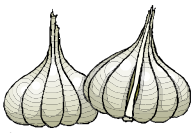
独立した調査データの融合
～汎用的データフュージョンシステムの開発～

2010/11/19



田村玄
株式会社ビデオリサーチ

本日お話しする内容

1. データフュージョンとは？
2. 開発の経緯[~2006年度]
3. 汎用的システムの開発[2007年度]

1. データフュージョンとは？

データフュージョン研究の背景

- シングルソースデータを安価に入手したい
  - シングルソースデータを取得するには非常にコストがかかる

データフュージョンとは？

- 独立して取得した2つの調査データを、サンプルの類似度に基づいて融合(fusion)すること
  - 融合したデータ(フューズドデータ)はサンプルレベルで紐づいているので、擬似シングルソースデータとして活用することができる
- 低コストで(擬似)シングルソースデータを提供できる

データフュージョン概念図

調査データA

標本番号	性別	年齢	質問a
a1	1	20	1
a2	2	30	0
a3	2	40	1

調査データB

標本番号	性別	年齢	質問b
b1	2	43	0
b2	1	22	0
b3	2	29	1

類似度が高いサンプル同士を同じサンプルとみなす

フューズドデータ

	調査データA				調査データB			
標本番号	標本番号	性別	年齢	質問a	(標本番号)	(性別)	(年齢)	質問b
f1	a1	1	20	1	b2	1	22	0
f2	a2	2	30	0	b3	2	29	1
f3	a3	2	40	1	b1	2	43	0

未調査の質問bの情報が得られる

2. 開発の経緯[~2006年度]

2.開発の経緯[~2006年度]

- フュージョン手法について
  - 「制約なし統計的マッチング」
  - 「制約付き統計的マッチング」
- フュージョンの目指すところ
  - 「制約付き」によるフュージョン
  - “集計レベル”での一致
- フュージョンシステムの開発
  - ACR(2001年度)専用にチューニング
  - 「“距離優先”制約付き」を搭載

「制約なし統計的マッチング」

調査データA

標本番号		性別	質問a
a1		1	1
a2		1	0
a3		2	0
			33.3%

調査データB

標本番号		性別	質問b
b1		1	1
b2		2	0
			50.0%

性別を類似度とみなす

フューズドデータ

標本番号		調査データA 標本番号	性別	質問a	調査データB (標本番号)	(性別)	質問b
f1		a1	1	1	b1	1	1
f2		a2	1	0	b1	1	1
f3		a3	2	0	b2	2	0
				33.3%			66.7%

「制約付き統計的マッチング」

調査データA

標本番号	ウェイト	性別	質問a
a1	2	1	1
a2	2	1	0
a3	2	2	0
			33.3%

調査データB

標本番号	ウェイト	性別	質問b
b1	3	1	1
b2	3	2	0
			50.0%

性別を類似度とみなす

フューズドデータ

標本番号	ウェイト	調査データA 標本番号	性別	質問a	調査データB (標本番号)	(性別)	質問b
f1	2	a1	1	1	b1	1	1
f2	1	a2	1	0	b1	1	1
f3	1	a2	1	0	b2	2	0
f4	2	a3	2	0	b2	2	0
				33.3%			50.0%

「制約なし」と「制約付き」

- 「制約なし統計的マッチング」
  - 基データが持つ情報がフューズドデータに引き継がれるとは限らない
- 「制約付き統計的マッチング」
  - 基データが持つ情報がフューズドデータに引き継がれる
    - ➢ *基データの平均と分散を保持するよう制約を付ける*

データフュージョンの目指すところ

- フューズドデータは、同一項目が存在する場合、“個人レベル” で一致していたほうが望ましいが･･･
- “個人レベル”での一致は困難
  - Soong, R. and M. de Montigny (2001), “The Anatomy of Data Fusion”, Session Papers of 10th Worldwide Readership Research Symposium in Venice October 2001, 87-109

