ISSN 0913-1442

理論と方法 36

Sociological Theory and Methods Vol.19 No.2 2004

特集　非定型データ分析の可能性

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—
大隅 昇・保田 明夫

計算機による新聞記事の計量的分析
—『毎日新聞』に見る「サラリーマン」を題材に—　樋口 耕一

ルールベース手法と機械学習による自由回答の分類
—職業コーディング自動化の方法—　高橋 和子・高村 大也・奥村 学

リレーショナル・データベースによる定型データの作成
—宗門改帳の統計分析のために—　中里 英樹

職歴パターンの分析 —最適マッチング分析による分析—　渡邊 勉

原著論文

「近い国・遠い国」 —多次元尺度構成法による世界認知構造の研究—　田辺 俊介

移動機会格差の変動分析：ロジスティック回帰モデルの応用　鹿又 伸夫

書評

『社会ネットワーク分析の基礎』
『講座社会変動　流動化と社会格差』
『フリーターという生き方』
『現代高校生の規範意識』

数理社会学会　　発売：ハーベスト社

理論と方法 36

Sociological Theory and Methods

Vol.19 No.2

目 次

特集 非定型データ分析の可能性 131

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－ 大隅 昇・保田 明夫 135

計算機による新聞記事の計量的分析
－『毎日新聞に見る『サラリーマン』を題材に－ 樋口 耕一 161

ルールベース手法と機械学習による自由回答の分類
－職業コーディング自動化の方法－ 高橋 和子・高村 大也・奥村 学 177

リレーショナル・データベースによる定型データの作成
－宗門改帳の統計分析のために－ 中里 英樹 197

職歴パターンの分析 －最適マッチング分析による分析－ 渡邊 勉 213

原著論文

「近い国・遠い国」－多次元尺度構成法による世界認知構造の研究－ 田辺 俊介 235

移動機会格差の変動分析：ロジスティック回帰モデルの応用 鹿又 伸夫 251

書評

『社会ネットワーク分析の基礎』金光 淳著 佐藤 嘉倫 265

『講座社会変動 流動化と社会格差』原 純輔 編著 中井 美樹 266

『フリーターという生き方』小杉礼子著 久木元 真吾 269

『現代高校生の規範意識』友枝敏雄・鈴木 譲著 佐藤 香 271

特集論文

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 －

大隅 昇
（統計数理研究所）

保田 明夫
（平和情報センター）

【要旨】

ここではまず，テキスト・マイニング（TM）あるいはテキスト型データのマイニング（TDM）の特徴を俯瞰すると同時に，これに関わる技術的な諸要素，諸事項について総合的に報告する．つぎに，現状考えられる TM を実際データの分析に用いるうえでの諸問題を整理する．とくに，その適用可能性について，データ科学の視点から問題解決を図ることの重要性について触れ，さらに具体的な TM 応用ソフトを紹介する．また，筆者等が独自に行った Web 調査データによる分析例を通じ，どのような使い方ができるかの要点，留意事項を示す．ここでは，自由回答設問で得た情報と通常の選択肢型設問との併用による定性型情報の計量的評価の例として示すが，これは TM のごく一部の具現化に過ぎず，本来の TM のあるべき姿，目標はこれだけではない．このようなことから TM の今後の進むべき道あるいは期待される方向は何かについての私見を述べる．

キーワード：テキスト・マイニング，テキスト型データのマイニング，データ・マイニング，知識発見とデータ・マイニング，定性調査，データ科学，自由回答設問，Web 調査，単語頻度クロス表の対応分析

1. まえがき

マーケティングや市場調査の分野では，CRM/eCRM を支援する有力な方法の一つとして，定性情報の有効活用が注目されている．とくに，顧客満足度の評価やコールセンターのシステム実装化過程で「顧客の生の声」「消費者の本音を知る」ための手段として，テキスト・マイニング（TM：text mining）あるいはテキスト型データのマイニング（TDM：textual data mining）が急速に拡がりつつある．また，社会調査（意識調査，態度調査）や，いわゆるアンケートに関わる利用者や研究者にとっても TM に高い関心が集まっている．

