レコメンデーションに誘導されやすい顧客の抽出方式の拡張と評価

太田 光雄† 高山 毅† 村田 嘉利† 佐藤 永欣† 松本 謙治†
岩手県立大学ソフトウェア情報学部†

1 はじめに

近年, 売り上げの向上を目的とする「レコメンデーション」への注目が高まっている. これは, 商品や顧客ごとの特性に注目して, 購入される可能性が相対的に高い商品を, 店側からプッシュ型サービスとして顧客側へお勧めするものである [1]. 著者らの研究グループでは, レコメンデーションに誘導されやすい顧客の抽出方式を検討している. 具体的には, 頻出時系列パターン A→B に基づき A を既購入の顧客に B をレコメンドする際, 吉兆度という尺度を用いることを提案している [2]. 本稿では, 文献 2)の方式を拡張してレコメンド精度の向上を目指す.

2 文献 2)の方式

2.1 分析単位

本研究では, データの単位として「品番」という概念を採用している. 一部例外はあるが, 品番とはデパート内の個々のお店, 売場と考えて良い. 本研究では, 共同研究中のデパート X 社との協議より, 個々の商品単位ではなく品番単位で議論を進めている.

2.2 参考期と誘導期

頻出する時系列パターン A→B は, 時期によって変動し得る. そこで, パターンの前側として考える期間を「参考期」, 後ろ側として考える期間を「誘導期」と呼んでいる. 例えば, 参考期を 3/1～5/31, 誘導期を 7/1～8/31 としたときに, 過年に頻出するパターン A→B を使うと, 以下のことが考えられる. すなわち, 夏期に直近の春期の購入履歴を基にしてレコメンデーションを実施することである.

2.3 偏り比 $D_{A\to B}$ と正規化偏り比 $ND_{A\to B}$

パターン A→B の順方向, 逆方向の各発生回数 $m_{A\to B}$, $m_{B\to A}$ を用いて, 著者らは, 偏り比 $D_{A\to B}$ と正規化偏り比 $ND_{A\to B}$ を以下のように定義している.

$$D_{A\to B} = \frac{m_{A\to B}}{m_{B\to A}} \quad \cdots (1)$$

$$ND_{A\to B} = \frac{(m_{A\to B} - m_{B\to A})}{\max(m_{A\to B}, m_{B\to A})} \quad \cdots (2)$$

そして, $D_{A\to B}$ や $ND_{A\to B}$ に基づき, 順方向が逆方向より相対的に多い, 方向の偏りが強めの頻出時系列パターンを用いたレコメンデーションに関する検討を進めている.

2.4 吉兆度

一般に, A→B が頻出時系列パターンであるとき, 各顧客は参考期に A に加えて A, B 以外の品番 H_i(i=1, …, n-2; n は品番の総数)のいずれかでも購入していることが少なくない. そこで文献 2)では, 参考期に A に加えてどの品番 H_i でも購入していると, 誘導期に B で購入しやすいかを示す尺度として, 以下の通り「吉兆度」を提案している. すなわち, ある参考期に品番 A かつ H_i で購入した顧客が, 対応する誘導期に品番 B で

- 購入したという事実が発生した回数: Y_num(A→B, H_i)
- 購入しなかったという事実が発生した回数: N_num(A→B, H_i)

とした上で, パターン A→B に対する品番ごとの吉兆度 K(A→B, H_i)は,

$$K(A\to B, H_i) = \frac{Y\_num(A\to B, H_i)}{N\_num(A\to B, H_i)} \quad \cdots (3)$$

ただし, 右辺の分母分子がともに 0 の場合には,

$$K(A\to B, H_i) = 0 \quad \cdots (4)$$

分母のみ 0 の場合には,

$$K(A\to B, H_i) = \infty \quad \cdots (5)$$

2.5 顧客抽出とレコメンデーション

文献 2)では, 吉兆度が相対的に上位にあり, レコメンデーション時に吉兆と見なして利用する品番のことを, 「吉兆品番」と呼んでいる. そして, 誘導期の B での購入のレコメンデーションを, 参考期に A に加え吉兆品番で購入している顧客へ行なうことを提案している.

3 本稿での拡張

3.1 パターン発生数のカウント法の拡張

単一の顧客の参考期に A が count(A)回, 誘導期に B が count(B)回登場する場合, パターン A→B の発生回数のカウント法としては, 以下の四通りが考えられる.

