NIKKEI-R

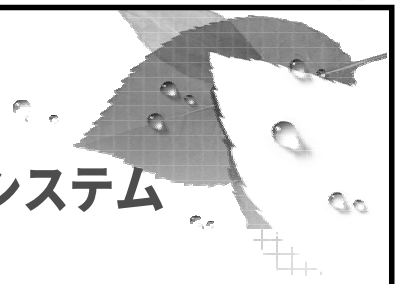
クロス集計の新しい形

ネットワーク図による
データ可視化システムの開発

2008/9/3
日本行動計量学会 第36回大会　好みの計量 セッション
（株）日経リサーチ 佐藤 邦弘

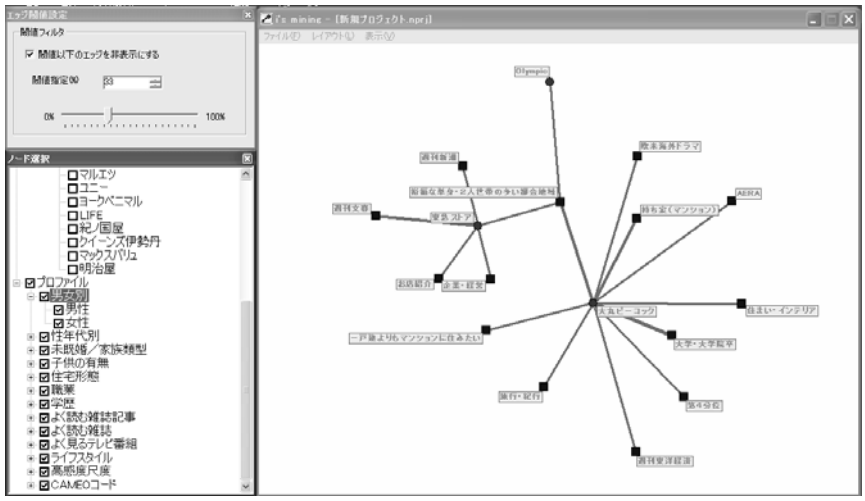
株式会社日経リサーチ

概要



❖作ったシステム

- ネットワークによるデータ可視化システム
- “i”s Mining の開発



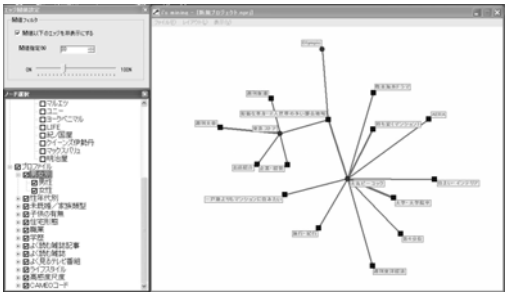
※ (株) 日経リサーチが数理システムの協力を得て開発

株式会社日経リサーチ

概要

❖“i”s Mining とは。

“i”s Mining



=

クロス集計表

Q1.以下の店舗・サービスブランドについておうかがいします。(いくつでも)
○○○

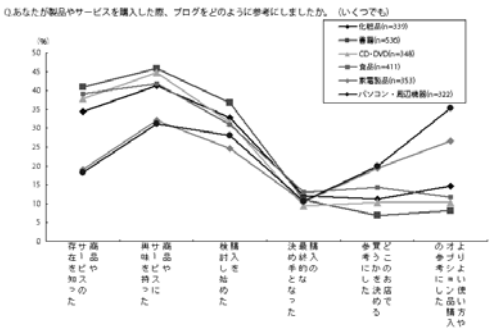
	回答者数	名前を知っている	利用したことがある	今後または今後とも利用したい	今より多少値段があがっても利用したい	他の人に薦めたことがある、または薦めたい
全体	619	96.6	66.7	36.5	4.8	4.8
■性別						
男性	311	97.7	69.5	37	4.2	3.2
女性	308	95.5	64	36	5.5	6.5
■年代別						
29歳以下	149	99.3	73.2	40.9	7.4	8.7
30歳代	160	100	71.3	36.9	3.8	3.1
40歳代	155	98.1	67.7	38.1	4.5	5.2
50歳以上	155	89	54.8	30.3	3.9	2.6
■性年代別						
男性29歳以	69	98.6	76.8	34.8	4.3	4.3
男性30歳代	80	100	75	43.8	6.3	3.8
男性40歳代	80	98.8	72.5	38.8	1.3	3.8
男性50歳以	82	93.9	54.9	30.5	4.9	1.2
女性29歳以	80	100	70	46.3	10	12.5
女性30歳代	80	100	67.5	30	1.3	2.5
女性40歳代	75	97.3	62.7	37.3	8	6.7
女性50歳以	73	83.6	54.8	30.1	2.7	4.1
■地域						
北海道・東	49	93.9	40.8	20.4	4.1	4.1
首都圏	252	100	84.1	52.8	7.9	7.1
関東・甲信	40	97.5	65	32.5	7.5	5
東海・北陸	71	91.5	38	11.3	1.4	1.4
関西	131	94.7	69.5	34.4	3.1	5.3
中国・四国	40	90	32.5	10	0	0
九州・沖縄	36	100	66.7	36.1	0	0
■職業						
男性勤めノ	198	99	72.7	38.4	5.1	3.5
女性勤めノ	71	98.6	76.1	36.6	8.5	4.2
パート・アノ	83	95.2	60.2	30.1	3.6	4.8
自営業・自	59	100	61	32.2	3.4	1.7
学生	41	100	80.5	51.2	7.3	12.2
専業主婦	126	92.1	60.3	34.9	3.2	5.6
無職	28	89.3	53.6	35.7	7.1	7.1
その他職ま	12	91.7	33.3	33.3	0	8.3
■高感度人間						
第一分位	110	99.1	60.9	31.8	2.7	1.8
第二分位	144	96.5	64.6	32.6	6.3	4.2
第三分位	224	95.5	67.4	36.2	3.6	5.8
第四分位	141	96.5	72.3	44.7	7.1	6.4

