一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
DE2018-10（2018-09）

食べログの再訪問レビューの発見

白髪　宙海†　　村本　直樹†　　大島　裕明†

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

あらまし　本研究では，飲食店のレビュー文書が初訪問のものであるか再訪問のものであるかを自動的に分類する問題に取り組んだ．ウェブ上には飲食店のレビューを書き込むことができるレビューサイトが数多く存在しており，多くの飲食店を対象として多くのユーザがレビューを書き込んでいる．飲食店の評価は，価格帯，味，サービスなどで行うことができるようになっており，レビューサイトではこれらの評価値を入力することができるようになっている．飲食店に対する評価の観点は，これらに限られるものではない．たとえば，リピーターの多さ，すなわち，再訪問率も重要な評価の観点であると考えられる．しかし，飲食店に対するレビュー文書を精査すると，初訪問であると考えられるものと，再訪問であると考えられるものが存在している．これらを自動的に分類することが，飲食店の再訪問率を求めることにつながり，より多面的な飲食店の評価につながるものと考える．

キーワード　再訪問推定，レビュー情報，データマイニング

Hiromi SHIRAGA†, Naoki MURAMOTO†, and Hiroaki OHSHIMA†

† Graduate School of Applied Informatics, University of Hyogo
7-1-28 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650–0047, Japan
E-mail: †{aa17j506,aa18c508,ohshima}@ai.u-hyogo.ac.jp

1. はじめに

近年，あらゆる商品やサービスについてのレビューが共有されるようになっている．その中でも，飲食店のレビューは，飲食店を選択する意思決定の場面において非常に重要な役割を果たすようになってきている．飲食店レビューサイトとしては，食べログ，ぐるなび，ホットペッパー，Retty などがあげられる．これらの検索サービスでは店名，メニュー，電話番号，住所（地図）のような飲食店そのものの情報が得られるほか，実際に飲食店を利用したユーザの投稿を見ることができる．

ユーザの投稿では，飲食店は，価格帯，味，サービス，コストパフォーマンスなどの評価項目で評価されている．また，同時に，自由記述のレビューがテキストで書かれる．飲食店は，これらの評価項目のように，多様な側面から評価されているが，これら以外の評価尺度でも評価することは可能である．そのうちの一つが，リピーターの多さ，すなわち，再訪問率であると考える．飲食店に対するレビュー文書を精査してみると，初訪問であると考えられるものと，再訪問であると考えられるものが存在している．これらを自動的に分類することは，飲食店の再訪問率を求めることにつながり，より多面的な飲食店の評価につながるものと考える．そこで，本研究では，飲食店レビューサイトに投稿されたレビュー文書が初訪問のものであるか再訪問のものであるかを自動的に分類する問題に取り組む．

飲食店レビューサイトの一つである食べログでは，ユーザが同一の飲食店に対して複数回のレビューを書くことが可能となっている．図 1 のとおり，各レビューにはそれが何回目のレビューであるかが表示される．このような場合には，少なくとも 2 回目以降のレビューは再訪問した上でのレビューであると