しかし，TM とは実はかなり曖昧な概念である．類似の言葉にデータ・マイニング（DM：data mining）がある．これも流行り言葉であり TM 同様に分かったようで漠としたものだが，関連書籍は無数にありコンピュータ・ソフトも多数登場している．これはまた，従来の統計的データ解析の各種方法論との違いがいまひとつ明らかではない．TM についても似たような事情にあると思われる．

ここでは，TM の特徴を俯瞰すると同時に，これに関わる技術的な諸要素，諸事項につい

て総合的に報告する．また，TM が何を行うか，ある調査データを用いた簡単な例を挙げる．なお，あくまで要約であるから個々の要素について詳しく述べることが目標ではなく，紙幅の都合もあるので大まかに概観する．また断るまでもなく，筆者等が考えるところを述べるものである．

1.1 データ・マイニングとテキスト・マイニング

TM は DM から派生した方法論であるとの記述がある．「鉱脈探し」（mining）という共通語からの類推であろう．確かに TM のある部分，とくにデータ処理や解析部エンジン（解析手法やそのアルゴリズム）については DM に類似したものがある．ここでどう類似し，異なるのかを知るには，まず DM とは何かを知る必要がある．前述のように DM については無数の研究報告や書冊がありこれらの個々の技法や方法論のすべてに言及することは難しい．ここでは DM の概念を概観し，続いて TM とは何かをみる．

DM は知識発見（KD : Knowledge Discovery）に関連づけて議論されることが多い．人工知能研究の一つの支流として，80 年代後半から 90 年代に入って登場した狭義の KDD（Knowledge Discovery in Databases）では，データベース上から知識発見を行う過程の中で，知識発見の方法論の集合体として DM が提唱されてきた．ここで KDD とは「データに潜在的に内在する，確かな，しかし予期しなかったような特徴の把握，有用で理解可能なパターンを特定化するプロセス」をいう．この狭義の KDD にデータ・マイニング（DM : Data Mining）が加わり今日の KDD （Knowledge Discovery and Data Mining）がある．つまり DM とは，知識発見過程において，データ解析，探索・知識発見操作（アルゴリズム）に相当する処理過程，また，検証，発見，予測，記述などの関連諸技法の集合体であり広義の KDD プロセスにおける解析部のエンジンの役割を果たすものである（Fayyad, Piatetsky-Shapiro 他（1996））．

ここで，従来からの統計的手法，統計的データ解析を知る者には，KDD との考え方の違いが見えてこない．DM の多くの関連書では，その違いを「統計的な分布の仮定がない，母集団概念などが不要」「扱うデータの規模・ボリュームが異なる」「整備されたデータベース機能やデータベース上のデータウェアハウスを用いる」等にあると主張する．しかし最近の統計的方法論は，これらに対する解決策はかなり提供されており，この主張だけで DM を特徴付けることは説得力がない．

膨大なデータセットの中から“金の鉱脈”を的確に探し当てる方法があるならそれに越したことはないが，現状の DM あるいは KDD 過程には思わぬ落とし穴がある．DM の多くの書に「ゴミを入れればゴミが出る」（GIGO : garbage in garbage out）とあるが，改めて考えると「ゴミではないデータはどこにあるか」「現象解明に有用なデータとは何か」との疑問にたどり着く．しかし我々は，DM の多くの方法論はこれへの答えはないと考える．“十分な量の適切で良質なデータ”があればとの前提で議論が展開されるだけでは真の現象解析からはほど遠いと考えざるを得ない．

一方，古典的な統計学では母集団を想定し実験計画や調査計画を厳密に構築し，サンプリング操作により分析対象（標本）を用意する．この厳密さゆえに，かえって現象解析に適した現実的なデータ取得環境が作れず，結果として数理の枠内の些末な議論となることもある．

