

LDAの適用によるレビュー記事からの価値観モデリングに関する考察

Consideration of Personal-value Modeling from Review Articles Using LDA

子田 周平 Yu-Shen Chen 高間 康史
Shuhei Koda Yasuhumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

This paper proposes a method for constructing personal value-based models from review articles using LDA (Latent Dirichlet Allocation). While existing modeling method requires attribute-level ratings, the proposed method treats topics extracted from review texts as attributes. This paper applies the proposed method on actual reviews and discusses the possibility of the modeling.

1. はじめに

本稿では LDA(Latent Dirichlet Allocation)[上田 04] を適用し、レビュー記事から価値観モデリングを行う手法を提案する。価値観とは物事の優先順位、重み付けの体系であり。価値観モデリングとは、ユーザの価値観をモデル化し情報推薦に利用する手法である [服部 13]。属性に対する評価極性と総合評価の極性が一致する割合を評価一致率として求め、これが高い属性に対しユーザがこだわりを持つと判断する。しかし既存手法では、アイテムが持つ属性(評価属性)毎の評価値が明示的に与えられる必要があった。

そこで本稿では代表的なトピックモデルである LDA[上田 04] を用いて、レビューで言及されている評価属性を抽出する手法を提案する。提案手法ではレビューから抽出したトピックを評価属性と見なし、その極性を評価極性辞書を用いて決定する。各レビューに対応した総合評価の極性と、抽出トピック(評価属性)の極性に基づき評価一致率を計算する。映画レビューサイトである Yahoo!映画^{*1} に投稿されたレビュー集合に対して提案手法を適用し、抽出された評価属性の性質などについて考察する。

2. 関連研究

2.1 価値観モデリング

インターネット上の大量の情報からユーザが必要とする情報を取り出す技術として情報フィルタリングと呼ばれる手法が研究されている [神島 07]。その中の主要な技術の一つが情報推薦である [神島 07]。情報推薦の手法の一つとして価値観モデリングが研究されている。価値観とは物事の優先順位、重み付けの体系である。服部らは「ユーザがアイテムを評価する際に、重視する属性にユーザの価値観が反映される」という考えに基づき価値観モデリングを提案している [服部 13]。ユーザ u がアイテム i に対して行った評価において、評価が一致した回数を $O(u, j)$ 、不一致の回数を $Q(u, j)$ としたとき、属性 j の評価一致率 (Rating Matching Rate) $RMrate(u, j)$ を式 (1) として定義する。ユーザは評価一致率が高い属性にこだわりを持つと考える。

$$RMrate(u, j) = \frac{O(u, j)}{O(u, j) + Q(u, j)} \quad (1)$$

価値観をモデル化することは、ユーザの消費行動推定に役立つ。またどの属性を重視しているかをモデル化することができれば、より少ない情報からユーザへの適切な推薦を行うことが可能になると考えられる。

2.2 レビュー分析に関する自然言語処理

EC サイトには大量のレビューが投稿されており、これを解析することを目的とした評判分析に関する研究が広く行われている [乾 06]。評判分析の一種である感情分析とは、ユーザのアイテムに対する印象を抽出する手法であり、レビュー記事の内容が好評か不評かを判断する極性判定があげられる [乾 06]。文章評価の判断材料としては単語の感情極性がある [東山 08]。これは単語自体がポジティブな意味かネガティブな意味を指す指標である。

またレビュー集合からトピックを抽出するトピックモデルに関する研究も広く行われている。トピックとはテキスト内で述べられているテーマのことである。複数のテキストからなるデータセットを解析することでテキスト内のトピックをモデル化する手法が提案されている [上田 04]。代表的なトピックモデルとして LDA があげられる。LDA では一つの文書が複数のトピックを含むということを想定する。トピックの分布をディリクレ分布と仮定しギブスサンプリングなどによりトピックを推定する。

3. レビューテキストからの価値観モデリング

3.1 提案手法概要

提案手法ではレビュー集合から単語リストを作成する。これはトピックモデルを適用しやすい形に変換し、ノイズとなる不要な単語を削除するためである。得られた単語リストに LDA を適用することでトピックを抽出する。

