商品の出荷パターンに基づく集約を用いた商品の潜在売行予測方法の検討

○廣沢柊平　後藤裕介（芝浦工業大学）

A Study on Potential Sales Prediction Method of Merchandise Using Aggregation Based on Merchandise Shipping Patterns

*S. Hirosawa and Y. Goto(Shibaura Institute of Technology)

概要－　本研究では，製パン企業の工場出荷データを用いてパン商品を推薦するシステム開発の前段階として既存の推薦システムを用いて潜在売行予測の研究を行う．製パンは多種多様な商品が取り扱われているため，購買が似ているユーザーを参考にする一般的な推薦システムである協調フィルタリングでは，多種多様なパン商品を単品で推薦するのは難しい．そこで，本研究では商品に対して出荷パターンという潜在的類似性を考慮したクラスタリングを行い商品の種類数を減らすことで，出荷数量を評価値とした既存の推薦手法による潜在的な売行を予測する手法の検討を行う．

キーワード:　クラスタリング，協調フィルタリング，推薦システム

1.　はじめに

パン商品は主にスーパーマーケット等で販売され複数企業の製品が置かれている．そのため製パン企業にとって他の製パン企業が競争相手であり，パンの価格と品質，品揃えによって市場競争が行われている[1]．市場競争により，パン商品は日々多数の新商品が開発されるが定着する商品はごく僅かであるため種類が多種多様であり，商品の特徴による辞書的な分類が難しく曖昧なものとなっている．その結果，商品を販売する小売店は店舗に並べる商品を決める作業に時間がかかってしまう．そこで，小売店に対しセールの目玉として新たに売り出す商品を決定する際や商品の入れ替えを行う際に，類似している店舗がどの程度売れていて，自分の店舗で販売した時の潜在的な売行が分かる様な商品発見を補助する推薦システムを実装すれば小売店の発注作業を効率化できると考えた．また，商品推薦の導出過程が明確な推薦システムを実装することで推薦の理由を理解することができ，小売店にとって売上が期待できる商品を高い信頼度で推薦することができる．

推薦システムは一般的に購買が類似したユーザーを導出し，類似ユーザーが高評価した商品を推薦するという協調フィルタリングによって行うが，中にはあまり購入されていない商品などが存在するため，多種多様な商品を単品で推薦するのは難しいとされている．そのため，協調フィルタリングの推薦に適するように商品の類似性による商品集約をする必要がある．

そこで本研究では，商品集約が推薦システムにどのような効果を与えるかを検証するため，商品集約を実行したデータを用いて推薦システムを実行し，潜在売行を予測する方法の検討を行う．

2.　関連研究

先行研究では製パン企業の工場出荷データを用いて店舗と商品の関係をネットワークで表し，Network Embedding手法により類似店舗を導出することで未取扱商品の売上予測手法を開発した[2]．しかし，この手法はネットワーク埋め込み処理にベクトルが使用されるため，小売店にとって導出過程を理解することが難しいと言える．

神嶌の研究では，推薦システムの1つとしてユーザー間型メモリーベース法協調フィルタリング(UMCF)が紹介されている[3]．この手法は各ユーザーの各商品への評価をまとめた評価データから推薦するユーザーと評価の仕方が似ている類似ユーザーを探し，類似ユーザが高評価した商品を推薦する手法である．この手法は導出過程が明確であり，他手法と比べ商品の特徴が不要，セレンディピティに優れているといった特長があるため，本研究では推薦システムとしてUMCFを使用する．

また，この手法は蓄積された評価データを使用するため多種多様な商品から推薦を行う際，あまり購入されていない商品，つまり評価数が少ない商品がある疎なデータや高次元なビッグデータに適用すると推薦の精度が下がることが指摘されている[4]．加えて，推薦の度に評価データを調査し直すためデータ数が多いほど推薦速度が遅くなるという欠点がある．そのため，近年では評価データの次元を減らすことで推薦の精度を向上させることに焦点が当たっている[5]．

