CF-002

行列分解を用いた再購買周期の推定とレコメンダーへの応用

Estimating Repurchase Cycles based on Matrix Factorization and Application to Recommender System

趙漢哲* Han-Cheol Cho
佐藤哲† Tetsu Sato

1 はじめに

情報推薦手法の一つであるレコメンダーシステムは 2000 年前後から Amazon を先頭とする多くの e コマース企業によって幅広く活用され始めた。その対象は書籍 (Amazon)[6], 音楽 (Spotify), 映画 (Netflix)[1] のような商品に留まらず，近来には宿泊 (Airbnb) や交通 (Uber) サービスまで拡大されてきた。今は社会の様々な分野で人々の意思決定を支える欠かせない重要な技術と言っても過言ではない。

レコメンダーシステムが推薦情報を生成する方法は大きく二つに分類される。一つは商品に付与されている情報を元に顧客がすでに購入または観覧した商品と類似する商品を推薦するコンテンツベースアプローチ [7] である。もう一つの方法は顧客の行動パターンに基づいて類似する顧客群を判定し，そこから推薦情報を生成する協調フィルタリングアプローチ [3, 4, 6, 8, 9] である。協調フィルタリング方式はコンテンツベース方式と比べて新規顧客への情報推薦が難しい問題点 (cold-start problem)[11] を抱えているが，学習データの構築にドメイン知識を必要としないため幅広い分野で活用できる長所を持っている。

本研究では消耗品のように顧客が繰り返し購買する商品を再購買周期に合わせて推薦する再購買レコメンダーシステムを研究対象とする。再購買レコメンダーは主に新たな商品の発見・推薦を目的とする既存のレコメンダーと相互補完的な機能を持ち，簡単に競争者のサイトを訪問し同じ商品を購入することが可能な e コマースの世界 [10] で顧客の満足度を向上させ離脱を防止する機能を強化させることが可能だと考えられる。本論文で提案する再購買レコメンダーは再購買周期推定器と再購買日予測器の二つの部分で構成されている。再購買周期は，顧客が商品を 2 回以上購買した場合，連続する購買イベントの時間差で定義する。私たちはこの再購買周期データに含まれているノイズを除去することで再購買日の予測をもっと的確に行うことが可能だと考えた。その手法としては画像・音声分野でデータに含まれているノイズを除去する手法としてよく使われている行列分解 [2, 14] を利用する。再購買日の予測は，最終購買日から再購買周期分の時間が経った日として定義する。

以下，2節では時間情報を活用したレコメンダーシステムと行列分解を用いてデータからノイズを除去する手法に関する関連研究を紹介する。3節では提案システムを詳しく説明し，4節では美容関連 e コマースサイトの購買データを利用した実験結果と分析内容を報告する。5節では本研究の貢献と今後の課題に関して議論する。

2 関連研究

顧客の嗜好や商品の人気は時間と共に変化する。レコメンダーがより適切な情報を提供するためにはこのような顧客・商品の変化を考慮する必要がある。Lathia et al.は時間的多様性 (temporal diversity) という概念がレコメンダーの評価に重要 [5] と主張し，その評価方法及び実現手法を提案した。続けて多くの関連研究が行われ始め，商品の購入順番の依存性を考慮した研究 [16] や顧客嗜好の変化を取り入れた研究 [15]，そして商品の季節性 [13] やライフサイクル [17] に焦点を当てた研究などが行われた。また，商品の再購買周期を考慮した研究 [12] も存在する。本研究と異なる部分は，再購買周期の推定が商品単位であり顧客を考慮しないこと，周期の単位が短中長期の三段階であり日単位ではないこと，そして再購買周期情報は伝統的なレコメンダーの結果をリランキングするために使われることである。

スパースな特徴空間でも堅固に動作する特徴を持ち協調フィルタリングの一手法として幅広く使われている行列分解は，音声・画像分野では次元圧縮の特徴を活かしたノイズ除去手法 [2, 14] としてよく使われている。本研究では顧客・商品ペアに対して再購買周期を予測する時，はずれ値を含むノイズを除去するために導入している。

