商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案

原田玲央　　伊藤聡志　　菊池浩明

明治大学総合数理学部 先端メディアサイエンス学科

背景: PWSCUP2016

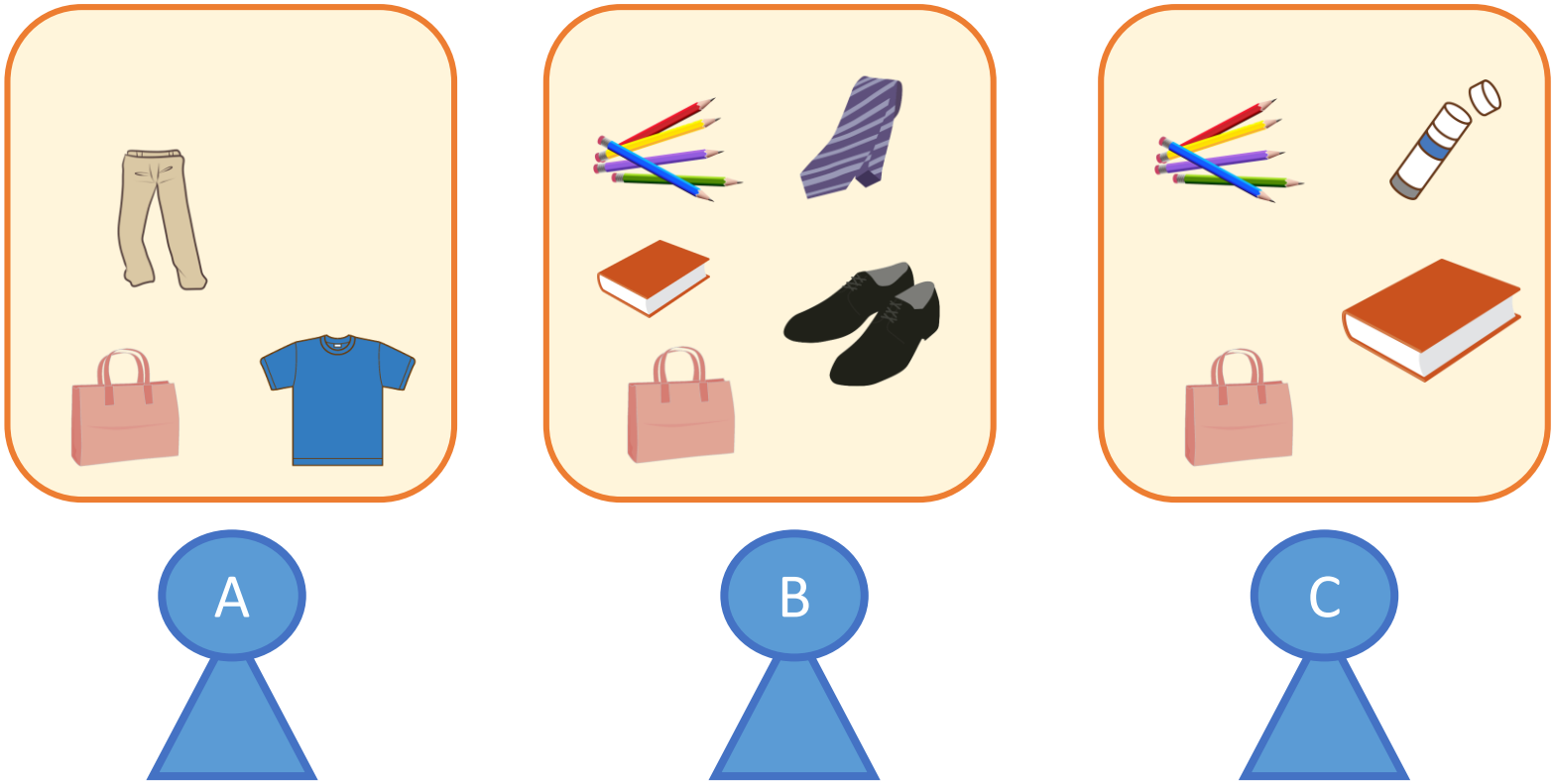
- 2016年10月, 第二回「匿名加工・再識別コンテスト」が開催
  - 英国での1年間に実際に取引された購買履歴データ(動的データ)

1. コンテストで用いた**再識別手法の解説**
2. コンテストでの手法を拡張した**匿名加工手法の提案**

チームステテコ伊藤2
再識別賞

1. 再識別手法

着目点: 購入商品の特徴量



購買履歴データ= 顧客ごとに購入商品の特徴を有する

個人の特定リスクにつながる情報

Jaccard再識別アルゴリズム

Jaccard類似度

2人が購入している商品のうち共通している商品の割合

$$J(u'_1, u_1) = \frac{|\{g_1, g_2\}|}{|\{g_1, g_2, g_3, g_4, g_5\}|} = \frac{2}{5}$$

Jaccard再識別

顧客が購入した商品リストについて,

Jaccard類似度が近いレコードについて再識別を行う.

