補足資料 2

よくある質問とそれへの回答

大隅　昇
テキスト・マイニング研究会代表
統計数理研究所・名誉教授

ここにいままで活用セミナーにご参加いただいた方々からいただいたご意見やアンケートの内容を要約してあります．これをおおよそ次のように分けたうえで，それぞれに回答あるいはコメントを付けてあります．WordMiner をご利用いただくうえでご参考になれば幸いです．

1） 総合的なご意見へのコメント
2） セミナー開催の要領について
3） WordMiner で用いている方法論に関連したご質問
4） WordMiner の基本原理と設計指針は？
5） 構成要素数の観察と閾値の決め方
6） 有意性テストについて
7） 他の統計ソフトウェアとのデータ互換性について
8） 分かち書き処理と品詞分類の必要性
9） 論文引用，著作権，情報開示について
10） 研究分野，ジョブ課題との関連
11） その他のご意見へのコメント
12） セミナー開催者，WordMiner 開発者からの印象　—余録として—

以下，この順にご説明します．

1．総合的なご意見へのコメント

＜ご意見＞

・ 操作，使い方がよく分からない
・ 利用用語に慣れない，厄介・混乱する
・ 操作が煩雑，ボタン一つで実行などを期待する
・ 習得が大変

＜回答＞

前向きのご意見，ご理解もいただけましたが，厳しいご指摘も多々ありました．

とくに，操作性については，上に記したように，使い方が難しい，煩雑である，用いる用語に慣れていない，といったご意見がありました．「操作性」については，他の統計ソフトウェアや類似ソフトを併用し，客観的に評価せねば何とも申せませんが，我々（とくに大隅）の今までの体験・経験から次のように考えております．またこうした統計ソフトウェアの長年の利用経験も WordMiner の設計指針に反映させるよう努めてまいりました．

① いわゆる SPSS，SAS，JMP（SAS 社の PC 用デスクトップ・ソフト），Minitab，DataDesk，StatView，StatWorks（日科技研）といった統計ソフトウェアは，いわゆる量的データの処理，あるいは質的データであっても“数値化されたデータの処理”に適している．
② テキスト型データをそのまま扱うことから生じる種々の制約に対して，とくに扱い情報処理量が圧倒的に増えることから生じる制約をクリアせねばならない．こうしたことから，インターフェースの設計などがやや重くなる傾向にある．
③ テキスト型データを扱うことから，登場する用語・語句には，若干の「方言」が入ることはお許しいただきたいことがある．
④ WordMiner 特有の用語としては，構成要素，閾値（しきいち），追加処理，…などがあるが，これはマニュアルやセミナー配付資料を参考としていただき，ご理解をいただきたいこと（注 1 参照）．
⑤ 一方，「成分スコア」「固有値，寄与率」「寄与度」などは，多次元データ解析（多変量解析）の一般的な基礎用語であるのでなるべく関連書等を参考に慣れていただきたい．

（注1） セミナー時にコピーで配付の資料「WordMiner 事例集：導入編」をご覧いただくと，あちこちに「用語解説」のページがあります.

（注2） ここでとくに以下の用語の説明を付けておきます.

i） **構成要素：**インポートしたテキスト型データを，分かち書きし半角空白で分割した単位要素のことを示します．おおよそ単語・語句と考えてよいのですが，分かち書き結果が必ずしも「意味を持った単語・語句とは限らない」こと，記号類やいわゆるフェイスマークなど，また数字のそれを文字記号と読み替えればこれもテキスト型データということになります．WordMiner では言語解析的に分かち書きを扱うのではなく，より一般的にデータ解析の対象として扱うことから，構成要素と名付けております.

ii） **閾値：**構成要素を抽出する際の限界値のことです．デフォルト値（初期設定値）は「2」となっております．つまり，「2 回以上現れた構成要素」に相当します．したがって，閾値=1 と指定すると「すべての構成要素数（総構成要素数）」となります．これはその時の課題で扱う全コーパスと考えることができます．「補足資料 1」に記載の情報も参考にしてください.

iii） **異なり構成要素と異なり構成要素数：**異なり構成要素とは，同じ構成要素を重複なしとして計数したときの構成要素数のことです．例えば，助詞「が」が何十回も登場したとき，これを延べで計数するのが構成要素数であり，何度登場しても「1 回」と計数するのが異なり構成要素数です．多次元データ解析における「（サンプル・回答）×（構成要素変数）」「（構成要素変数）×（質的変数）」のデータ表では異なり構成要素数が基礎となります．ただし，有意性テストなどでは構成要素数，異なり構成要素数のいずれも利用します.

