F-020

商品の比較履歴とユーザーレビューに基づく推薦手法に関する一考察

A Study on Recommend Method using Customer reviews and Comparative logs

榮枝隼人† 三川健太† 後藤正幸‡
Hayato Sakaeda Kenta Mikawa Masayuki Goto

1 はじめに

近年，情報通信技術の発展や，消費者の購買行動の多様化に伴い，数多くの電子商取引（Electronic commerce：EC）サイトが存在しており，その用途も多様化している．しかし，これら EC サイト上には膨大な量の商品が掲載されているため，ユーザーが掲載されている商品を網羅して，商品の選択を行うことは困難である．そこで，ユーザーの商品選択を支援するため，多くの EC サイトで購買履歴やユーザーによる評価値からユーザーの嗜好を特定し，商品を推薦するシステムが実装されている [1]，[2].

一方，EC サイト上には，過去に商品を購買したユーザーによるレビューが掲載されており，消費者の商品，サービスに対する感想，評価，要望などが多く含まれていることより,「第三者の購買意思決定に影響を与える」,「ユーザーの製品，サービスに対する意見が取得できる」などの理由から，非常に重要なマーケティング情報となっている．このようなテキストデータに対しては，テキストクラスタリング，代表意見抽出や評判分析など計算機を用いた様々な分析手法が提案され，現在も盛んに研究が進められており [3]，ユーザーによるレビュー情報を積極的に活用することの有効性は高いと考えられる．

さらに，EC サイトには，Webclip と呼ばれるユーザーが注目，検討した商品を一時的に保存できる機能が実装されているものも存在する．この機能を用いることでユーザーは興味をもった商品間の比較をよりスムーズに行うことができるため，購買時の意思決定における補助ツールとして活用されている．Webclip 上に登録された商品は，アクティブユーザーが購買意思決定前に，まさに比較検討している商品情報を示しており，現時点においてユーザーが探索しているものである．したがって，ユーザーの過去の購入履歴等から推定されるユーザーの嗜好だけでなく，この情報を活用することは意義があると考えられる．例えば，過去の購買履歴において，ビジネスホテルを多く選択していても，購買する際に温泉旅館を探している場合は，ビジネスホテルの推薦は意味をなさない．このように，アクティブユーザーの過去の嗜好に合致しているからといって，その商品情報が常に有用な推薦商品となり得る訳ではない．そこで，アクティブユーザーが，購買意思決定時にどのような商品を求めているのかという情報は，Webclip に保存された商品を抽出することで把握できるのではと考えられる．そのため，その有効活用は消費者の購買意思決定に大きな影響を与えることが期待できる．

これらの点から，本研究では，テキストデータ，Webclip 情報の有効活用という視点から新たな推薦システムに対する枠組みを与える．この枠組みの下，テキストデータを用いた商品特性の抽出手法，Webclip の情報を用いたユーザーの嗜好抽出手法を提案する．さらに，テキストと Webclip から抽出した商品特性と消費者の嗜好を用いて，評価点のみを利用した方法とは異なった商品推薦を行う．また，代表的な EC サイトの一つである「じゃらん.net」[4] を事例として分析を行ない，本研究の可能性を示す．

2 準備

2.1 ユーザーレビュー

ユーザーレビューとは，ユーザーが購入・使用した製品やサービスに対して，点数やテキストデータでその評価を与えたものである．これらは EC サイト上に多く投稿されており，ユーザーは他のユーザーの属性や，与えた評価点と共にレビューを閲覧することが可能であり，購買意思決定の補助ツールとして用いられている．ユーザーレビューの一例として，実際に「じゃらん.net」[4] に投稿された宿泊施設利用者のレビューを図 1 に示す．

