Department of Social Systems and Management

Discussion Paper Series

No. 1177

アクセスログデータに基づく重要顧客の“ リアルタイム ”判別

by

小池雄平, 菅谷健人, 住田潮, 高橋一樹, 平野智章, 山本浩平

July 2007

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

アクセスログデータに基づく重要顧客の“リアルタイム”判別

小池雄平 * 菅谷健人 * 住田潮 * 高橋一樹 * 平野智章 * 山本浩平 †
* 筑波大学大学院 システム情報工学研究科 〒305-8573 茨城県つくば市天王台 1-1-1
† (株) 経営科学センター 〒102-0074 東京都千代田区九段南 2-4-11 パシフィックスクエア九段南 4F

2007 年 7 月

1 はじめに

インターネットの登場により, 消費者が直接店舗に足を運ぶことなく, Web サイトを通じて商品やサービスを購入する機会が増大している. Web サイトのページアクセスに関しては, サイト訪問の日時, どの検索エンジンやキーワードを用いて訪れてきたか, 閲覧ページの時系列データ等の情報が, Web サーバにログデータとして蓄積される. このログデータを活用することにより, 閲覧顧客のページアクセスの時間的変化をリアルタイムで捉えることが可能となり, サイト訪問の進行中に顧客の行動特性を解析し, その結果に基づいてマーケティングの観点から個別的な対応を実施する道が拓かれた. e-ビジネスにおいては, このようなリアルタイム・リコメンデーションエンジンを開発し実装することが極めて重要になってきている.

EC(e-Commerce) サイトにおいて商品・サービスが繰り返し購買される場合には, 閲覧顧客の購買履歴や属性データに基づいて顧客の行動特性を把握することが可能であり, リアルタイム・リコメンデーションエンジン開発の基礎となる情報源も豊富となる. それに対し, 商品・サービスが一度しか購買されない場合には, 蓄積された豊富なデータから閲覧顧客の行動特性を把握することが不可能となり, 現在継続中のセッション履歴からのみ重要顧客であるか否かの判別を行う必要がある. このような環境におけるリアルタイム・リコメンデーションエンジンの開発は極めて困難であり, 本論文の目的はこの難題に挑戦することにある.

Web サイトを立ち上げ e-ビジネスを展開する際の経営効果を測る指標の 1 つとして, コンバージョンを挙げることができる. コンバージョンとは, 閲覧者のアクションが『何らかの成果』をもたらすことを指し, オンラインショッピングサイトであれば商品購入, 情報提供サイトやコミュニティサイトならば会員登録がコンバージョンにあたる. 商品・サービスが一度しか購買されない e-ビジネスにおいては, 限られた情報に基づいて閲覧顧客を効率的にコンバージョンへと誘導することが重要となり, そのためには, コンバージョンを起こす確率が高いと考えられる閲覧顧客をアクセス進行中にリアルタイムで判別し, 個別的なプロモーションを展開することが不可欠となる.

閲覧顧客を判別する研究は, 様々な視点からなされてきた. Gao and Sheng [1] は, ユーザーの閲覧したページ集合の中から, そのユーザーに固有のページ集合を抽出することで, 同一のプロキシサーバからアクセスしてくる不特定多数のユーザーを判別する手法を確立した. Moe and Fader [2] は, アクセスログデータの購買履歴や閲覧回数から各ユーザーがコンバージョンを起こす確率を推定する手法を提案し, Sarwar [3] は, 購入された商品の組み合わせの頻出パターンと顧客特性の類似性という 2 つの視点から, EC サイト内における商品のリコメンデーションシステムを構築することに成功した. これらの研究は, ユーザーの履歴データの蓄積に基づく判別が基礎となっており, 筆者らの知る範囲では, ページ・アクセス経路情報のみに基づいて, サイトアクセス進行中にリアルタイムで顧客のクラス判別を行う研究はなされていない.

本研究では, 現在サイトにアクセスしている閲覧顧客のこれまでの閲覧ページの時系列パターンに基づき, リアルタイムで重要顧客を判別するアルゴリズムを提唱する. サイトを閲覧する顧客を,

- C_1 : 商品に関心を持たず, 購入確率が極めて低いと思われる閲覧顧客
- C_2 : 商品に関心はあるが, それ程強い興味を持たないと思われる閲覧顧客
- C_3 : コンバージョンを起こす確率が比較的高いと思われる閲覧顧客
- C_4 : コンバージョンを起こす確率が非常に高いと思われる閲覧顧客

の 4 つに分類し, 最長 6 ページ目までのアクセスパターンによって閲覧顧客がどのクラスに属するかを判別する. 情報検索の分野で広く用いられている『サポート』と『コンフィデンス』という概念を導入し, C_1 から C_4 へ向けて次第に篩にかけるように判別する方式を採用する. アルゴリズム開発に当たっては, 平成 18 年度データ解析コンペティションにおいて提供されたアクセスログデータを用いてパラメータ設定等を行っているが, 考え方そのものは他のアクセスログデータにも適用可能である.

本論文は, 以下のように構成されている. 第 2 節では分析の対象となるアクセスログデータの概要について述べ, 第 3 節で閲覧顧客の判別アルゴリズムを開発する. 第 4 節では, テストデータに基づく検証結果を論じ, 第 5 節で結論をまとめる.

