商品の比較履歴とユーザレビューに基づく推薦手法に関する研究

情報数理応用研究　5210C014-4　榮枝隼人
指導教員　後藤正幸

A Study on Recommender Method based on Customer reviews and Comparative Logs

SAKAEDA Hayato

1　はじめに

近年，多数の電子商取引（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．これら EC サイト上には膨大な量の商品や口コミ（ユーザレビュー）が掲載されているが，ユーザの興味に合致した商品を発見できないために購買が成立せず，機会損失が発生している可能性がある．ユーザの購買活動を補助し，売上を向上させるため，多くの EC サイトで購買履歴などから被推薦ユーザの嗜好を考慮し，購買確率が高いと予測される商品を推薦するシステムが実装されている [1].

一般に，推薦システムにおける推薦商品の予測には，商品の特徴に基づく内容ベースフィルタリング（content-based filtering:CB）[2] と，ユーザ同士の類似性に基づく協調フィルタリング（collaborative filtering:CF）[2] の 2 つの方法がある．その中でも特に，CB では被推薦ユーザの過去の購買履歴や評価履歴からユーザの嗜好を予測し，その嗜好に近いとされる特徴をもつ商品を推薦する手法である．

しかし，「ユーザの商品選考基準が，購入時ごとに大きく異なる可能性がある商品」あるいは「購入頻度が低く，購買履歴データが蓄積されない商品」のように，過去の購買履歴を前提とした従来手法の適用が難しい商品カテゴリも存在する．例えば，宿泊施設やパーティー会場の選考では，シチュエーションごとに目的（嗜好）が変わるため，過去の購買履歴からでは被推薦ユーザの購買時の嗜好の推定が困難であるという問題がある．一方，多くの EC サイトでは，注目商品情報を一時保存（WebClip）する機能が実装されている．WebClip された情報はユーザの現時点での興味を表しているため，購買意思決定時のユーザの嗜好を推定することに活用することができると考えられる．

また，WebClip からは嗜好情報が抽出可能であるものの，CB では抽出された嗜好と商品をマッチングさせるため，各商品の特性を適切に定量化することが重要であり，性能に大きな影響を与える．これに対し，テキストデータであるユーザレビューからユーザの商品に対する様々な意見など，定性的な評価（特徴）を得ることができる [3]．したがって，これらの情報を有効活用することにより，従来の推薦手法では困難であった商品カテゴリに対する推薦ができる可能性がある．

これらの点から，本研究では WebClip 情報，テキストデータを有効活用することで，従来の推薦手法では困難であった商品カテゴリに対する新たな推薦システムを与える．そのため，本研究では，WebClip を用いた商品選考時のユーザ嗜好の抽出と，ユーザレビューを基に商品の特徴を細分化し，ユーザの嗜好に近い商品を推薦する手法を提案する．抽出したユーザの嗜好と商品特性を用いることで，従来の推薦手法では推薦を行えないような商品に対する推薦を行う．また，代表的な EC サイトの一つである「じゃらん.net」[4] を事例としたユーザ実験を行い，本研究の有効性を示す．

2　従来手法と本研究への展開

推薦システムでは現状，購買履歴のみを用いた推薦が行われている．以下では，従来の推薦手法についての概要と共に，ユーザレビューと WebClip の推薦への適用方法を述べる．

2.1　推薦システム

推薦システムとは，ユーザの購買履歴を用いて，ユーザの嗜好を判断し，そのユーザ嗜好に適した商品を推薦するシステムである．推薦システムにおける推薦商品の予測手法として，協調フィルタリング（CF）と内容ベースフィルタリング（CB）の代表的な 2 つの手法がある．

CF では，購買履歴が類似したユーザ同士は，今後購買する商品もまた類似しているという仮定の下，被推薦ユーザとの類似ユーザが購買した商品情報を基に推薦する．ユーザ間の類似性を測るため，一般には相関係数などが用いられる．