3 提案手法

図1は本研究で提案する再購買レコメンダーシステムの構造を表している。入力データは購買時間，顧客，商品，商品数量で構成された購買データである。図の左側では各顧客に対して 2 回以上購入された商品の再購買周期の推定を行う。右側では再購買周期データと購買データから抽出された最終購買日データを元に次回の購入日を予測する。続けて各モジュールを詳細に説明する。

3.1 再購買周期計算モジュール

購買時間，顧客，商品，商品数量で構成された入力データから各顧客・商品ペアの再購買周期 ($pc_{u,i}$) を式1を利用して計算する。

$$pc_{u,i} = f_{mean}(\frac{date_{u,i,k} - date_{u,i,k-1}}{amount_{u,i,k-1}}) \quad (1)$$

$date_{u,i,k}$ は顧客 u が商品 i を k 番目に購入した時間であり，$amount_{u,i,k-1}$ は前回の注文で購入した商品の数量を表す。3 回以上購入された場合は複数の再購買周期が生成される。この

* NHN Techorus 株式会社データサイエンスチーム
† NHN Techorus 株式会社データサイエンスチーム

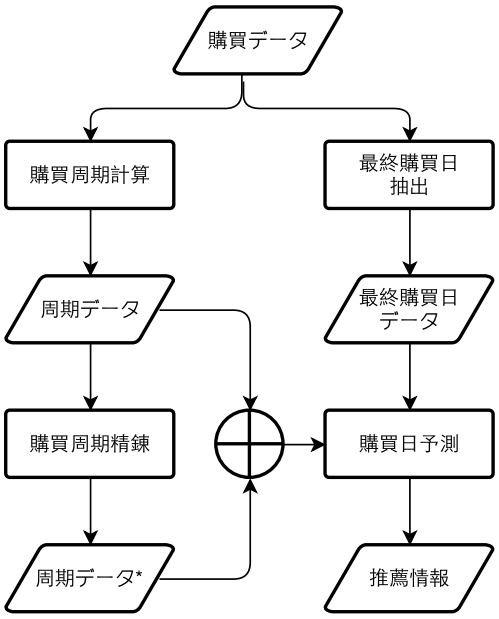

図1　再購買レコメンダーシステムのフローチャート。再購買周期は購買データから得られたもの (周期データ) または行列分解法で精錬されたもの (周期データ *) を使うことが可能。

場合，平均値 (f_{mean}) を利用することで再購買周期の揺れ (ノイズ) を軽減させることが可能だと判断した。ここで取得した再購買周期を利用したレコメンダーは実験時にベースラインモデルとして使われた。

3.2　再購買周期精錬モジュール

式1で求めた顧客・商品ペアの再購買周期は再購買回数がすくないと数値の信頼性が担保できない問題点を抱えている。また，再購買周期データはロングテールな特徴を持つため多くデータを収集するだけではこの問題を解決することは難しい。

私たちは多くの顧客が類似する再購買周期パターンを持つことに着目し，類似するパターンを持つ顧客群の情報を活用することでこの問題を緩和することが可能だと判断した。その実現のために，次元圧縮の機能を持つ行列分解を導入し下記の二段階プロセスを考案した。

1. 高次元の再購買周期 (C) 行列を低次元の行列の積 ($U \times I$) として分解
2. 低次元の行列の積 ($U \times I$) で精錬された再購買周期 (C') を再推定

ステップ 1 では，再購買周期データを $m \times n$ サイズの行列 (C) として表す。m は学習データに含まれている顧客の数であり，n は商品の数である。この行列の (u, i) 位置に存在する数値は顧客 u と商品 i に対するの再購買周期である。続けて，$r \ll m, n$ を満足する r(行列の階数) を選択し，元の再購買周期行列 (C) を $m \times r$ サイズの行列 U と $r \times n$ サイズの行列 I の積として表す。行列 C と UI の誤差を最小限する U と I を求めることでステップ 1 は終了する。

ステップ 2 では，行列分解によって生成された U と I の積を計算することで精錬された再購買周期データを得る。十分に小さな r を選択することで再推定された周期データから外れ値のようなノイズを除去することが可能である。