2 データ概要

本論文では, 平成 18 年度データ解析コンペティションで提供されたアクセスログデータを用いて, ページ・アクセス経路情報に基づき, サイトアクセス進行中にリアルタイムで閲覧顧客を判別するアルゴリズムを開発する. 対象となるアクセスログデータは, ASP[1]型アクセスログ解析ソフト[2] 「シビラ」を販売するサイトの Web サーバー情報を記録したもので, データの集計期間は 2006 年 1 月〜6 月の 6ヶ月間である. 以降, この 6ヶ月間をデータ期間と呼ぶ. データ内容としては, リクエスト URL[3], UNIX 時間[4], リファラ URL[5], ユーザ ID, セッション ID 等が与えられている.

シビラサイトは 86 ページから構成され, 分析の便宜上, 内容を基にこれらのページを 16 の大カテゴリに分類し, P00 から P15 で表す. 次いで重要と思われる中カテゴリを導入し, 最終的に 30 クラスに細分化した. 表 2.1 に, これら 30 クラスをまとめて置く. トップページは, インデックス (P00) である. シビラ導入 (P04) は, 購入手続き, 料金表, 導入事例等のクラスから成り, 商品に関心を持つ顧客が閲覧する可能性が高い. 目的ページ (P11) は, シビラ購入またはシビラ体験キャンペーンへの申し込みと問い合わせに関するクラスを含み, ここを閲覧する顧客はシビラに相当強い関心を持つと考えられる. 目的ページ (P11) から入力内容確認ページ (P12) を経て送信完了ページ (P13) へこの順序でアクセスされた場合は, 目的ページに含まれるクラスの 1 つが実現したことを意味し, 従って, P11 ⇒ P12 ⇒ P13 のアクセスパターンをコンバージョンと呼ぶことにする. 但し, ⇒ は順序だけを意味し, 必ずしも連続的にアクセスされる必要はないことを注意して置く.

表 2.1: 大カテゴリ（上）と中カテゴリ（下）

インデックス:P00	最新情報:P01	シビラについて:P02	シビラの機能紹介:P03
.com/index.html:P0000	.com/news/(What's New) :P0101 .com/press/(プレスリリース):P0102	.com/info/:P0200	.com/site/(サイト解析):P0301 .com/page/(ページ解析):P0302 .com/path/(経路解析) :P0303 .com/etc/(その他機能) :P0304 .com/adv/(広告効果測定):P0305 .com/seo/(SEO効果測定):P0306
シビラ導入:P04	オプションサービス:P05	その他サービス:P06	用語集:P07
.com/flow/(お申し込みの流れ):P0401 .com/price/(料金表):P0402 .com/ex/(導入事例):P0403	.com/report/(アクセス解析レポート):P0501 .com/consult/(コンサルティング):P0502	.com/suport(サポート) :P0601 .com/column(コラム):P0602 .com/mail/(メルマガ):P0603 .com/faq/(よくあるご質問):P0604 .com/cooperate/(提携サービス):P0605 .com/s_partner/(パートナー募集):P0606	.com/words/:P0700
基本情報:P08	キャンペーン:P09	アフィリエイト:P10	目的ページ:P11
.com/company.html(会社情報):P0800 .com/law.html(特定商取引法の表示):P0800 .com/privacy.html(プライバシーポリシー):P0800 .com/sitemap.html(サイトマップ):P0800 .com/contact/index.html(お問い合わせ):P0800	.com/top.html:P0900	.com/affiliate:P1000	.com/contact/apply.html(申し込み):P1100 .com/topform.html(キャンペーン申し込み):P1100 .com/contact/index(お問い合わせ):P1100
入力内容確認:P12	送信完了:P13	解約フォーム:P14	解約完了:P15
.com/cgi-bin/formmail_a8.cgi:P1200 .com/cgi-bin/formmail.cgi:P1200	.com/test/formthx.html :P1300 .com/cgi-bin/formmail_a8.cgi?thanx:P1300	.com/cancel/sibulla_cancel.html:P1400	.com/test/formthx_act.html:P1500

閲覧ページの時系列データを上記 30 クラスのアクセスパスとして表現し, 前節における閲覧顧客の 4 類型 C_1, C_2, C_3, C_4 をそれぞれ以下のように定義する.

C_1 : 目的ページ (P11) とシビラ導入 (P04) に属する料金表にアクセスしない閲覧顧客の集合

C_2 : 目的ページ (P11) を閲覧しないが料金表にアクセスする閲覧顧客の集合

C_3 : 目的ページ (P11) にアクセスするがコンバージョンを起こさない閲覧顧客の集合

C_4 : コンバージョンを起こす閲覧顧客の集合

重要顧客判別アルゴリズムの開発に際しては, データ期間を前半 (2006 年 1 月〜3 月) と後半 (2006 年 4 月〜6 月) に分ける. 前半データを学習データとし, アルゴリズムの確立とパラメータ値決定に用いる. 更に, 後半データをテストデータとして活用し, その有効性を検証する. 本研究の目的は, アクセス進行中にリアルタイムで閲覧顧客を判別することにあり, データ期間内に 1PV(ページ・ビュー) しか閲覧していない顧客は対象データから除外した. 学習データ・テストデータの PV 数, 閲覧顧客数等の詳細を表 2.2 にまとめる.

