

JMPの統計解析機能を活かした テキスト・マイニング

・ WordMiner™との相互補完的活用法 ・

JMPer's Meeting

2005年10月4日

テキスト・マイニング研究会・代表
統計数理研究所

大隅 昇
ohsumi@ss.iij4u.or.jp

テキスト・マイニング研究会・事務局長
(株)平和情報センター
プロダクト営業部 担当部長

保田 明夫
yasuda@hic.co.jp

All rights reserved. Copyright by Noboru Ohsumi, ISM Professor Emeritus.

本日のトーク

- JMPer's Meetingではあるが、始めに「まえおき・背景」として、テキスト・マイニング(TM)についての私見を若干述べる。
- JMPの優れた統計解析機能の、テキスト・マイニング(テキスト型データの統計解析)への適用可能性を探る。
- とくにJMPの探索的データ解析機能の援用(初動探査)。
- version 6.0になって文字情報(テキスト型データ)の処理の自由度が格段に増したことの紹介(有効利用が可能)。
- テキスト型データ・マイニングを行う専用ソフト: WordMinerとJMPの相互補完利用の可能性を探ること。
- 主にインターネット調査(Web調査)で取得のデータで確認。

結論は、...

- JMPは数値型データ(量的データ, 質的データ)の解析に適した解析用ソフトであること(再認識).
 - 注: 量的データ: 区間尺度, 比例尺度; 質的データ: 名義尺度, 順序尺度
- とくに, 初動探査, 探索的利用に効果的である.
- ボリュームが大きく, また非等質・非構造的であるテキスト型データを, 計量化・数量化, あるいは指標化した後の分析には有効と思われる.
- 変数名の文字数制限, テキスト型データの文字制限が無くなったことはきわめて便利である.
- テキスト型データのボリュームが増えると負荷が増える.
- グラフィカル表現の機能は有効であるが, 出力表示に若干の難がある.
 - 文字情報の表示の可読性
 - 文字数, 要素数(表示の個数)が増えたときの対応 何か工夫があるか?
- 成分スコアのような加工済み・分析結果データの利用に若干の制約があるようにみえる(実際のデータ解析では, 加工済みデータの二次分析の機会も多い).

3

現状のテキスト・マイニング(概要俯瞰)

- 定性情報, 質的情報の利用場面が加速的に増え, かつ多様化してきた(あるいは様相が変わった).
- 文字情報(テキスト型データ)だけでなく, 一般に定性情報, 質的情報の電子的取得が容易となってきた.
- データ取得・収集法の技術改善は大きいが, 取得情報の分析方法が満足な環境にないと思われる.
- 同時に, 分析・解析のためのソフトウェア環境が整いつつあるようだ(器・ウェアはあるが利用のための知識が弱い).
- 統計ソフトウェア開発各社が競ってデータ・マイニング(DM)やテキスト・マイニング(TM)向けのツールのリリース.
- テキスト・マイニング対応をうたう多数のソフトの登場.
- 多くは, データ・マイニング・ツールの亜種に見える.

4

テキスト・マイニングとは: 定義をいくつか, ...

定義1:

- データベース等に蓄積された大量のテキスト, 文書(ドキュメント)情報の中から, 目的にあったテキストや文書を検索収集し, それらの間に潜在する関連性を分析し, 類型化し, さらにその内容や情報を計量化し, またその探査の推移を把握することから, 新たな知見を得る一連の接近方法をいう.
- 技術的には, 大量のテキスト, 文書を数値化データと同様に自由にハンドリングして(データ処理), 潜在する隠れた事実や関連性を発見することを目的とし, 原始テキスト型データを直接扱うこと.

5

定義2:

- 未発見の鉱山, 鉱脈(mine)である大規模なテキスト・コーポラを想定して, どこに有用な情報(宝の山, 金鉱)があるかを探し, 予想もできなかったような情報や知見を発見すること.
- テキスト・マイニング・ツールを用いてテキスト・コーポラの内容を俯瞰し, 現象の明解な読み解きのきっかけとなる情報をユーザに提供すること, 隠れた意味ある類似性を発見すること, 関連情報の類似性を探索すること, それらを要約, 視覚化し, 理解可能な情報に変換すること, などを行うことをいう.

6

定義3:

- 自然文や自然言語テキスト(言葉の表記体), 文書の集合体の中にある規則性, パターン, 傾向を探索することである. また, 通常は, これらテキストを特定な目的をもって科学的に分析・解析することを行う.
- 例えば, 高度に構造化されたデータベースから, 顕著なパターンを発見する, データマイニング技法に基づく, あるいはその援用を受けたテキスト・マイニング手法により非構造的なテキストから, 有用な知識, 知見を引き出すことを目的とする.

7

「共通項」が見える

- 大量の文書, テキストの処理を行うこと
- 大規模データベース, ドキュメント・ウェアハウスを用いること
- テキスト・コーパス(コーポラ)を想定
- 規則性, 類似性, パターンの探索, 特徴付けを目指す
- 関連情報(関連性)やそれらの連鎖を発見すること
- 例外的なもの, 変則的なものに目星を付けること
- 構造化データと非構造化データ
- データ処理, データ解析
- 情報検索と情報管理
- 情報, とくに大量なテキスト情報の視覚化・可視化
- 情報の知識化, 知識の発見と取得
- 「テキスト」という用語がなければDMにほとんど同じようにみえる
- TMIはDMの亜種と言われる所以である

8

TMの適用範囲の増大, 関連する最近の話題

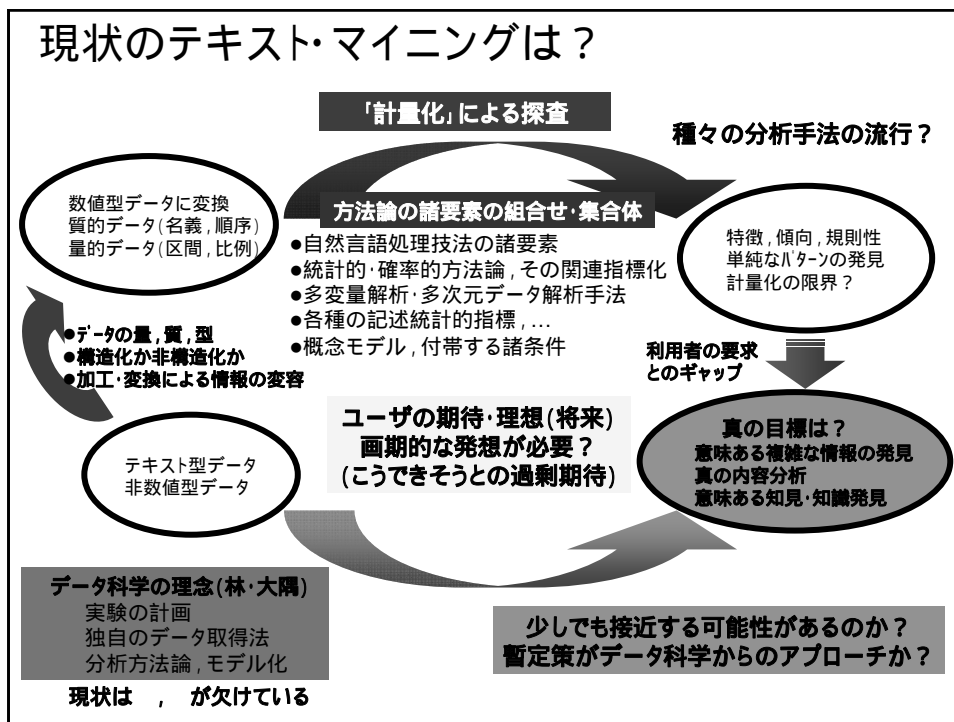
- 定性情報の分析法・質的研究 (Qualitative Research) の見直し
 - 市場調査, 社会調査 (意識調査, 態度調査), ...
 - エスノグラフィー (集団観察他)
 - 福祉・看護, 介護問題, ... (日記, 聞き取り調査, ...)
 - 質的心理学研究, ...
- 内容分析 (Content Analysis)
 - 研究の長い歴史がある分野
 - CACA (Computer-assisted Content Analysis) として進展
- コーディング処理
- データ取得環境の観点からは電子的取得, 多様化がある
 - 調査の自由回答 (open-ended questions), FG・OFG, ...
 - ディスコース分析 (発語・発話・言説分析)
 - 日記形式, ブログ, 聞き取り調査, ...
 - コールセンター, コンタクトセンター, ...

9

「データの様相」が多様化, これをどう考える

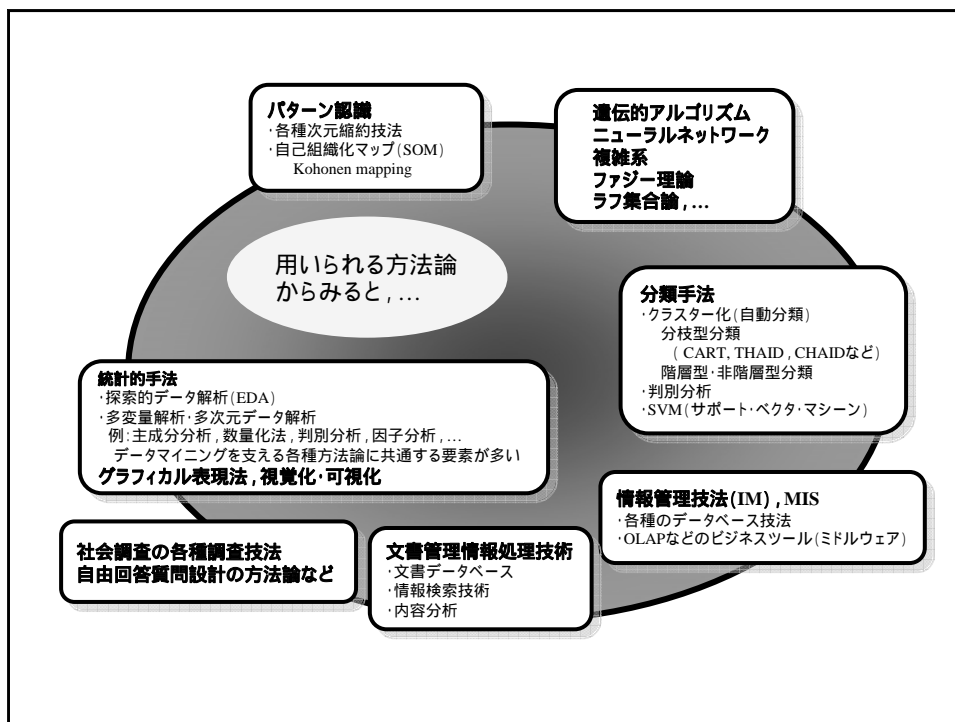
- 電子的取得 何でも集まる, 集められる (そうみえることが問題)
- 入力データの多様性から, TMを一元的に議論できない
- 「意図的に集める」vs 「既に集められたもの / 集まってくる」
- 「構造的」vs 「非構造的」
 - 構造的 RDB, DWH, ... (実はなかなか得られない, 整備が不十分)
 - 非構造的 E-mail・Web, コール / コンタクト・センター, ...
 - 非構造的 大きく二つの意味がある
- 「等質」vs 「非等質」
 - 統計学では原則, 等質的データを想定している
 - 母集団と標本, 無作為などの設定要件があること
- 日頃のデータ解析体験から理想条件を満たすデータは皆無
- とくにTMでは, データの様相と分析方法のミスマッチが顕著
- どう取得したか (収集方式), 何のためか (目的), データ履歴の透明性

10



関連研究分野, 方法論として, ...

