商品階層を自動構造化するトピックモデルの提案と小売販売データ分析への適用

1X20C093-1　藤田柊子
指導教員　後藤正幸

1. 研究背景と目的

市場における商品別の販売データを用いて消費傾向を分析することは，小売業者だけでなく，メーカーやデータ収集機関にとっても有用である．ここで，分析の要となるのが商品のカテゴリ情報であるが，多くのデータは，一般的に用いられる大分類・小分類程度の商品カテゴリしか付与されていない．また，データ収集期間が長い場合，度重なる商品の新発売・販売停止・リニューアルに起因して，大量の類似商品が異なる商品 ID で記録されていることが多い．このようなデータを，商品 ID 単位で，各商品の需要や販売傾向の分析をしても，有用な結果を得ることはできない．

そこで本研究では，目的に合わせた適切な粒度での分析を可能とするために，商品名や説明文などの文書を入力として，階層的なカテゴリを自動で付与することを考える．そのような階層的なカテゴリ (トピック) を推定可能な手法として hierarchical Latent Dirichlet Allocation(hLDA) [1] が考えられる．しかし，hLDA はある程度長い文書データを対象としており，本研究が対象とする，短い文書データに対しての頑健なトピック推定は困難である．

以上を踏まえ，本研究では短い文書データから階層的なトピック構造を推定可能な hierarchical Biterm Topic Model(hBTM) を提案する．提案手法により，各商品に対して，新たな階層のカテゴリを自動で生成する．これにより，様々な分析目的に合わせて，必要な粒度の階層を用いた商品分析が可能になる．最後に，実データを用いた評価実験，及び，得られた結果を活用した分析の例を複数示し，提案手法の有効性・有用性を示す．

2. 準備

2.1. hierarchical Latent Dirichlet Allocation

hLDA は，Latent Dirichlet Allocation(LDA) の拡張手法であり，文書や単語の所属するトピックの背後に階層構造を仮定し，トピックを確率的に推定する手法である．hLDA におけるトピックの階層は，nested Chinese Restaurant Process(nCRP) に基づいて生成される木構造で表現される．また，トピック数が生成過程で自動的に調整されることが利点である．

2.2. Biterm Topic Model

Biterm Topic Model(BTM) [2] は，同じ文書内で共起する単語対 (biterm) が同一トピックから生起すると仮定し，biterm を対象としてトピックを推定する手法である．BTM は，コーパス全体で単語の共起を捉えるため，短い文書に対しても頑健なトピック推定が可能である．

3. 提案手法

hLDA は，トピックの階層構造を自動で構築できる点と，分析者によるトピック数の設定が不要である点から，階層性を考慮した商品分類に有用であると言える．しかし，hLDA は長い文書への適用が想定されているため，本研究で対象とする商品名のような短い文字列からは上手くトピックを学習できない．そこで，短い文書に適した BTM を hLDA に援用した，hBTM を提案する．hBTM は，biterm がある同一階層の同一トピックから生起することを仮定したモデルである．hBTM により，短い文書データからも頑健なトピック (階層構造) の推定が期待される．

hBTM における各文書の生成過程を以下に示す．ここで，c_l は第 l 階層で選択したノード (トピック)，$\boldsymbol{\alpha}$ はディリクレ分布のパラメータ，$\boldsymbol{\beta}_c$ はノード c の単語分布，$\{w_{n,1}, w_{n,2}\}$ と z_n はそれぞれ n 番目の biterm b_n に含まれる 2 単語，所属する階層を表す．また，図 1 に hBTM のグラフィカルモデルを示す．

1. c_1 を根ノードと設定
2. 第 l 階層について，c_l を c_{l-1} の子ノード集合から nCRP に基づいて選択 ($l \in \{2, \ldots, L\}$)
3. 階層の多項分布パラメータ $\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ を選択
4. 文書中の各 biterm $n \in \{1, \ldots, B\}$ について，
   (a) 階層 $z_n \sim \mathrm{Multinomial}(\boldsymbol{\theta})$ を選択
   (b) biterm $w_{n,1}, w_{n,2} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{c_{z_n}})$ を選択