一方，CB では，被推薦ユーザの購買履歴やアンケート情報を基にユーザの嗜好を予測し，その嗜好と特徴が類似した未購買商品を推薦する．商品の特徴とユーザの嗜好を基に推薦を行うため，他のユーザとの購買傾向の類似性を考慮した CF と比べて，よりユーザの嗜好に近い商品を推薦できるという利点がある．

本研究では，シチュエーションによりユーザの選考基準が大きく異なる商品や，定量的な情報では詳細がわからない商品を対象とした推薦を行うため，WebClip とユーザレビューを活用し，CB を用いた推薦システムについて述べる．

2.2　ユーザレビュー・WebClip 情報の推薦への適用

ユーザレビューとは，ユーザが購入・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．EC サイト上にはこれらの情報が多く投稿されており，ユーザは他のユーザの属性や，商品に与えた評価点と共にレビューを閲覧することが可能となる．他のユーザの体験談などを把握することが可能であり，購買意思決定時に大きな影響を与えるようになっている．

いま，あるレビューを $\boldsymbol{d}_i$ としたとき，ユーザによるレビュー集合を $\Delta = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_D\}$，$D$ を総レビュー数とする．また，本研究で推薦対象となる商品集合を $\mathcal{A} = \{A_1, A_2, \cdots, A_M\}$ で表す．ここで，M は商品の種類数を表す．Δ は，各商品アイテム m について書かれたレビュー集合 Δ_m に分割できる．ただし，$\Delta = \bigcup_m \Delta_m$ かつ $\Delta_m \cap \Delta_{m'} = \emptyset$ である．また，レビュー集合 Δ で使用されている単語の集合を $\Sigma = \{w_1, w_2, \cdots, w_J\}$ で表す．J は全ユーザレビューに出現する総異なり単語数を表す．さらに，レビュー $\boldsymbol{d}_i$ に含まれる単語集合 Σ の各要素 w_j の出現有無を v_{ij} を用いて，レビューベクトル $\boldsymbol{d}_i = (v_{i1}, v_{i2}, \cdots, v_{iJ})$ を定義する．但し v_{ij} は 0,1 の 2 値をとる要素であり，単語 w_j がレビュー $\boldsymbol{d}_i$ に出現する時に $v_{ij} = 1$ となり，出現しないときは $v_{ij} = 0$ と

なる．なお，以下ではユーザレビューを文単位に分割して扱うが，同様にレビューと表記する．$\boldsymbol{d}_i$ の要素はそれぞれ，商品の特徴を表していると考えられる．

一方で，WebClip 機能とは，EC サイトにおいてユーザが注目した商品を一時的に保存できる（Clip できる）機能である．購買行動を行うユーザをアクティブユーザと定義し，そのアクティブユーザが購買行動時に注目した商品情報を保存しておくことで，商品選択時にそれらの情報を比較することができる．WebClip 情報の取得により，ユーザの購買時の商品選定基準（嗜好）を得ることができる．

これらの情報を推薦に用いて，商品の特徴を詳細に推定し，アクティブユーザの嗜好と近い特徴を持つ商品を抽出し，推薦する．

3　提案手法

3.1　提案手法概要

ユーザの嗜好に合わせた新たな推薦システムに対する枠組みを与えるため，テキストデータ，WebClip 情報を有効活用した推薦システムについて提案する．本研究ではユーザに宿を推薦するため，

1. ユーザレビューを用いた商品情報（特徴）の定量化
2. WebClip の履歴を用いたユーザ嗜好の抽出

を行い，ユーザの嗜好に最も近い特徴を持つ商品を候補としてユーザに推薦する．なお，ユーザの嗜好は，ユーザが WebClip した商品の特徴を集約することで推定される．以下では，宿泊施設を例として本手法について説明を行うものとする．図 1 に推薦システムのイメージを示す．

