論文・事例研究

関心度と忘却度に基づくレコメンド手法
―単調性制約付きレコメンドモデルの構築―

岩永　二郎，鍋谷　昴一，梶原　悠，五十嵐　健太

1. はじめに

マーケティングにおいてユーザとアイテムの関係をさまざまな視点から分析する方法が知られている．例えば，RFM 分析 [1, 2] では，Recency（最新購入日），Frequency（購買頻度），Monetary（購買金額）に基づいて，ユーザのセグメンテーションを行い，セグメントごとに適切な施策を行うことを目的としている．本研究では関心度と忘却度という新しい指標を利用することでアイテムのセグメンテーションを行う方法を提案する．提案する関心度と忘却度の指標は RFM 分析における Frequency，Recency と類似の概念であり，ユーザのアイテムに対するアプローチの度合いを定量化した指標となっている．関心度はアイテムへの興味に対して正の相関を持ち，忘却度はアイテムへの興味に対して負の相関を持つものとなる．

RFM 分析はユーザ視点の分析であり，ユーザのセグメンテーションに利用される．一方，提案手法はアイテム視点の分析であり，アイテムのセグメンテーションを行うものである．この各セグメントに適当なスコアを与えることでレコメンドに利用する．

本論文では，平成 24 年度データ解析コンペティションの課題で提供された不動産賃貸ポータルサイトのアクセスログに提案手法を適用し，サイトに訪れたユーザのアクションを予測することで，提案手法の有用性を示す．

2. 提案手法概要

提案手法は，ユーザのアイテムに対するアクションを時系列で取得できるデータから，ユーザにアイテムをレコメンドする手法である．ここでは，データとしてアクセスログを想定しており，アクションとしてサイト内におけるアイテムの閲覧行動を考える．

本論文では 2 つの提案をする．第 1 の提案は，アイテムを関心度と忘却度に基づいてセグメンテーションし，セグメントごとに算出される再閲覧確率をもとに，ユーザが過去に閲覧したアイテムからレコメンドを行う方法である．ここでセグメントは，関心度と忘却度を正の整数値として定義するので，これらのペアで定める．ユーザが閲覧したすべてのアイテムには関心度と忘却度を算出できるように定義するので，いずれかのセグメントに分類が可能となる．また，学習データからセグメントごとにアイテムの再閲覧確率を計算してテーブルを作成しておく．関心度と忘却度に対して再閲覧確率を紐付ける二次元のテーブルを再閲覧確率テーブルと呼ぶことにする．各セグメントにおける再閲覧確率の信頼性はデータ数に依存するため，データ規模が大きくなるほど信頼性の観点から望ましい．なお，本手法は，レコメンドモデル構築時，適用時ともにスケーラビリティを担保した処理が可能となっており大規模データに適している．

再閲覧確率テーブルは，次の性質が期待される．

性質（単調性制約）

- 関心度が大きい物件ほど再閲覧確率が高い
- 忘却度が小さい物件ほど再閲覧確率が高い

しかし，実際に実データから作成した再閲覧確率テーブルは，単調性制約が成り立たないセグメントも多く，サンプル数が確保できていないセグメントで顕著である．

第 2 の提案は，この問題に対し，凸二次計画法を利用して，再閲覧確率を単調性制約を満たすようにスムージングし，再閲覧確率を推定することである．推定した再閲覧確率テーブルは，学習データへの過学習を回避しており，より高精度なレコメンドモデルが構築できると期待できる．

いわなが　じろう，なべたに　こういち，かじわら　ゆう，いがらし　けんた
（株）NTT データ数理システム
〒 160–0016 東京都新宿区信濃町 35 番地信濃町レンガ館 1 階

3. 問題設定

平成 24 年度データ解析コンペティションでは，不動産賃貸ポータルサイトのアクセスログ（株式会社リクルート住まいカンパニー提供）を題材に，サイトに訪れたユーザのアクションを予測するという課題が設けられた．予測期間は 1 週間であり，予測すべきアクションは，ユーザの物件資料請求 (CV) と物件詳細ページ閲覧 (PV) の 2 種類が定められた．以下では評価指標を単に CV，PV と呼ぶことにする．具体的な課題の設定は，予測期間に CV，PV の形跡があるユーザを対象に 5 個の物件をレコメンドし，その的中率で評価するというものであった．なお，レコメンド可能な物件は予測期間に掲載されていた物件に限定された．スコアは各ユーザに対して的中した CV と PV の個数によって計算され，表 1 のとおりに定められた．

例えば，あるユーザに対して CV 物件を 2 個的中させると $30 + 12 = 42$ 点を獲得し，PV 物件を 3 個的中させると $1 + 1 + 1 = 3$ 点を獲得することができる．

本論文で扱うアクセスログの期間は 2012 年 5 月 16 日～2012 年 7 月 24 日までの 10 週間であり，全データサイズは約 10 GB になる．アクセスログは前半 9 週間と後半 1 週間に分割し，後半 1 週間分を評価用データとした．評価用データに上記スコアリング方法を適用すると 17,803 ユーザに対してスコアの最大値は 76,017 点であり，その内訳は CV によるスコアが 13,317 点，PV によるスコアが 62,700 点であった．スコアの最大値に対して，レコメンド手法を適用して得られたスコアの割合を精度と呼ぶことにする．

