするけれども文章はそれほど悪い，良い文章を書いているわけではないということがわかる．

表 3 P/N 分類精度表 参照:電子掲示板からの  
評価表現および評判情報抽出

10 段階評価				
星数	宿泊者	アンケート	7.5	7.9
Manhattan Hotel				
3.5	7	4.85	6	6.3
Sawasdee Sukhumvit Inn Hotel				
2	6.4	7.12	7	6.6

## 6 まとめ

実際にホテルのレビューを 30 件スコアリングした結果、実際にゲスト評価が悪い，良いけれどもレビュー自体には悪い，良いことが書かれていないことがわかる．しかしながら評価表現辞書の作成においてレビューには省略スペリングや英語圏でない人の書き込みのミススペリング，形態素解析辞書内にはない言葉などがありこれらの問題を解決することでさらに正確なスコアリングができるようになるだろう．今後さらに多くの国とホテルを評価表現辞書作成時に参考し，ホテルの情報をデータベースに追加する事で実際に web 上のホテル予約サイトの宿泊者レビューの信憑性を持たせたり宿泊者に頼らない平等なホテル評価をすることも可能である．さらに同じ星のホテルでもこの宿泊者レビューを用いたスコアリングを使うことにより，ホテルの星数が同じ場合にもどのホテルの評判が良いのかなどがわかるようになる．

## 参考文献

- [1] 藤村滋, 豊田正史, 喜連川優, “電子掲示板からの評価表現および評判情報の抽出,” 人工知能学会第 18 回全国大会, 3F1-03, 2004.
- [2] 櫻井敦規, “地理的抱合関係を用いた自動ニュースマップの実装,” 南山大学数理情報学科 2008 年度卒業論文, 2009.
- [3] 佐藤誠也, 祖父江達師, 稲垣諭, “評価表現に基づく飲食店評判マルチファセット検索システム,” 南山大学数理情報学科 2005 年度卒業論文, 2006.
- [4] CPAN, Alvis::Treetagger, <http://search.cpan.org/~wbuntine/Alvis-QueryFilter-0.3/lib/Alvis/Treetagger.pm>, 参照 Nov, 2010,