- テキスト・マイニングが対象とする“目標”, “焦点”は?
- 見方・視点によって様々である
- どの(研究)分野に軸足を置くか, どこに焦点をあてるかで, テキスト・マイニングの考え方は様々あり得る(多様)
- データ・マイニングに通底するものがある(共通要素)
- 例えば, ...
 - 関連研究分野から
 - 用いられる方法論から, ...



欧米のTM研究は課題が広範囲(TMを広く解釈)

- テキスト・カテゴリーゼーション(text categorization)
- ドキュメント分類(document clustering, document classification)
- ルール探索, ルール発見(rule mining from text)
- 概念抽出, 関係の発見(concept, relationship mining from text)
- 情報の統合化, 有機的統合化(information integration)
- 特定なトピックスの検出(topic detection)
- テキスト・文書要約化(summarization, analysis of text collection)
- 知識取得・獲得と理解(knowledge acquisition/capture & understanding)
- テキスト・ナビゲーション, 視覚化(text navigation, visualization)
- Webへの応用(Webマイニング, テキスト学習, 知的エージェント化)
- 生物情報学への応用(ゲノム解析, 生物文献情報処理など)
- ビジネスへの応用(CRM, 意見のマイニング)
- 調査データの分析への応用(自由回答, 自由記述)
- テキスト検索・全文検索(full text documents search, document retrieval)

15

「データ科学」の理念: データをどう考えるのか

- データ科学(data science)とは?
 - WordMiner設計指針の基本理念
 - 現象解析の基本は「データ」にあると考える(data-driven)
 - つまり「データによる現象理解」を前提とする
 - 統計的データ解析は常に仮説発見的かつ探索的である(仮説検証的)
 - 主要なツールが多変量解析・多次元データ解析他のデータ解析関連手法を用いる統合的アプローチ
- **基本の理念**
 - 実験計画: データをいかにして計画的に取得するか(experimental design)
 - データ収集法: データをいかに集めるのか(data collection mode)
 - 解析法: 問題とする現象解明に適した独自の解析法はどうあるべきか(analyzing)
 - ~ を有機的かつ探索的に“行きつ戻りつ”する過程をいう
 - 当たり前のことのようにだが実現が厄介なこと(とくにデータ取得環境の確保)

16

とくに市場調査，社会調査における適用可能性に注目

- 市場調査，社会調査等では定性情報 (qualitative information) の利用場面が多々ある。
- 数量・数値として取得できる定量的情報ばかりではない。
- 世の中の流通情報が全体に定性的情報に移行(非数値情報処理が一般的となりつつある)。
- とくに(社会)調査における自由回答・自由記述データの取得。
- グループインタビュー(GI)やフォーカス・グループ(FGI)等で取得の定性型データ(とくに，オンライン・フォーカス・グループ: OFG)。
- 自由回答・自由記述等のテキスト型データの利用法がテキスト・マイニングとの理解(狭義 ただし，これはTMの一部にすぎない)。
- 広義のテキスト・マイニング(ドキュメント・マイニング，Webマイニングなどを含む)と意識的に分けて考える必要がある。

17

参考情報

- 大隅昇，保田明夫(2004): テキスト型データのマイニング - 定性情報におけるテキスト・マイニングをどう考えるか - ，理論と方法，第19巻，第2号，135-159。
- テキスト・マイニング研究会活用セミナー資料(下記HPからダウンロード可)
 - WordMinerの概要
 - 対応分析法の数理，分かち書き・辞書編集の方法，...
- テキスト・マイニング研究会ホームページ
<http://wordminer.comquest.co.jp/>
- 上記の論文やサイトから論文，参考文献，関連サイトへのリンク情報が得られる。

18

自由回答データの分かち書き処理後のイメージ

< 表 2 >

		分かち書きで得られる構成要素 (単語, 語句, キーワード...)							
「回答者・サンプル」 あるいは 「質的変数・属性」	1	$w_1^{(1)}$	$w_2^{(1)}$...	$w_j^{(1)}$...	$w_k^{(1)}$...	
	2	$w_1^{(2)}$	$w_2^{(2)}$...	$w_j^{(2)}$...	$w_k^{(2)}$...	
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	i	$w_1^{(i)}$	$w_2^{(i)}$...	$w_j^{(i)}$	$w_l^{(i)}$...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	n	$w_1^{(n)}$	$w_2^{(n)}$...	$w_j^{(n)}$...			

(回答者・サンプル) × (構成要素), (質的変数) × (構成要素) のイメージ図

注:

分かち書き処理 = 区切りのない日本語テキスト型データを分割すること

構成要素 (fragments) = 分かち書き処理で得た単語, 語句, キーワードなどの分析処理単位

21

例 1 : (サンプル) × (構成要素) の一例

< 表 3 >

サンプル	構成要素 (キーワードを用いたとき)
1	為, しらべ, 利用, 家族, 遊園地, 公園, 宿へ物屋, 情報収集, 調査
2	あまり, セキュリティ, 必要, ミーティング, 世間話, 仕事
3	新製品, スベック, 価格, お店
4	役所, 証明書発行, 受け取り
5	旅行, 計画, 観光地, チェック, お店, 情報収集
6	情報収集, 調査, メール, 座席予約, 航空機, 列車, オークション
7	地図検索, 鉄道, 乗り換え, 検索, その他, 時々, 必要, 情報検索
8	通信販売, 申し込み, 旅行, 情報収集
9	情報ツール
10	あまり, ぶつ, 店舗, 販売, 商品, 販売店, ショッピング, 建築図面作成用, CADデータ, ダウンロード
11	自分, 興味, 事柄, 容易, 公式, 専門家, 情報
12	日常生活, 中, 帰省時, 飛行機, 時刻表, 育児, 経験談, アドバイス, 仕事, 必要, 情報, 特定人物, 活動, 著書
13	情報収集
14	電話, 手紙, かわり
15	仕事上, 事, 出張, 際, ホテル, 情報, 等
16	パソコン, 周辺機器, 仕様, 価格, 懸賞, 応募, ドライバ, ダウンロード, ゲーム
17	掲示板, 一つ, 場所, みんな, 話
18	調べ物, ショッピング, オークション
19	情報, 収集, 自己, PR
20	ニュース, 天気, 行楽情報, 仕事, 情報
21	映画, 書籍, 情報入手, 求人検索, 単語, 等, 検索, メール
22	専門的, 事柄, 情報収集
23	メール, 一番, 仕事, 不明瞭, 確認, 美術館, 博物館, 映画, その他, 催し物, 情報収集, たまに, オークション, お食事, 電車, 時刻表, 経路
24	調べ物, ホームページ, サイト
25	友人, 知人, 連絡
26	百科事典
27	趣味, 人, 交流, 勉強, 場所, 交通機関, 時間
28	天気予報, 道路状況, 気象情報, 等, 行楽, 情報収集, 辞書, 新聞
29	自分, 知識, 情報, 時間, 辞書, 新聞, 地図, 最近, ネット, 使用, 事
	< 以下, 省略 >

(ここでは, 構成要素として
キーワードをあげた.)

データファイルの例

22

[(サンプル) × (構成要素)]の解析データ表(2元クロス表)

(WordMinerにより生成, 一部を切り出した例)

< 表 4 >

サンプルID	SEQ	行和	HP	いろいろ	いろいろな	いろんな	お店	その他	ときに	やり	やりとり	アーティ	イベン	インター	オーク	オンライ	ゲーム
	列和	3378	18	6	18	10	19	7	9	17	24	7	7	20	28	7	9
36	00000042	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
778	00000846	24	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
716	00000773	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	00000040	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
558	00000602	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	00000058	14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
509	00000548	14	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
759	00000824	14	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
98	00000107	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
139	00000154	13	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
310	00000338	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
401	00000432	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
407	00000438	13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
379	00000409	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
502	00000540	12	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
515	00000554	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
639	00000688	12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
51	00000059	11	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
52	00000060	11	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
89	00000098	11	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
157	00000174	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
303	00000330	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
484	00000520	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
564	00000608	11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
676	00000728	11	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
801	00000873	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	00000030	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

行側が回答・サンプル, 列側が構成要素群(構成要素変数), 非常に疎になる。

23

例2: 複数の項目を含むデータ表の例

< 表 5 >

サンプル	性別	年齢区分	性年齢区分	未婚	職業	構成要素(ここではキーワード)
1	男性	4_35才~39才	男性/4_35才~39才	既婚	営業職	為しらべ 利用 家族 遊園地 公園 賞へ物屋 情報収集 調査
2	男性	5_40才~44才	男性/5_40才~44才	既婚	研究開発職	あまり セキュリティ 必要 ミーティング 世間話 仕事
3	女性	5_40才~44才	女性/5_40才~44才	既婚	主婦専業	新製品 スベック 価格 お店
4	男性	5_40才~44才	男性/5_40才~44才	既婚	労務職	役所 証明書発行 受け取り
5	女性	2_25才~29才	女性/2_25才~29才	既婚	主婦専業	旅行 計画 観光地 チェック お店 情報収集
6	男性	5_40才~44才	男性/5_40才~44才	既婚	研究開発職	情報収集 調査 メール 座席予約 航空機 列車 オークション
7	女性	4_35才~39才	女性/4_35才~39才	既婚	無職・その他	地図検索 鉄道 乗り換え 検索 その他 時々 必要 情報検索
8	女性	2_25才~29才	女性/2_25才~29才	既婚	主婦専業	通信販売 申し込み 旅行 情報収集
9	男性	3_30才~34才	男性/3_30才~34才	既婚	自営業とその家族	情報ツール
10	男性	6_45才~49才	男性/6_45才~49才	既婚	専門職	あまり ぶつつ 店舗 販売 商品 販売店 ショッピング 建築図面作成
11	男性	3_30才~34才	男性/3_30才~34才	未婚	無職・その他	用 CADデータ ダウンロード
12	女性	3_30才~34才	女性/3_30才~34才	既婚	専門職	自分 興味 事柄 容易 公式 専門家 情報
13	男性	9_60才~64才	男性/9_60才~64才	既婚	無職・その他	日常生活 中 帰省時 飛行機 時刻表 育児 経験談 アドバイス 仕事
14	男性	8_55才~59才	男性/8_55才~59才	既婚	管理職	必要 情報 特定人物 活動 著書
15	男性	7_50才~54才	男性/7_50才~54才	既婚	販売・保安・サービス	情報収集 電話 手紙 かわり
16	男性	5_40才~44才	男性/5_40才~44才	既婚	営業職	仕事 上 出張 際 ホテル 情報 等
17	男性	5_40才~44才	男性/5_40才~44才	既婚	技能職	パソコン 周辺機器 仕様 価格 懸賞 応募 ドライバ ダウンロード
18	女性	3_30才~34才	女性/3_30才~34才	既婚	パート・アルバイト	ゲーム
19	女性	1_25才未満	女性/1_25才未満	未婚	自由業	掲示板 一つ 場所 みんな 話
20	女性	3_30才~34才	女性/3_30才~34才	既婚	技術職	調べ物 ショッピング オークション
						情報 収集 自己 PR
						ニュース 天気 行楽情報 仕事 情報