行列分解タスクには大規模のデータが扱える Apache Spark[*1] とそれに附属している機械学習ライブラリを使用した。

3.3　購買日予測モジュール

3.1章または3.2章で得られた再購買周期と 2 回以上の購入履歴が存在する顧客・商品ペアに対して式2を適用することで次回の購買日 ($pd_{u,i}$) を予測する。

$$pd_{u,i} = \lfloor date_{u,i,latest} + (pc_{u,i} \times amount_{u,i,latest}) \rfloor \quad (2)$$

$date_{u,i,latest}$ は顧客 u が商品 i を購入した最後の日を，$pc_{u,i}$ は顧客 u の商品 i に対しての再購買周期を，そして $amount_{u,i,latest}$ は前回の注文時に購入した商品の数を表す。予測された購買日は床関数によって日単位に変換される。

4　実験・結果分析

本研究で提案した再購買周期の精錬方法の有効性を実験を通じて確認する。実験には美容関連 e コマース分野の 3 年間の購買データ, 約 19 万件が使われた。データに含まれている顧客数約 1.7 万人で，個人よりは小規美容店の担当者が多い。商品数は約 7 千種類が存在する。レコメンダーの評価は，推薦対象日を含む前後一週間の間に再購買が見込まれる商品を予測し，推薦対象日から一週間以内に購買された場合を正例と判断して評価を行う。

比較対象になるベースラインモデルとして提案手法を適用しない再購買周期データ (3.1章を参照) を利用するレコメンダーを構築した。再購買周期は，テスト対象日直前の 2 年分のデータを元に計算した。テスト対象日は，2012 年 3 月 2 日から 15 日までの 2 週間[*2]のデータを利用した。また，顧客・商品ペアに対して複数の再購買周期が存在する時は平均値を使用した。

続けて提案手法を適用したモデルを構築した。この時，精錬結果に大きく影響を与える二つのハイパーパラメーターをグリッド探索を利用して決定した。各パラメーターとテストされた数値の範囲は下記に示されている。

- 行列の階数: $2^n, n = 4, 5, 6, 7$
- 最大繰り返し数: $8, 12, 16$

図2はこの二つのパラメーターの組み合わせ (総 12 種類) をテストした結果である。行列の階数が 32 そして最大繰り返し数は 12 の時に最大精度が得られた。今後の実験ではこの数値を利用する。また，分解された二つの行列の要素が負値を持たない制限 (非負値行列分解) を与えると学習がうまく行かない問題が発生したため今回は使用しないようにした。

提案手法を適用したモデルとベースラインモデルの再購買予測性能を比較した結果が図3(適合率) と図4(再現率) である。この結果では適合率は平均 5.5% 下降し，再現率が 0.6% 上昇したことが確認できる。適合率が大きく下降した主な原因としては，提案手法を適用したモデルがベースラインモデルと