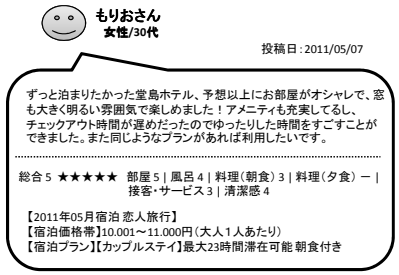

図 1． ユーザーレビューの一例

このレビューから得られる情報は，宿泊者の性別，年代，投稿日などの「属性情報」と，宿泊施設の内容及び宿泊施設への意見や感想などの「内容情報」である．

2.2 推薦システム

推薦システムとは，ユーザーの行動履歴を用いて，ユーザーの嗜好を判断し，そのユーザーの嗜好に適した商品を推薦するシステムである．

ユーザーによるテキスト集合を $\Delta = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_D\}$，宿泊施設集合を $\mathcal{A} = \{A_1, A_2, \cdots, A_M\}$, 評価項目の種類数を P としたとき，m 番目の宿泊施設 A_m に対する評価点を $\boldsymbol{E}_m = (e_{m1}, e_{m2}, \cdots, e_{mP})$ で表す．但し，$(1 \leq e_{mp} \leq v)$ である．また，宿泊施設 A_m のテキスト集合を $\mathcal{A}_m$ で表す．

$\boldsymbol{E_m}$ や Δ の下，ユーザーに対して $\mathcal{A}$ の中から好むと想定されるものを予測し，推薦する．

†早稲田大学大学院創造理工学研究科
‡早稲田大学理工学術院

2.3 相関係数法

推薦システムは，類似したユーザー同士は，購買する商品もまた類似しているという仮定の下，類似ユーザーの購買した商品のうち被推薦ユーザーが未購買のものを推薦するしくみである．このようなユーザー間の類似性を測る代表的な手法として，相関係数法が用いられている．入力データを $\boldsymbol{x} = (x_1, x_2, \cdots, x_H)$，$\boldsymbol{y} = (y_1, y_2, \cdots, y_H)$，その平均を $\bar{x}$，$\bar{y}$ とすると相関係数は

$$C = \frac{\sum_{h=1}^{H}(x_h - \overline{x})(y_h - \overline{y})}{\sqrt{\sum_{h=1}^{H}(x_h - \overline{x})^2}\sqrt{\sum_{h=1}^{H}(y_h - \overline{y})^2}}, \quad (1)$$

として表すことができる．C は 2 つの座標間の類似性の度合いを示す統計学的指標であり，-1 から 1 の間の実数値をとる．これが 1 に近いときは 2 つの確率変数には正の相関があるといい -1 に近ければ負の相関があるという．0 に近いときはもとの確率変数の相関は弱い. 正の相関が強いものを類似性が強いと判断する．

2.4 従来研究

テキストデータのマーケティングへの活用を行った研究として，上田らによる研究 [1] がある．上田らは定性的な情報であるテキストデータを定量的に扱い，他のメタデータと組み合わせることで，有用な情報の抽出を行った．また，本研究同様に宿泊施設サイトを対象とした田邊らの研究 [5] がある．この研究では，テキスト情報を活用し，宿泊施設の特徴を分析することで，企業の戦略に対して有効な知見を得ることを目的としている．しかしながら，テキストの分類には人手を介しており，分析の自動化やユーザーに対する推薦に関しては対象としていない．

また，筆者らは Web 上のユーザーコメントを対象に，宿泊施設に対するユーザーコメントに含まれる単語と，宿泊施設に対するユーザーの満足度などのメタデータの関係性を CHAID 分析を用いて明示化し，両者の関係を抽出する手法 [6] の提案を行っている．以下では，ここから得られた知見の一部を用いて，テキストデータの集約を行う．

3 提案手法

3.1 問題設定

テキスト，Webclip 情報の有効活用という視点から，本研究では推薦システムに対する新たな枠組みを与える．ユーザーに宿を推薦するため，

1. 定性的な情報であるユーザーの感想，評価，要望などが含まれている，宿に投稿されているユーザーレビューを用いた宿の特徴の集約
2. 購買意思決定時の，ユーザーの嗜好を表す Webclip の履歴を用いたユーザーの特徴の集約

を行い，ユーザー，宿の特徴をそれぞれ抽出し，両者の類似度を算出する．さらに，これらの類似度が高いものを候補として推薦する．過去のユーザー情報を示す購買履歴ではなく，購買意思決定時のユーザー情報を表す Webclip を活用することで，ユーザーの現状に適した宿を推薦することが可能となる．図 2 に推薦システムのイメージを示す．