いずれにせよ問題とする現象解明のための“目的に合ったデータ取得法”が必要であり，それを前提とした“データ主導型”の解析過程が必要であり，この点で統計的データ解析は明確なパラダイムがある．とくに「データ科学」(data science) はこれを発展的に考えるものである（林 2001)）．ここでは，現象解析の基本は「データ」にあり，「データを通じた現象理解」を前提とし，統計学，分類操作，その他の関連手法を背景に，統合的に現象解明を進める発展的な探索的データ解析（EDA）が重要との立場に立っている．要はデータから離れた議論では真の現象解明にはほど遠いのである．

2. テキスト・マイニングの背景

2.1 テキスト・マイニングとは？

TM のもっとも安易な定義は DM の亜種という見方である．人工知能研究の支流の一つとして DM が登場し，これらと言語学研究，自然言語処理研究などが融合して TM という支流が生まれたと考える．ステロタイプな言い方だが，ひとつ定義を挙げると以下のようなことである（Neri, Nahm, Ye (2003), Sullivan (2001) 他）．

【定義】

■ 大量のテキスト，自然文や自然言語テキスト（言葉の表記体），文書の集合体，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，隠れた意味のある類似性を発見し類型化する．またそれらを要約，視覚化し，理解可能な情報に変換するなどを行う一連の操作をいう．さらにその内容や情報を計量化し，その探査の推移を把握することから，新たな知見・知識を得る一連の接近法をいう．

■ 大量のテキスト，文書を数値化データと同様に自由に操作し（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．

■ 高度に構造化されたデータベースやデータ・ウェアハウス，ドキュメント・ウェアハウスから，顕著なパターンを発見するため，データ・マイニング技法，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，有用な知識，知見を引き出すことを目的とする．

定義はいろいろあるが，また表現は微妙に異なるが，以下のような“共通項”があることが TM の特徴である．

- 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[1]（コーポラ）の利用
- 規則性，類似性，パターンの探査，特徴付け
- 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
- 有用なパターンの発見

- 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
- 情報の知識化，知識の発見と取得

Hearst（1999）によると，TM のゴールはデータから新たな情報を発見し，データセット間のパターンを探査し，あるいはまた，ノイズから信号を分離することであるという．しかもその本質は，単に自然言語処理技術やテキスト要約，分類技術にあるのではなく，それらを利用した「探索的データの解析」に意味があるとし，ここでも事の本質が“探索的アプローチ”にあると主張している．

2.2 TM と関連する分野，方法論，そして適用の範囲

TM が対象とする“目標”はどの研究分野や関連分野に軸足をおくか，どこに焦点をあてるかで様々である．また学際的かつ広範な分野にまたがり，これといった厳密な制約や境界もない．例えばここで，関連研究分野から眺め，また TM で利用される方法論から眺めよう．

（1）関連研究分野からの観察

まず関連する主な研究分野として，自然言語処理，計算機言語学，人工知能（AI），エキスパートシステム，知識獲得・知識工学，情報検索（IR），情報処理，計量言語学，コーパス言語学，計量文献学，言語学，社会学，行動科学，記号論，テクスト論，カテゴリー論，意味論，内容分析・テキスト分析等，多彩である．さらにそれぞれの分野の諸要素が含まれしかも相互に絡み合っている．

研究の長い歴史がある「内容分析」（content analysis）も同様である．コンピュータ利用の内容分析（CACA： computer-assisted content analysis）が登場したのは半世紀近くも前だがそれ以前も様々な研究が行われてきた．文書情報管理・検索機能は重要で，例えば KWIC（keyword in context），コンコーダンス（concordance）により，語句の文章内での使い方や共起の関係を調べ，また共起語，コーパス頻度，共起頻度の閲覧や統計的指標などを観察する．CACA に関連した多数の（主に英語）コーパスやコンピュータ・ソフトがありこれを用いた言語情報処理が盛んである（中村（2003），Popping（2000），Neuendorf（2002））．CACA の成果，とくに KWIC やコンコーダンスによる語句の使い方や共起の関係などコンピュータ対応の諸機能等は TM を考えるうえで無視できない．