質的変数として「選択肢型質問」, 「属性」を含むデータ表の例。

24

[(年齢区分) × (構成要素)]の解析データ表(2元クロス表)
(WordMinerにより生成, 一部を切り出した例)

<表6 >

通番	列和	行和	1_25才未満	2_25才～29才	3_30才～34才	4_35才～39才	5_40才～44才	6_45才～49才	7_50才～54才	8_55才～59才	9_60才～64才
117	情報	270	39	42	41	36	45	26	19	7	8
121	情報収集	130	11	19	21	27	20	14	10	2	5
109	趣味	99	15	14	19	15	17	6	7	1	3
33	メール	95	12	13	11	19	17	4	10	5	4
66	読書	79	12	9	14	11	11	5	9	2	5
84	仕事	74	8	5	14	14	11	9	8	3	1
162	友人	60	7	6	9	12	9	9	4	1	2
145	夢	58	6	11	10	8	5	6	5	1	4
149	入手	58	7	8	4	10	12	4	6	2	3
166	旅行	56	1	8	9	10	9	3	7	2	2
55	活用	55	11	8	11	9	7	3	3	0	1
91	書	54	11	10	16	5	1	5	2	1	2
99	自分	49	9	6	15	3	6	3	5	1	0
18	ショッピング	48	3	7	12	8	3	7	3	3	1
150	買い物	46	3	11	9	9	7	2	2	0	1
170	連絡	46	4	5	6	6	9	5	3	3	3
24	ニュース	43	7	10	3	4	8	3	3	0	4
94	時	43	2	5	9	10	6	6	3	1	0
164	予約	42	2	8	6	7	6	7	1	0	5
135	調べ物	40	8	0	13	4	10	2	2	0	1
110	収集	36	3	6	4	6	6	8	2	0	0
31	ホームページ	34	6	6	5	6	3	1	6	0	1
128	人	34	11	10	6	4	2	0	1	0	0
93	事柄	33	6	3	7	4	4	2	2	4	1
165	利用	33	1	4	7	4	8	5	1	2	1
16	コミュニケーション	29	7	4	5	8	3	1	1	0	0
76	購入	29	3	5	2	4	5	4	3	0	3
13	アクション	28	3	5	5	6	5	0	2	0	2
113	商品	28	3	5	5	4	5	1	2	0	2
153	必要	28	3	4	3	5	5	5	1	1	1

ここで行側が構成要素群(構成要素変数), 列側が質的変数(性年齢区分), 頻度でソート.

25

対応分析法: 2元クロス表と成分スコアの関係

<図1 >

		項目 J					項目 I の成分スコア								
		1	2	...	j	...	n	1	2	...	k	...	k'	...	K
項目 I	1	f_{11}	f_{12}	...	f_{1j}	...	f_{1n}	z_{11}	z_{12}	...	z_{1k}	...	$z_{1k'}$...	z_{1K}
	2	f_{21}	f_{22}	...	f_{2j}	...	f_{2n}	z_{21}	z_{22}	...	z_{2k}	...	$z_{2k'}$...	z_{2K}

	i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{in}	z_{i1}	z_{i2}	...	z_{ik}	...	$z_{ik'}$...	z_{iK}
...	
m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{mn}	z_{m1}	z_{m2}	...	z_{mk}	...	$z_{mk'}$...	z_{mK}	
項目 J の成分スコア	1	z_{11}^*	z_{12}^*	...	z_{j1}^*	...	z_{n1}^*	↑ 行の項目 I の選択肢の成分スコア ← 列の項目 J の選択肢の成分スコア							
	2	z_{12}^*	z_{22}^*	...	z_{j2}^*	...	z_{n2}^*								
								
	k	z_{1k}^*	z_{2k}^*	...	z_{jk}^*	...	z_{nk}^*								
	k'	$z_{1k'}^*$	$z_{2k'}^*$...	$z_{jk'}^*$...	$z_{nk'}^*$								
K	z_{1K}^*	z_{2K}^*	...	z_{jK}^*	...	z_{nK}^*									

項目 I の成分スコアの行列

項目 I の成分スコアの行列

•項目 I と J との 2 組の成分スコアの行列が得られる。
• $K = \min\{m, n\} - 1$ となる。

数値例(人工データ): 次の2つの質問を取り上げる

質問I: 次に挙げるレストランのうち, あなたがお気に入りのレストランは次のどれですか?

- | | | | |
|---------|---------|----------|---------|
| 1. さとみ | 2. バッハ | 3. ムガール | 4. いりふね |
| 5. コルシカ | 6. クラーク | 7. ロゴスキー | 8. きくみ |
| 9. ラ・マレ | 10. かりや | | |

質問J: その選択時の評価基準は次の3つのうちのどれでしょうか?

- | | | |
|------|------|------------|
| 1. 味 | 2. 量 | 3. 工夫・サービス |
|------|------|------------|

典型的な質的変数(名義尺度)の例

27

数値例: 原データ表とクロス表

<表7>

項目 回答者	I (レストラン)	J (評価基準)
1	バッハ	味
2	ムガール	量
3	さとみ	量
4	ラ・マレ	工夫・サービス
5	きくみ	味
⋮	⋮	⋮
⋮	⋮	⋮
N	いりふね	量

N=1,284(回答者数)

原データ表
「(サンプル) × (項目, 変数)」の場合

<表8>

項目 I 項目 J	1. 味	2. 量	3. 工夫・サービス	行和
1. さとみ	46	7	42	95
2. バッハ	76	18	48	142
3. ムガール	44	16	49	109
4. いりふね	25	32	98	155
5. コルシカ	77	13	32	122
6. クラーク	14	54	34	102
7. ロゴスキー	35	42	48	125
8. きくみ	8	67	35	110
9. ラ・マレ	82	15	49	146
10. かりや	35	38	105	176
列 和	442	302	540	1,284

原データ表から作った2元クロス表

行と列のプロフィール(相対確率)の分布

評価項目	1. 味	2. 量	3. 工夫 サービス	行和
レストラン				
1. さとみ	0.484	0.074	0.442	1.000
2. パツハ	0.535	0.127	0.338	1.000
3. ムガール	0.404	0.147	0.450	1.000
4. いりふね	0.161	0.206	0.632	1.000
5. コルシカ	0.631	0.107	0.262	1.000
6. クラーク	0.137	0.529	0.333	1.000
7. ロゴスキー	0.280	0.336	0.384	1.000
8. きくみ	0.073	0.609	0.318	1.000
9. ラ・マレ	0.562	0.103	0.336	1.000
10. かりや	0.197	0.213	0.590	1.000
列の平均ベクトル	0.344	0.235	0.421	1.000

行のプロフィール

- プロフィールとは比率のパターンのこと、つまり相対確率のこと。
- は2次元空間内の10個の点の分布を、は9次元空間内の3個の点の分布を、それぞれ表わすと考える。
- よって は2次元空間内に布置できる。
- またこの例は三角図(重心座標系)で描ける。

<表9>

<図2>へ

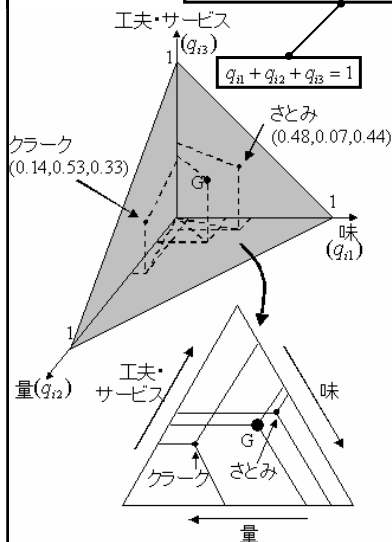
<表10>

列のプロフィール

評価項目	1. 味	2. 量	3. 工夫 サービス	行の平均 ベクトル
レストラン				
1. さとみ	0.104	0.023	0.078	0.078
2. パツハ	0.172	0.060	0.089	0.089
3. ムガール	0.100	0.053	0.091	0.091
4. いりふね	0.057	0.106	0.181	0.181
5. コルシカ	0.174	0.043	0.059	0.059
6. クラーク	0.032	0.179	0.063	0.063
7. ロゴスキー	0.079	0.139	0.089	0.089
8. きくみ	0.018	0.222	0.065	0.065
9. ラ・マレ	0.186	0.050	0.091	0.091
10. かりや	0.079	0.126	0.194	0.194
列和	1.000	1.000	1.000	1.000

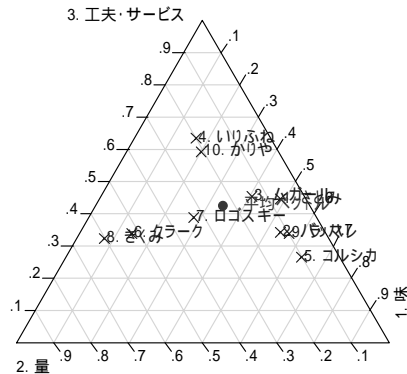
行のプロフィールの意味を図で確認

表9で行和 = 1の制約があるので
2次元平面内に布置



<図2>

三角図



重心座標系の布置図
(JMPの三角図機能を利用)

「レストランと評価基準」の例で成分スコアを確認

2項目への成分スコア

<表11>

項目と選択肢		成分スコア	
		第1成分スコア	第2成分スコア
成分		z_{i1}	z_{i2}
項目 /	さとみ	0.40067	-0.09077
	パツハ	0.39656	0.12200
	ムガール	0.19686	-0.08210
	いりふね	-0.20169	-0.40820
	コルシカ	0.54972	0.25857
	クラーク	-0.66717	0.25584
	ロゴスキー	-0.21980	0.10024
	きくみ	-0.85898	0.30915
	ラ・マレ	0.46355	0.11909
	かりや	-0.16472	-0.32610
成分		z_{j1}^*	z_{j2}^*
項目 √	味量	0.52347	0.17643
	工夫・サービス	-0.65787	0.25247

固有値と寄与率

<表12>

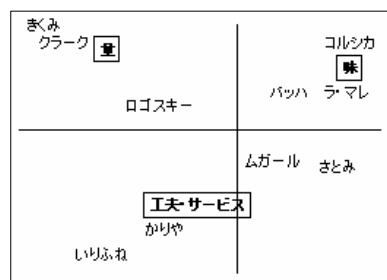
主成分 k	固有値 λ_k	寄与率(%)
1	0.19766	76.71
2	0.06002	23.29

- <図1>の成分スコアに相当の表.
- 固有値の数は $K = \min\{m, n\} - 1 = 2$ となる.
- 2成分に対する成分スコアが算出される.
- レストラン(行), 評価基準(列)のそれぞれに成分スコアがある.
- 成分スコアの布置図, 同時布置図が描ける.