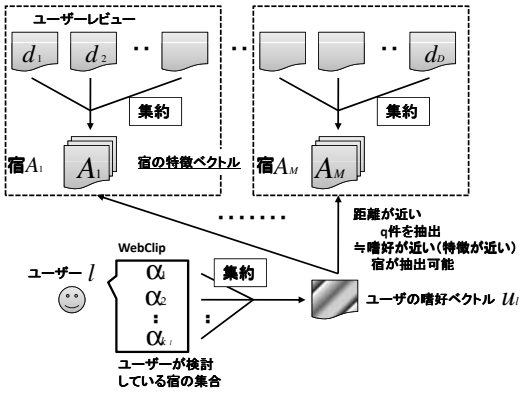

図 2． 推薦システムのイメージ

3.2 モデルの構成

レビュー集合を $\Delta = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_D\}$ で表し，宿泊施設集合を $\mathcal{A} = \{A_1, A_2, \cdots, A_M\}$ で表す．ここで D と M は総異なりレビュー数と全異なり宿泊施設の数を表す．レビュー集合 Δ で使用されている単語を単語集合 $\sum = \{w_1, w_2, \cdots, w_W\}$ で表す．W は全ユーザーレビューに存在する異なり単語数を表す．

ここで，レビュー $\boldsymbol{d}_i$ に含まれる単語集合 $\sum$ の各要素 W_w の出現有無を v_w^i を用いて，$\boldsymbol{d}_i = (v_1^i, v_2^i, \cdots, v_W^i)$ で定義する．但し v_w^i は 0-1 の 2 値をとる要素である．レビューはどれか一つの宿泊施設に対して投稿されている．宿泊施設 A_m のテキスト集合を $\mathcal{A}_m$ で表す．このとき，宿の特性の代表ベクトルは，

$$\boldsymbol{g}_m = \frac{1}{|\mathcal{A}_m|}\sum_{\boldsymbol{d}_i \in \mathcal{A}_m} \boldsymbol{d}_i, \quad (2)$$

$$= \frac{1}{|\mathcal{A}_m|}\sum_{\boldsymbol{d}_i \in \mathcal{A}_m} (v_1^i, v_2^i, \cdots, v_W^i), \quad (3)$$

$$= (g_{m1}, g_{m2}, \cdots, g_{mW}), \quad (4)$$

のように表すことができる．ここで $|\mathcal{A}_m|$ は宿 A_m に投稿されたレビュー数を示す．但し，v_w^{i*} は v_w^i を重み付けしたものである．さらに，情報検索で用いられる代表的な重み付け手法である *idf* により基準化した代表ベクトルを，

$$\boldsymbol{g}_m^* = \frac{1}{|\mathcal{A}_m|}\sum_{\boldsymbol{d}_i \in \mathcal{A}_m} (v_1^{i*}, v_2^{i*}, \cdots, v_W^{i*}), \quad (5)$$

$$= (g_{m1}^*, g_{m2}^*, \cdots, g_{mW}^*), \quad (6)$$

のように定義する．宿の特徴ベクトルとすることで，$\boldsymbol{g}_m^*$ はレビューを集約して宿 A_m の特性を表していると考えられる．単語頻度ベクトルのままで距離を測ると，相対的に出現頻度が高い単語に依存してしまう傾向にある．そこで，単語同士の出現頻度の違いによる影響度の大きさをを平準化するため，*idf* により基準化し，宿の特徴ベクトルとする．(1) から (4) 式の宿泊施設の特徴ベクトルを作る過程のイメージを図 3 に示す．

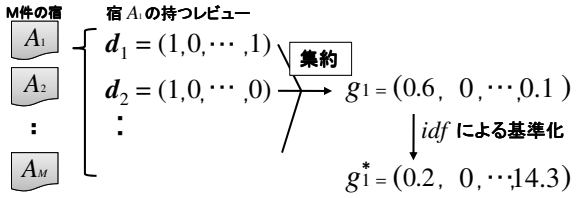

図 3． 特徴ベクトルの作成過程

また，宿泊施設 A_m に対応する評価点ベクトル $\boldsymbol{E}_m = (e_1, e_2, \cdots, e_P)$ とし，単語出現ベクトル $\boldsymbol{g}_m^*$ を持ち，これら二つを宿泊施設の特徴として用いる．但し，$(1 \leq e_p \leq 5)$ である．また，購買行動を行なうユーザーをアクティブユーザーと定義する．そのアクティブユーザー l が K_l 件の宿を Webclip したとし，$\boldsymbol{g}_{k_l}^*$ はユーザー l が選んだ任意の宿 A_m の特徴ベクトルと定義する．ユーザー特性の一部を表す嗜好ベクトルをアクティブユーザーが選んだ K_l 件の宿の特徴ベクトルの平均，