元データ

行	顧客ID	商品の集合
1	u_1	$\{g_1, g_2, g_5\}$
2	u_2	$\{g_1, g_5, g_6\}$

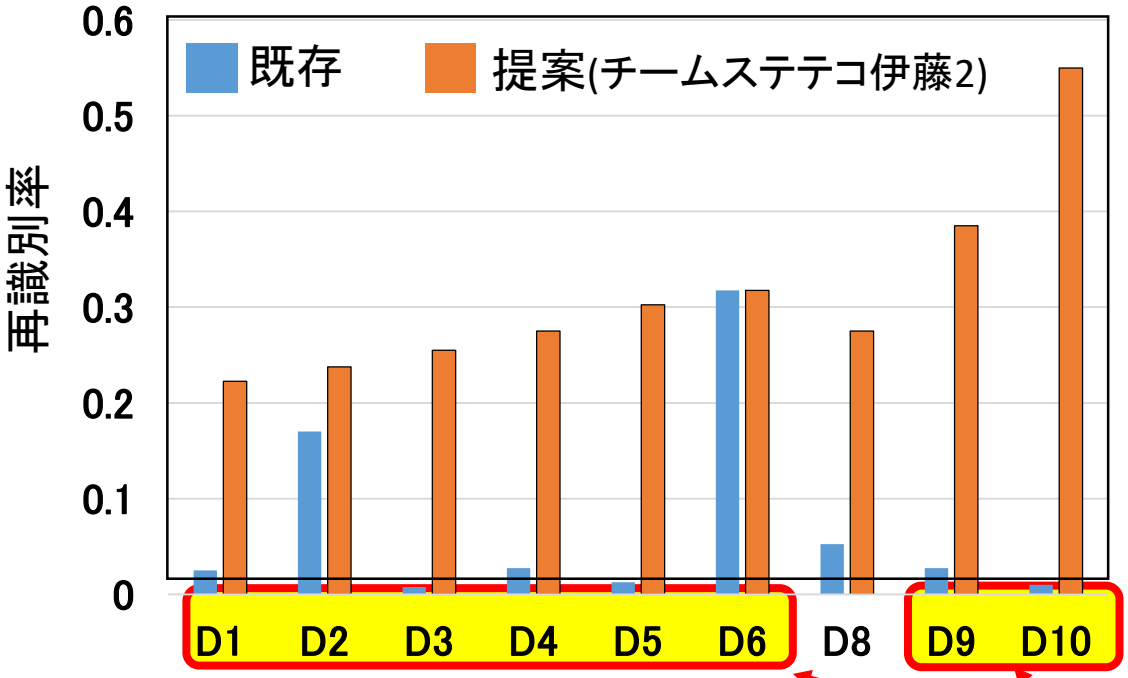
$2/5$

$1/6$

加工データ

P	仮ID	商品の集合
1	u'_1	$\{g_1, g_2, g_3, g_4\}$

提案アルゴリズムの再識別率



匿名加工データ
(コンテスト上位チーム)

最大識別率を記録したデータ

提案アルゴリズムがほとんどのデータで最大識別率を記録

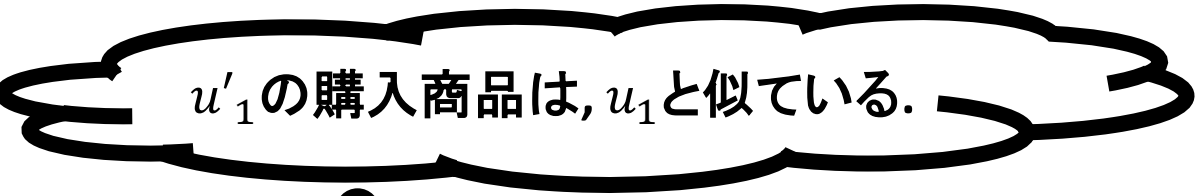
購買履歴データは，商品の特徴による特定リスクが存在

2. 匿名加工手法の提案

- コンテストで用いた手法の拡張

目的

Jaccard再識別に対して有効な匿名加工手法の提案



加工データ

仮ID	購入商品
u'_1	$\{g_1, g_2, g_3\}$
u'_2	$\{g_1, g_3\}$
u'_3	$\{g_5\}$
u'_4	$\{g_4, g_6\}$

類似

元データ

顧客ID	購入商品
u_1	$\{g_1, g_2\}$
u_2	$\{g_1, g_3, g_4\}$
u_3	$\{g_5, g_6\}$
u_4	$\{g_4, g_5, g_6\}$

目的

Jaccard再識別に対して有効な匿名加工手法の提案

加工データ

仮ID	購入商品
u'_1	$\{g_1, g_2, g_3, g_4\}$
u'_2	$\{g_1, g_2, g_3, g_4\}$
u'_3	$\{g_4, g_5, g_6\}$
u'_4	$\{g_4, g_5, g_6\}$

元データ

顧客ID	購入商品
u_1	$\{g_1, g_2\}$
u_2	$\{g_1, g_3, g_4\}$
u_3	$\{g_5, g_6\}$
u_4	$\{g_4, g_5, g_6\}$

耐Jaccard匿名加工手法

元データ

顧客ID	購入商品
u_1	g_1, g_2
u_2	g_3

加工データ

仮ID	購入商品
u'_1	g_1, g_2, g_3
u'_2	g_1, g_2, g_3

統一する顧客のグループをどうやって作るか?

