4K-06

レビューデータにおける推薦のための属性抽出

安部小百合† 小林一郎‡
†お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻情報科学コース
‡お茶の水女子大学 基幹研究院自然科学系

1 はじめに

レビューデータは商品選択の際に有用な情報となるが，膨大な量がある場合すべてに目を通すことは不可能で，その解決のために様々な試みが行われている．本研究では推薦に着目した．レビューデータにおいて推薦をする際，ユーザを特定する情報を取得していない場合，単一データのみがそのユーザの情報となり，評価や属性等の限られた情報を用いた類似度の算出となる．推薦の際に使用できる各レビューのユーザや商品の属性について検討し，属性の抽出を行う．

2 化粧品レビューの推薦

ユーザがレビューを探す際の目的として，ある化粧品について情報収集をする，新しい化粧品を探す，使用している化粧品に合う化粧品を探す，化粧品の使い方を調べる等がある．本稿はこのような様々な目的を満たす推薦システムの構築を目指す．

2.1 使用データ

今回使用したデータは，@cosme(アットコスメ)[3] のレビュー文である．アットコスメは日本最大級の化粧品レビューサイトである．各レビューは商品ごとに書かれており，商品にはそれぞれブランドとメーカーがある．各ブランドの商品は一つとは限らず，様々なアイテムの商品を複数持つことが多い．アットコスメのデータにおいては，それらは商品 ID・商品名で分類される．

2.2 データの前処理

レビューにおけるユーザに関する素性として，肌質，年齢，購入場所，ブランド名を使用する．また，情報抽出によりレビュー文から得られる情報も素性として加える．

本研究でのレビュー同士の関連性算出におけるベースラインとして，ブランド名・肌質・購入場所・年齢のコサイン類似度で算出したものを設定する．

ユーザの類似度を通じてレビュー同士の関係性を可視化して示す．可視化はオープンソースライブラリである arbor.js[4] を使用した．可視化は各レビューをノードとし，類似度を持つノード同士に無向エッジを追加した．

3 実験

3.1 属性に基づくユーザの類似度

アットコスメのデータ中カテゴリ名洗顔料，2010 年 2 月 1 日から 2011 年 1 月 31 日までの 8790 件を用いた．

その中からランダムに 20 件のレビューを抽出し，グラフの構築を行った．グラフを構築した結果を図 1 に示す．

図 1: 類似度の可視化結果

3.2 レビュー文書の情報付加

次に，レビュー文書から情報獲得し，ユーザ属性に加えることで類似度の計算をする．

文書内におけるブランド名の参照を使用するために，ブランド名辞書を用いてレビュー文書中に他のブランド名が出現するか否かを調べた．

ブランド名を抽出するためにブランド名を登録した辞書を構築する．多くのブランド名には表記の揺れが存在し，化粧品ブランドにおいても略語や表記揺れが多くみられる．例えば，”デジャヴュ”というブランド名を”デジャブ”，”デジャヴ”といった英単語をカタカナ

An Approach to Attribute Extraction from Review Data for Product Recommendation
†Sayuri ANBE(anbe.sayuri@is.ocha.ac.jp),
‡Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.

に直した際の日本語特有の表記揺れもある．他にも，ー（ハイフン）－（ダッシュ）といった線が横に伸びている記号は見た目が同じでもコンピュータでの検索では合致しない．よってこれらを正規化するためのパターンを作成した．

ブランド名辞書の有効性の確認のため，レビューにおける他のブランド名の出現数を確認する実験を行った．本研究では同アイテム間での検索を行った．これらの処理の結果，検索において一致するブランド名が微小であるが増加した．

また，ブランド名には表記ゆれだけでなく略語や別称もある．(例：”マジョリカマジョルカ”→”マジョマジョ”)

正規化パターンの拡張だけでなく略語パターンや人手の構築による辞書拡張は今後の課題となる．

3.3 文書データ中の情報を追加

ベースラインとしたデータに他のブランド名が出現するか否かのデータを付与する．文書中に他のブランドが出現しているということは，ブランドの比較等を考えているユーザであるとみなすことができ，それらのレビューに類似性があると考えられる．他のブランド名が出現する場合はダミー変数の該当のブランド名に 1 を加算することで重みをつける．

3.4 考察

類似しているとされたクチコミ本文ノードの内容を見たところ，ある程度類似している内容のものもあるが，全く違う内容のものもあった．また，データ数に偏りがあり，多い年齢層，数の多いブランド等のデータ同士のみエッジが張られ，数の少ないデータはノードが孤立してしまうケースも見られた．ブランド名情報を加味したグラフに特に変化はみられなかった．カテゴリ名洗顔料において全体の中でのブランド名辞書中の他のブランド名の出現は 21%であり，20 件という少ないデータ中では出現しなかったと思われる．

4 評価視点の抽出

前項まではユーザ情報における類似度を算出したが，レビューによって内容は異なっており，ユーザの求めている情報を得られているとは言い難い結果となった．よって，ユーザの興味に基づいた推薦をするために評価視点の抽出を行う．以下に例を示す．

レビュー	評価視点
値段 は少し高いかも	値段
洗い上がり はつっぱります	洗い上がり

アイテム名洗顔料のレビュー 100 件から人手で評価視点の抽出を行った．評価視点は一単語とは限らず，洗い ＋ 上がり のように複数の単語から成る句で構成されるものもあるとする．また，単一のレビュー文書には複数の文があり，評価視点もレビューごとに複数存在する．

評価視点には，洗い上がり と 洗いあがり，にきび と ニキビ のように読みが同じものや 値段 と 価格 と 安さ のように評価視点として同義の語彙がある．これらを同一のものとして扱う処理を行う．

抽出された評価視点の中で読みが同一のものを辞書を用いて統一した．本稿では IPA 辞書 [5] を用いた．さらに，評価視点として同義であるものを人手で分類した．

その結果，評価視点は 157 件，読みで統一した結果 144 件となった．同義であるものを分類した結果 52 件となった．誤りとして，読みが同じ中で 高価 と 効果 等意味は違うが読みが同じものが同一とされていた．

5 おわりに

本稿では化粧品レビューの推薦に用いる属性について検討した．年齢肌質等のユーザデータのみでなくレビュー内におけるブランドや評価表現の出現，またそれらの抽出を行った．

ブランド名を用いた情報は出現数が少なく，各レビューごとでは得られる情報が限られていることがわかった．ブランド名を用いるには同一製品の他の情報も活用することを検討したい．

評価視点の抽出ではブランド名と比較してレビュー文書ごとに出現する数が多く，また分類した結果 52 件であったため比較的少なく，レビュー推薦において有用である可能性が高いと考えられる．

今回抽出した評価視点等をアイテムごとに活用することで推薦の効率の向上を検討したい．

謝辞

本研究では，株式会社アイスタイル様よりデータを提供していただきました．ここに感謝の意を表します．

参考文献

[1] Feldman R, Fresko M, Goldenberg J, Netzer O, Ungar L, *Analyzing Product Comparisons on Discussion Boards, Language, Culture, Computation.* Computing - Theory and Technology, pp 399-408, 2014.
[2] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, et al. *Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis.* SIGIR, 2014.
[3] http://www.cosme.net/
[4] http://arborjs.org/
[5] https://osdn.jp/projects/ipadic/

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.