Vol.2013-NL-213 No.8
2013/9/12

Random Forest を用いた類似レビュアーの推薦手法の検討

徳田祐貴† 梅澤猛† 大澤範高†

Web 上の商品レビューは，多様な好みや感覚を持ったレビュアーによって書かれている．そのため，ユーザは多くのレビューに目を通すことで自らが興味のある情報を探し出さなければならない．そこで本研究では，ユーザ自身もレビューを書いていることを前提とし，ユーザと類似したレビュアーを推薦することでレビュー閲覧の負荷軽減を図る手法を提案する．ユーザとレビュアーのレビュー群に出現する単語の tf-idf 値，文章の長さや文字種の割合などを素性とする Random Forest を用いてユーザとレビュアーの類似度を求め，それに基づいた推薦する手法およびその評価方法を検討する．

A Study on Similar Reviewer Recommendation using Random Forest Algorithm

YUUKI TOKUDA† TAKESHI UMEZAWA†
NORITAKA OSAWA†

Customer reviews on the Web are written by reviewers with various preferences. A user often wants to know reviewers with similar preferences to find interesting reviews for him/her efficiently. It is usually tedious and time-consuming to find interesting reports from a large set of reviews. Therefore, we investigate methods of similar reviewer recommendation to find useful information efficiently. Similarity between a user and reviewers is computed using Random Forest Algorithm, in which features are , tf-idf values of words in reviews, the length of a sentence and occurrence ratios of character type. On the basis of similarity between reviewers, we study reviewer recommendation methods and evaluation of the methods.

1. はじめに

Web 上のレビューサイトなどでは誰でも簡単に情報発信できるため膨大な量の口コミ情報が存在している．また，情報を発信するレビュアーは多様であり，様々な好みや感覚を持ったレビュアーが存在する．あるアイテムに対して書かれたレビューが大量に存在していると，ユーザはその中から自分が興味のある情報を探し出さなければならない．

そこで，ユーザに類似した嗜好を持ったレビュアーを推薦することで，レビュー閲覧の負荷軽減を図ることができると考え，ユーザにレビュアーを推薦する手法を検討する．本研究では，ユーザ自身もレビューを書いているレビュアーであることを前提とし，そのユーザとレビュアーのレビュー文書を利用する．そのレビュー文書群の出現単語の tf-idf 値，文章の長さ，文字種の割合といった特徴と，データの欠損値に頑健な学習アルゴリズムである Random Forest を用いて識別モデルを作成し，そのモデルからユーザとレビュアーの類似度を求める検討を行った．

2. 関連研究

協調フィルタリング（Collaborative Filtering）[1][2][3]は，ユーザ毎に嗜好情報を蓄積し，あるユーザと類似した嗜好をもつ他のユーザの情報を利用して商品の推薦を行う手法である．嗜好情報としては，商品に対する評価値，購入や閲覧などの行動履歴，年齢・性別などの人口統計情報が利用される．ユーザ間類似度の指標としては，コサイン類似度やピアソン相関係数が用いられることが多い．ユーザ u とユーザ v のコサイン類似度 $sim(u,v)$ は，全ユーザの集合を U，全商品の集合を I，ユーザ u の商品 i に対する評価を $r_{u,i}$ とすると，式（1）で表せる．

$$sim(u,v) = \sum_{i \in I} r_{u,i}\, r_{v,i} \Big/ \left(\sqrt{\sum_{u \in U} {r_{u,i}}^2} \sqrt{\sum_{v \in U} {r_{v,i}}^2} \right) \quad (1)$$

共通の商品に対するユーザの評価値や履歴の情報を用いてコサイン類似度を求めることで，類似したレビュアーを見つけることができる．しかし，共通の商品を評価，購入または閲覧していることが前提となっており，それ以外の要素は考慮されない．実際には，商品の総数は非常に大きく，ユーザの評価値や履歴などの情報データベースにおいては，データに NULL 値が多数点在し，データベースが疎である問題がある．したがって，共通に評価がついている商品の数がわずかな場合には，この手法で算出される類似度は適切ではない恐れがある．