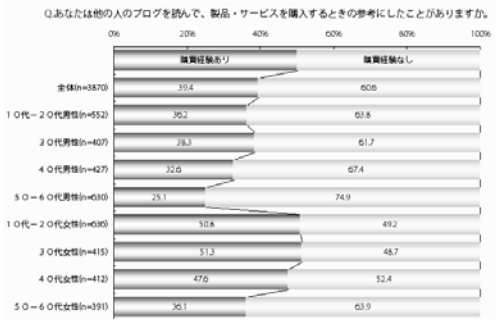
クロス集計表を
ネットワーク図で可視化

背景

❖調査業界の現状

アウトプットのクロス集計表への依存度の高さ





商品ジャンルによる
情報収集経路の違い

ブログ情報の影響度の
性年代別の違い

日経リサーチ データシグナル - 女性の半数がブログを参考に商品購入 - 調査結果より

株式会社日経リサーチ

クロス集計のメリット･デメリット

メリット

•分析者に依存しない。
•***判断が入らない。***

•説明のしやすさ。
•***相手に数学の知識を求めない。***

デメリット

•結果の読み取り精度の低下
• *ex. **全体**:50% **での** +10%**の差** と、**全体**:20% **での** +10%**の差も同じ**10%*

•要約の度合いの低さ
•***莫大な表を読み取る。コスト高、見逃し。***

株式会社日経リサーチ

開発の目的

❖ **クロス集計の「分かりやすく」「人に依存しない」部分を残しつつ、「統計的な怪しさ」「情報集約度の低さ」を削減する。**

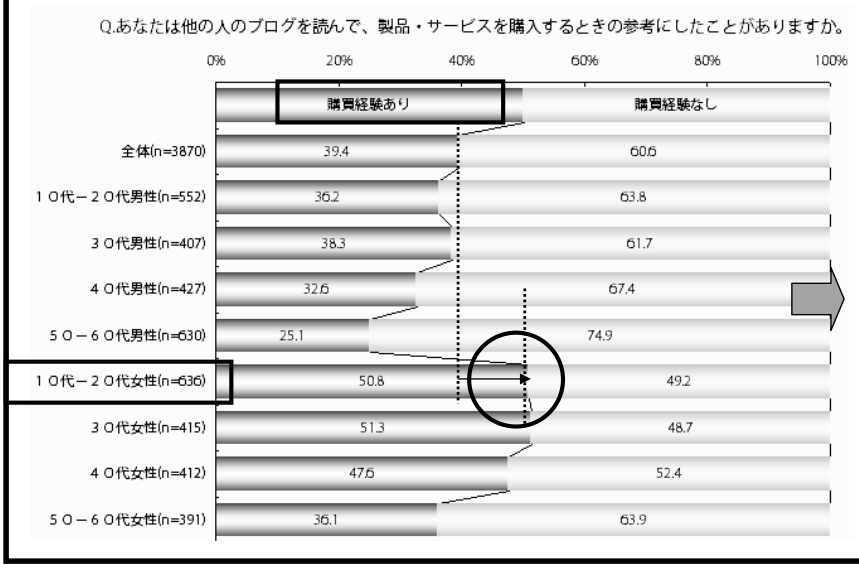
❖ **これにより、データの「読み取りのスピードアップ」と「読み間違い･見逃し」を減らす。**