2. セミナー開催の要領について

＜ご意見＞

以下のようなご意見がありました.

- セミナーの情報ボリュームが多すぎる，内容が多すぎる
- 課題を絞って欲しいこと，話題を限定してほしいこと
- 知識追いつかず
- 内容を初級，中級，…と段階別に分けるとよい
- 分かり易く聞き手の立場であった，説明分かり易く勉強になった
- 地方でのセミナー希望する

＜回答＞

要は「十分な時間と平易な情報を用意すること」に尽きるかと思います．セミナーの開催形式と内容について，とくに「初級コース」「中級コース」「上級コース」と分けることへの要望が従来から多いことは承知しております．一方，現状のように十分なテキスト，資料を用意することが必要と考えておりますが，無料のセミナーとしてテキスト・マイニング研究会がこれに十分に支援・対応できる範囲も限られます．今後，以下のような対応策の検討を進める所存です.

① かって「初級」「中級」と分けて考えたこともあるが，そもそも何をもって「初級」「中級」他を分けるかの基準が設定しがたいこともみえてきたこと（参加者の知識の範囲，分散が大きいので即断出来かねること）.
② そこで少なくとも「操作説明と演習体験」と「数理的な内容の紹介」は，できれば日を別にして，例えば２日間の日程とし，分けて考えたいこと（検討中）.
③ その場合, 参加者にはどちらを選ぶか, あるいは両日参加とするかの選択肢を設けること.
④ ただし，現状の研究会スタッフのマンパワーでは東京以外での開催には無理があるので，しばらくは東京開催でご勘弁いただくこと.
⑤ そのための対応策として常時開設している「無料相談コーナー」を積極的にご利用いただきたいこと（ホームページからアクセス可：http://wordminer.comquest.co.jp/）.
⑥ また．お許しいただければ，セミナー配付資料については，若干の資料代をいただくことを検討したいこと.

3. WordMiner で用いている方法論に関連したご質問

＜ご意見，ご質問＞

成分スコアの意味，クラスタリング（クラスター化法）について知りたい，因子分析との関係などについて知りたいこと．

＜回答＞

対応分析法・数量化法 III 類については，なるべく具体的な解説・説明となるような資料を用意したつもりです．しかし，やはりあるところからは聞いていただくだけでは理解が徹底しないでしょう．そこでお奨めしたいことは，簡単な模擬データやミニチュア・データ，人工データ，それもテキスト型データの構造が見えるようなデータを用意して，実際に WordMiner を動かしてみることです．

このテキストに示した「レストランの評価」例や，その他の簡単な数値例を参考としてください．この他，WordMiner のプログラム CD 内に添付のサンプルデータ（実際の Web 調査の加工データ他）をご利用いただけます．説教めいて恐縮ですが，何事も「習うより慣れよ」ということでしょう．

なお対応分析法とは，（セミナーで何度も申したように）クロス表から誘導されるプロフィル（相対比率）のある種の変換データの主成分分析法に類似しております．また「因子分析法」とはまったく異なる方法論とお考えください．蛇足ですが，調査で得たような質的データに因子分析を適用することも（一般には無節操に使っているようですが）十分に注意せねばなりません（本来は使ってはならないと考えます，誤解があるようです）．

数理的は，いずれも「固有値問題」や「特異値分解」に帰着しますので，その意味では類似方法とは言えますが，対象とするデータの種類（つまり質的データか量的データか）で，用途が異なります．念のために記します．

