オペレーションズ・リサーチ学会
秋季発表会

関心度(Frequency)と忘却度(Recency)に基づくレコメンド手法
-サンプリングでは対応できないビッグデータの活用-

2013 年 9 月 12 日
株式会社 NTTデータ数理システム
*岩永二郎　鍋谷昴一　梶原悠　五十嵐健太

お知らせ

社名変更

2013年9月1日をもって
「数理システム」から「NTTデータ数理システム」に
社名変更しました.

移転

2013年9月1日をもって
「東京都新宿区新宿２丁目４－３フォーシーズンビル１０階」
から
「東京都新宿区信濃町３５番地　信濃町煉瓦館１階」
に移転しました.

近くにお越しの際には是非ともお立ち寄りください

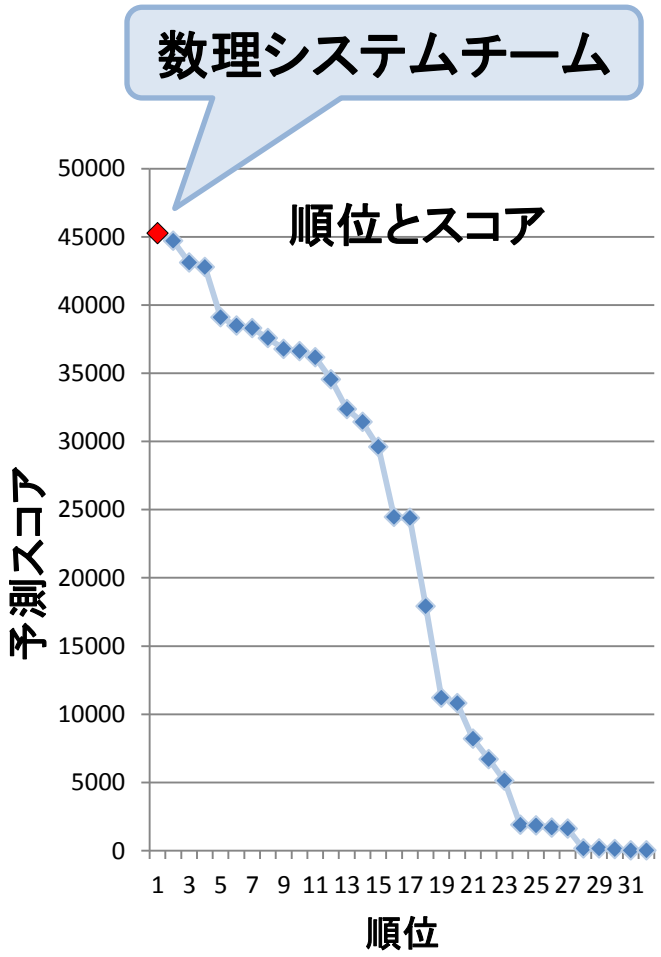
本日と内容

1. はじめに
2. 課題の紹介
3. 分析の概要
4. 関心度と忘却度に基づくレコメンド手法
5. 過学習の回避
6. まとめ

1. はじめに

1.1. データ解析コンペテション

- 第１９回 データ解析コンペティション
  - 76チームがエントリー・総勢400名が参加
- 課題設定部門（32チーム参加）
  - 評価方法　：予測スコアと分析内容
  - データ　：不動産賃貸ポータルサイト
- 数理システムチーム
  - チーム名　：明日分かることは今日予測しない
  - 代表者　：岩永二郎
  - メンバー　：鍋谷昴一・梶原悠・五十嵐健太
- 結果
  - 予選　：殊勲賞（１位）受賞
  - 本戦　：最優秀賞（１位）受賞



1.2. コンペの成果紹介

- マーケティングの事例

  頻度（Frequency）と直近さ（Recency）に基づいて顧客をセグメンテーションする手法が知られている.