白井らの研究では，JANコードの分類が過度に詳細なことなどに着目し，商品名によらない商品分類手法の開発を行った[6]．白井らは商品の併売状況という相互依存する商品群のEMクラスタリングを再帰的に繰り返すことで，最終的な分類結果を得る「ローテーションクラスタリング」を提案し，商品の併売状況という潜在的な類似性による商品分類ができた．本研究では商品の名前や特徴によらない潜在的類似性による商品分類に着目し，商品の集約・分類手法として製パン企業の工場出荷データから読み取れる各小売店へどの商品を何個出荷したかという出荷パターンという潜在的な類似性を考慮したクラスタリング処理をすることで，辞書的な分類が難しい多種多様なパン商品を意味の理解できる分類をすることができるのではないかと考えられる．

3. 研究目的と研究方法

本研究では商品の出荷パターンに基づく集約を用いた商品の潜在売行予測手法についての検討を行い，以下の様に研究を進めていく．

I. 潜在売行予測手法の提案・開発
II. 実際の製パン工場の事例に適用
III. 提案手法の有効性の評価

4. 提案手法

研究提案手法の概要を以下のFig.1に示す．

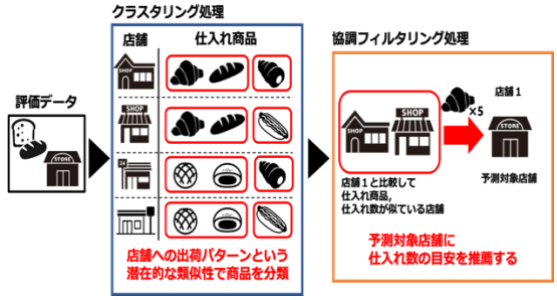

Fig.1：提案手法の概要図

Fig.1のように，製パン工場出荷データより評価データを作成し，そのデータを入力としてクラスタリング処理によって店舗への商品という潜在的な類似性で商品を集約・分類を行う．商品集約を行なった評価データを入力として協調フィルタリングによって潜在的な売行である出荷数を予測する．

4.1. 評価データの作成

提案手法では，製パン工場出荷データを店舗と商品の組み合わせに関して1年間の出荷数量の中央値を評価値とした評価データを作成する．その評価データを入力としてクラスタリング処理を行う．

4.2. クラスタリング処理

クラスタリング処理について，商品に関する群平均法による階層型クラスタリングを使用する．階層型クラスタリングはデータを一つ一つ比較し，類似しているものを同じグループにする手法である．また，群平均法とは2つのクラスタを構成するデータのすべての組み合わせの距離を求め，その平均をクラスタ間の距離とする手法である．

クラスタリング処理を行うため，評価データを店舗と商品の関係を表すマトリックスデータを作成する．マトリックスデータの例を以下のTable 1 に示す．

Table 1: マトリックスデータの例

店舗ID	5678	5679	5680	5681	5682	…
商品ID						
1234	5	8	0	0	10	…
1235	10	12	9	9	0	…
1236	0	0	0	0	8	…
1237	5	5	5	5	5	…
1238	0	0	7	0	9	…
	⋮	⋮	⋮	⋮	⋮	

Table 1 の様なマトリックスデータを行方向の数列を特徴ベクトルとして商品をクラスタリングすることで，各小売店へ出荷した商品とその出荷数という出荷パターンによって商品を集約・分類ができる．また，Table1を見ると商品ID 1236の様に，あまり購入されていない商品があるためこのマトリックスデータは疎なデータと言える．一般的なクラスタリングではデータのすべての組み合わせの距離を測る際にユークリッド距離を使用するが，疎なデータを扱う際には，適切な距離が測れないため，今回はコサイン距離を使用する．

4.1で作成した評価データに対し，以上の処理を行い，各店舗への出荷パターンによる潜在的な類似性を用いて商品を集約し新たな商品分類を行う．その結果を反映させた評価データを協調フィルタリング処理への入力とする．

評価データにて購入していない商品は欠損値となるため，クラスタリング処理にて，同じクラスタ同士の中央値を取ることで欠損値を補う．以上のように，商品集約を実行することで評価数を増加させ，疎なデータを改善し，全体のデータ数を減少させることで先述した協調フィルタリングの問題を解決できると考えた．

4.3. 協調フィルタリング処理

協調フィルタリング処理では，先述した通りUMCFを使用する．UMCFは予測を行うユーザーとその他のユーザーとでピアソンの相関係数を使って類似度を算出する．ピアソンの相関係数はMを全商品の集合とし，xを評価値を予測するユーザー，yを類似度を計算するユーザー，$r_{x,m}$をユーザーxの商品mへの評価値，$\bar{r_x}$をユーザーxの平均評価値とすると類似度$w_{x,y}$は式(1)の様に計算される．

