DEIM Forum 2019 F3-5

飲食店レビューにおける再訪問ユーザの行動の分析

白髪　宙海†　　村本　直樹†　　高橋　克郎†　　大島　裕明†

† 兵庫県立大学 応用情報科学研究科　〒 650–0047 兵庫県神戸市中央区港島南町 7–1–28
E-mail: †{aa17j506,aa18c508,ab18y501,ohshima}@ai.u-hyogo.ac.jp

あらまし　本研究では，飲食店レビューサイトにおいてユーザが投稿したレビュー文書に注目し，初訪問なのか再訪問なのかを分類する問題に取り組む．Web 上には商品やサービスを利用したユーザが情報を共有するレビューサイトがある．なかでも飲食店レビューサイトは広く一般に利用されており，味，価格，サービス，レビュー文書などによる個別の飲食店のレビューが共有されている．テキストで記述されたレビュー文書を精査すると，初訪問のものと再訪問のものが混在していることが分かる．これらを自動的に分類することによって，飲食店の再訪問率の推定や，始めて訪れたユーザと，何度も訪れるユーザによる評価の違いの分析などが可能になる．

キーワード　再訪問の推定，レビュー分析，テキストマイニング

1. はじめに

本研究では，飲食店レビューサイトにおいてユーザが投稿したレビュー文書が，初訪問におけるものなのか再訪問におけるものなのかを分類する手法を提案する．Web 上では商品やサービスを利用したユーザや，利用を検討しているユーザが情報を共有するレビューサイトがある．そのなかでも飲食店のレビューが得られるサイトとしては複数のサイトがあり，人々に広く利用されている．それぞれのサイトでは店名，食事のジャンル，営業時間，定休日，メニュー，住所（地図）の情報が得られるほか，実際に飲食店を利用したユーザが投稿したレビューをみることができる．ユーザのレビューには，数値で表現される味・価格，接客・サービス，雰囲気，コストパフォーマンスなどの評価がある．さらに，飲食店を利用した感想などを文書で書いたレビュー文書が存在する．飲食店は，このように多数の項目で評価されている．飲食店レビューサイトには様々な評価指標がある一方で，人々が飲食店を評価する観点はこれらに限定されるものではない．たとえば，初めてでも入りやすい飲食店なのかどうかや，常連客に愛されている飲食店なのかどうかなどは，飲食店の評価として興味深い観点であると考えられる．そのような評価を行うためには，あるレビューが初訪問におけるものなのか再訪問におけるものなのかということが分析の軸として必要になると考えられる．

たとえば，飲食店レビューサイトのひとつである食べログでは利用した飲食店に対して，複数回レビューを投稿することが可能となっている．投稿されたレビューにはユーザ情報のほかにレビューの投稿回数や味やサービスに対するスコア，利用した時間帯，価格帯などのスコアを見ることができる．図 1 の例では，投稿回数の数値を見ることで 2 回はその飲食店を利用したことが確認できる．しかし，投稿回数が 1 回のレビューの情報は本当に初訪問のものであるかは実際に読まなければ判断できない．図 2 におけるレビュー文書を読むと，「以前から喫茶店として利用していましたが，今回初めてハンバーグを食べに行ってきました.」と記述されている．実際に投稿されたレビューの文書を読むことで，過去にもその飲食店を利用していると推測できる．このように文書から初訪問または再訪問であるかを特徴付ける特徴量を取得することにより，自動分類が実現できると考えられる．

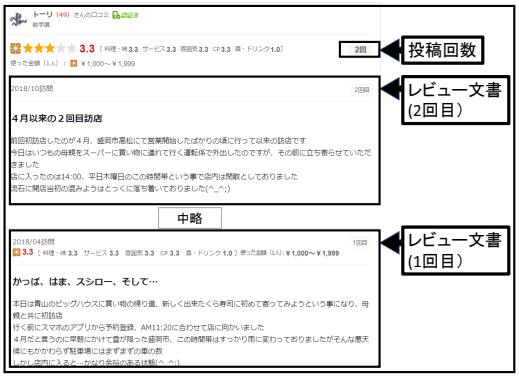

図 1　飲食店レビューサイトのページの例

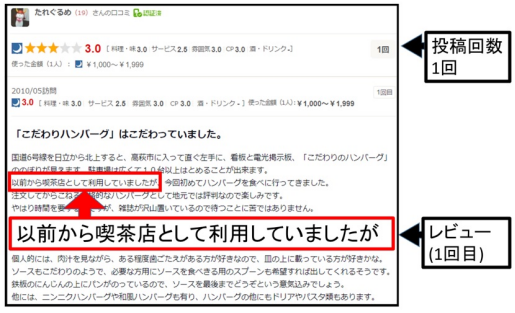

図 2　投稿回数 1 回だが再訪問であると推測できる例

本研究ではレビュー文書の特徴ベクトルを基本的な特徴量として用いながら，初訪問や再訪問のレビューで現れやすい表現や語などに着目した特徴量と，文書ベクトルで表現される特徴量を用いた評価を行い，それぞれの特徴量が分類に貢献していることを明らかにした．