岡田らは，協調フィルタリングを行う際のユーザ間の類似度を，ユーザのレビュー文書を用いて求める手法[4]を提案している．ユーザのレビュー文書に出現する各単語の tf-idf 値を特徴ベクトルとし，それに基づいたコサイン類似度によりユーザの類似度を求めている．tf-idf は，ある文

† 千葉大学大学院融合科学研究科
Chiba University, Graduate School of Advanced Integration Science

書において出現頻度が高く（tf：Term Frequency），他のドキュメントにはあまり出現していない（idf：Inverse Document Frequency）単語を特徴的な単語としてとらえるための指標であり，情報探索やテキストマイニングなどの分野で広く利用されている[5]．tf-idf 値 $w_{i,j}$ は式（2）によって定義される．

$$w_{i,j} = tf_{i,j} \times \log \frac{N_D}{df_i} \quad (2)$$

ここで $tf_{i,j}$ は，文書 D_j 中に現れる単語 i の出現回数を文書 D_j の総単語数で割ったものであり，N_D は総文書数，df_i は単語 i を含む文書数である．

岡田らの実験では，ユーザの評価値を用いてコサイン類似度を算出し，協調フィルタリングを行う従来の手法と，ユーザのレビュー文書に出現する全単語の tf-idf 値を用いてコサイン類似度を算出して協調フィルタリングを行う提案手法を比較している．しかし，従来の手法と岡田らの手法を比較した結果，性能の向上が見られなかった．一方で，ユーザのレビュー文書から感情を表す単語を抽出し，その tf-idf 値を用いてコサイン類似度を算出して協調フィルタリングを行った場合には性能の向上がみられたことが報告されている．このように，感情を表す単語を抽出するなどの加工を行ってはいるが，レビュアーのレビュー文書を利用することで協調フィルタリングの性能の向上することに成功している．

3. Random Forest

Random Forest[6][7]は，複数の決定木を弱識別器とする集団学習アルゴリズムであり，データマイニングなどの分野で広く利用されており，欠損値を持つデータでも有効な動作が可能，高精度な分類が可能，多変数であっても計算可能，変数の寄与率が算出可能といった特徴がある．ここで，寄与率は重要度に関係した値であり，モデルを用いて分類を行う際に識別性能に寄与している割合のことである．

Random Forest では，まず学習データ学習サンプルから T 個のブートストラップサンプル $I_1 \ldots I_t$ を生成し，それぞれのブートストラップサンプル I_k における M 個の変数の中からランダムに m 個*を選択する．つぎに，その中から最もよい変数を分岐ノードとし，未剪定の T 個の決定木を構築する．そして，作成したモデルを用いて未知のサンプルデータの予測を行う際に，学習で構築された T 個の決定木それぞれで未知のサンプルデータを判定し，T 個の決定木から得られた結果を統合することにより最終的な出力結果とする．

Random Forest のアルゴリズムにより作成された判別モデル（決定木）を用いて，N 個のサンプル間の $N \times N$ の類似度行列を求めることができる．類似度は 0 から 1 の値を取り，1 に近いほどデータ同士が類似していることを意味する．類似度行列は対称行列であり，行列の対角要素は同じサンプルデータの類似度のため全て 1 となる．類似度行列は，つぎの手順で求めることができる．

1. 成分が全て 0 の $N \times N$ の行列の行列を用意する
2. サンプル S_r とサンプル S_c がある決定木において同じ終端ノードにたどり着いた場合（r, c）成分に 1 を加える
3. 全ての決定木において手順 2.を行い，最後に決定木の本数で行列全体を割る

サンプル S_r とサンプル S_c の類似度 $prox(S_r, S_c)$ は，この類似度行列の（r, c）成分である．このように，生成されたそれぞれの決定木において，ある 2 つのサンプル同士がどのくらい同じ葉ノード（終端ノード）に属したかに基づいて類似度が求められる．