$$\boldsymbol{u}_l = \frac{1}{K_l}\sum_{k_l=1}^{K_l} \boldsymbol{g}_{k_l}^*, \tag{7}$$

$$= (u_{l1}^*, u_{l2}^*, \cdots, u_{lW}^*), \tag{8}$$

と定義する．また，同様に評価点に対しても平均を算出し，ユーザーの嗜好評価点ベクトルとする．E_{k_l} はユーザー l が選んだ任意の宿 A_m の評価点ベクトルと定義すると，アクティブユーザーの嗜好評価点ベクトルは，

$$\boldsymbol{E}_l = \frac{1}{K_l}\sum_{k=1}^{K_l} \boldsymbol{E}_k, \tag{9}$$

$$= (e_{l1}, e_{l2}, \cdots, e_{lP}), \tag{10}$$

と定義する．

3.3 類似度算出

提案手法では，宿特性とユーザー特性の類似度を算出する方法として，相関係数法を利用する．相関係数は，評価点ベクトルと宿の特徴ベクトルの両者について算出する．

$$C_E = \frac{\sum_{p=1}^{P}(e_{mp} - \overline{e}_m)(e_{lp} - \overline{e}_l)}{\sqrt{\sum_{p=1}^{P}(e_{mp} - \overline{e}_m)^2}\sqrt{\sum_{p=1}^{P}(e_{lp} - \overline{e}_l)^2}}, \tag{11}$$

$$C_W = \frac{\sum_{w=1}^{W}(g_{mw}^* - \overline{g}_m^*)(u_{lw} - \overline{u}_l)}{\sqrt{\sum_{w=1}^{W}(g_{mw}^* - \overline{g}_m^*)^2}\sqrt{\sum_{w=1}^{W}(u_{lw} - \overline{u}_l)^2}}. \tag{12}$$

C_E は，アクティブユーザーが Webclip に登録した宿の平均評価点と，各宿の全ユーザーによる平均評価点の相関係数，C_W は特徴ベクトルによる同様の相関係数を示す．本研究ではこれら両者が共に高いものを上位 10 件推薦する．

3.4 学習・予測アルゴリズム

提案手法は以下のアルゴリズムで学習を行い，推薦商品を予測する．

Step0) 宿泊予約サイトからユーザーレビューを収集する．

Step1) テキストデータに対する前処理を行う（形態素解析・不要語除去・単語出現有無ベクトルに変換）．

Step2) 式 (3) を用いて，宿ごとの特徴を示す宿の特徴ベクトルを作成する．

Step3) 式 (6) を用いて単語頻度ベクトルを *idf* により，基準化する．

Step4) 式 (7),(9) より，アクティブユーザーが Webclip した宿からその嗜好ベクトルを算出する．

Step5) 式 (11),(12) の相関係数法を用いて，アクティブユーザーの嗜好と近い特徴をもつ宿泊施設の推薦を行う．

(Step5) において，宿泊施設は評価点ベクトル $\boldsymbol{E}_m$ と宿の特徴ベクトル $\boldsymbol{g}_m^*$ を持つため，この二つの特徴量を単独で相関係数法を行った結果と，両方を組み合わせ相関係数法を行った結果で，ユーザーの嗜好ベクトルに近い宿泊施設を推薦し，評価点と特徴ベクトルを組み合わせる有用性を示す．