2. 関 連 研 究

Web 上には，飲食店やそのほかの様々な商品に対するレビューが存在している．それらを多様な観点から分析する研究が数多く行われている．

Dave ら [1] は，Amazon などスコア付けされた商品レビューを使い，商品への意見が「ネガティブ・ノーマル・ポジティブ」のものであるかを出力する研究を行っている．Ding ら [2] は，Amazon の商品レビューで使われる文脈に注目し，そのレビューが商品に対してポジティブな意見かネガティブな意見かを判定する研究を行っている．Osman ら [3] は，ブログの文書から意見を抽出する研究において，抽出する精度の高いコーパスを発見する比較と検証の研究を行っている．Wijaya ら [4] は，Web 上の映画のレビューを使い，使われている形容詞の規則性でポジティブ・ネガティブのスコア付けをし，レビューから新たな映画のランキングを作る手法を提案した．Gilbert ら [5] は，Amazon の商品レビューでは先に投稿したレビューを参照し投稿する追記型のレビューが多いことに注目し，追記型のレビューを検出するために似た文書をまとめる研究を行っている．Fayazi ら [6] は, Amazon の商品レビューであからさまに高得点をつける欺瞞的なレビューを分類する研究を行っている．Mullick ら [7] は，新聞や Twitter や YouTube で投稿される文書の内容が「意見」「事実」のどちらかを分類する二値分類問題の研究を行っている．Cheng ら [8] は，Web 上の記事やスパムメールなどのリークされたデータセットを用いフェイクニュースを発見する研究を行っている．Zagal ら [9] は，ゲームのレビューサイトを利用し，どのようにゲームを楽しんだのかやユーザからの開発者への提案などのトピックを分類する研究を行っている．Mahony ら [10] は，TripAdvisor で書かれたレビューから，ホテルの選択に参考となったレビューだけを抽出する分類ベースの推奨システムを研究している．Lappas ら [11] は，特定のレビューを扱う上で新たなコーパスを作成しなければいけない課題に対し，レビューに使える最適なコーパスを数値化する研究をしている．Vu ら [12] は，モバイルアプリのレビューにおけるフレーズに注目しユーザの意見を抽出する研究を行っている．

3. 再訪問レビューと初訪問レビューの分類

本研究で取り組む問題は，テキスト文書の三値分類問題である．分類するテキスト文書は飲食店レビューサイト上のレビューである．このレビュー文書を入力として受け取り，初訪問であるか再訪問であるか不明であるかを出力する．

飲食店レビューサイトの食べログでは，同一の飲食店に対し複数回のレビューの投稿が可能である．レビューは図 1 のように投稿した投稿回数と，それぞれのレビューに対して何回目に投稿されたものかが表示される．本研究では投稿されたそれぞれのレビューの文書を**レビュー文書**とよぶ．図 1 のレビューであれば，レビューの中に 1 回目に投稿されたレビュー文書と，2 回目に投稿されたレビュー文書があることとなる．これらのレビュー文書において，初訪問時に投稿されているものを**初訪問レビュー**とし，再訪問時に投稿されているものを**再訪問レビュー**とする．なお，2 回目以降に投稿されたレビュー文書はすべて再訪問レビューとなるが，本研究では 1 回目に投稿されたレビュー文書のみを分類の対象とし，分類器を構築する．そして，分類器の精度を評価する．なお，分類器には三値分類に適した分類器を探すテストをおこないこれを決定する．実験で作成される分類器は，**レビュー文書**を入力として受け取り，初訪問であるか再訪問であるか不明であるかを出力する．評価には F 値を用いた．

4. レビュー文書を用いた特徴ベクトルの生成

三値分類器を構築するためには、レビュー文書から特徴ベクトルを作成する必要がある．特徴ベクトルは，以下の 3 種類の素性から作成する．

（1） TF-IDF に基づく特徴量
（2） 再訪問や初訪問を表す語などに基づく特徴量
（3） 根拠文との類似性に基づく特徴量

1 つ目の素性は TF-IDF に基づく特徴量である．TF-IDF とは平均的な文の単語分布からの乖離を KL 情報量で測り，語それぞれの貢献度をベクトルで表したものである．この特徴量は分類問題に比較的にもちいられるため，本研究の評価の軸として機能することを目的に，この特徴量を得た．

2 つ目の素性は再訪問や初訪問を表す語などに基づく特徴量である．再訪問レビューや，初訪問レビューには，初訪問や再訪問を特徴付ける語や言い回しが含まれている可能性がある．そこで，それらのように判断する根拠となった文を分析することにより，再訪問レビューや，初訪問レビューであらわれる語や語の使われ方による特徴量を得た．

3 つ目の素性は再訪問レビューや，初訪問レビューと判断した文には類似性があることに注目した特徴量である．しかし，文の類似性だけに着目した場合は, 否定語などが入ることで文意が異なる可能性がある．本研究では，根拠とした文のクラスタリングをおこない，抽象的な表現にすることで, これらの不都合な点の緩和を試みた特徴量である．

これら 3 つの特徴量を作成するため，実際にレビュー文書を読み，分析をおこなった．人がレビュー文書を読み，再訪問レビューまたは初訪問レビューまたは不明レビューかを判断するとき，そのレビュー文書中に判断の根拠となる文が存在することが多い．この文には初訪問や再訪問を特徴付ける語や言い回しが含まれている可能性がある．そこで，この文を分析することによって，そのような語などを発見することを試みた．本研究で使う言葉の説明とともに，根拠となる 1 文の例を以下にあげる．なお，本研究での 1 文は「。」「！」「改行」で区切った文とする．