（2）利用される方法論からの観察

次に利用される方法論から TM を考えよう．当然，関連分野と方法論とは不可分の関係にあって厳密には分けられない．しかし「TM の解析部」の核となる方法・手法として考えると，パターン認識の各種方法論，各種統計的手法（特に，多変量解析，多次元データ解析諸手法），分類手法（判別，クラスター化，自動分類），社会調査の各種調査技法，自由回答設

問設計等，情報管理技法（IM），情報管理システム（MIS），文書管理情報処理技術（データベース技法，情報検索技術等），各種の視覚化・可視化の技法，グラフィカル表現法等がある．この他，遺伝的アルゴリズム，ニューラル・ネットワーク，複雑系，ファジイ理論，ラフ集合と，様々な方法論が利用され，これも実に多様である．

つまり，多様な分野の“技術要素の集合体”がTMの特徴であり，この点ではDMに同様である．TMという特定な方法論があってそれを用いるのではなく，諸分野の利用技術の特色を活かし，また方法論の利点を分析目的に応じてどう使いこなすかという「使い方」がTMを活用するための鍵である．「どんな方法を使うか」が先にあるのではなく，分析対象に応じてどのように「使いこなすか」が肝要である．

（3）適用範囲，応用の範囲からの観察

TMが関与する適用範囲も多彩である．報告書，研究論文に登場するアイテムを列記すると，テキスト・カテゴリーゼーション，ドキュメント分類，ルール探索と発見，概念抽出，関係の発見，テキスト分割，テキスト・文書の要約化，知識取得と理解，テキスト・ナビゲーション，Webへの応用（Webマイニング，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，顧客意見のマイニング），調査データの分析への応用（自由回答，自由記述），全文検索などあらゆることが対象となる．このように，本来のTMの応用分野は実に様々な分野に拡がっている．とくに，“構造化された”（structured）膨大な文書データベース，ドキュメント・ウェアハウス，コーパスを用いた知識発見のツールとしてTMがあり，これが一般にいうテキスト・マイニングであろう（Sullivan（2001），Ye（2003））．

しかし日本国内では，とくに市場調査，社会調査の分野では，調査データ（自由回答設問）の分析やコール・センターやコンタクト・センター等で収集した“非構造的なデータ”（unstructured data）など，限定された範囲の利用が多い．TM本来の利用法である大規模文書データベース，ドキュメント・ウェアハウスからのルール探索や発見，概念抽出，関係の探査といったアプローチは，研究として散見されてもビジネスや応用面での利用は少ない．

3. テキスト・マイニングをどう活用するか

3.1 定性調査におけるテキスト・マイニングの適用可能性

つまりTMの本来の目標は，大量の文書・テキストからの“有用な情報・知識発掘”にある．しかしここでは，我々が日頃関心を持ちまた守備範囲とする定性調査における適用可能性，例えば自由回答・自由記述データ，グループ・インタビューやフォーカス・グループ，談話分析（discourse analysis）等の定性情報から有効な知見を得る方法としての適用可能性を考える．対象をこの分野に限定して考えたとき，どのような視点で取り組めばよいか．つまり，「利用上の留意事項」「調査における利用法，活用法」「調査における自由回答設問方式」等はどうあるべきかを考える．このことは，国内の現状のTMツールが許容できる範囲（どこまで分析可能か）を考えることでもある．例えば，我々は現状のTMの守備範囲を次のよ

うに考えている.

① 当面の関心事は日本語の自然言語処理や，その関連研究にあるのではないこと
② 自然言語処理技法はデータ解析のために必要な前処理であり，必要最小限の力を注入すべきであること
③ 日本語の品詞[2]分類特定の正確性，語義の曖昧性の解消，正確な要約や分類までを求めない，あるいは現時点でそこまでを要求してもただちに達成が難しいと考えられること
④ テキストの意味のニュアンスの違いなどへはあまり拘泥しない，つまり高度な意味論的アプローチには限界があるし，本当に必要かを分析コスト，作業量から考慮すべきこと
⑤ しかし有用な知見や情報を得るためには，解析結果に客観的，科学的な解釈を与える必要性があること
⑥ そのためには，そもそもの「データ取得計画，取得法の研究」が重要であること（素性の分からぬデータセットでは，分かることにも限界がある），例えば，自由回答は何でも聞けばよいではなく，調査目的に合った構造化した設問構成の工夫が必要であること，さらには調査の企画設計までも考慮すべきこと