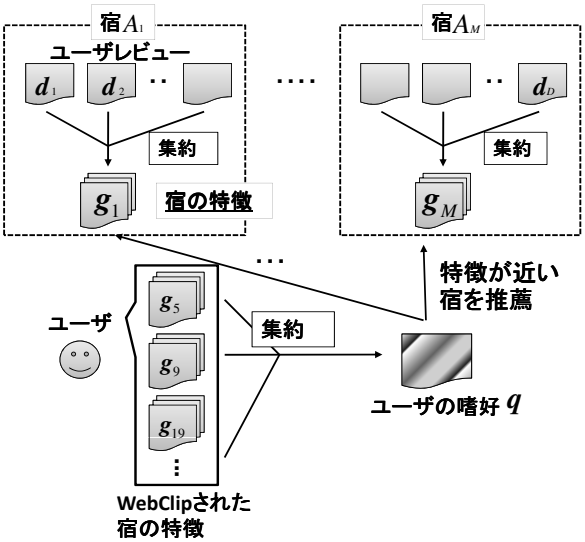

図 1．推薦システムのイメージ

一般に，CB では，ユーザプロファイルと商品ベクトルの定量化の方法が性能を大きく左右する．テキストデータを用いる場合，商品アイテム m について書かれたレビュー Δ_m 内の単語頻度を数え，1 つの J 次元ベクトルで表現する方法も考えられる．

しかし，宿泊施設に投稿されたユーザレビューには，一般に「食事」や「風呂」等，様々な内容について述べられている．このようにユーザレビューのコメント内容は多岐に渡るため，ユーザレビューのコメント対象を考慮せず，すべて単一の集合として扱うことは適切ではない可能性が高い．例えば,「食事」に関する単語が多く出現するレビューが多数を占めるとき,「風呂」に関係する単語が少数出現したとすると，宿の特徴とユーザの嗜好が非常に類似していても,「食事」に関する単語の影響が大きいため,「風呂」に関する単語の情報は過少評価され，推薦対象とならないことがある．

そこで本研究では，この単語の影響の度合いが適切に評価されない問題を回避するため，ユーザレビューを分類することで，レビューが示す特徴間の比較を細分化し，同項目ごとに比較を行う．そのことで，各項目ごとに嗜好の比較を行うことが可能となり，ユーザの嗜好をより反映させた商品の推薦を行うことができる．一般に EC サイトなどに蓄積されているユーザレビュー数は膨大なため，細分化作業には機械学習で用いられている文書分類の手法を用いる．それにより，ユーザレビューを細分化項目に分類し，分類された項目の情報をユーザレビューに付与することで，項目ごとに類似性を計算することができるようになる．

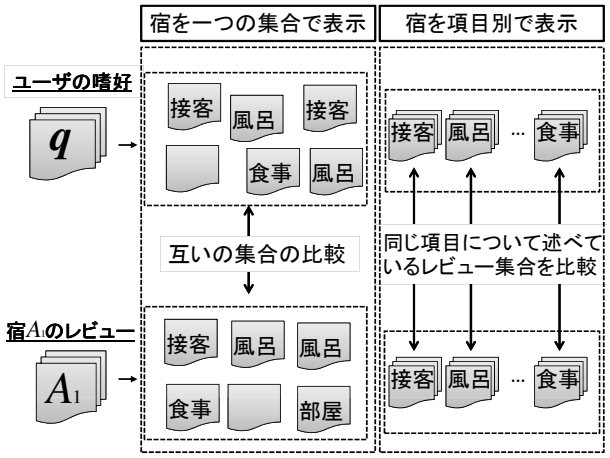

図 2．項目別にベクトルを比較するイメージ

これら WebClip 情報とユーザレビューから宿の特徴とユーザの嗜好の抽出を行う．さらに，ユーザレビューを項目別に分類するという，2 つの視点から提案のモデルを構成する．