> 昭和の雰囲気です～．創業５１年!! 凄いです笑．昭和の映画に出てくるような定食屋さんでございます．**会社から車で数分なもんでたまに行かせていただいてます**．今日は豚汁定食８５０円也！豚汁の旨さにはビックリです．コスパは高めかなぁ～って思いますが味が

美味しいので満足です！

この例では，**「会社から車で数分なもんでたまに行かせていただいてます.」**という文より，すでに訪問したことがあることがわかり，このレビュー文書が再訪問レビューであることがわかる．このように 1 文だけで初訪問か再訪問かを示せるものを**根拠文**とする．

根拠文には再訪問や初訪問を特徴づける語や言い回しなど，特有の表現が含まれている可能性がある．本研究ではこれらの特有の表現を**特有表現**と呼ぶ. 特有表現に注目して作成する特徴量については 4.2 節で述べる．

4.1 TF-IDF に基づく特徴量

レビュー文書を Bag-of-words とみなし，TF-IDF を重みづけとして特徴ベクトルを作成する．TF-IDF を用いるために，レビュー文書を形態素解析し，すべての語を基本形にして分かち書きを行う．形態素解析器には MeCab を用いる．ここで，DF を求めるための文書集合は 5.1 で説明するラベル付けされたレビュー文書の 2425 件である．TF-IDF の実装は，scikit-learn の TfidfVectorizer を用いる．パラメータとしては，全文書中で 1 つの文書にしか現れない語を無視する，min_df=2 を用いる．語の総種類数は $TF = 11,889$ であった．次に，TF-IDF で表現された特徴ベクトルを LSA [13] に相当する手法で 300 次元に圧縮した．このようにし, 文書から TF-IDF に基づく 300 次元の特徴量が得られた．

4.2 特有表現に基づく特徴量

レビュー文書を初訪問または再訪問であるかと判断した根拠文には，それらを判断するにいたった特徴的な語や言い回しが含まれている可能性がある．本研究ではこれらを**特有表現**と呼ぶこととする．そこで根拠文を分析することによって，そのような語や語の使われ方などを発見することを試みた．この根拠文の分析には，本実験で使用するデータとは異なる，618 件のレビュー文書より取得した根拠文を用い特徴ベクトルを作成する．ベクトルを作成する流れは以下のようになる，

（1） 根拠文を観察し特徴的な語や表現を人手で探す

（2） 取得された特徴的な語などのそれぞれに対して，一つの素性を得る

まず, 根拠文より取得された特有表現が，レビュー文書内に存在するか存在しないかの 0/1 で特徴ベクトルを作成する．つまり，用意する特有表現の個数が n 個のときに，0/1 で表される特徴ベクトルが n 次元作成される．

ここで，特有表現を取得する手法について述べる．まず，根拠文を形態素化し，語を基本形にする．そのようにして得られた語について，再訪問レビューないしは初訪問レビューのいずれかの根拠文にのみよく現れる語を観察し，特徴的な語を取得した．このようにして取得した再訪問レビューにおける特有表現を表 1 に，初訪問レビューにおける特有表現を表 2 にまとめる．

「初訪」「再訪」「ログ」といった表現は文中にその語が含まれているだけで特徴的な表現と考えられる．たとえば,「初訪」は,「今回が初訪でした。」や「ランチタイムに初訪問です。」といった表現である．これらの表現は初めてその飲食店を利用したと考えられる．

表 1　再訪問レビューにおける特有表現の例

特有表現（基本形）	例文
再訪 (再訪問)	12 月**再訪**.
久しぶり	2015 年 5 月**久しぶり**に行きました。
年ぶり	約 1 **年ぶり**となるランチ訪問です。
回目	2 **回目**の訪問です！
回数	訪問**回数** 3 回
毎回	ここでは**毎回** 750 円のミックスフライ定食。
よく利用	私は**よく利用**しています。
前回は	**前回**は家族できました。
相変わらず	**相変わらず**何を食べても美味

表 2　初訪問レビューにおける特有表現の例

特有表現（基本形）	例文
初訪（初訪問）	今回が**初訪**でした.
ログ	ブログや食べ**ログ**などを見て来店.
入る	**入る**きっかけがありませんでしたが...
近く	**近く**にあるこのお店に来ました.
見つける	また素敵なお店を**見つける**ことが出来ました。
思った	**思った**以上に具沢山♪
たまたま	**たまたま**物色中に発見しました。

また，同じように,「ログ」という語がある，これらの例は「その時過去ブログを思い出した。そうりの食卓！」「食べログの評価を見て決めただけに益々期待が高まります。」「神戸のランチブログで、こちらの『冷やしきつねうどん』を知って、気になっていました。」といったような記述である．これらの文は，他のレビューを参照し飲食店を決定したことを示しており，そのようにして訪れるのはたいていの場合が初訪問であると考えられる．そこで,「初訪」「ログ」という語が出現する場合の素性を作成する．

「再訪」は，初訪問レビューにおいても存在することがある．例えば,「是非再訪したいお店です」といったような記述は，どちらかというと初訪問を示す特徴と考えられる．一方で,「【再訪】」や「再訪。」のように,「再訪」という語が括弧でくくられていたり，体言止めで現れたりする場合には，再訪問のレビューであることが多いと考えられる．そこで,「再訪」が直後に日本語にはない文字をともなって出現する場合に 1 となり，そうでない場合には 0 となる素性を作成した．初訪問や再訪問を表す語や表現に基づき，16 次元の特徴量が得られた．

