◆◆ 質問への回答とコメント ◆◆

今回のセミナー（2006 年 9 月 7 日開催）でも，参加者の皆様から色々なご質問をいただきました．これに対して，セミナー会場で簡単に説明いたしましたが，限られた時間内でのことでもあって，十分ではありませんでした．そこでここに改めて要約を作りましたので，ご一読いただき，分析時にご利用いただけると幸いです．

なお，具体的にどのような課題を抱えておられるか，あるいはどんな対象に関心があるかも，いろいろお書きいただきました．これらについては，ここに表記してよいかどうかの判断に迷いましたので掲載を控えました．しかし，より具体的な情報やコメントを必要とされる場合は，個別にお問い合わせいただければ対応させていただきます（「無料相談コーナー」をご利用ください：http://wordminer.comquest.co.jp/support/consulting.html）．

なお，セミナー時に以下のテキスト，資料を配付いたしました．回答の記述に際して，これらを引用しますので，以下に資料番号を付与して列記しておきます．

資料１：テキスト型データのマイニング（テキスト）
資料２：WordMiner 事例集（導入編）
資料３：WordMiner における多次元データ解析－設計指針と主な特徴の紹介－（スライド・コピー資料）
資料４：WordMiner における多次元データ解析－ミニチュアデータによる対応分析法の仕組みの解読－（スライド・コピー資料）
資料５：テキスト・マイニング研究会 活用セミナー（セミナー当日配布のスライド 8 枚の資料）

◆ お断りとお詫び ◆

配付資料の「資料３」「資料４」は今までのセミナーで用いてきた資料に同じ内容となっております．

この中で「資料１」内の図表番号やページを引用しております．しかし今回はテキスト（資料１）の全面的な再編集を行ったため，「資料３」「資料４」で引用のページ番号とテキストのページとが一致しておりません（セミナーまでにスライド資料の再編集が間に合いませんでした）．お読みいただく場合，図表番号で「資料１」との対応をご確認ください．

訂正：

テキストを何回見直しても細かいミスが残ってしまいます．ここで「資料１」のテキストについて，現時点で気付いた箇所を「訂正」として示します．

①100 ページ，表 18 の中で，行列の対角ブロック内の要素（頻度数）で右下にある「176」を「178」に変更．これはその上の表 16 のクロス表の行和の最下段の「178」に相当する．
②140 ページ，↓1 行目の「表 9」とある 2 箇所はいずれも「表 6」に変更する．

その他，細かい誤記，分かりにくい記述や書き手の勘違いなどもあるかもしれません．お気づきのことがありましたら作成者までお問い合わせ，あるいはご連絡ください（ohsumi@ss.iij4u.or.jp）．

◆ いただいた質問への回答 ◆

質問の内容が多岐にわたりますが，それぞれを関連したグループに括って，またご質問の意味を再確認（書き替える，読み替える）などして，以下に順にお答えします．

1．一般的な質問

WordMiner 利用に限らず一般的な事項と思われるご質問について記します．

Q1:具体的なプレゼンテーションを行う方法は？

A1:ご質問の意味がやや曖昧なのですが，回答者の理解の範囲でお答えします．やや大まかな言い方になります．

① WordMiner の実行内容をプレゼンテーション時にリアルタイムに引用表示する場合，例えば PowerPoint を使って，「ハイパーリンク」としてリンク設定を行えば表示は可能です．他の応用ソフトと同じです．

② 発表資料などで WordMiner の分析結果を利用したい場合は，セミナーの説明にあったように，必要情報（画面）をエクスポート機能で外部ファイル（csv ファイル）として出力し，それをエクセル（Excel）などで編集します．

③ 布置図も同様で，コピー／ペースト機能で Word ファイルなどに貼付することができます．

④ エクスポートした分析結果を他の統計ソフトウェアに移し替えて再分析，再編集することも可能です．JMP（SAS 社）がお奨めする統計ソフトウェアの一つです（http://www.jmp.com/japan/）．JMP はバージョン 6 になってから，扱い文字数の制限も緩やかになり，自由記述，テキスト型データの処理が楽になりました．

⑤ WordMiner のエクスポート出力ファイルを再加工するエクセル対応マクロプログラムも提供しております．構成要素数分布のグラフ表示，簡易集計機能，有意性テストの表内の単語・語句をカラリングする機能などがあります．テキスト・マイニング研究会（TM 研）のホームページからダウンロードできます（シェアウェアとして 5,000 円で配付，TM 研ホームページの「WordMiner Tips」をクリックして「シェアウェア」を参照：http://wordminer.comquest.co.jp/wmtips/shareware.html）．

Q2：以下の 3 つの質問には共通した部分があると思われるので，まとめて回答します．
Q2-1：本人の語り（方言や感情）を分析する場合に使えるだろうか？
Q2-2：アンケートの自由回答の分析
Q2-3：聞き取り調査（本人の語り）の質的データ分析

A2：WordMiner はそもそも，社会調査やアンケートなどで取得したデータ解析用ツールとして開発されております．したがって一般的な質的データ・質的調査データの分析に適しております．しかし，セミナーでも申し上げたように，WordMiner はデータ解析支援ツールとして機能するものです．Q2-1，Q2-3 のような課題については，やはり「**データ取得をどう行うか**」（**データ収集方式**：data collection mode と言います）が決定的な要素であると理解しております．適切なデータ収集が行われれば，WordMiner はそれに応じた分析結果を与えると考えてください．データ収集がいい加減であれば，当然「闇に鉄砲」となって，期待する結果は得られません．一般に，マイニング・ツールは分析対象データが適切に取得されたかの客観的な評価を行えるほど知的（intelligent）ではありません，マイニングという言葉に惑わされないよう適切なデータ取得計画，調査設計を行うべきです．とくにインタビュー形式による聞き取り調査やフォーカス・グループ（FG）によるデータ収集などでは，聞き取り方・インタビューの手順などの標準化に注意すべきです．ここらのコンサルティングも TM 研として行っております．

（†）資料 1 の第 1 章，第 2 章あたり，および資料 5 を合わせてご覧ください．

Q3:看護用語コーパスの作成は可能か？（「看護用語コーパス」の記述に対して）

A3:これへの回答も，一部は上に述べたことに関連するでしょう．この分野（≒看護・福祉他）での用語集はもとより，専門的に用いる用語，日常的に生じる「患者対看護側」のやり取りなどで，どのようなデータ（＝用語，単語・語句）が体系的に収集できるか，ということでしょう．

10 数年前頃に「エキスパート・システム」が流行となりました．しかし最近はめっきりこの言葉を聞かなくなりました．とくに，統計ソフトウェアと連携して統計エキスパート・システムの萌芽的研究や試作システムが多数登場しましたが，現在残っているものは皆無に近いでしょう．理由の一つは「知識ベース，知識ルール」の構築に問題があったと考えております．そもそもこうした知識情報が体系的に組織化されるなら問題は解決するはずで，それが難しいからこそ，データ解析のノウハウが必要となるのです．最近のデータマイニング（DM）やテキスト・マイニング（TM）もこれと同じような陥穽に落ちないかと懸念されます．

ある特定分野のコーパス（それを集めたコーポラ）の作成も，（コンピュータ利用に関わりなく昔から）言語学などでの長い歴史や膨大な研究があるように，かなりの力仕事となる課題です．

