

自然言語処理を用いたレビュー分析による 因数分解モデル型推薦

岡村 航大[†] 秦野 亮[†] 西山 裕之[†]

東京理科大学 理工学研究科[†]

1 背景・目的

今日の情報化社会では、膨大な情報の集積、発信の高速化が可能である。一方で人々は情報過多により欲しい情報を見つけられない事態に直面している。このような背景から近年、価値のある情報を素早く提供し意思決定を支援する推薦システムが注目を集めている。実例として Amazon で購入する製品の35%、Netflix で視聴の80%は推薦システム経由であることが報告されている。このような電子取引 (EC) 市場規模は年々増加傾向にあり、商品ごとのフィードバックであるレビューが公開されている。総務省情報通信白書によると全体の約8割のユーザーが評価レビューを参考に売買を行っており、商品レビューは意思決定に重要であることが示されている。そこで、商品レビューはユーザーの購買体験を反映するものであり、レビューに含まれる潜在情報を抽出し、ユーザー嗜好の類似評価に組み込むことで情報推薦におけるユーザーの嗜好とアイテムの特徴との関係の推定につながると考えた。

本研究ではユーザー購買体験を反映させたレビューを推薦アルゴリズムの計算に組み込むことで EC サイトにおけるより良い推薦方法の構築を目的とする。

2 関連研究

既存の推薦システムの多くはユーザーアイテム間の評価値行列の内積計算により予測されている。Koren[1]らはアイテム間とユーザー間のベクトル内積計算により推薦を行う際、得られた評価行列のスパース性を考慮し次元削減を行うことで類似性をより明確に表現する Matrix Factorization (MF) を提案した。しかし、この手法は入力データとしてアイテムやユーザーの属性情報を使えないという問題点がある。

MF の推薦方法とデータのスパース性に起因する推薦方法の問題に対し Steffen[2]の提案した Factorization Machine (FM) はパラメータの行列を因数分解して表現することで、上記の問題の解消が可能である。また、入力情報に制限がなくこれまでの評価値だけを用いた評価計算と比べて属性情報を考慮できる。本研究ではこの FM のアルゴリズムの補助情報にユーザーの購買体験を反映させた商品レビューを用いた特徴量を加えることで推薦を向上させるモデルを提案する。この補助情報には自然言語処理 (NLP) を用いるが、これは日常で使用される文章や話し言葉における各単語の出現確率や、ある単語の後に別の単語が出現する確率を用いる分析手法である。人間のコミュニケーションは複雑であるが近年の NLP の発展によりロコミなど不規則な情報も分析することが可能になった。特に BERT[3]は深層学習手法のモデルの一つで2018年にGoogleにより発表された NLP モデルである。学習に使用することができるデータが大量に存在し、与える様々なタスクに対して転移学習を援用することで柔軟な対応が可能という特徴がある。

3 提案手法

提案手法の概要を図1に示す。まず、入手した EC サイトデータに対し前処理を施す。次に評価ユーザーアイテム情報とレビュー情報にデータを分ける。本研究では商品レビューを客の嗜好データを含む補助情報として活用する。そこで FM で扱える数値データに変換するために自然言語処理を用いる。ここでは補助情報の作成としてレビューの極性分類とトピック分類を行う。これにより得られた補助情報とユーザーとアイテム情報をもとに予測・評価を行う。

4 実験

Amazon の EC サイト商品データ (120,000 件) を用いて提案手法の実験を行う。まず初めに補助情報の作成のため商品レビューの極性分類を行い商品に対しユーザーがどのような感情を抱いてい

Factorization Machines for Recommendation using review analysis with natural language processing