	A						B				
標本番号	商品A	c1	c2	c3	c4	c5	c1	c2	c3	c4	c5
1	利用	1	1	0	1	1	0	0	0	1	1
2	利用	0	0	1	1	0	1	1	1	1	0
3	非利用	1	0	1	1	1	0	0	1	0	0
4	非利用	1	1	1	0	0	0	1	1	1	1

個人レベルの一致は5割(10/20)

データフュージョンの目指すところ

- “個人レベル”で一致していなくとも、“集計レベル”での一致を目指す
  - なぜなら、知りたいのは
    - こっち　⇒商品A利用者の集計値(率)
    - こっちではない⇒(特定)サンプルのナマ回答(ON／OFF)

	A						B				
	商品A	c1	c2	c3	c4	c5	c1	c2	c3	c4	c5
	利用	0.5	0.5	0.5	1	0.5	0.5	0.5	0.5	1	0.5
	非利用	1	0.5	1	0.5	0.5	0	0.5	1	0.5	0.5

率の一致は9割(9/10)

“集計レベル”を一致させるには

- “集計レベル”はさまざまあるので･･･
  - 商品A,B,C,･･･
- 各商品ごとの“集計レベル”が一致するよう、各商品ごとにフュージョンするのではなく･･･
- (フェイス項目など)基本的な“集計レベル”が一致するようフュージョンする
  - 得られたフューズドデータに対して、各商品ごとの“集計レベル”が一致しているかどうかを確認する必要がある

ACRフュージョンシステムの開発

- ACR(2001)に特定データをフュージョンする
  - ACR(Audience and Consumer Report)
  - 498項目について･･･
  - 基本的“集計レベル”が一致するようフュージョン
    - (12)基本的“集計レベル” = (2)男女×(6)10~60代

	ACR						data				
	集計レベル	c1	c2	c3	…	c498	c1	c2	c3	…	c498
	1										
	…										
	12										

ACRフュージョンシステムの開発

- 「“距離優先”制約付き統計的マッチング」
  - 「制約付き」は計算時間がかかるため、「“距離優先”制約付き」を搭載
  - 精度よりも時間を優先
  - 「制約付き」とほぼ変わらぬ精度を検証済み

ACRフュージョンシステムの開発

- 「サンプル間類似度」
  - 「(“距離優先”)制約付き」に必要
  - 算出式をシステム外で探索して搭載
    - 両データに共通する項目をピックアップ
      - 当然、ACR(2001)と特定データに共通する項目
      - 【例】性別と年齢
    - “最適”な係数を探索
      - “最適” =「498項目×基本的“集計レベル” が一致する」
      - 【例】α×(性別の違い)＋β×(年齢の違い)

ACRフュージョンシステムの開発　~　まとめ

- 498項目について、“基本的集計レベル”の集計値がなるべく一致するようフュージョンした
- (“基本的集計レベル”以外の)商品A,B,C･･･利用者での集計値も一致する傾向が見られた

3. 汎用的システムの開発 [2007年度]

データフュージョンに関する当時の動向

- 2007年5月、Nielsenがデータフュージョンなどさまざまな手法を用いたデータ統合サービス「Nielsen Combine」を発表
  - 今後、問い合わせや引き合いが想定される
  - 当社の競争力強化のためにも、フュージョン案件に対応できるようにしておく必要がある

ACRフュージョンシステムの問題点

- 「サンプル間類似度」の非汎用性
  - 「サンプル間類似度」の算出にはACR(2001)と共通する項目が必要
  - 「サンプル間類似度」を計算する式は、既述の498項目×基本的“集計レベル”にのみ最適化

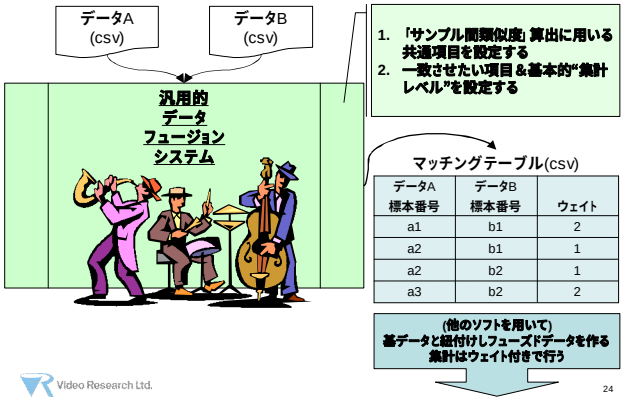
汎用的データフュージョンシステムの開発

- 「制約付き」を搭載
  - 「制約付き」は計算時間がかかると言われていたが、昨今のハードウェアおよびアルゴリズムの進歩により問題ないことが判明
    - “輸送問題”の解法を適用する