4. 提案手法

本研究では，ユーザ間の類似度を求める手法として，レビュー文書を用い，欠損値に強い学習アルゴリズムとして Random Forest を適用することを提案する．提案手法による類似レビュアー推薦の概要を図 1 に示す．

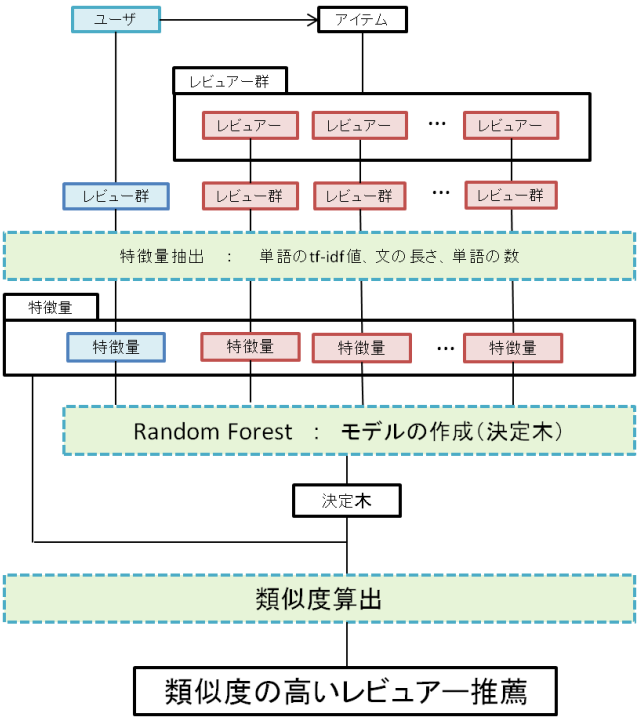

図 1　提案システム概要図

Figure 1　Overview of proposed system.

まず，ユーザがレビューを参照したいアイテムを指定する．システムは，指定されたアイテムにレビューを書いて

* m < M であり，$\sqrt{M}$ が常用されている

いる各レビュアーについて，公開している全レビューを収集する．つぎに，レビュー文書を言語解析して得られた特徴量を変数として Random Forest に適用し，識別モデルを得る．そして，得られたモデルからユーザと各レビュアーの類似度をそれぞれ算出し，値の高いものを類似レビュアーとして推薦する．

なお, Random Forest に変数として入力する特徴量の候補としては，つぎのようなものが挙げられる．

- 単語の tf-idf 値（出現文書数が上位の単語）
- レビューの長さ（単語数，文字数）
- 文字種や単語種の数（品詞，記号，句読点）

4.1 類似度の算出

ユーザとレビュアー間の類似度は, Random Forest における類似度行列の平均値と定義する．ユーザを a，別のレビュアーを b，それぞれのレビュー文書の集合をR_a，R_b，その要素のレビューを$r_{a,di}$，$r_{b,dj}$とし，文書間類似度を$prox(r_{a,di},r_{b,dj})$とすると, ユーザ a とレビュアーb 間の類似度$sim(a,b)$ は式(3)のと定義する．なお，n はレビュアー一人あたりのレビュー文書数である．

$$sim(a,b) = (\sum_{dj=1}^{n}\sum_{di=1}^{n} prox(r_{a,di},r_{b,dj}))/n^2 \quad (3)$$

ここで，作成した決定木の本数を T，ある単語 i の含まれる文書数を df_iとし，$df_i \geq N_{df}$の単語全ての tf-idf 値を特徴量として用いて算出した類似度を$sim(a,b,T,N_{df})$と表記する．N_{df}を最低出現文書数と呼ぶ．

5. 実験

提案手法の有効性を検証するために，Web 上に公開されている実際のレビュー情報を用いて実験を行った．まず，グルメレビューサイトより，50 以上のレビューを公開しているレビュアーを無作為に 30 人選び出し, それぞれについて 50 文書ずつレビュー文書を収集した（30 人×50 文書＝1,500 文書）. つぎに, 集めたレビュー文書群を言語解析し，単語ごとの tf-idf 値を算出した．そして，算出した値を特徴量とし，Random Forest を用いて 100～3,000 本の決定木を作成した．特徴量としては，$N_{df} \geq 10$ 以上の全ての単語 i の tf-idf 値を用いた. 最低出現文書数N_{df}と単語 i の種類の関係を表 1 に示す.