4. レコメンド手法の方針

ユーザに物件をレコメンドする際に，全物件をレコメンドの対象とすると計算コストが非常に大きくなる．そのため，ユーザごとにレコメンドする物件のドメインを限定する必要があり，次のドメイン限定方法について調査した．

(1) ユーザが過去に閲覧した物件をドメインとする
(2) ユーザが過去に閲覧した物件と，共起して閲覧されたことがある物件をドメインとする
(3) ユーザが検索したことがある沿線・駅の物件をドメインとする
(4) ユーザが検索したことがある地域の物件をドメインとする
(5) 全物件をドメインとする

対象とするアクセスログデータでは (1)～(5) の昇順でドメインが大きくなるほど，一つ一つの物件が予測期間に閲覧される確率が低下するため，(1)～(5) の昇順で物件の的中率も低くなる．そこで，次の 2 つのアプローチに分類した．

(A) ユーザが過去に閲覧した物件から，再閲覧する物件を割り当てる
(B) ユーザが過去に閲覧した物件以外からも，物件を割り当てる

アプローチ (A) が (1) のドメイン定義で，アプローチ (B) が (2) (3) (4) (5) のドメイン定義である．アプローチ (A) は，的中率の高い物件にドメインが限定されているというメリットがある一方で，過去に閲覧した物件が 5 物件未満の場合には，本アプローチのみでは解決できないため，アプローチ (B) と組み合わせてレコメンド処理を構築する必要がある．構築したレコメンドロジックではアプローチ (A) を適用した後にアプローチ (B) を適用した．実際にアプローチ (A) を適用すると，全予測対象の 80.5%を割り当てることができ，残り 19.5%にアプローチ (B) を適用した．以下では，本手法の核となるアプローチ (A) について詳述し，7.1 節にてアプローチ (A) の補足を，7.2 節にてアプローチ (B) の補足を行う．

表 1 CV と PV の正解個数とスコア

	1 個目	2 個目	3 個目	4 個目	5 個目
CV	30	12	9	6	3
PV	1	1	1	1	1

5. 関心度と忘却度に基づくレコメンド手法

第 1 の提案手法である関心度と忘却度に基づくレコメンド手法について述べる．特に，前章で述べた「アプローチ（A) ユーザが過去に閲覧した物件から，再閲覧する物件を割り当てる」方法について詳細に述べる．

5.1 特徴量の作成

アクセスログにはユーザがサーバーにアクセスするたびに蓄積される行動履歴データがある．行動履歴データの代表的な項目は表 2 のとおりである．ここで，セッションとは，ユーザがサイトに訪れてから離脱するまでの行動を指し，一連のサイト内行動には同一セッション ID が振られる．

2012 年 5 月 16 日～2012 年 7 月 24 日までの 10 週間の行動履歴データを加工して物件プロファイルを作成した．物件プロファイルは，全ユーザが本期間に閲覧した物件 1 つ 1 つをレコードとするデータで，前半

表 2　行動履歴データの代表的な項目

項目名	説明
セッション ID	セッションごとに付与される ID
リクエスト受付時刻	リクエストを受け付けた時刻
ユーザ ID	ユニークユーザの識別をするための ID
リクエスト ID	検索沿線・検索駅，検索市区郡，検索物件を紐付ける ID
ページ滞在時間（秒）	次のリクエストまでの間隔，セッションの最終リクエストは 0
コンバージョンフラグ	CV，PV の情報
検索情報（92 種類）	賃料，間取り，占有面積，築年数など詳細な検索情報

表 3　ユーザの物件への興味を表す特徴量

特徴量名	分類名	説明
閲覧回数	F-A	物件を閲覧した回数
セッション登場回数	F-B	物件を閲覧したセッション数
閲覧時間	F-C	物件を閲覧した合計時間
閲覧順番	R-A	物件を閲覧した直近からの順番
セッション順番	R-B	物件を閲覧した直近からのセッションの順番
経過日数	R-C	物件を閲覧した日と予測期間までの間隔

9 週間（アプローチ期間）から特徴量を作成し，後半 1 週間（結果期間）から CV，PV の情報を付与した．ただし，結果期間に掲載されていない物件は CV，PV が不可能であるためあらかじめ除外している．実際には，行動履歴データから 489,865 件のレコードを持つ物件プロファイルが作成され，CV は 179 件，PV は 32,338 件含まれていた．また，CV の件数が極端に少ないため CV と PV を同一視し，新しく CV&PV フラグを作成した．CV&PV フラグは 32,353 件となった．物件プロファイルで作成した特徴量はユーザのアイテムに対するアプローチの度合いを正の整数値として定量化したものである．ユーザごとに過去に閲覧した物件からレコメンドするため，過去閲覧物件同士の相対的な優先順位を表すような指標となっている．表 3 は実際に作成した特徴量と分類名，簡単な説明である．