量的データ（区間尺度，比例尺度）では　⇒　主成分分析法，因子分析が適用可
質的データ（名義尺度，順序尺度）では　⇒　対応分析法・数量化法 III 類が適用可

対応分析法は数量化法 III 類（正確には提唱者である林知己夫氏は「パターン分類」と呼んでおります）と同等手法と紹介いたしました．数量化法にはこの他に（飽戸弘氏が命名の）「数量化法 I 類，数量化法 II 類，数量化法 IV 類，数量化法 V 類，VI 類」まであります．

4. WordMiner の基本原理と設計指針は？

これは，どの範囲で，どのようにお答えすべきか，なかなかの難問です．

＜回答＞

まず，WordMiner は高度の言語解析や意味分析，構文解析（統語解析）などを行うツールではありません．また高度な内容分析も少し違った方向にあります．あくまでも，テキスト型データの探索的な多次元データ解析ツールです．多次元データ解析手法としては，セミナーで紹介されたように，基本要素として「対応分析法」と「クラスター化法」のみを含みます．また言語解析ツールとしては「分かち書き処理」のみ行います．品詞の分類や特定化は行いません（いわゆる形態素解析の細かい処理までは行いません）．

素性のよく分かった熟成した方法論を核として，周辺に分析に有効な機能，例えば有意性テスト，構成要素頻度情報の確認，種々の検索機能，コンコーダンス機能（KWIC）などを用意しました．

これらが具体的に何を行うかについては，セミナーで配布の資料類，それにテキスト・マイニング研究会ホームページにサイトアップされた情報をご利用いただくことで，ある程度はご理解いただけるものと考えております．

なお，立ち入った数理の詳細については，WordMiner を利用する以前の一般的な初等統計

学や多変量解析の入門的な知識を必要とすることがあります．この点では，多くの他のソフトと同様で，やはりある程度の基礎情報を入手・習得されることを期待し希望するものです．ここでは，こうした知識取得に必要な情報源（書籍）をいくつか挙げておきます．また，テキスト・マイニング研究会ホームページにも文献情報としていろいろ引用しておりますので，ご参照いただきご利用ください．

① ホーエル著，浅井晃・村上正康訳，「初等統計学」，培風館．
② 岩坪秀一著，「数量化法の基礎」，朝倉書店．
③ 大隅昇，ルバール他著，「記述的多変量解析」，日科技連．
④ Lebart 他著，*Exploring Textual Data*, Kluwer Academic Publishers.
⑤ Greenacre 著，*Theory and Applications of Correspondence Analysis*, Academic Press.

統計学の初等的な知識として，最小限①の内容程度は必要となります．またここで，②，③は既に絶版なのですが図書館などでご覧いただけると思います．対応分析法・数量化法 III 類についてはこれらに詳しい記述があります．④は，WordMiner のソースとなったプログラムの基本概念と応用例を説明した書です．⑤は対応分析法の一般的な数理を平易に説明した書です．この他，セミナーで配布のテキストにある文献や論文に目を通されることをお奨めします．

5. 構成要素数の観察と閾値の決め方

＜ご質問＞

・単語の出現頻度を知りたい．
・自由回答からのキーワード抽出は可能か？

＜回答＞

このご質問に対しては，すべて対処が可能です．しかし，重要なこととして，WordMiner は他のテキスト・マイニングを行うとするソフトとは異なり，

- 元のテキスト型データと分かち書き処理後で得た加工データの比較が完全に可能であること
- つまり総構成要素数，異なり構成要素数の分布を分析し完全に把握できること

があります．換言すると，他のソフトでは，分かち書きをどう行って，どのような分かち書き結果が得られたか，全情報を出力しないソフトが多い，加工済みの一部情報の提供しかないことが多い，ということです．