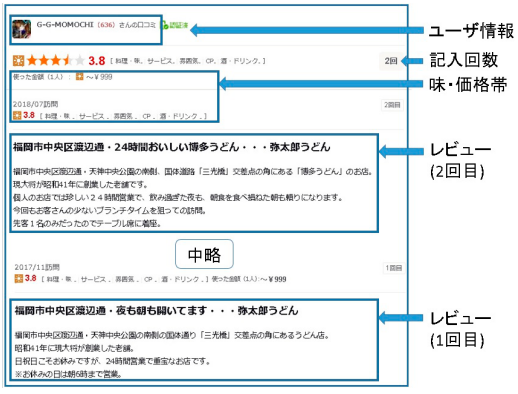

図 1　食べログのある飲食店のレビューのページ

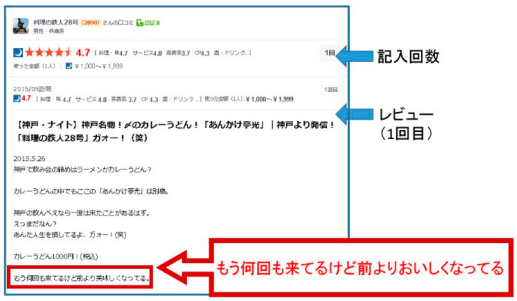

図 2　再訪問のレビューの例

判断できる．しかし，1 回目のレビューは本当に初訪問時のものである場合もあれば，再訪問時のものである場合もある．また，多くのユーザは同一の飲食店にレビューを 1 回だけ書いており，そちらにも初訪問時のものと再訪問時のものが混在している．たとえば，図 2 では，あるユーザがある飲食店に対して 1 回レビューを書いている．レビュー文書を精査すると，「もう何回も来てるけど前よりおいしくなってる」という記述が存在しており，過去にもこの飲食店を利用していることが推測できる．このように，書かれた内容から，初訪問や再訪問を特徴付ける特徴量を取得することで，自動分類が実現できると考えられる．

本研究では，レビュー文書の文書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特徴的な語などに着目した特徴量，初訪問や再訪問を表す文との類似性に考慮した特徴量を用いて分類器を作成することを提案する．テスト用のデータを用いて評価を行い，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

ウェブ上には，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．

浅野ら [1] は，食べログとぐるなびで投稿されたレビューがどのような観点から行われたものかを分類する研究を行った．提案手法では，あるレビュー文書に含まれるそれぞれの文が，「メニュー単品」「メニュー全体」「雰囲気」のいずれの観点に着目しているかを自動的に分類する．ぐるなびと食べログのレビューの性質の違いに着目し，ある種のラベル伝播を行うことによって分類を実現している．萩原ら [2] は，楽天市場のレビューの中から実際に商品を使用した人の体験に基づいたレビューを発見する研究を行った．彼らは，体験情報の記述には，特徴的な助詞や助動詞の出現があることに着目した．レビュー文書の係り受け解析を行い，主格を表す助詞を含む文節から，助動詞「た」が含まれる文節への係り受けが存在する場合に，体験情報であると判定する手法を提案した．西川ら [3] は，旅行ポータルサイトの TripAdvisor に投稿されたレビューの研究を行った．TripAdvisor では，レビューの投稿時には「ビジネス，カップル，家族旅行，友達，一人旅」の 5 種類の利用の種別を指定することになっている．彼らの研究では，この利用の種別を正解ラベルとして，それらをレビュー文書から推測するという問題に取り組んだ．提案手法はレビュー文書を，文字 N-gram と単語 N-gram を用いて特徴量を作り，SVM で分類するものであった．

3. 問題定義とデータセット

3.1 問 題 定 義

本研究で取り組む問題は，テキスト文書の 2 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，初訪問であるか再訪問であるかを出力する．

先述したとおり，食べログでは，あるユーザがあるレストランについて複数回レビューを書くことが可能であるが，本稿では，そのそれぞれの回の文書のことを**レビュー文書**と呼ぶこととする．すなわち，図 1 では 2 件のレビュー文書があり，図 2 では 1 件のレビュー文書があるということになる．

評価は，適合率，再現率，F1 スコアを用いることとする．

3.2 分析と学習のためのデータセットの収集

本研究では，はじめに，ある程度の量の食べログのレビューを収集して分析を行い，そのデータを用いて分類器を構成した．そのデータ収集の方法について説明する．

まず，6 つの地域と 6 つのジャンルを選択した．選択にあたっては，ある程度のばらつきがあることと，飲食店の件数がある程度あること，それらの飲食店に対するレビュー件数がある程度あることを考慮した．

飲食店の 6 つの地域は以下のとおり，都市部から 3 地域，郊外から 3 地域である．

都市部

- 兵庫県神戸市中央区
- 北海道札幌市中央区
- 福岡県福岡市中央区

郊外

- 兵庫県宝塚市
- 北海道帯広市
- 福岡県春日市

飲食店の 6 つのジャンルは以下のとおりであり，特別な日などに利用される 3 つのジャンルと，日常で利用される 3 つのジャンルである．

特別な日などに利用されるジャンル

- フレンチ
- 寿司
- ハンバーグ

日常で利用されるジャンル

- うどん
- パン・サンドウィッチ
- 定食・食堂

地域とジャンルの組み合わせは全 36 通りある．それぞれの組み合わせで飲食店を 5 店舗ずつ，計 180 店舗選択した．次に，飲食店ごとに 5 名のユーザを選択し，レビュー情報を収集した．これにより，総計 900 件のレビュー情報が収集された．