行動履歴データから物件プロファイルを作成する例を述べる．表 4 はあるユーザの行動履歴データを表形式に表したものである．日付である 6/12，6/23，7/2 はアプローチ期間であり，7/18 は結果期間である．例えば物件コード 2 の物件はアプローチ期間の 6/12 のセッションで 2 回閲覧（閲覧時間 82 秒）し，6/23 のセッションで 1 回閲覧（閲覧時間 71 秒）して，結果期間の 7/18 に閲覧を行っていることを表す．行動履歴データ表 4 から物件プロファイル表 5 を作成することができる．

作成した特徴量は相関を考慮して関心度（F-A，F-B，F-C）と忘却度（R-A，R-B，R-C）に分類した．実際に表 6 で相関関係を確認することができる．関心度である F-A，F-B，F-C は互いに相関があり，忘却度 R-A，R-B，R-C も互いに相関があることがわかる．一方，関心度と忘却度の関係は無相関であることも確認できる．

5.2　特徴量の選択

前節の関心度と忘却度が無相関であることに注目して，関心度と忘却度から 1 つずつ特徴量を選出することを考える．本分析では，次の理由から関心度として閲覧回数を，忘却度としてセッション順番を採用した．

実際，関心度と忘却度の特徴量を絞り込むために物件プロファイルの CV&PV フラグを目的変数として決定木分析を行った．表 7，表 8 は決定木分析における二分木の第一分岐の情報利得比の表である．情報利得比は数値が高いほどうまく目的変数を分類していることを表す．関心度では閲覧回数・セッション登場回数が良い特徴量になっており，忘却度ではセッション順番が良い特徴量となっている．また，第一分岐を閲覧回数かセッション登場回数のどちらで分岐しても第二分岐では，セッション順番が最も高い情報利得比で分岐された．ここで，セッション登場回数と閲覧回数は

表 4　あるユーザの行動履歴

閲覧物件（閲覧順）	6/12	6/23	7/2	7/18
物件コード 1	PV（5 秒）			
物件コード 2	2PV（82 秒）	PV（71 秒）		PV
物件コード 3	2PV（57 秒）		3PV（103 秒）	PV
物件コード 4			2PV（277 秒）	

表 5　あるユーザから生成された物件プロファイル

閲覧物件	F-A	F-B	F-C	R-A	R-B	R-C	CV&PV フラグ
物件コード 1	1	1	5	4	3	36	0
物件コード 2	3	2	153	3	2	25	1
物件コード 3	5	2	160	2	1	25	1
物件コード 4	2	1	277	1	1	16	0

表 6　特徴量とピアソンの相関係数

特徴量分類名	F-A	F-B	F-C	R-A	R-B	R-C
F-A	1.00	0.80	0.58	−0.04	−0.01	−0.10
F-B		1.00	0.47	−0.03	−0.01	−0.12
F-C			1.00	−0.06	0.01	−0.06
R-A				1.00	0.57	0.23
R-B					1.00	0.31
R-C						1.00

表 7　関心度

特徴量	情報利得比
閲覧回数	0.0273
セッション登場回数	0.0245
閲覧総時間	0.0103

表 8　忘却度

特徴量	情報利得比
セッション順番	0.0137
閲覧順番	0.0124
経過日数	0.0120

定義の仕方から「セッション登場回数 $\leq$ 閲覧回数」が成り立つ．これは閲覧回数のほうがセッション登場回数よりも優先順位をより細かくつけることができることを表す．上記理由より，関心度としてセッション登場回数ではなく，閲覧回数を採用した．

特徴量の選択は，目的変数に効いている特徴量を選択するだけでなく，次節で述べる再閲覧確率テーブルのセグメント粒度を考慮して適切なセグメント数になるような特徴量を選択することが重要である．

5.3　再閲覧確率テーブルの作成

再閲覧確率テーブルは，物件プロファイルからセグメントごとに再閲覧された件数と再閲覧されなかった件数を集計して作成した．計算式は次のとおりである．

- n_{ij}：関心度セグメント i，忘却度セグメント j の物件が再閲覧された件数
- m_{ij}：関心度セグメント i，忘却度セグメント j の物件が再閲覧されなかった件数
- $p_{ij} = \frac{n_{ij}}{n_{ij}+m_{ij}}$：関心度セグメント i，忘却度セグメント j の物件の再閲覧確率

ただし，サンプルのないセグメント，すなわち $n_{ij} = 0, m_{ij} = 0$ の場合は $p_{ij} = 0$ と定義した．再閲覧確率テーブルは，評価実験の予測期間を考慮して 2012 年 5 月 16 日〜2012 年 7 月 17 日までの 9 週間の行動履歴データから加工した物件プロファイルより作成される．前半 8 週間をアプローチ期間，後半 1 週間を結果期間としている．表 9 は実際に作成した再閲覧確率テーブ