抽出したトピックごとに、当該レビューにおける極性を判定し、属性評価とみなす。トピックに対する評価の極性と総合評価の極性を比較することで評価一致率を求める。提案手法はユーザモデル、アイテムモデル構築のどちらにも適用可能であるが、本稿では各アイテムがユーザー(レビュアー)にどのように評価されているかを表すアイテムモデルを対象とする。

連絡先: 高間康史, 首都大学東京大学院システムデザイン研究科, 〒191-0065, 東京都日野市旭が丘 6-6, ytakama@tmu.ac.jp

*1 <http://movies.yahoo.co.jp>

3.2 レビューからのトピック抽出

アイテム別のレビュー集合を処理対象とする。各レビューを形態素解析し、単語リストに変換する。形態素解析には MeCab を利用した。単語リストに含まれる単語は形容詞、動詞、名詞のみとし、接頭辞、接尾辞、数詞、記号といった単語は削除する。また「ない」「する」といった登場頻度が著しく高い単語は stop-words として削除する。

得られた単語リストを単語と TF-IDF 特徴量の組の集合からなるコーパスに変換し、LDA モデルを適用することでトピックを抽出する。トピック数は 20 個とした。トピックモデルの抽出にはライブラリとして gensim を利用した。各トピックについて、生起確率の上位 10 単語がその確率とともに出力され、これを代表語とする。

3.3 トピックごとの極性判定

抽出された各トピックについて極性を判定する。各トピックの代表語に係る単語の極性値の合計をトピックの極性とする。係り受け解析には CaboCha、辞書は感情極性辞書 [東山 08] を使用した。感情極性辞書は単語とその単語の意味がネガティブなのかポジティブなのかという感情極性情報を含んでいる。提案手法ではネガティブな単語の極性値を -1、ポジティブな単語の極性値を +1 としてトピックの極性判定をする。

3.4 評価一致率の算出

各トピックについて、それが含まれるレビューに対する総合評価の極性と、そのレビューにおけるトピックの極性に基づき式 (1) を用いて計算する。ここでは属性評価があるレビューのうち、総合評価が一致しているレビューの割合を評価一致率とする。なお総合評価が 3 以上かつトピックの極性値が 0 より大きいとき好評一致、総合評価が 3 未満かつトピックの極性値が 0 より小さいとき不評一致とする。トピックの極性値が 0 だった場合は好評、不評のどちらでも無いと見なし、評価一致率の計算には含めない。

4. 実験

本稿では以下の二点について明らかにするために実験を行った。

- レビュー集合に LDA を適用することでトピックを評価属性として抽出することができるか
- レビュー集合から価値観モデルを構築できるか

映画レビューサイトである Yahoo!映画から抽出したレビューデータをデータセットとする。各レビューは 5 段階の総合評価、およびレビュー本文からなる。各アイテムに対するレビュー集合に提案手法を適用し、トピックとそれぞれの評価一致率を求める。このとき、映画に関する単語で MeCab に登録されていない単語、「3D」「2D」などはユーザ辞書に登録して解析を行った。

映画「アバター」のレビューにトピック数 20 として提案手法を適用し、抽出されたトピックを表 1 に示す。表には、各トピックの評価一致率、好評あるいは不評の場合のみについて計算した評価一致率、生起確率上位 5 件の単語を示している。「アバター」の総レビュー数は 3040 件、評価ごとのレビュー数は低評価のものからそれぞれ 228 件、352 件、576 件、920 件、964 件であった。

評価一致率は全てのトピックにおいて 0.69 を超えている。一般に、評価一致率は 0.7 以上がこだわりの強い属性の目安とされており [服部 13]、抽出された殆どの属性についてこだわ

りが強いと判断できる。これはユーザがレビュー記事を書く際、こだわりのあるトピックだけについて言及する傾向にあるためと考える。また評価一致率の高いトピックに「3D」「映像」など映像の美麗さに関する単語が出現する傾向にあることがわかる。「アバター」は映像の美しさや 3D 効果などで話題になったため、この結果は妥当と考える。一方「ストーリー」「中だるみ」といった脚本に関する単語も出現しているが、映像関連のトピックに比べると少ないことがわかる。

