レビュー文書の比較に基づく記述対象の把握への取り組み

坂梨 優 [a]　　小林 一郎 [b]

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

a) *sakanashi.yu@is.ocha.ac.jp*　b) *koba@is.ocha.ac.jp*

概要　ブログ，口コミサイトなどの普及により消費者から商品の評価に関する有益な情報が提供される機会が増えている．本来，自分が商品のレビュー全てに目を通し，どの商品が自分が本当に欲するものかを検討できることが望ましいが，レビュー文全てに目を通すことは困難であり，また，自分が知りたい観点から商品ごとに比較を行うことは難しい．もし，知りたい観点について書かれた情報を自動的に抽出し，比較することができれば，沢山の商品から自分に合ったものを選ぶ際の助けになる．本研究では化粧品のレビュー文を対象にし，LDA-DF を用いて制約知識を入れたトピック分類をした後，商品ごとにレビュー文の比較をする手法の提案を行う．

キーワード　ディリクレ森分布，制約知識，レビュー文書

1　はじめに

近年, ブログや口コミサイトなどの Web サービスの増加により，Web 上で商品に対する意見を発信したり，他人の口コミを読んで商品を購入する際の参考にする消費者が増加している．その数は膨大であり，全てを読み比べることは不可能である．口コミ情報は消費者が商品を選ぶ際だけでなく，商品を生産，販売している企業にとっても重要なものであり，口コミ情報の収集したり，評価情報を抽出のため，多くの研究がおこなわれている．本研究では，潜在トピックでレビュー文の分類を行った後に，同一トピック内で商品ごとに文を分類し，さらに文の持つ特徴に基づいて，比較する文同士を決定する比較手法を提案する．特に潜在トピックで分類する際に，商品を選ぶ消費者の視点に基づきトピックを分類するために，トピック分類に対する制約知識を導入したディリクレ森分布を用いた潜在的なディリクレ配分法 (LDA)[1] を導入する．

2　提案手法

ユーザーの視点から口コミ文書を解釈するため，ディリクレ森分布を用いてトピック抽出を行い，そのトピックの下で比較を行う．使用する化粧品のレビュー文書から抽出されるトピックはある程度予想することができ，また，ある決まったおおまかな視点でトピック分類ができることが望ましいと考えたため，同じトピックに入ることが妥当と考えられる単語を制約知識として導入できるディリクレ森分布を使用する．また，制約知識の数を潜在的なトピックの数と限定することなく，その他の潜在的なトピックの抽出も期待し，トピック数は制約知識で構築したグループの数より多く設定する．

2.1　ディリクレ森分布を用いた LDA

LDA を利用し，制約を組み込むことで潜在的なトピックの分類を行うために，ディリクレ森分布 [2] を用いる．ディリクレ森分布とはディリクレ分布を階層化したものであり，これにより，LDA により同じトピックに入る単語を制御することが可能となる．

2.2　与える制約知識の決定

制約知識は，消費者が商品を選ぶ視点に基づき人手により用意し構成する．与える制約が本文中にない場合，日本語 WordNet[3] により制約単語と文書中の単語との類似度を測り，用意した制約単語との類似度が最も高い語を本文中から探し，置き換えて制約単語とする．

3　実験

3.1　使用データ

対象とするレビュー文書に株式会社アイスタイルの化粧品クチコミサイト@cosme のレビュー文書を用いる．
期間：2010 年 2 月 1 日から 2011 年 1 月 31 日
商品：期間内の上位 20 位以内にランキングされた商品
カテゴリ：口紅・グロス・リップライナー
文数：24,037（文書数：2,800）

3.2　制約知識

今回与えた制約は表 1 に示す 6 つのグループとなる．

表 1 制約知識で構成したグループ

トピックのグループ	トピックを構成する単語
色	(色，発色，肌)
ツヤ感	(ツヤ，潤い)
乾燥	(荒れる，乾燥，剥ける，皮)
持ち	(持ち，時間，食事，落ちる)
ラメ・パール感	(ラメ，パール)
香り	(香り，匂い)