株式会社日経リサーチ

現状分析

❖クロス集計で行っていることを抽象化

- 質問軸と分析軸の組み合わせから
- 差を読み取る



「１０代－２０代女性」
は、
全体と比べて、
「ブログを参考に購入したこと」が
「多い」
傾向がある。

株式会社日経リサーチ

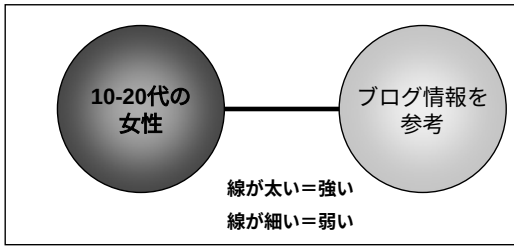
現状分析

❖クロス集計で行っていることを抽象化

「１０代－２０代女性」
は、
「ブログを参考に購入したこと」が
「全体と比べて、多い」
傾向がある。

３つの要素
見ているのは「関係性」

ネットワークでの可視化が向いてる？



株式会社日経リサーチ

© 2008 Nikkei Research Inc. All Rights Reserved.

日本行動計量学会 第36回大会 － 好みの計量 セッション

手続き

❖方針

- クロス集計をネットワークで可視化する

⇩

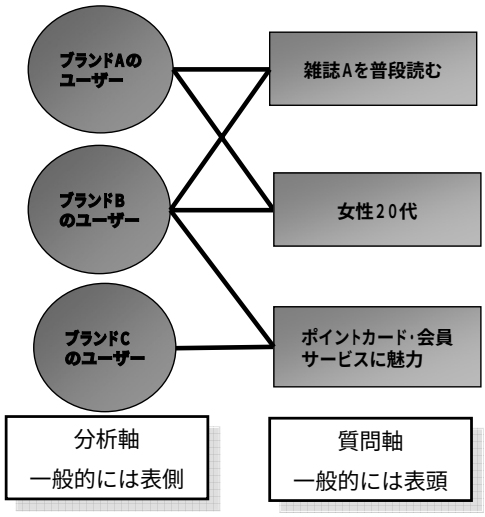
- 2つの工夫
  - ネットワークの構造（変換ルール）・・・ Structure
  - 関係の強さを測る尺度　・・・ Metric

株式会社日経リサーチ

2つの工夫：① 構造 - structure

❖ ノードを2種類に分けたグラフ
(二部グラフ、bi-partite)として表す。

Ex. 分析軸をブランド毎のユーザーの特性を知りたい場合



関係性は、２部グラフで共有関係をみるべき。

ワッツ他 (2004)
『スモールワールドネットワーク』, 阪急コミュニケーションズ より

株式会社日経リサーチ

2つの工夫：② 測度 - metric

❖ では、次に、特徴の強さをどうやって測るのか？

- **満したい要件**

  - **クロス集計表から差が多いところを抽出できるようなもの。**
  - **関係が「ある」 / 「ない」の2値ではなく、関係の強さを連続的に表せるもの。**
  - **n数の異なる分析軸間でも、フラットに比較できるもの。**
    - Ex)
      ブランドAのユーザー 200人 にみられる特徴の強さ。
      ブランドBのユーザー10人にみられる特徴の強さ。

2つの工夫：② 測度 - metric

❖ シャノンの情報量

$$I = -\log P$$

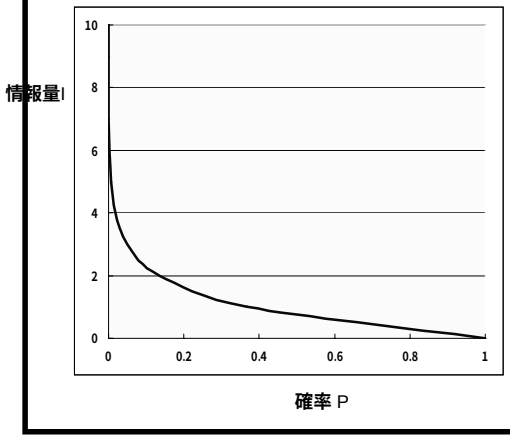
$$= -\log\left(\frac{W_{after}}{W_{before}}\right)$$

$$= \log W_{before} - \log W_{after}$$

$$= \text{分からなさ度合い}_{before} - \text{分からなさ度合い}_{after}$$

※ W：取り得る場合の数

情報エントロピーの差分



・3通りが3通りのまま(100%)

