4M1-1

トピックモデルを用いた時系列分析に基づく潜在トピック推移の抽出
Extraction of Latent Topic Transition using Dynamic Topic Model

江本 守[*1]
Mamoru Emoto

大澤 幸生[*1]
Yukio Ohsawa

[*1] 東京大学大学院工学系研究科
School of Engineering, The University of Tokyo

In this study, we focus on extraction of latent topic transition from POS data. POS analysis is conducted to obtain frequent patterns of customer's behavior. In the fundamental method for POS analysis such as market basket analysis we can extract sets of products often bought at the same time. In market basket analysis, however, the effect of time series is hardly considered. We conducted the experiment based on two hypotheses. One is that each product has several topics. The other is that the proportion of each product on a topic changes as the time period changes. To extract topics and their changes, we used Dynamic Topic Model (DTM), an extended model of Latent Dirihlet Allocation (LDA). As a result, we obtained the change of the word distribution on each topic. Different topics have different characters, but seem to have a certain co-relationship correlation according to the correlation analysis we executed for several items.

1. はじめに

小売業において POS (Point Of Sales)システムにより収集される顧客購買データ分析は重要である.商品の売れ行きや地域ごとの売れ筋商品の差の分析,同時購買されやすい商品の特定を行うために POS 分析は行われており,店舗のタイムセールの実施品目,陳列商品の位置,店舗に陳列する商品の決定に役立てられている.また,近年は ID-POS と呼ばれる,会員 ID,クレジットカード番号などが紐付いた POS データの収集も積極的に進められている.ID-POS と機械学習の手法を組み合わせることで各顧客に合わせた商品の推薦が可能となり,またペルソナ分析を行うことで,年代,来店時間,性別などから商品購買パターンの分析を行うことが可能となる.

小売業における POS 分析手法の一つにマーケットバスケット分析がある.これはアソシエーション分析の一つであり,スーパーやコンビニエンスストアなどの買い物カゴ(バスケット)の中で同時購買確率の高い商品を分析する手法である.各購買バスケット内の商品の共起性を計算し,Confidence(信頼度),Support(支持度)を求めた後に Lift(リフト値)を計算することが実業界では多い.リフト値とは,「条件 X の時の事象 Y の割合」を「全体での事象 Y の割合」で割ったものである.これを小売業の文脈に置き換えると,「商品 X が買われた時に商品 Y も買われる確率」を「全体で商品 Y が買われる確率」で割ったものと解釈されるため,商品 X,Y の同時購買傾向を表す指標となる.しかし,マーケットバスケット分析においては,バスケットの時系列性は考慮していない.また,実際の購買行動において,商品は幾つかのトピックを有しており,時系列によりトピック内でのトピックらしさは変化すると考えられる.例を一つ挙げると,惣菜コーナーで売られている唐揚げは,昼間は弁当として購買されることが多い(弁当トピックらしさが高い)一方で,夜間はビールのおつまみとして購買されることが多い(おつまみトピックらしさが高い),といった具合である.商品のトピックらしさの時系列的推移を考慮することで,リフト値のような 2 商品間の関係だけでなく,購買トピックを考慮したマーケティング戦略を策定に役立つと考える.

本研究においては,自然言語処理で多く用いられるトピックモデルの中でも,Latent Dirichlet Allocation(LDA) [Blei 03]の拡張モデルで,時系列性を考慮した Dynamic Topic Model(DTM) [Blei 06]を POS データに適用し,時系列変化した時にトピック遷移を抽出した.

2. 関連研究

筆者は,レシピ共有サイトを題材として料理間の分類と特徴抽出を行い,レシピ共有サイトで利用されるレシピの季節変化についてテキストマイニングの手法により実験を行った[江本 15].サイトでは,レシピコンテンツと,そのレシピに対する調理後のレビューが投稿されている.各投稿レシピ,レビューに対しては投稿された日時データが紐づけられている.レシピレビューのテキストデータに対して形態素解析を行い,ワードカウントの手法により健康,家族,時短,美味しさの 4 要素から成るベクトルを作成した.この 4 つのトピックはトピックモデルを用いて作成したものではなく,早矢仕ら[Hayashi 13]が提唱するアクション・プランニングという手法を用いて,スーパーマーケットにおけるレシピ推薦アプリの開発についてのアクションプランを作成した際,多く議論がなされたトピックから定めた.分析の結果,各要素値が大きい月が存在することを特定した.