看護に関しては「ある患者さん，その看護側，関係者」といった小規模の限定した範囲内でまず，試験的に小さなコーパスを作成し，それをコアとして試行錯誤的に他の患者さんを対象に拡げることが考えられます．つまり初期の「コーパスの範囲」をどのように設定するかでしょう．WordMiner で試験的に辞書群を作って（ここでのポイントは，なるべく分類・カテゴリー化して複数の辞書を作ることかもしれません）試用することでしょう．「分類」「語彙群の考え方（コーパス）」「類語・関連語群（シソーラス）」の設計が一つのキーワードかと思います．この段階での自動分類法（クラスター化法）への過大な期待は避けるべきでしょう（その分野の経験則，体験などを計量化することが先決，より重要でしょう）．

2．分かち書き処理と構成要素

Q4：構成要素数の計数の仕方と解釈は？

A4：これを一言で述べることはやや困難です．構成要素数の（統計的な）分布の特徴，計量的評価やその解釈などは，古典的な研究を含め（例：Zipf の法則他）資料１の補足資料１を参照してください（1 節とくに 1.3～1.6 節）．また WordMiner の出力する結果の解釈，観察方法についても，資料 1 にいろいろと示しました．ここは我々がいう**初動探査**が重要な操作となります．

Q5：構成要素変数を作るとき，一度削除したものをもう一度戻して使えるか？

A5：残念ながら「変数情報の確認（変数名変更・削除など）」にある変数一覧情報から一旦削除した変数は復活できません．なお，構成要素変数については，編集前の構成要素変数あるいは編集過程で別名称で保存しておいた構成要素変数は残っております．削除した変数の元通りの復活とはなりませんが中間作業の状態には戻れます．WordMiner には基本的には「Undo 機能がない」のでくれぐれもご注意ください．

対応策としては，中間作業状態で別名の変数を作成し保存しておくことでしょう．また作業履歴は WordMiner デスクトップの左フレーム内の「実行の履歴」で確認できます．

Q6：分かち書き処理を行わないで構成要素を抽出する方法はあるのか？

A6：他の質問（Q7，Q8）にも関連しているようです．このご質問の意味が少し曖昧ですが，以下のような場面であろうと想定してお答えします．

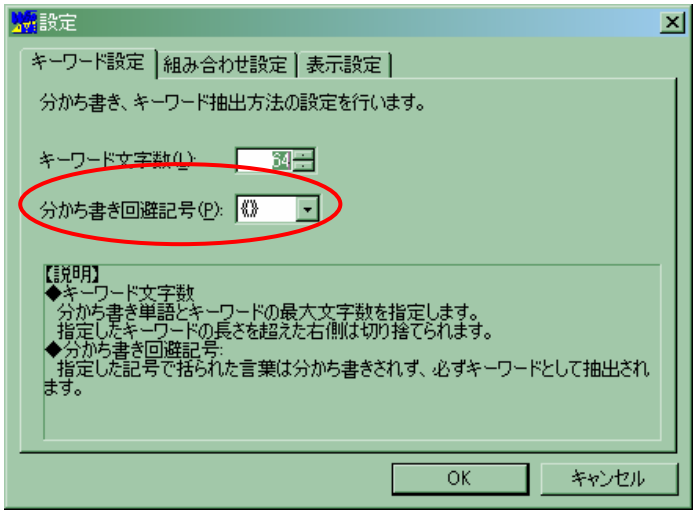
場合１：いわゆる定量型調査・量的調査（選択肢を設けて回答を得るような形式の調査）で，その選択肢を文字データで表記し，インポートしたような場合．

これについては「資料１」にいくつか例を挙げてあります．例えば 92 ページの表 5，94 ページの表 7，95 ページの表 8 などがそれに相当します．こうした例では，文字表記した質問文の選択肢を含む（原始）変数を質的変数に変換することで質的変数として用いることもできますし，またそれを構成要素変数として用いることも可能です．

場合２：「資料１」の 92 ページ，表 5 のような場合，つまり多項選択型質問（MA：multiple answer）あるいは自由回答連想で列記してもらうような自由記述文の場合があります．この例のようにベバリッジ名を適当な区切子（セパレータ）を使って，ここではカンマ「,」となっておりますが，これを原始変数としてインポートしたあと，いったん分かち書き処理を行い，各ベバリッジ名を単語（構成要素）として抽出することが可能です．つまり「分かち書き処理は必要」ですが，選択肢型質問文の処理に使えるということです．

場合３：「資料１」の 94 ページの表 7，95 ページの表 8 のように，いわゆる選択肢型質問の選択肢文をそのまま構成要素変数として用いることはできます．

（†）なお，分かち書きして欲しくない語句・文節などは，「回避記号」を用いて分かち書きを抑制することができます．例えば，「お寺詣りをよくする」という選択肢があったとして，これを分かち書きしないで，1 語の構成要素として使いたいときは，回避記号を使って「《お寺詣りをよくする》」としておくとこれは分かち書きされずにこの形で保存されます．後は辞書で記号「《》」を削除します．回避記号としては何種類かが用意されております．画面例を下に示します．



この例では分かち書き回避記号として「《》」を選んであります．

場合４：自由回答に限らず長い文字数のテキスト型データを分かち書きせずに取り込む別例としては「他の分かち書きツールを使った結果」をそのまま読み込むという使い方があります．換言すると「WordMiner で行う分かち書き処理とは異なる分かち書き処理結果を使いたい」という場合に相当します．これについては「資料１」の「補足資料２」の 250 ページあたりに説明があります．形態素解析による品詞分類を行った結果を WordMiner で分析したいというような場合にもこの方法で対応可能です（WordMiner 対応のインポート・ファイルを作ることも可能です）．

Q7：原始変数データをそのまま構成要素として用いる方法は？

A7：ここのご質問の意味が少し曖昧ですが，長いテキスト型データを含む変数（原始変数）をそのまま構成要素変数として指定することは“形式的”には可能です．しかし長い文字数の，しかも個々の意味が違う無数の構成要素が生成されますので，実用的見地からは意味がない分析となります．あるいは上の質問 **Q6** のような意味であれば，上に回答した通りです．

Q8：分かち書き処理で，長い文章を扱うとき，分かち書きされることで全体の文脈が見えづらくなる．データ入力段階で先にコーディングしてから WordMiner で分析することは可能か？

A8：このご質問の前半と後半では少し違った意味内容となっています（複数の意味が含まれております）．まず「分かち書きにより全体の文脈が見えづらくなる」という点ですが，これは（部分的には）確かにその通りです．とくに自由記述などは長さに制限がないことが多いので画面上ではどのような表示方法を用いても無理があります．また WordMiner は構文解析（統語解析）や係り受け処理などを行いませんから，このような意味での分析はできません．しかし分解した個々の単語・語句（構成要素）の相互の関連性（というか類似性）を対応分析によりある程度評価できるということも指摘しておきます．

次に後者，つまり「データ入力段階で先にコーディングしてから WordMiner で分析可能か」ですが，これは可能です．ご質問の意図はこちらにあったようですので，これについてもコメントします．社会調査などではこうした操作を**アフターコーディング**（あるいは**コーディング**，ポストコーディング）と言います．選択肢などを設けて前もってコードを決めておくこと（プレコーディング）に対してこういう言い方をします．このような場合，良く知られているようにコーディングを行う作業者（コーダー）によってそのアフターコーディングの結果に差異が出ることがあります（当然です）．作業者の感じるところ，重きをおいた語句・表現などが異なること，つまり自由回答の内容や解釈は多様ですので，読み手（作業者）の反応，つまりアフターコーディングの結果に揺らぎが生じます．

大抵は事前に「コードブック」を作っておいて，（それを辞書のように使って）なるべくコーディング結果がそろうように努めても（標準化しても），複数の作業者がいる場合は異なるコーディング結果となる可能性が高いでしょう（またかなりのスキル，リテラシーを必要とするとされております）．