つまり，WordMiner では，得られたすべての構成要素情報の分析をユーザに委ねることから，構成要素の分析を始め辞書編集などの操作がやや煩雑となるきらいがあります．操作を簡便化し出力結果を容易に得るのか，多少の煩雑性があっても処理内容・過程を透明化し情報を詳しく示すこととするか，このトレードオフをどう均衡させるかは，設計指針にあります．WordMiner 開発時にも議論のあったことです．結果として，WordMiner は後者を採用，つまり構成要素（≒単語・語句）に関する処理情報をすべてオープンにするという指針を採用しております．

次に，構成要素数，異なり構成要素数の分布の観察方法ですが，これは補足資料 1「よくある質問へのヒント」として用意しました．しかし原則として以下にご留意ください．

① 閾値の法則的な決め方というのはない．
② 構成要素数，異なり構成要素数（率）の分布を観察すること，これらがどのような挙動をするかの例を補足資料 1 にいくつか示した，こうした情報を自ら作ってみるとよい．
③ 異なり構成要素数を一旦決めて分析（多次元データ解析など）を行ったら，それで一つの

解を得たと考えること．

④ 換言すると，異なり構成要素数を変えて分析するということは，別のデータセットを対象とした別の分析であるということ．

⑤ しかし，異なり構成要素数を変えても（布置図の構造や成分スコアの分布に）ある類似の構造が見えるならそれがそのデータセットの中にある潜在的な構造（規則性）があると考えられること（つまり情報のマイニング）．

⑥ そのようなことから，閾値を大きめにとって，徐々に少なくする，つまり構成要素数を少ない方から多い方へと変化させることが探査方法としてはよいかもしれないこと．

⑦ この意味では「探索的かつ発見的」であって，入学試験問題のように特定な決まった解があるようなわけにはならないこと．

これらについては「**補足資料 1**」として用意してありますので，ご通読いただきご活用ください．

6. 有意性テストについて

有意性テストについても補足資料 1「**よくある質問へのヒント**」に記述しました．従来からこれについての質問が多いからです．

有意性テストには「頻度による有意性テスト」と「距離による有意性テスト」があります．補足資料 2 に，主に「頻度による有意性テスト」について紹介しております．「距離の有意性テスト」については別の機会にあらためて資料を用意いたします．

「頻度による有意性テスト」については，簡単な数値例を入れて，各出力情報との関係を示しましたので，お読みいただくとおおよそのことがお分かりいただけます．もちろん，基礎となるいくつかの初等統計知識については統計関連書を併せてご確認ください．またここで示した例示は電卓とエクセルを使うことで簡単に追試できます．ご自分のデータによる分析結果についても，是非とも同様の方法でフォローアップすることをお奨めします．

7. 他の統計ソフトウェアとのデータ互換性について

これについてのご質問がありました．以下に要約します．

＜回答＞

① WordMiner では，セミナーでの説明でお分かりのように，ほとんどの表示画面でマウス右ボタン操作により結果データをエクスポートすることが可能です．これはテキスト・ファイル（csv 形式）ですので，他のソフトウェアでほとんど可読です．

② 他の統計ソフトウェアから WordMiner にデータをインポートしたい場合も，ほぼ同様に対応可能です．多くの統計ソフトウェアが，外部出力・エクスポート用のオプションを持っておりますので，テキスト・ファイル（csv 形式，タブ区切り形式）でエクスポート後に WordMiner にインポートできます．

③ WordMiner には分析対象データ，変数名，その他の扱い情報のほとんどに「文字数制限」がありません．一方，多くの統計ソフトウェアにはこの文字数制限があります．よって WordMiner のエクスポート・ファイルを他の統計ソフトウェアにインポートするときには十分な注意が必要です．

④ データ構造の比較的相性が良いソフトウェアとして SAS 社の JMP（ジャンプ）をお奨めします．JMP は最新のバージョン 6 となってから，扱い文字数の制約がかなり緩やかになって，WordMiner とのデータ授受が比較的円滑にできます．この情報については SAS 社のホームページをご覧ください（http://www.jmp.com/japan/corp/index.shtml）．JMP をご利用の場合，ファイル形式としてテキスト・ファイル（csv，タブ区切り）の他，エクセル・ファイルも直接授受できます．