表 2.2: 学習データとテストデータ

	学習データ	テストデータ	総データ
PV数	20373	24231	44968
閲覧顧客数	3315	3741	6747
C_3顧客数	365	350	663
C_4顧客数	72	114	159
C_3の平均セッション数	2.1	2.5	2.5
C_4の平均セッション数	4.2	3.3	4.3
C_3の1セッション当たり平均PV数	6.6	6.7	6.6
C_4の1セッション当たり平均PV数	6.8	6.3	6.5

3 分析手法

本章では, 現在進行中のページ履歴のみに基づき, その顧客が, 前述した顧客クラス C_1〜C_4 のどれに所属するかを判別するアルゴリズムを開発する. 先ず, 学習データから, 各顧客クラスに固有のアクセスパターンを抽出する方法を確立し, このアクセスパターンを『特性パターン』として定義する. 次いで, 現在閲覧進行中の顧客のアクセスパターンと特性パターンを照合することにより, 閲覧顧客の所属する顧客クラスを判別するアルゴリズムを構築する.

3.1 特性パターン

本節では, 学習データに基づいて顧客クラス C_1〜C_4 の特性パターンを確立する. 閲覧顧客がコンバージョンを起こす前に重要顧客か否かを判別することが主要目的であり, 特性パターンを構成するページ時系列には, 目的ページ (P11), 入力内容確認ページ (P12), 送信完了ページ (P13) は含まれないことを注意して置く. また, 表 2.2 によれば, C_3 顧客と C_4 顧客の 1 セッション当たり平均 PV 数はそれぞれ 6.6 と 6.8 であり, これに基づいて最大 6PV までに判別を終えることとする. 判別は, 連続する 3PV の履歴を基本単位とし, これを 1〜3, 2〜4, 3〜5, 4〜6 という具合にずらし, 最大 4 回の判別で C_1 から C_4 へ向けて篩にかけるように所属する顧客クラスを特定していくことになる.

[1]Application Service Provider

[2]分析対象ページに JavaScript 等のプログラムを埋め込むことにより, アクセス情報をリアルタイムで集計・解析するソフトウェアで, Web 上でツールとしてレンタルすることも可能である.

[3]閲覧要求されたページの URL

[4]1970 年 1 月 1 日午前 0 時 00 分 00 秒からの経過秒数を基準とした時間

[5]ある Web ページのリンクをクリックして別のページに移動したときの, リンク元のページの URL

学習データに現れる閲覧顧客の T 番目の PV から $(T+2)$ 番目の PV までのページ履歴を

$$\underline{p}^{\top}(T,j) \stackrel{\text{def}}{=} [p_1(T,j), p_2(T,j), p_3(T,j)], \quad j = 1, \cdots, K(T) \tag{3.1}$$

と書く. ここで $T = 1, 2, 3, 4$ であり, $K(T)$ は 3PV で可能なアクセスパターン数 30^3 の内, 学習データに現れた実際の数である. $\underline{p}(T,j)$ の集合を

$$\mathcal{P}(T) \stackrel{\text{def}}{=} \{\underline{p}(T,1), \underline{p}(T,2), \cdots, \underline{p}(T,K(T))\} \tag{3.2}$$

と定義する. また, 閲覧顧客 c の T 番目から $(T+2)$ 番目までのページ履歴を $\underline{c}(T)$ で表し,

$$\underline{c}^{\top}(T) \stackrel{\text{def}}{=} [c_1(T), c_2(T), c_3(T)] \tag{3.3}$$

と書く. $\underline{p}(T,j)$, $\underline{c}(T)$ において, $m = 2, 3$ に対し, $(T+m-1)$ 番目の閲覧ページが存在しない場合には $p_m(T,j) = c_m(T) = \phi$ とする.

$\delta_{\{S\}}$ を命題 S が真のとき 1, それ以外は 0 の値をとる命題関数とし, 学習データの内 $\underline{p}(T,j)$ を実現した C_i 顧客数を

$$n_i(T,j) \stackrel{\text{def}}{=} \sum_{c \in C_i} \delta_{\{\underline{c}(T) = \underline{p}(T,j)\}} \tag{3.4}$$

と定義する. この $n_i(T,j)$ に対し, サポートとコンフィデンスを

$$Supp_i(T,j) \stackrel{\text{def}}{=} \frac{n_i(T,j)}{\sum_{j=1}^{K(T)} n_i(T,j)}, \tag{3.5}$$

$$Conf_i(T,j) \stackrel{\text{def}}{=} \frac{n_i(T,j)}{\sum_{i=1}^{4} n_i(T,j)} \tag{3.6}$$

で与える. $Supp_i(T,j)$ は T〜$(T+2)$PV 間において, C_i 顧客のうち $\underline{p}(T,j)$ を実現した顧客の割合であり, この値が高い場合, $\underline{p}(T,j)$ が C_i 顧客の特性を示す可能性がある. しかし, $Supp_i(T,j)$ は C_i 顧客内部における $\underline{p}(T,j)$ の実現頻度のみを問題にしており, 全顧客集合を対象とした場合, $\underline{p}(T,j)$ の実現に関し, 他の顧客クラスに対する C_i 顧客の重要性を保障するものではない. この観点から顧客クラス間の相対的重要度を指標化したものが $Conf_i(T,j)$ で, $\underline{p}(T,j)$ を実現した全顧客のうち C_i 顧客の占める割合を示す. $\underline{p}(T,j)$ が C_i の特性を表わすためには, $Supp_i(T,j)$ と $Conf_i(T,j)$ が共に高い値を持つ必要がある. すなわち, α_T, β_T をパラメータとするとき, $Supp_i(T,j) \geqslant \alpha_T$, $Conf_i(T,j) \geqslant \beta_T$ を満たす $\underline{p}(T,j)$ を T〜$(T+2)$PV 間における C_i の特性パターンとし, T〜$(T+2)$PV 間における C_i の特性パターン集合を

$$\mathcal{CP}_{i:T}(\alpha_T, \beta_T) \stackrel{\text{def}}{=} \{\underline{p}(T,j) \mid Supp_i(T,j) \geqslant \alpha_T,\ Conf_i(T,j) \geqslant \beta_T\} \tag{3.7}$$