表 9　関心度と忘却度セグメントに対する再閲覧確率テーブル（実績値）

関心度 \ 忘却度	1	2	3	4	5	6	7	8	9	10	11	12
1	8.7%	6.6%	5.5%	4.1%	3.7%	3.1%	3.0%	2.8%	2.5%	2.4%	2.1%	2.3%
2	17.3%	12.8%	10.3%	8.8%	7.1%	6.5%	5.5%	5.4%	4.9%	4.9%	3.8%	4.3%
3	24.4%	20.0%	15.3%	13.8%	10.0%	9.3%	8.1%	8.3%	8.4%	7.5%	5.6%	5.1%
4	29.9%	22.7%	18.3%	16.3%	14.0%	12.2%	8.9%	10.3%	10.2%	10.5%	12.7%	8.5%
5	35.3%	29.6%	21.1%	19.9%	14.3%	11.5%	12.6%	12.0%	8.3%	16.6%	7.1%	8.6%
6	37.4%	27.7%	23.3%	20.6%	18.0%	11.8%	10.9%	19.6%	9.2%	11.5%	12.3%	19.1%
7	38.7%	29.3%	23.6%	27.8%	18.4%	14.2%	16.9%	11.2%	19.0%	26.0%	22.6%	12.1%
8	47.3%	30.8%	28.7%	21.4%	11.8%	16.9%	8.1%	20.0%	13.5%	12.5%	18.4%	8.3%
9	46.4%	31.5%	38.9%	23.1%	14.0%	20.0%	19.3%	15.6%	20.8%	9.1%	9.1%	0.0%
10	52.4%	38.0%	29.5%	11.7%	20.4%	16.7%	0.0%	29.2%	11.1%	20.8%	0.0%	27.3%
11	51.6%	33.3%	29.7%	29.4%	25.0%	16.2%	23.3%	16.7%	33.3%	22.2%	0.0%	7.1%
12	58.1%	41.6%	28.0%	17.9%	26.8%	28.0%	47.1%	12.5%	7.7%	40.0%	25.0%	7.1%
13	47.7%	45.9%	42.1%	27.3%	20.0%	27.8%	20.0%	14.3%	10.0%	37.5%	0.0%	33.3%
14	65.6%	35.3%	21.2%	13.3%	11.8%	27.8%	22.2%	8.3%	0.0%	0.0%	0.0%	50.0%
15	59.3%	34.2%	22.2%	17.6%	43.8%	26.7%	23.1%	22.2%	0.0%	20.0%	20.0%	0.0%
16	53.7%	52.0%	35.7%	50.0%	0.0%	37.5%	33.3%	0.0%	20.0%	0.0%	33.3%	100%
17	58.0%	50.0%	76.5%	30.0%	44.4%	12.5%	16.7%	22.2%	0.0%	0.0%	0.0%	0.0%
18	72.4%	20.0%	42.9%	46.7%	0.0%	50.0%	0.0%	0.0%	14.3%	0.0%	25.0%	0.0%
19	65.6%	40.0%	54.5%	0.0%	9.1%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	25.0%
20	61.5%	70.6%	23.1%	40.0%	50.0%	0.0%	40.0%	0.0%	0.0%	0.0%	0.0%	25.0%

ルである．なお，関心度と忘却度は正の整数値を範囲とするが，大部分の物件をレコメンドするために十分な大きさを確保し，関心度（1〜20），忘却度（1〜12）の範囲を定めた．

関心度と忘却度について単調性制約がおおよそ満たされていることが確認できる．本テーブルを利用することで，任意の 2 つの物件に対して優先順位をつけることができる．例えば，関心度セグメント 2，忘却度セグメント 2 の物件（再閲覧確率 12.8%）と関心度セグメント 3，忘却度セグメント 3 の物件（再閲覧確率 15.3%）を比べると，後者の物件のほうが優先順位が高いことがわかる．

5.4　レコメンド手法

レコメンド手法の処理ステップについて述べる．ここでは，再閲覧確率テーブルがすでに作成されていることを前提とする．

STEP1　レコメンド対象ユーザが過去に閲覧した物件を列挙する

STEP2　過去閲覧物件の関心度と忘却度を算出し，どのセグメントに入るか調べる

STEP3　再閲覧確率テーブルを参照して，過去閲覧物件に再閲覧確率を対応づける

STEP4　過去閲覧物件の再閲覧確率が高い物件から順番にレコメンドする

本レコメンド手法は，静的に保持している再閲覧確率テーブルを利用することで，ユーザごとに独立にレコメンドを行うことができる．

実際にレコメンドを行う場合，物件の関心度と忘却度を算出すると作成した再閲覧確率テーブルのセグメント外となるケースが現れる．例えば，本論文で行った分析では再閲覧確率テーブルとして関心度は 1〜20，忘却度は 1〜12 のセグメントを用意したが，テーブルの定義外のセグメントとなる物件が現れて再閲覧確率の参照が不可能なケースが生じる．忘却度がテーブルの定義外となる場合には，再閲覧確率が 0 に近づくのでレコメンド対象として無視しても予測精度として問題は少ない．一方，関心度がテーブルの定義外に出る場合には，再閲覧確率が 1 に近づくのでレコメンド対象として無視することができない．7.3 節で述べるレコメンド手法の評価実験では，関心度 21，忘却度 1 の物件は関心度 20，忘却度 1 の再閲覧確率を参照するといった例外的な処理を行っている．