表 3　実験結果

商品 A	商品 B	商品 C	商品 D
• グロスを使えば問題ないのですが… 私は唇が非常に荒れやすく、リップクリームも市販のものもほぼ全滅なのですが、これは荒れずに使えてます。 • リップクリームを下地に付けても厳しい（すぐにリップクリーム塗りたくなる） 付けると皮がむけたり、唇が痒くなったり。 • 私これだけ塗ると凄い縦じわが目立つので、ベースをしっかり整えるか、グロスをたっぷり塗るかしないと、汚い仕上がりに。	• 何を使っても荒れてしょうがなかった口紅もグロスも コレさえ下地にしとけば、唇の荒れとは無縁に快適で使える。 • この保湿力のおかげで、乾燥した唇にでも綺麗にのりますし、一塗りでぷっくりリップが簡単に作れます。 • 敏感肌で唇がすぐあれるのでいつもリップはほとんどつけていませんでしたがこのグロスは平気でした。	• 口紅は皮剥けするし、縦皺ができちゃうのに、これは塗りやすくて時間たってもぷっくりうるうるです。 • チップが使いやすく口角まで簡単に綺麗にラインがとれ，けばけばも柔らかくて全く刺激がありません。 • 今までは他社のリップグロスを付けていて皮ムケがひどかったのですが、こちらは皮ムケしない。	• 下にリップクリーム、上にはうるおい成分のあるグロスをつければなんとか乾燥は防げますが皮むけは防げません。 • でもちょっと縦シワが目立ちます … なのでたっぷーりリップクリームを塗ったりして、唇を万全の体調にしないとつけれないのが難点。 • 使ってないけど (笑) 唇のコンディションが悪い日は、皮が剥けて唇がブヨブヨに白くなりました (笑) だから日頃から商品 X でケアしてます。

3.3　潜在的なトピックの推定

各文のトピックを推定するとして，口紅・グロス・リップライナーのカテゴリ内のレビュー文書を全文書集合とする．これにディリクレ森分布を使用した LDA を適用することによりレビュー文書内に含まれるトピックを抽出する．用意した制約のグループを 6 つ用意し，潜在的なトピックも抽出できるよう，トピック数を 10 とする．ディリクレ森分布を使用した LDA によるトピック推定の後，トピックごとに文におけるトピック分布の値が高い順にソートし，各トピックについて商品ごとに分類し，比較を行った．

4　結果と考察

4.1　実験結果

各トピックの上位 10 単語を表 2 に示す．

表 2 各トピックを構成する出現確率上位 10 単語

トピック 0	トピック 1	トピック 6	トピック 7	トピック 9
唇	口紅	色	ピンク	良い
荒れる	好き	唇	ベージュ	持つ
リップ	重ねる	発色	色	発色
皮	香り	良い	パール	いい
乾燥	感じ	肌	オレンジ	落ちる
口紅	マット	感じ	青み	色
つける	使う	赤み	強い	ストック
クリーム	色	ピンク	感じる	ピーチ
保	発色	いい	ラメ	リピ
剥ける	いい	見える	赤	よい

表 3 に，制約乾燥グループの単語を含むトピックに分類された口コミ文を商品ごとに示す．

4.2　考察

表 2 はディリクレ森分布を用いた LDA により抽出された各トピックを構成する上位 10 単語である．予め用意した制約知識で同じグループとした語彙が，トピックの構成に反映されているのがわかる．しかし，人手で構成した制約では，文書内の出現回数にばらつきがあるため，文書中に含まれる制約知識とした単語の数が少ない場合，与えた制約が効かない可能性があると考えられる．トピックを構成しやすい単語を同じグループに設定するには，共起する単語で制約知識を作るなどの方法も考えられる．

また，トピック数は用意したグループより少し多い，10 と設定したが，通常の LDA で予備実験した際の，パープレキシティを指標とした最適なトピック数は 19 であり，トピック数の設定方法も検討すべきである．

表 3 では，3 つ目の制約乾燥グループの単語を含むトピックに分類された文を示す．商品 B，C では保湿力があり，唇が荒れることなく潤い，一方商品 A，D では唇が荒れやすくなるという内容の文がみられた．また他の発色グループを構成する単語が表れたトピックでは，商品 A は色が薄く B はほんのり色づき，商品 C は肌の色に馴染む，そして商品 D は発色がよいという内容が示された．

ディリクレ森分布を用いた LDA により，ある決まった視点でのおおまかなトピック分類ができたと言える．

5　まとめと今後の課題

本研究では，レビュー文の決まった視点での分類を行うため，ディリクレ森分布を用いた LDA によって，制約知識を組み込んだトピック抽出を行った．

今後はトピック内で，文章中に現れる評価情報や，商品を特徴づけている言語的特徴で文同士の比較を行う．具体的には，文の極性判定により意見を分析し評価情報を抽出する，オノマトペによりトピック内で商品がどのような特徴を持っているかを判定するなど，商品の比較を行いたいと考えている．

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] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. : Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3:993-1002(2003)

[2] Minka, T. P.: The Dirichlet-tree distribution (Technical Report) http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirtree.pdf, 1999.

[3] http://nlpwww.nict.go.jp/wn-ja/