と定義する. 後の議論の便宜上, C_i の特性パターン集合の内, C_i にのみ現れる特性パターンの集合を

$$\mathcal{UCP}_{i:T}(\alpha_T, \beta_T) \stackrel{\text{def}}{=} \{\underline{p}(T,j) \mid \underline{p}(T,j) \in \mathcal{CP}_{i:T}(\alpha_T, \beta_T),\ \underline{p}(T,j) \notin \bigcup_{k \neq i} \mathcal{CP}_{k:T}(\alpha_T, \beta_T)\} \tag{3.8}$$

とする.

具体例として, 学習データから抽出した T〜$(T+2)$PV 間におけるアクセスパターンが高々4 つしか存在しない場合を考えてみる. 表 3.1 は $\underline{p}(T,j)(j = 1,2,3,4)$ の 4 つのアクセスパターンについて, $n_i(T,j)$ を集計したものであり, 表 3.2, 表 3.3 は表 3.1 を基に算出された $Supp_i(T,j)$, $Conf_i(T,j)$ の値である. ここで, $(\alpha_T, \beta_T) = (0.5, 0.5)$ とする.

表 3.1: アクセスパターン抽出の例

顧客クラス＼アクセスパターン	$\underline{p}(T,1)$	$\underline{p}(T,2)$	$\underline{p}(T,3)$	$\underline{p}(T,4)$	計
C_1	10	5	2	3	20
C_2	15	4	5	10	34
C_3	2	1	4	5	12
C_4	10	5	18	1	34
計	37	15	29	19	100

表 3.2: 表 3.1 から算出されたサポート

顧客クラス＼アクセスパターン	$\underline{p}(T,1)$	$\underline{p}(T,2)$	$\underline{p}(T,3)$	$\underline{p}(T,4)$	計
C_1	0.50	0.25	0.10	0.15	1
C_2	0.44	0.12	0.15	0.29	1
C_3	0.17	0.08	0.33	0.42	1
C_4	0.29	0.15	0.53	0.03	1

表 3.3: 表 3.1 から算出されたコンフィデンス

顧客クラス＼アクセスパターン	$\underline{p}(T,1)$	$\underline{p}(T,2)$	$\underline{p}(T,3)$	$\underline{p}(T,4)$
C_1	0.27	0.33	0.07	0.16
C_2	0.41	0.27	0.17	0.53
C_3	0.05	0.07	0.14	0.26
C_4	0.27	0.33	0.62	0.05
計	1	1	1	1

C_1 における $\underline{p}(T,1)$ について見ると, C_1 顧客は他のアクセスパターンに比べて $\underline{p}(T,1)$ を起こす確率が高い $(Supp_1(T,1) = 0.50)$ が, C_2 顧客の方が $\underline{p}(T,1)$ を起こす確率が高い $(Conf_2(T,1) = 0.41 > 0.27 = Conf_1(T,1))$. よって, $\underline{p}(T,1)$ は C_1 固有のアクセスパターンとは言えない. C_2 の $\underline{p}(T,4)$ に着目すると, C_2 顧客は他のクラスの顧客に比べ, $\underline{p}(T,4)$ を起こす確率が高い $(Conf_2(T,4) = 0.53)$ 一方, C_2 自身の他のアクセスパターンに比べてそれ程傑出した値であるとは言えず, 実際, C_2 顧客は $\underline{p}(T,4)$ よりも $\underline{p}(T,1)$ を起こす確率がより高い値を示している $(Supp_2(T,4) = 0.29 < 0.44 = Supp_2(T,1))$. よって, $\underline{p}(T,2)$ は C_2 顧客に固有のアクセスパターンであるとは言えない. C_4 顧客の $\underline{p}(T,3)$ を考えると, 他のアクセスパターンに比べて $\underline{p}(T,3)$ を起こす確率が高く $(Supp_4(T,3) = 0.53)$, 且つ他の顧客クラスと比べても $\underline{p}(T,3)$ を起こす確率が高い $(Conf_4(T,3) = 0.62)$. よって, $\underline{p}(T,3)$ は C_4 顧客固有のアクセスパターンであると判断される.

以上の議論より, $Supp_i(T,j) \geqslant \alpha_T$, $Conf_i(T,j) \geqslant \beta_T$ を満たすアクセスパターンを抽出することで, 特性パターンを得ることができる. 後の議論の便宜上, $\mathcal{CP}_{i:T}(\alpha_T, \beta_T)$ を抽出するアルゴリズムを以下にまとめて置く.

Algorithm 1　特性パターンの抽出

[入力] $T, C_i(i=1,2,3,4), \alpha_T, \beta_T$

[出力] $\mathcal{CP}_{i:T}(\alpha_T, \beta_T)$ $(i=1,2,3,4)$

[1] $i \leftarrow 1$

[2] $\mathcal{P}(T)$ を (3.2) に基づいて抽出する

[3] LOOP: $n_i(T,j)$ を算出し, (3.5), (3.6) を用いて $Supp_i(T,j)$, $Conf_i(T,j)$ を計算する $(j = 1,2,\cdots,K(T))$

[4] (3.7) に従って $\mathcal{CP}_{i:T}(\alpha_T, \beta_T)$ を抽出する

[5] $\rightarrow (4 \geqslant i \leftarrow i+1)/LOOP$

学習データから最適パラメータセット (α_T^*, β_T^*) を特定する方法については, 3.3 節で詳しく述べる.

