レビュー文書の分析に基づく商品推薦手法の開発

理学専攻 情報科学コース
安部小百合

1 はじめに

近年，CGM(Consumer Generated Media) の発達により一般ユーザの意見が活発に発信されるようになった．レビュー文には多くの評判情報が含まれている．中にはその商品のみの評判情報だけでなく，他の商品との比較や同時使用に関する情報が書かれている場合も多くあり，商品を選択し購入する大きな動機づけになっている．しかし，これらの文書量は膨大であり，全てに目を通して判断することは時間や労力の面から不可能である．この問題解決として，レビュー文の分類や情報抽出，要約の手法等が活発に研究されている．[1][2]

本研究では情報の抽出や分類に基づくレビューの推薦に着目をした．ユーザがレビューを探す際に推薦されるレビューは複数の目的に応じて変更されるべきだと考えられる．このことから，目的に応じて推薦における特徴量の重みが変化する手法が望まれる．

以上を踏まえて，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を収集する，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)[1] 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 データの分析

カテゴリ名が洗顔料の 8,790 件のレビューについて，分析を行った．年齢は最年少が 7 歳で最年長が 63 歳，平均値が 28.371 歳である．ブランドに対する商品の数は最小が 63 件，最大 2038 件であり，ブランドにより大きく偏りがあることがわかる．クチコミ件数は多くのユーザが 1 件程度だが，2,000 件以上を投稿しているヘビーユーザーもいる．購入場所については最小が訪問販売の 25 件，最多のものが通販化粧品・コスメであり，これはインターネット環境でのレビュー収集であるという特性によるものであると考えられる．肌質は最小がアトピーの 204 件であるのに対し最大が混合肌の 3,386 件であり，こちらも大きく偏りがある．

2.3 属性に基づくユーザ類似度の発見

レビューにおけるユーザに関する属性として，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属性として加える．

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，上記ユーザ属性のコサイン類似度で算出したものを設定する．これによりレビューの関連性を捉える．

3 実験

3.1 ユーザの類似度に基づくレビュー同士の関係性の可視化

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[2] 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．

データとして，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，グラフの構築による可視化を行った．可視化の結果を図 1 に示す．

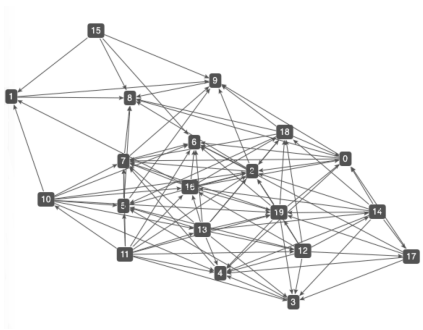

図 1: レビューの関係性の可視化結果

3.2 ブランド名辞書による表記ゆれの解消

文書内におけるブランド名の参照情報を使用するために，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

[1] http://www.cosme.net/

[2] http://arborjs.org/

合致しない．よってこれらを正規化するため以下のパターンを作成した．

- カタカナ・ひらがな＋濁点，半濁点が 2 文字の場合は一文字に統一
  例：ガ゛→ガ　か゛→が　ハ゜→パ
- ヴュ，ヴ，ブはブに統一
- カタカナ・英字間の中点やビュレットは削除
  例：マリ・クレール → マリクレール
- ひらがなおよびカタカナ間のスペースは削除
  例：マリ　クレール → マリクレール
- ハイフン，マイナス，ダッシュ類は全角長音符に統一
- 英字は小文字に統一

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．本研究では同一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わずかであるが増加した．

アイテム名「乳液・美容液」などで正規化した結果を以下の表に示す．

レビュー数	ブランド数	前	後	増加率
11,309 件	15	3,436	3,451	0.31%

3.3 文書データ中の情報に基づく関連性の発見

ブランド名辞書の構築においては，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が該当するブランド名に 1 を加算することで重みをつける．

3.4 考察

3.1 で構築したグラフにおいて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[3] 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．

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．

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の活用として、同一アイテムの評価視点を各レビューの属性に加味する手法を用いることで推薦精度の向上を検討したい．

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] https://osdn.jp/projects/ipadic/