3.2　モデルの構成

EC サイト上で宿に対する評価項目として活用されている「食事」「部屋」「風呂」「接客」「清潔感」「その他」などの項目を $\mathcal{C} = \{c_1, \cdots, c_k, \cdots, c_K\}$ で表し，これらを予め与えられたカテゴリとする．これらに対し，文書分類を行い，分類器によってレビューが述べている項目が c_k に分類されたとき，その文を $\boldsymbol{d}_i^k = (v_{i1}^k, v_{i2}^k, \cdots, v_{iJ}^k)$ と表す．このとき，宿の特徴は，投稿されているユーザレビューの平均をとり，以下のように表現する．

$$\boldsymbol{g}_m^k = \frac{1}{|\Delta_m^k|} \sum_{\boldsymbol{d}_i^k \in \Delta_m^k} \boldsymbol{d}_i^k, \tag{1}$$

$$= \frac{1}{|\Delta_m^k|} \sum_{\boldsymbol{d}_i \in \Delta_m^k} (v_{i1}^k, v_{i2}^k, \cdots, v_{iJ}^k), \tag{2}$$

$$= (g_{m1}^k, g_{m2}^k, \cdots, g_{mJ}^k). \tag{3}$$

ただし，$g^k_{mw} = \dfrac{1}{|\Delta^k_m|} \sum_{\boldsymbol{d}^k_m \in \Delta^k_m} v^k_{mj}$ とする．ここで Δ^k_m は宿 A_m の項目 k について述べられたレビュー集合であり，$|\Delta^k_m|$ はそのレビュー数を示す．$\boldsymbol{g}^k_m$ は宿の特徴ベクトルであり，宿 A_m の項目 k における特徴を表す．(2) から (4) 式における宿泊施設の特徴ベクトルの作成過程のイメージを図 3 に示す．

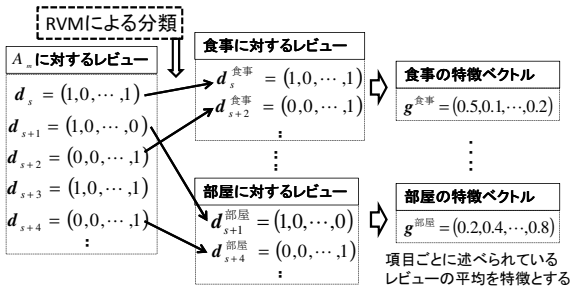

図 3． 特徴ベクトルの作成過程

ユーザの嗜好を表現するために，以下では嗜好ベクトルを定義する．嗜好ベクトルは WebClip した宿の特徴ベクトルを用いて作成する．今，アクティブユーザが U 件の宿を WebClip したとし，$\boldsymbol{g}^k_u$ はユーザが選んだ任意の宿 A_u の項目 k における特徴ベクトルと定義する．ここで，A_u の項目 k におけるレビュー集合を Δ^k_u とし，Δ^k_u に属する宿泊施設の特徴ベクトル $\boldsymbol{g}^k_u$ の平均をユーザの項目 k における嗜好ベクトル $\boldsymbol{q}^k$ とし，

$$\boldsymbol{q}^k = \frac{1}{U} \sum_{\boldsymbol{g}^k_u \in \Delta^k_u} \boldsymbol{g}^k_u, \tag{4}$$

$$= \frac{1}{U} \sum_{\boldsymbol{g}^k_u \in \Delta^k_u} (g^k_{u1}, g^k_{u2}, \cdots, g^k_{uJ}), \tag{5}$$

$$= (q^k_1, q^k_2, \cdots, q^k_J), \tag{6}$$

と定義する．ただし，$q^k_j = \dfrac{1}{U} \sum_{\boldsymbol{g}^k_u \in \Delta^k_u} g^k_{uj}$ とする．

3.3 類似度算出

提案手法では，宿特性とユーザ特性の近さを算出する方法として，類似度を測るための手法の一つである，相関係数法を利用する．

$$C^k_m = \frac{\sum_{j=1}^{J}(g^k_{mj} - \overline{g}^k_m)(q^k_j - \overline{q}^k)}{\sqrt{\sum_{j=1}^{J}(g^k_{mj} - \overline{g}^k_m)^2}\sqrt{\sum_{j=1}^{J}(q^k_j - \overline{q}^k)^2}}, \tag{7}$$

$$\bar{C}_m = \max_k C^k_m. \tag{8}$$

$$\widehat{A} = \arg\max_m \bar{C}_m. \tag{9}$$

ただし，$\overline{g}^k_m = 1/J \sum_{j=1}^{J} g^k_{mj}$，$\overline{q}^k = 1/J \sum_{j=1}^{J} q^k_j$ とする．C^k_m 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．アクティブユーザが WebClip に登録した宿の平均単語頻度と，各宿の全ユーザによる平均単語頻度の相関係数を示し，本研究では WebClip された宿以外で，$\bar{C}_m$ が高いものを上位 N 件推薦するものとする．