3.2　重要顧客の判別アルゴリズム

3.1 節で抽出した特性パターンを基に, 閲覧顧客が属する顧客クラスを最大で 6PV 以内に判別するアルゴリズムを開発する. 判別は連続する 3PV の履歴を基本単位とし, これを 1〜3, 2〜4, 3〜5, 4〜6 という具合にずらし, 最大 4 回の判別で C_1 から C_4 へ向けて所属する顧客クラスを篩にかけるように特定する. すなわち, $T=1$ 時点では, 白紙の状態から各顧客が C_1, C_2, C_3 の何れの顧客クラスに属するのかを判別する. コンバージョンは 目的ページ (P11) ⇒ 入力内容確認ページ (P12) ⇒ 送信完了ページ (P13) というアクセスパターンで定義されており, その実現には少なくとも 3PV を必要とする. 最初に目的ページ (P11) にアクセスする顧客は実質的に存在せず, $T=1$ 時点で C_4 を実現してしまう顧客は皆無である. 従って, C_4 であるか否かの判断は, $T=2$ 時点以降に持ち越されていることに注意して置く. 次いで, ある顧客 c が時点 T で C_i に属すると判断されているとき, $T+1$ 時点以降に C_i に留まるかそれより上の顧客クラスに移行するかを判断し, $T=4$ 時点での判断を最終判断とする. 但し, C_4 ではないと判断されている顧客が 6PV 前に C_4 を実現した場合には, 判別アルゴリズムの誤りとして処置し, それ以後, その顧客を考察の対象から排除する.

記法の便宜上, T〜$(T+2)$PV 間の 3 ページが全て記録されたアクセスパターンの集合を REC_T, T〜$(T+2)$PV 間の 3 ページが全て記録されていないアクセスパターンの集合を EMP_T, 料金表ページ (P0402) を含むアクセスパターンの集合を $PRICE_T$, 目的ページ (P11) を含むアクセスパターンの集合を OBJ_T, 送信完了ページ (P13) を含むアクセスパターンの集合を $CONV_T$ と定義する. すなわち, $\mathcal{M} = \{1,2,3\}$ とすると,

$$REC_T \stackrel{\text{def}}{=} \{\underline{c}(T) | \forall m \in \mathcal{M},\ c_m(T) \neq \phi\}, \tag{3.9}$$

$$EMP_T \stackrel{\text{def}}{=} \{\underline{c}(T) | \forall m \in \mathcal{M},\ c_m(T) = \phi\}, \tag{3.10}$$

$$PRICE_T \stackrel{\text{def}}{=} \{\underline{c}(T) | \exists m \in \mathcal{M},\ c_m(T) = P_{0402}\}, \tag{3.11}$$

$$OBJ_T \stackrel{\text{def}}{=} \{\underline{c}(T) | \exists m \in \mathcal{M},\ c_m(T) = P_{11}\}, \tag{3.12}$$

$$CONV_T \stackrel{\text{def}}{=} \{\underline{c}(T) | \exists m \in \mathcal{M},\ c_m(T) = P_{13}\} \tag{3.13}$$

である. アルゴリズムは, 以下の各ステップで閲覧顧客 c の T〜$(T+2)$PV 間のアクセスパターン $\underline{c}(T)$ に対し, ページ欠損の有無 (すなわち $\underline{c}(T) \in REC_T$ か否か) を調べることから始まる. ■ で判別終了を表わす.

Algorithm 2　顧客クラス判別

STEP 1 $(T=1)$:

分析対象データに現れる閲覧顧客の集合を $\mathcal{C}$ とし, 全ての $c \in \mathcal{C}$ について判別を行う.

Case I : $\underline{c}(1) \in REC_1$ の場合

i) $\underline{c}(1) \in \bigcup_{i=1}^{3} \mathcal{CP}_{i:1}(\alpha_1, \beta_1)$ ならば ii) へ; そうでない場合は iii) へ

ii) $\underline{c}(1) \in \mathcal{UCP}_{i:1}(\alpha_1, \beta_1)$ ならば $c \in C_i$ ■; そうでない場合は, 複数の顧客クラスの特性パターン集合に属しており, 学習データから得られるこのアクセスパターンのサポート比に基づいて, 該当する顧客クラスに振り分ける ■

iii) $\underline{c}(1) \in OBJ_1$ ならば $c \in C_3$ ■; そうでない場合は iv) へ

iv) $\underline{c}(1) \in CONV_1$ ならば入力ミスと見なし, 以後この c を考察の対象から排除する ■; そうでない場合は, 学習データから得られるこのアクセスパターンの閲覧顧客数の比に基づいて, C_1, C_2, C_3 に振り分ける ■

Case II : $\underline{c}(1) \notin REC_1$ の場合

i) $\underline{c}(1) \in OBJ_1$ ならば $c \in C_3$ ■; そうでない場合は ii) へ

ii) $\underline{c}(1) \in CONV_1$ ならば入力ミスと見なし, 以後この c を考察の対象から排除する ■; そうでない場合は, 学習データから得られるこのアクセスパターンの閲覧顧客数の比に基づいて, C_1, C_2, C_3 に振り分ける ■

STEP 2 $(T = 2, 3, 4)$:

$c \in CUS(T) \stackrel{\text{def}}{=} \{c \mid T$ 時点で $c \in \bigcup_{i=1}^{3} C_i\}$ についてのみ判別を行う