情報量 ; -log(1)=log(3)-log(3)=0

『赤か青か黄の線のどれかだ！』

・3通りを1通りに減らす (33.3%)

情報量 ; -log(1/3)=log(3)–log (1) =log3

『赤の線だ！』

画像引用元： 株式会社バンプレスト
起床装置 DANGERBOMBCLOCK

2つの工夫：② 測度 - metric

NIKKEI-R

- クロス集計で考えてみると？

	20代	not 20代	計
ソニー愛好者	45	5	50
全体	60	40	100

「ソニー愛好者には、20代が多い」といえるか？

→『+30%』の差を 情報量（確率）の考え方で考える。

確率の定義：

100 個のレコードから、ランダムに50個のレコードを抽出したとき、

20代が、45人いる確率を計算する。

もし、確率が高ければ、ソニー愛好者にとって、20代のいる人数は、

ランダムとほとんど見分けがつかない。

→情報の価値は、ない。とみなす。逆なら、あるとみなす。

日経リサーチ

2つの工夫：② 測度 - metric

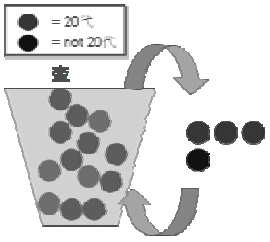
NIKKEI-R

- 全データは有限か無限か？
  無限 → いくら玉を取り出しても、20代の玉は、常に60％で出現

	20代	not 20代	計
ソニー愛好者	45	5	50
全体	60	40	100

$$
\begin{aligned}
I &= -\log(P) \\
&= -\log\left\{{}_{50}C_{45}\,0.6^{45}0.4^{5}\right\} \\
&= -\left[\log {}_{50}C_{45} + 45\log 0.6 + 5\log 0.4\right] \\
&= -\left[\log\frac{50!}{45!\cdot 5!} + 45\log 0.6 + 5\log 0.4\right] \\
&= -\left[\sum_{i=1}^{50}\log i - \left(\sum_{i=1}^{45}\log i + \sum_{i=1}^{5}\log i\right) + 45\log 0.6 + 5\log 0.4\right] \\
&= 6.42
\end{aligned}
$$

復元抽出



これがテキストマイニングなどで
使われる一般的な情報量らしい・・・。

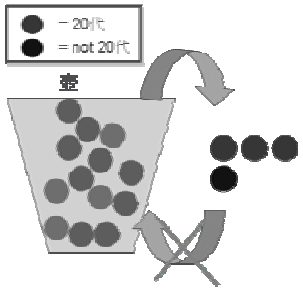
2つの工夫：② 測度 - metric

- 全体の数は、有限か無限か？
  ：有限 →20代の玉は最初は、60%で出現。あとは、残りの玉数によって変化。

	20代	not 20代	計
ソニー愛好者	45	5	50
全体	60	40	100

$$
\begin{aligned}
I &= -\log P \\
&= -\log\left(\frac{{}_{60}C_{45}\times{}_{40}C_{5}}{{}_{100}C_{50}}\right) \\
&= -\left(\log\frac{60!}{45!\cdot 15!}+\log\frac{40!}{5!\cdot 35!}-\log\frac{100!}{50!\cdot 50!}\right) \\
&= 21.78
\end{aligned}
$$

非復元抽出



株式会社日経リサーチ

2つの工夫：② 測度 - metric

- 無限を仮定した場合と、有限を仮定した場合では、（セルに小さい結果が有るときは特に）、結果が全く変わってくる。

	20代	not 20代	計
ソニー愛好者	45	5	50
全体	60	40	100

無限を仮定：6.42

P=exp(-6.42)=0.00163=1.63 x 10^{-3}

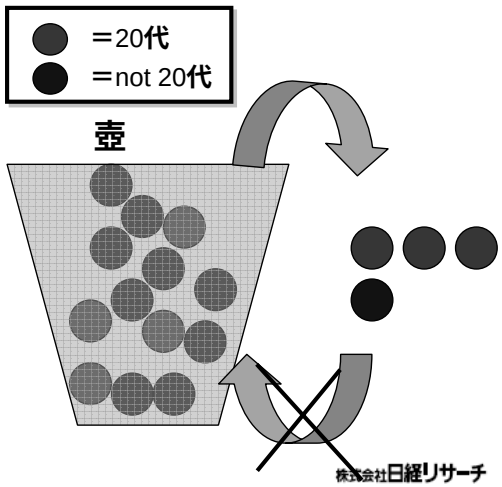
確率に７桁の差！

有限を仮定：21.78

P=exp(-21.78) =3.48 x 10^{-10}

玉を戻さないと、赤玉が出るほど、
壺の中の赤玉が減ってくるため、なかなか出なくなる。

確率が小さくなる→情報量大



株式会社日経リサーチ

2つの工夫：② 測度 - metric

❖ 有限 or 無限?