高橋らは,本研究にも使用する DTM 分析を時系列ニュースに対して実施し,情報集約を行うための二種類の方式として,バースト解析とトピックモデルの 2 つの手法を組み合わせ,時系列ニュースから推定されたトピックに関するバーストを検出する手法を提案している[高橋 11].

POS データにトピックモデルを用いた研究として石垣らは,PLSI(確率的潜在意味解析)を ID-POS データに対して行い,顧客の購買パターンに基づいた自動的なトピック生成をし,顧客と商品を同時分類し,その後にカテゴリに加えて,天気,購買時間帯,バーゲンなどの状況変数を ID 付き POS データに付与し,ベイジアンネットワークにより各変数間関係を確率的構造モデルで表現し,変数間関係の抽出法を提案している[石垣 11].また,ID-

連絡先:江本守, 東京大学大学院 工学系研究科 システム創成学専攻, mamoemo.0130@gmail.com
大澤幸生,東京大学大学院 工学系研究科 システム創成学専攻,ohsawa@sys.t.u-tokyo.ac.jp
This research was supported by JST, CREST
本研究に使用した POS データを提供してくださった株式会社カスミの皆様に感謝申し上げます.

POS データに対して PLSI を行い,顧客と商品を同時分類した後に,顧客アンケートデータから顧客の消費,生活因子を抽出し,PLSI の結果と統合することにより,顧客パーソナリティやライフスタイルを考慮した顧客と商品の同時カテゴリ分類も行っている[石垣 10].いずれの研究においても,ID-POS に対してトピックモデルを用いることで,顧客の特徴を抽出し,クラスタリングを行うことに着目しており,店舗全体の時系列トピック遷移については考慮されていない.本研究では,顧客ベースではなく,店舗ベースのトピック遷移を抽出するために Dynamic Topic Model を用いた.

3. Dynamic Topic Model

Dynamic Topic Model(DTM)は,LDA の時系列拡張モデルである.トピック分布とトピックごとの単語分布に時間マルコフ性を取り入れたモデルとして提案され,トピックごとの単語分布,トピック分布の時間発展を捉えることが可能である.確率的生成モデルの視覚的表現であるグラフィカルモデルを用いて DTM を表した図を図 1 に,生成モデルを図 2 に示す.

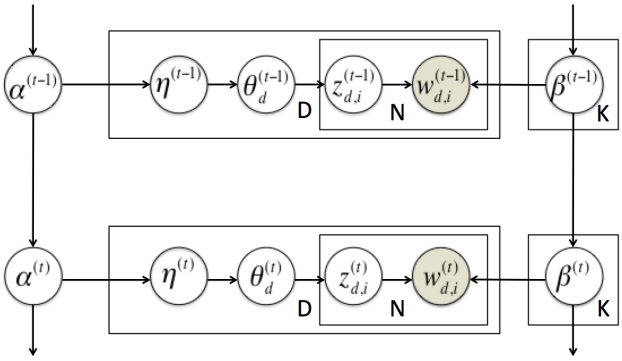

図 1 DTM グラフィカルモデル

for t = 1,2,…,T :
1.Draw topics
$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2\mathbf{I})$
2.Draw αt
$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2\mathbf{I})$
for each document :
3.topic proportion
$\eta_{t,d}|\alpha_t \sim N(\alpha_t, a^2\mathbf{I})$
$\theta_{t,d}|\eta_{t,d} \sim \pi(|\eta_{t,d})$
for each word:
4.topic-word assignment
$z_{t,d,n}|\theta_{t,d} \sim Multinomial(\theta_{t,d})$
5.word observation
$w_{d,n}|z_{d,n}\{\beta_{t,k}\} \sim Multinomial(\pi(\beta_{t,z_{d,n}}))$

図 2 DTM 生成モデル

図 1 の K はトピック数,N は時刻 t における語彙数,D は時刻 t における文書数である.α は,文書-トピック分布を生成するためのハイパーパラメータであり,時刻 t-1 のトピック分布を反映して時刻 t の α が生成されることを示している. 単語の出現分布は連続値をとる確率変数であるため,時系列モデリングにおいては状態空間モデル(state space model)を LDA に適用することで DTM に拡張されている.状態空間モデルは正規分布を仮定するため,状態空間モデルによるモデル化では,実数ベクトルが生成されるため,多項分布に用いるために soft-max 関数(正規化指数関数) π を利用して変換する. η は文書-トピック分布を生成するための実ベクトルであり, θ は η を soft-max 関数を用いて変換し,生成される文書-トピック分布である. β は時刻 t に生成されるトピックとして生成される実ベクトルであり,各トピックの単語分布が時系列変化することを示している. β を soft-max 関数を用いて変換することでトピック-単語分布 ϕ を生成する.