表 1　最低出現文書数N_{df}と単語数の関係
Table 1　N_{df} and the number of words which occur in more than N_{df} documents.

N_{df}	500	200	100	50	30	20	10
単語数	57	186	389	808	1330	1902	3266

全 1,500 文書に含まれる総単語数は 20,363 であった．以上のような環境において，つぎの 2 つの項目を調べる実験を行った．

- 誤判別率

まず，類似度を計算するにあたって用いることになる決定木が，与えられた文書を正しいレビュアーに分類することができるかどうかを確かめるために，決定木の誤判別率を求めた．その際，決定木を作成する際に用いる単語の最低出現文書数N_{df}の値を変化させ，N_{df}ごとの誤判別率を求めた．誤判別率とは，決定木により正しく判別されなかったレビュー文章数を，全てのレビュー文書数で割った値のことである．

- レビュアー同士の類似度

類似度の数値がどのような傾向で現れるかを調べるために，レビュアーの中から任意の 1 人選び，そのレビュアーを a とし, その a をユーザとして他の 29 名のレビュアーb_1~b_{29}との類似度$sim(a,b,T,N_{df})$を算出した．

5.1 誤判別率

誤判別率と決定木の本数の関係のグラフを図 2 に示す．

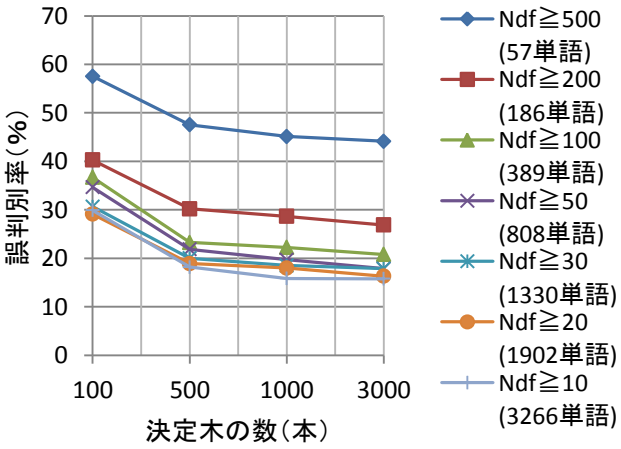

図 2　単語の出現文書数ごとの誤判別率
Figure 2　Error rate vs. number of tree.

図 2 から，決定木の本数が 100 から 500 にかけて誤判別率が 10%ほど下がっているのに対し, 500 から 3000 にかけては 1~3%程度しか変わっておらず，決定木の本数が増えるほど精度の向上率は下がっていることが分かる．また，少なくともN_{df}が 10～50 であるモデルの精度は全て 17%~18%となっており，大きな差は見られない．単語数として約 800 単語と約 3300 単語を用いて作成したモデルはほぼ同じ性能であることが読み取れる．このことから，1,500 文書程度であれば，1,500 文書のうち少なくとも 50 文書以上に出現する単語を特徴として用いれば，それ以上の多くの単語を用いるのと同程度の精度が得られると考えられる．今回の 1,500 文書に出現した全単語は 20363 単語であり，この全ての単語を Random Forest に適用しなくてもよいと考えられるため，余計な学習をせずにすむと考えられる．

また，今回一人当たり 50 レビューを用いて実験を行い，1,500 文書の 3.3%である 50 レビューに出現する単語 808 単語（全単語の約 4%）を用いれば十分な精度が得られるという結果になったが，一回りスケールを大きくして同様に誤判別率や必要な単語数を調べることで，さらにデータが大規模になった際のどの程度の学習が必要であるかを推測することができると考える．