[*1] https://spark.apache.org/

[*2] 評価では予測された再購買商品がテスト対象日から一週間後まで購入されたら正例とするため実際は 3 週間のデータが使われる。

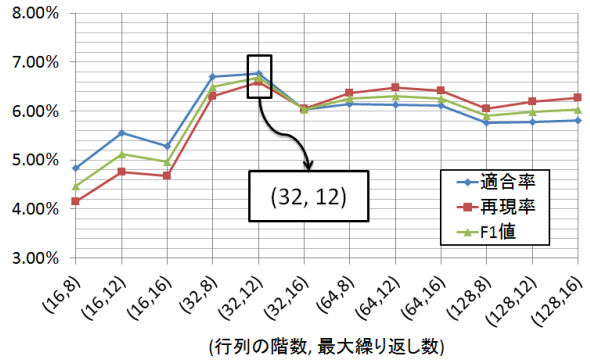

図2　ハイパーパラメータのグリッド探索結果。

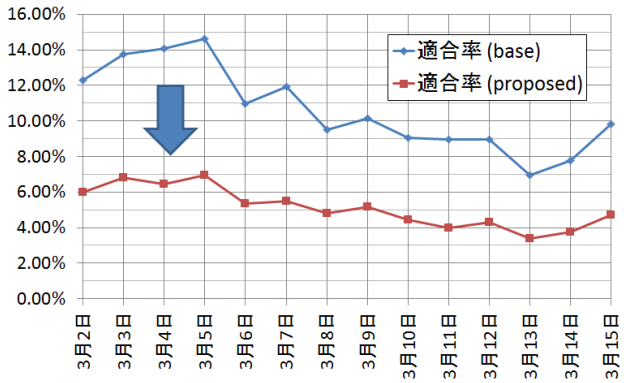

図3　適合率の比較。予測結果が約 2.5 倍になったため適合率の下降が目立つ。

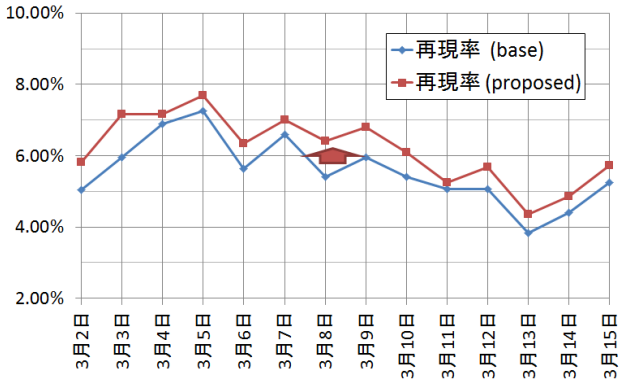

図4　再現率の比較。前区間での上昇が確認できる。

日付	ベースラインモデル			提案モデル		
	TP	FP	FN	TP	FP	FN
3 月 2 日	193	1375	3636	223	3511	3606
3 月 3 日	224	1407	3541	270	3686	3495
3 月 4 日	246	1502	3323	256	3710	3313
3 月 5 日	257	1503	3290	273	3669	3274
3 月 6 日	192	1557	3217	216	3836	3193
3 月 7 日	212	1567	3005	225	3889	2992
3 月 8 日	171	1630	2993	203	4022	2961
3 月 9 日	187	1653	2949	213	3920	2923
3 月 10 日	167	1678	2920	188	4038	2899
3 月 11 日	159	1619	2986	165	3967	2980
3 月 12 日	159	1614	2979	178	3975	2960
3 月 13 日	122	1632	3064	139	3947	3047
3 月 14 日	136	1611	2954	150	3849	2940
3 月 15 日	172	1582	3106	188	3800	3090

表1　実験結果の詳細 (TP，FP，FN はそれぞれ True Positive，False Positive，False Negative を示す)。ベースラインモデルと比較して予測結果の数 (TP+FP) が多く増えていることが確認できる。

比較して約 2.3 倍の予測結果を生成したことである。表1にある詳細な結果を確認すると，提案手法の適用によって正しい予測結果 (TP 数: 2597 → 2887) も増えているが，全予測結果数 (TP+FP 数: 24527 → 56706) がそれを上回る増加を見せたため適合率が大きく低下したことが分かる。この問題を改善する一手法として，購買日予測モジュールをを確率モデルまたはランキングモデルに変更することが考えられる。再購入が予測された商品であってもその可能性が高くないものは除外することで適合率の低下を防ぐことが可能である。

適合率と比べて再現率が低いのは評価データの構築方法が影響を与えている可能性が大きい。たとえば，評価に使われたデータには約 2400 の商品が含まれているが，この商品らが過去 2 年間顧客によって最大何回再購入されたかを調べた結果が図5である。どの顧客からも 2 回以上再購入されたことがない商品が約 14% を超えて，5 回以上再購入されたことがない商品は半数を超える。このような特徴を持つ商品としては初期段階で数回購入されその後はほぼ再購入されない耐久材[*3]のようなものがある。評価データの構築及び再購買推薦候補の選択で使われている現在のロジック (過去 2 回以上購入された商品) を改善することで再現率を適切に評価することができると考えられる。

