F-032

出現頻度に基づく自由回答文の格付け推定
楽天 GORA レビューデータへの応用

早坂　透†　　川村　秀憲†　　鈴木　恵二†
Toru Hayasaka　Hidenori Kawamura　Keiji Suzuki
†北海道大学大学院情報科学研究科

1. 概要

近年、チャット、Weblog 等の普及により、個人が自由に情報発信できる環境が整っており、その利用者は年々増加する傾向にあると考えられる。個人が発信する情報にはしばしば、自分が経験したある対象に関する意見、感想、評価などが含まれているため、そういった自由回答文から有用な情報を収集したいというニーズがある。そこで今回、数値データである 5 段階評価と自由回答文からなる楽天 GORA のレビューデータを対象として自由文の格付けを試みる。

2. 目的

ゴルフ場に関するレビューデータはユーザにとって選択の支援となっている。ユーザがゴルフ場の比較をする際、5 段階評価は重要な指標になると考えられる。楽天 GORA レビューデータから 5 段階評価と自由回答文に関連性を見出し、ルールを作成出来れば、web 上にあるゴルフに関連するクチコミやアンケート等の自由回答文からルールを適用することで定量的な情報を得られると考えられる。本研究ではその推定手法を提示する。

3. 楽天 GORA

楽天 GORA はゴルフ場の予約、検索、コースの閲覧などが可能であり、ゴルフ場を予約した利用者はプレー後にゴルフ場のコース、サービス、食事、道具の感想をクチコミとして投稿することができる。このクチコミが今回の研究対象のレビューデータであり、1（悪い）～5（良い）の 5 段階評価が可能な 8 項目（総合評価、コストパフォーマンス、スタッフ接客、コース・戦略性、食事が美味しい、設備が充実、フェアウェイが広い、距離が長い）に加えて、コメント文とそのタイトルを自由に記述する形式となっている(表 1)。

項目	内容
クチコミID	
コースID	
クチコミ投稿者名	「USER000」のようにマスクされた
都道府県	
年齢	10歳代ごとの選択
平均スコア	
オススメ目的	カップル、接待／高級、エンジョイ／カジュアル、アスリート
オススメタイプ	女性、上級者、初心者、中級者から択、複数選択も可
利用回数	
総合評価	1：悪い－5：良い
コストパフォーマンス	1：悪い－5：良い
スタッフ接客	1：悪い－5：良い
コース／戦略性	1：悪い－5：良い
食事が美味しい	1：悪い－5：良い
設備が充実	1：悪い－5：良い
フェアウェイが広い	1：悪い－5：良い
距離が長い	1：悪い－5：良い
タイトル	自由回答
コメント	自由回答
クチコミ投稿日	
プレー日	

表 1　レビューデータの詳細

4. 格付け推定手法

楽天 GORA のレビューデータから 5 段階評価とそれに付随する自由回答文との関連性を取りだす。関連性のイメージは図 1 のようになる。第一に、レビューデータの自由回答文から出現頻度の高い名詞を抽出する。次に出現頻度の高い名詞を含む自由回答文から総合評価を除いた 5 段階評価の各項目（(a)コストパフォーマンス、(b)スタッフ接客、(c)コース・戦略性、(d)食事が美味しい、(e)設備が充実、(f)フェアウェイが広い、(g)距離が長い）それぞれに対して関連する名詞集合 $N_i(i \in \{a,b,c,\ldots,g\})$を抽出する。この関連する名詞集合の要素 $n_{ij} \in N_i$ それぞれに対して頻出する係り受け関係集合 $L(n_{ij})$を抽出する。次に抽出された「名詞－係り受け」の組 $l_k \in L(n_{ij})$毎に 5 段階評価と出現頻度のヒストグラム $H(l_k)$を作成し、それを正規化したものを頻度ベクトル $D(l_k)$とする。同様のヒストグラムと頻度ベクトルを関連する名詞が出現しないデータからも作成しこれを$\tilde{H}$、$\tilde{D}$とする。最後にこれらのベクトルを足し合わせたヒストグラム（式(1)）を作成し、これを正規化したものを D_i とする。ここでαは重み係数とする。