  Frequency と Recency を具体的に定量化して
  レコメンドロジックとして実装した事例報告

- ビッグデータの事例

  “ビッグデータを利用して○○した”という宣伝はよく聞くが・・・

  - 実際，どのように利用したのか不明
  - サンプリングで良かったのでは？という疑問

  大規模データの特性を活かした手法の事例報告

2. 課題の紹介

2. 題材とデータ

- 題材：不動産賃貸ポータルサイトのアクセスログ

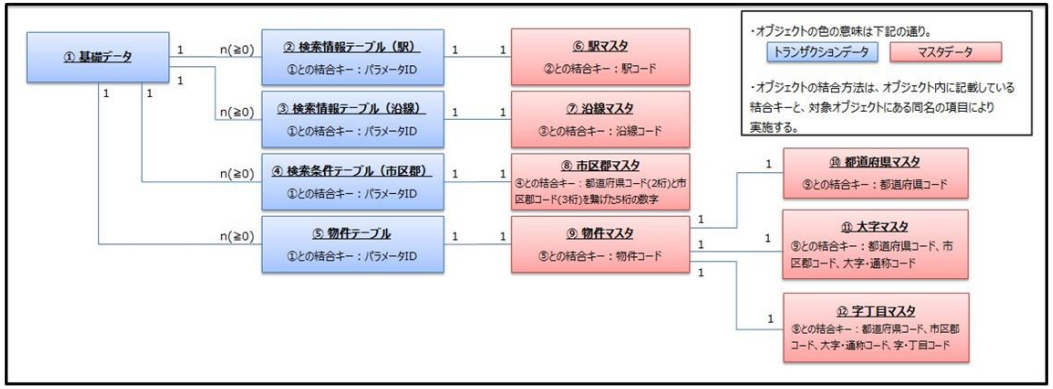
  ポータルサイト上のユーザの活動を観察

  1. サイトへの流入
  2. 物件の検索
  3. 物件の詳細閲覧（**PV：ページビュー**）
  4. 物件の資料請求（**CV：コンバージョン**）
  5. サイトからの離脱

  予測

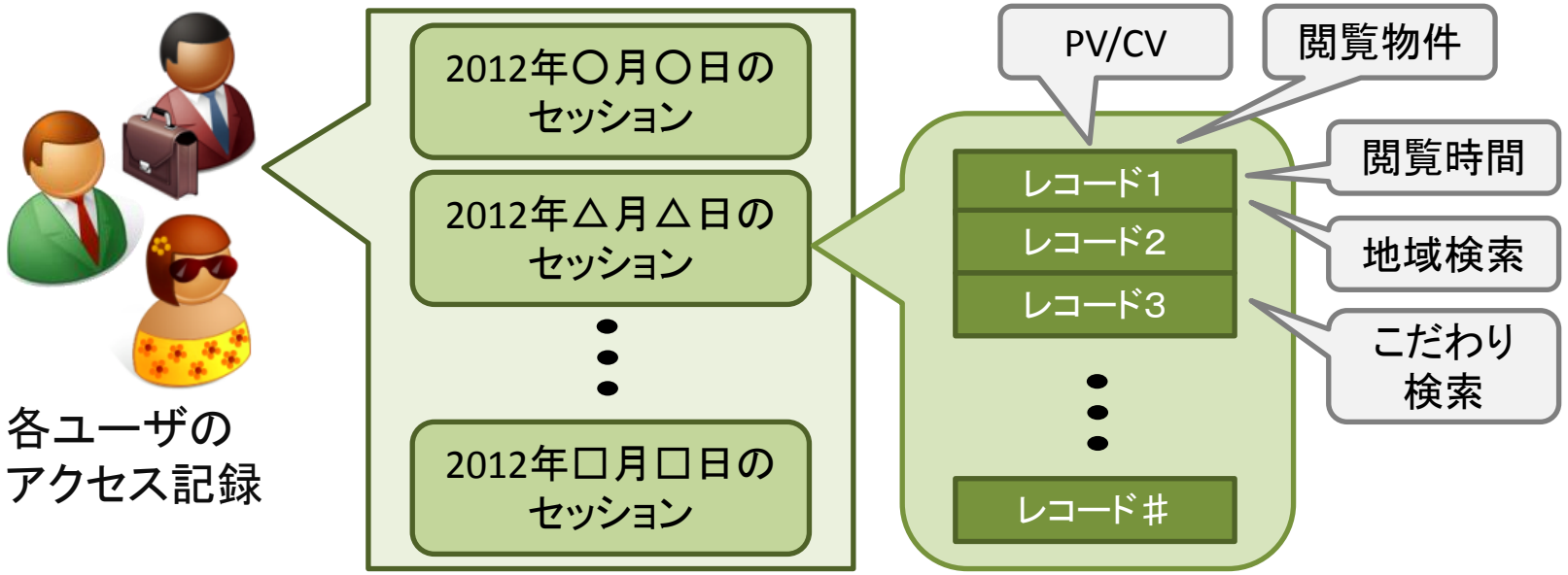
- データ
  - トランザクションデータ
    - 分析用データ
    - 本番用データ
  - マスタデータ

