ブランド名に基づく化粧品レビュー文書からの情報抽出への取り組み

安部 小百合　(指導教員：小林 一郎)

1　はじめに

近年，インターネットにおける口コミサイト，SNS，ブログに例えられる CGM(Consumer Generated Media) の発展により，ユーザの意見が多数発信されるようになった．CGM は消費者の声を多数有しており，特に商品に関する意見や評判の文書データは他の消費者の商品選択の際に有益な情報となりうる．しかし，これら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ため，効率的に情報収集を自動化をできる解決法が求められている．この問題へのアプローチとして，レビュー文書の分類や情報抽出，要約の手法等が活発に研究されている．

本研究では化粧品レビュー文書からの情報抽出に着目する．化粧品における商品のレビュー文には，商品間の比較や併用，ブランドの転向等が，様々な状況下で評価が現れている．化粧品のブランド間の関係を特定の条件の下で判断することができれば，ユーザに対し，適切な化粧品を推薦できる知識の抽出へとつながる．まずは，実際の化粧品レビュー文書を用いて化粧品の特徴に基づいたブランド間の関係を抽出する基礎的な検討を行う．

2　化粧品使用における特徴

商品としての化粧品には以下の特徴がある．

1. 数ヶ月単位で購入し続けるもの
2. 消耗品であり，安価である
3. 他の化粧品に乗り換えるハードルが低い
4. 同時に多くの化粧品を使用する
5. 組み合わせに相性がある

これらの特徴を加味して化粧品に特化した分析をすることにより，一般的な商品と違った化粧品独自の評価基準に関する情報を得られる可能性がある．本研究では，特に他のブランド名が出現するレビュー文における対象ブランドと他のブランドの関係に着目する．

本研究では，化粧品レビュー文書からの情報抽出へのアプローチを 2 種類考察する．

3　実験準備

今回使用したデータは，化粧品レビューサイトアットコスメ [7] のレビュー文 69,437 件である．データには，アイテム名，ブランド名，商品名，レビュー文章，おすすめ度，投稿ユーザの属性等が含まれている．この中で，ブランド名とレビュー文章を使用した．

初めにブランド名の項目から 93 件のブランド名が含まれる辞書を作成した．これを用いてレビュー文章全体から他のブランド名が含まれる文章を抽出したところ，8606 件のデータが得られた．

4　分類器による抽出手法

機械学習を用いて文書の分類を行う．

4.1　概要

他のブランド名が出現するレビュー文における関係は大きく以下の 2 つに分けることができると仮定する．

1. **比較：** 購入を検討する際に比較した物や使用したことがあり品質や色み，香り等を比べているもの
   例：「今まではヘレナルビンスタインのマスカラを使っていました.」
   「私にはランコムのほうが合っていました.」
   「シャネルのものと迷いましたが，こちらにしました.」
2. **併用：** 下地のように重ねて使ったもの，メイク落としとメイク用品のように併用して使われる関係にあるもの
   例：「DHC の下地を塗ってからこれを使っています.」
   「クレンジングはクリニークを使うとよく落ちます.」

上記，2 つの関係性を持つ文書にはそれぞれ違った特徴があると考えられるため，それぞれのカテゴリの文書から商品推薦に有益な情報を抽出するアプローチは異なると考えられる．本手法はまずこれら 2 つのカテゴリ分類に取り組む．

4.2　分類手法

機械学習を用いた手法には，ナイーブベイズ分類器，決定木，ニューラルネットワーク等が挙げられる．本研究ではサポートベクターマシン (SVM) とナイーブベイズ分類器を使用して文書を分類する．それぞれ教師あり学習のアルゴリズムであり，SVM は与えられたデータの超平面を生成することで 2 値分類を行い，ナイーブベイズ分類器はベイズの定理を独立性を仮定して用いることで分類を行う．ここでは SVM のライブラリとして LIBSVM[5] を用いた．

4.3　素性

ここでは素性として，Gamon[4] のアプローチと同じく単語の頻度ベクトルを用いる．また，分類の際一般的に精度に寄与しないとされる副詞や助詞を除き，さらに品詞ごとの特徴を把握するために名詞，動詞，形容詞それぞれを素性にした場合の実験を行う．形態素解析には形態素解析器 MeCab[6] を用いる．文章は区切り文字によって区切られ，他のブランドが出現する文のみを使用した．上記の素性により分類器を構築し，それらの精度を検証するために評価実験を行った．

4.4　実験

4.4.1　実験仕様

実験準備により得られた他のブランド名を含むレビュー文書を用いた．ランダムに抜き出した 400 件をあらかじめ人手で 2 つのブランド間の関係を「比較」または「併用」でラベル付けした．比較と分類されたものは 276 件，併用と分類されたものは 118 件であった．この 2 カテゴリに該当しなかったものとしては，購