対応策しては，まず原文（元の自由回答文）を活かして WordMiner で処理する，複数の作業者による結果を比較分析する，その両者の結果を比較するなどが考えられます．こうした一連の処理分析操に WordMiner は利用できます．実際に，回答者も過去に何例か経験したことがあります．

Q9：分かち書き処理で得られる分かち書き結果とキーワード結果の違いは何か？

A9：ここはセミナー演習で体験されたことでご理解いただけたかもしれません．分かち書き結果は「元の文章・テキストを**分かち書き単位＝構成要素**に分解した状態」を言います（形態素とは限りません）．「キーワード」とは，分かち書きした結果から，主に名詞的な主要単語・語句を拾い出したものです．正確な意味での名詞的語句を選んだわけではありません．また重複語は除外してあります．つまり（サンプル単位でみたときに）あるサンプルで使われた複数回の構成要素は 1 語として置きかえられます．例えば，「私」があるサンプルで多数回使われてもキーワード抽出では「私」は 1 回と見なされます．実は TM ソフトによって，ここらの処理手順が異なり，しかも手順が公開されないことが多いので注意を要します．

3．対応分析法，クラスター化法

対応分析法に関する説明は「資料 1 」の 85 ページから基本的な情報（WordMiner 利用上，

必要最小限の情報）として記述しました．またクラスター化法については「資料３」に簡単な説明があります．なおクラスター化法の詳しい説明は現在作成中です．今後開催のセミナーで順次加える予定です．資料１に挙げた参考文献などを併せてお読みください．

Q10：対応分析，クラスター化法以外の解析手法はあるのか？

A10：答えは「ない」です．理由はいろいろありますが，そのいくつかを記します．

まず第１に，汎用統計ソフトウェアとは異なり多数の技法・手法を採用することは意図的に避けたことがあります．第２に，定性情報・質的データの計量化・数量化を行う方法論として（おそらくは）対応分析法が優れた手法であるということです．もちろん対応分析法の変形（亜種）は無数にありますが，数理的特性が明示的に分かる熟成された手法ということで採用しております（また WordMiner を開発した研究者はこの分野の専門家です）．第３に，テキスト型データ，つまり定性情報・質的情報は何らかの形で計量化・数量化を行わないと，また基本的に多次元構造のデータの情報を何らかの形で縮約化（節約原理による次元縮小操作）を行わないと視認化や客観的な解釈が容易ではないからです．

クラスター化法（クラスター分析，クラスタリング，自動分類法）の多くは，数量化された情報（例：いわゆる量的データ＝区間尺度・比例尺度などの加減乗除が可能なデータ）に変換しないと適用できない手法が多いことがあります．良く利用される**ウォード法**（Ward's method : Wishart のアルゴリズムによるウォード規準による分類法）や***k*-平均法**（*k*-means 法：後ろで改めてコメントします）などは，原則として量的データにしか適用できません．こうした基本原則も守っているかどうか怪しいソフトが市場に流通しておりますので十分な注意が必要です．WordMiner ではいわゆる**ハイブリッド法**（階層的分類法，非階層的分類法の併用）を用いております（「資料３」参照）．

WordMiner ではこの２つも手法（対応分析法，クラスター化法）を中核に，実用上有効であると思われる周辺の機能をいろいろと加えたことにあります．有意性テスト（頻度，距離），追加処理機能(†)，その他いろいろ設けてあります（セミナーではほんの一部を紹介しました）．テキスト型データを解析するうえで，一体どのような情報が必要かを研究面と実用体験・経験則から考えて設計したということです．

（†）追加処理機能については「資料１」の 131 ページ（6 節）に簡単な紹介あります．

つまりこのご質問への回答は，ソフトウェアはその設計者・開発者の設計指針（design policy）に依存するということです．また，随所にエクスポート機能を設けた理由は，他の統計ソフトウェアに（WordMiner の出力結果を移して）別の分析を行う機会を提供するという設計指針があるからです．他のソフトでできることはそちらに任せよう，その分だけ負荷を減らしたかったということでもあります（テキスト型データを扱うというだけで負荷が非常に大きくなる）．

Q11：対応分析のクロス表で頻度ゼロの発生時の対応は？

A11：分析対象として生成したクロス表の行和，列和のどこかに「ゼロ（頻度）」が生じた場合のことを指しているとしてお答えします．答えは「問題なし」です．計算上はこうした列・行は除外されます．このような例を実際に計算されて観察していただくとご理解いただけると思います．実際，構成要素数の閾値（あるいは「いきち」と言うこともある）を変更して用いる構成要素変数（単語・語句群，つまりそのときのコーパス）を変えると当然このような状況が生じますが適切な手当がなされております．

Q12：対応分析で得られる軸（成分軸）の解釈について

A12：いわゆる因子分析などの影響があって，「因子≒軸あるいは成分」といった（ある種の）誤った情報が流布しているようです．蛇足ですが「軸の回転，因子の回転」操作についても同様の傾向があります（誤った利用法）．

まず，対応分析で得られる成分（軸）は，いわば主成分分析型手法で得られる成分（主成分）に類似したことと考えてください．例えば主成分分析ではデータ（の多次元空間内）の変動・チラバリの総量を共分散行列で表しこの総情報（＝共分散行列の跡和・トレース）を固有値で分解するというものです（少々専門的な記述ですが，多次元空間内の変数間の関連・相関を考慮しながら個々の変数の変動の大きさを，見方を変えて固有値で評価するのです）．

対応分析法では（ピアソンのカイ二乗統計量による）**クロス表の独立性検定**と重要な関係にあり，やや粗っぽい言い方ですが，このカイ二乗統計量を固有値で分解する操作に相当します．分かりにくいでしょうが資料１の第 3 章「対応分析法・数量化 III 類の考え方」に関連する基礎的な情報を記述してありますのでご覧ください．なお今後のセミナーで「数理編（仮称）」を設けて解説を行う予定です．また「資料４」にミニチュアデータを使った例示がありますので，資料１の記述と併せてご覧いただくとよいでしょう．

さてここでの質問への回答は以下のようになります．少々短絡的な端折った書き方となります．

① まず，布置図で言えば，その図で見ている次元（成分軸）について，図内の各要素（構成要素やサンプル）の相対的な位置関係，つまり同じ位置にあるか，互いに離れているか，離れ方が反対方向か，直角の方向か，その図内（観察している成分軸）で，中央あたりに位置するか，はずれた位置にあるか，…といった観察が有効です．

② とくに，眺めている図（の成分軸）で，中央当たりに位置するということはその成分については平均的ということ，はずれた位置にあることは，特徴的な意味を持つか，あるいははずれ値であるか，といった見方をすることです．換言すると布置図では周辺に布置された要素（構成要素，サンプルなど）から観察するというのがコツです．

③ 成分スコア（数量化得点），布置図の見方については「資料１」の 111 ページ～114 ページあたりに記述しました．

④ その他，WordMiner が出力する情報として，寄与度（絶対寄与度，相対寄与度）なども参考になります．図でみた（主観的な）印象を，布置要素の相対的な位置関係だけでなく客観的な重要度を知るための指標として用いるものです．

Q13：クラスター化で用いている手法は？

A13：階層的分類法（凝集型階層的分類法）の一つであるウォード法（Ward’s method）と，非階層的分類法とくに分割最適化型手法の一つである *k*-means 法（*k*-平均法）を用いております．簡単な図解が「資料４」にあります（セミナーでも要点を説明しました）．