  全データサイズ：16GB



2.2. アクセスログのイメージ

- アクセスログの内容



2.3. 問題設定

予測課題

アクセスログ 10 週間を分析し，その後 1 週間のユーザの CV/PV を予測

課題

ユーザ 51364 人に対して，5個の物件をレコメンドする

スコアリング方法

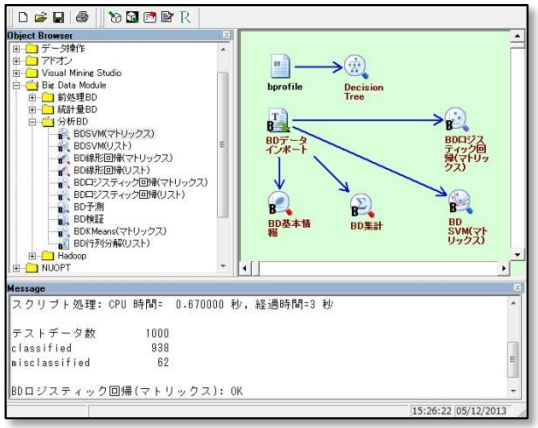
正解 CV/PV の得点は次の通り.

	正解数				
	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

3. 分析の概要

3.1. 分析のレシピ

- 分析の環境
  - ＣＰＵ：Intel Core-i7 3930K 3.20GHz（6コア）
  - メモリ：32.0 GB
- 分析の道具
  - Python（前処理・レコメンドロジック実装）
  - sqlite3（データベース）
  - R（基礎集計・グラフ描画）
  - Visual Mining Studio（決定木分析）
  - Big Data Module（ロジスティック回帰・SVM）
  - NUOPT（信頼領域内点法）



- 分析の流れ
  - ① 分析準備（クレンジング・分析用DB構築）
  - ② 分析と割当ロジックの検討
  - ③ 実験と検証

3.2. レコメンド方針

- アプローチ

  ユーザの“過去閲覧物件” から再閲覧する物件をレコメンドする

- 物件のスコアリング関数の構築

  物件プロファイル（特徴量ベクトル）に対して，閲覧確率を紐付ける

  ① ユーザが過去に閲覧した物件を列挙

  ② 各物件の特徴量を算出

  ③ 各物件の再閲覧確率を算出

閲覧物件	特徴量1	特徴量2	特徴量3	・・・	再閲覧確率
物件コード1	1	34	False	・・・ →	6%
物件コード2	5	67	True	・・・ →	19%
・・・				→	・・・

- 分析のタスク
  - 特徴量の作成と選択
  - 再閲覧確率の計算

4. 関心度と忘却度に基づくレコメンド手法

4.1. 特徴量の作成

- ユーザの閲覧物件に特徴量を与える

分析期間 / 予測期間

閲覧物件	5/28	6/12	6/23	7/02	7/25	7/28
物件コード1	PV					
物件コード2	2 PV	PV				
物件コード3	2 PV		3 PV		CV	
物件コード4			2 PV	PV	CV	PV
物件コード5				PV		PV
物件コード6				PV		

直近から3セッション

閲覧回数2

物件-セッションテーブル

閲覧物件	特徴量①	特徴量②	特徴量③	・・・	CV・PVフラグ
物件コード1	1	1	4		0
物件コード2	3	2	3		0
物件コード3	5	2	2		1
物件コード4	3	2	1		1
物件コード5	1	1	1		1
物件コード6	1	1	1		0

物件プロファイル

- 作成した特徴量グループ
  - A) ユーザに関する特徴量
  - B) 物件に関する特徴量
  - C) ユーザの物件への興味を表す特徴量

4.2. 特徴量の抽出と分類

MATHEMATICAL SYSTEMS INC.