3.4 学習・予測アルゴリズム

提案手法は以下の手順で行い，推薦商品を予測する．

Step1) ユーザレビューを文単位に分割し，単語出現有無ベクトルに変換する

Step2) ユーザレビューの単語頻度ベクトルを基に RVM[5] を用いて分類し，レビューに項目を付与する

Step3) ユーザレビューを宿ごとに集約し，宿の特徴ベクトルを作成する

Step4) アクティブユーザが WebClip した宿の特徴ベクトルからユーザの嗜好ベクトルを作成する

Step5) 項目ごとに嗜好ベクトルと宿の特徴ベクトルの相関係数を計算する

Step6) 相関係数の値が大きい宿をユーザへ推薦する

上記のアルゴリズムでは，Step1) から Step3) までで宿の特徴ベクトルを作成し，Step4) から Step6) でアクティブユーザの嗜好を推定し，ユーザに推薦すべき宿を予測する．一般に文をベクトル表現する際には高次元・スパースなベクトルとなるが，本研究では文単位に分析を行うため，異なり単語数をその次元とするベクトル空間において，ベクトル表現した各ユーザレビューは，よりスパースなベクトルとなってしまうことが想定される．そこで，スパースなベクトルに対して高精度の分類が可能な，RVM を使用した．Step5) の推薦を行うイメージを図 4 で示す．

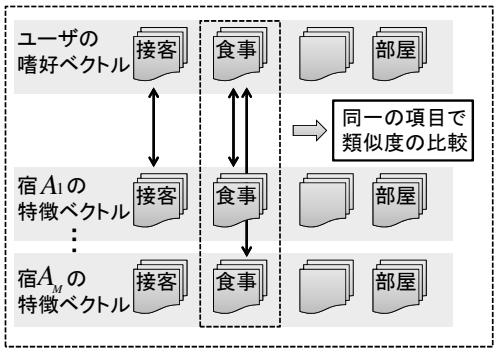

図 4． 相関係数による推薦

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

4 実験及び考察

4.1 実験条件

提案手法の有効性を示すため，宿泊予約サイト「じゃらん.net」[4] 内のユーザレビューを用いた実験を行った．分析対象は,「じゃらん.net」内から抽出した 15,098 件の宿泊施設に対する，合計 1,685,220 件のユーザレビューとする．このユーザレビューには，総合・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，推薦された宿に関する情報が有益かどうかを定性的に評価するため，各手法で抽出したそれぞれの宿の特徴を示す単語を比較する．さらに，20 人のユーザに実際に宿泊施設を選択する場面を想定し，その目的に則した 1〜5 件の WebClip をしてもらい，比較 1,

比較 2，提案手法を WebClip 情報を用いて，ユーザに推薦した宿を「目的と合致しているか？」,「宿に対して好感がもてるか？」の 2 点について 10 段階で評価を行った．一般に，EC サイト上に掲載されている宿泊施設には各項目に対する 1〜5 点の評価点（ユーザの採点の平均）が与えられている．以下では，評価点を用いた手法と，ユーザレビューを用いて作成した宿の特徴の 2 つを用いて実験を行う．

4.2 実験結果

実験では，(1) 評価点のみを用いて推薦を行う方法（比較手法 1），(2) 宿の特徴ベクトルをそのまま用いる方法（比較手法 2）[6]，(3) レビュー項目別宿の特徴ベクトルを利用した方法（提案手法），の 3 パターンに対する推薦における推薦結果を示す．

4.2.1 定性的実験

あるユーザが「貸切風呂」「食事にこだわり有」という条件の下，3 件の宿を WebClip したとする．WebClip した宿のユーザレビューに特化して出現する単語を表 1 へ，各手法を用いて推薦された宿泊施設のレビューに特化して出現し，宿の特徴を示していると考えられる単語を表 2〜表 4 に示す．