ここに指摘の各項は，当たり前のことのようにも見えるが，実は現状の TM を考えるときにもっとも欠落している部分と我々は考えている．何が，どこまでできるのか，そもそも TM が適用可能なデータであるのか，といった基本的な議論のないまま，TM の技術的側面ばかりが強調されているように感じている．

3.2 テキスト・マイニングが行うこと

ここで“日本語”テキスト型データの解析に TM を適用する際の考慮点について考える．既述のように，筆者等は TM で重要なことは，対象とする事象の解明に適したデータ取得法の設計と併せて考えることが肝要と主張したい．これを前提に TM プロセスで留意すべき事項は何かを要約する．

(1) 初動探査と前処理

データ解析すべてに共通することであるが，収集データセットの事前処理や初動探査，例えばデータランドリ(クリーニング)，論理チェック，単純集計による探査処理が必要である．また，必要に応じて大量データセットから一部データを抽出する情報検索機能やサンプリング操作を用いる．統計手法の利点はデータに内在する規則性や法則性の探査にあるが，一方，例外やはずれ値的データを見抜くことが不得手である．TM の課題として，ここをどう処理できるかに留意すべきである．

(2) 形態素解析と統計処理

日本語テキスト型データ処理の最大の課題は「分かち書き処理」である．日本語は言語類型論により形態的特徴で区分すると膠着語とされる．膠着語とは単語の前後にさらに別の単

語を付けることができるということで，単に連なって切れ目のない語の並び，いわゆる「べた書き」という意味だけではない．切れ目がないという意味では中国語もそうであるが，中国語は孤立語に分類される．

現代日本語の特徴の一つは，漢字，仮名（カタカナ，ひらかな）交じりで記述されることである．混用は「くぎり」を示す役割を果たすので視認により意味解釈の誤解が避けられるという利点もある．しかしコンピュータにとっては，この「くぎり」が難問となる．欧米語と異なり，語句・単語が連なった「べた書き」は，解析時の処理単位が明らかでなくそのまま扱うことができない．欧米で開発されたTMツールが日本語処理にそのまま転用できない理由の一つがここにある．そこで，ある要素単位に区分する「分かち書き処理」が必要となる．さらに必要に応じて形態素解析を行う．形態素（morpheme）とは「意味をもつ最小の言語単位」をいう．形態素解析とは，所与のテキスト（文）を形態素に相当する要素単位に分解し，その個々の要素の文法的属性（品詞や活用など）を，辞書を用いて特定することをいう．その結果を用いて，語句・単語の頻度別集計，異なり単語数の集計，品詞分類集計などの統計処理を行う．分かち書き処理を含む形態素解析のツールは多数あって処理方式も様々である[3]．つまり“同じテキストを用いても形態素解析の結果は同じとはならない”．また完全な分かち書き処理（正確に形態素分解する）ができるとは限らない．つまり出発点が異なるデータセットを用いたデータ解析から同じ解が得られるとは限らないことに留意すべきである．多くの場合，TMの分析結果に，これら基礎情報の説明がないことは結果解釈の信頼性を損なうものであり，分析者は報告に際してこれら情報を明らかにする必要がある．必要に応じて，形態素解析だけでなく言語的知識（辞書，語彙，文法）と非言語的知識（一般常識，専門知識，スキルなどのセマンティックな要素集合）との支援を受けて，統語解析（構文解析），文脈解析なども行う．TMはこうした技法体系の一部を利用している．