好評・不評の評価一致率を比較すると、好評の場合が 10 倍以上高い値となっている。これは好評のレビューの件数が多いこと、ストーリー上の様々な点に対して言及される傾向にあったため、ネガティブな意見はトピックとして抽出されなかったと考える。

表 1: 抽出したトピックの評価一致率と生起確率上位の単語

| 順位 | 評価一致率 | | | 生起確率上位の単語 |
|----|-------|-------|-------|----------------------|
| | 総合 | 好評 | 不評 | |
| 1 | 0.756 | 0.716 | 0.049 | 一言、良い、スカッ、ハンター、モンスター |
| 2 | 0.751 | 0.712 | 0.039 | 見る、驚かす、世界、対応、無駄 |
| 3 | 0.750 | 0.713 | 0.037 | 見る、すごい、CG、観る、映像 |
| 4 | 0.750 | 0.711 | 0.039 | 観る、ストーリー、見る、映像、3D |
| 5 | 0.746 | 0.707 | 0.039 | すごい、見る、観る、3D、映像 |
| 6 | 0.745 | 0.710 | 0.035 | 観る、見る、楽しめる、3D、映像 |
| 7 | 0.742 | 0.703 | 0.039 | 観る、作品、3D、ある、見る |
| 8 | 0.736 | 0.692 | 0.044 | サウンド、吹替え、最高、観る、見る |
| 9 | 0.735 | 0.699 | 0.036 | メガネ、見る、3D、字幕、映像 |
| 10 | 0.735 | 0.694 | 0.041 | めがね、見る、内容、3D、凄い |
| 11 | 0.735 | 0.689 | 0.046 | 観る、アメリカ、綺麗、映像、思う |
| 12 | 0.731 | 0.695 | 0.036 | 癖、満足、中だるみ、観る、ながい |
| 13 | 0.731 | 0.672 | 0.059 | 得、見る、楽しむ、中間、ストーリー |
| 14 | 0.727 | 0.683 | 0.044 | 観る、凄い、良い、作品、思う |
| 15 | 0.721 | 0.679 | 0.042 | 評判、可、不可、綺麗、見る |
| 16 | 0.709 | 0.666 | 0.043 | リアリティ、間延び、低い、見る、展開 |
| 17 | 0.709 | 0.660 | 0.049 | 見る、大佐、アバター、アメリカ、作品 |
| 18 | 0.703 | 0.658 | 0.045 | 感じ、観る、見る、あまり、期待 |
| 19 | 0.701 | 0.682 | 0.019 | 遊園、ナウ、シカ、のける、つらい |
| 20 | 0.695 | 0.643 | 0.052 | 一見、観る、作品、心情、愛 |

5. おわりに

本稿では LDA を適用し、レビュー記事から価値観モデリングを行う手法を提案した。実際のレビュー記事に提案手法を適用した結果、評価属性に該当するトピックを抽出可能であることを示した。これにより、属性別評価が与えられていないレビュー集合からも価値観モデルを構築することが期待できる。今後は抽出した評価属性に基づくアイテム推薦の可能性について検証することや、ネガティブな意見に対応する評価属性の抽出方法について検討することが課題として挙げられる。

参考文献

- [神鷹 07] 神鷹敏弘, 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.
- [服部 13] 服部俊一, 高間康史, 価値観に基づくユーザモデルを用いた情報推薦手法に関する検討, 第 3 回インタラクティブ情報アクセスと可視化マイニング研究会, pp.1-6, 2013.
- [上田 04] 上田修功, 齋藤和巳, 多重トピックテキストの確率モデル-テキストモデル研究の最新線-, 情報処理, Vol.45, No.2, pp.184-190, 2004.
- [乾 06] 乾孝司, 奥村学, テキストを対象とした評価情報の分析に関する研究動向, 自然言語処理, Vol.13, No.3, pp.201-241, 2006.
- [東山 08] 東山昌彦, 乾健太郎, 松本裕治, 述語の選択好性に着目した名詞評価極性の獲得, 言語処理学会第 14 回年次大会, pp.584-587, 2008.