表 1． WebClip した宿の特徴

宿泊施設 A 1	宿泊施設 A 2	宿泊施設 A 3
貸切	露天風呂	貸切
焼きたて	貸切	和洋室
子供	ペット	魚料理

表 2． 比較手法 1 の推薦した宿の特徴

宿泊施設 B 1	宿泊施設 B 2	宿泊施設 B 3
尾瀬	家族湯	掃除
ボリューム	柴犬	味
天ぷら	ボリューム	露天風呂

評価点のみを利用した従来手法では「貸切温泉」などのニーズを掴むことができなかった．

表 3． 比較手法 2 の推薦した宿の特徴

宿泊施設 C 1	宿泊施設 C 2	宿泊施設 C 3
露天風呂	家族風呂	貸切
貸切	子供	自然
清潔	部屋食	刺身

評価点のみを利用した比較手法 1 に比べて，単語頻度ベクトルを利用することで,「貸切温泉」というユーザの嗜好を考慮した推薦を行うことができている．

表 4． 提案手法の推薦した宿の特徴

宿泊施設 D 1	宿泊施設 D 2	宿泊施設 D 3
家族風呂	ペット	貸切
子供	焼きたて	自然
部屋食	貸切	刺身

提案手法では,「ペット」等の，レビュー全体においては，あまり出現頻度の高くない特徴語を得られた．

4.2.2 ユーザ実験

ユーザ実験の結果を表 5 に示す．

表 5． ユーザによる推薦結果に対する評価

評価項目／手法	比較 1	比較 2	提案
目的との一致度	3.82	5.76	7.18
宿に対する好感度	5.18	6.31	7.22

表 5 はユーザへのアンケート結果の平均となる．ユーザの「目的との一致度」「宿に対する好感度」の双方の視点において，提案手法の結果が比較手法に比べて良い結果が得られた．

4.3 考察

4.3.1 定性実験考察

提案手法ではテキスト情報の活用により，比較手法 1 と比較して「貸切温泉」などのニーズを掴むことが可能となった．またユーザが意図的に選択していない「子供」や「ペット」といった特徴の抽出もできていた．さらに，提案手法では項目別に分類を行うことで,「焼きたて」等の，特定の項目に関する特徴の抽出を行うことができた．しかし，単語を単独で利用しているため，単語間の関係性を考慮できず，否定語などに対応できていない．その改善として，単語間の関係性を示す係り受けを考慮するなどが考えられる．

4.3.2 ユーザ実験考察

ユーザ実験のアンケート項目である「目的との一致度」において，提案手法の結果が比較手法に比べて特に良い結果が得られた．このことから，定性的な情報であるテキスト情報を活用することで，ユーザの宿泊施設の選択における目的を抽出することができ，本研究で対象とした宿泊施設などの商品の推薦に適した推薦手法であると考えられる．

5 結論及び今後の課題

本研究ではユーザレビューと WebClip 情報を活用し，従来手法が想定していない商品カテゴリに対する推薦方法を提案した. また，宿泊施設を対象とした実験を行うことで提案手法の有効性を確認した．

本研究で対象とした宿泊施設の数，並びにレビュー数は非常に多く，推薦結果を算出するまでに計算量がかかるという問題点がある．今後の課題として，この削減が考えられる．

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, 2005.

[2] 神嶌敏弘, “推薦システムのアルゴリズム（2）,” 人工知能学会誌 ,23(1), pp. 89–103, 2008.

[3] 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レビュー分析に関する一考察,” 日本経営工学会平成 *22* 年度秋季研究大会予稿集, pp. 192–193, 2010.

[4] じゃらん.net : http://www.jaran.net/

[5] M.E.Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” *Journal of Machine Learning Reserch*, pp.211–244, 2001.

[6] 榮枝隼人, 三川健太, 後藤正幸, “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,” 第 *10* 回情報科学技術フォーラム, pp. 451–454, 2011.