4.3 語や表現に基づく特徴量

1 文だけで初訪問か再訪問かを示せる**根拠文**には不都合が存在する．それは，形式上よく似た文であっても意味が異なる場合である．その不都合の例を以下にあげる．

形容詞による対照的な表現

文中の単語がほとんど同じでも，評価する形容詞などが対照的な表現の場合がある．対照的な形容詞は「美味しい（おいしい)」ならば形容詞「まずい」である．たとえば「値段のわりに美味しい」と「値段のわりにまずい」は，ともに 5 個の単語で構成されており，最後の形容詞のみが異なる．前者は再訪問レ

表 3　根拠文におけるタイプ例

タイプ	例文
直接記入	12 月 初訪問 再訪問です
紹介・発見	食べログを見て来たいと思っていました 友人の紹介で知りました
常連	いつ食べても飽きません いつも美味しくて満足です

ビューで気に入った飲食店の評価の表現で観察でき，後者は初訪問レビューで飲食店の批判をするときに観察できた．

肯定語と否定語で文意が異なる表現

文が肯定語だけで構成される場合と，文に一語だけ否定語が含まれている場合に逆の特徴となってしまう．たとえば，肯定語の「ある」と否定語の「ない」ならば,「行ったことがある」と「行ったことがない」である．この 2 文は 5 分の 4 の割合で単語が同じであるが根拠文として逆の素性になってしまう．このような不都合は，人が分類をした場合は生じない．しかし，文の構成に着目した数値表現での類似度を分類基準として使う場合，文意をくみ取らずに分類される可能性がある．この問題点を緩和するためにより抽象化した根拠文の表現から得られる素性を成分とする特徴ベクトルを作成する．以下に特徴ベクトル作成の過程を示す．

（1） ラベルの付いたレビュー文書をベクトル化
（2） 根拠文ごとに K-means 法を用いてクラスタリング
（3） 分類するレビュー文書の各文をベクトル化
（4） 各文と各クラスタとのユークリッド距離を計算
（5） 距離が最も近いクラスタを特徴ベクトルとして出力

はじめに，レビュー文書中の各文を TfidfVectorizer を用いてそれぞれベクトル化する．同様に，再訪問レビューと初訪問レビューにおけるそれぞれの根拠文をベクトル化する．これを根拠文ベクトルと呼ぶ．根拠文はある程度タイプ分けが可能であると考えられる．根拠文のタイプ例を表 3 に示す．したがって，根拠文のクラスタリングが可能であると考えられる．表 3 のように，根拠文にはいくつかのタイプが存在するという仮定のもと，K-means 法を用いてクラスタリングする．根拠文ベクトルを再訪問レビューと初訪問レビューそれぞれ 15 ずつのクラスタにクラスタリングする．得られた各クラスタの重心を，各クラスタの代表ベクトルとする．したがって，30 個の根拠文の代表ベクトルが作成される．

この代表ベクトルを用いて，分類対象のレビュー文書に対する特徴ベクトルを作成する．この特徴ベクトルはレビュー文書の 1 成分と 1 代表ベクトルの間の特徴に対応するような 30 次元のベクトルである．この特徴ベクトルはレビュー文書の 1 文と根拠文の間のユークリッド距離を求めた類似度ベクトルを，一定のしきい値で二値化したものである．特徴ベクトルの作成手法の詳細を以下に記す．

とある未知のレビュー文書 R が与えられた場合，まず，その文書を文のリストとみなし，$R = \{r_1, r_2, \cdots, r_n\}$ と表す．ここで，n はレビュー文書 R における文の数を示す．r_j は R における j 番目の文を表す．v_j を文 r_j を TfidfVectolizer で表現したベクトルとし，レビュー文書ベクトルと呼ぶ．次に，根拠文各クラスタの代表ベクトルのリストを $E = \{e_1, e_2, \ldots, e_{30}\}$ と表す．$v(e_i)$ は E の i 番目の根拠文クラスタの代表ベクトルを表す．レビュー文書 R の i 番目の文と，根拠文クラスタの代表ベクトル E の i 番目をそれぞれ，$x = r_i$，$y = e_i$ とする．この x と y のユークリッド距離 $d(x, y)$ を

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} \qquad (1)$$

とする．すなわち，とある根拠文クラスタの代表ベクトル $v(e_i)$ と，とあるレビュー文書の 1 文のベクトル $r_j (1 \leqq j \leqq n)$ とのユークリッド距離を計算したベクトルを，類似度ベクトルとよぶ．さらに，このようにして得られた類似度ベクトルを以下のように二値化して特徴ベクトルを作成する．本実験では 30 個の成分のうち，しきい値以下を 1 とし，それ以外を 0 とする．すなわち，

$$s_i = \begin{cases} 1 & (d_i \leqq threshold) \\ 0 & (otherwise) \end{cases} \qquad (2)$$

である．この二値化された距離ベクトルを特徴ベクトルと呼ぶ．特徴ベクトルの次元は類似度ベクトルと同じ 30 次元である．以上の根拠文に基づく特徴ベクトルの作成の流れについて，まとめたものを図 3 に示す．なお，しきい値については 5.2 節で述べる．