表 1　レビュー文書 900 件のラベル付けの結果

ラベル	レビュー文書数
1（初訪問）	418
2	79
3	176
4	24
5（再訪問）	203

収集された 900 件のレビュー情報から「1 回目」のレビュー文書を取得した．2 回目以降のレビュー文書については取得しない．これにより，総計 900 件のレビュー文書が収集された．

3.3 「初訪問」「再訪問」ラベルと根拠文

それぞれのレビュー文書を，初訪問におけるものか，再訪問におけるものかという観点から，ラベル付けを行った．ラベル付けは以下の 5 段階で行った．

（1） 確実に初訪問
（2） 初訪問だろうと思われる
（3） 不明
（4） 再訪問だろうと思われる
（5） 確実に再訪問

ラベル付けは，第一著者が行った．以下に，レビュー文書の例をあげる．

> 「**以前から気になっていた神戸の精肉店**．中山手通にあります．店内のお肉の一部は 30% 追加で調理してもらえ店内でいただけるということ．今回はサンドウィッチなどをいただくことにしました．（中略）**気になっていたので行けて良かったです♪**お肉を是非焼いてもらいたくなりました！ブログでも紹介しています．（後略）」

このレビュー文書の冒頭部には「以前から気になっていた神戸の精肉店」，後半部には「気になっていたので行けて良かったです♪」という記述がみられる．通常，これらのような記述は，過去に訪問したことがあるユーザが書くとは考えられない．そのため，このレビュー文書は初訪問のものであると判断し，「1」というラベルを付与する．

900 件のレビュー文書のラベル付けの結果は表 1 の通りである．ラベル「5」が付与されたレビュー文書が 900 件中 203 件存在しており，あるユーザがある飲食店に初めて書いたレビューが必ずしも初訪問ではないということが確認できた．

3.4 根拠文の取得

先ほど示したレビュー文書の中には，「以前から気になっていた神戸の精肉店」という記述が現れており，この部分はこのレビュー文書が初訪問のものであると判断した根拠となっている．このように，レビュー文書には判断の根拠となった文が一つないしは複数存在すると考えられる．そのような判断の根拠となった文をここでは**根拠文**と呼ぶこととする．初訪問か再訪問かが不明であるラベル「3」の場合には，根拠文は得られないと考えられるが，確実に初訪問のラベル「1」，ないしは，確実に再訪問のラベル「5」が付与できる場合の多くでは，根拠文が得られると考えられる．根拠文には，初訪問や再訪問を特徴付ける語などが現れている可能性があるため，900 件のレビュー文書中から取得した．

根拠文は，初訪問のラベル「1」に対して 419 件，再訪問のラベル「5」に対して 199 件取得できた．根拠文の例を表 2 に示す．

一つのレビュー文書から複数の根拠文が取得できる場合もあれば，得られない場合も存在した．一例としては，食べログにおける 1 回のレビュー文書に，明らかに複数回の訪問についてのレビューを書いている場合があげられる．このような場合，ラベル付けは「5」と行った．しかし，このようなときには，レビュー文書中には再訪問を明確に表す文が存在せず，根拠文が得られないことがあった．

3.5 テスト用データセットの収集

3.2 節で収集したデータセットは，分析して，どのような素性を取得すればよいかを検討することなどに利用した．そのため，別に，テスト用のデータセットを用意する必要がある．

そこで，以下の 2 地域を対象として，別のレビュー文書を収集した．

- 千葉県千葉市
- 千葉県浦安市

飲食店のジャンルは 3.2 節で述べた 6 つのジャンルを用いる．

飲食店の地域とジャンルの組み合わせは全 12 通りとなる．それぞれで 2 店舗ずつを選択し，計 24 店舗をテスト用データセットの対象とした．それぞれの飲食店で 5 名のユーザを選択し，「1 回目」のレビュー文書を取得した．これにより，総計 120 件のレビュー文書が収集された．