6.　単調性制約付きレコメンドモデルの構築

第 2 の提案手法である単調性制約付きのレコメンドモデルの構築方法について述べる．

6.1　再閲覧確率テーブルの問題点

実データを集計して作成した再閲覧確率テーブルは，サンプル数が確保できていないセグメントで単調性制約を満たさないことが多く，それらのセグメントの再

表 10 関心度と忘却度セグメントのサンプル数

関心度 \ 忘却度	1	2	3	4	5	6	7	8	9	10	11	12
1	67,515	48,666	36,903	30,231	23,853	19,312	16,345	13,837	11,974	10,399	9,190	7,872
2	14,701	10,242	7,719	6,069	5,025	4,040	3,592	2,985	2,563	2,121	1,846	1,762
3	5,282	3,607	2,588	2,076	1,748	1,355	1,224	979	898	791	638	568
4	2,612	1,689	1,239	965	744	581	496	448	332	353	268	235
5	1,514	886	615	508	385	323	302	251	204	169	112	128
6	871	545	374	282	244	204	174	163	120	104	73	68
7	608	358	212	180	163	113	124	89	79	77	62	58
8	412	214	174	117	102	89	74	60	52	48	38	24
9	332	143	108	78	57	50	57	45	24	33	33	17
10	227	150	88	60	54	36	37	24	27	24	16	11
11	157	87	64	51	28	37	30	24	9	9	9	14
12	124	77	50	28	41	25	17	16	13	5	8	14
13	109	61	38	33	20	18	20	14	10	8	3	3
14	90	51	33	15	17	18	9	12	6	7	7	4
15	54	38	27	17	16	15	13	9	5	5	5	6
16	54	25	14	16	5	8	6	14	5	5	3	1
17	50	18	17	10	9	8	6	9	3	2	1	3
18	58	15	7	15	5	4	1	2	7	2	4	3
19	32	15	11	6	11	3	3	4	2	4	3	4
20	39	17	13	5	6	0	5	2	3	1	1	4

閲覧確率の信頼性は低い．表 10 は再閲覧確率を計算する際の各セグメントのサンプル数である．

例えば，表 9 では関心度 16 忘却度 12 のセグメントの再閲覧確率は 100%であるが，表 10 を参照するとサンプル数 1 件から計算された確率であり，明らかに学習データを過学習していることがわかる．

関心度と忘却度は値が大きくなるほどサンプル数が少なくなるため，学習データの規模をいくら大きくしても常にその境界ではサンプル数が少なくなり再閲覧確率の信頼性が低下するという問題が起きる．

6.2 再閲覧確率テーブルの推定

前節で述べた問題を再閲覧確率テーブルを推定することで解決する．そこで，単調性制約を満たす再閲覧確率テーブルを推定する問題を凸二次計画問題として定式化した．推定する再閲覧確率テーブルの要件は次の 3 つである．

- 単調性制約を満たす
- サンプル数が多いセグメントの再閲覧確率ほど信頼する
- 任意の 2 つのセグメントにおける再閲覧確率の順序ができるだけ保存されている

上記を考慮してモデリングしたのが次の定式化である．

定式化（単調性制約を満たす再閲覧確率テーブルの推定）

- 集合

  $I(=\{1,2,3,\ldots,I_{\max}\})$：関心度のセグメント

  $J(=\{1,2,3,\ldots,J_{\max}\})$：忘却度のセグメント

  ここで，$I_{\max}$，$J_{\max}$ はそれぞれ事前に定めた関心度，忘却度のセグメント数である．
- 定数

  p_{ij} $(i \in I, j \in J)$：関心度 i，忘却度 j のセグメントの再閲覧確率

  w_{ij} $(i \in I, j \in J)$：関心度 i，忘却度 j のセグメントのサンプル数

  ϵ：適当な微小な正の値
- 変数

  $x_{ij} \in \mathbb{R}$ $(i \in I, j \in J)$：関心度 i，忘却度 j のセグメントの推定する再閲覧確率
- 制約

  $0 \leq x_{ij} \leq 1$ $(i \in I, j \in J)$：確率の定義

  $x_{i_1 j} + \epsilon \leq x_{i_2 j}$ $(i_1 < i_2 (\in I), j \in J)$：関心度について推定する再閲覧確率は狭義単調増加

  $x_{ij_2} + \epsilon \leq x_{ij_1}$ $(i \in I, j_1 < j_2 (\in J))$：忘却度について推定する再閲覧確率は狭義単調減少
- 目的関数（minimize）

  $\sum_{i \in I, j \in J} (w_{ij}^2 \cdot (p_{ij} - x_{ij})^2)$：実績確率と推定確率との差の重み付き二乗誤差最小化