入した店舗の名前がブランド名と同一だったものや，商品名の一部に他のブランド名が含まれているものであった．

データの数に偏りがあるため，比較ラベルのデータを 118 件に減らし，計 336 件のデータで学習と評価を行った．評価は 2 分割交差検定を用いて行った．

4.4.2 実験結果

表 1 に実験結果を示す．それぞれの項目は，形態素解析した単語の中から素性として使われた単語の品詞である．F 値とは，適合率と再現率の調和平均によって求められる尺度である．

表 1: 実験結果 (数字は F 値)

品詞	SVM	ナイーブベイズ
全て	0.7801	-
名詞	0.8095	0.6788
形容詞	0.2972	0.8098
動詞	0.8020	0.8235
名詞＋動詞＋形容詞	0.8400	0.7292

4.4.3 考察

実験の結果，分類器によって品詞ごとにばらつきがみられることがわかった．また，動詞に関してともに高い精度を示した．今回データが少ないこと，レビュー文は略語や表記のゆれが多く，データ自体にノイズが多いことから，データを増やすこと，次元削減に取り組むなどでノイズを減らすことでさらに精度の向上が見込まれる．

5 ルールベースによる抽出手法

佐藤ら [2] は “[名詞] の方が”, “[名詞] より”のような比較特有表現のルールを人手で作成し，blog から比較関係情報を抽出している．本研究では，化粧品の “数ヶ月単位で購入し続ける”,“他の化粧品に乗り換えるハードルが低い”という特徴に着目した．ユーザはレビュー文書の対象の化粧品以外にも使用した経験や，購入予定がある場合がほとんどであり,「前は…を使っていました」「現在は…」「次は…を使おうと思います」等の表現を用いている．それらを前後表現と呼ぶこととし，抽出するための新たなルールを作成した．

5.1 実験

実験準備により得られた文章より，他のブランド名の比較表現を含んでいる文書を取得する．クエリとしてブランド名リストと各表現の組を全て作成した．以下にブランド名 (NE) の比較関係を抽出するクエリを列挙する．

5.2 検索クエリー覧

- 比較特有表現
  NE より，NE のほうが，NE のほうは，NE と比べ，NE に比較
  より NE，よりも NE，比べ NE，比べて NE
- 前後表現
  今は NE，以前は NE，前は NE，最近 NE，次は NE，普段は NE，現在 NE

5.3 実験結果

データ 8606 件中，比較特有表現のみを用いた場合 271 件，前後表現のみを使用した場合 465 件の文章を取得することができた．また，比較特有表現と前後表現をともに用いた場合，731 件の文章が得られた．比較特有表現，前後表現どちらも有している文章は 5 件のみであった．

比較/併用に人手で分類した 400 件のデータにこれを適用したところ，比較表現のみの場合 14 件，前後表現のみの場合 20 件，どちらも有している文書は存在しなかった．また，前後表現を用いた抽出の中に “併用”ラベルが付いているものがあったが，以下のように併用している状況を説明しているものであった．

> 普段は DHC のまつげ美容液を塗ってからこれを重ねてます.

5.4 考察

比較特有表現に加えて前後表現を用いることで，比較関係にあるブランドをより多く取得することができることがわかった．また，抽出された文章が比較ラベルのものであるとは限らないということがわかった．

6 おわりに

他のブランド名が出現する化粧品レビュー文において，2 種類のアプローチを用いてブランド名同士の関係の抽出に取り組んだ．

今後取り組むそれぞれの手法の課題を挙げる．機械学習を用いたアプローチでは現状比較/併用の 2 クラスタにのみ分類をしており，より詳細な情報を得るために多くのクラスタへ分類することが望まれる．ルールベースのアプローチでは，今回は有差の比較特有表現，前後表現を用いた．さらにルールを強化するために，同等・最上のルールを組み込んでいく．また，係り受け解析を用いたルールも適用することでより多くの情報を抽出する．最終的には 2 つの手法を統合する必要があると考えている．中山ら [1] の手法のようにルールを分類器の素性として組み込む方法を考えている．

参考文献

[1] 中山 祐輝，藤井 敦，レビューテキストを対象とした評価条件の抽出手法，言語処理学会第 19 回年次大会，A4-2，2013.

[2] 佐藤敏紀, 奥村学. blog からの比較関係抽出. 情報処理学会自然言語処理研究会, pp.7-14, 2007

[3] 大塚 裕子，乾 孝司，奥村学，意見分析エンジン-計算言語学と社会学の接点-，コロナ社，2007.

[4] Gamon, M., “Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis.” In Proceedings of the 20th International Conference on Computational Linguistics, 2004.

[5] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[6] http://code.google.com/p/mecab/

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