これらのデータセットにおいても，3.3 節と同様に，初訪問におけるものか，再訪問におけるものかという観点から，ラベル付けを行った．

120 件のレビュー文書のラベル付けの結果は表 3 の通りである．これらのうち，ラベル「1」または「2」が付与されたレビュー文書を初訪問のレビュー文書とし，ラベル「4」または「5」が付与されたレビュー文書を再訪問のレビュー文書として，評価に用いる．ラベル「3」が付与されたレビュー文書は評価に用いないこととする．これにより，初訪問のレビュー文書が 66 件，再訪問のレビュー文書が 30 件用意できた．

4. レビュー文書の特徴ベクトルの生成

本研究では，レビュー文書を分類器を用いて分類する．そのために，まず，レビュー文書を特徴ベクトルとして表現する．特徴ベクトルは，3 種類の素性から作成した．本節ではそれらについて説明する．

4.1 TF-IDF に基づく特徴量

まず，レビュー文書を Bag-of-words とみなして，TF-IDF を重みづけとして特徴ベクトルを作成した．レビュー文書を形態素解析し，すべての語を基本形とする．形態素解析器には，Janome(注1) を用いる．ここで，DF を求めるための文書集合

(注1)：Janome Web サイト: http://mocobeta.github.io/janome/

表 2　レビュー文書から得られた初訪問の根拠文と再訪問の根拠文の例

初訪問	うどん県に生まれ育った私にとっては初めて目にしました.
	何度か前を通り，入店しようとしましたが満席で入れませんでした.
	人気店ですので高いかと思いましたら，意外にお手頃・・.
	何度か利用したことがありましたが，天神のこちらのお店は初訪問.
	美味しい定食屋さん『タカチホキッチン』があると聞き訪問しました.
	食べログのおかげで知らなかったパン屋さんに出会えたこと，本当に先行レビュアーさんに感謝なの.
	ステーキを食べに行くお店としては選択外かな.
再訪問	再訪．もう，何回来たか分からんぐらいなのに，お昼には初訪問.
	ひっさし振りに?コチラのうどんが食べたくなりまして?の訪問.
	一人で福岡に行った時は食べるが頻度としては 10 年に 1 度くらいもの.
	高校時代によくお世話になりました.
	ここは堅焼き美味しいんですよねー.
	既に 2〜3 回利用していまして.
	家族での会食に利用させていただいています.

表 3　テスト用データセット 120 件のラベル付けの結果

ラベル	レビュー文書数
1（初訪問）	56
2	10
3	24
4	2
5（再訪問）	28

は，3.2 節で説明した 900 件のレビュー文書の中で，ラベルが不明とされた「3」以外の 724 件である．TF-IDF の実装は，scikit-learn の TfidfVectorizer を用いた．パラメータとしては，全文書中 1 つの文書にしか現れない語を無視する，min_df=2 を用いた．語の総種類数は 7,436 語であった．

次に，TF-IDF で表現された特徴ベクトルを LSA [4] に相当する手法で 200 次元に圧縮した．

以上のように，文書から TF-IDF に基づく 200 次元の素性が得られた．

4.2　初訪問や再訪問を表す語などに基づく特徴量

3.4 節で得られた根拠文には，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，根拠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．

まず，根拠文を Janome で形態素解析し，すべての語を基本形にする．語の品詞が名詞，動詞，形容詞，副詞のもののみを残し，また，数値は取り除いた．そのようにして得られた語について，全根拠文における出現回数を数え，根拠文において頻出する語のリストを得た．その中から，初訪問ないしは再訪問のいずれかの根拠文にのみよく現れる語を中心として，観察を行い，特徴的な語などを取得した．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．最終的には，17 次元の特徴量を取得した．

たとえば，顕著に初訪問を示す語として「初訪」があげられる．この語に対応して，レビュー文書中に単純に「初訪」が含まれる場合には 1 として，そうでない場合には 0 とする素性を一つ得た．初訪問を示す他の語としては「ログ」という語が得られた．この語は，他の食べログユーザや他のグルメブログを参照している場合によく現れる語である．たとえば，「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」といったような記述である．このように，他のレビューを参照としている場合には，それを飲食店を選択するために利用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初訪問であると考えられる．そこで，「ログ」という語が含まれているかいないかを表す素性を作成する．