単純に手元のデータの
特性だけをしらべたい。

- 手元のデータをそのまま母集団とする（サンプルをそもそも考えない。）
  →有限 ：100のデータのうち60が男性データ

- データはその裏にある母集団から抽出したサンプルと考える
  →無限：無限データのうち60%が男性データと推測

調査で得られたのは、無限大のデータ
から得られたサンプル結果とし、
全データを仮定し推計したい。

株式会社日経リサーチ

2つの工夫：② 測度 - metric

❖ まとめ

- 満したい要件

  - クロス集計表から差が多いところを抽出できるようなもの。
    →○
  - 関係が「ある」 /「ない」の2値ではなく、関係の強さを連続的に表せるもの。
    →○
  - n数の異なる分析軸間でも、フラットに比較できるもの。
    →○

    例） 全数 1,000 人 のうち、
    マイナーな雑誌 Aの読者が、10人　いたとする。

    そのとき、
    あるブランドのユーザー 10人のうち、8人が、その読者

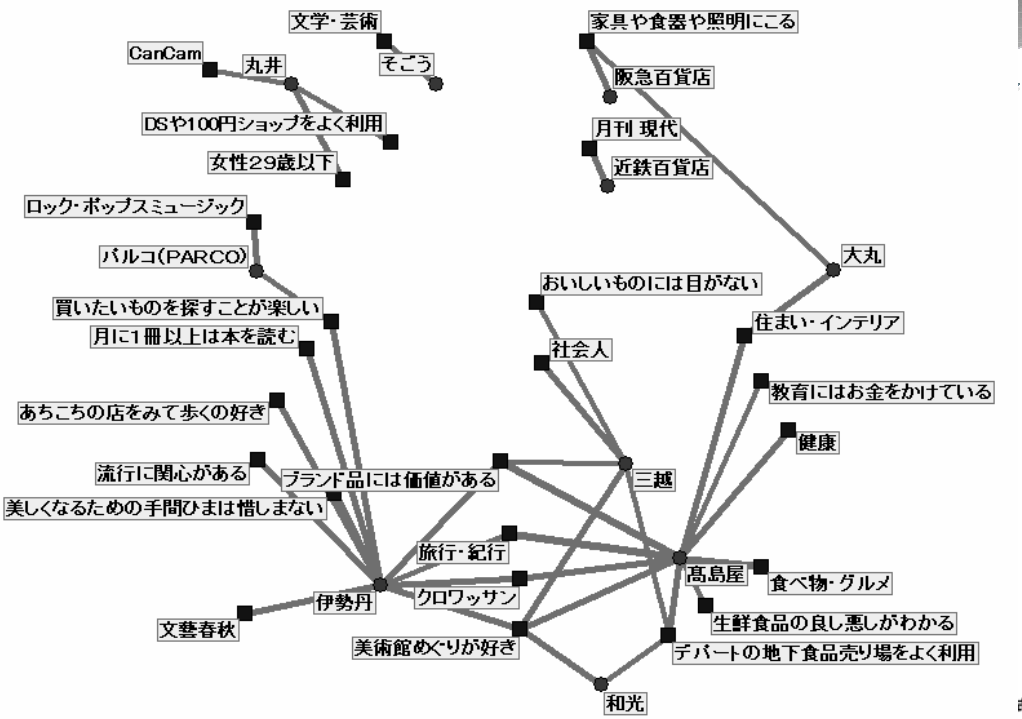