次の質問，**Q14**，**Q15** にも関連しますので，ここで留意事項を要約しておきます．

① まず，ウォード法ですが，これは，無数にある階層的分類法の一つです．正確にはウォード規準による Wishart（ウィシャート）のアルゴリズムによる階層的分類法です，ウォード規準とは分割で得たクラスターの級内分散（クラスター内分散）の和を最小化するような規準を言いますが，これを階層的に用いる，つまり級間分散（クラスター間分散＝一種の距離になります）の小さいクラスターから併合を（階層的に）順に繰り返すというアルゴリズムを Wishart が考えたことで広く普及しました．初期で用いる距離として原則として平方ユークリッド距離を用いますが，ここらがいい加減なソフトが多いので注意が必要です．

② *k*-means 法とは，こうした一つの確定的な手法があるわけではありません．正しくは，*k*-means 型（ルールによる）手法の集まりを総称しているというべきでしょう．クラスター化最適化規準として何を用いるか，配置・再配置（更新作業）を行う際の更新時期とそのルール，初期化手順（初期クラスターの生成方法），…と様々な条件設定が必要なこと，またプログラムを作成した人のさじ加減で調整されるパラメータが多いことが特徴です．

換言すると，これらの諸条件の設定によって，用いたプログラムによって同じデータセットを用いても結果が異なることがあるということです. このことは階層的分類法を用いる場合にも同じことが言えます. であるからどのような手順で分析処理を行ったかを暗箱化せずに明らかにする必要があるのです.

③ クラスター化法全般に共通することとして，様々な（研究，応用）分野から，様々な手法が登場したことで，同じ方法に異なる名称が付いていることが多いこと，つまり方言（jargon）が多いことがあります．また，用いるクラスター化規準とそれを具体化するアルゴリズムが多様ですから，この組合せを考えただけでも膨大な組合せ数つまり処理手順があるわけです．要はクラスター化法（クラスター分析，クラスタリング，自動分類法）という一つの手法があるわけではない，ということを承知したうえで利用することです. 報告書，論文等作成時に「クラスタリングを行ったら／クラスター分析によると○○の結果を得て…」といった記述はあってはならない誤記あるいは理解されていない表記です. 少なくとも適用した手法名，用いたパラメータの設定条件，用いたソフトの名称などをはっきりと記述すべきです.

Q14：クラスター数の決定方法は？

A14：これへの回答は「今はない」です．これでは身も蓋もないので，もう少し補足して述べます．実はこの質問の背景にはそもそも「**クラスターとは何か**」（どのように定義するのか）という基本的な問題があります．多くのクラスター分析関連の研究書やペーパーの始めに大抵は書かれていることです．ここでは詳しいことは記しませんが，以下の参考文献を挙げておきます.

(1) 大隅昇（2000），多次元データ解析における分類手法の役割 –分けて知ることの効用と難しさ–，エストレーラ，10 月号，79 号，10-20.
(2) 大隅昇，保田明夫（2004），テキスト型データのマイニング－定性調査におけるテキスト･マイニングをどう考えるか－，理論と方法（数理社会学会誌），Vol. 19，No. 2，135-159.
(3) Boris Mirkin (2005), *Clustering for Data Mining – A Data Recovery Approach* -, Chapman & Hall/CRC.

クラスターとは何か，つまりどのような場合をクラスターと考えるか（**クラスターの定義**）によって，どのようなクラスターが作られるのか，つまりアルゴリズムが決まると考えればよいでしょう．クラスタリングとはある約束したクラスターをアルゴリズムによって生成すること（作ること）です．よって「**クラスター化**」と呼ぶことが相応しいのです．つまりクラスター化は利用者の要求する目的に応じて条件設定がなされるものです.

一方，分類対象には構造的にいわゆるクラスター（cluster：房状）の塊が散在するのか，あるいは雲のようにもやっとした境目もはっきりしない姿なのか，それは様々です．しかし，どのようなクラスターを想定するか（＝**クラスター化規準**）を決めないことには分類処理（アルゴリズムの実行）ができないことになります．ウォード法や *k*-means 法は，例えばクラスター内分散（の和）の最小化を規準とする場合で（規準はこれだけではありません），どちらかというとバラツキが似たような大きさのクラスターを“作る”傾向にあります．換言するとはずれ値の検出はやや苦手なのですが，WordMiner ではこれらをハイブリッド化することで，比較的妥当な分類結果が得られるようになっております．大切なことは，多くの場合，とくに対応分析で得た成分スコアなどの場合，その分布は雲状になって塊状のような（多くのクラスター化法にとって）理想的なクラスターは存在しないことが多いということです.

このようなことでクラスター数の決め方についても適切な指標はないといってよいでしょう．この状況は丁度ヒストグラム（度数分布）の級の数（セル数）を決める決定的な方法がないことに類似します（最適層別化問題）．いろいろな指標が提案されてはいますが，これという決定的な方法はないでしょう.

WordMiner では，クラスター化過程における（階層化過程における）クラスター内分散の

変動の履歴を追跡し，その大きさが大きく変化する位置を棒グラフとして表し目安とするという方式を採用しております．これについて以下にいくつか例を挙げておきます．要は「クラスターの生成情報（構成要素クラスターの生成情報，サンプルクラスターの生成情報，カテゴリークラスターの生成情報)」「布置図」「成分スコア」それぞれを対比照合して分析することです．