(Step4)(Step5) の推薦を行うイメージを図 4 で示す．

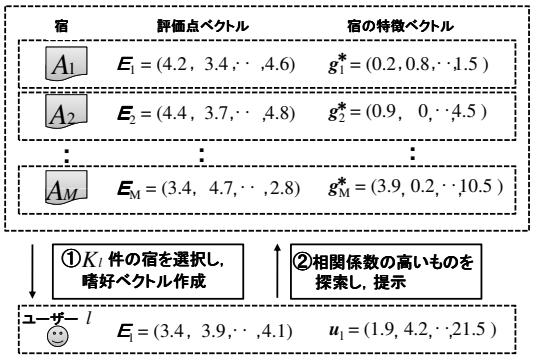

図 4． 嗜好の特定と推薦

図４のように，ユーザーが選択した複数の宿により，ユーザーの嗜好を判定する．そのユーザーの嗜好と，類似したものを全宿泊施設集合の中から推薦する．

4 実験

4.1 実験条件

本研究では，宿泊予約サイト「じゃらん.net」[4] 内のユーザーレビューを用いる．分析対象は,「じゃらん.net」内からランダムに抽出した 5800 件の宿泊施設に対する，合計 195,000 件のユーザーレビューとする．このユーザーレビューには，部屋・風呂・朝食・夕食・サービス・清潔感の各項目に対して，1〜5 までの評価点情報とテキスト情報が含まれている．以降では，評価点とテキスト情報を用いて，宿泊施設の特徴とユーザーの嗜好を明らかにし，ユーザーに対して推薦を行う．

4.2 実験結果

本研究において，宿泊施設は評価点ベクトルと単語頻度ベクトルの 2 つを持つ．実験では，(1) 評価点ベクトルのみを利用，(2) 単語頻度ベクトルのみを利用 ，(3) 評価点ベクトルと単語頻度ベクトルの両方を利用，の 3 パターンに対する推薦を考え，それぞれの場合における推薦結果を示す．

あるユーザーが「良い風呂に入りたい」「京都に行きたい」という条件の下，複数宿を Webclip し，その際に推薦された宿泊施設の実験結果を図 5〜図 7 に示す．

推薦順序	評価点ベクトルのみ	g_{mw} が大きな値をとる単語	C_E
1位	十津川温泉　ホテル昴	加湿器・入浴剤	0.934
2位	旅の宿　ぼっぽ庵	湯布院・風景	0.933
3位	グリーンランド中洲店	マッサージ・忘年会	0.927
4位	牧水苑	塩焼き・山々	0.923
5位	ゆ宿　美や川	家具・硫黄	0.922
6位	玉峰館	家具・かけ流し	0.919
7位	錦水館	京都旅行・飲み屋	0.916
8位	志賀パークホテル	電車・駅連結	0.911
9位	花の宿　にしき園	アロマ・お花	0.906
10位	日杵　湯の里	越後温泉・日本酒	0.905

図 5． 評価点ベクトル単独結果

推薦順序	特徴ベクトルのみ	g_{mw} が大きな値をとる単語	C_W
1位	錦水館	京都旅行・飲み屋	0.707
2位	礼文島　三井観光ホテル	暖炉・自家製	0.600
3位	はなやホテル	せせらぎ・お寿司	0.378
4位	四季育む宿　然林房	京都旅行・鮎	0.307
5位	アランヴェールホテル京都	京都・展望	0.280
6位	京都ホテルオークラ	京都旅行・シャワールーム	0.273
7位	柚子屋旅館	京都旅行・日本酒	0.272
8位	長楽寺宿坊　遊行庵	京都旅行・お粥	0.270
9位	ペンション　ブモリ	暖炉・ゲレンデ	0.259
10位	京都東急ホテル	京都駅・シャトルバス	0.250

図 6． 単語頻度ベクトル単独結果

推薦順序	評価点と特徴ベクトル	g_{mw} が大きな値をとる単語	$C_E + C_W$
1位	錦水館	京都旅行・飲み屋	1.623
2位	はなやホテル	せせらぎ・お寿司	1.176
3位	奥屈斜路温泉　ランプの宿　森つべつ	暖炉・フレンドリー	0.975
4位	箱根強羅温泉　コージーイン 箱根の山	暖炉・鳥	0.967
5位	旅の宿　ぼっぽ庵	湯布院・風景	0.946
6位	海とエステと美肌風呂の宿　天使のイルカ！	日の出・フルコース	0.933
7位	南欧ホテル　エズ・ヴィラージュ	自家製・外国	0.927
8位	十津川温泉　ホテル昴	加湿器・入浴剤	0.927
9位	玉峰館	家具・かけ流し	0.918
10位	グリーンランド中洲店	マッサージ・忘年会	0.917