表 1 単一の顧客ごとのパターン発生回数のカウント法
*1 binary(x)は, x があれば回数によらずに 1. なければ 0.
*2 count(x) = 品番 x での購入回数. トランザクション単位でカウントする.
*3 binary(A) = 1 の場合が以下の各値. binary(A) = 0 の場合, パターン A→B の発生回数は 0.

カウント法	参考期側評価値	誘導期側評価値	パターン A→B の発生回数*3
両側圧縮法	binary(A)*1	binary(B)	binary(B)
価値考慮法	count(A)*2	count(B)	$\frac{count(B)}{count(A)}$
参考期圧縮法	binary(A)	count(B)	count(B)
誘導期圧縮法	count(A)	binary(B)	$\frac{binary(B)}{count(A)}$

両側圧縮法は, 参考期の品番 A, 誘導期の品番 B の購入の有無のみを考慮する. 価値考慮法は, 参考期の A での購入 1 回が, 誘導期での B での何回の購入に結びつくかを考慮する. 参考期圧縮法, 誘導期圧縮法はいずれも, 両側圧縮法と価値考慮法の中間的なカウント法である. また, 全顧客を通じてのパターン A→B の発生回数は, 顧客ごとの発生回数の総和とする. 文献 2)では参考期圧縮法を採用していたが, 本稿では両側圧縮法と価値考慮法も導入し, 三者相互の差異を検討する. なお, 誘導期圧縮法は, X 社との打合せに基づき, レコメンデーション上の利用価値は高くないと考え, 検討の対象外とする.

3.2 吉兆度を算出する粒度の改良

文献 2)では品番単位で吉兆度を算出しているが, 細粒度過ぎてデータ量が充分となりにくい. そこで本稿では, 品番群 G_j という単位を用いて粗粒度化し吉兆度を算出する. 品番群とは, デパート側の組織構成に基づき, 同一カテゴリに属する品番を 1 グループにまとめたものである.

3.3 吉兆度の算出式の改良

2.4 項の(5)式は, Y_num(A→B, G_j)の値の大小による吉兆度の差異を表現できない. そこで吉兆度の算出式を

Extension and Evaluation of Expectable Customers Picking up Method in a Recommendation
†M. Ota, T. Takayama, Y. Murata, N. Sato, and K. Matsumoto(Faculty of Software and Information Science, Iwate Prefectural University)

$$K(A\rightarrow B, G_j) = \frac{Y\_num(A\rightarrow B, G_j)}{Y\_num(A\rightarrow B, G_j) + N\_num(A\rightarrow B, G_j)} \cdots (6)$$

と変更する. ここで Y_num(A→B, G_j), N_num(A→B, G_j)は, パターン発生数のカウント法により変化させる. 紙幅の都合により導出過程の詳細は省くが, 表 2 の通りとする.

表 2 単一の顧客ごとのパターン A→B に対する品番群 G_j の Y_num と N_num

	Y_num(A→B, G_j)	N_num(A→B, G_j)
両側圧縮法	1	1
価値考慮法	$\frac{count(B)}{(count(A)+count(G_j))\times\frac{1}{2}}$	$\frac{count(A)}{(count(A)+count(G_j))\times\frac{1}{2}}$
参考期圧縮法	count(B)	1

なお, 全顧客を通じての Y_num と N_num は顧客ごとの値の総和とする.

4 実データを用いた分析

レコメンデーションの精度向上には, ルール抽出した過年の傾向が, レコメンデーションの実施年にも再現されることが望ましい.

4.1 頻出時系列パターンの再現性の分析

頻出時系列パターンのランキング作成時に, サポート値をどう設定すれば過年の傾向が再現されやすいのかは, 明らかではない. そこで, 3.1 項で提案したカウント法ごとに, 表 3 の三期間で「サポート値－3 年共通パターン数」特性を分析し平均した(図 1). 具体的には, 以下の手順で分析を行った.

Step1) サポート値を S_1 とし, 偏り比と正規化偏り比のランキングを 2006～8 年で年ごとに 100 位まで作成;
Step2) 3 年間に共通するパターン数を数える;
上記のサポート値 S_1 を種々変えてみる.