[†] Tokyo University of Science, Faculty of Science and Engineering

るかポジティブ、ネガティブ、ニュートラルの3つのラベルに基づく分類器を作成した。結果は表1及び図2の混同行列で示す。

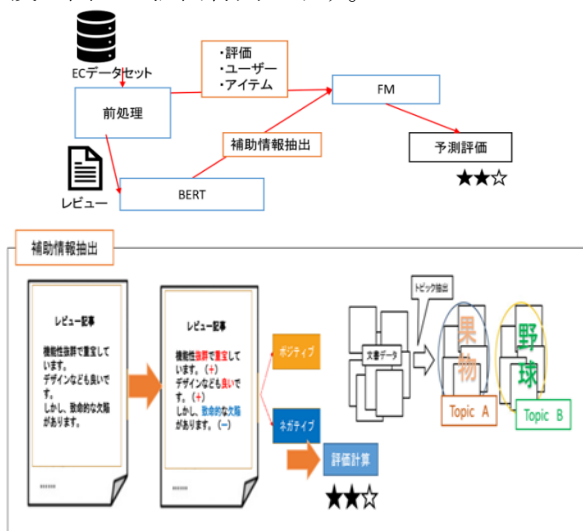


図1. 提案手法

結果からポジティブとネガティブについて90%近く判別できていることがわかる。しかしながらニュートラルについては72%となり判別が困難であることがわかる。

表1. 極性分類の実験結果

epoc	trainloss	Validationloss	Accuracy	F1
1	0.473	0.400	0.838	0.837
2	0.342	0.427	0.838	0.836

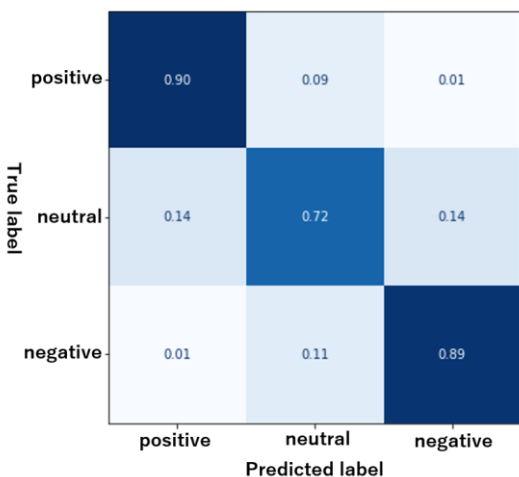


図2. 極性分類の混同行列

また、次の図3は商品レビューに対してトピック分類を行った結果であり、レビューから導き出したトピック間の距離を示す。図3を見るとapple製品と保護フィルムのような商品が近い位置にあることがわかる。この位置関係の近いものをクラスターとして扱いトピック情報とする。



図3. トピック間距離とクラスタリング

さらに、得られた補助情報と商品データを用い提案手法により予測を行った。今回は比較のためにRMSEを用いて評価する。実験結果は次の表2に示す。提案手法で最小のRMSEを算出することができた。しかしながらトピック分類のみの場合はMFや補助情報なしのFMと変わらない精度であった。これはトピックが似ているからと言って購入してくれるユーザーが同じでないことを示している。また、極性情報は商品の良し悪しに影響があるため予測誤差が小さくなるという直観を反映する結果となった。

表2. 提案手法の分析結果

手法	RMSE
MF	0.6579
FM (レビュー補助情報無し)	0.6721
FM (トピック付与)	0.6968
FM (極性付与)	0.0041
FM (極性情報及びトピック付与)	0.0037

4 今後の課題

本研究ではFMの補助情報としてレビューを用い評価予測を行ったことでユーザーが良い評価を付けそうな商品を推薦できると言える。しかし現状では嗜好をどれほど反映できるかを定量的評価できておらず、万人受けしそうな商品だけを提示する可能性も検討すべきである。今後、作成したモデルに対しランキング指標を用いた性能評価やオンライン実験を行う必要が考察される。

5 参考文献

[1] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. Computer 42.8 (2009): 30-37.
 [2] R. Steffen. Factorization machines. In: 2010 IEEE International conference on data mining. IEEE, pp. 995-1000, 2010.
 [3] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 1 (2019): 4171-4186.