なお，目的関数にてセグメントのサンプル数で重み付けをしているため，サンプルがないセグメントの確率（5.3 節で 0 と定義）は目的関数に影響を与えなくなり，テーブルの前後左右のセグメントの確率と単調性制約

表 11　関心度と忘却度セグメントに対する再閲覧確率テーブル（推定値）

関心度 \ 忘却度	1	2	3	4	5	6	7	8	9	10	11	12
1	8.7%	6.6%	5.5%	4.1%	3.7%	3.1%	3.0%	2.8%	2.5%	2.4%	2.2%	2.2%
2	17.3%	12.8%	10.3%	8.8%	7.1%	6.5%	5.5%	5.4%	4.9%	4.9%	4.0%	4.0%
3	24.4%	20.0%	15.3%	13.8%	10.0%	9.3%	8.2%	8.2%	8.2%	7.5%	5.6%	5.1%
4	29.9%	22.7%	18.3%	16.3%	14.0%	12.1%	10.0%	10.0%	10.0%	10.0%	10.0%	8.5%
5	35.3%	29.1%	21.1%	19.9%	14.3%	12.1%	12.1%	12.0%	11.7%	11.6%	10.0%	8.6%
6	37.4%	29.1%	23.3%	20.6%	17.4%	13.6%	13.6%	13.6%	11.7%	11.7%	11.6%	11.6%
7	38.7%	29.3%	24.2%	24.2%	17.4%	15.9%	15.9%	15.8%	15.8%	15.8%	15.8%	11.6%
8	47.0%	30.8%	28.7%	24.2%	17.4%	16.9%	15.9%	15.9%	15.8%	15.8%	15.8%	11.7%
9	47.0%	32.5%	32.5%	24.2%	17.4%	17.4%	15.9%	15.9%	15.9%	15.8%	15.8%	11.7%
10	52.2%	36.8%	32.5%	24.2%	20.4%	17.4%	15.9%	15.9%	15.9%	15.9%	15.8%	11.9%
11	52.2%	36.8%	32.5%	26.2%	25.0%	19.0%	19.0%	15.9%	15.9%	15.9%	15.9%	11.9%
12	53.5%	40.8%	32.6%	26.2%	26.0%	25.9%	25.9%	15.9%	15.9%	15.9%	15.9%	11.9%
13	53.6%	40.8%	32.6%	26.2%	26.0%	26.0%	25.9%	15.9%	15.9%	15.9%	15.9%	14.9%
14	61.2%	40.8%	32.6%	26.2%	26.0%	26.0%	26.0%	15.9%	15.9%	15.9%	15.9%	14.9%
15	61.2%	40.8%	32.6%	26.5%	26.5%	26.1%	26.0%	15.9%	15.9%	15.9%	15.9%	14.9%
16	61.2%	47.8%	41.0%	40.9%	26.5%	26.1%	26.0%	15.9%	15.9%	15.9%	15.9%	14.9%
17	61.2%	47.8%	47.8%	41.0%	26.6%	26.1%	26.0%	21.0%	15.9%	15.9%	15.9%	15.0%
18	68.5%	47.8%	47.8%	41.0%	26.6%	26.5%	26.0%	21.0%	16.0%	15.9%	15.9%	15.0%
19	68.5%	47.8%	47.8%	41.0%	26.6%	26.6%	26.0%	21.0%	16.0%	15.9%	15.9%	15.9%
20	68.6%	68.5%	47.8%	45.9%	45.9%	43.9%	40.0%	21.0%	16.2%	16.2%	16.2%	16.2%

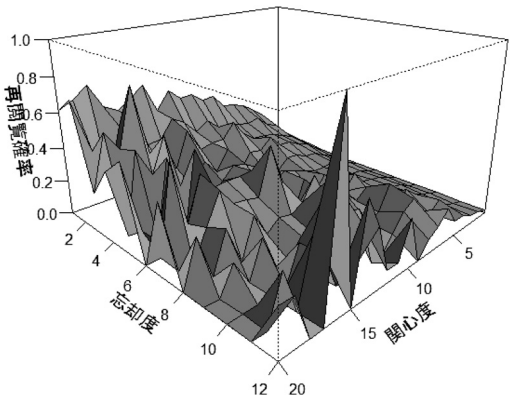

図 1　関心度と忘却度セグメントに対する再閲覧確率（実績値）

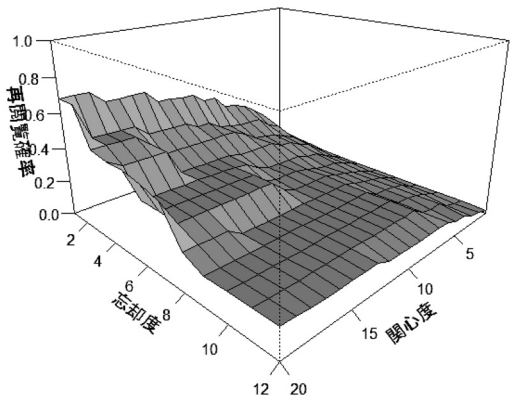

図 2　関心度と忘却度セグメントに対する再閲覧確率（推定値）

によって確率が補間される．

求解には株式会社 NTT データ数理システム開発の数理計画ソルバー Numerical Optimizer 略称 (NUOPT) を利用し，アルゴリズムは信頼領域内点法を用いた．上記の凸二次計画問題を解いて得た再閲覧確率テーブルが表 11 である（小数点第二位を切り捨て）．