$\underline{c}(T) \in CONV_T$ 且つ $1 \leqslant \tau \leqslant T-1$ を満たすある τ について $\underline{c}(\tau) \in OBJ_\tau$ ならば判別アルゴリズムの誤り (C_4 顧客の見逃し) として処置し, 以後この c を考察の対象から排除する ■; $\underline{c}(T) \in EMP_T$ ならば, $(T-1)$ 時点における判別結果に基づいて c の属する顧客クラスを確定する ■; 上記の何れにも該当しない場合は, 以下を $i = 1, 2, 3$ について繰り返す

Case I : $\underline{c}(T) \in REC_T(c \in C_i)$ の場合

i) $\underline{c}(T) \in \bigcup_{r=i}^{4} \mathcal{CP}_{r:T}(\alpha_T, \beta_T)$ ならば ii) へ ; そうでない場合は iii) へ

ii) $\underline{c}(T) \in \mathcal{UCP}_{r:T}(\alpha_T, \beta_T)$ ならば $c \in C_r$ ■; そうでない場合は複数の顧客クラスの特性パターン集合に属しており, 学習データから得られるこのアクセスパターンのサポート比に基づいて, 該当する顧客クラスに振り分ける ■

iii) $\underline{c}(T) \in OBJ_T$ ならば $c \in C_3$ ■; そうでない場合は iv) へ

iv) $\underline{c}(T) \in CONV_T$ ならば入力ミスと見なし, 以後この c を考察の対象から排除する ■; そうでない場合は, 学習データから得られるこのアクセスパターンの閲覧顧客数の比に基づいて, C_r $(i \leqslant r \leqslant 4)$ に振り分ける ■

Case II : $\underline{c}(T) \notin REC_T(c \in C_i)$ の場合

i) $\underline{c}(T) \in OBJ_T$ ならば $c \in C_3$ ■; そうでない場合は ii) へ

ii) $\underline{c}(T) \in CONV_T$ ならば入力ミスと見なし, 以後この c を考察の対象から排除する ■; そうでない場合は学習データから得られるこのアクセスパターンの閲覧顧客数の比に基づいて, C_r $(i \leqslant r \leqslant 4)$ に振り分ける ■

以上を, STEP 1 と STEP 2 に分けてフローチャートとしてまとめて置く.

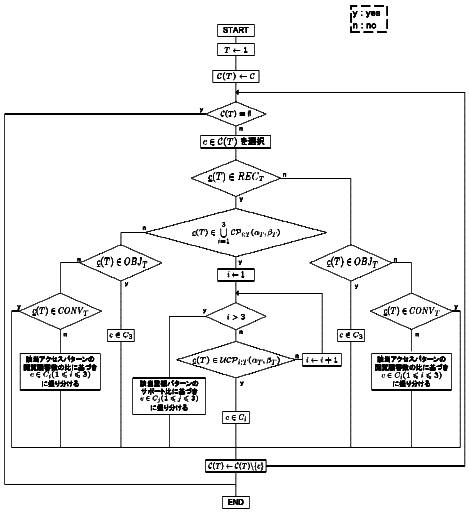

図 3.1: STEP 1 のフローチャート

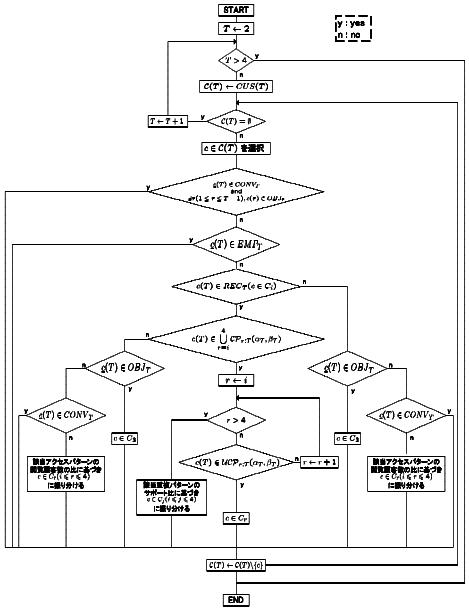

図 3.2: STEP 2 のフローチャート

とすると, 判別結果は表 3.4 のようにまとめられる. ここで,

$$X_i \stackrel{\text{def}}{=} \sum_{j=1}^{4} x_{ij} \ , \ Y_j \stackrel{\text{def}}{=} \sum_{i=1}^{4} x_{ij},$$

$$Precision_i(\alpha_T, \beta_T) \stackrel{\text{def}}{=} x_{ii}/X_i \ ,$$

$$Recall_j(\alpha_T, \beta_T) \stackrel{\text{def}}{=} x_{jj}/Y_j$$

と定義する. X_i は C_i と判別された顧客数を表わし, Y_j は実際に C_j に属する顧客数である. 従って, $Precision_i(\alpha_T, \beta_T)$ は, C_i であると判別された顧客の内, 実際に C_i に属する顧客の占める割合であり, $Recall_j(\alpha_T, \beta_T)$ は, C_j 顧客の内, 正しく C_j と判別された顧客の占める割合を表す.

表 3.5 は, $C_1 \cup C_2$ と $C_3 \cup C_4$ の集合について表 3.4 を

3.3 最適パラメータの特定

本節では, 判別アルゴリズムの評価指標として広く用いられている 2 つの概念『Recall』と『Precision』を導入し, これらの指標に基づいて学習データから最適パラメータ集合 (α_T^*, β_T^*) を特定する方法を論じる.