例１：

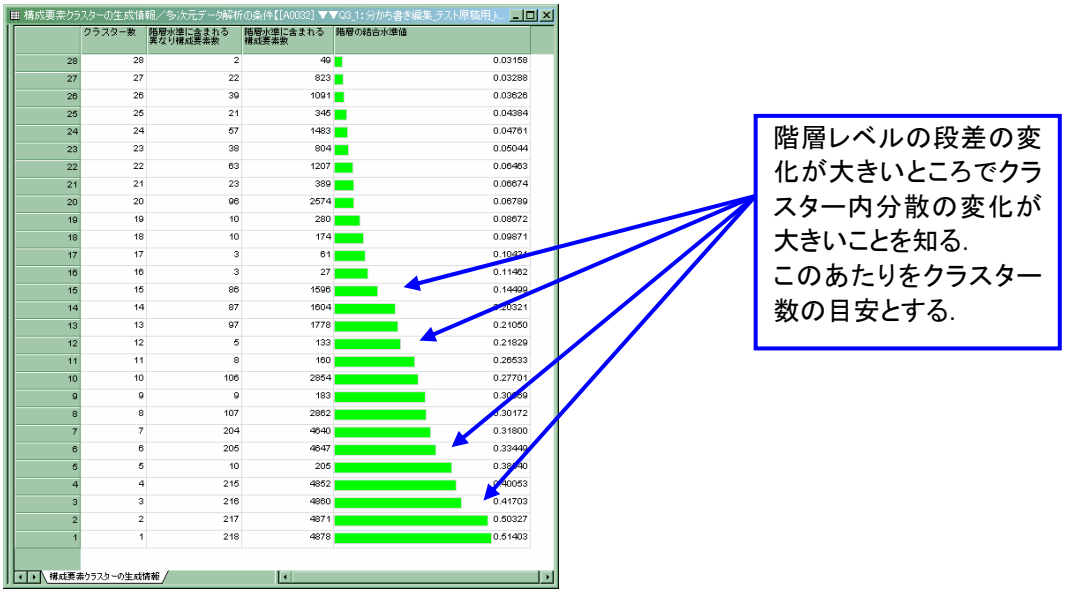


図1　構成要素クラスターの生成情報例（1）

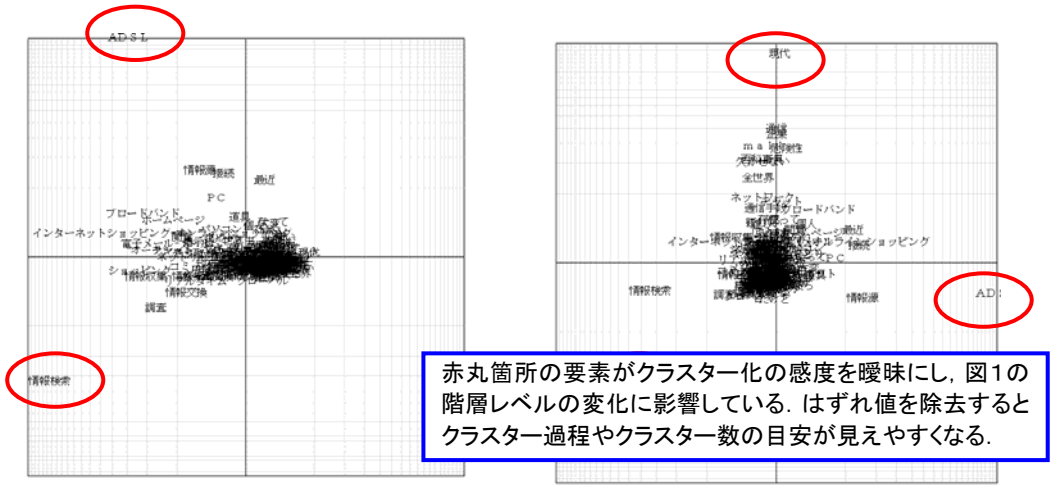
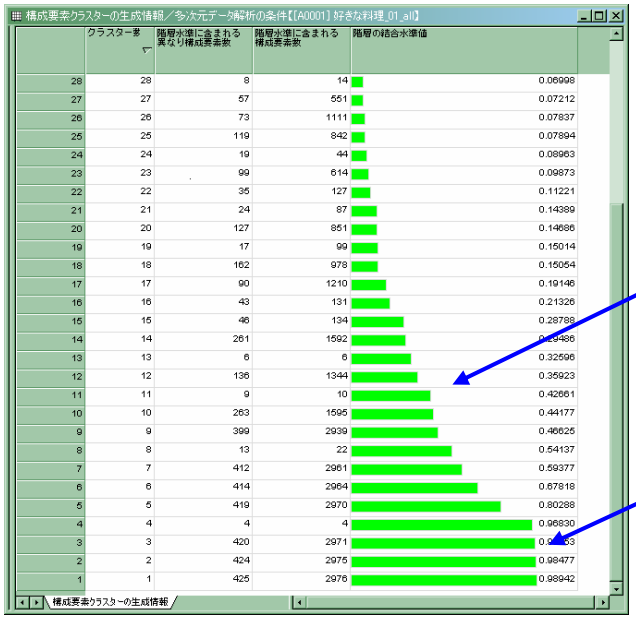


図2　図1に対応する構成要素布置図，左が(1,2)成分，右が(1,3)成分

この例では，いくつかの成分軸ではずれ値が見られるので，これらを検出（検索で確認）したあとに必要に応じて置換や削除を行うことになります．また中央部の雲上の部分ついては，その後に細分類・再分類を行うことがよいでしょう．次の例 2 も似たような傾向にありますが，クラスター化履歴のグラフからみるように，少数のはずれ値の影響を大きく受けているので，これの処置を行う必要があります．これらはいずれも対応分析で得られる成分スコアの特徴を示すものであり，同時に多くの場合にこうした布置図が得られることが多いのです．はずれ値への対応は処理条件を変えながら探索的に根気よく観察を続けることです．

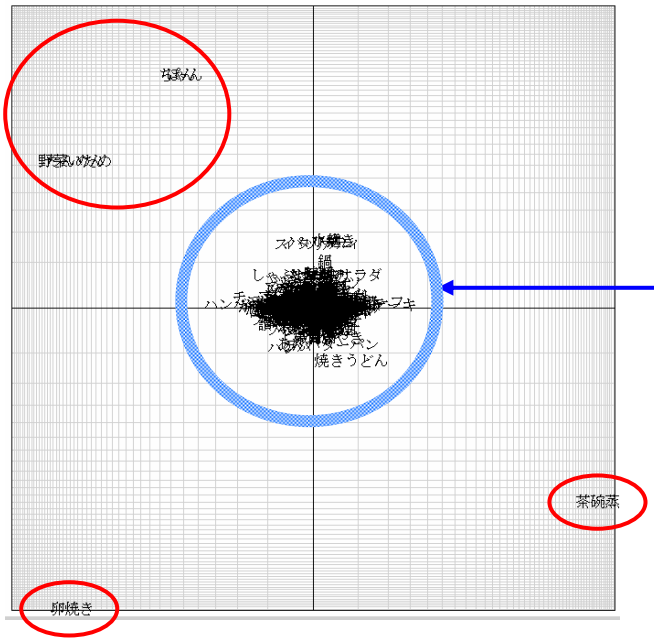
例２：



クラスターの存在を示すクラスター内分散の大きな変化はなく棒グラフは滑らかに変化する．図４の中央部分（雲状部）に対応する．

クラスター数が少なくなるとはずれ値の影響を受けて段差，つまりクラスター内分散が大きく変化する．図 3 の周辺のはずれ値の影響がここに見える．

図3　構成要素クラスターの生成情報例（2）



赤丸で囲った要素がはずれ値としてクラスター化過程に影響する．
中央部の他の多くの要素の特徴をはずれ値が抑制している．はずれ値相当の構成要素を成分スコアの分布を見るなどして，必要に応じて辞書編集（削除，置換）で調整する．

この中を改めてクラスター化する（再分類・細分類する）．

図4　図3に対応する構成要素の成分スコアの布置図，ここは（1,2）成分の図

例３：

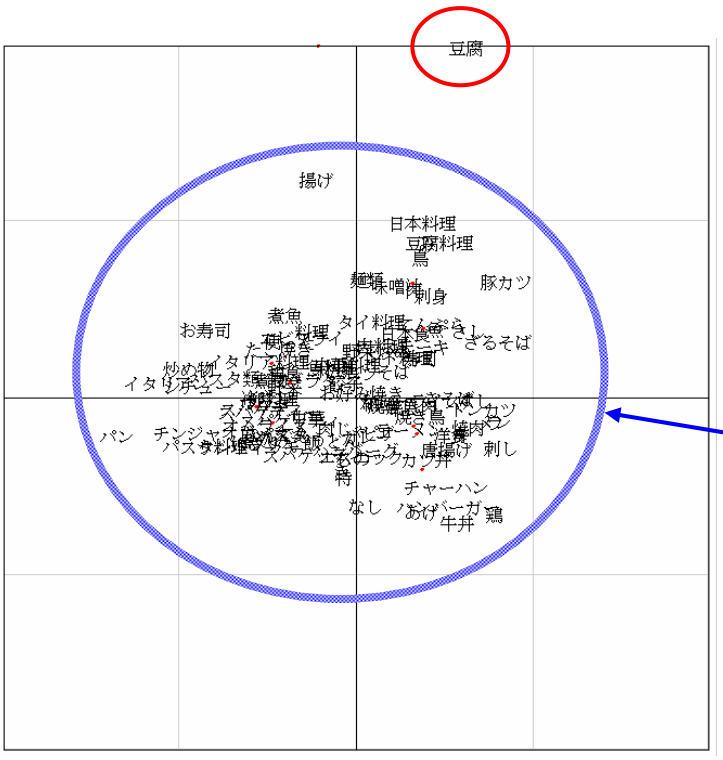
構成要素クラスターの生成情報／多次元データ解析の条件【[A0015] 未処理テスト_t=5】