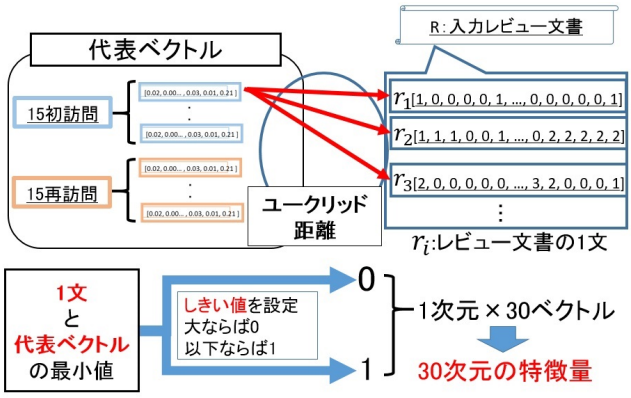

図 3　とある 1 文と代表ベクトルとの類似度による特徴量

5. 評価実験

本実験で使用するデータは食べログから収集したレビュー文書である．このレビュー文書に対しラベル付けをおこないデータセットを作成する．このデータセットを用いて分類機を構築し精度を評価する．本章は以下のながれで説明する．

- データセットの作成について
- 三値分類のための訓練データのクラス分け
- 評価実験

5.1 データセット

本研究で取り組む問題は，テキスト文書の三値分類問題である．作成される分類器は，レビュー文書を入力として受け取り，

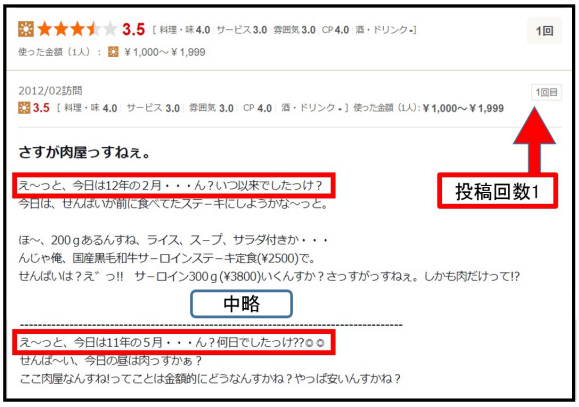

図 4　同一レビュー文書内で追記されている例

初訪問であるか再訪問であるか不明であるかを出力する．この分類器の構築にはレビュー文書をベクトルとして表現する必要がある．本研究でははじめに，データセットとして 2538 件の食べログのレビュー文書を収集して分析を行った．そして，本研究に適さない 113 件のレビューを除く 2425 件のレビュー文書を対象とした．そのデータの収集の方法と，使用したレビュー文書について説明する．

食べログ上で各飲食店に対し，あるユーザが投稿したレビュー文書のうち，1 回目に投稿されたレビュー文書を対象として収集する．つまり同一ユーザが同一店舗に 2 回以上投稿している場合は，2 回目以降のレビュー文書は対象としない．

ここで，収集する際に地域と飲食店のジャンルで条件を指定する．地域は 47 都道府県として，食べログでクラス分けされている 6 つのジャンルを選択する．選択した地域と飲食店のジャンルは表 4 の通りである．

つまり 47 都道府県において 6 つのジャンルの飲食店を使用するため，地域とジャンルの組み合わせは全 282 通りとなる．そして，それぞれの組み合わせで飲食店を 3 店舗ずつ，計 846 店舗を選択する．次に，選択した 846 店舗の飲食店ごとに，投稿されたレビュー文書から，ランダムに 3 名のユーザを選択し，前章の図 2 のような，1 回目に投稿されたレビュー文書を収集する．これにより，846 店舗に対しそれぞれ 3 件のレビュー文書の総計 2,538 件のレビュー文書が収集された．

ここで，1 つのレビュー文書内で複数回にわたり追記しているレビュー文書は，純粋な初訪問とはみなせないため対象外とした．例を図 4 に示す．追記部分については下記のようになっている．

> 「え～っと，今日は 12 年の 2 月 … ん？いつ以来でしたっけ？(中略) え～っと，今日は 11 年の 5 月 … ん？何日でしたっけ??(後略)」

このように，新たなレビュー文書を作成せずに，以前に作成したレビュー文書に対して追記しているものは，103 件存在した．これらは初訪問と再訪問が混在しているため使用しない．

また，図 5 のように一回目の投稿が画像のみの場合や，図 6 のようにスコアのみの場合や，投稿を取り下げている場合など，

表 4　レビュー文書を収集する地域とジャンル

地域	47 都道府県		
ジャンル	フレンチ	寿司	ステーキ・ハンバーグ
	うどん	定食・食堂	パン・サンドウィッチ

図 5　画像のみでレビュー文書が存在しない例

図 6　評価欄のみでレビュー文書が存在しない例

表 5　ラベル付け前に除くレビュー文書の数

除く項目	該当数
追記型	103
画像のみ	4
評価欄のみ	6

表 6　評価者ごとのラベル付けの結果

ラベル	1	2	3	4	5
評価者 X	1250	343	278	56	498
評価者 Y	1220	311	444	54	396
評価者 Z	1288	381	197	67	492

評価欄のみで，レビュー文書が存在しないもの 10 件を除く．除くレビュー文書の総数は表 5 に示すとおり 113 件となり，これらのレビュー文書を除くと 2425 件となった．この 2425 件のレビュー文書に対しラベル付けを行う．