5.2 レビュアー同士の類似度

決定木の本数は 3000 とし，類似度$sim(a, b_k, 3000, N_{df})$を算出した．$sim(a, b_k, 3000, N_{df})$において類似度が大きい順に昇順に並べ変えたグラフを図 3 に示す．また，N_{df}＝10 とし，決定木の本数 T を変化させた類似度$sim(a, b_k, T, 10)$を，算出したものを図 4 に示す．

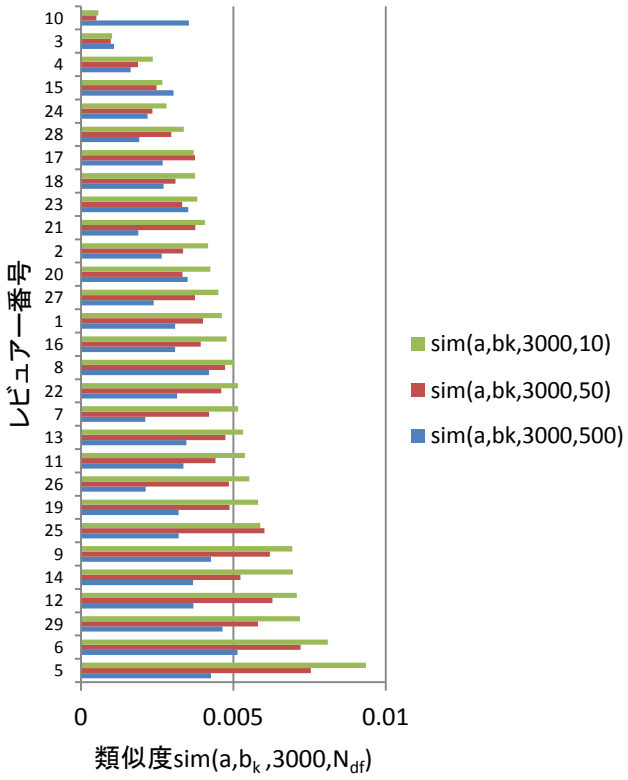

図 3　類似度$sim(a, b_k 3000, N_{df})$

Figure 3　Similarity between a and b_k：$sim(a, b_k 3000, N_{df})$

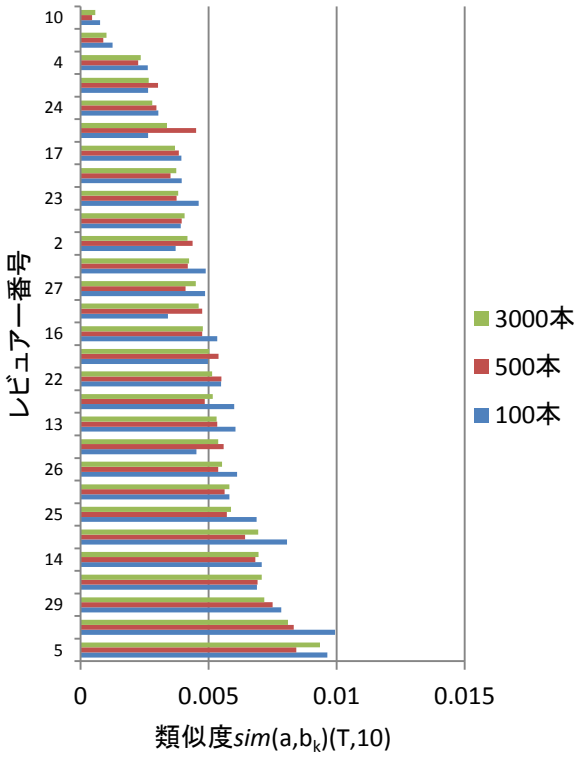

図 4　類似度$sim(a, b_k, T, 10)$

Figure 4　similarity between a and b_k：$sim(a, b_k, T, 10)$

図 3 から，誤判別率が高い場合と低い場合で類似度に大きなばらつきが現れていることが分かる．N_{df}＝10 とN_{df}＝30 の類似度は，数値の順序は全く同じわけではないが傾向は一致している．しかし，N_{df}＝500 の場合は，N_{df}＝10 の類似度が高くなっているレビュアーの部分では若干偏っているが，N_{df}＝10 との類似度に相関がほとんど表れていない．また，図 4 から，N_{df}＝10 の場合では木の本数 T が少なくても多くても類似度にほぼ同じ傾向が現れていることが分かる．これらのことから，誤判別率が低い決定木を用いた場合ではそれぞれ同様の類似度傾向が表れるというように考えられる．また，このように類似度に傾向は現れたものの，類似度の最大値が 1 であるにあるのに対し，今回算出された類似度はいずれも 0.01 以下であり，小さな値となった．今回は類似度の数値を式(3)のように全文書データの平均で定義したためである可能性がある．例えば，図 3 より a と b_5の類似度$sim(a, b_5, 3000, 10)$は約 0.009 であるが，a と b_5 の文書の類似度行列を確認したところ，類似度$prox(a_r, b_{5_c})$が 0.1 以上である文書が複数存在していた．この類似度の定義が正しいかどうかは定かではないが，類似している上位数件の文書で類似度の計算を行うなど，他の定義で類似度を求めると異なる結果になる可能性がある．

6. まとめ

ユーザに類似しているレビュアーを推薦することで，レビュー閲覧の負荷軽減を図ることができると考え，レビュ