⑤ WordMiner のデータ併合機能（コンカチネート機能），JMP の同様の機能を使って，複数のファイルの併合処理なども相互活用できます．

⑥ セミナーで申したように，数値データも（とくに質的データの数値化情報を），文字データと読み替えることでテキスト型データとして処理が可能，ということを念頭にデータ処理を考えるとよいでしょう．

8. 分かち書き処理と品詞分類の必要性

＜質問＞

WordMiner で品詞分類ができるのだろうか？

＜回答＞

前述のように，WordMiner では分かち書き処理の後の品詞分類までは行いません．もし所与のテキスト型データについて「品詞分類や品詞特定化」あるいは「その結果ファイル」を欲しいという場合は以下のような対応で解決できます．

① 形態素解析を行うツールを用いて，その結果情報を再編集し，WordMiner にインポートする場合．例えば，「茶筌（ちゃせん）」などの形態素解析を用いると品詞分類が可能です．その結果ファイル（テキスト・ファイル）をご自分で編集し，WordMiner で利用するというオプションがあります．

② この中間作業を効率化した「コーディング・ツール」にフリーウェアとして開示されている KH Coder を使うオプションがあります．KH Coder は樋口耕一氏の開発したツールです．ここでも形態素解析には茶筌を利用しておりますが，データを加工して WordMiner でインポート可能なフォーマットで出力するという機能を備えております（樋口耕一氏がそのように改良してくださいました）．これについては，下記のホームページをアクセスすると情報が得られます．

http://khc.sourceforge.net　または　　http://koichi.nihon.to/psnl

③ ここでの留意事項を挙げておきます．

- 品詞分類は一意的かつ確定的に厳密にはできないこと（未確定語の扱い，活用形の扱いをどうするかなど）．
- 分類結果がかなり細かくまた複雑になるので，再編集をどう考えるかがある（WordMiner にインポートしてからも）．
- WordMiner による分かち書きとは異なる分かち書き処理とその結果を利用できるという利点がある．
- つまり，一般に「分かち書き処理の結果」は使ったソフトに依存し，一意性がないということにも注意する．

9. 論文引用，著作権，情報開示について

研究者の方からのお問い合わせの頻度が高い事項に，この論文引用の方法があります．これについては，原則として以下のようにお願いいたします．

＜回答＞

① まず，テキスト・マイニング研究会ホームページの参考文献コーナーをご覧いただき，WordMiner 利用の論文などをご覧いただく（大隅，保田，その他のペーパー，報告があります）．

② それでも十分でないときは，セミナー配布の資料を引用していただく．とくにテクニカルな内容は，我々にとっては論文の種とはならないので，ホームページにサイトアップの研究会資料や技術解説などでの公開となります（いわゆるホワイト・ペーパーに近いでしょう）．これらを引用願います．

③ しかし，内容的に不明な箇所が出てきたときには，やはりお問い合わせいただくことがよ

いかと思います.

④ ソフトウェア内部の技術的なノウハウの部分についての開示は行うことはできません. 例えば，どうして規模の大きい非常に疎なデータ行列の計算が出来るのか，大量データの高速クラスタリングがどうして可能か，といったようなことがこれに該当します．ここらはセミナーでお話しする範囲に限定されます.

10. 研究分野，ジョブ課題との関連

具体的な分析課題名や研究分野とその対象課題，業務上の課題などについて多数のご意見をいただいております．しかしこれらについてここに具体的に記すことは適切ではないと考え，コメントは控えさせていただきます（いろいろ解決策はあります）.

ただし皆様が個々に抱えておられる課題の分析について，何らかのサポートが必要である場合は，「**無料相談コーナー**」をご利用ください．また我々にとっては，いただいたご意見はユーザの皆様がどのようにお考えになっているかを知る有用かつ重要な手がかりとなります.