仮ID	購入商品
u_1	g_1
u_1	g_2
u_2	g_3

レコード追加

仮ID	購入商品	
u'_1	g_1	m
u'_1	g_2	
u'_2	g_3	
u'_1	g_3	擬似レコード数 Δm
u'_2	g_1	
u'_2	g_2	

問題点

コンテストのデータセットは,

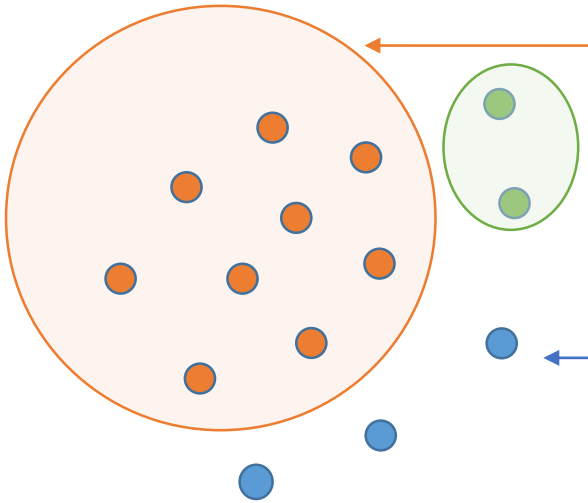
- 高次元データ
- 類似する顧客が少ない

高次元

	g_1	g_2	g_3	…	g_{2781}
u_1	1	1	0	…	0
u_2	1	0	1	…	1
u_3	0	0	0	…	1

似ていない

ナイーブなクラスタリングの例



問題点1:「巨大なクラスタ」

⇒ Δmが膨大になる

問題点2:「独立したクラスタ」

⇒ 一意に再識別される

先行研究

- 行列分解 [長谷川,菊池, 正木, 浜田, 2016]
  - 高次元データを行列分解により低次元化し, k-匿名化

- クラスタリング E2DCKM [緒方, 遠藤, 2013]
  - 静的な人工データをクラスタリング手法
  - クラスタサイズを均等にする

解決方法

方式1: TF-IDF

	g_1	g_2	g_3
u_1	1	1	0
u_2	1	0	1
u_3	0	1	1
u_4	0	1	0

TF
顧客内における商品の出現頻度

IDF
商品の希少性

TF-IDF

	g_1	g_2	g_3
u_1	0.7	0.6	0
u_2	0.7	0	0.7
u_3	0	0.6	0.7
u_4	0	1.1	0

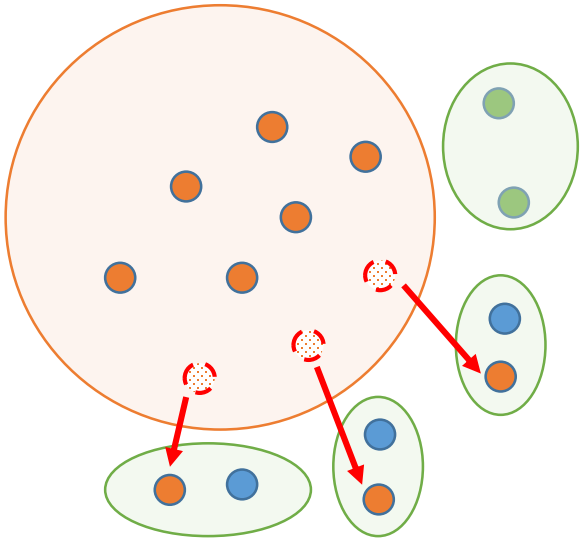
TF　IDF

$$\frac{1}{2}\left(\log\frac{4}{2}+1\right)=0.7$$

k-means

u_1 u_2 | u_3 u_4

方式2: 最大クラスタの分配



最大クラスタサイズ $s_{max}=9\rightarrow 6$

最小クラスタサイズ $s_{min}=1\rightarrow 2$

クラスタ数は変化しない

評価

- TF-IDFの効果(方式1)
- 最大クラスタ分配の効果(方式2)
- 有用性
- 安全性
- 最適クラスタ数

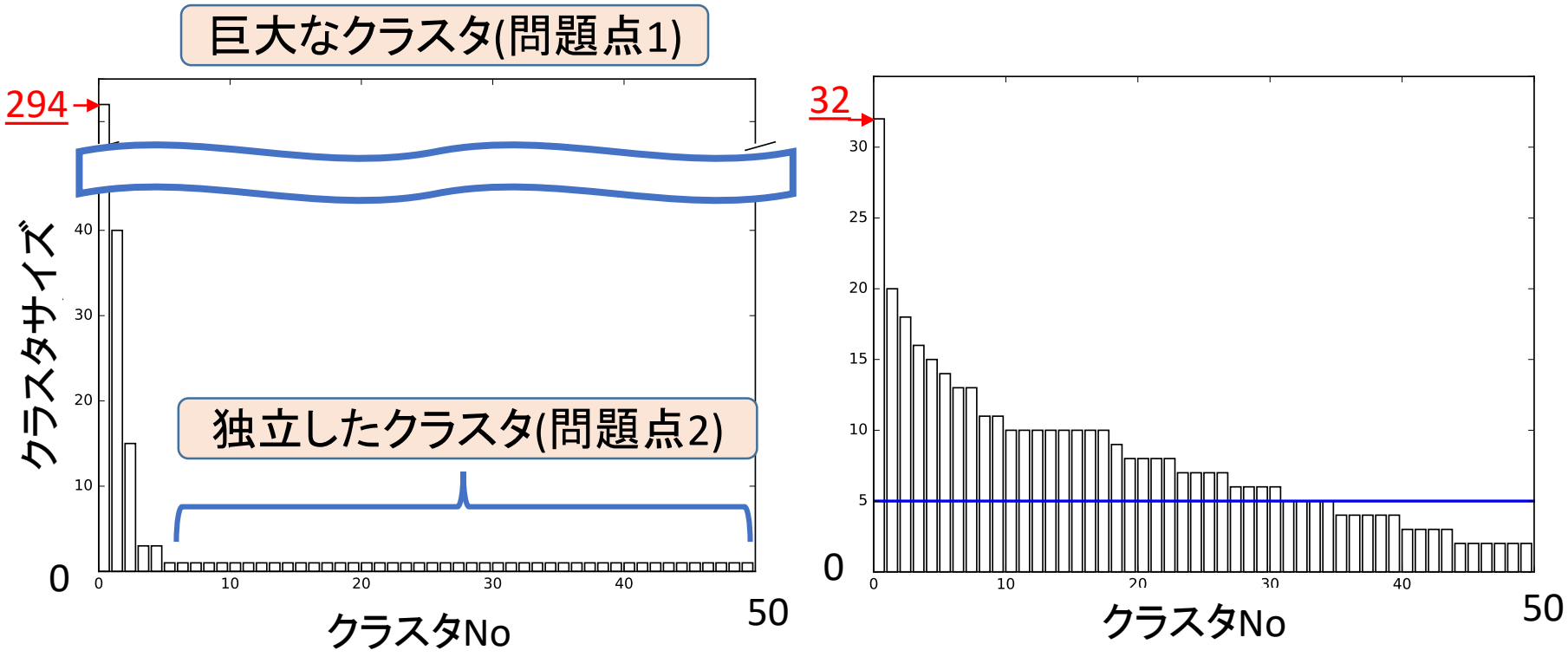
  - 実験データ
    - 顧客数$n = 400$, レコード数$\mathrm{m} = 38067$のコンテストで使われたデータセットを使用

TF-IDFによる効果 (方式1)