一文章と Random Forest を用いてユーザにレビュアーを推薦する手法を検討した．今回は，レビュアーのレビューに出現する単語のうち出現頻度が上位の単語の tf-idf 値と Random Forest を用いて決定木を作成し，その分類精度とその決定木を用いてレビュアーごとの類似度を算出した．1,500 文書規模の場合，800 単語程度の特徴があればそれ以上の数の特徴を用いた場合と同等の精度が得られることが分かった．今後スケールを変えて実験を行うことにより効率的な決定木の構築ができるようにしていきたいと考えている．

また，レビュアー同士の類似度がどのように表れ，どのような傾向が現れるのかを確認できた．類似度の定義についてもう少し調査する必要はあるが，類似しているレビュアーの候補を提示することが可能になった．

今後，実際にアンケートによる主観評価実験を行い，推薦されたユーザが適切なものであるかどうか，どのような点で類似しているのかということを調査する予定である．そして，レビューの長さや単語の種類ごとの数などの素性をモデル構築に用いることでのモデルの分類精度の実験も行う予定である．そして，Random Forest でモデルを構築する際の素性の寄与率を調べ，どのような特徴が分類の要となっているのかも調査していく．

参考文献

[1] Xiaoyuan Su, Taghi M. Khoshgoftaar, “A survey of collaborative filtering techniques, ” Journal Advances in Artificial Intelligence archive Volume 2009, Article No. 4 (2009.1)
[2] 神嶌敏弘，“推薦システムのアルゴリズム(1), ”人工知能学会誌，Vol. 22，No. 6，pp.826-837 (2007.11)
[3] 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，Vol. 23，No.1，pp.89-103 (2008.1)
[4] 岡田瑞穂，藤井敦，“レビューテキスト間の類似度を用いた協調フィルタリング, ”言語処理学会第 18 回年次大会発表論文集, pp.711-714 (2012.3)
[5] 相澤彰子，“語と文書の共起に基づく特徴度の数量的表現について, ”情報処理学会論文誌, Vol. 41，No.12，pp.3332-3343 (2000.12)
[6] Leo Breiman, “Random Forests, Machine Learning, ” 45, pp.5-32 (2001.8)
[7] 波部斉，“ランダムフォレスト, ”情報処理学会研究報告 コンピュータビジョンとイメージメディア（CVIM）, Vol. 2012-CVIM-182，No.31，pp.1-8 (2012.5)