4. DTM の購買トピックの抽出

4.1 分析フローと実験条件

本研究においては,DTM をスーパーマーケットにて取得された POS データに対して適用し,特に各トピック内での単語分布の時系列変化(β の変化)に着目して分析を行う.以下の分析フローに従って実験を実施した.表 1 に POS データのメタデータを示す.

- POS データをトランザクションごとにまとめ,バスケットを作成
- 購入点数 5 点未満のバスケットをフィルターにより除去,入力ファイルを作成
- 前ステップで作成したファイルを入力として,DTM により分析
- 出力結果からトピック内単語分布の時系列変化を抽出

表 1 分析対象 POS データ

収集期間	2014/09/01-2014/11/30
バスケット数(購入点数 5 点以上)	212439
語彙数(商品種類数)	17162

トピックモデルの対象の多くはテキストデータ,つまり文書であり,Bag of Words として文書内に単語が含まれているものを入力とし推定を行う.本研究においては,文書と購買バスケット,単語とバスケット内商品という対応関係の類似性に基づき POS データに対してトピックモデルを適用している.DTM による推定においては,C++で作成されたプログラムを用いた.また,DTM の初期条件は,事前実験による調整の結果,トピック数 k=5, α =0.1 と設定した.バスケットは日ごとに分割し,各日付をタイムピリオドとしており,t=0,1,2,…,90 である.

4.2 実験結果

DTM により生成された 5 つのトピックについてカテゴリと単語分布の上位 10 件のアイテム(t=0)を表 2 に示す.表 2 に示している上位アイテムは t=0 の時の上位アイテムであるから,DTM 解析により時刻 t の変化に伴う上位アイテムの変化が発生するが,全体としては各トピックに関して単語分布が上位の商品から以下のような特徴が見られた.

- Topic0:菓子パン,飲料水,スナック
- Topic1:惣菜,炭酸,お酒
- Topic2:野菜,肉(料理の材料)
- Topic3:惣菜,サラダ,出来合い品
- Topic4:朝食,肉,乳製品

表 2 Topic 内上位アイテム(t=0)

Topic	Item
Topic0	いばらく あじわい牧場牛乳 1000ml, コカ・コーラ 森の水だより 2L, WILLバナナ,

	第一パンオールドファッションチョコD5個， ミックスサラダ，もやし200g(富士)， トマト L，カルビー夏ポテト対馬の浜御塩味,コロッケ 98 (代表･店舗使用コード)， 東ハト ポテコ うましお味
Topic1	オリジン商品，やきとりバラ販売， 大きな若鶏唐揚げ， コロッケ 98 (代表･店舗使用コード)， コカコーラ ゼロ 500ML， 野菜かき揚，綾鷹 525ML， アサヒ ウィルキンソン炭酸 500ML， 明和 緑茶 500ML， キリン 午後の紅茶おいしい無糖500ML
Topic2	もやし200g(富士),にら(束),キャベツ,オーエムツーミート豚肉スライス(切落し)， きゅうり(バラ),ミニトマト大， 舞茸(100gパック)，キャベツ(カット) えのき茸(200g)，長葱(1本)
Topic3	東立商事 中華惣菜，オリジン商品， 芝武 焼き魚，オーエムツーミート惣菜その他，オーエムツーミート 米飯，コロッケ 98 (代表･店舗使用コード)，オーエムツーミー 揚物，おつまみ枝豆， ほたての磯辺天ぷら， オーエムツーミート 温惣菜
Topic4	キャベツ,いばらく あじわい牧場牛乳 1000ml， 信州高山食品 美味しいきぬとうふ3個， 日本ハム シャウエッセン2ケ束， おもてなしたまご 白10個,もやし200g(富士),コロッケ 98 (代表･店舗使用コード),おかめ ふわりんやわらか納豆 40GX3，おかめ 旨味かつお ミニ3， 利根 米油を使った油揚げ 5枚 2便

また,Topic1,2,3 について単語分布の値の時系列平均の上位アイテム 4 件を抽出し,グラフにまとめた結果を図 3 から図 5 に示す.