顧客n=400人を50のクラスタに分類

(問題点)
高次元データのクラスタリング

(方式1)
TF-IDFを使ったクラスタリング

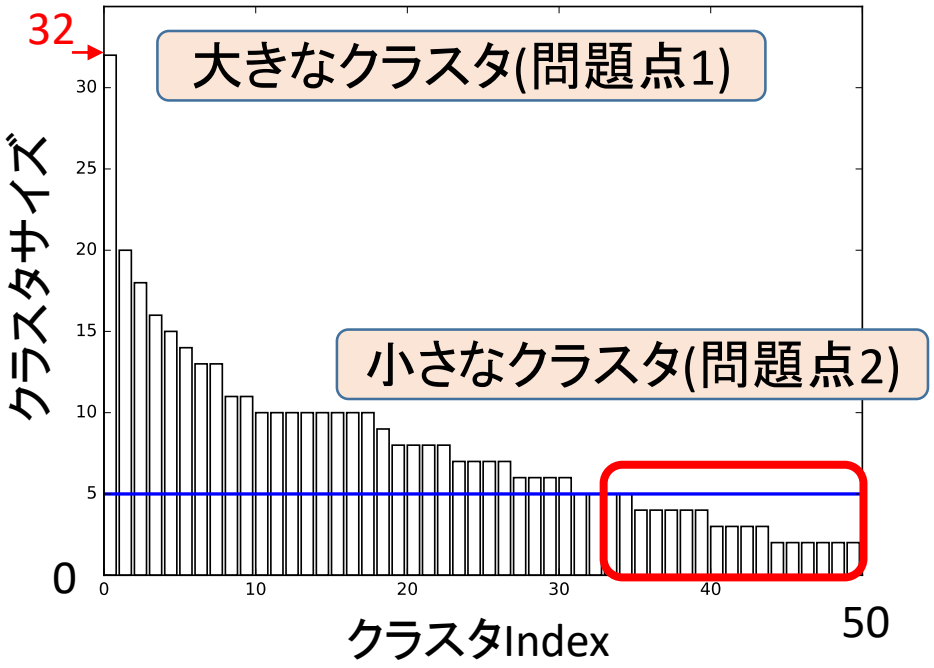


クラスタサイズの偏りを改善

クラスタの分配による効果(方式2)