	クラスター数	階層水準に含まれる	階層水準に含まれる	階層の結合水準値
23	23	6	71	0.00559
22	22	7	86	0.00579
21	21	9	123	0.00647
20	20	10	148	0.00653
19	19	12	159	0.00656
18	18	9	315	0.00668
17	17	6	76	0.00686
16	16	11	344	0.00728
15	15	5	246	0.00785
14	14	3	19	0.00792
13	13	13	482	0.01049
12	12	21	474	0.01281
11	11	19	558	0.01449
10	10	6	62	0.01462
9	9	11	317	0.01698
8	8	13	157	0.01896
7	7	23	620	0.01935
6	6	31	622	0.01969
5	5	24	501	0.02162
4	4	35	818	0.02827
3	3	54	1242	0.04080
2	2	60	1304	0.05498
1	1	95	2122	0.11410

構成要素クラスターの生成情報

この箇所（2 群～3 群）での変化が大きく，その後は階層レベルにとくに大きな変化が見られない．

図5　構成要素クラスターの生成情報例（3）



赤丸で囲った一つの要素がはずれ値としてあるが，他の要素はまとまってとくに塊状のクラスターは観察されない（少なくとも，この成分軸では）．
このことが図5のクラスター履歴の階層レベルにも現れて棒グラフの段差は比較的滑らかに変化し顕著な段差は見られない（つまり塊状のクラスターは検出されない）．

多くの場合，このような布置の状況でクラスター化を行うことになる．つまり塊状のクラスターなどは存在しないことが多い．よってクラスターを生成するのである（クラスター化）．

図6　図5に対応する構成要素の成分スコアの布置図，ここは（1,2）成分の図

Q15：以前，「茶筌（ちゃせん）」で学生のレポートを分析して，SPSS でクラスターしたが，WordMiner でやった場合と結果が違うか？

A15：回答は「（おそらく）異なる」です．理由は既に上の **A13** に記した通りです．大抵の統計ソフトウェア（例：SAS, SPSS, JMP, Minitab, S-Plus,…）にはクラスター化機能が搭載されておりますが，マニュアルなどをみてもあまり詳しい説明がありません．十分に注意して利用することが肝要です．

利用上のヒントとして，以下を挙げておきます．

① まず，どんな名称の誰が提唱した方法であるかを確認すること．
② 階層的分類法であれば，どのような類似度・非類似度（距離）を使うのか，アルゴリズムは何か（誰の提唱した方法か），それらの組み合わせはオプションとしてどう規定されるのか，出発データに何かの加工を行うのか否か（例：標準化処理の有無），…などを確認する．
③ 非階層的分類法，とくに分割最適化型手法であれば，初期化の方法（初期クラスター・核の作り方），例えばランダム配置・ユーザ指定など，配置・再配置（最適化規準の更新）の方法と手順，収束判定をいかに行うか，などが分かること．
④ 分割最適化型手法などは，完全最適解を求めるのではなく**局所最適解**を求めることになるので，初期設定パラメータによって分類結果が異なる．よって，対象データをランダム分割して（無作為抽出して）何回も分類操作を試みるなどの手当が必要とされること．
⑤ WordMiner はここらの操作をある程度緩和するために，クラスター初期化に階層的分類法（ウォード法）を用い，クラスターの核（種：seed point）を作ったあと，*k*-means 型手法により反復再配置・更新を行いクラスター化規準の最適化を行うという方式を採用している．しかしこれとても局所最適であることには変わりはない．
⑥ 前述のようにクラスタリングとは設定した条件の下にクラスターを生成して情報を見易くする（圧縮化する）操作の一つ（クラスター化）と考えるべきである．
⑦ なお **A6** に示しましたが「茶筌」の処理結果を（ほぼ）自動的に WordMiner にインポート可能なファイルを作るツールが用意されておりますので，これをご利用いただくと別の比較ができます（KH Coder というツールを使います）．

4．有意性テスト関連

Q16：有意性テストの方法は？

A16：WordMiner でいう有意性テストにはいくつかの方式があります．構成要素の出現頻度によるテスト，距離（カイ二乗距離：χ^2 距離）によるテスト，要約化情報の有意性テスト，…です．これらについてはいずれ詳しい説明が必要でしょうが，質問の多い「頻度による有意性テスト」については「資料１」の「補足資料１」にかなりのページをさいて例題を用いて示しましたので，これをご覧ください（資料 1 の 231 ページ～243 ページ）．「距離による有意性テスト」についての記述も若干ありますので参考としてください．

有意性テストは WordMiner の特徴的な機能の一つです．WordMiner の布置図もそうですが，多くの TM ツールでは何らかの図を示す，あるいは視覚化情報を描画することで理解を容易にするような錯覚を与えます．いわゆる SOM（自己組織化マップ；別名，コホーネン・マップ）もそれらしい結果を見せてくれます．しかし良く考えると人が視認できる範囲，図をみて理解できる範囲は（経験的にも）**せいぜい数百語くらいまで**です．数千語，数万語となった構成要素（単語・語句）の関連を視認するには，さらなる情報の圧縮化操作を必要とするわけです．例えば，クラスター化で得たクラスター変数（つまり類似グループ）の内容を観察して適当なネーミングを行うなどを行うわけです（マーケティング分野でよく利用される方法）．WordMiner は構成要素数がかなり多くなっても，有意性テストを用いることで，視点を変えて観察しようという設計指針があります．例えば，（構成要素変数）×（質的変数，ク

ラスター変数）のオプションを用いれば，構成要素数に関係なく，その構成要素群がどのような質的変数，クラスター変数と有意であるのかを探索的に検証することができます．また（回答・サンプル）×（構成要素変数）の場合も同様です．この利点を有効に活用する操作をぜひ習得してください．これと布置図に観察を合わせれば，おおよそのデータの特徴は把握できます．

なお，有意性テストは対応分析の成分スコアの布置図とは直接は関係はありません（もちろんまったく無関係でもありません）．布置図はあくまでも出発行列としたクロス表の行（表側）と列（表頭）の関連性（つまり対応）を少数次元内に圧縮した情報を観察するツールです．一方やや粗っぽい言い方ですが，有意性テスト（頻度，距離）は分析時点で確定した構成要素群（その課題のコーパスと考えてよい）の内容を分析するツールです．全構成要素の分布と，特定な層内（＝質的変数の選択肢やクラスター変数の各クラスター）内にある構成要素の分布を比較する操作です．これをその層内で，例えば頻度検定であれば，ある性年齢区分（質的変数）に登場する構成要素群が，全構成要素群に比べて多く登場するか少ないかを有意性テストの結果として「上位（全体からみて多い），下位（全体からみて少ない）」として表すものです．

5. 辞書，類語・関連語，語彙群，など

類語・関連語（シソーラス），語彙群（コーパス）などに関連して，辞書の考え方は重要です．また TM ツールが避けて通れない課題でもあります．

WordMiner は，言語解析を行うツールではなく，あくまでもテキスト型データの（統計的）解析に軸足をおくソフトです．よって言語解析的な機能，例えば，構文解析（統語解析），意味解析，などは行いません．またいわゆる形態素解析による品詞特定化や品詞分類，活用の分類などは行いません．