31

対応分析で得られる布置図と三角図の比較

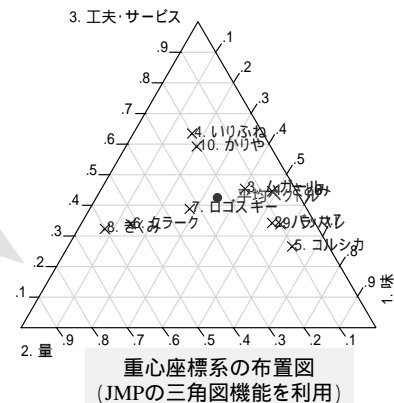
成分スコアの同時布置図



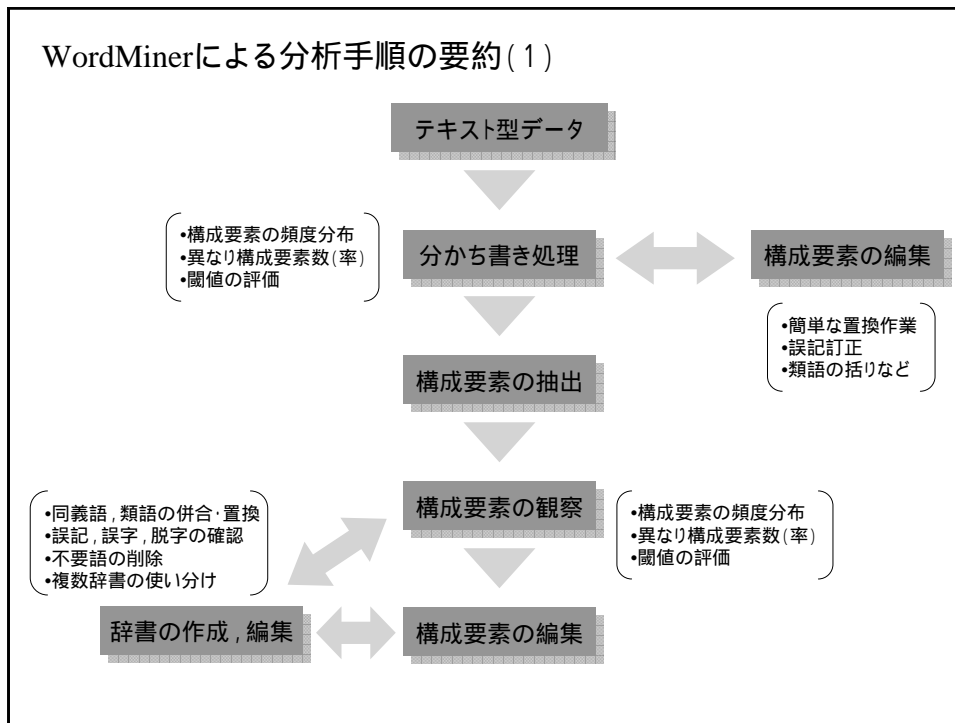
もとの2次元平面上的10のレストランの三角図布置が上の成分スコアとして再現される(2成分スコアとして)

<図3>

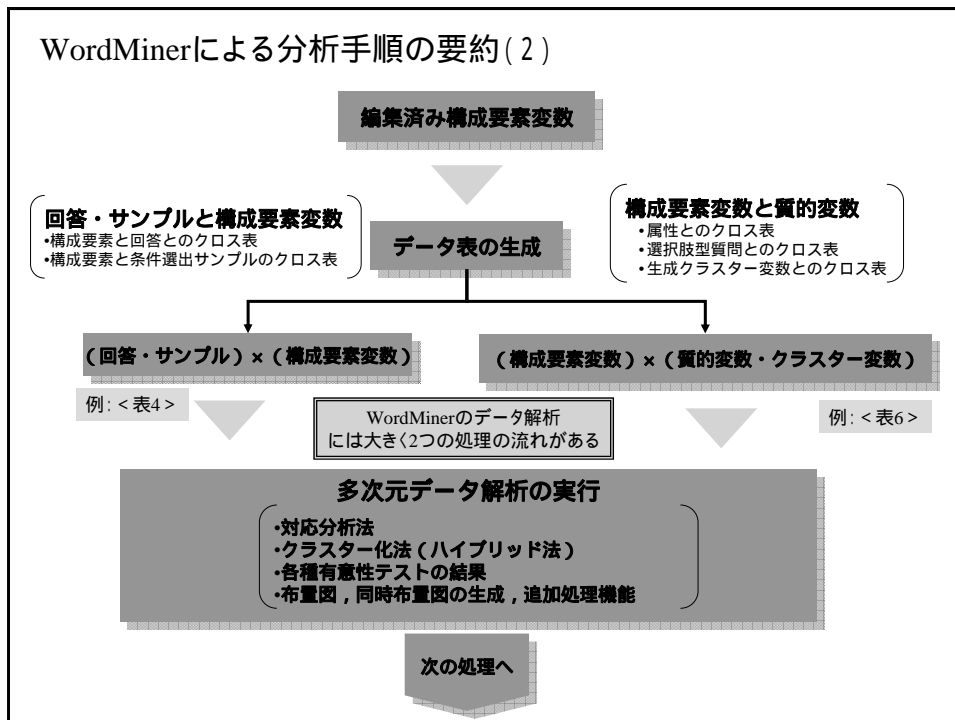
三角図



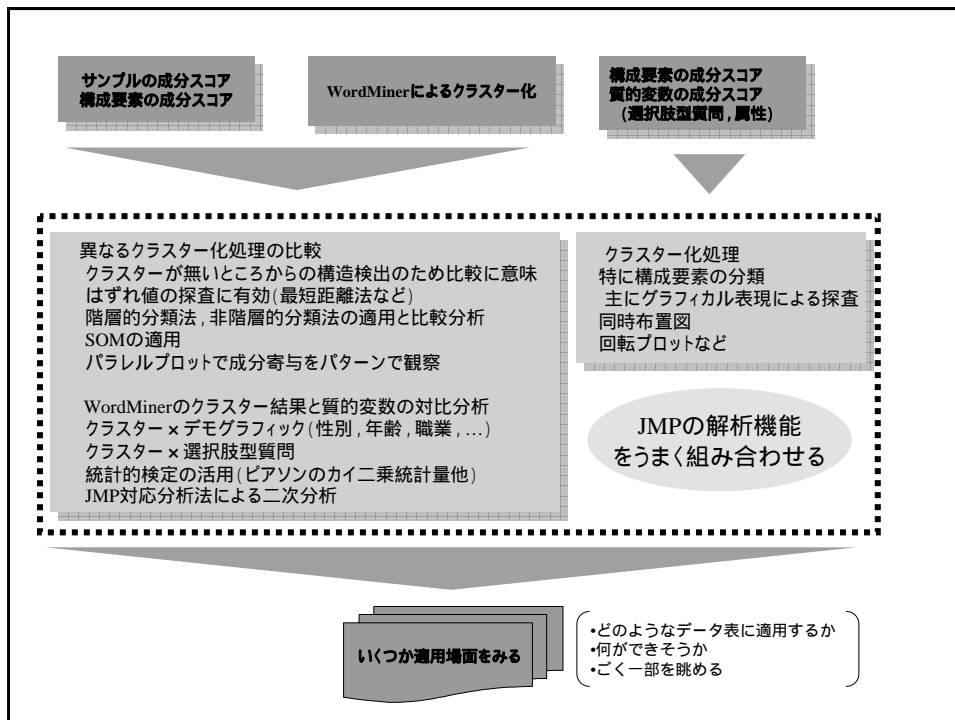
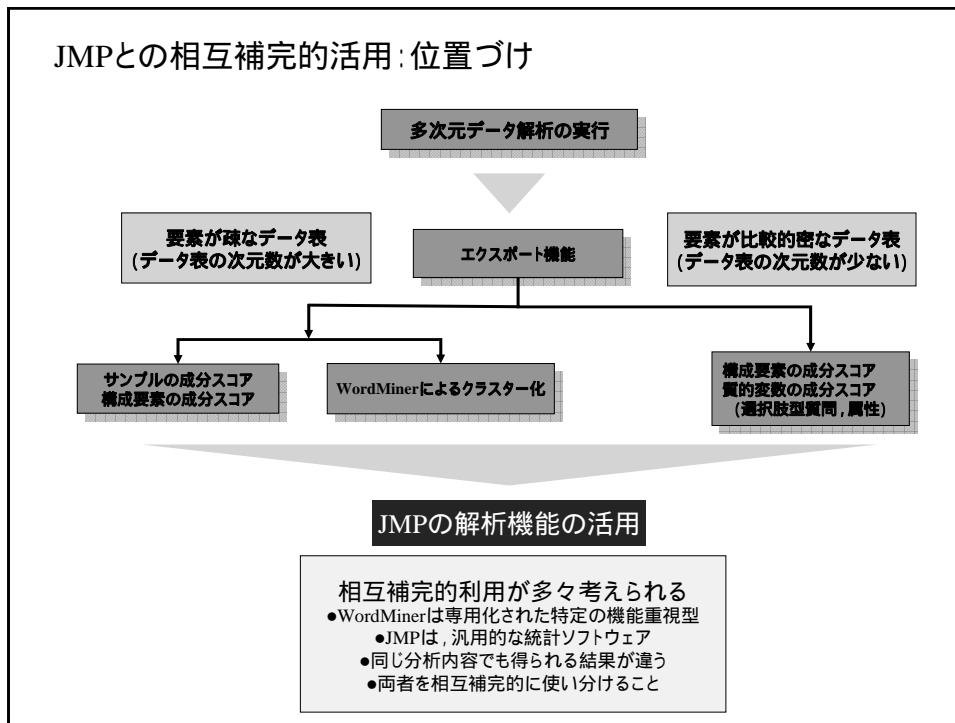
WordMinerによる分析手順の要約(1)



WordMinerによる分析手順の要約(2)



JMPとの相互補完的活用:位置づけ



事例分析に用いる質問 (Web調査)