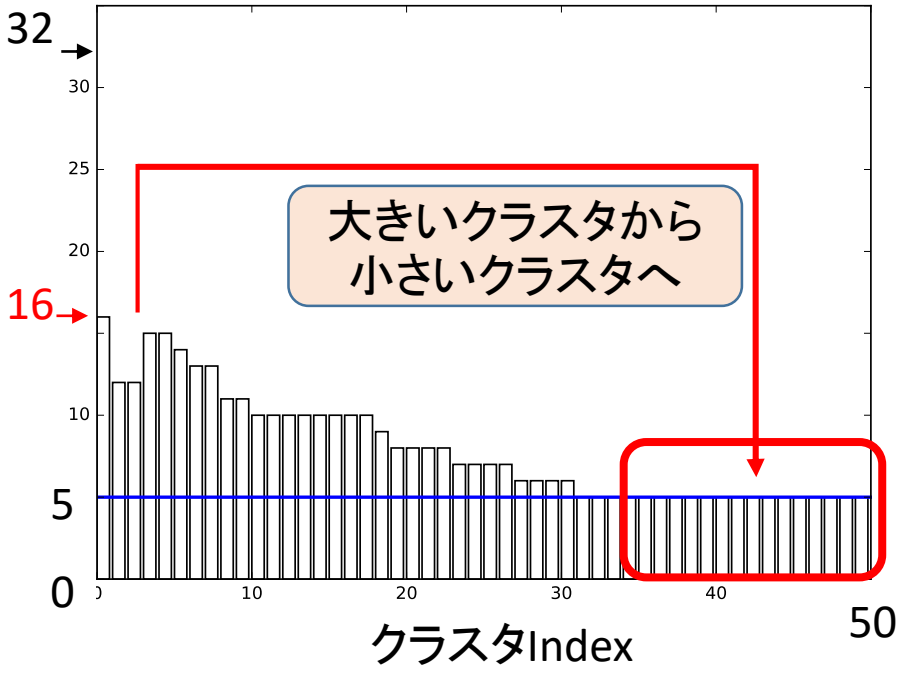
顧客n=400人を50のクラスタに分類, 下限値5にした例

- 方式1



$\Delta m = 182{,}897$

- 方式2

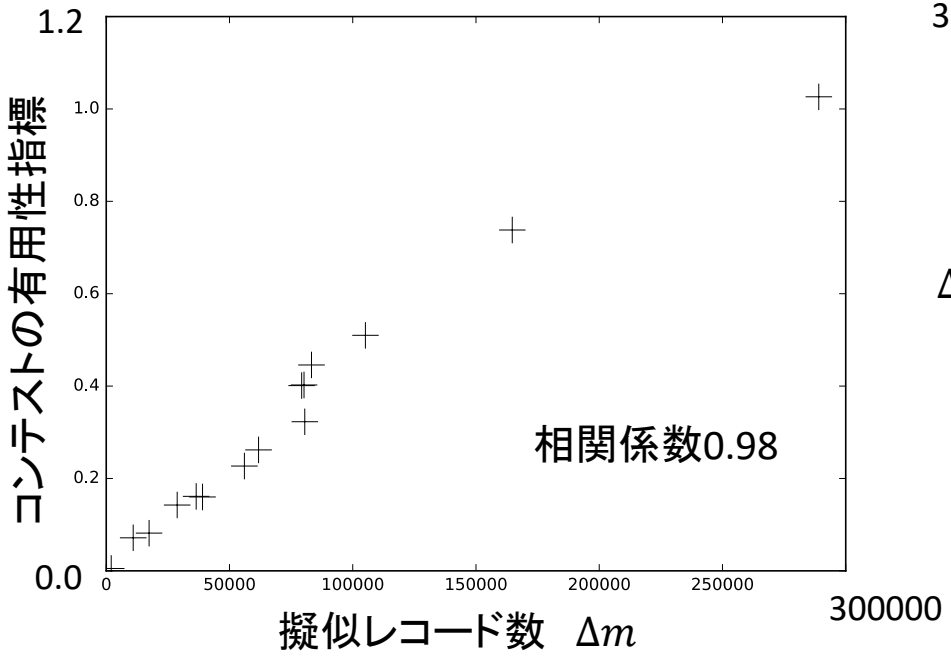


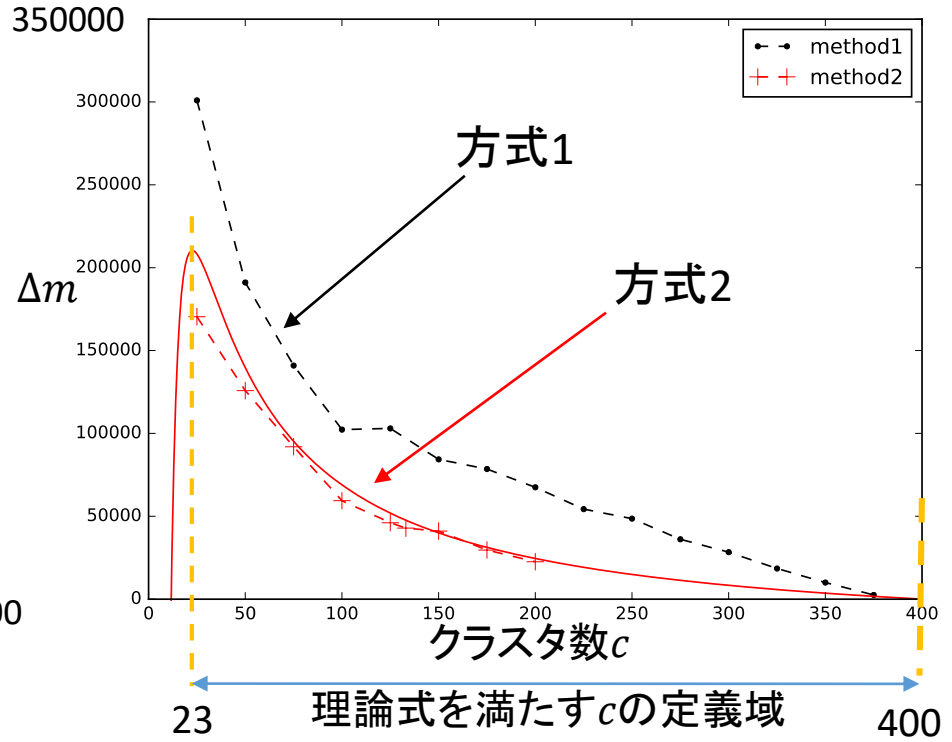
$\Delta m = 145{,}747$

独立したクラスタの解消

擬似レコード数の削減

有用性





有用性はクラスタ数に依存

データセットの統計量
- b:平均購買商品の種類数
- h:ある2顧客間が共通して購入した商品数
- n:顧客数

$$E(\Delta m) = -\frac{hn^3}{2c^2} + \underbrace{(b + \frac{h}{2})\frac{n^2}{c}}_{\text{支配項}} - bn$$

安全性

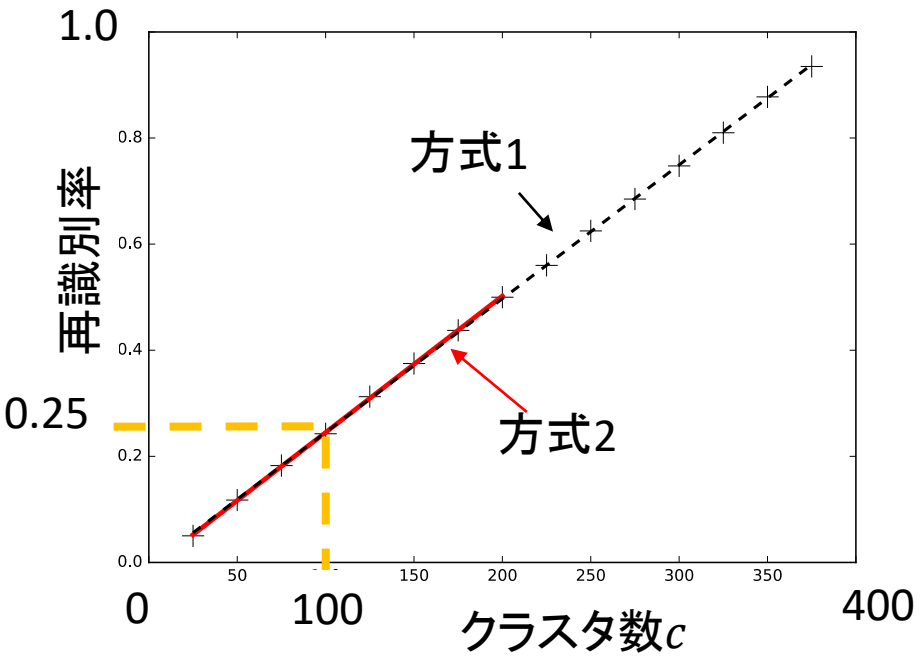
元データ

ID	商品の集合
u_1	$\{g_1, g_2\}$
u_2	$\{g_1, g_3, g_4\}$
u_3	$\{g_2, g_3\}$

Jaccard再識別

加工データ

仮ID	商品の集合
u'_1	$\{g_1, g_2, g_3, g_4\}$
u'_2	$\{g_1, g_2, g_3, g_4\}$
u'_3	$\{g_1, g_2, g_3, g_4\}$