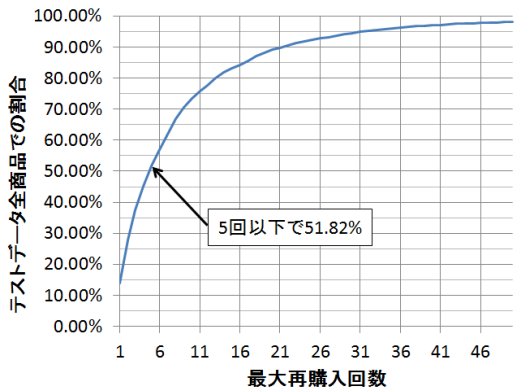

図5　テストデータ商品に対して過去 2 年間の最大再購買回数の累積分布。

[*3] 美容関係ではシャンプボール，椅子，各種機器などがある。

5 まとめ

本研究では顧客が繰り返し購入する商品を再購買時期に合わせて推薦する再購買レコメンダーシステムを対象として，行列分解を用いることで再購買時期の推定をより正確に行う一手法を提案した。実際の e コマースデータを利用して検証を行った結果，提案手法を適用してない時と比べて多くの再購入を予測することが可能であった。しかし，再購買の予測結果が多く増加することで適合率が下降する問題点も発見した。

再購買レコメンダーをさらに改善していくためにはまだ多くの課題が残されているが，その中でも幾つか重要だと考えられるものが存在する。一つは評価データの問題である。この研究では顧客が商品を 2 回以上購入したデータをすべて再購買可能性があるものと見なして使っているが，実際に再購買される可能性が低い商品が多く含まれている可能性もある。正しい評価・比較を行うためには適切な評価データの設計・構築が必要である。もう一つは多重周期の問題である。本研究では一つの顧客・商品ペアに対して一つの再購買周期が存在すると仮定しているが，実際は複数の周期に依存する場合も存在することが実験から分かってきた。例えば，季節の果物 (スイカ，みかん) や消耗品 (花火，傘) の再購入は季節という長い周期だけではなく，その中で実際繰り返して購入する短い周期にも依存する。

参考文献

[1] J. Bennett, S. Lanning, et al. The netflix prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, page 35, 2007.

[2] R. Farouk and H. Khalil. Image denoising based on sparse representation and non-negative matrix factorization. *Life Science Journal*, 9(1):337–341, 2012.

[3] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[4] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 263–272. IEEE, 2008.

[5] N. Lathia, S. Hailes, L. Capra, and X. Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217. ACM, 2010.

[6] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[7] M. Pazzani and D. Billsus. Content-based recommendation systems. *The adaptive web*, pages 325–341, 2007.

[8] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of dimensionality reduction in recommender system-a case study. In *ACM WebKDD 2000 Workshop*. ACM, 2000.

[9] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on WWW*, pages 285–295. ACM, 2001.

[10] J. B. Schafer, J. A. Konstan, and J. Riedl. E-commerce recommendation applications. In *Applications of Data Mining to Electronic Commerce*, pages 115–153. Springer, 2001.

[11] G. Shani and A. Gunawardana. *Evaluating Recommendation Systems*, pages 257–297. Springer US, Boston, MA, 2011. ISBN 978-0-387-85820-3.

[12] M. Song, X. Zhou, E. Haihong, and Z. Ou. A recommender system model based on commodity-purchase-cycle classification. In *Proceedings of the 9th EAI International Conference on Mobile Multimedia Communications*, pages 48–53, 2016.

[13] H. Stormer. Improving e-commerce recommender systems by the identification of seasonal products. In *AAAI Workshop*, pages 92–99, 2007.

[14] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 4029–4032. IEEE, 2008.

[15] Y. K. Wu, Y. Wang, and Z. H. Tang. A collaborative filtering recommendation algorithm based on interest forgetting curve. *International Journal of Advancements in Computing Technology*, 4(10):148–157, 2012.

[16] G. Zhao, M. L. Lee, W. Hsu, and W. Chen. Increasing temporal diversity with purchase intervals. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 165–174, 2012.

[17] C. L. Zheng, K. R. Hao, and Y. S. Ding. A collaborative filtering recommendation algorithm incorporated with life cycle. In *Advanced Materials Research*, volume 765, pages 630–633, 2013.