問3. 次に、あなたと「インターネット」とのかかわりについてお伺いします。

Q3 - 1. あなたご自身にとって「インターネット」は、どのようなことから活用できると思いますか。どんなことでも結構ですので、以下になるべく具体的にご記入ください。

Q3 - 2. では、一般的に「インターネット」は、どのようなことから活用できると思いますか。なるべく、他にはないような活用法を、どんなことでも結構ですので、以下になるべく具体的にご記入ください。

問3. あなたと「インターネット」とのかかわりについてお伺いします。

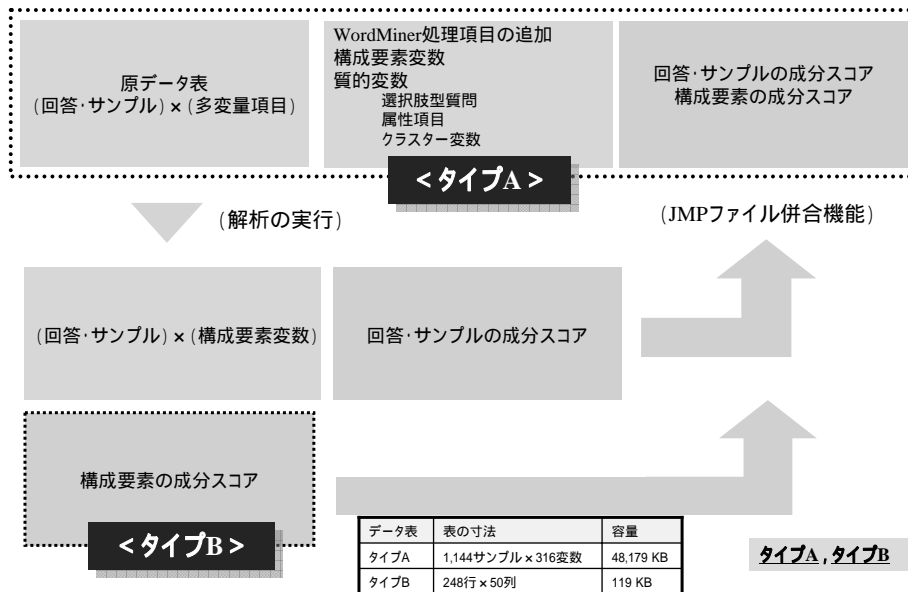
1. あなたご自身にとって「インターネット」は、どのようなことから活用できると思いますか。どんなことでも結構ですので、以下になるべく具体的にご記入ください。

2. では、一般的に「インターネット」は、どのようなことから活用できると思いますか。これもどんなことでも結構ですので、以下になるべく具体的にご記入ください。

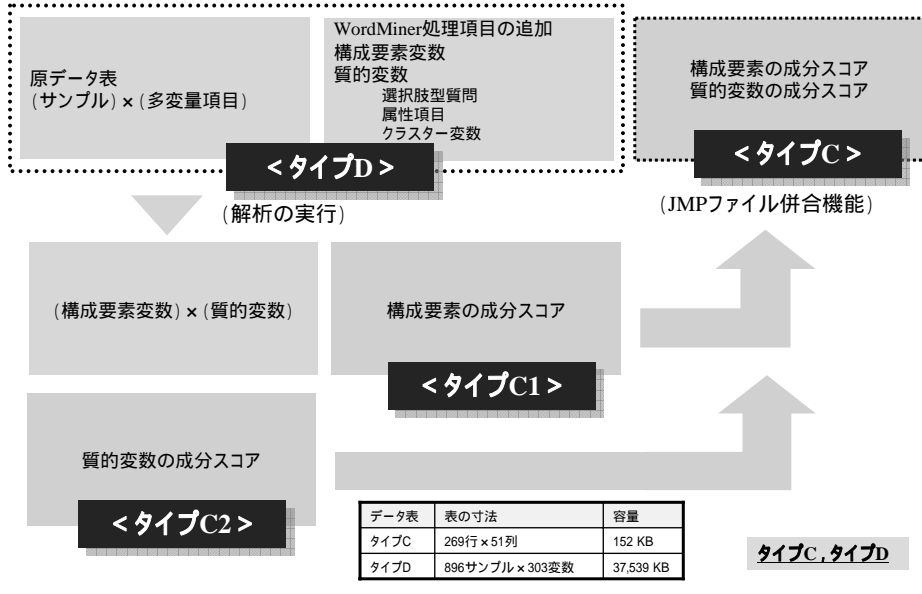
(電子調査票の例)

37

JMP用ファイル生成： (回答・サンプル) × (構成要素変数) から



JMP用ファイル生成:
(構成要素変数) × (質的変数・クラスター変数) から



WordMinerエクスポート・データとJMPの併用(1)

(回答・サンプル) × (構成要素変数)

<表13>

用いるデータ表	分析内容, 目標	WordMinerの対応	JMPの対応
<タイプA> <タイプB>	-1: 回答・サンプル成分スコアのクラスター化[サンプル・クラスター] -2: 構成要素成分スコアのクラスター化[構成要素クラスター]	ハイブリッド法 ●専用化されている ●階層的・非階層的分類法の併用 ●カイニ乗距離の適用 ●大量データ処理が可能 ●はずれ値手当を重視 ●はずれ値の一時除去と再配置	一般的な分類法 ●階層的分類法(ワード法, 最短・最長距離法他) ●樹形図(デンドログラム) ●非階層的分類法(k-平均法, SOM他) ●パラレルプロット ●成分スコアの布置図(散布図), ヒストグラム, 箱髭図など ●クロス表の利用(分類間の比較, 一致度)
<タイプA>	-3: 回答・サンプル成分スコアと構成要素成分スコアの比較	比較分析 ●布置図・同時布置図(伸縮機能) ●はずれ値探査 ●はずれ値の一時除去と再配置	比較分析 ●布置図・同時布置図(尺度調整機能) ●回転プロット(3次元布置図) ●はずれ値探査
<タイプA>	-3: サンプル・クラスターと構成要素クラスターの布置図[生成したクラスター変数, 成分スコアの利用]	比較分析 布置図・同時布置図(伸縮)	比較分析 ●布置図・同時布置図(尺度調整機能) ●回転プロット
<タイプA>	-4: (サンプル・クラスター) × (質的変数) [-1と選択肢型質問, 属性などの比較]	対応分析法 ●独自の有意性テストによる構成要素の探査	対応分析法 ●カイニ乗統計量検定他による探査

WordMinerエクスポート・データとJMPの併用(2)

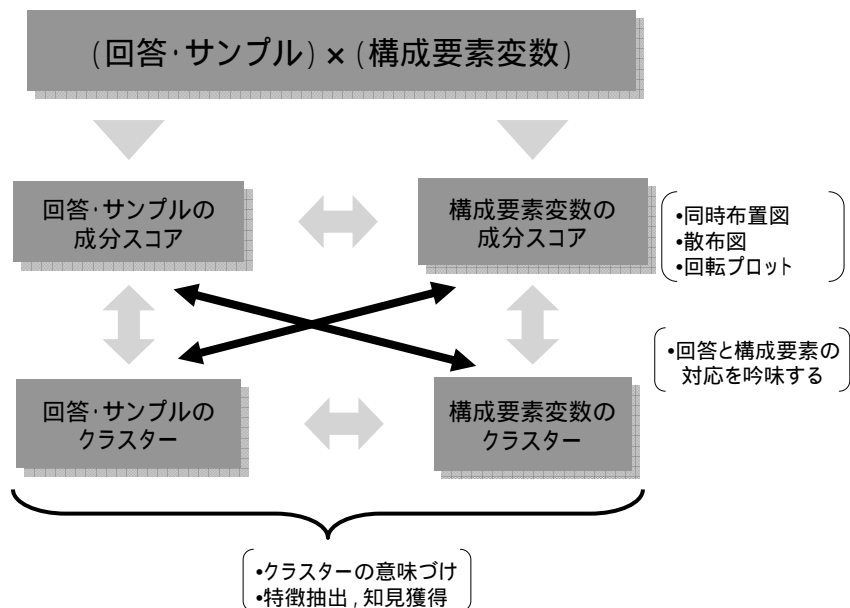
(構成要素変数) × (質的変数: 選択肢型質問, 属性他)

< 表14 >

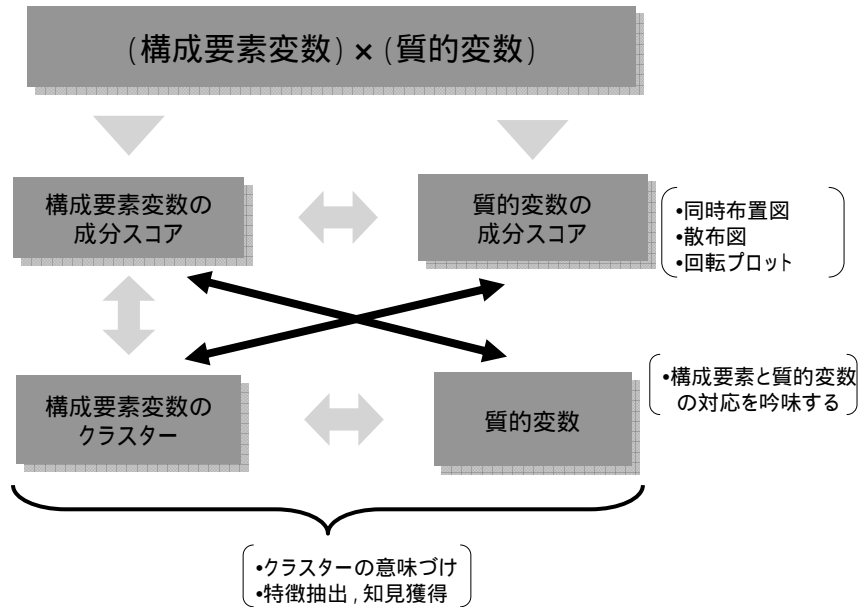
用いるデータ表	分析内容, 目標	WordMinerの対応	JMPの対応
<タイプC1>	-1: 構成要素変数成分スコア	ハイブリッド法 ●専用化されている ●階層的・非階層的の分類法の併用 ●カイ二乗距離の適用 ●はずれ値手当を重視	一般的な分類法 ●階層的の分類法(ワード法, 最短・最長距離法他) ●樹形図(デンドログラム) ●非階層的の分類法(k-平均法, SOM他) ●成分スコアの布置図(散布図), ヒストグラムなど ●クロス表の利用(分類間の比較, 一致度)
<タイプC2>	-2: 質的変数成分スコア 注: 通常は, データ表の次元数が小さいので分類操作は不要, 布置図で十分		
<タイプC>	-3: 構成要素変数成分スコア + 質的変数成分スコア[併合ファイル]	比較分析 布置図・同時布置図(伸縮)	比較分析 ●布置図・同時布置図(尺度調整) ●回転プロット
<タイプA> <タイプD>	-4: 質的変数に対する一般的な統計解析もあり得る	該当機能はない	様々な質的変数・質的データの分析機能

量的データとなった成分スコアを扱うので適用する手法はほぼ同種となる。結果解釈の過程で意味が異なってくる。

分析内容の要約(1)



分析内容の要約(2)



ここでみるJMP分析課題(要約)

- 分析課題1: 成分スコアの観察
 - ・ -3: 回答・サンプルと構成要素, それぞれの成分スコアの観察
 - ・ 布置図, 同時布置図 散布図, 回転プロット, 多変量連関図
 - 分析課題2: クラスター化機能を用いる例
 - ・ -1: 回答・サンプル成分スコアのクラスター化
 - ・ -2: 構成要素成分スコアのクラスター化
 - ・ -1, -2: 異なる分類手法による分類結果の比較, 樹形図(デンドログラム)の比較
 - 分析課題3: クラスター化情報と質的変数の関係
 - ・ -4: (サンプル・クラスター) × (質的変数)
 - 分析課題4: 構成要素変数, 質的変数の成分スコアの同時観察
 - ・ -3: 布置図, 同時布置図により, 変数間の関連性の探査
 - ・ 構成要素変数に有意な質的変数の探査
- 実は大半の課題はWordMinerでも解析が可能だが, JMPのインタラクティブ性を活かした分析にも十分に意味がある.