評価者 3 名にて，レビュー文書が，初訪問におけるものか，再訪問におけるものかのラベル付けを行った．ラベル付けは「確実に初訪問」「初訪問だろうと思われる」「不明」「再訪問だろうと思われる」「確実に再訪問」の 5 段階とした．

評価者ごとのラベル付けの結果を表 6 に示す．また，ラベル付けと同時に根拠文の収集を行った．ここで収集する根拠文は評価者が，ラベル「1」「5」を選択したときに，レビュー文書内において出現する根拠文のうち，一番最初に出現した根拠文とする．5 段階評価の流れは以下になる．

（1） レビュー文書を読み初訪問，再訪問の推測をする
（2） 初訪問，再訪問と判断がつかない場合は「3」とする

表 7　根拠文の例

初訪問	たまたま通りがかりに見つけて寄ってみました． こちらのお店がオープンしてからずっと行きたいなぁと思っていました． どこか行きたくて色々検索してたら，良さそうな所で空きがあったので． 〜〜実店舗初訪問 (2015．3)〜〜
再訪問	小さい頃から、祖父母に連れられてよく来ていたお店。 学生時代の一時期，道後に滞在したことがあり，よく行っていました． 久しぶりに，ここ弥太郎寿司に行ってきました． 9 月に 2 回目の訪問．

表 8　5 段階のラベルから三値した結果

初訪問	不明	再訪問
1249	739	437

（3） 初訪問，再訪問と推測した場合はその根拠文をさがす

（4） 根拠文がある場合，初訪問と思われるならば「1」，再訪問と思われるならば「5」とする

（5） 根拠文がない場合，初訪問と思われるならば「2」，再訪問と思われるならば「4」とする

このラベル付けにより集められた根拠文の一例を表 7 に示す．

また，評価者 3 名によるラベル付けの結果を評価するため，

- 「1」「2」のラベルを初訪問
- 「3」のラベルを不明
- 「4」「5」のラベルを再訪問

とし，これらの一致度を計るために，Fleise [14] の κ 係数を導出した．κ 係数の値は 0.58 となり，評価者 3 名のラベル付けの結果は，中程度の一致であった．

次に三値分類に用いるデータについて述べる．実験で使用するデータは文書の無いレビューを除く 2425 件のレビュー文書である．これらのレビュー文書を以下のルールに基づき 5 段階のラベルから三値に変換した．そのレビュー文書の数は表 8 になる．また，三値に変換する条件は，

- 5 段階のラベル付けにて初訪問「1」で 2 名以上のラベルが一致したレビュー文書を「初訪問（-1)」
- 5 段階のラベル付けにて再訪問「5」で 2 名以上のラベルが一致したレビュー文書を「再訪問（+1)」
- それ以外のものを「不明（0)」

とした．

5.2 評 価 実 験

本研究の目的は三値に分類する分類器を構築することである．その前段階としてレビュー文書の三値分類に適した分類器を決定するための実験を行った．次に実験の結果をうけ，スコアのよい分類器を用い本研究で作成した素性の評価実験を行った．

全素性を用いた場合の分類器のスコア

本研究で提案する特徴量の検証に用いる分類器を選択するために，

- Naive Bayes
- 決定木
- ロジスティック
- SVM（サポートベクターマシン）
- KNN（k 近傍法）
- Ada Boost
- RandomForest

の 7 種の分類器に対し交差検証を行った．

テストは，交差検証を 5 分割でおこない，評価スコアは F 値とした．分類器の構築時には，毎回グリッドサーチをおこない，*accuracy* を最大にするようなパラメータの決定をおこなった．グリッドサーチを行う際の各種パラメータは分類器ごとに設定した．パラメータは以下のとおりである．

$$\text{kernel} = \{\text{linear}, \text{rbf}\} \quad (3)$$

$$\gamma(\text{svm}) = \{0.1, 0.01, 0.001\} \quad (4)$$

$$\text{C(svm)} = \{1, 10, 100\} \quad (5)$$

$$\text{C(LogisticRegression)} = \{0.01, 0.1, 1, 10, 100, 1000\} \quad (6)$$

$$\text{KNeighbors} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \quad (7)$$

カーネルは線形と RBF カーネルを使用した．線形、RBF カーネルともに機械学習における正則化の強さを表す定数 C を上式のとおり．RBF カーネルの訓練データの影響範囲の半径の逆数を γ の値は上式のとおりである．またテストをするにあたり，初訪問と再訪問と不明のデータに偏りが生じているため，アンダーサンプリングのある場合と，アンダーサンプリングの無い場合のそれぞれでテストをする．7 つの分類器ごとに評価に用いる F 値が高いものを表 11 に示す．

テストでは，はじめに 4.3 節の (2) 式におけるユークリッド距離のしきい値 d の決定を行い，7 つの分類器の精度の比較をし，実験に使用する分類器の決定を行った．比較に使う d の値は 0.6 から 2.0 まで 0.1 刻みで確認し，結果 1.8 のスコアがもっとも良かったためこれを用いた．この分類器ごとのスコアは表 11 になる．この結果より，7 つの分類器の中で最もスコアが高かった SVM とロジスティック回帰を用い，本研究の提案する 3 素性の検証をする．