あるパラメータ集合 (α_T, β_T) に対し, 前節で述べた判別アルゴリズムを学習データに適用したとき, C_i であると判別された顧客の内, 実際は C_j であった顧客数を x_{ij}

表 3.4: 判別結果の集計表

		Truth				
		C_1	C_2	C_3	C_4	
Judgement	C_1	x_{11}	x_{12}	x_{13}	x_{14}	X_1
	C_2	x_{21}	x_{22}	x_{23}	x_{24}	X_2
	C_3	x_{31}	x_{32}	x_{33}	x_{34}	X_3
	C_4	x_{41}	x_{42}	x_{43}	x_{44}	X_4
		Y_1	Y_2	Y_3	Y_4	

書き改めたものである. 実際は重要顧客 $C_3 \cup C_4$ であるのに誤って非重要顧客 $C_1 \cup C_2$ と判断した場合を Type I Error(第 1 種の過誤), 重要顧客 $C_3 \cup C_4$ と判断したが実際は非重要顧客 $C_1 \cup C_2$ であった場合を Type II Error(第 2 種の過誤) と定義する. 顧客判別に基づいて商品プロモーションを行うとする場合, Type I Error は『本来, 得ることのできたはずの収益を取り逃がしてしまう機会損失』に繋がり, Type II Error は『プロモーションを展開するための投資に対する利益効果の低減』を意味する. Precision は費用対効果と正の相関があり, Recall は機会損失と負の相関がある. よって, Precision と Recall を正しく把握することは, マーケティングにおける投資効果を測る上で重要である (Mizuno, Saji, Sumita and Suzuki [4] を参照のこと).

表 3.5: Type I Error と Type II Error

		Truth	
		$C_1 \cup C_2$	$C_3 \cup C_4$
Judgement	$C_1 \cup C_2$		Type I Error
	$C_3 \cup C_4$	Type II Error	

パラメータ $\alpha_T, \beta_T (T = 1,2,3,4)$ が与えられたとき, Algorithm 1〜2 を実行し, $Precision_1(\alpha_T, \beta_T)$, $Precision_4(\alpha_T, \beta_T)$ を計算することができる. 最適パラメータ (α_T^*, β_T^*) を特定するに際しては, $T = 1$ 時点においては, C_1 と判断した顧客が実際は重要顧客であった際の機会損失を最小にすることを目的とする. よって, $Precision_1(\alpha_1, \beta_1)$ が最大となる (α_1^*, β_1^*) を求める. 第 T 段階 $(T = 2,3,4)$ では, $C_3 \cup C_4$ と判別された顧客からコンバージョンを起こす確率の高い C_4 顧客を抽出し重点的にプロモーションを展開することが必要であり, $Precision_4(\alpha_T, \beta_T)$ を最大化する (α_T^*, β_T^*) を求める. 最適化に当たっては, パラメータ (α_T, β_T) に関して Precision の値を格子探索することを柱とする. 本論文では, 格子幅を大きく取り最適値を含む領域を大局的に特定することから出発し, 次第に刻み幅を小さくして解の精度を上げる 3 段階方式で格子探索を行う. 以下に, 学習データから最適パラメータ集合 (α_T^*, β_T^*) を特定するアルゴリズムを Algorithm 3. としてまとめる. 値 v から値 w まで刻み幅 c で格子探索することを $[v, w](c)$ によって表す.

Algorithm 3 最適パラメータ特定

第 1 段階: $\alpha_T, \beta_T (T = 1,2,3,4)$ に関して $[0.1, 0.9](0.1)$ の全ての組合せについて計算を実行する. $T = 1$ のとき, $Precision_1(\alpha_1, \beta_1)$ を最大にする $(\widetilde{\alpha}_1, \widetilde{\beta}_1)$, $T = 2,3,4$ のとき, $Precision_4(\alpha_T, \beta_T)$ を最大にする $(\widetilde{\alpha}_T, \widetilde{\beta}_T)$ を特定する.

第 2 段階: $\widetilde{f} = \widetilde{\alpha}_T, \widetilde{\beta}_T (T = 1,2,3,4)$ のそれぞれについて, $[\widetilde{f} - 0.1, \widetilde{f} + 0.1](0.01)$ で発生するあらゆる組合せに対し, 計算を実行する. $T = 1$ のとき, $Precision_1(\alpha_1, \beta_1)$ を最大にする $(\widehat{\alpha}_1, \beta_1^*)$, $T = 2,3,4$ のとき, $Precision_4(\alpha_T, \beta_T)$ を最大にする $(\widehat{\alpha}_T, \beta_T^*)$ を特定する.

第 3 段階: $\beta_T^* (T = 1,2,3,4)$ を除き, $\widehat{\alpha}_T (T = 1,2,3,4)$ について, $[\widehat{\alpha}_T - 0.01, \widehat{\alpha}_T + 0.01](0.001)$ で発生するあらゆる組合せに対し, 計算を実行する. $T = 1$ のとき, $Precision_1(\alpha_1, \beta_1^*)$ を最大にする (α_1^*, β_1^*), $T = 2,3,4$ のとき, $Precision_4(\alpha_T, \beta_T^*)$ を最大にする (α_T^*, β_T^*) を特定する.