    ランダムでは、あり得ない➔ 大きな値となる。

    マニアックなクルマのユーザーなど、データ数が少なくても、結構うまくでる。

アウトプット例1

(株) 日経リサーチ ストア＆サービスブランド調査2006 より

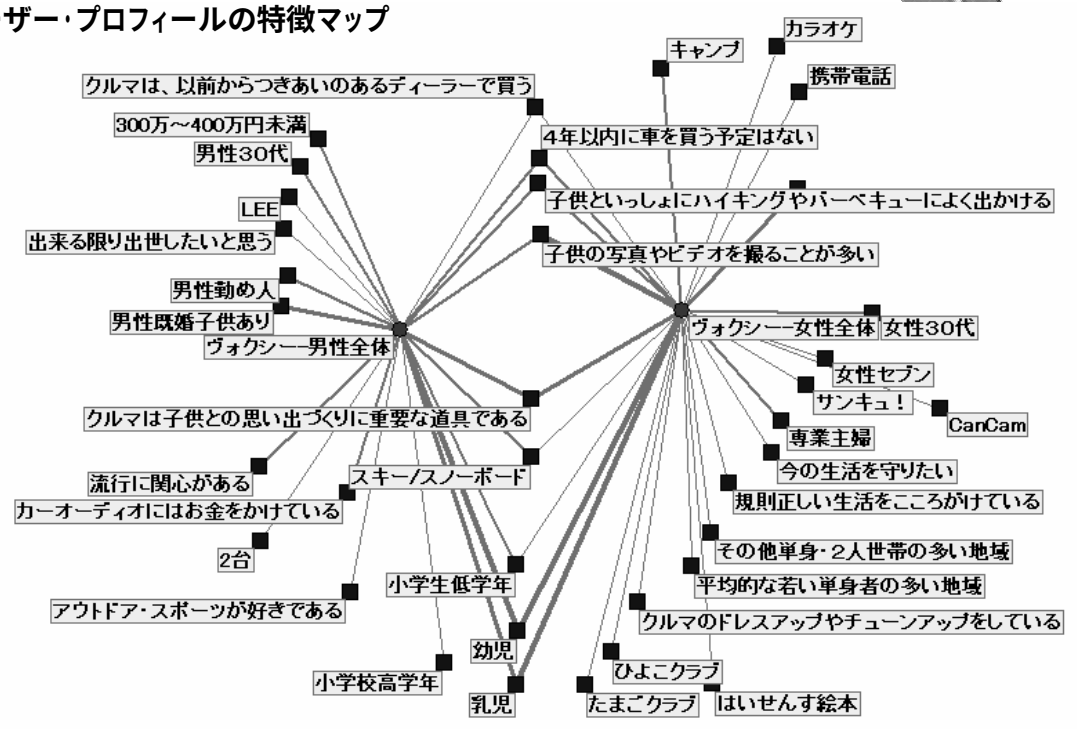
店舗毎のロイヤル層 (今より高い価格を払っても利用したい) の特徴マップ



アウトプット例2

(株) 日経リサーチ カーブランディング調査2007 より

ユーザー・プロフィールの特徴マップ



アウトプット例3

ポルシェユーザー（n=28）の特徴ランキング

販売物の元データのため、削除しています。

株式会社日経リサーチ

少ない方の特徴でみると‥？（こんなポルシェユーザーは嫌だ！）

販売物の元データのため、削除しています。

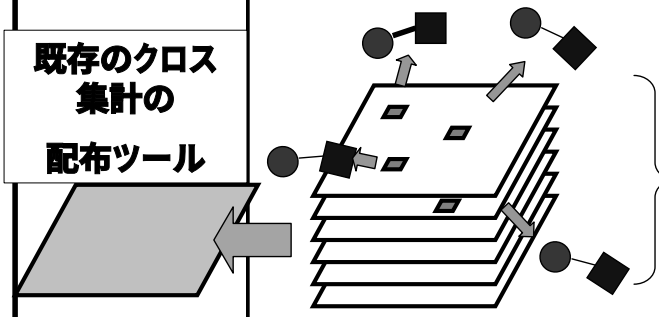
株式会社日経リサーチ

© 2008 Nikkei Research Inc. All Rights Reserved.