- 特徴量の抽出処理
  - STEP１：特徴量の加工
  - STEP２：CV/PVとの相関・クロス集計
  - STEP３：決定木分析・SVM・ロジスティック回帰分析
- STEP2による絞り込み
  - C グループ(ユーザの物件への興味を表す特徴量)のCV/PVへの貢献が大きい
  - C グループを関心度と忘却度グループに分類
    - 関心度（閲覧回数・セッション登場回数・総閲覧時間）
    - 忘却度（物件の閲覧順番・セッション順番・経過日数）

Frequency & Recency

- STEP3による選択
  - gini係数・information gain ratio，回帰係数
  - およびセグメンテーションの粒度に考慮して次の指標を選択
    - 関心度：閲覧回数
    - 忘却度：セッション順番

4.3. 関心度と忘却度の分類(相関係数)

MATHEMATICAL SYSTEMS INC.

- ピアソンの相関係数

	分類	関心度A	関心度B	関心度C	忘却度A	忘却度B	忘却度C
閲覧回数	関心度A	1	0.80	0.58	-0.04	-0.01	-0.10
セッション登場回数	関心度B		1	0.47	-0.03	-0.01	-0.12
閲覧総時間	関心度C			1	-0.06	0.01	-0.06
閲覧順番	忘却度A				1	0.57	0.23
セッション順番	忘却度B					1	0.31
経過日数	忘却度C						1

＊セッション順番：最終セッションから数えて，何セッション目に物件を閲覧したか

関心度と忘却度が無相関

⇒ 関心度と忘却度から１つずつ特徴量を選抜

4.4. 関心度と忘却度の選択(決定木分析)

MATHEMATICAL SYSTEMS INC.

- 二分木における gini 係数

関心度グループ

特徴量	gini係数値
閲覧回数	**0.0034**
セッション登場回数	**0.0033**
閲覧総時間	0.0016

忘却度グループ

特徴量	gini係数値
セッション順番	**0.0024**
閲覧順番	**0.0023**
経過日数	0.0020

- 二分木における information gain ratio

関心度グループ

特徴量	info gain ratio
閲覧回数	**0.0273**
セッション登場回数	0.0245
閲覧総時間	0.0103

忘却度グループ

特徴量	info gain ratio
セッション順番	**0.0137**
閲覧順番	0.0124
経過日数	0.0120

4.5. 再閲覧確率テーブル構築

MATHEMATICAL SYSTEMS INC.