4 分析結果

学習データに対して Algorithm 1〜3 を実行し, 最適パラメータ集合 $(\alpha_T^*, \beta_T^*)(T = 1,2,3,4)$ を得る. それらの最適値を用いて, テストデータに対して Algorithm 2 を $T = 1,2,3,4$ について実行した結果を表 4.1 にまとめる. ここで『Original』の行は, テストデータ中, コンバージョンを 6PV 以内に達成した 10 人を除く全顧客数 3,731 人に占める C_i 顧客の割合である.

もし全ての顧客を C_i と判断すると, その顧客クラスに対する Recall は 100% となり, Precision は対応する『Original』の値となる. 一般的に, Recall の値を減じながらどこまで Precision の値を伸ばせるかが判別アルゴリズム開発の要諦となる.

表 4.1: 分析結果

		Truth						
		C_1	C_2	C_3	C_4	$C_3 \cup C_4$	Total	Precision
Judgement	C_1	1480	29	16	9	25	1534	96.48%
	C_2	140	579	30	10	40	759	76.28%
	C_3	145	204	221	42	263	612	36.11%
	C_4	437	263	83	43	126	826	5.21%
	$C_3 \cup C_4$	582	467	304	85	389	1827	21.29%
	Total	2202	1075	350	104	454	3731	
	Recall	67.21%	53.86%	63.14%	41.35%	85.68%		
	Original	59.02%	28.81%	9.38%	2.79%	12.17%		

先ず, 非優良顧客に着目すると, Precision の値が C_1 (96.48%) と C_2 (76.28%) で非常に高く, 非優良顧客を高い精度で選別できていることが分かる. すなわち, 6PV 以内で商品に関心を示さないと判別した顧客は, ほぼ正確に非優良顧客であった. Recall の値を見ると, C_1 (67.21%) と C_2 (53.86%) であり, 『Original』の値 C_1 (59.02%) と C_2 (28.81%) に比較してかなり高くなっている.

次いで優良顧客に目を向けると, Precision の値 C_3 (36.11%) と C_4 (5.21%) は, 一見して低く思える. しかし, 612 人 (C_3) と 826 人 (C_4) という判別顧客数に対し, 221 人 (C_3) と 43 人 (C_4) を正しく判別している. しかも, Recall 値から分かるように, それぞれのクラスに属する人数 350 人 (C_3) と 104 人 (C_4) の内, 63.14%(C_3) と 41.35%(C_4) が正しく拾い出されている. これらの数字は, 『Original』の値 C_3 (9.38%) と C_4 (2.79%) に比較して遙かに高い水準を達成しており, 判別アルゴリズムとして極めて優れていると言っても過言ではない.

この判別アルゴリズムの相対的優秀性を示すため, 例えば, ランダムに顧客クラスを割り振ることを考える. 判別アルゴリズムが C_4 と判断した同じ人数 $N = 826$ 人の顧客を無作為に抽出したとき, それに含まれる正しい C_4 顧客数を K とすれば, K は $N = 826$, $\theta = C_4$の『Original』の値 $= 0.0279$ の二項分布に従う. ここで N は十分大きいため正規近似し, それを標準化することで $K \geqslant 43$ となる確率を求めると, その値は 0.00001 となる. この事実からも, C_4 に対する Precision の値 $5.21\% = 0.0521$ を達成する本論文の判別アルゴリズムの優秀さが, 明瞭に理解されるであろう.

5 結論

本研究では, EC サイトを閲覧する顧客を重要度によって類型化し, 閲覧顧客がどの顧客クラスに属するかを, 進行中のページ・アクセス履歴のみに基づき, 最大でも 6PV 以内に判別するアルゴリズムを構築した. これにより, アクセス継続中にリアルタイムで閲覧顧客を判別し, 早い段階で効果的にコンバージョンへ導く個別的なプロモーションを展開することが可能となる. 通常の EC サイトはもとより, 商品・サービスが一度しか購入されないといった『購買履歴や属性データに基づいて顧客の行動特性を把握することが不可能な EC サイト』においても通用する, リアルタイム・リコメンデーションエンジン開発への道が拓かれたことになる.

閲覧顧客数の比に基いてランダムに顧客を判別するとすれば, 顧客の重要度を正しく推定する確率は限りなく 0 に近くなる. 本論文の判別アルゴリズムを適用することにより, EC サイトの閲覧顧客の重要度を高い精度で判別し, 個別的なプロモーションをリアルタイムで展開し, 費用対効果の増大と機会損失の減少を同時に実現することが可能となる.

謝辞 データ解析コンペティションを通じお世話になった日本オペレーションズ・リサーチ学会マーケティング・インテリジェンス研究部会の皆様に深く感謝申し上げる. 特に, 多くの質問に丁寧に対応してくださり有意義なアドバイスを頂いた生田目崇先生には, 心からの謝意を表明する. 本研究は文部科学省科学研究費補助金 (基礎研究 (C)17510114) の助成を受けている.

参考文献

[1] Gao. W and Sheng. O: “ Mining Characteristic Patterns to Identify Web Users, ”
Fordham University Research Working Paper, (2004)

[2] Moe. W.W and Fader. P.S: “ Dynamic Conversion Behavior at E-Commerce Sites, ”
Management Science, Vol.50, No.3, pp.326-335 (2004)

[3] Sarwar. B, Karypis. G, Konstan. J and Riedl. J: “ Analysis of Recommendation Algorithms for E-Commerce,”In Proceedings of the ACM EC’00 Conference. Minneapolis, MN.158-167 (2000)

[4] Mizuno. M, Saji. A, Sumita. U and Suzuki. H: “ Optimal threshold analysis of segmentation methods for identifying target customers, ”
European Journal of Operational Research(2007) to appear