また，類似したユーザーの集合をUとすると各ユーザーの平均評価値を算出し，類似度と平均評価値を重みとした加重平均をとることで予測評価値として潜在出荷数 $\widehat{r_{x,m}}$ を式(2)のように算出する．

$$w_{x,y} = \frac{\sum_{m\in M}(r_{x,m}-\bar{r_x})(r_{y,m}-\bar{r_y})}{\sqrt{\sum_{m\in M}(r_{x,m}-\bar{r_x})^2}\sqrt{\sum_{m\in M}(r_{y,m}-\bar{r_y})^2}} \quad (1)$$

$$\widehat{r_{x,m}} = \bar{r_x} + \frac{\sum_{y\in U} w_{x,y}(r_{y,m}-\bar{r_y})}{\sum_{y\in U}|w_{x,y}|} \quad (2)$$

4.2で作成したクラスタリングを反映させた評価データを入力として以上の処理を行い，商品を仕入れる場合に何個仕入れるべきかという潜在出荷数を評価値として予測する．

5. 事例への適用

本研究では，製パン企業様ご提供の工場出荷データを用いて研究を行う．

5.1. 使用データ

工場出荷データは日時のデータであり，時間と店舗と商品を1レコードとしている．工場出荷データについて内容の一部を以下のTable 2 にまとめる．ただし実際に存在する店舗の情報があるため例を用いる．

Table 2 : 工場出荷データの例

出荷日	数量	価格	店舗ID	店舗名	商品ID	商品名
2020/1/1	8	300	1234	○○　××店	5678	あんぱん
2020/1/1	5	350	1234	○○　××店	8765	チョコパン
2020/1/1	12	300	4321	□□　△△店	5678	あんぱん
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2020/12/31	10	350	4321	□□　△△店	8765	チョコパン

5.2. 評価データの作成

本研究で使用しているデータは期間を2020年1月1日-2020年12月31日の1年間とし，対象商品を最も売上が高い「菓子パン群」に限定したものであり，この工場出荷データに対して期間限定商品や小売店でのセールといった通常の出荷から逸脱した異常値を除去するため，以下のようにデータクリーニング処理を行った．

- 各レコードの出荷数が正規分布の80%信頼区間から外れているレコードを削除
- 店舗毎，商品毎の出荷期間が7日以上の組み合わせに限定
- 小売店での一般的な販売個数である出荷数5〜12個のデータに限定

以上のデータクリーニング処理の結果，レコード数1,406,735，店舗の種類数3,218 ，商品の種類数409となった．

データクリーニング処理を行った工場出荷データを4.1で述べたように，店舗と商品の組み合わせに関して1年間の出荷数量の中央値を評価値として評価データを作成する．評価データの例を以下のTable 3 にまとめる．

Table 3 : 評価データの例

店舗ID	商品ID	評価値
1234	5678	8
1234	8765	5
⋮	⋮	⋮
4321	5678	10
4321	8765	10

クラスタリング処理をした評価データを各小売店で出荷期間が長い商品の5件をテストデータ，それ以外のデータを学習データに分割し学習データをUMCFへ入力，実行した後に予測評価値$\widehat{r_{x,m}}$とテストデータの評価値を$r_{x,m}$とし，評価値予測を行ったデータ数をn，全店舗の集合をXとして以下の式(3)で定義されるMAE(平均絶対誤差)にて精度評価を行う．

$$MAE = \frac{1}{n}\sum_{x\in X}\sum_{m\in M} |\widehat{r_{x,m}} - r_{x,m}| \quad (3)$$

5.3. 工場出荷データの基礎集計

工場出荷データに関して基礎集計した結果を説明する．まず，各店舗が 1 年間で何種類の商品を取り扱うかというラインナップ数について Fig.2 にまとめる．

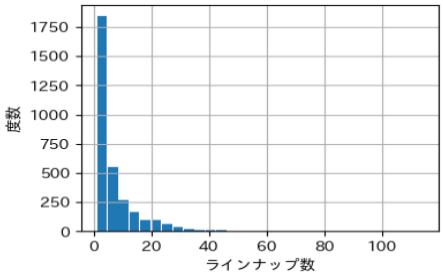

Fig.2 : 店舗ごとの年間ラインナップ数

Fig.2より，各店舗が1年間で取り扱う商品の種類数は 1~20 程度の店舗がほとんどであると言うことが分かる．次に1日の平均ラインナップ数について Fig.3 にまとめる．