汎用的データフュージョンシステムの開発

- 「サンプル間類似度」に汎用性を持たせる
  - 「サンプル間類似度」の算出に用いる項目は自由
  - 「サンプル間類似度」を計算する式は、様々な項目×様々な“基本的”集計レベルに最適化するよう、その都度システム内で生成する
    - “割り当て問題”の解法を適用してサンプルを“基本的”集計レベルに割り当てる
    - 共通項目を用いて、割り当て結果を判別する式を作り、判別値の大きさを類似度とする

システムの概要



システムのデモンストレーション

- データ
  - データA
    - 約3,000サンプル
  - データB
    - 約1,000サンプル

【計算時間】
設定した反復計算時間＋約17分
※高スペックPC(2007年夏現在)使用

システムのデモンストレーション

- 設定項目
  - 「サンプル間類似度」算出に用いる共通項目
    - 171項目
  - 一致させたい項目＆基本的“集計レベル”
    - 498項目
    - 8レベル
      - M12~19才,M1,M2,M3,F12~19才,F1,F2,F3

フューズドデータの検証1

- フューズドデータが基データの情報を維持しているかどうかを確認する
  - ➢フューズドデータの平均と分散が、基データの平均と分散と同じかどうかの確認
  - 498項目について確認した

検証1~概念図

B

標本番号	c1	c2	…	c498
B1			…	
B2			…	
…	…	…	…	…
Bn			…	
平均	XXX	XXX	…	XXX

Aにフュージョン

A

標本番号		…	(B) c1	c2	…	c498
A1		…			…	
A2		…			…	
…	…	…	…	…	…	…
An		…			…	
平均			XXX	XXX	…	XXX

ここを確認

フューズドデータの検証2

- 設定した「一致させたい項目＆基本的“集計レベル”」が、フューズドデータに反映されているかどうかを確認する
  - ➢例えば「男女ごとに特定の項目」が一致するように設定した場合に、得られたフューズドデータがその旨を反映しているかどうかの確認
  - 設定した項目＆基本的“集計レベル”
    - 498項目
    - 8レベル
      - M12~19才,M1,M2,M3,F12~19才,F1,F2,F3

検証2~概念図

フューズドデータを集計

A

標本番号	集計レベル	…	c1	c2	…	c498
A1		…			…	
A2		…			…	
…	…	…	…	…	…	…
An		…			…	

B

c1	c2	…	c498
		…	
		…	
…	…	…	…
		…	

A

集計レベル	…	c1	c2	…	c498
M~19	…			…	
M1	…			…	
…	…	…	…	…	…
F3	…			…	

B

c1	c2	…	c498
		…	
		…	
…	…	…	…
		…	

ここを確認

フューズドデータの検証3

- 設定した「基本的“集計レベル”」以外の“集計レベル”における一致度合いを確認する
  - ➢例えば「男女ごとの特定の項目」が一致するように設定した場合の、「商品A(B,C,･･･)利用者の特定の項目」の一致度合いの確認

検証3~概念図

フューズドデータを集計

	A						B			
標本番号	基本的～	商品利用	c1	c2	…	c498	c1	c2	…	c498
A1					…				…	
A2					…				…	
…	…	…	…	…	…	…	…	…	…	…
An					…				…	

		A					B			
		商品利用	c1	c2	…	c498	c1	c2	…	c498
		商品A			…				…	
		商品B			…				…	
		…	…	…	…	…	…	…	…	…
		商品Z			…				…	

ここを確認

おしまい