（†）ここらは「資料２」の「補足資料２」も参照してください．

Q17：置換辞書の使用方法

A17：辞書作成の方法については「資料１」の第 4，5 章に若干の説明があります．そもそも辞書編集などは行わないで済ませられればそれにこしたことはないのですが，日本語の特性，つまり表記・表意・表現が多様なことから，避けて通れない作業にもなってます．TM ソフトウェアによっては，こうした部分を見せないようにして，つまり適当なルールで形態素解析や分かち書き処理を行った結果をそのまま容認することで次の分析ステップに移行する方式をとっているものもあります（つまり分析内容が不透明なことが多いのです）．一方，WordMiner は原則として，利用者に辞書編集の有無の判断を委ねるという方針をとっております．よって回答は以下のようになります．

① 利用者がとくに複雑な辞書編集を必要としないと判断すればそれで分析を進めればよい．記号，句読点程度の削除を行って，また閾値をやや多めにとって出現頻度の少ない構成要素を除外して，データの基本的な構造の探査を行う．

② 類似語・関連語の併合・置換などが必要と判断するなら，必要な範囲で置換辞書を作成する．共通テーマの研究課題を持つとか，特定の製品の顧客クレームの分析を行うなどの例では，共有できる置換辞書を一種のコーパスとして作ることが良いかもしれない．

③ 要点は，一つの大きな辞書（置換，削除）を作るのではなく，分析対象データ（テキスト型データ）の内容をよく吟味してカテゴリー化を行って複数の（目的別）辞書に分けることがよいかもしれない．元々非構造的（unstructured）なテキスト型データを部分的に加工して半構造化（semi-structured）して用いるということです．欧米の TM ソフトにはこうしたカテゴリー化，タグ化といったドキュメント・マイニング的な機能までを含む例があるようです．

Q18：自分の作った辞書が果たして正しいかどうか分からない．
加工の仕方によって結果が変わってくる可能性がある（信頼性の問題）
人によって結果が変わってくるはず（これは無視してもよいのか）
同じデータを用いても…

A18：非常に重要なご指摘であり質問です．またこのことは WordMiner に限らず TM ツールに共通した問題でしょう．

まず「(辞書の）加工の仕方で（処理）内容が変わる」ほその通りであり当然です．通常の選択肢型質問を用いる場合などと異なり，計量的に比較する，例えば回答比率%で比較することができませんから，確かにこうした問題は起こります．

（信頼性とありますが）これは一般に「妥当性，再現性」などの問題として議論されることでしょうが，自由回答を始めテキスト型データの分析では常に頭のすみに置いておかねばならないことです．

要は TM とは数式を解いて確定的な解を得るというような確たる話しではなく，曖昧性をもった，しかしそこに含まれているであろう，ある種の知見・発見的要素を探査するツールであると考えることです．

また「辞書が正しいかどうか」は実は誰も決めることはできません．分析の当事者が「これでよしとして」用いることです．ここらはコーパスやシソーラスにも関連することで，これらの確定情報を作ることができない限りは議論できないことでもあります（解はないということになります)．これといった解がないからこそ，長い研究の歴史がある課題でもあるわけです．

① 辞書作成の履歴をしっかりと把握し，分析をどのような辞書で行ったか，つまり作成したコーパスを確定させたうえで分析を行う．
② 辞書の内容を何回も編集することが起こるだろうが，最終的な分析で利用した辞書（群）をその課題（プロジェクトや多次元データ解析の試行）でのコーパスとして確定し内容を確認する．
③ 「人によっても結果が変わる」こともその通りです，よって辞書作成を標準化するとか，いったん作成した辞書を確定したコーパスとして標準辞書として用いる(お互いに共有する）などを決める必要もあるかもしれません．コーパス化をどう進めるか，どのように考えるかがポイントと考えます．
④ ついでにここで指摘しますが，分かち書き処理に用いるソフトによる分かち書き結果にも一意性はありません．つまり用いる分かち書きツールで分かち書き結果が変わります．これについての実験例が「資料１」の 63 ページ～79 ページにありますので参考にしてください．
⑤ セミナーでも確認した，閾値を変えて構成要素数を変えた場合にも同様のことが起こります．選んだ構成要素群が変われば分析内容結果は当然変わります．
⑥ よって，同じデータセットを用いても分析処理の結果は異なります．
⑦ 繰り返し指摘するように，ここらの議論はデータ収集法にも大きく関連することです．つまり過去の経験では（科学的に立証されているわけではありませんが)，構成要素数の変化，異なる分かち書き結果の利用，（若干）異なる内容の辞書を使った場合でも，分析対象とするテキスト型データに確たる潜在的な構造，あるいは何らかの規則性や特徴がある場合には，かなり類似した分析結果が得られることが多いことも事実です（ここに TM の利点があります)．例えば，セミナーで例として用いた「あなたにとって大切なものは」という質問は，調査方式（モード)，調査サイト（実施機関)，調査実施時期などの諸条件を変えて何回も実験調査した結果から，いずれも非常に類似した傾向にあることが分かっています．こうした積み上げで妥当性・再現性などを検証するということです．どのような研究分野，適用分野でも同様の問題が生じると考えております．
⑧ むしろ，いったんコーパスを確定したあと，それに関連度の高い他の諸要因（質的変数，

デモグラフィック要因，クラスター変数，その他の要因・変数情報など）のどれが有意に働くかを探索的に調べ，いわば「仮説発見的」に用いることが有効かもしれません．WordMiner はこれを目的とした解析ツールです．

Q19：辞書の作成が最も時間がかかり難しい（自由記述の量が多いので）．

A19：辞書作成を行うことは日本語の特性・特徴を考えると避けられない作業です．しかし，以下のような留意事項を念頭に分析を進めてはいかがでしょうか．

① まず大まかな分析を行って対象データに本当に期待するような構造，知見が含まれているのだろうかを探査する．つまり大まかな見当を付ける．このためには，構成要素数の分布で閾値を観察して比較的大きい閾値でフィルタリングして（出現頻度の少ない構成要素は除外して）その構成要素群を用いる．

② そしてデータ内に何らかの潜在的特徴が見られるようなら，丁寧に辞書作りを行ったうえで再分析する．

③ いったん作成した辞書は，類似課題（類似プロジェクト）あるいは類似の解析対象を分析する場合にも転写して使い回しを行う（辞書を共有化する）．この操作は WordMiner では簡単にできるので問題はない．

④ プロジェクトの内容に応じて，つまり似たようなプロジェクトについては，やはり共通辞書を作ってそれを再編集して対応する．

これに関連した若干の記述が「資料１」の「補足資料２」にありますので，それらもご覧ください．

Q20：置換辞書ソフトがあるというが具体的に何があるのか？

A20：残念ながら置換辞書ソフトはないと思います．ただし，経験的には以下のような手順を試みたことはあります．

① 市販の CD 化された類語辞典，シソーラス辞典などを用いること，ここから必要に応じて辞書内に語句をコピーする．

② 我々が過去の分析において作成した置換・削除辞書の流用が可能な課題内容なら，それを提供することは可能です．例えば，セミナーの例示に用いた，インターネット関連の質問，食べ物の好み（好きな食べ物），大切なものなどのような例を言います．

③ 「資料１」の 18～19 ページ，80～81 ページなどの参考文献欄に若干の情報がありますが，以下にいくつか挙げておきます．これらは一般の辞書ではなくいわゆるシソーラス，コーパス的な辞典です．

④ ここで「デジタル類語辞典」「日本語大シソーラス−類語検索大辞典−」は結構便利な辞典です．

- ジャングル（2006）：デジタル類語辞典（第 5 版）（類語・シソーラス辞典ソフト）．
- 国立国語研究所（1964，2004）：分類語彙表，国立国語研究所資料集，大日本図書．
- 山口翼編．2003．日本語大シソーラス−類語検索大辞典−，大修館書店（CD-ROM 版）．
- NTT コミュニケーションズ科学基礎研究所監修（1999）：日本語語彙大系，岩波書店（CD-ROM 版）．
- 柴田武，山田進編（2002）：類語大辞典，講談社（CD-ROM はなし）．
- 松井栄一編（2005）：日本語新辞典，小学館（CD-ROM はなし）．