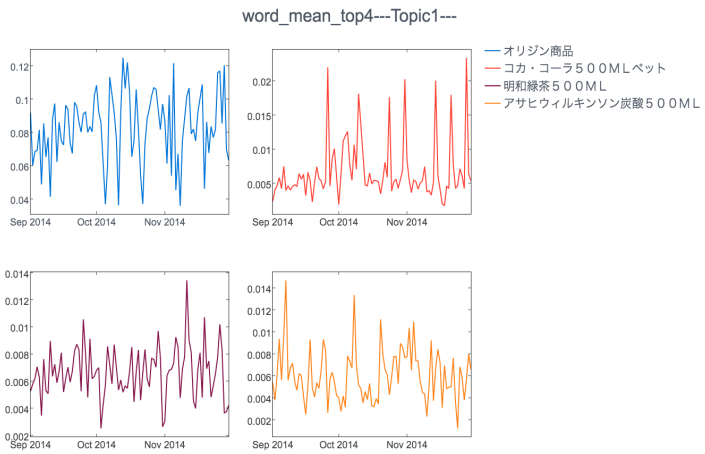

図 3 Topic1 単語分布時系列平均上位アイテム

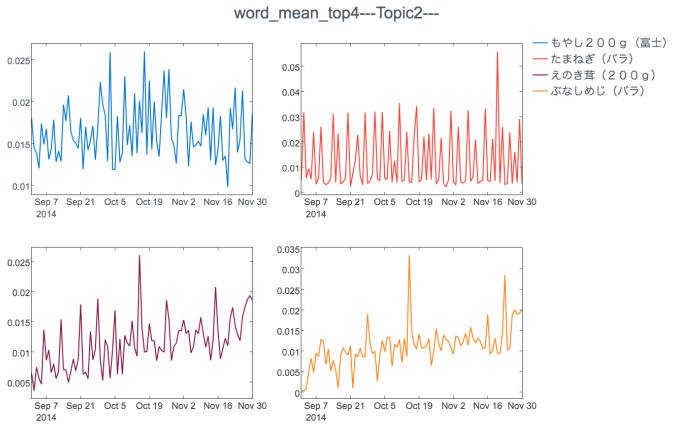

図 4 Topic2 単語分布時系列平均上位アイテム

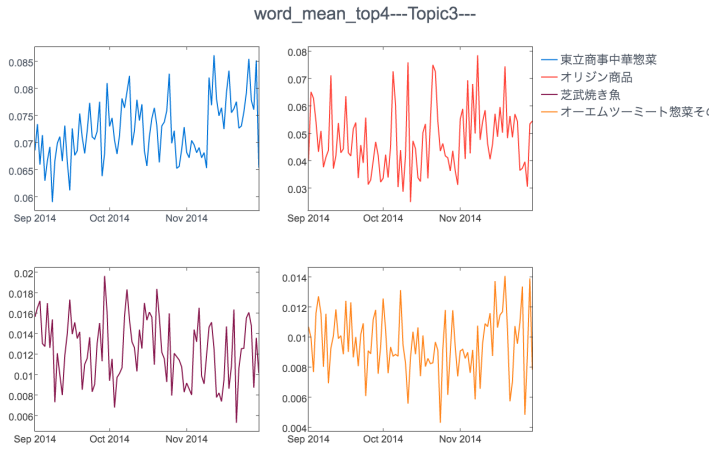

図 5 Topic3 単語分布時系列平均上位アイテム

5. 考察

本分析では,あるスーパーマーケットにおいて 2014 年 9 月 1 日から 2014 年 11 月 30 日に収集された POS データについて Dynamic Topic Model(DTM)を適用し,各トピックの単語分布の時系列遷移を抽出した.図 3-図 5 の結果から,各アイテムの単語分布の時系列変化には,恒常的に大きな値を示しているアイテムと,局所的に大きな値を持つ時系列変化を行うアイテムが存在することがわかる.特に図 3 で示された Topic1 と図 5 で示された Topic3 では,オリジン商品(惣菜),中華惣菜が常に高い値を示している.図 4 で示された Topic2(野菜,肉(料理の材料)の多いトピック)は,局所的に高い値を複数の商品が同日に示していることが特徴として挙げられる.一週間に一回(土曜日,日曜日)程度の周期で複数の商品が高い値を示すことがあり,調理材料としての食品を多く購入する顧客が多い日が存在していることが分かる.