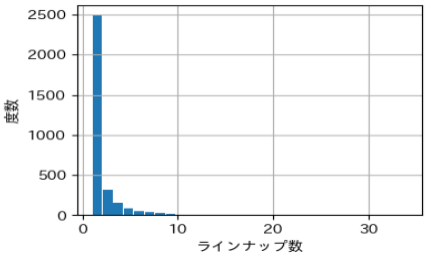

Fig.3 : 店舗ごとの1日あたりのラインナップ数

Fig.3より，各店舗が1日に取り扱う商品は 1~10種類程度がほとんどであるということが分かる．

以上のことから，小売店は菓子パン群だけでも1年間で400種類以上ある商品の中から20種類程の商品を選ぶ必要があることが分かる．一方で製パン工場にとってはあまり購入されていない商品が数多く存在しているということが分かる．

6. クラスタリング処理の適用

4.2で述べたように，実際の製パン工場出荷データでマトリックスデータを作成し，クラスタリングを実行する．階層型クラスタリングでは，同じクラス

タと判定するコサイン類似度の閾値を変化させることで生成されるクラスタ数を変化させることができるため，コサイン類似度の閾値を変化させ生成されるクラスタ数を調査する．その結果について以下のFig.4にまとめる．

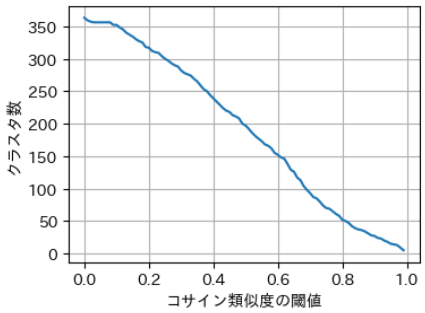

Fig.4：同クラスタ判定のコサイン類似度の閾値と生成されるクラスタ数の関係

Fig.4を見ると，類似度の閾値が大きくなるほど1次関数のように生成されるクラスタ数が少なくなることが分かる．

次に，4.2で述べた様にクラスタリング処理を行うことで欠損値を補い，以上のように，各商品への評価数を増加させることができたのかを調査する．そこで同クラスタ判定のコサイン類似度の閾値を変化させた場合各商品への評価数の平均値の変化について以下のFig.5にまとめる．

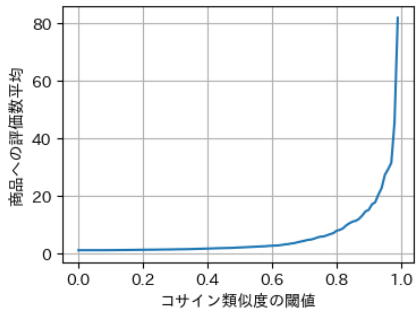

Fig.5：同クラスタ判定のコサイン類似度の閾値と各商品への評価数平均の関係

Fig.5を見ると，コサイン類似度を0.8未満では各商品への評価数の増加の効果は薄く，0.8を超えると評価数増加の効果が急激に現れることが分かった．

次に，各クラスタに割り当てられた商品数とその内容について調査する．ここではFig.6にて評価数増加の効果が見られたコサイン類似度の閾値0.8でクラスタリング処理を行なった場合で説明する．各クラスタに割り当てられた商品数について以下のFig.6にまとめる．

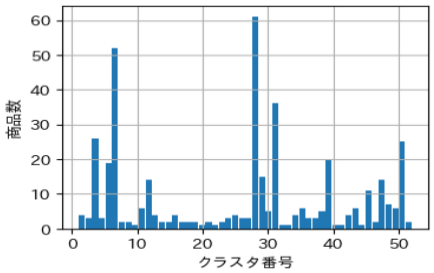

Fig.6：各クラスタに割り当てられた商品数

Fig.6を見ると，各クラスタで均等に割り当てられているわけではなく，クラスタによって商品数が大きく違うことがわかる．また，割り当てられた商品の内容を調査すると，商品の内容が類似しているものがまとめられているクラスタも存在するが，特定のコンビニエンスストア専用の商品などが同じクラスタに割り当てられているものが目立った．以上のことから，特定の店舗のみに販売される，発注のされ方が類似しているといった商品の出荷パターンが類似しているものが同じクラスタに割り当てられたと言える．