顕著に再訪問を示す語としては，「毎回」「追記」「年」「回数」があげられる．これらについても同様に素性を作成した．また，「回目」という語も再訪問を示す語であると考えられた．ただし，「1 回目」や「一回目」という記述での出現の場合は，初訪問であるか再訪問であるかは分からないことが多く，特徴量として用いることは適切ではないと考えられる．そこで，「1 回目」や「一回目」という記述以外での「回目」という出現について，同様に素性を作成する．

「再訪」という語は，初訪問の場合にも出現する語である．たとえば，「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，再訪が文末に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感じされた．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 となる素性を作成する．

レビュー文書中における，過去の経験を表す表現が再訪問であることを示す場合がある．たとえば，「学生時代によく利用していた」のような場合．現在形の「利用する」や過去形の「利用した」といった表現は，初訪問でも再訪問でも使われると考えられる．「利用しています」のように現在の経験を表す表現は，ある程度は再訪問で使われると思われるが，今回は対象としなかった．今回利用したのは，過去の経験を表す表現である「利用していた」である．動詞表現であり，変化するため，形態素解析を行って原型を調べることで，変化に対応してこの表現の有無を取得し，素性を作成した．「利用していた」と同様に，「行っていた」「食べていた」「通っていた」という 3 つの表現

についてもそれぞれ素性を作成した．

たとえば，「約 1 年ぶりとなるランチ訪問です」といったような記述は，過去に来店したことを示す特徴と考えられる．数値や単位は変化させることができるため，「1 年ぶり」という表現以外にも，「二週間ぶり」といった表現が考えられる．これらに同様の表現を取得するために，形態素解析を行い，数助詞に続いて「ぶり」が現れる場合を見つけることで素性を作成した．

「前回」という語は前に来店したという表現で使われることが多い．たとえば，「前回は奥側の大将の前で」という記述は，前に来店したという表現であると考えらる．そこで「前に」「前は」「前回は」「前回に」のいずれかが含まれているかどうかを 1 つの素性とした．

初訪問を表すと考えられる語としては，「見つける」「近く」「入る」があり，これらの語のそれぞれについて，含まれているかどうかを素性とした．「見つける」という語は，たとえば「また素敵なお店を見つけることが出来ました」というように，初訪問のレビュー文で使われることが多いと考えられる．「近く」の語の例としては「近くに所用で来た折に」「ホテル近くに」「近くのパン屋を検索すると」など他の目的で近くの飲食店を利用するきっかけの表現での記述が多く確認できた．このような場合，総じて初訪問のケースが多かった．「入る」は初めて訪れる飲食店のレビューの始めに使われることが多い．たとえば，「門を入るとすぐにパンのショーケースがあり」「大きな一軒家のお店の一階ホールに入ると」といったようなものが見られた．

以上，初訪問や再訪問を表す語などに基いて，17 次元の特徴量が得られた．

4.3 根拠文との類似性に基づく特徴量

3.3 節で述べたとおり，根拠文は，初訪問ラベル「1」に対して 419 件，再訪問ラベル「5」に対して 199 件の合計 618 件取得された．ある未知のレビュー文書の中に，これらの根拠文と類似する文が出現することが，初訪問や再訪問の特徴を表す可能性があると考えられる．そこで，レビュー文書に現れる文と根拠文の類似性に基づく特徴量を取得する．最終的に，根拠文の数と同じ 618 次元の素性を作成する．

文どうしの類似性を計算するため，Doc2Vec [5] を用いる．Doc2Vec のモデル生成のために，楽天データ公開 [6] にて提供されているデータを用いる．これは，今回収集したデータは量が十分ではないと考えられるためである．レビューを扱っていることから，楽天市場のレビューデータのうち，2010 年 3 月に投稿された全 1,280,237 件のレビューを利用した．Doc2Vec で表現される文書の特徴ベクトルの次元数は 300 次元とした．ウインドウサイズは 5，エポックは 5 とした．モデル生成を行った後，そのモデルに新規文書として文を与えると，300 次元のベクトルが得られる．これにより，根拠文や未知のレビュー文を 300 次元のベクトルにすることが可能であり，これらの文どうしの類似度を計算することができるようになる．