期待値 $E(Reid) = {}^{c}/_{n}$

n:顧客数, c:クラスタ数

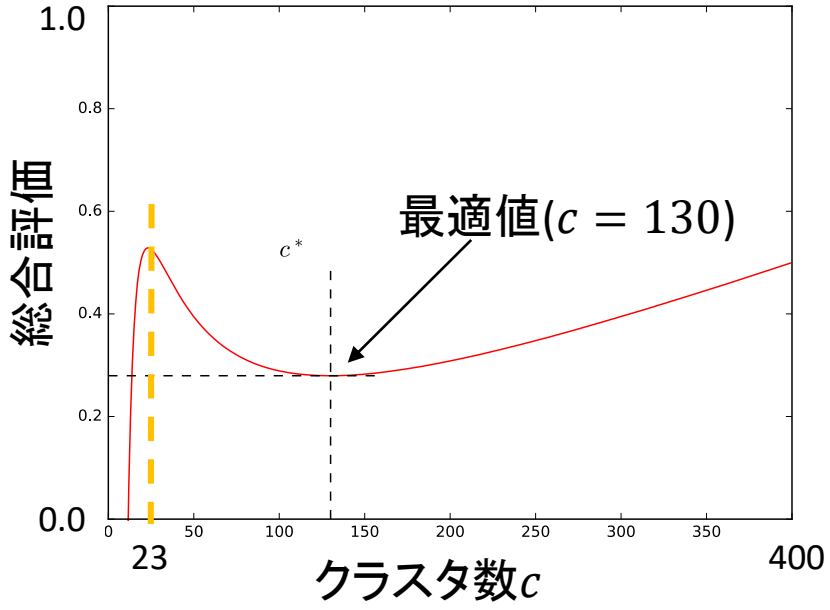
各クラスタで一人が識別

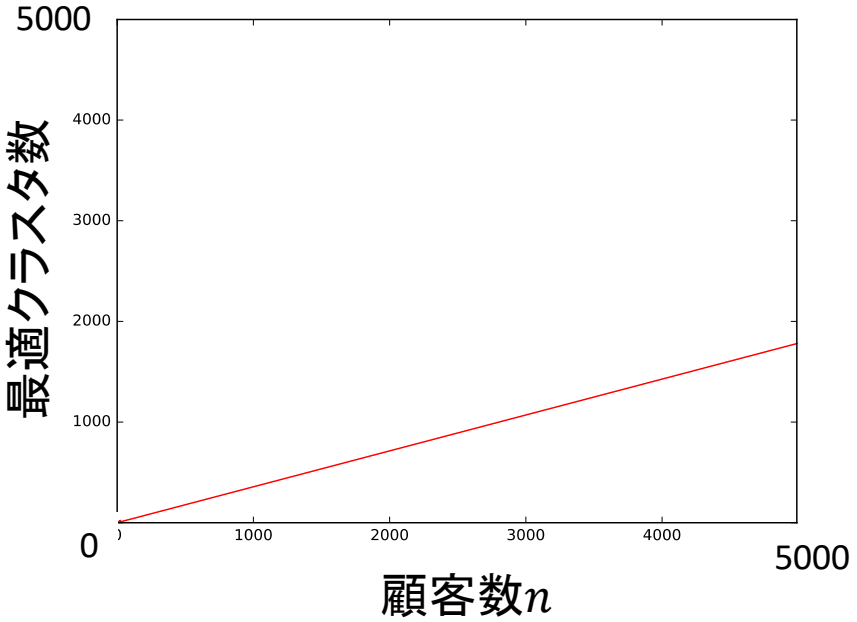
最適クラスタ数

有用性(正規化)　　安全性

総合評価指標: $\dfrac{\alpha E(\Delta m)+E(Reid)}{2}$

$n = 400$





データセットの統計量から本手法適用による最適値を定義

まとめと今後の課題

- まとめ
  - 購買履歴データに対する有効な再識別手法
  - TF-IDFを利用した匿名加工手法の提案
  - TF-IDFによる分類などの評価

- 今後の課題
  - クラスタリングの精度評価
  - レコード削除や書き換えによる手法

(参考)

ID	商品の集合
u_1	$\{g_1, g_2, g_5\}$
u_2	$\{g_1, g_3, g_4\}$
u_3	$\{g_2, g_3, g_6\}$

→

仮ID	商品の集合
u'_1	$\{g_1, g_2, g_3, g_4, g_5, g_6\}$
u'_2	$\{g_1, g_2, g_3, g_4, g_5, g_6\}$
u'_3	$\{g_1, g_2, g_3, g_4, g_5, g_6\}$

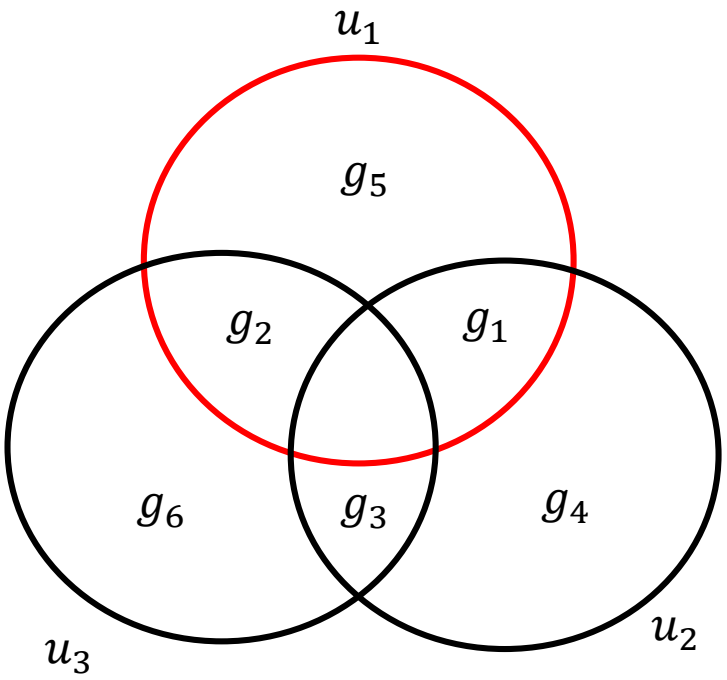
1 1 1 2 2 2 = 9 (Δm)

$1 \cdot 3 \cdot 1 + 2 \cdot 3 \cdot 1 = 9$

エリア内の商品数

エリアの数

他の顧客人数



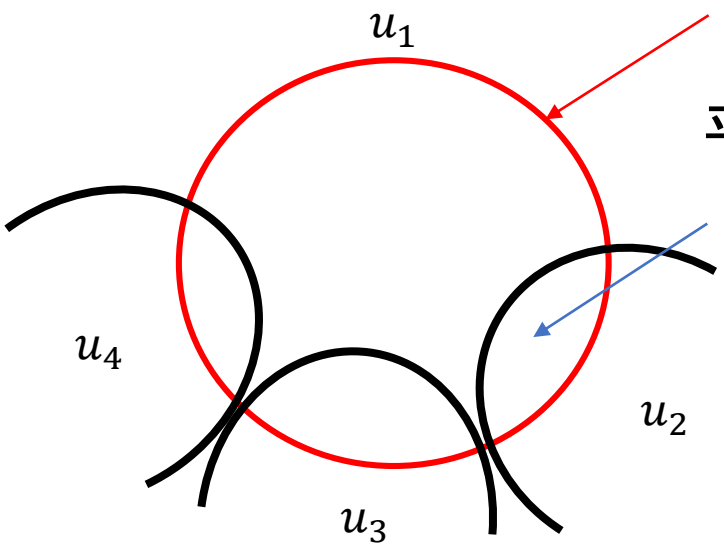
(参考)