7. 有効性の検証

商品の出荷パターンに基づく階層型クラスタリングを実行した結果について，精度が最大のもの，クラスタリング処理を行なっていないもの，Network Embedding 手法による売上予測を行なった先行研究[2]の３つの誤差分布を以下のFig.7-9に示し比較する．ここで，先行研究では使用データは同じであるがデータ範囲や適用している手法が異なるためデータ数などに違いがある．

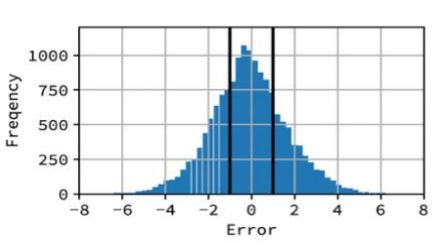

Fig.7：先行研究の誤差分布[1]

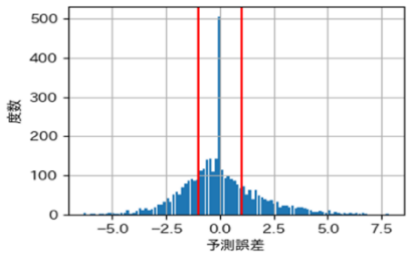

Fig.8：クラスタリング処理前のUMCFの誤差分布

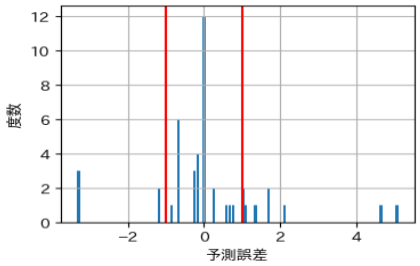

Fig.9 : クラスタリング処理後の
UMCFの誤差分布

比較の結果，先行研究はMAE 1.363 , クラスタリング処理前のUMCFはMAE 1.339，クラスタリング処理後はMAE 0.910となり，クラスタリング処理を行なった場合のUMCFが精度が最も良い結果となった．また，Fig.9を見ると評価数が大きく減少しているが，その他と比較して誤差の範囲が狭まっていることがわかる．

次に，Fig.9のようにクラスタリング処理を行うことで精度が向上した要因は蓄積された評価データを使用するUMCFの特徴から各商品の評価数の変化が関係していると考え，クラスタリング前のテストデータに含まれる各商品について何店舗が同じ商品を取り扱っているかという評価数と商品ごとのMAEを分析した結果を以下のFig.10 に示す．各商品のMAEとは小売店と商品の組み合わせに対し予測される評価値を商品毎で合計し商品の種類数で割り，平均を取ったものである．

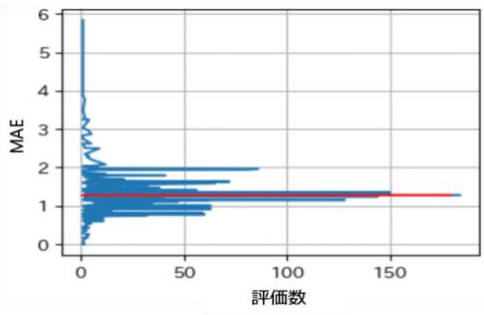

Fig.10 : テストデータに含まれる各商品の
評価数とMAEの関係

Fig.10 から，MAEの大きい商品は評価数が少ない傾向にあることが判明した．クラスタリング前のデータでは予測を行う商品への評価を行っている店舗が少なく疎なデータであり，クラスタリング処理をすることで欠損値が補完され，評価数が多くなる可能性が向上することで，予測精度も向上したのだと考えられる．

次に，Fig.10の結果より評価数が多くなる可能性の向上がUMCFの精度とどのような関係があるか調査する．Fig.4より，同クラスタ判定のコサイン類似度の閾値を大きくすることでクラスタ数が減少することが分かっており，クラスタ数を少なくするほど各クラスタで割り当てられる商品数が増え，評価数が多くなる可能性の向上するため，同じクラスタと判定する類似度の閾値を変化させ，作成される商品クラスタの数とMAEの関係について以下のFig.11に示す．