分析課題1:成分スコアの観察

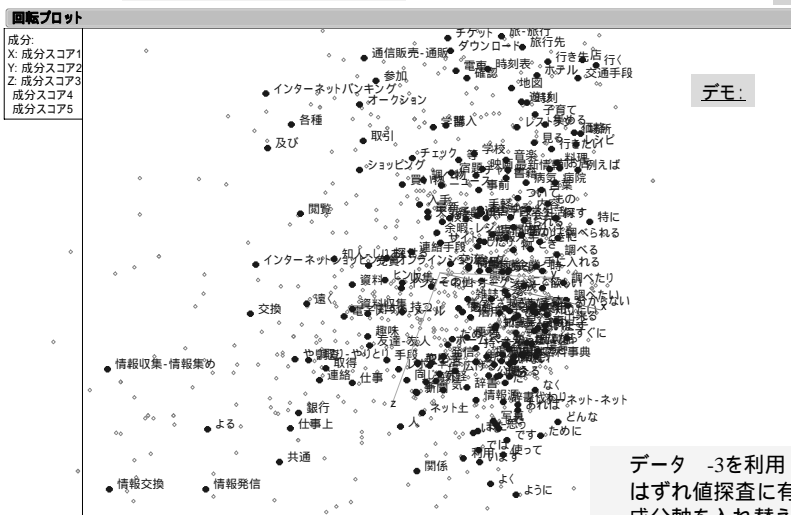
- 回答・サンプルと構成要素,それぞれの成分スコアの観察
- 散布図,回転プロット,多変量連関図を利用
- これらを成分スコアの布置図,同時布置図として用いる
- はずれ値の探査と特定化,除外やマーキングが有効
- 一般に,データ数(サンプル数,構成要素数いずれも)多数となるので,図が煩雑になる
- 理想的なクラスター化構造も,まず一般的にはあり得ない
- 成分スコアの“分布”の特徴を知って,クラスター化を併用し構造を強調化する(課題1で行ったように)

45

回答・サンプルと構成要素の各成分スコアの観察(1)

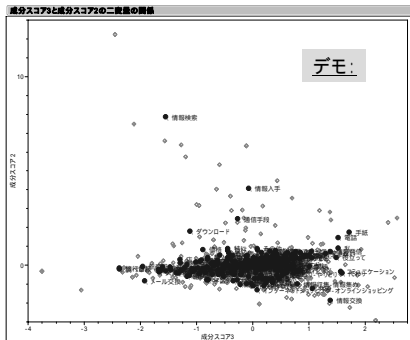
回転プロット(同時布置図)

< 図8 >



データ -3を利用
はずれ値探査に有効
成分軸を入れ替えながら観察
WordMiner出力のスコア検定値
他と比較

回答・サンプルと構成要素の各成分スコアの観察(2)

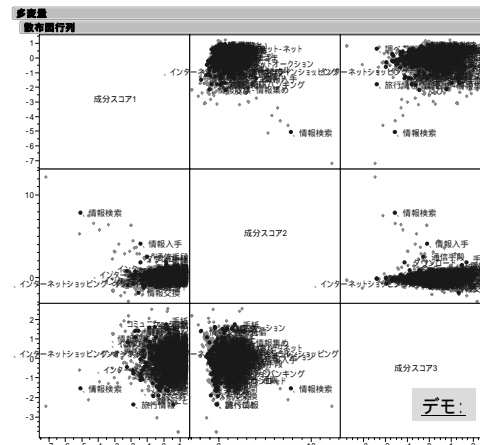


同時布置図(散布図)

同じく、データ-3を利用
 多くの場合成分スコアはこのような“雲状”(nuage)の分布となる
 よってはずれ値探査が重要
 同時にここでも分類手法による類型化が重要

< 図9 >

多変量連関図(同時布置図)



分析課題2: クラスター化機能を用いる例

(回答・サンプル成分スコアのクラスター化)

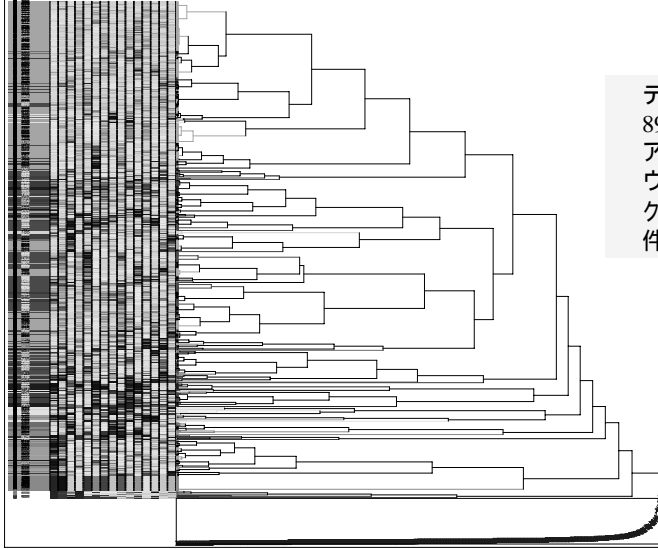
- 回答・サンプル成分スコアのクラスター化処理
 - 階層的分類法(ウォード法, 最短・最長距離法他)
 - 樹形図(デンドログラム)の観察
 - 非階層的分類法(k-平均法, SOM他)
 - 成分スコアの布置図(散布図)
 - ヒストグラム, 箱髭図によるデータチェック(はずれ値など)
 - パラレルプロットによるパターンの観察
 - クラスター化間の比較: クロス表の利用(分類結果間の比較, 一致度)
- 分類手法の特性を活かす(JMPは操作性がよい)
 - そもそも, 過去経験の多くの事例では, いわゆる「房状のクラスター」などは存在しない
 - よって, 意図的にクラスター化を行う, つまり「クラスターを生成」する
 - よって複数の手法の適用・比較が必要となることがある
 - 最短距離法(単連結法)によりはずれ値を検出(クラスター化の利点の一つ)
 - ウォード法でまとまりのよいクラスターを生成(クラスター化評価基準の問題)
 - k-平均法のオプションを使い分けて「最適化の様子」を観察し, はずれ値の影響を観察, ウォード法とも比較
 - 分類手法間の分類結果の相互比較(対応分析法などで二次分析を行う)

回答・サンプル成分スコアのクラスター化(ワード法)

階層型クラスター分析

手法 = Ward法

樹形図



< 図4 >

データ -1を利用.
896サンプルの成分スコア, 15成分を使って分類.
ワード法, 15群を生成.
クラスター数他の設定条件は試行錯誤で決定.

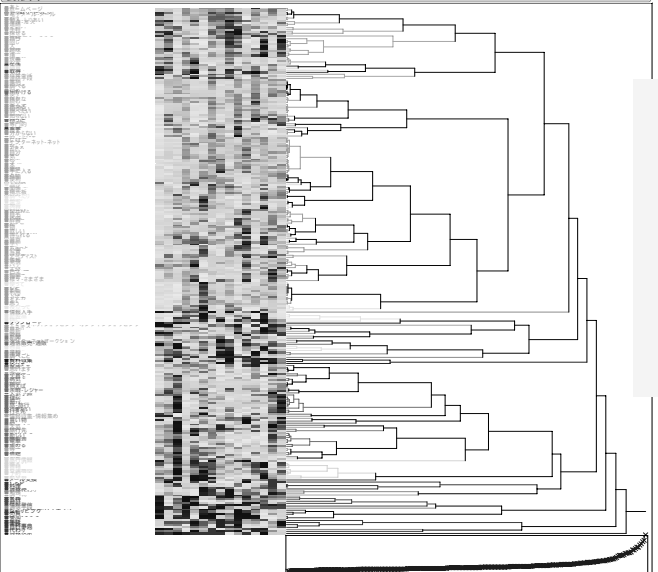
デモ:ワード法

構成要素成分スコアのクラスター化(ワード法)

階層型クラスター分析

手法 = Ward法

樹形図



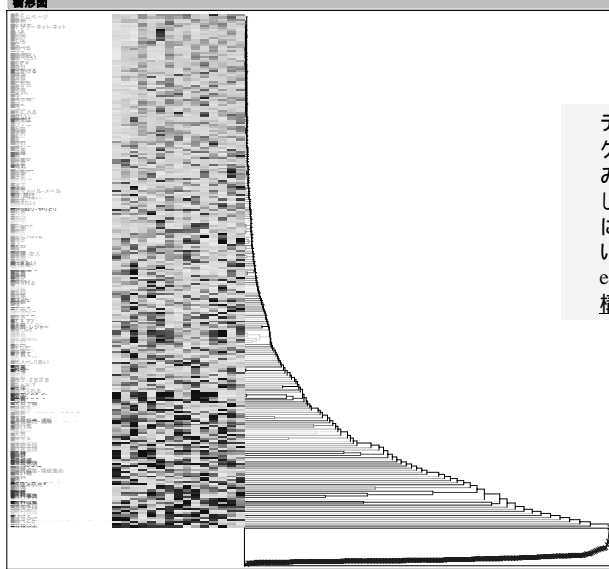
< 図5 >

データ -2を利用
ある種のクラスター構造があるように見える(そのようなクラスターを生成した).
はずれ値的な構成要素が一部に集まった.
ラベルとした構成要素(語句)の可読性が良くないのが欠点.
構成要素数は248語, 成分スコア数は15成分.