ある未知のレビュー文書 R が与えられたとする．まず，その文書を文集合とみなし，$R = \{r_1, r_2, \cdots, r_n\}$ と表すこととする．ここで，n は当該レビュー文書における文の数であり，r_j は j 番目の文を表す．次に，根拠文のリストを $E = [e_1, e_2, \ldots, e_{618}]$ と表すこととする．ここで，e_i は i 番目の根拠文を表す．

このとき，レビュー文書 R の根拠文 e_i に対応するベクトルの重み w_i を，以下の式で求める．

$$w_i = \max_j(\cos(e_i, r_j)) \tag{1}$$

この式は，ある根拠文と，レビュー文書中のすべての文とのコサイン類似度を計算し，その中での最大値をその根拠文に対応する次元の重みとするものである．次に，このようにして得られたベクトルの各次元を，さらに，以下のように二値化する．

$$w_i' = \begin{cases} 1 & (w_i \geq \theta) \\ 0 & (otherwise) \end{cases} \tag{2}$$

ここでは，値があるしきい値 θ 以上の場合には，重みを 1 とし，それよりも小さい場合には 0 としている．以上のように，根拠文との類似性に基づいて 618 次元の素性が得られる．

5. 評価実験

本節では，前節で述べた特徴ベクトルを用いて，レビュー文書を分類する手法とその評価実験について述べる．

5.1 実験方法

学習用のデータとして，3.3 節で述べた 900 件のレビュー文書のうち，初訪問か再訪問かの区別がつかなかった「3」のラベルが付与されたものを除く 724 件のビュー文書を用いた．初訪問を表すラベル「1」「2」が付与されたレビュー文書を負例（0）とし，再訪問を表すラベル「4」「5」が付与されたレビュー文書を正例（1）とした．それぞれの件数は，負例が 497 件，正例が 227 件となった．同様に，テスト用のデータには 3.3 節で述べた，負例 66 件と正例 30 件の計 96 件のレビュー文書を用いた．

分類器の構築には SVM を用い，分類器構築時には毎回，グリッドサーチで F1 を最大化するように最適なパラメータの決定を行った．グリッドサーチを行う際の各種パラメータは以下のとおりである．

$$\text{カーネル} = \{\text{linear}, \text{rbf}\} \tag{3}$$

$$C = \{1, 10, 100, 1000, 10000\} \tag{4}$$

$$\gamma = \{0.1, 0.01, 0.001, 0.0001\} \tag{5}$$

γ は RBF カーネルのパラメータである．

4 節で述べたように，特徴ベクトルは次の 3 種類の素性から作成される．

- TF-IDF：TF-IDF による特徴量
- 特有表現：初訪問や再訪問を表す語などに基づく特徴量
- 根拠文類似：根拠文との類似性に基づく特徴量

評価実験では，はじめに 4.3 節の (2) 式におけるしきい値 θ の決定を行う．θ を 0.1 から 0.9 の範囲で，0.1 ずつ変化させながら，全素性を使用した分類器を構築し，F1 が最も良くなる θ を確認する．さらに，得られた θ を用いて，以下の素性の組み合わせについて分類器を構築する．それぞれの分類精度について比較を行い，各素性が分類に貢献しているかを確認した．

表 4　θ を変化させたときの全特徴量を使用した分類器の精度

θ	適合率	再現率	F1	カーネル	C	γ
0.1	0.74	0.74	0.74	linear	1	
0.2	0.65	0.66	0.65	rbf	1000	0.001
0.3	0.68	0.67	0.67	rbf	1000	0.001
0.4	0.66	0.65	0.65	linear	10	
0.5	0.65	0.66	0.65	rbf	10000	0.0001
0.6	0.65	0.65	0.65	rbf	10000	0.0001
0.7	0.65	0.67	0.66	linear	1	
0.8	**0.78**	**0.78**	**0.78**	rbf	100	0.01
0.9	0.67	0.67	0.67	rbf	1000	0.01