6. 検索機能

Q21：検索段階で単語・語句の編集は可能か？

A21：残念ながら検索機能には編集機能は含まれません．

Q22：検索機能で「and，or」検索は可能か？

A22：これもできません．

7．その他の操作上の質問

Q23：コード形式の MA を質的変数にしたときの使い方を知りたい（MA 化したきにばらせないのですが）．

A23：これは既に **A6** で述べた通りです．ある程度のデータ加工を行うことで対応可能です．

Q24：変数生成後にデータビューアを見る方法は？

A24：データビューアはいつでも閲覧可能です．WordMiner では何か作業を行ったあと，結果はすべて新たに書き加えられます（変数生成，辞書処理後の変数，クラスター変数，質的変数，など）．ただしウィンドウ内に表示の分析結果はエクスポートで外部出力されます．データビューアで閲覧可能な範囲は無制限ではありません．

なお，データビューア上の内容はすべてテキスト・ファイル（csv ファイル）としてエクスポートが可能です（通常マウス右ボタンを操作）．つまり分析加工を行った結果を別の統計ソフトウェアに移して再分析を行うことが可能です．

Q25：表示フォントのポイント数が制御できない．

A25：これも多くの画面（ウィザード，ウィンドウ）上では制御できません．ただし成分スコアの布置図では表示のフォント･サイズやフォントの色，タイプを変更することができます．

Q26：行でわかれたテキストデータを複数同時に合成せずに分析はできるか？

A26：質問の意味が少し分からないのですが，以下のような操作をいうのでしょうか．

① 自由記述・テキスト型データ部が列（カラム）として複数あるとき，それを個別の分析する．例えば，調査データで自由回答質問が複数（質問 A，質問 B，質問 C，…）あるとします．このとき，個々の質問（A, B, C, …）を同時に分かち書き処理し，結果を個別に分析することは当然できます．また，複数の質問を併合して，例えば「質問 A と質問 C を併合」して，新たな一つの変数を作って分析することも可能です．

② 異なる変数，例えば上の例で「質問 A と質問 B」を同時的につまり分析の並列処理はできません．

Q27：データがファイルとして保存できない．他の PC に入っている WordMiner で見ることができない．

A27：マニュアル等の説明が不十分なようです．重要なご指摘です．WordMiner では他のソフトと異なり，拡張子「.wsp」のファイルをコピーするだけでは動作しません．しかし以下の手順で対応できます．

① あるプロジェクトを新規作成するとその名前のフォルダが作られる．例えば「プロジェクト A」を作るとその名称のフォルダ「プロジェクト A」が作られる．

② 同時にそのフォルダ内にフォルダ名と同じ名称で拡張子「.wsp」のファイル「プロジェクト A.wsp」と他のアーカイブ・ファイルが生成される（フォルダ NAA，NGG，NVV その他のファイル）．

③ 他の PC 上にコピーしたい場合はプロジェクト名のフォルダ，上の例ならフォルダ「プロジェクト A」全体をコピーする必要がある．このフォルダのサイズがかなり大きくなるの

で，コピー時には適当なファイル圧縮化ソフト（例：lhaca，StuffIt）あるいは Windows の圧縮化処理機能（zip 形式）を使って一度圧縮化しそれをコピーする．

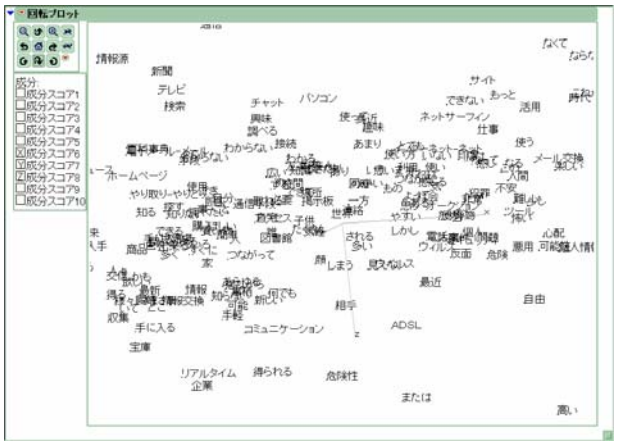
Q28：（布置図に単語を）マッピングしたときに単語・語句が重なってしまうので PowerPoint（に貼付）でグループ化を解除して単語・語句をばらしているが，その時の配置の仕方と読み込み（評価の仕方）はどう行うのか？

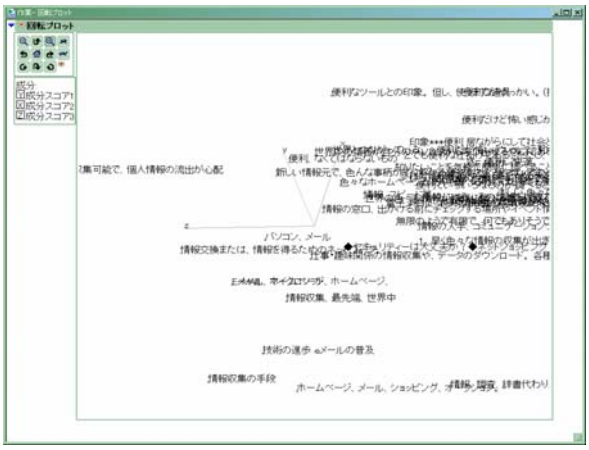
A28：具体的な操作が分からないのですが，確かにこのような操作が可能なようです．WordMiner では布置上の打点は次のようになっております．ここでは以下を回答とします．

① 主に成分スコアの布置図となっているので，その成分スコアが同じ値なら，どんな加工を行っても（WordMiner 上では）打点位置を変えることはできない．

② WordMiner のエクスポート機能を用いて成分スコアを外部ファイルとして出力し，それを他のソフトを使って加工編集する．その際に，同じ位置にある打点を意図的に加工して少しだけ位置をずらして布置図を作る．

③ 成分スコアに出力ファイルを用いて新たに図を描くときに，JMP（SAS 社）の利用をお奨めしております．散布図（2 次元布置図），回転プロット図（3 次元散布図，立体散布図）などを描くことが可能です．WordMiner の結果をエクスポートし，JMP で再分析した簡単な例を挙げておきます．しかしどのようなソフトを用いても，無数の文字の表示はやはり煩雑で視認性が悪くなります．





Q29：布置図で得られた結果を島状に分けた布置図に囲み（例：〇で囲むなど）を入れた図は得られるのか？

A29：類似する構成要素群を囲むとか，同じ区域にあるサンプルを囲むといった操作を言うと思いますが，WordMiner では対応できません．上の **A28** に示したように他のソフトを使って二次加工を行えば可能です．

なお，他社の TM ソフトには，対応分析法の結果やグラフィカル表現の中で，単語群を括ったり網掛けを行った図を描くなどのオプションがあるようです．これはこれで有効ではありますが，「大量のテキスト型データからマイニングする」という TM の本来の目標は満たされないのではないかとも考えております．どの程度の規模のデータセットが客観的に分析可能かということは，ソフトの性能やパーフォーマンスを判断する重要なキーワードです．

資料作成
テキスト･マイニング研究会　代表
大隅　昇（E-mail：ohsumi@ss.iij4u.or.jp）