5.3 SVM とロジスティック回帰と 3 素性の組合せの検証

SVM とロジスティック回帰を用い，以下の素性の組み合わせについて分類器を構築する．

素性 1　TF-IDF に基づく特徴量

素性 2　初訪問や再訪問を表す語などに基づく特徴量

素性 3　根拠文とのユークリッド距離に基づく特徴量

それぞれの分類精度について比較を行い，各素性がどの程度分類に貢献しているかを確認した．SVM とロジスティック回帰による素性の組合せの結果を表 9 と表 10 に示す．

表 9　SVM に素性を組み込んだ場合の F 値の結果

SVM						
	アンダーサンプリング無し			アンサーサンプリング有り		
本研究の素性の組合せ	適合率	再現率	F1	適合率	再現率	F1
(1)：TF-IDF に基づく特徴量	0.5039	0.5562	0.4882	0.5337	0.4989	0.5076
(2)：語や表現に基づく特徴量	0.3844	0.5414	0.4217	0.5430	0.4499	0.4390
(3)：ユークリッド距離に基づく特徴量	0.2755	0.5121	0.3506	0.4101	0.3694	0.3472
(1)+(2)	0.5769	0.5752	0.5267	0.5610	0.5294	0.5365
(1)+(3)	0.5014	0.5538	0.4934	0.5263	0.4960	0.5046
(2)+(3)	0.3851	0.5414	0.4217	0.5236	0.4503	0.4460
(1)+(2)+(3)	**0.5791**	**0.5785**	**0.5387**	0.5572	0.5253	0.5321

表 10　ロジスティック回帰に素性を組み込んだ場合の F 値の結果

ロジスティック回帰						
	アンダーサンプリング無し			アンダーサンプリング有り		
本研究の素性の組合せ	適合率	再現率	F1	適合率	再現率	F1
(1)：TF-IDF に基づく特徴量	0.5299	0.5628	0.4977	0.5385	0.4952	0.5050
(2)：語や表現に基づく特徴量	0.3913	0.5463	0.4251	0.5396	0.4507	0.4438
(3)：ユークリッド距離に基づく特徴量	0.2966	0.5142	0.3520	0.4175	0.3868	0.3845
(1)+(2)	0.5901	0.5909	0.5504	0.5767	0.5414	0.5487
(1)+(3)	0.5326	0.5653	0.5056	0.5371	0.4940	0.5035
(2)+(3)	0.4917	0.5360	0.4365	0.5141	0.4515	0.4497
(1)+(2)+(3)	**0.5965**	**0.5987**	**0.5697**	0.5639	0.5270	0.5332

分類器には 7 つの分類器のうち，SVM とロジスティック回帰を用い本研究の素性の評価を行った．評価の結果は表 9 と表 10 である．実験の結果，SVM とロジスティック回帰の両分類器ともに素性 1，素性 2，素性 3 の 3 つを用いた場合がもっとも F 値が高い結果となった．それぞれの F 値は，SVM は 0.5387，ロジスティック回帰は 0.5697 となり分類には疑問の残る結果となった．実験の結果よりロジスティック回帰の適合率は 0.5965 であり，飲食店の再訪問の割合を出力するならば微少ながら有効であると確認できた．

また，混同行列にて実験結果の分析を行った．混同行列の結果を表 12 と表 13 に示す．この表より，SVM とロジスティック回帰ともに,「再訪問」と「不明」が「初訪問」に偏って分類されていることが確認できる．また，本研究で提案した分類器で再訪問レビューを集めた場合と，ランダムで再訪問レビューを集めた場合の比較をする．本研究で扱ったレビュー数は 2425 件である．評価者によって得られた再訪問レビューは 437 件である．全レビューからランダムで選択した場合は 437/2425 件となり約 18%で収集される．対して，混同行列の結果から，分類器が再訪問と判断した件数は 193 件，予測と正解が一致したものは 124 件であった．これより分類器が再訪問レビューと分類した場合の精度は 124/193 件，やく 64%である．このことより，再訪問レビューを収集するさいに本研究の分類器を用いることは有用であると述べられる．

6.　まとめと今後の課題

本研究では飲食店のレビュー文書から，初訪問のレビュー文書であるか，再訪問のレビュー文書であるか，どちらでもない不明のレビュー文書であるかを自動的に分類する手法を提案した．

表 11　7 つの分類器に全素性を用いた場合の F 値のスコア

分類器	u	適合率	再現率	F1
Naive Bayes	1	0.6160	0.3558	0.2472
決定木	0	0.2472	0.4659	0.4551
ロジスティック回帰	0	0.5965	0.5987	**0.5697**
SVM	0	0.5791	0.5785	**0.5387**
kNN	1	0.6160	0.3558	0.2472
AdaBoost	0	0.5288	0.5439	0.5028
RandomForest	0	0.4962	0.5183	0.4103

本研究ではレビュー文書を 2425 件集め，それらのレビュー文書を，評価者 3 名にて初訪問レビューか再訪問レビューか不明レビューであるかラベル付けを行った．このラベル付けの結果より，人ならばレビュー文書を読めば，ある程度は一致して分類することが可能であることを示した．また，このラベル付けにおいて，初めてレビューを投稿したユーザであっても，その飲食店を再び利用していると評価者 3 名中 2 名が一致して判断したレビュー文書は，全レビュー文書 2425 件中に 437 件あり，本研究の課題は適当であることを確認できた．