- 再閲覧確率テーブルとは

  関心度と忘却度のセグメントに再閲覧確率を対応付けたテーブル

- 再閲覧確率の計算式

n_{ij} ： 関心度 i，忘却度 j のセグメントの物件が閲覧された件数

m_{ij} ： 関心度 i，忘却度 j のセグメントの物件が再閲覧されなかった件数

$\dfrac{n_{ij}}{n_{ij} + m_{ij}}$ ： 再閲覧確率

忘却度

集計 閲覧確率 テーブル	1	2	3	4	5	6	7	8	9	10	11	12
1	6%	4%	3%	2%	2%	2%	1%	1%	1%	1%	1%	1%
2	13%	9%	6%	5%	5%	4%	3%	3%	3%	3%	2%	2%
3	19%	13%	9%	7%	7%	5%	5%	4%	4%	4%	4%	4%
4	24%	17%	12%	10%	9%	8%	6%	7%	7%	5%	2%	4%
5	28%	19%	15%	11%	9%	9%	7%	6%	4%	3%	5%	5%
6	33%	22%	18%	12%	12%	8%	5%	11%	5%	7%	2%	5%
7	36%	17%	16%	14%	11%	9%	8%	6%	7%	6%	10%	6%
8	35%	28%	15%	14%	17%	15%	9%	9%	4%	8%	4%	6%
9	38%	24%	18%	14%	15%	10%	11%	7%	13%	6%	6%	0%
10	45%	27%	19%	15%	18%	16%	13%	7%	5%	5%	3%	10%
11	41%	23%	19%	20%	14%	19%	14%	4%	6%	0%	10%	16%
12	36%	37%	26%	14%	13%	14%	12%	20%	20%	8%	5%	10%
13	52%	27%	27%	16%	6%	16%	9%	18%	6%	3%	0%	7%
14	49%	35%	22%	29%	24%	22%	0%	19%	7%	0%	17%	0%
15	69%	35%	24%	24%	27%	13%	7%	13%	9%	0%	20%	11%
16	47%	42%	40%	12%	25%	17%	8%	8%	0%	33%	0%	14%
17	36%	33%	24%	23%	13%	22%	0%	10%	10%	100%	0%	17%
18	67%	35%	24%	13%	10%	10%	10%	11%	9%	0%	50%	0%
19	68%	39%	57%	31%	25%	33%	40%	17%	100%	0%	0%	0%
20	54%	25%	27%	8%	29%	15%	40%	20%	50%	0%	0%	0%

関心度

データの規模が大きいほど確率の信頼性が上がる

4.6. レコメンドロジック

MATHEMATICAL SYSTEMS INC.

物件プロファイル × 再閲覧確率テーブル

再閲覧確率の高い順に物件をレコメンド

物件プロファイル

閲覧物件	忘却度	関心度	閲覧確率
物件コード1	1	1	6%
物件コード2	1	3	19%
物件コード3	1	2	12%
物件コード4	2	2	9%
物件コード5	2	2	9%
物件コード6	3	1	3%
物件コード7	4	2	5%
物件コード8	4	4	10%

参照

再閲覧確率テーブル（実績値）

閲覧確率テーブル	1	2	3	4
1	6%	4%	3%	2%
2	12%	9%	6%	5%
3	19%	13%	9%	7%
4	24%	17%	11%	10%

忘却度

関心度

関心度と忘却度のトレードオフを考慮したレコメンドを実現

5. 過学習の回避

5.1. レコメンド手法の改善

- 関心度と忘却度に成り立つ“単調性制約”
  - 関心度が大きい物件ほど再閲覧する
  - 忘却度が小さい物件ほど再閲覧する

再閲覧確率テーブルで単調性制約が満たされないセグメントが存在

閲覧確率テーブル	1	2	3
1	6%	4%	3%
2	13%	9%	6%
3	19%	13%	9%
4	24%	17%	12%
5	28%	19%	15%
6	33%	22%	18%
7	36%	17%	16%
8	35%	28%	15%
9	38%	24%	18%

忘却度

関心度

閲覧確率テーブル	5	6	7	8
12	13%	14%	12%	20%
13	6%	16%	9%	18%
14	24%	22%	0%	19%
15	27%	13%	7%	13%

忘却度

関心度

- 原因
  - 学習データとして十分な量を確保できていない
  - 業務上の施策の影響が反映されてしまっている

過学習を回避した再閲覧確率テーブルの推定をしたい

5.2. 数理モデルの構築

- 推定する再閲覧確率テーブルの要件
  - 単調性制約を満たす
  - データ件数が多いセグメントの再閲覧確率ほど信頼する
- 凸二次計画問題に定式化して最適化パッケージ **NUOPT** で求解