日本行動計量学会 第36回大会 ー 好みの計量 セッション

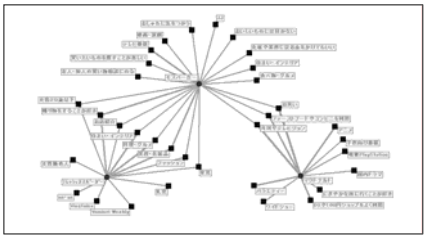
まとめ

既存のクロス集計の配布ツール



1枚を選択

集計表



意味のあるところだけを抜き出して、1枚のネットワーク図に

- 質問を串刺しでみることで、知見の抽出が速い。
- 効く質問と効かない質問がひとめで分かる。
- １０万～１００万の２ｘ２のクロス集計を見落とし無く見られる。

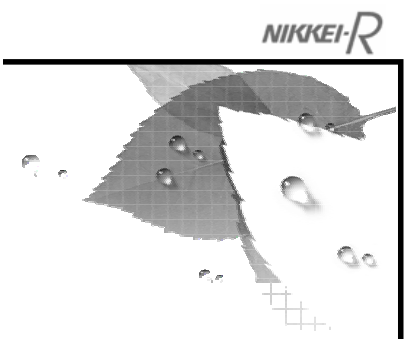
株式会社日経リサーチ

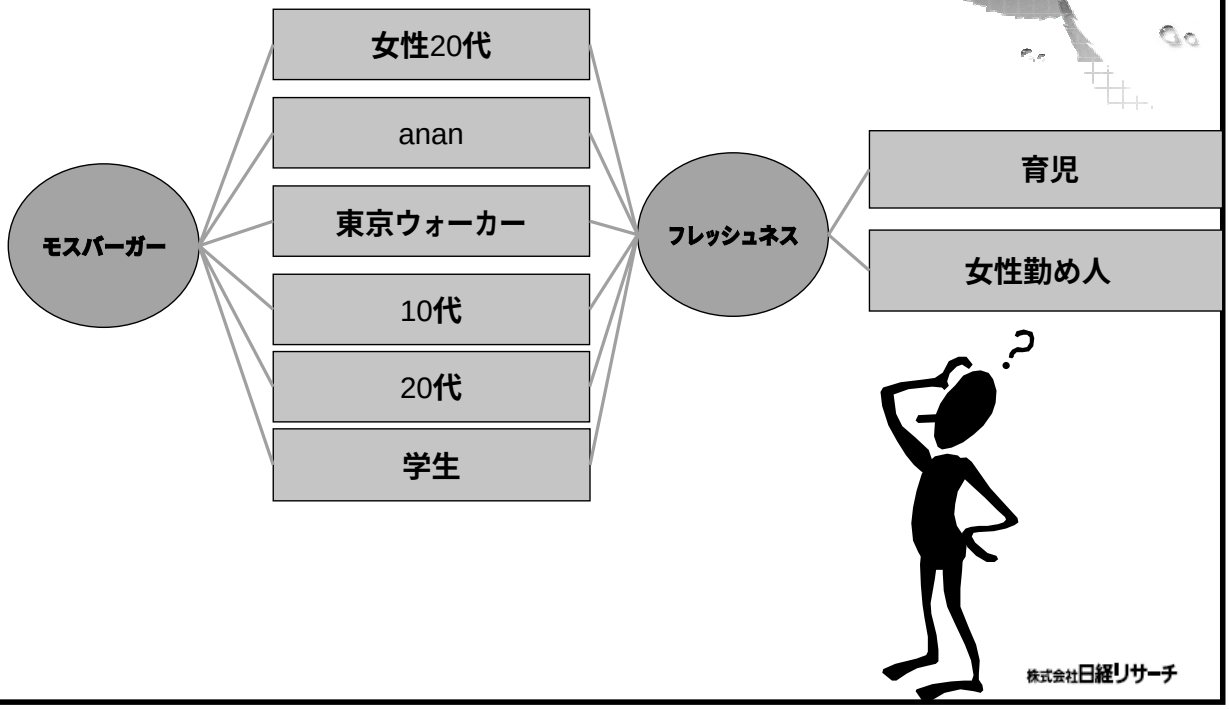
おまけ：探索的データ解析

おまけ：探索的データ解析

株式会社日経リサーチ

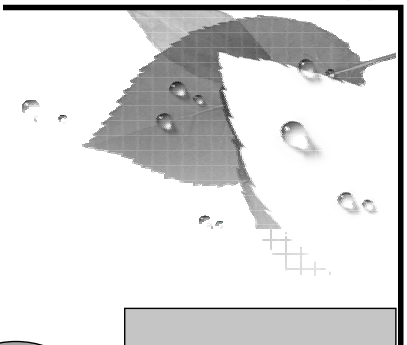
探索的データ解析

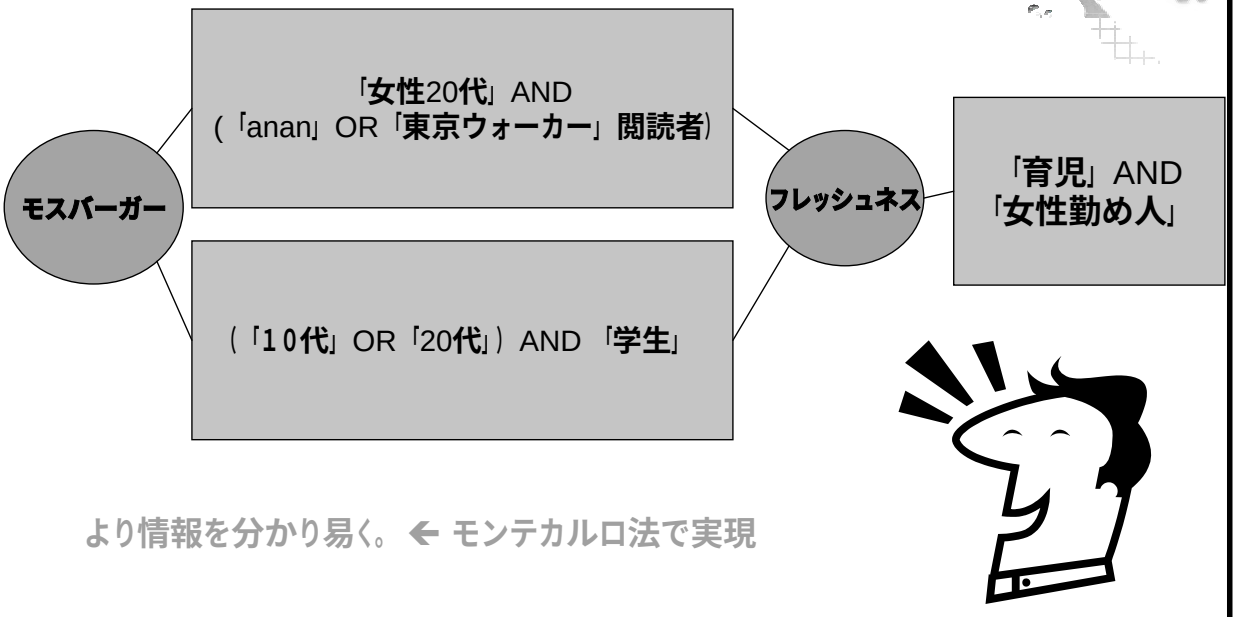




株式会社日経リサーチ

探索的データ解析





より情報を分かり易く。← モンテカルロ法で実現

株式会社日経リサーチ

探索的データ解析

❖ 手法

- **モンテカルロ法。**
  - ① 同一質問内で、ランダムに2つの属性ノードを選択。
  - ② OR 条件で繋げて、合成変数を作成。
    Ex)「週刊新潮」or「FRIDAY」閲読者 など。
  - ③ 個別で見たときよりも
  - ④ 情報量が高い ➔ 採用。
    情報量が低 い➔ 不採用。

LOOP

  - ⑤ 別質問間で、ランダムに2つの属性ノードを選択。
  - ⑥ AND 条件で繋げて合成変数を作成。
    Ex)「男性」and「車が好き」
  - ⑦ 個別で見たときよりも、
    情報量が高い ➔ 採用。
    情報量が低 い➔不採用。

LOOP

Bottom Up ﾒ やみくもに結合を実施する CHAID のようなものと考えればよい。