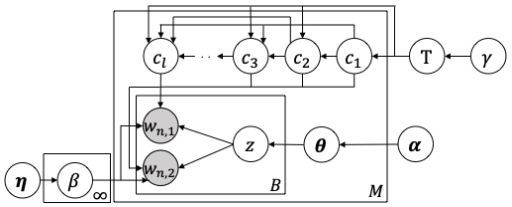

図 1: hBTM のグラフィカルモデル

ここで，$\boldsymbol{\eta}$ はディリクレ分布のパラメータ，γ は nCRP において文書が新しいトピックに割り当てられる頻度を制御するパラメータ，M は文書数，T は nCRP によって構築されたトピックの階層関係を表す木構造を示す．

4. 実験

4.1. 実験条件

本研究では，経営科学系研究部会連合協議会主催の令和 5 年度データ解析コンペティションで提供された，日経 POS データを用いて提案手法の有効性を確認する．対象データは，2013 年 7 月〜2022 年 6 月の約 10 年間に収集された，96,396 種類の飲料商品の売上と販売個数である．各商品には，28 種の大分類 (L)，128 種の小分類 (S) のうち，各一つずつのカテゴリが付与されている．

提案モデルの評価実験には，売上上位の「発泡酒風飲料」「缶入りビール」「チューハイ」の 3 カテゴリ (S) に属する 3,436 種類の商品データを用いる．hLDA，hLDA の入力形式を単純に biterm に変換した (生成過程には変更を加えていない)hLDA(bi)，提案手法である hBTM の 3 つの手法を，手動でラベリングした正解ペア (150 ペア) がそれぞれ同一トピックに分類されるかを再現率によって評価する．

4.2. 実験結果と考察

初めに，再現率を比較した結果を表 1 に示す．表 1 より，全カテゴリの全階層 (小小分類 (2S)・小小小分類 (3S)) において，hBTM が最も高い再現率を示した．このことから，提案手法は商品の階層構造を構築する上で有効であることが示唆される．

表 1: カテゴリ・階層ごとの再現率比較 (2S/3S(%))

手法	発泡酒風飲料	缶入りビール	チューハイ
hLDA	30.77/23.08	27.91/25.58	34.38/32.29
hLDA(bi)	38.46/38.46	72.09/62.79	71.88/69.79
hBTM	**100.00/92.31**	**76.74/76.74**	**85.42/85.42**

次に，発泡酒風飲料の商品群に提案手法を適用し，カテゴリの階層を 2S, 3S まで拡張した結果の一部を図 2 に示す．図 2 より，提案手法により得られた新たな階層において，メーカーや商品のシリーズ，特性などでまとまった分類になっていることが分かる．

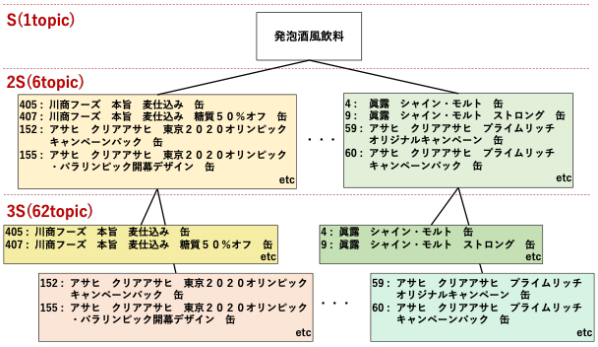

図 2: hBTM により階層を拡張した結果 (発泡酒風飲料)