そしてそれらのレビュー文書を分類するため，本研究では分類する分類器の構築のための特徴量を 3 つ作成した．

1 つ目の特徴量は，レビュー文書を Bag-of-words として扱い，TF-IDF に基づく素性を作成した．2 つ目の特徴量は，他のデータから取得した根拠文を分析し，レビュー文書で特徴的に表れる，人の感覚に基づく，初訪問や再訪問を表す語や表現などの特徴量を見つけ，ルールベースにて 16 次元の素性を作成した．3 つ目の特徴量は人が初訪問レビューや再訪問レビュー

表 12　SVM かつ全素性を用いた場合の混同行列

		予測		
		初訪問と識別	不明と識別	再訪問と識別
正解	初訪問	1067	144	38
	不明	472	245	22
	再訪問	279	67	91

表 13　ロジスティック回帰かつ全素性を用いた場合の混同行列

		予測		
		初訪問と識別	不明と識別	再訪問と識別
正解	初訪問	1056	157	36
	不明	434	272	33
	再訪問	235	78	124

と判断した文に注目したものである．人がレビュー文書を判断するのは，レビュー文書全体からではなく，レビュー文書中にある根拠文が主な要因であると推測した．そしてそれらの根拠文には一定の類似性があることに注目し素性を作成した．素性の作成にあたり，根拠文の類似度が高くとも，否定語が入るなどで意味の違いが生じる不都合を緩和するため，根拠文をクラスタリングし類似性を緩和した代表ベクトルを作成した．クラスタリングには K-means 法で初訪問と再訪問のそれぞれでクラスタ数を 15 個と指定し，ユークリッド距離にて 30 個のクラスタを得て 30 次元の素性を作成した．そして，これら 3 つの特徴量を付与し分類器を構築した．

分類器には 7 つの分類器のうち，F 値が高かった SVM とロジスティック回帰を用い本研究の素性の評価を行った．評価の結果は表 9 と表 10 である．実験の結果，SVM とロジスティック回帰の両分類器ともに，TF-IDF による特徴量と，特徴的な語の使われ方による特徴量を用いた場合がもっとも高い結果となった．それぞれの F 値は，SVM は 0.53652，ロジスティック回帰は 0.5697 となり分類には疑問の残る結果となった．実験の結果よりロジスティック回帰の適合率は 0.5965 であり，飲食店の再訪問の割合を出力するならば微少ながら有効であると確認できた．また SVM とロジスティック回帰の混同行列は表 12 と表 13 になる．

分類問題において強い特徴量は分類器の精度をあげるために重要である．本研究では人が気づける特徴からのアプローチと，人が気づきにくい特徴を発見するアプローチからレビュー文書の分類を試みた．実験 1 では「初訪問」「再訪問」レビュー文書を分類する分類器の構築の実現可能性を導きだせた．実験 2 では「初訪問」「再訪問」「不明」のラベル付けにおいて複数人の意見に似た兆候が確認できた．このことより，再訪問と初訪問の分類の特徴が存在すること，そして分類器を構築することが可能であると確信が持てた．今後も，分類の精度を上げるため，新たな特徴量を求める研究をおこなってゆきたい．

謝　　辞

本研究の一部は JSPS 科学研究費助成事業 JP16H02906，JP17H00762，JP18H03243，JP18H03244，JP18H03494 による助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] K. Dave, S. Lawrence, and D.M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” Proceedings of the 12th International Conference on World Wide Web, pp.519–528, 2003.

[2] X. Ding and B. Liu, “The utility of linguistic rules in opinion mining,” Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.811–812, 2007.

[3] D. Osman, J. Yearwood, and P. Vamplew, “Using corpus analysis to inform research into opinion detection in blogs,” Proceedings of the Sixth Australasian Conference on Data Mining and Analytics - Volume 70, pp.65–75, 2007.

[4] D.T. Wijaya and S. Bressan, “A random walk on the red carpet: Rating movies with user reviews and pagerank,” Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp.951–960, 2008.

[5] E. Gilbert and K. Karahalios, “Understanding deja reviewers,” Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, pp.225–228, 2010.

[6] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, “Uncovering crowdsourced manipulation of online reviews,” Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.233–242, 2015.

[7] A. Mullick, S. Ghosh D, S. Maheswari, S. Sahoo, S.K. Maity, S. C, and P. Goyal, “Identifying opinion and fact subcategories from the social web,” Proceedings of the 2018 ACM Conference on Supporting Groupwork, pp.145–149, 2018.

[8] L.-C. Cheng, J.C.R. Tseng, and T.-Y. Chung, “Case study of fake web reviews,” Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp.706–709, 2017.

[9] J.P. Zagal, A. Ladd, and T. Johnson, “Characterizing and understanding game reviews,” Proceedings of the 4th International Conference on Foundations of Digital Games, pp.215–222, 2009.

[10] M.P. O’Mahony and B. Smyth, “Learning to recommend helpful hotel reviews,” Proceedings of the Third ACM Conference on Recommender Systems, pp.305–308, RecSys ’09, 2009.

[11] T. Lappas, M. Crovella, and E. Terzi, “Selecting a characteristic set of reviews,” Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.832–840, 2012.

[12] P.M. Vu, H.V. Pham, T.T. Nguyen, and T.T. Nguyen, “Phrase-based extraction of user opinions in mobile app reviews,” Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, pp.726–731, 2016.

[13] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” Journal of the American society for information science, vol.41, no.6, pp.391–407, 2013.

[14] J.L. Fleiss, “Measuring nominal scale agreement among many raters.,” Psychological bulletin, vol.76, no.5, p.378, 1971.