実は既に，かなりの頻度で，この無料相談コーナーをご利用いただいております．具体的な調査設計を始め，WordMiner 出力結果の読み方，論文作成時のコメントや添削などがありました．このようなことで，忌憚なく，またお気軽に「無料相談コーナー」をご利用いただければ幸いです．とくに遠方の方の，このコーナーのご利用を歓迎いたします．これについてはテキスト・マイニング研究会のホームページをご覧ください.

11. その他のご意見へのコメント

その他，以下のようなご質問がありました.

＜ご質問＞

男女のように２カテゴリーのとき、一次元に圧縮可能だろうか？

＜回答＞

可能です．実際に試みてください.

＜ご質問＞

因子分析との違いはあるのか？

＜回答＞

上に若干指摘しましたが，対応分析と因子分析は異なる手法です.

＜ご質問＞

テキスト型データからの言葉の分類可能だろうか？

＜回答＞

「言葉の分類」の意味が若干曖昧ですので，的外れの回答・意見かもしれませんが，意味解析や（本当の意味での）内容分析などを指すのであれば，WordMiner が対象とする分野としては不得手といえるでしょう.

12. セミナー開催者，WordMiner 開発者からの印象　―余録として―

「知は力なり」というベーコンの言葉があります．ここで知＝knowledge とは日本語で言う知識よりはむしろ「**知恵**」と読み替えるべきではないでしょうか．流行り言葉のデータ・マイニングも含め，テキスト・マイニングなどの言葉に惑わされず，現象解析，現象解明のために本当に必要なことは何かを考える必要があるでしょう．我々の WordMiner 開発時の設計指針のスケルトンとして「**データ科学**（data science）」を掲げてきました．データ科学の３要素とは，以下を言います.

〇当該対象とする現象解明に必要な「実験の計画，調査計画」（experimental design）
〇それに合った「データ収集方式」（いわゆる data collection mode）の検討
〇そして取得データの分析のための方法論と支援ツールの開発（tool for analyzing）

ここで WordMiner がお役に立てるのはあくまでも分析ツールとしてのソフト支援環境の提供です．うまく使いこなしていただくための「知恵」は，上の３要素が一体となって始めて達成されると考えております．ごく平易な例を挙げますと，意識調査を行う場面で，適切な調査設計（実験計画），調査データ取得の方式（調査方式・調査モード）の検討，そして取得データ（回収データ）の適切な分析方法の三位一体が肝要ということで，このどれが欠けても適切な「マイニング」つまり宝の探査と発掘とはならないでしょう．

また，単語・語句さらには意味までを含めて，あるいは心理学的な意味での言葉の分類は，日本語言語学研究に様々なアプローチがあります．換言すると，確定的な方法論がないという現状では，高度の内容分析，意味解析，構文解析などはかなり難しい課題です．もちろん自分の仮説にあったような主観的な結論に導けばよい，というような立場もあるでしょうが，これではいかにも非科学的です．とにかく透明化されたデータ取得環境から得たテキスト型データを使って「事実」を客観的に探査するための支援ツールとするという WordMiner の設計指針がありますので，これを逸脱するような使い方は難しいでしょう．

インターネット，IT 礼賛の時代にあって，電子化されたテキスト型データとしてボリュームが多くなる傾向にあり，一見すると情報が豊富なように見えます．しかし量は情報の質を保証するものではありません．調査方式を例にとると，Web 調査が他の調査方式に比べて情報量が多いといった根拠のない意見があります．しかし，これに疑問を持って，ごく簡単な調査方式間比較や質問形式の比較実験を行うと，これが誤った意見ということが見えてきます．要は，何が「**事実**」であり，どれが「**正確な情報**」であるかを見極めることで，テキスト・マイニングに限らず何事にも慎重な対応が求められるということでしょう．すべては「データ」にある，では信頼できるデータをいかに取得するのかを常に考えること，WordMiner のキャッチコピーを“*WordMiner™ for Analyzing and Exploring Textual Data*”としている理由はこうしたことにあります．

いささか釈迦に説法となりましたが，今後も WordMiner を皆様のお仕事の良きアシスタントとして，ご利用ただけるよう重ねてお願いいたします．

テキスト・マイニング研究会・代表
大隅　昇