3人以上共通している商品は存在しないと仮定
→ b,h,nを使って表現するため

クラスタサイズ
$s = 4$

平均購買数
$b = 65$

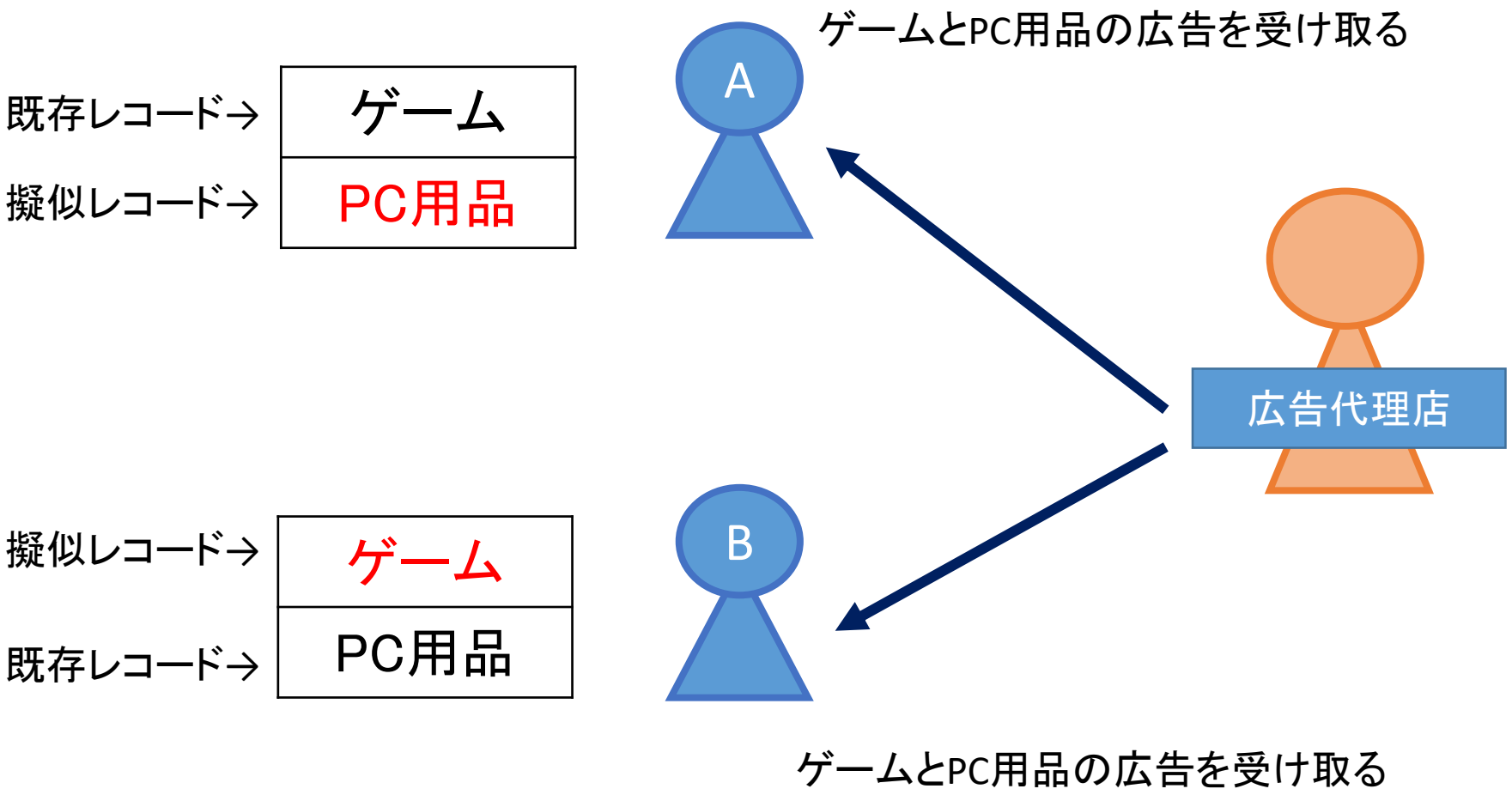
平均共通購買数
$h = 4$



条件式: $b > h \times s$

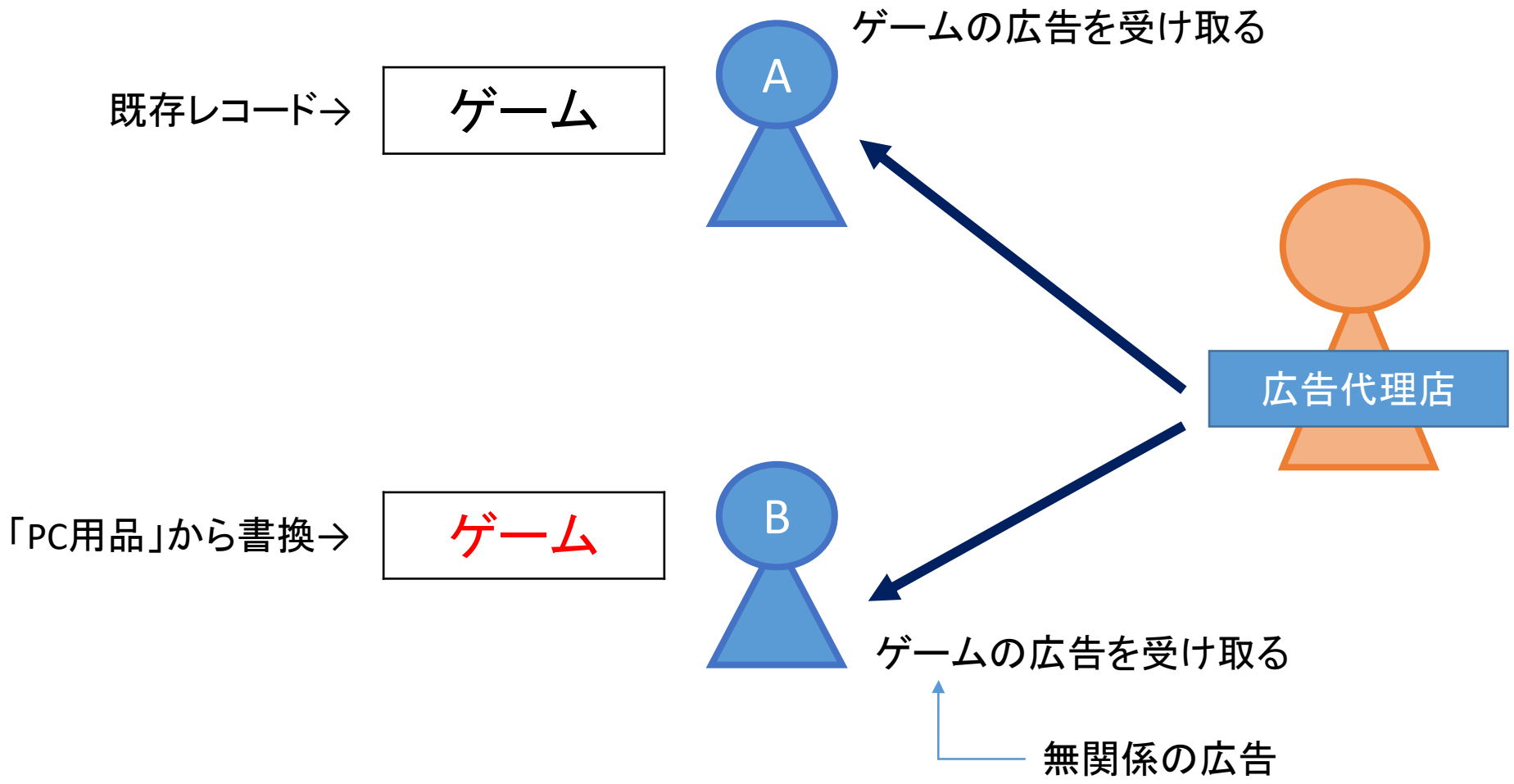
(参考)