図 7． 評価点と単語頻度ベクトル両方結果

(3) の宿の評価点ベクトルを利用した方法，(2) の宿の特徴ベクトルを利用した方法のそれぞれ単独で示した結果は，お互いに大きく異なっているが，(3) の双方を利用した結果は，(1)(2) で出てきた結果と重複する部分が存在した．

4.3 考察

評価点ベクトルのみを利用した推薦結果においては，ユーザーの選択した宿は「風呂」と「夕食」の項目において評価点が高いものが多かったので，推薦された結果を見ても，同様に 2 項目の評価点が高いものが多かった．上位 10 件においては全て相関係数が 0.9 を越え，評価点はかなり類似しているものが選択されている．評価点は特徴量が少ないので，高い値をとりやすいが，単語を見ると共通するものは少ない．

特徴ベクトルのみを利用した推薦結果においては，ユーザが選択した宿で「京都旅行」という単語を利用されているものが多く，推薦結果も「錦水館」や「四季育む宿　然林房」などの京都にあり「京都旅行」という単語が多く使用されている宿泊施設が選択された．全レビューにおいて単語頻度が高い単語よりも，全体では単語出現が少ないが，特定の宿のみにおいて単語出現件数が多い単語の方が，推薦結果に大きな影響を与えることがわかった．評価点と比較すると，ベクトルの要素数も多く，要素の値が特に定められていないため，相関係数は 1 位のものでも 0.7 前後，2 位のものでは 0.6 前後と，評価点ベクトルによる相関係数の値よりも小さかった．

宿の評価点と特徴ベクトルの両方を利用した結果では，双方の手法により上位に出現した宿泊施設と同様のものが複数件推薦されている．しかし，評価点ベクトルと単語出現ベクトルの双方でバランス良く高い相関係数を示した宿泊施設が推薦されるなど，単独の手法では得られない推薦結果を得ることができた．そのことで，ユーザーの「良い風呂に入りたい」「京都に行きたい」という定量的な希望と定性的な希望の両方を考慮した結果が得られたと考えられる．

5 結論及び今後の課題

本研究ではユーザレビューを集約して宿泊施設の特徴を明らかにした．また，ユーザの選択履歴からユーザの嗜好を推測し，その嗜好に合った特徴を持った宿泊施設の推薦を行った．評価点及びテキスト情報が得られたので，それぞれ単独で利用したものと組み合わせて推薦を行った結果を示した．このことで，テキスト，Webclip 情報の有効活用という視点から，新たな推薦システムの枠組みを与えた．

今後の課題として，ユーザーテストによる推薦結果の評価や，テキスト情報のより有効な活用方法の考案が考えられる．また，ユーザーの性別や年代などの情報を活用して，ユーザの属性情報に適した商品の推薦手法の提案などが考えられる．さらに，本研究では類似性を測る際に相関係数法を利用したが，今後は別の手法を用いて比較実験を行うことで，更なる精度の向上を目指す．

参考文献

[1] 上田隆穂, 黒岩祥太, 戸谷圭子, 豊田裕貴, “テキストマイニングによるマーケティング調査,” 講談社サイエンティフィク, July 2005.

[2] 島松千春, 御手洗秀一, 伊東栄典, 廣川佐千男, “クチコミ情報からの関連商品マイニング,” 電子情報通信学会信学技報, DEWS2008, March 2008.

[3] 乾孝司, 奥村学, “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,” 言語処理学会, Vol. 13, pp. 201–241, July 2006.

[4] じゃらん.net：http://www.jaran.net/

[5] 田邊亘, 後藤正幸, “宿泊施設の戦略構築を支援するユーザーレビュー分析に関する一考察,” 武蔵工業大学環境情報学部情報メディアセンタージャーナル, Vol. 9, pp. 91–101, July 2008.

[6] 榮枝隼人, 三川健太, 後藤正幸, “宿泊施設を対象とした評価サイトにおけるユーザーレビュー分析に関する一考察,” 日本経営工学会, Vol. 103, pp. 192–193, October 2010.