$$H_i=\sum^{k} D(l_k)+\alpha \tilde{D} \qquad \text{式(1)}$$

以下 8 項目ある 5 段階評価の中の「コストパフォーマンス」を例として具体的な手順を踏む。「コストパフォーマンス」という評価項目に対して関連する名詞としては N_a=「値段」「コストパフォーマンス」「料金」「価格」「お値段」が抽出され、「値段」と係り受け関係にある名詞として抽出されたのは出現回数が多い順で、「安い」「考える」「高い」「リーズナブル」等が得られた。l_1=「値段－安い」、l_2=「値段－考える」、l_3=「値段－高い」、l_4=「値段－リーズナブル」、等のそれぞれに組に対して 5 段階評価と出現頻度のヒストグラムと頻度ベクトルを作成する(表 2)。他の名詞集合の要素「コストパフォーマンス」「料金」「価格」「お値段」の係り受け関係にある名詞を抽出し、同様のヒストグラムと頻度ベクトルを作成する。また「値段」「コストパフォーマンス」「料金」「価格」「お値段」について言及されていないデータからも同様のヒストグラムとベクトルを作成する。以上の頻度ベクトルを全て足し合わせ、新たなヒストグラムを作成し、正規化を行い、頻度ベクトルを作成する。これを「コストパフォーマンス」に関する推定頻度ベクトル D_a と呼ぶ。最後にαを決定するため実際の 5 段階評価ベクトル R_a と比較する。$\alpha(0.1 \leqq \alpha \leqq 1.0)$とし、$R_a$-$D_a$ は $\alpha=0.7$ のときに最小となることから、Σ

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.

$k_{D(l_k)}+0.7\tilde{D}$を推定頻度ベクトルとするルールが得られる。

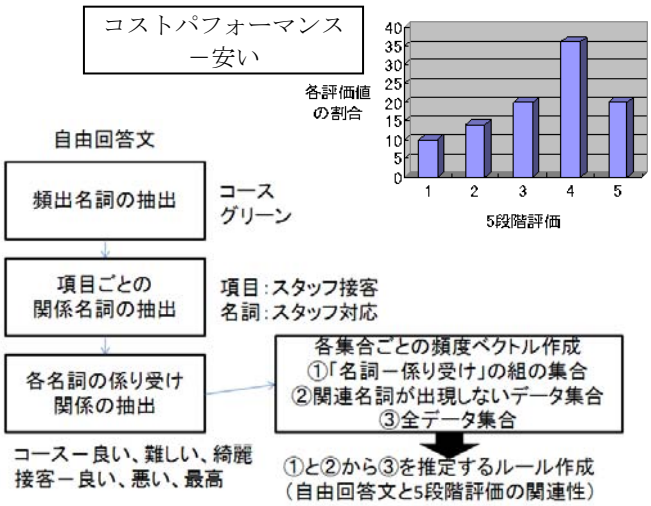

図 1. 自由回答文と 5 段階評価の関連性

単語	出現回数	1	2	3	4	5
安い	2419	0.006201	0.035965	0.200909	0.346424	0.4105
考える	586	0.001706	0.010239	0.156997	0.389078	0.44198
高い	302	0.089404	0.271523	0.34106	0.172185	0.125828
リーズナブ	87	0	0	0.16092	0.448276	0.390805
相応	82	0	0.012195	0.390244	0.402439	0.195122
満足	66	0	0.015152	0.181818	0.393939	0.409091
なる	65	0.015385	0.030769	0.446154	0.307692	0.2
いい	50	0.08	0.06	0.18	0.38	0.3
良い	50	0.02	0.02	0.18	0.38	0.4
仕方ない	29	0.034483	0	0.241379	0.37931	0.344828
やすい	28	0	0.035714	0.25	0.392857	0.321429
ある	22	0.090909	0.136364	0.181818	0.318182	0.272727
言う	17	0	0	0.117647	0.235294	0.647059
よい	15	0	0	0.333333	0	0.666667
行く	13	0	0.076923	0.307692	0.230769	0.384615
見合う	12	0	0	0.25	0.583333	0.166667