Topic1 と Topic3 では,上位アイテムに惣菜が多く出現する.両方のトピックで単語分布が高い値を示すアイテムについて,各トピックでの値を比較する.ここでは例として,「レタスとコーンのサラダ」,「オリジン商品」に関してグラフを作成し,図 6,図 7 に示す

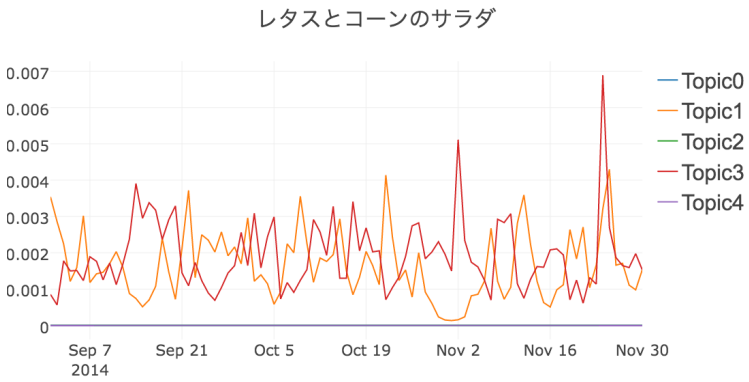

図 6　時系列トピック-単語分布値[レタスとコーンのサラダ]

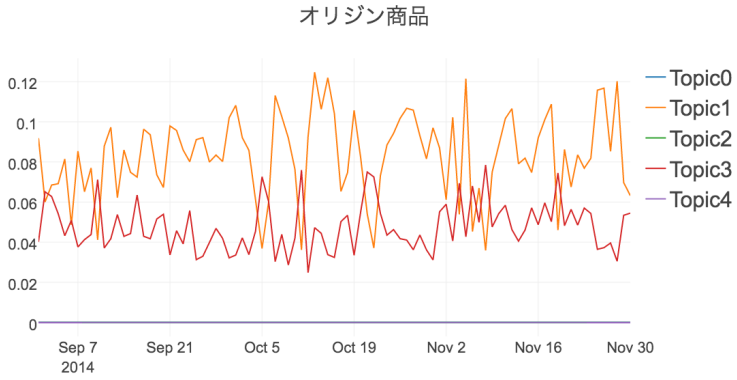

図 7　時系列トピック-単語分布値[オリジン商品]

各アイテムについて,Topic1 と Topic3 の値についての相関分析の結果,「レタスとコーンのサラダ」では-0.345442,「オリジン商品」では-0.755951 となり,負の相関が示唆された.Topic1 の単語分布においては惣菜とともに炭酸飲料,お酒が高い値を示していることから Topic1 の購買パターンでは,惣菜はおつまみ要素が強いことが想定される.

以上の考察を踏まえ,今後の課題としては,惣菜商品の購買性質の変化について曜日,購買時間帯との関連性などを調査することを挙げたい.また,より長期的な POS データを対象とすること,タイムピリオドの粒度を細かくし,時間単位での分析を行うことも挙げられる.時間単位での購買トピックの推移を抽出することで,マス的な視点で,タイムセールの戦略,宣伝方法の改良を行う際に役立つ.さらに現在は折れ線グラフによる可視化のみであるが,DTM により単語分布の時系列変化が行列として取得可能であるから,より直感的に時間推移が理解できるような可視化手法の研究も展望としてあげられる.スーパーマーケットの購買は料理と密接に関係しているため,関連研究で述べたレシピ共有サイトを対象として DTM 解析を行い,テキストデータに基づいて食材の季節変動などを抽出し,POS データとの比較を行うことで,消費者の購買パターンの抽出を行うことも試みたい.

参考文献

[Blei 03]　Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol.3, pp.993–1022 (2003).

[Blei 06]　Blei, D.M. and Lafferty, J.D.: Dynamic Topic Models, Proc. 23rd ICML, pp.113– 120 (2006).

[江本 15]　江本　守,大澤　幸夫:レシピ共有サイトにおける料理間分類と特徴抽出,IEICE technical report, Vol.115, No.337, pp73-77(2015).

[Hayashi 13] Hayashi, T. et al: Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game Plus Action Planning, International Journal of Knowledge and Systems Science, Vol. 4(3), pp. 14-38, (2013).

[高橋 11]　高橋　佑介,横本大輔,宇津呂　武仁,吉岡　真治:ニュースにおけるトピックのバースト特性の分析，情報処理学会研究報告, Vol.2011-NL-204, No.6,pp1-6(2011).

[石垣　11]　石垣　司,竹中　毅, 本村　陽一: 百貨店 ID 付き POS データからのカテゴリ別状況依存的変数間関係の自動抽出法，日本オペレーションズリサーチ学会誌，Vol.56, No.2, pp77-83(2011).

[石垣　10]　石垣　司,竹中　毅, 本村　陽一: 確率的潜在意味解析を用いた大規模 ID-POS と顧客アンケートの統合利用による顧客-商品の同時カテゴリ分類，電子情報通信学会誌，Vol. 109, No.461, pp425-430(2010)