◆ 集合　I ：関心度のセグメント　J ：忘却度のセグメント

◆ パラメータ　p_{ij} $(i \in I, j \in J)$ ：各セグメントの閲覧確率（実績値）

w_{ij} $(i \in I, j \in J)$ ：各セグメントのデータ数

◆ 変数　$x_{ij} \in [0,1]$ $(i \in I, j \in J)$ ：各セグメントの推定する閲覧確率

◆ 制約　$x_{ij} + \varepsilon \le x_{i'j}$ $(i < i' (\in I))$ ：関心度について狭義単調増加

$x_{ij} \ge x_{ij'} + \varepsilon$ $(j < j' (\in J))$ ：忘却度について狭義単調減少

（ε ：適当な微小な値）

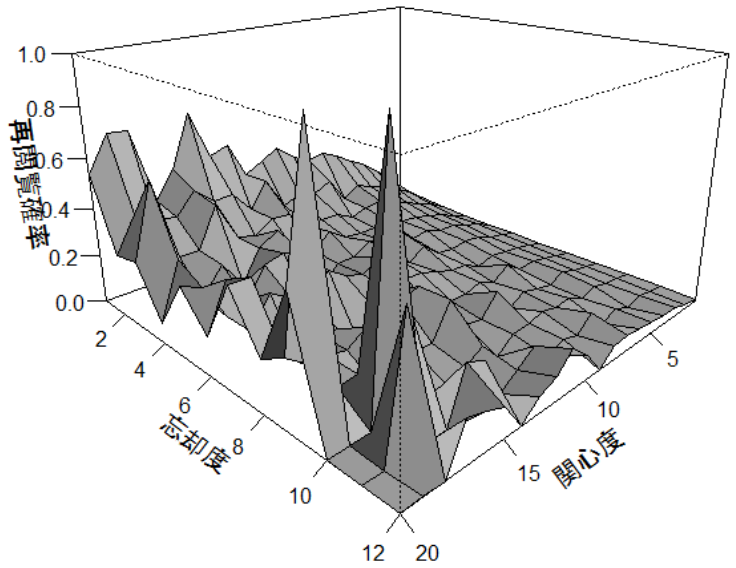
◆ 目的関数　$\mathrm{minimize} \sum_{i \in I, j \in J} w_{ij}^2 \cdot \left(x_{ij} - p_{ij} \right)^2$