表 2. 値段と係り受け関係にある名詞と
その 5 段階評価の分布

5. 結果と考察

推定頻度ベクトル D_a を作成したデータとは別のレビューデータ群(正解用データ)の自由回答文にルールを適用し、推定の精度を検証した。この際、正解用データからもルール作成時と同様に、「コストパフォーマンス」と関連性のある名詞集合を抽出し、各々から係り受け関係の単語を抽出し、「名詞ー係り受け」の組毎に、出現回数を分析する。図 2 にあるゴルフ場における推定頻度ベクトルと正解評価ベクトルを、図 3 にゴルフ場 10 施設の誤差を示す。図 3 より推定平均値と正解平均値の誤差は最大でも 1.0 未満であった。さらに 10 施設のうち 8 施設に関して誤差は 0.6 未満であった。また誤差の最低値は 0.069 で、平均値は 0.28 であった。これより厳密な推定を必要しない場合、かつ大規模なデータから全体としての評価を知りたい場合には、この推定ルールを用いて自由回答文から 5 段階評価を自動的に推定することが可能であると言える。

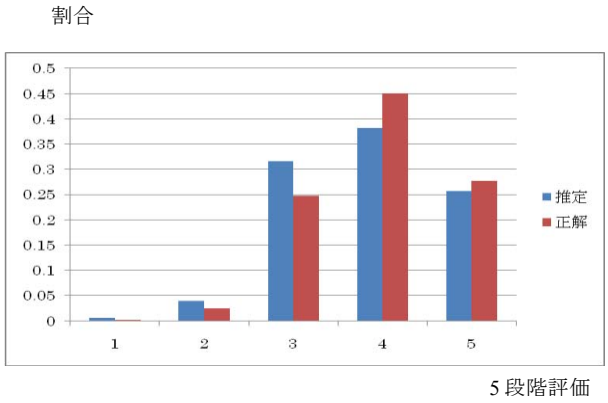

図 2. あるゴルフ場における推定頻度ベクトルと正解評価ベクトル

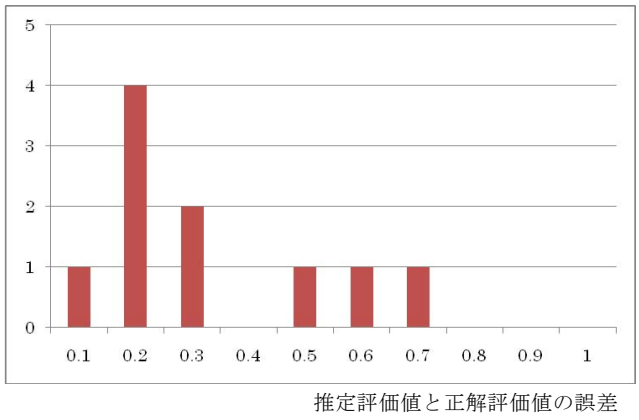

図 3. ゴルフ場 10 施設の誤差

6. おわりに

本提案により、自由回答文から 5 段階評価への推定が可能となった。

今後は推定ルールの重みづけを工夫することで精度が上がると考えられる。また係り受け関係の抽出単語全てを推定に用いたが、その中には 5 段階評価とは関係ないものも含まれているので、それらの単語を選別することで正確さが上がると考えられる。

【謝辞】本稿執筆にあたり、分析の対象データとして株式会社より楽天 GORA のレビューデータをご提供いただきましたことを深く感謝いたします。また、株式会社日立東日本ソリューションズよりテキストマイニングシステム CoreExplorer をご提供いただき、分析に関してご支援いただきましたことを深く感謝いたします。

【参考文献】[1] 福井知子、川村秀憲、鈴木恵二“楽天 GORA のレビューデータを対象としたテキスト分析”観光情報学会　第 2 回研究発表会、pp. 33-36（2010）
[2] 福井知子、川村秀憲、鈴木恵二、“楽天 GORA のレビューデータに関する研究”電子情報通信学会技術研究報告、AI2010-58－AI2010 -63、pp.17-20（2011）

Copyright © 2011 by Information Processing Society of Japan and
The Instiute of Electronics, Information and Communication Engineers
All rights reserved.