デモ:ワード法

構成要素成分スコアのクラスター化(最短距離法)

階層型クラスター分析
手法 = 最短距離法



< 図6 >

データ -2を利用
クラスター構造が無いように
みえる。
しかし、カラーマップでは下方
に特徴的なパターンが見える。
いわゆる連鎖現象(chaining
effect)があり、はずれ値的な
構成要素の検出に役立つ。

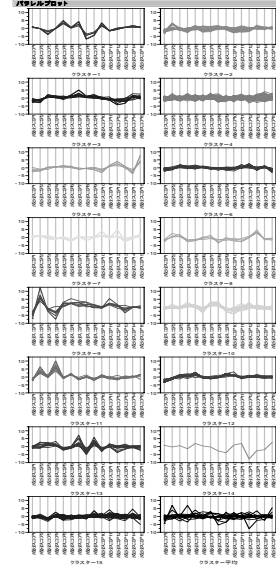
デモ:最短距離法

k-平均法後のパラレルプロット表示

階層型クラスター分析

構成要素

パラレルプロット



▼ クラスター要約

ステップ	基準		
10	0		
クラスター	度数	最大距離	事前距離
1	7	9.11359408	5.48876283
2	25	11.1415627	5.26108036
3	12	9.50622018	5.3948926
4	40	9.19148363	5.17239979
5	5	7.91219445	5.54498217
6	14	8.24204148	5.36731089
7	4	8.22472813	5.58105444
8	4	7.35851267	5.58105444
9	7	15.2029389	5.48876283
10	12	9.13501545	5.3948926
11	8	8.92001242	5.46591174
12	675	9.683839	4.59523945
13	9	12.8104417	5.44552476
14	1	5.77454702	5.77454702
15	18	10.7709805	5.32172735

< 図7 >

クラスター・サイズ
が不揃い、特定なク
ラスタに集中してい
ることに注意

データ -1を利用

回答・サンプル成分スコアのk-平均法で15群に分類

そのパラレルプロット図で成分スコアのパターンを観察

成分スコアの特徴がよく現れた

この後、二次分析として、WordMinerで出力した成分スコアの

検定値を利用して意味解釈

デモ:k-平均法

分析課題3：(サンプル・クラスター) × (質的変数)

- WordMiner生成クラスター(クラスター変数)と質的変数の関連探査
- 生成クラスターにどの質的変数が有意に関係するか
- WordMinerでは独自の有意性テストを用いる
 - 構成要素(単語, 語句)の出現頻度によるテスト(構成要素の特徴評価)
 - カイ二乗距離によるテスト(グループのまとまりの良さを評価)
- JMPを用いて別の評価を行えるか
 - 生成クラスターに複数の質的変数のどれが有意で関係ありそうかを探査
 - 用いた事例データで以下の4種の組合せを比較分析
 - 「クラスター変数」×「性年齢区分」 高度に有意
 - 「クラスター変数」×「職業」 高度に有意
 - 「クラスター変数」×「現在所有のメールアドレス数は？」 有意でない
 - 「クラスター変数」×「インターネットを利用した調査への参加頻度は？」 有意でない
 - この例では「職業」が「性別」と相関があるので、これによるブレイクダウンを行った分析も必要なことを示唆している。

53

サンプル・クラスターと職業区分の対比



クラスターと職業区分の比較:特徴的な構成要素の探査(1)

< 表15 >

関連職業	営業職	販売・保安・サービス専門職	無職・その他	技術職				
順位	サンプルクラスター01 サンプル数:78 異なり構成要素数:175	サンプルクラスター08 サンプル数:52 異なり構成要素数:104	サンプルクラスター12 サンプル数:18 異なり構成要素数:36	サンプルクラスター06 サンプル数:75 異なり構成要素数:94	サンプルクラスター10 サンプル数:20 異なり構成要素数:48	サンプルクラスター13 サンプル数:65 異なり構成要素数:49	サンプルクラスター14 サンプル数:36 異なり構成要素数:73	サンプルクラスター16 サンプル数:5 異なり構成要素数:4
上位1	して	ニュース	調べごと	入手	情報交換	情報収集-情報集め	予約	代わり
上位2	いる	インターネットショッピング	もの	情報	閲覧	仕事上	購入	情報交換
上位3	います	ゲーム	懸賞	収集	代わり	銀行	チケット	雑誌
上位4	利用	見る	簡単	仕事	情報収集-情報集め	趣味	航空券	新聞
上位5	時間	天気	行く	必要	関係する	電子メール-メール	メール交換	
上位6	です	旅行先		交換	買い物	通信販売-通販	検索	
上位7	また	特に		連絡			交通機関	
上位8	使って	等		取得			ホテル	
上位9	私	ダウンロード		余暇-レジャー			会社	
上位10	では	見たり		子育て			書籍	
上位11	役立って	チャット		取引			懸賞	
上位12	活用	テレビ		知人-知りあい			地図	
上位13	ために	天気予報		友達-友人			商品	
上位14	新製品	電車		旅行情報			確認	
上位15	インターネット	最新		新しい			内容	
上位16	よく	時刻表		趣味				
上位17	その	価格		子育て				
上位18	今	地図		調査				
上位19	調べたり	イベント		調書				
上位20	写真	インターネットオークション		調べ				
上位21	時	インターネットオークション		インターネットバンキング				
上位22	どんな	遠く						
上位23	この	事前						
上位24	開	探せる						
上位25	強	映画						
上位26	ない	情報収集-情報集め						
上位27	には	調べられる						
上位28	できない							
上位29	あれは							
上位30	あれは							
上位31	価格							
上位32	したり							
上位33	どこ							
上位34	色々							

クラスター情報(クラスター変数)と職業区分の関連づけを試みる
 クラスターによって、特徴的な語句と職種が対応することがみえる
 職業に性別との関連が埋め込まれている傾向がある
 「職業×性別」のクロス表分析も必要

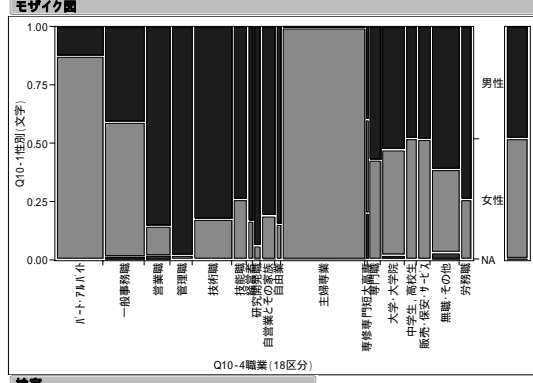
クラスターと職業区分の比較:特徴的な構成要素の探査(2)

< 表15(つづき) >

技能職/ IT・アルバイト/一般事務職/主婦専業	自営業とその家族/管理職	自由業/研究開発職/経営者	専修専門/大学・大学院/中学生・高校生
サンプルクラスター02 サンプル数:99 異なり構成要素数:209	サンプルクラスター07 サンプル数:144 異なり構成要素数:209	サンプルクラスター09 サンプル数:35 異なり構成要素数:64	サンプルクラスター15 サンプル数:10 異なり構成要素数:19
サンプルクラスター03 サンプル数:66 異なり構成要素数:143	サンプルクラスター04 サンプル数:139 異なり構成要素数:24	サンプルクラスター17 サンプル数:8 異なり構成要素数:8	サンプルクラスター18 サンプル数:16 異なり構成要素数:54
サンプルクラスター05 サンプル数:144 異なり構成要素数:209	サンプルクラスター06 サンプル数:144 異なり構成要素数:209	サンプルクラスター11 サンプル数:30 異なり構成要素数:54	サンプルクラスター12 サンプル数:30 異なり構成要素数:54
コミュニティ	情報	旅行	レンタル
通べる	人	行く	よる
知りた	同じ	好きな	オークション
ずくに	速く	できる	情報発信
について	友達-友人	する	ショッピング
調べたい	趣味	手に入る	生活
分からない	持つ	行く	情報源
知る	手	居ながら	表裏
調べられる	気軽	調べたり	電子メール-メール
出来る	欲しい	行き先	手紙
興味	探せる	時間	電話
事柄	自分	病気	仕事上
ある	自分	下調べ	早
言葉	た	家族	連絡手段
知らない	発信	確認	最新
とき	商品	前	電子メール-メール
持った	ときに	例えば	通信手段
広げる	物	余暇-レジャー	画
専門的	自宅	電車	買い物
探す	や取り-やりと	買い物	ない
自分	ホームページ	ゲームページ	
できる	便利	及び	
とに	本	イベント	
場	いる早	チェック	
	手に入る	病院	
	色々な	家	
	音楽	地図	
	学校	図	
	雑誌	際	
	なる		
	ある		
	インターネット-ネット		
	最新情報		
	交流		

併せて「職業×性別」の関係を確認

Q10-4職業(18区分)とQ10-1性別(文字)の分割表に対する分析



< 図11 >

検定

要因	自由度	(-1)*対数尤度	R2乗(U)
モデル	34	254.87129	0.3940
緑差	852	392.00296	
全体(修正済み)	886	646.87425	
N	888		

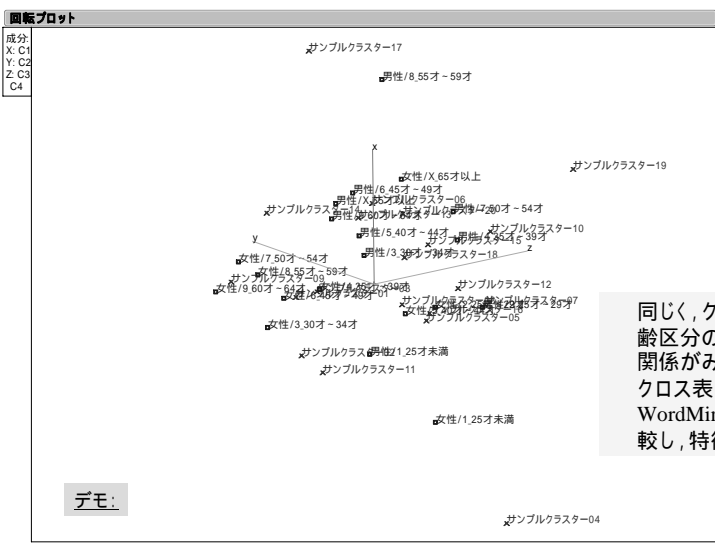
検定	カイ2乗	p値(Prob>ChiSq)
尤度比	509.743	<.0001*
Pearson	441.650	<.0001*

警告:
セルのうち20%の期待度数が5未満です。カイ2乗に問題がある可能性があります。

職業と性別間に明らかに関係がある、
よってブレイクダウンし、再分析を行う必要あり。
二次分析として対応分析法を行う(有意)。

デモ:

同じように、サンプル・クラスターと性年齢区分の対比

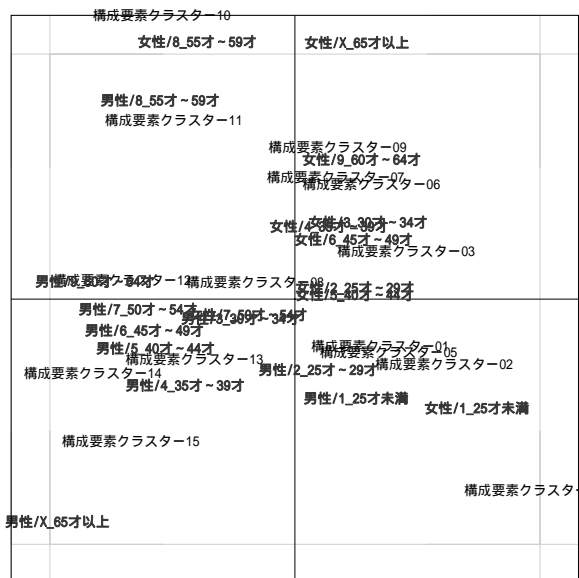


< 図12 >

同じく、クラスター変数と性年齢区分の関連づけを試みる。
関係が見える(20群に分類)。
クロス表, 検定で有意となる。
WordMinerの有意テストと比較し, 特徴抽出を行う。

デモ:

参考: WordMinerによる同時布置図の例



< 図14 >

ここでは、構成要素変数のクラスター化で得た構成要素クラスターと性年齢区分の関係を探索。視認性がよくなる。ラベル名のオン・オフ機能。拡大・縮小機能(中心部、周辺部)。他の変数のオーバーレイ機能(重ね描き)、など。

(WordMinerへ)

(構成要素変数) × (性年齢区分) とクラスターの対応 (1)

対応するクラスター	クラスター-C1	クラスター-C2	クラスター-C13	クラスター-C13	クラスター-C14
順位	男性/1_25才未満 サンプル数: 58 変数構成要素数	男性/2_25才~29才 サンプル数: 42 変数構成要素数	男性/3_30才~34才 サンプル数: 58 変数構成要素数	男性/4_35才~39才 サンプル数: 59 変数構成要素数	男性/5_40才~44才 サンプル数: 64 変数構成要素数
上位1	事-こと	する	公開	情報検索	よる
上位2	人	交流	売買	インターネットショッピング	仕事上
上位3	サイト	人	情報源	情報収集-情報集め	調査
上位4	速く	買い物	情報収集-情報集め	コミュニケーション	情報収集-情報集め
上位5	ように	字音	新聞	仕事	特に
上位6	できる	連絡手段	なった	した	確認
上位7	掲示板	ときに	ちょっと	売買	通信販売-通販
上位8	レポート	同じ	インターネットバンキング	入手	では
上位9	簡単	なる	関係	交通機関	あらゆる
上位10	には	使って	知らない	等	仕事
上位12	ネット上	手に入る	地図	趣味	ホテル
上位13	ある	利用	連絡手段	及び	利用
上位14	いち早く	取得	やり	出かける	航空券
上位15	取得	連絡手段	調べ物	情報発信	連絡
上位16	音楽		仕事	同じ	です
上位17					
上位18					
下位4		検索			
下位3		旅行-旅行	友達-友人		ショッピング
下位2		いる	入手	旅行-旅行	ホームページ
下位1					
対応するクラスター	クラスター-C14	クラスター-C14	クラスター-C11	クラスター-C12	クラスター-C15
順位	男性/6_45才~49才 サンプル数: 38 変数構成要素数: 93	男性/7_50才~54才 サンプル数: 31 変数構成要素数: 99	男性/8_55才~59才 サンプル数: 18 変数構成要素数: 51	男性/9_60才~64才 サンプル数: 21 変数構成要素数: 65	男性/10_65才以上 サンプル数: 14 変数構成要素数: 45
上位1	収集	仕事上	手紙	予約	よる
上位2	情報収集-情報集め	各種	調査	航空券	旅行先
上位3	仕事	発信	情報入手	購入	情報交換
上位4	意見	使える	メール交換	参加	入手
上位5	航空券	新しい	余暇-レジャー	会社	地図
上位6	等	連絡手段	知らない	書籍	ない
上位7	閲覧	百科事典	電話	聞いた	インターネット-ネット
上位8	活用	情報発信	調べたり	ホテル	情報
上位9	予約	情報入手	事務	チケット	その他
上位10	います	ポイント	確認	確認	公開
上位11	仕事上	情報	価格	確認	取引
上位12	色々な	代わり	見たり	その	早く
上位13	必要	に	情報収集-情報集め	仕事	会社
上位14	ショッピング				交通機関
上位15	購入				航空券
上位16	には				取手
上位17	手に入る				手に入る
上位18	内容				内容
下位4					
下位3		できる			
下位2		調べる			
下位1	事-こと	事-こと			事-こと

< 表16 > (男性)

< 図12 > はWordMinerの(回答・サンプル) × (構成要素変数) から得た回答・サンプル成分スコアのクラスター化で得た情報。ここでは、WordMinerによる(構成要素変数) × (性年齢区分)の分析結果を観察。有意性テスト結果一覧ここで、性年齢区分を特徴づける構成要素(語句群)が分かる。さらにJMPを用いることで、両者の関係を推論する情報が得られる。< 図14 >のクラスターと性年齢区分の対応

(構成要素変数) × (性年齢区分) とクラスターの対応 (2)

対応するクラスター	クラスター-C5	クラスター-C8	クラスター-C10	クラスター-C9	クラスター-C15
階位	女性/6.45才-49才 サンプル数: 31 異なり構成要素数: 92	女性/7.50才-54才 サンプル数: 29 異なり構成要素数: 89	女性/8.55才-59才 サンプル数: 11 異なり構成要素数: 57	女性/9.60才-64才 サンプル数: 11 異なり構成要素数: 40	女性/10.65才以上 サンプル数: 6 異なり構成要素数: 20
上位 1	友達・友人	とる	病院	病氣	旅・旅行
上位 2	使って	病状	私	知る	知人・知りあい
上位 3	利用	出かける	私	ために	買い物
上位 4	問する	電報	連絡	行き先	交換
上位 5	いれる	広げる	電話	ニュース	使用
上位 6	連絡	電子メール・メール	情報入手	交通機関	日常生活
上位 7	日常生活	時刻表	時間	病院	物
上位 8	気軽	育児	生活	よく	使う
上位 9	メール交換	知識	事情	時刻	お店
上位 10	子供	色々	いるいる	どこ	時刻表
上位 11	情報	役立って	メール交換	料理	料理
上位 12	会報・レジャー	読む		情報入手	興味
上位 13	必要			興味	情報交換
上位 14	分からない				
上位 15	ため・為				
上位 16	時				
上位 17	レシビ				
下位 3	できる				
下位 2	物	もの			
下位 1	様系	趣味			

<表16> (女性)
下段表から上段表へ移る



対応するクラスター	クラスター-C2, C4	クラスター-C3	クラスター-C6	クラスター-C7	クラスター-C5
階位	女性/12.25才未満 サンプル数: 65 異なり構成要素数: 101	女性/12.25才-28才 サンプル数: 71 異なり構成要素数: 80	女性/13.30才-34才 サンプル数: 60 異なり構成要素数: 66	女性/14.35才-38才 サンプル数: 66 異なり構成要素数: 66	女性/15.40才-44才 サンプル数: 60 異なり構成要素数: 66
上位 1	情報収集	すべて	子育て	勉強	心切り
上位 2	レポート	書籍	いる	その	勉強
上位 3	学校	好きな	ついで	今	勉強
上位 4	友達・友人	洋館	インターネット・オアシ・どんな	雑誌	雑誌
上位 5	読める	読める	下調べ	電報	多変
上位 6	興味	ある	知らない	時刻	時刻
上位 7	知る	地味	知りた	読たり	時刻
上位 8	いませ	向	関係	場所	場所
上位 9	集める	ホームページ	興味	場所	場所
上位 10	聞きたい	アーティスト	どこ	手に入る	手に入る
上位 11	だ	探せる	雑誌	すく	すく
上位 12	行	読べる	電子メール・メール	連絡	連絡
上位 13	電子メール	電子メール	電子メール	連絡	連絡
上位 14	テレビ	分からない	写真	交通手段	交通手段
上位 15	物	物	物	立て	立て
上位 16	思います	企業	家	趣味	趣味
上位 17	知る	趣味	趣味	趣味	趣味
上位 18	趣味	趣味	趣味	趣味	趣味
上位 19	趣味	趣味	趣味	趣味	趣味
上位 20	趣味	趣味	趣味	趣味	趣味
上位 21	趣味	趣味	趣味	趣味	趣味
上位 22	趣味	趣味	趣味	趣味	趣味
上位 23	趣味	趣味	趣味	趣味	趣味
上位 24	趣味	趣味	趣味	趣味	趣味
上位 25	趣味	趣味	趣味	趣味	趣味
上位 26	趣味	趣味	趣味	趣味	趣味
上位 27	趣味	趣味	趣味	趣味	趣味
下位 15	仕事	ショッピング	ショッピング	連絡	連絡
下位 12	仕事	仕事	仕事	仕事	仕事
下位 11	仕事	仕事	仕事	仕事	仕事
下位 10	仕事	仕事	仕事	仕事	仕事
下位 9	仕事	仕事	仕事	仕事	仕事
下位 8	仕事	仕事	仕事	仕事	仕事
下位 7	仕事	仕事	仕事	仕事	仕事
下位 6	仕事	仕事	仕事	仕事	仕事
下位 5	仕事	仕事	仕事	仕事	仕事
下位 4	仕事	仕事	仕事	仕事	仕事
下位 3	仕事	仕事	仕事	仕事	仕事
下位 2	仕事	仕事	仕事	仕事	仕事
下位 1	仕事	仕事	仕事	仕事	仕事

おわりに: JMP適用上の留意点

- JMPバージョン6.0となり、テキスト型データ処理の自由度が高まった
 - エクスポート・データのインポートや併合処理が容易となった
 - 変数名の文字制限解除 (例: 調査票の質問文や選択肢をそのまま入力)
 - 注: WordMinerには扱い文字数制限はない
- テキスト原データ表内の変数に直接JMP適用は難しい(質的データの分析法を除いて)。
- テキスト型データは計量化・数量化が必要となる。
- 質的変数も場合に応じて数量化が必要となる。
- 数量化で得た量的データ(成分スコア)に種々の手法を適用することが効果的。
- 成分スコアとは用いた変数(テキスト型データ、質問など)の一種の加重平均(合成指標)である。
- 成分の寄与を的確に知悉して処理を行うこと、とくにクラスター化処理。
 - 固有値(あるいは特異値)、寄与率、寄与度の観察が必須要件。
 - 二次分析に用いる成分数(成分スコアの数)の吟味が重要。
- 各手法の特性、性質を習熟した上で利用することが肝要。