関心度について再閲覧確率が単調増加し，忘却度について再閲覧確率が単調減少していることを確認できる．

図 1，図 2 は実績値の再閲覧確率テーブルと推定値の再閲覧確率テーブルを視覚化したものである．再閲覧確率がスムージングされていることが確認できる．

7.　レコメンド手法の評価

レコメンドするうえで提案手法のみでは情報が不足する場合があるため，その補足をした後に評価結果について述べる．

7.1　アプローチ (A) の補足

関心度として物件閲覧回数を，忘却度としてセッション順番を選択しているが，この 2 つの特徴量だけでは，物件のレコメンドが完了しないケースがある．例えば，あるユーザについて，過去に閲覧した物件のすべてが関心度 1，忘却度 1 のセグメントに分類されてしまっ

た場合，どの物件の再閲覧確率も等しいため優先順位をつけることができない．このような場合には，物件の閲覧時間が長い順に優先順位をつけている．物件の閲覧時間は，関心度として採用可能な特徴量であるが，次の理由で採用していない．

- アクセスログの仕組み上，正確な値がとれていない
- 粒度が細かい（単位が秒）ためカテゴリ化する必要があるが，適切なセグメント数になるような妥当なカテゴリ化のルールを定めることが難しい

しかし，基礎的な分析からほかの特徴量に比べて CV との相関が高かったため閲覧回数より優先順位を下げてレコメンドの指標に採用した．なお，物件閲覧総時間を利用しても物件の優先順位が決定しない場合には，閲覧順番の小さい値，すなわち直近で閲覧した物件をレコメンドするようにした．

7.2 アプローチ (B) の補足

アプローチ (A) で 5 個の物件をレコメンドできなかった場合，次の順番でレコメンド処理をしている．

1. ユーザにレコメンドした物件が 1〜4 個の場合
   ユーザが過去に閲覧した物件とレコメンド対象物件の間の共起確率の計算と，過去閲覧物件の再閲覧確率を計算する．共起確率と再閲覧確率を元に優先順位付け関数を構築し，優先度が高い物件から順番にレコメンドを行う．
2. 沿線・駅検索履歴がある場合
   対応する沿線・駅における人気度の高い物件をレコメンドする
3. 市区郡検索履歴がある場合
   対応する市区郡における人気度の高い物件をレコメンドする
4. その他のユーザ
   全物件で最も人気度の高い物件をレコメンドする

ここでは，共起確率を利用してレコメンドする方法について詳細な解説は与えないが，共起確率だけでレコメンドするよりも再閲覧確率を重みとして利用することで精度が向上することが観測されている．再閲覧確率テーブルは，精度の高いアプローチ (B) を実現するための特徴量として利用できることにも注意されたい．また，上記における人気度は，物件の閲覧回数と掲載期間を利用して算出している．

7.3 評価

次の 4 つのレコメンド手法を評価した．

- 比較手法 1：閲覧順番が直近の物件から順番にレコメンドする
- 比較手法 2：閲覧回数が多い物件から順番にレコメンドする
- 提案手法 1：関心度と忘却度に基づくレコメンドモデル（再閲覧確率テーブル：実績値）
- 提案手法 2：関心度と忘却度に基づくレコメンドモデル（再閲覧確率テーブル：推定値）

表 12 レコメンド手法の比較

	合計スコア	CVスコア	PVスコア	精度
比較手法 1	11,937	1,791	10,146	15.70%
比較手法 2	13,146	1,704	11,442	17.29%
提案手法 1	14,093	1,893	12,200	18.54%
提案手法 2	14,204	1,923	12,281	18.69%

なお，7.1 節，7.2 節で述べたレコメンド処理も実装したうえで手法を比較する．比較手法 1・2 をベースラインとし，関心度と忘却度に基づくレコメンドモデルを提案手法とした．また，提案手法は，実績値の再閲覧確率テーブルを利用した場合と推定値の再閲覧確率テーブルを利用した場合でそれぞれのスコアと精度を算出した．表 12 はレコメンド手法の比較結果である．

提案手法 1・2 はベースラインである比較手法 1・2 と比べて非常に高いスコアを獲得することができた．提案手法 1 と提案手法 2 を比較すると，凸二次計画法を利用して推定した再閲覧確率テーブルを利用したほうが，より精度が高いことを確認することができる．これは，実績値から作成した再閲覧確率テーブルでサンプル数が少ないセグメントの過学習を解消した結果であると考えられる．