表 5　実 験 結 果

素性の組み合わせ	適合率	再現率	F1	カーネル	C	γ
TF-IDF	0.70	0.72	0.71	linear	10	
TF-IDF+ 特有表現	0.72	0.73	0.72	linear	10	
TF-IDF+ 根拠文類似	0.66	0.65	0.65	rbf	10000	0.001
TF-IDF+ 特有表現 + 根拠文類似	**0.78**	**0.78**	**0.78**	rbf	100	0.01

- TF-IDF
- TF-IDF+ 特有表現
- TF-IDF+ 根拠文類似
- TF-IDF+ 特有表現 + 根拠文類似

5.2 実 験 結 果

表 4 が全素性を使用し，適合率，再現率，F1 がどのように変化したかを示している．結果は，θ が 0.8 のときに，適合率，再現率，F1 のすべてが最も高くなった．その時の SVM のパラメータは，RBF カーネル，C は 100，RBF カーネルのパラメータ γ は 0.01 であった．それぞれの θ で，最適な SVM のパラメータが異なるため単純に比較することはできないが，θ が 0.8 の時と，θ が 0.1 の時に F1 が 0.74 となった以外では，F1 が 0.65 から 0.67 程度であった．

表 4 より，θ が 0.8 のときに最も精度が高くなった．このしきい値 θ を用いて，先述した素性の組み合わせ 4 種類について，分類器を構築し，評価を行った．表 5 が，それぞれの素性の組み合わせにおける最適な SVM のパラメータと，適合率，再現率，F1 を示している．すべての特徴量を用いた場合に，最も良い精度が得られたことが分かった．TF-IDF に基づく特徴量を用いたときには F1 が 0.71 であった．それに，根拠文との類似性に基づく特徴量を加えた場合には，F1 が 0.65 に下がってしまう．しかし，それにさらに初訪問や再訪問を表す語などに基づく特徴量を加えた場合，すなわち，すべての特徴量を用いた場合に F1 が最も良くなっており，これらを組み合わせて用いることが必要であることが示唆された．

6. ま と め

本研究では飲食店のレビュー文書を，初訪問のものと再訪問のものに自動的に分類する手法を提案した．人がレビュー文書を読んだ際には，そのユーザが初訪問の時に書いたレビューか，再訪問の時に書いたレビューかをある程度見分けることが可能である．900 件のレビュー文書を収集し，その分析を行うことによって，レビュー文書に特徴的に現れる初訪問や再訪問を表す語などを発見した．さらに，レビュー文書中で初訪問か再訪問を判断する判断材料となった根拠文を 618 件取得した．分類のための特徴量としては，まず，レビュー文書を Bag-of-words として扱い TF-IDF に基づいて作成した素性を用いた．次に，初訪問や再訪問を表す語などが含まれているかどうかを表す素性を計 17 次元作成した．さらに，根拠文との類似性に基づく 618 次元の素性を作成した．新たにテスト用のデータセットを作成して，これらの特徴量を用いて SVM によって分類を行った場合の精度評価を行った．最適なパラメータでは，F1 が 0.78 となる結果が得られた．今後，本研究を基礎技術として用いて，飲食店の再訪問率を推定することなどを行っていきたい．

謝　　辞

本研究の一部は JSPS 科学研究費助成事業 JP16H02906, JP17H00762, JP18H03243, JP18H03244 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] 浅野祥汰，北山大輔，“レビューサイトの差異に基づくラベル伝播を用いたユーザレビュー分類手法，” 第 9 回データ工学と情報マネジメントに関するフォーラム（DEIM 2017），pp.D4–3，2017.

[2] 萩原一貴，大野一樹，波多野賢治，“品詞間の係り受けに着目した体験情報抽出の提案，” 第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM 2014），pp.E6–6，2014.

[3] 西川崇哉，岡田真，橋本喜代太，“レビュー文書の自動分類におけるテキストの前処理手法の検証，” 言語処理学会第 18 回年次大会発表論文集，pp.517–520，March 2012.

[4] 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.

[5] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” IInternational Conference on Machine Learning, pp.1188–1196, 2014.

[6] “楽天データ公開,” （2018 年 8 月 7 日アクセス）. https://rit.rakuten.co.jp/data_release_ja/.