：閲覧確率（実績値）との重み付き自乗誤差最小化

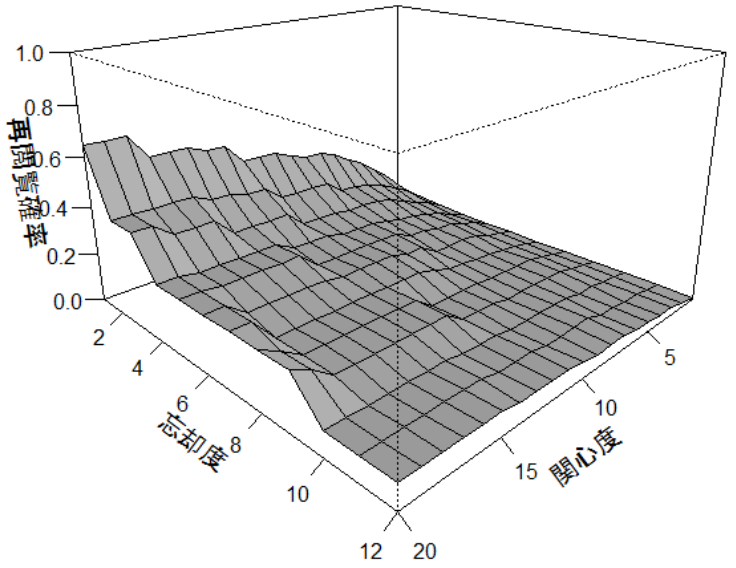
5.3. 推定した再閲覧確率テーブル

- 再閲覧確率テーブルの比較

再閲覧確率テーブル（実績値）



再閲覧確率テーブル（推定値）



スムージングによって過学習を回避

5.4. 実験と評価

評価用ツールの作成（分析用データ）

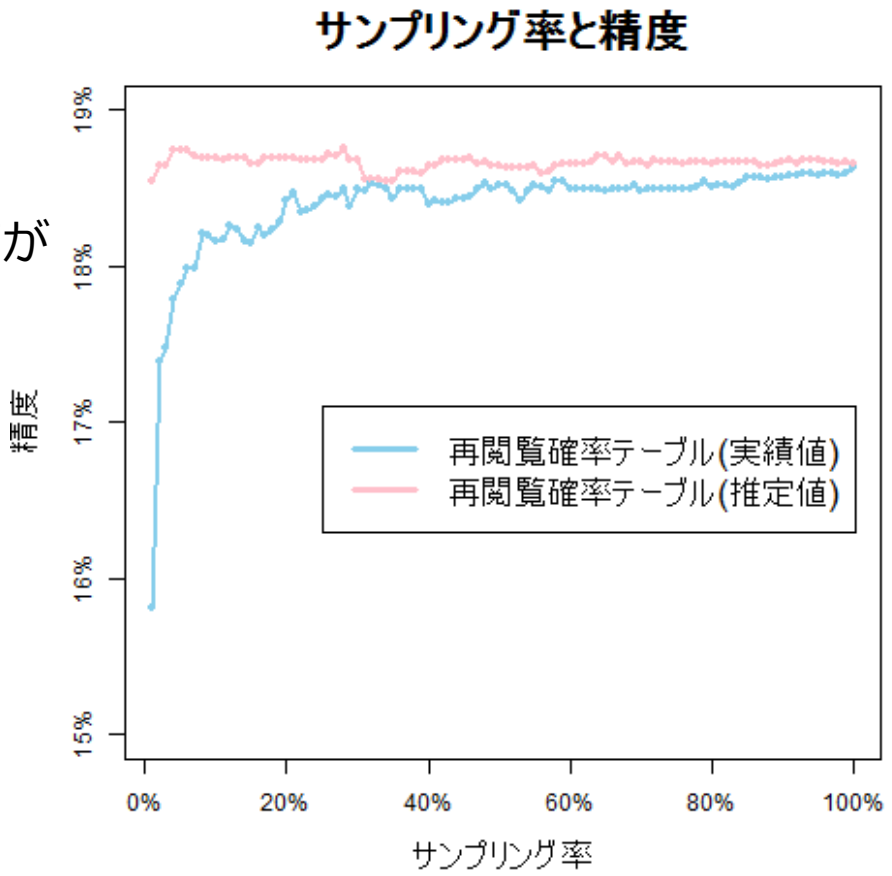
アクセスログの最終週を予測期間として，17803 ユーザを抽出

総スコア 76,017 点に対する得点率を予測精度としてレコメンド手法を評価

レコメンド手法	スコア	精度
比較手法①：閲覧が最新の物件から順にレコメンド	11,937	15.70 %
比較手法②：閲覧回数が多い物件から順にレコメンド	13,146	17.29 %
提案手法①：関心度と忘却度に基づくレコメンド（**実績値**）	14,181	18.66 %
提案手法②：関心度と忘却度に基づくレコメンド（**推定値**）	14,232	18.72 %

5.5. サンプリング実験

- 17,803 ユーザからサンプリング（1%～100%）
  実績値と推定値の２つの
  再閲覧確率テーブルを比較
  - 実績値より推定値の方が
    データ量に限らずレコメンド精度が
    良いことを確認
  - データ不足も解消可能



より詳細なセグメンテーションが可能

6. まとめ

6.1. まとめ

- マーケティングについて
  - 頻度（Frequency）と直近さ（Recency）を具体的に定量化して
    レコメンドロジックを構築
  - 予測精度は特徴量の作成と選択に尽きる
- ビッグデータについて
  - 大規模データの特性
    規模に比例して確率の信頼性が向上・詳細なセグメンテーションが可能
  - 過学習の回避＆データ不足の解消
    凸二次計画問題に定式化して再閲覧確率テーブルを推定
- ビジネスにおける実現性
  - スケーラビリティ
    - 再閲覧確率テーブルの作成（Hadoop 等の分散処理技術）
    - 再閲覧確率テーブルの推定（凸二次計画法：変数数 |I|×|J| ）
  - レコメンド時のリアルタイム性
    - 再閲覧確率テーブルの参照と確率のソート処理でレコメンド可能