K411　菊池研・斉藤研合同発表会　2017年2月4日

購入履歴データを用いた匿名加工データの最頻アイテムに注目した再識別手法freqItemの提案とその評価

菊池研B4　岡本健太郎

背景・目的

背景
ビッグデータの流行による匿名加工の需要増加

目的
匿名加工データのリスクの正しい評価をすること

用いたデータ“Online Retail”について

- オンラインショッピングの購買履歴データ
- 約4000人分の36万履歴の動的データ
- 顧客マスタと購買トランザクションの二つのデータがある

顧客マスタ

- 顧客のデータ
- ID、性別、生年月日、国籍
- ４００行

購買トランザクション

- 購入履歴データ
- 顧客ID、伝票ID、購買日時、購買時間、製品ID、単価、購買数
- ３８０６７行

用いたデータ“Online Retail”について

- 木口君と佐伊藤君が買い物をした
- 木口君はヒマワリの種を5袋とアヒルの卵を2パック、佐伊藤君はAndroid端末を10台とバーベキューの肉を購入

顧客マスタ

Name	Sex	Birthday	Nation
Kikuchi	男	1965/4/25	Japan
Saito	男	1988/6/6	Japan

購買トランザクション

Name	Bag	Item	Fee	Num
Kikuchi	1	Seed	100	5
Kikuchi	1	Duck’s egg	500	2
Saito	2	Android	20000	10
Saito	2	meat	1000	1

提案手法

- 最頻アイテムに注目した再識別手法freqItemの提案
- 最頻アイテム・・・ユーザーごとに最も購入したアイテムを単一に定めた値、複数ある場合はランダムで決定

	Name	Item
1	Abe	Doughnut
2	Abe	Eraser
3	Baba	Doughnut
4	Abe	Doughnut
5	Chiaki	Fork
6	Baba	Eraser
7	Chiaki	Eraser
8	Baba	Eraser

Name	freqItem
Abe	Doughnuts
Baba	Eraser
Chiaki	EraserまたはFolk

提案手法

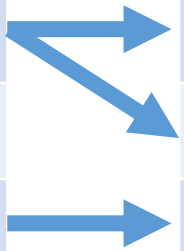
① 国籍と性別が同じユーザーでグループを作成

② トランザクションデータをもとにfreqItemを求める

③ 元データのユーザーのfreqItemを参照し、国籍性別freqItemの同一のユーザーを同一ユーザーとして再識別

※同一のユーザーが複数いた場合は全員をそのユーザーとする

Group	Name	freqItem
α	A	Doughnuts
α	B	Eraser
β	C	Folk



Group	freqItem	Name
α	Doughnuts	A
α	Doughnuts	A
β	Folk	C

実験方法

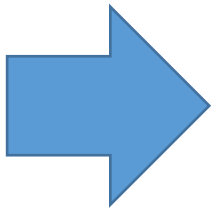
- PWSCUP2016に提出された加工データを提案手法とjaccard再識別で再識別しそれぞれに対する識別率を計算
- 識別率や加工手法を見ながら再識別が有効な加工データを判断

- Jaccard再識別とは…ユーザーごとに購入したアイテムで多重集合の類似度を計算し、加工の前後で最も類似度の高いユーザー同士を同一ユーザーと再識別する手法

結果

チーム	freqItem 再識別	jaccard 再識別	YA 匿名化	Ja 最適化	Ja ランダム化	CMAE1 最適化	CMAE2 最適化	RFM 最適化	仮名化	列統一	属性値追加	国スワップ	国&性別スワップ	購入スワップ
T	1.25	22.25			○	○	○		○	○				
K	0.75	25.50	○	○		○	○		○	○				○
J	9.50	27.50	○		○	○	○	○	○	○				○
B	14.75	30.25	○			○	○		○	○			○	○
N	15.25	27.50	○	○			○		○	○	○	○	○	○
M	13.00	38.50	○	○					○	○		○	○	
I	44.75	**88.50**	○						○	○				
平均	14.18	37.14	6	3	2	4	5	1	7	7	1	2	3	4

結果



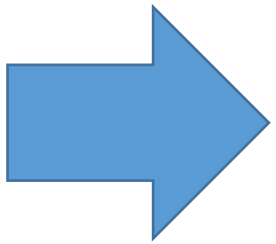
チーム	freqItem 再識別	Jaccard 再識別	YA 匿名化	Ja 最適化	Ja ランダム化	CMAE1 最適化	CMAE2 最適化	RFM 最適化	仮名化	列統一	属性値追加	国スワップ	国&性別スワップ	購入スワップ
T	1.25	22.25			○	○	○		○	○				
K	0.75	25.50	○	○		○	○		○	○				○
J	9.50	27.50	○		○	○	○	○	○	○				○
B	14.75	30.25	○			○	○		○	○			○	○
N	15.25	27.50	○	○			○		○	○	○	○	○	○
M	13.00	38.50	○	○					○	○		○	○	
I	44.75	88.50	○						○	○				
平均	14.18	27.50	6	3	2	4	5	1	7	7	1	2	3	4

考察

- YA-匿名化、列統一、仮名化などのメジャーな加工には強い
- jaccard最適化、jaccardランダム化などに極端に弱い
- 属性値追加されたデータに対してはjaccard再識別で下がった識別率が上がっていた
- チームTとKはさらに購入したアイテム集合を変えずに最頻アイテムの加工を行っていたと思われる

失敗要因

	ID	Item
1	Abe	Eraser
2	Abe	Doughnuts
3	Abe	Doughnuts
4	Baba	Eraser
5	Baba	Folk
6	Abe	Eraser
7	Baba	Eraser
8	Baba	Eraser
9	Abe	Folk
10	Baba	Doughnuts
11	Abe	Doughnuts



	ID	Item
1	Abe	Eraser
2	Abe	Doughnuts
3	Abe	POST
4	Baba	Eraser
5	Baba	Folk
6	Abe	POST
7	Baba	POST
8	Baba	POST
9	Abe	Folk
10	Baba	Doughnuts
11	Abe	POST

※POST：送料

まとめ

- 購入アイテムに注目した再識別手法freqItemを提案した
- 残念ながらfreqItem再識別はjaccard再識別よりも識別性能は劣っていた
- freqItem再識別とjaccard再識別を組み合わせることによってチームNに対してはjaccard再識別の精度の向上に繋がる

→jaccard再識別では手が届かないユーザーの再識別が可能