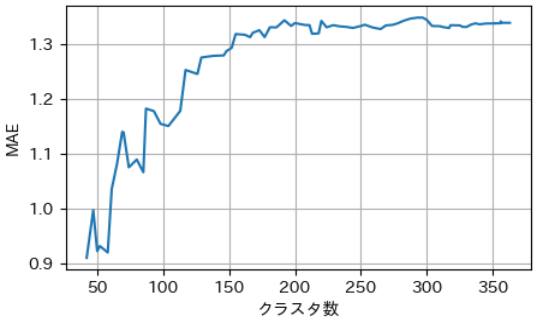

Fig.11 : クラスタ数とMAEの関係

ここではクラスタ数を40未満にした場合UMCFを実行する際に必要最低限のデータ数を下回ってしまうため，40未満のクラスタ数の結果はない．

Fig.11 を見ると，400以上ある商品を集約・分類する際，クラスタ数200まではUMCFの精度に与える効果が薄いことが分かる．しかし，クラスタ数を半分以下の200より少なくすると精度向上の効果があることが分かった．

8. 考察と今後の方針

本研究の結果から，商品の出荷パターンによるクラスタリング処理を用いることでUMCFの精度が向上することがわかった．これは，蓄積された評価データを用いるUMCFの特徴より，クラスタリング処理によって商品の種類数が減るため，商品について小売店が購入していないもの，つまり評価の無い欠損値が補完され，各商品の評価数が多くなる可能性が上がり，精度向上に繋がったと考えられる．

また，Fig.11の結果から商品分類をおおまかにすればするほど精度が向上する傾向があることが分かった．これは，先述の通り欠損値が補完され，各商品の評価数が多くなる可能性の向上の効果がより強く現れたた

めだと考えられる．しかし，商品分類をおおまかにしすぎると，蓄積された評価データを扱うUMCFの推薦の精度が低下し，最終的に小売店に推薦する意義が薄れてしまうため，適切なクラスタ数を見極める必要がある．

今後は本研究の結果を活かし，潜在的な類似性による商品分類の手法の開発とともに，UMCFに適用でき，小売店への推薦の意義のある適切なクラスタ数の見極めることで商品推薦手法の開発を行なっていく．

9. まとめ

パン商品は多種多様なため商品の集約を用いた商品推薦システムの実装が必要となる．本研究では推薦システム実装の前段階として，商品の出荷パターンという潜在的な類似性による商品集約を用いたデータに対し推薦システムを実行し商品の潜在売行である出荷数を予測することで，効果検証を行った．今後は本研究の結果を活かし，商品の潜在的な類似性による商品集約を用いた推薦手法の開発を目指す．

謝辞

本研究は，白石食品工業株式会社から工場出荷データを提供していただきました．厚く御礼を申し上げます．

参考文献

1) 堀内俊洋 ：パン産業の最近の構造についての一考察,早稲田政治経済学雑誌,372,39/54(2008)
2) K. Takahashi and Y. Goto : Embedding-based Potential Sales Forecasting of Bread Product, Journal of Advanced Computational Intelligence and Intelligent Informatics, **26**-2, 236/246 (2022)
3) 神嶌敏弘 ：推薦システムのアルゴリズム,人工知能学会, **22**-6, 826/837 (2007)
4) R. Chen, Q. Hua, Y. Chang, B. Wang, L. Zhang, and X. Kong : ASurvey of Collaborative Filtering-Based Recommender Systems:From Traditional Methods to Hybrid Methods Based on Social Networks,IEEE Access,**6**, 64301/64320 (2018)
5) Y. Takama, H. Shibata, and Y. Shiraishi : Matrix-Based Collaborative Filtering Employing Personal Values-Based Modeling and Model Relationship Learning, Journal of Advanced Computational Intelligence and Intelligent Informatics, **24**-6, 719/727 (2020)
6) 白井康之，森田裕之,後藤裕介 ：商品の潜在的類似性に基づくクラスタリング手法の提案,オペレーションズ・リサーチ, **62**-2, 91/99 (2017)