なお，提案手法 2 のスコアのアプローチ別の内訳は，アプローチ (A) によりレコメンドの 80.5%を行い，スコア 13,937 点を獲得しており，アプローチ (B) により残り 19.5%をレコメンドを行ってスコア 267 点（合計スコアの 1.9%）を得ている．また，アプローチ (B) では CV の的中によるスコア獲得は得られなかった．以上から，対象とするデータと問題設定に対しては，アプローチ (A) が有効であることが確認できる．

8. まとめ

不動産賃貸ポータルサイトにおける物件閲覧，および資料請求の予測という問題設定に対して再閲覧確率テーブルを用いたレコメンドモデルを構築した．機械学習を利用してレコメンドモデルを構築する場合，数百にもおよぶ特徴量を作成するケースがあるが，提案手法は閲覧回数（関心度）・セッション順番（忘却度）に加え，閲覧時間の 3 つの特徴量しか使っていない．

また，機械学習の適用時には過学習の回避が課題となるが，われわれの知る限りでは学習する際に単調性制約などの構造を組み入れることで過学習を回避する取り組みは知られていない．

再閲覧確率テーブルの推定において，セグメントごとに計算される再閲覧確率に単調性制約付きでフィッティングをしているが，連続関数でフィッティングを行ってもよい．連続関数でフィッティングすることができれば，作成する再閲覧確率テーブルのセグメントの範囲を考慮する必要がなくなり，セグメントの粒度も考慮する必要がないというメリットがある．しかし，一般に連続関数でフィッティングする際には対象とするデータや選択する特徴量によってフィッティングする関数を調べることが必要であり，適切な連続関数が存在する保証ができない．提案手法はデータの規模とセグメントの粒度に注意して特徴量を選択すれば，フィッティング関数を意識する必要がないという意味で汎用的であり，計算された確率をできるだけ直接扱うことで高精度なレコメンドを実現する手法である．

レコメンドに用いた2つの指標，関心度と忘却度はアクセスログに限らず，多くのユーザの購買行動を表現するための指標となりうる．本問題では，関心度として閲覧回数を，忘却度としてセッション順番を選択したが，問題に合わせて別の単調性制約を持つ特徴量を作成すればよい．提案手法をより一般的に解釈をすれば，次の2つのステップを踏むことが本質的である．まず，いくつかの特徴量に基づいてセグメント分割を行い，セグメントごとに再閲覧確率を計算する．次に，セグメント間の再閲覧確率に対して，制約を入れた数理計画問題を解くことで再閲覧確率の推定を行う．このとき，データの規模とセグメントの粒度を考慮した特徴量の作成が重要である．

最後に，実務への適用を見据え，処理のスケーラビリティについて触れる．提案手法には，レコメンドモデル構築時の 再閲覧確率の集計処理と再閲覧確率の推定処理，およびレコメンド時の再閲覧確率の高い物件の算出処理がある．Hadoop 等の分散処理技術を適用することで大規模なログ集計は可能である．一方，6.2 節の定式化において再閲覧確率の推定は変数数 $I_{\max} \times J_{\max}$ の凸二次計画問題として定式化されるため NUOPT 等の汎用ソルバーで実用的な時間内に解ける．実際に，本論文で取り扱った凸二次計画問題は，変数数 240 の規模であり，CPU：Intel Core i7-3930K 3.20 GHz, RAM：32 GB の実験環境で 0.56 秒で求解できた．関心度と忘却度の定義次第ではセグメント数が多くなり汎用ソルバーで求解困難なケースが予想されるが，関心度と忘却度の指標を適切なセグメント数になるようにまとめ上げることで求解可能な問題規模に落とすことが可能である．上記よりモデル構築時のスケーラビリティが確保されていることがわかる．また，レコメンド時の処理も再閲覧確率テーブルを静的なデータとして保持し，テーブル参照とソート処理だけでレコメンド物件を算出できる．すなわち，リアルタイム性が求められる場面でも適用可能である．

提案する手法はモデル構築時のスケーラビリティ，モデル適用時のリアルタイム性を担保できる大規模データにふさわしい手法であるといえる．

参考文献

[1] 阿部誠：CRM のデータ分析に理論とモデルを組み込む消費者行動理論にもとづいた RF 分析，*CIRJE-J-121* ワーキングペーパー，2004.

[2] 杉浦登：RFM 分析手法の提案，オペレーションズ・リサーチ学会 2008 年秋季研究発表会予稿集，56–57, 2008.

[3] Pedro Domingos: A Few Useful Things to Know about Machine Learning. *Communications of the ACM*, **55**(10), 78–87, 2012.

[4] 松本健，西郷彰：データ解析コンペティション課題設定部門―EC サイト顧客の顧客セグメントの予測―，オペレーションズ・リサーチ，**58**(2), 68–73, 2012.

[5] 正木俊行，伊豆永洋一，佐藤俊樹，鮭川矩義，石濱友裕，田中彰浩，中島雄基，舟橋史明：アクセスログデータ可視化の試み，オペレーションズ・リサーチ，**58**(2), 74–79, 2012.