株式会社日経リサーチ

探索的データ解析

❖ OUTPUT例 (フレッシュネスバーガー：再利用意向者)

http://www.obubu.com/blog/ archives/cat_cat1.html

株式会社日経リサーチ

探索的データ解析

フレッシュネスバーガー：再利用意向者：特徴ランキング

1	【子供】乳児（0～2歳）、【子供】子供はいない	20.6
2	贈り物をすることが好き	18.5
3	買い物	17.4
4	【雑誌記事】(育児、テレビ番組、その他) AND 自分の体や健康によい商品を買う	16.7
5	男性･29歳以下、男性･30代、女性･29歳以下、女性･30代	16.5
6	【雑誌記事】お店紹介、その他	14.3
7	(短大･高専･専門学校卒、大学･大学院（旧制高、旧制高専）卒) AND 【雑誌記事】ファッション	14.2
8	男性･30代、女性･29歳以下、女性･30代	13.9
9	【雑誌記事】育児、住まい･インテリア、映画･演劇、その他	13.7
10	お店紹介	13.7
11	【雑誌記事】住まい･インテリア、音楽、その他	13.5
12	大学･大学院（旧制高、旧制高専）卒 AND 女性･29歳以下	13.4
13	住まい･インテリア、その他	13.4
14	大学･大学院（旧制高、旧制高専）卒 AND (流行に関心がある、化粧や美容にはお金をかけてもい	13.1
15	【雑誌記事】(育児、テレビ番組、その他) AND 環境や自然にやさしい商品を買う	12.5
16	住まい･インテリア	12.3
17	育児、映画･演劇、その他	11.7
18	(男性･29歳以下、女性･29歳以下) AND 飲み会やパーティーに参加する	11.2
19	男性･29歳以下、女性･29歳以下、女性･30代	11.1
20	欧米海外ドラマ、食べ物･グルメ、その他	10.9

株式会社日経リサーチ

Thank You !

株式会社日経リサーチ