➢擬似レコード追加


既存レコード→
ゲーム
擬似レコード→
PC用品
A
ゲームとPC用品の広告を受け取る
広告代理店
擬似レコード→
ゲーム
既存レコード→
PC用品
B
ゲームとPC用品の広告を受け取る

(参考)

➢書き換えや削除


既存レコード→
ゲーム
A
ゲームの広告を受け取る
広告代理店
「PC用品」から書換→
ゲーム
B
ゲームの広告を受け取る
無関係の広告

(参考)TF-IDF

	g_1	g_2	g_3
u_1	1	1	0
u_2	1	0	1
u_3	0	1	1
u_4	0	1	0

TF=$\frac{1}{2}$

顧客u_1がg_1を購入している確率

IDF= $\log \frac{4}{2} + 1$

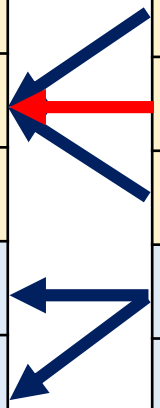
$$log \frac{\text{顧客数}}{g_1\text{を購入している顧客の数}} + 1$$

(参考)安全性

- クラスタ内の最大要素数の顧客=識別される
- 各クラスタに1人は識別 →: 再識別率の期待値$= \frac{c}{n}$
- クラスタ内の最大要素数が複数ある場合を除く

元データ

行	顧客ID	商品の集合
1	u_1	$\{g_1, g_2\}$
2	u_2	$\{g_1, g_3, g_4\}$
3	u_3	$\{g_2, g_3\}$
4	u_4	$\{g_2, g_5\}$
5	u_5	$\{g_4, g_5\}$



加工データ

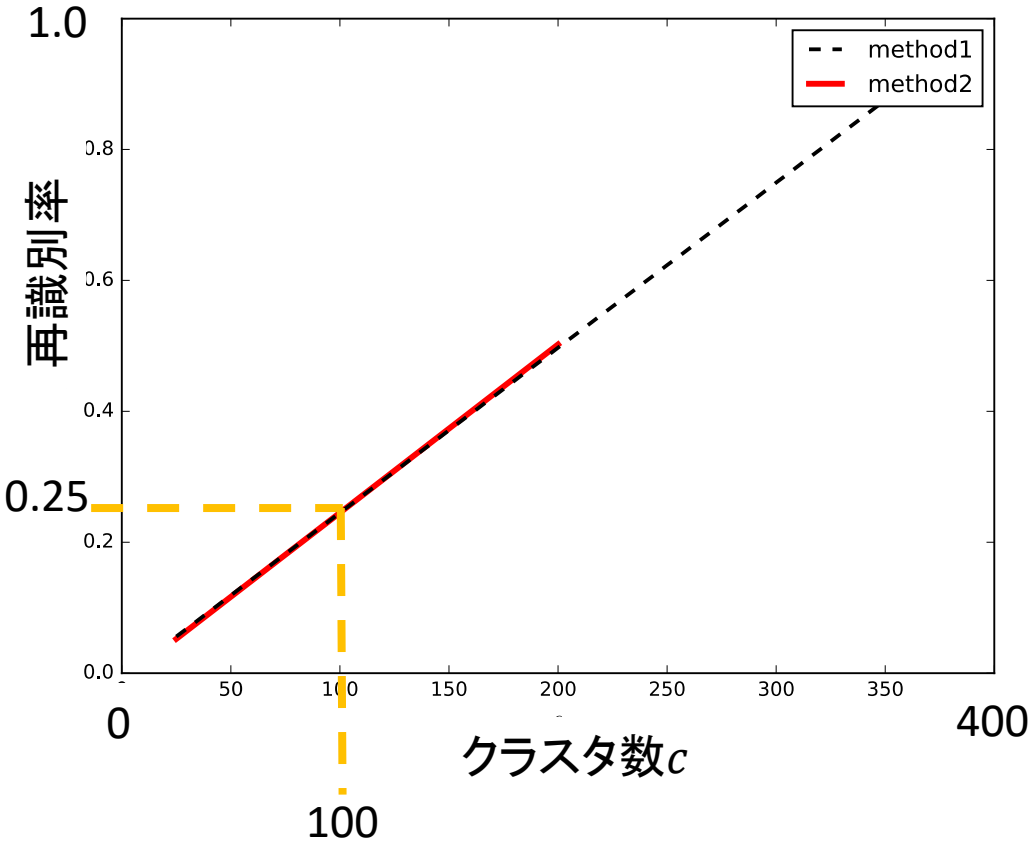
P	仮ID	商品の集合
1	u'_1	$\{g_1, g_2, g_3, g_4\}$
2	u'_2	$\{g_1, g_2, g_3, g_4\}$
3	u'_3	$\{g_1, g_2, g_3, g_4\}$
5	u'_5	$\{g_2, g_4, g_5\}$
4	u'_4	$\{g_2, g_4, g_5\}$

推定行番号

Q
2
2
2
4 or 5
4 or 5

(参考)安全性

期待値



実測値

	$c = 100$	
	Jaccard	再識別率
方式1	0.3060	0.2488
$s_{min} = 2$	0.3061	0.2475
$s_{min} = 3$	0.3041	0.2480
$s_{min} = 4$	0.3044	0.2465
期待値	-	0.2500

期待値 $E(Reid) = {}^{c}/_{n}$

→各クラスタで一人が識別される