5. 提案手法の応用

本章では，hBTM によって自動で拡張された各階層のカテゴリを活用した実データ分析手法について 2 事例を示す．

5.1. 小小分類 (2S) の NTF への適用

提案モデル (hBTM) から得られた 2S の商品分類により，非負値テンソル因子分解 (NTF) [3] を用いた商品の売上と気象条件の関係性分析が可能となる．NTF とは，3 次元以上の高次元データを低次元データの積で近似し，複雑な構造を持つデータを解釈がしやすい形に展開する手法である．本研究の対象データを直接 NTF に適用した場合，カテゴリの粒度が荒く，有用な結果が得られなかった．一方，NTF に入力する特徴量として提案手法から得られる 2S を採用することで，適切な粒度での分析によって有用な結果を得ることができる．NTF により得られた結果を表 2 に示す．

表 2: NTF によって得られたクラスタの解釈

クラスタ	解釈
1	ホップ商品が多い/降水量は少なく，価格帯は中〜高
2	徳岡ホールベルシリーズが多い/価格帯は低い
3	降水量は多く，気温は高い/高価格帯の商品が多い
4	商品名に「贅沢」が入る商品が多い/価格帯は低〜中
5	降水量は中程度で，気温は高め/価格帯は低め

表 2 より，例えば，クラスタ 1 には降水量が少なく，価格帯が高い商品が属していることがわかる．また，クラスタ 1 に属する商品はホップが多いことから，高価格帯のホップは降水量が少ない日に売れる傾向があると解釈できる．このように，hBTM から得られた新たな階層のカテゴリに対して NTF を活用することで，ビジネス上で有用な知見が得られることを確認できた．

5.2. 小小小分類 (3S) を用いた時系列予測

本節では，提案モデル (hBTM) から得られた 3S の活用例として，リニューアル商品やマイナーチェンジの多い商品に対しての需要予測モデル構築の例を示す．小売業者が新たな商品の取り扱いを検討する際，需要量は重要な意思決定のための情報となる．しかし，リニューアルやマイナーチェンジのたびに異なる ID が付与され，商品 ID 単位では十分な過去の販売データが得られないような商品に対して，需要予測は困難である．そこで，提案モデルの 3S を活用して先行・後継商品の紐付けを行うことでこの問題を解決する．具体的には，同じ 3S のカテゴリに属し，販売終了から新商品発売までの期間が 1 年以内である商品ペアのうち，商品名の最も類似するペアを紐付ける．その後，時系列予測モデルである ARIMA を用いて，紐づけたデータから後継商品の最後 3ヶ月分の需要を予測する．評価指標には，平均絶対誤差 (MAE) を用いた．提案により，需要予測の精度が向上したある商品に対する予測結果を図 3 に示す．

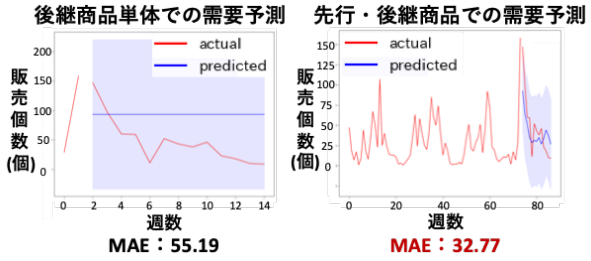

図 3: 3S を用いた需要予測の結果の例

このように，提案モデルで生成した 3S により，本質的に同一と見做せる商品をまとめることで，将来の需要予測モデルが構築できる．この結果は，小売業者が新たに取り扱う商品を検討する際に需要な参考情報となる．

6. まとめと今後の課題

本研究では，商品名 (短い文書データ) から商品の階層構造を自動で推定可能な hBTM を提案し，実データを用いた評価実験・結果の観察により有効性を示した．さらに，提案手法から自動で付与された各階層のカテゴリの多角的な活用例を示し，提案手法によるカテゴリ拡張の有用性を示した．今後の課題としては，分類精度の向上や，より深く拡張した階層の活用方法の検討などが挙げられる．

参考文献

[1] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proc. Int. Conf. NeurIPS*, pp. 17–24, 2003.

[2] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. In *Proc. Int. Conf. WWW*, pp. 1445–1456, 2013.

[3] A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proc. ICML*, pp. 792–799, 2005.