表 3 分析期間

期間番号	参考期	誘導期
1	3/1～5/31	7/19～8/31
2	3/1～5/31	6/19～7/31
3	4/1～6/30	7/19～8/31

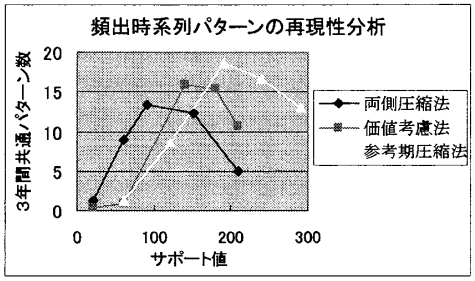

図 1 「サポート値－3 年間共通パターン数」特性

分析の結果, 三つのカウント法とも山型の形状になった. そして, 頂上になるときのサポート値は, いずれも 100 位までのランキングを作成できる限界のサポート値(以降, 「臨界サポート値」と呼ぶ)付近である. また, 参考期圧縮法は他の二つと比べて, サポート値, 3 年間共通パターン数とも高い値を取りやすい.

4.2 吉兆品番群の再現性の分析

吉兆品番群のランキング作成時に, サポート値をどう設定すれば過年の傾向が再現されやすいのかは, 明らかではない. そこで, 表 3 の期間番号 1 で, 「サポート値－順位再現の相関係数, 再現される品番群数, 再現されない品番群率」特性を分析した(図 2-3). 具体的には, 以下の手順で分析を行った.

Step1) 両側圧縮法と参考期圧縮法で, 臨界サポート値を用いた頻出時系列パターンランキングから, 3 年共通かつランキング上位のパターンを三つずつ選択;
Step2) サポート値を S_2 とした上で, 種々変える;
Step3) 2006,7 年学習データと 2008 年評価データの吉兆度順位を相対比較し, 順位再現の相関係数, 再現される品番群数, 再現されない品番群率を算出;

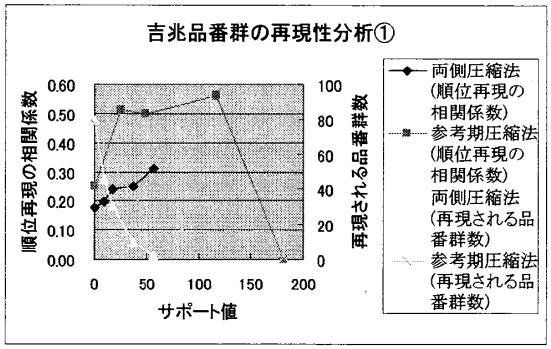

図 2 「サポート値－順位再現の相関係数, 再現される品番群数」特性

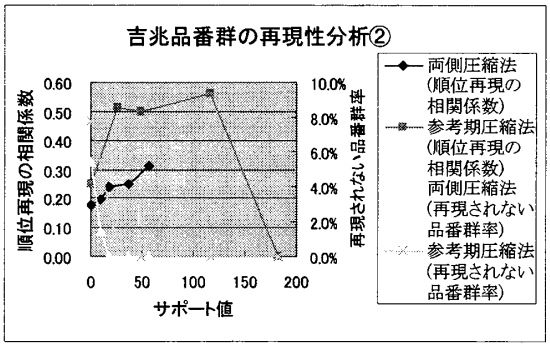

図 3 「サポート値－順位再現の相関係数, 再現されない品番群率」特性

「順位再現の相関係数」, 「再現される品番群数」, 「再現されない品番群率」の三つのバランスを考慮すると, 以下のことが言える. すなわち, 各品番群の Y_num の値の相加平均を avg(Y_num), Y_num の最大値を max(Y_num)とするとき,

$$\frac{1}{4}(\max(Y\_num) + 3 \times avg(Y\_num)) \cdots (7)$$

周辺をサポート値とすると, 比較的吉兆品番群が再現されやすい.

5 まとめと今後の展望

本稿では, レコメンデーションに誘導されやすい顧客を抽出する方式[2]を拡張した. そして, 抽出されるルールの再現状況を, デパートの実データを用いて分析した.

今後の展望として, 以下のことが考えられる: i) 価値考慮法の場合の吉兆品番群の再現性の分析, ii) レコメンデーションを実施して, 有効な結果を得ること.

参考文献

1) 土方嘉徳: 情報推薦・情報フィルタリングのためのユーザプロファイリング技術, 人工知能学会誌, Vol.19, No.3, pp.365-372, 2004.
2) 恵津森真仁, 高山毅, 村田嘉利, 佐藤永欣: レコメンデーションに誘導されやすい顧客の抽出方式と評価, 情報処理学会第 70 回全国大会 2T-7, 2008.