（3）多変量解析，多次元データ解析

DMと同様に，TMでは解析部の方法論にパターン認識や統計的手法（多変量解析，多次元データ解析）を多用するが，大抵はソフトの内容・分析手順が具体的に開示されることがないので正確なことは分からない．特異値分解（SVD）・スペクトル分解系のモデル（主成分分析，対応分析・数量化III類等），回帰分析型手法，多次元尺度構成法（MDS）等が利用される．扱うデータセットのサイズや項目数・変量数，語句数などは膨大かつ高次元となるから，次元縮約や節約の原理を目標とするこれら手法が有効とされるのである．

（4）分類手法（クラスター化，自動分類，判別手法）

クラスタリング手法はTMにとって必須である．各種クラスタリング手法（階層的，非階層的など教師なし分類），判別手法（あるいは教師あり分類），SVM（サポート・ベクター・マシーン）などが利用される．非階層的分類ではk-平均法（k-means法）やその変型手法が多用される．また，DMとの関係では，分岐型階層的分類法のCART（二進木解析）やCHAIDなども頻用される．

多変量解析や分類手法では，モデリングに関連しニューラル・ネットワーク，遺伝的アル

ゴリズムなどの利用も盛んである．統計ソフトウェア開発企業にとっては，既存の技術資源を核に，データベース機能や機械学習型機能を付加することでDMツールに衣替えして提供できる素地がある．例えばEnterprise（SAS社）やClementine（SPSS社），STATISTCA-Text Minerなどをみれば明らかである．［表1，表2も参照］

（5）情報の要約化と視覚化

これもTMにとって重要な機能である．かりに，テキスト型データに潜在する漠然とした特徴，傾向，関係，パターンを探査できたとして，それらを理解が容易な形に視覚化することは有効である．しかし，視覚化操作に過剰な期待を持つことには危険がある．視覚化した情報に“客観的な解釈”を与え知識抽出に有効な指針を“具体的に示す”ことがどこまで可能かを常に問うべきである．

これは統計ソフトウェアの視覚化情報と比べると分かり易い．統計ソフトウェアでは各種統計量指標の算出と同時にグラフィカル表現を用いて，統計指標の意味解釈の助けとする．一方，TMではテキスト情報を扱うことから，この視覚化と分析指標の対比や客観的解釈を与えるための手当が十分とはいえない（どのように計量化されたか，それがどう視覚化されたか）．ここをどう解決するかが今後の課題である．

Kohonenの提案したSOM（自己組織化マップ：Self-Organizing Maps）も良く利用される．SOMはテキスト型データだけを対象とした分析法ではないが，Webマイニング等に関連してSOMを適用する例が増えている（Lagus他（1996），川端・樋口（2003），Murtagh（1999），Sullivan（2001））．なお視覚化過程での検討事項として以下を挙げておこう．

- 視覚化情報に客観的な意味づけ，解釈を与えられること（意味ある視覚化とは）
- 数値情報あるいは計量化情報を的確にグラフィカル表現すること
- 本来は数値化できない仮想的あるいは概念的な情報を可視化すること
- 膨大なテキスト情報からいかにして適切な視覚化が可能か，例えば無数の単語の布置図を観察しても解釈は容易ではない（知識取得に即座に結びつかない）
- つまり情報縮約化や要約化を行った上で視覚化処理を行うべきである
- 要約や縮約化に伴う情報損失をどう客観的に評価するか，ここで縮約の方法を誤ると，誤った解釈を与えることになること

現状のTMツールは，視覚化の目的や意味解釈の方法が総じて明らかではない．これはTMの重要な目標であるのに，機能の設計指針が曖昧であり客観的な意味解釈を与える情報に乏しい．なお，後で適用の限界や注意点を示す視覚化の簡単な例を示そう．

（6）辞書の機能，その周辺の課題

これも日本語TMにとって重要な要素であるが扱いが厄介な事の一つである．多くの場合，形態素解析や分かち書き処理を行うために辞書を備えている．しかも分析対象には非構造的なテキストが多いことが問題である．

高度なコンピュータ化が進んだコーパス，構造化された文書データベースやドキュメント・ウェアハウスを利用できる場合は，かなり的確な分析結果が期待できる（例えば新聞記事情報のコーパスなど）．TM の本来の対象はこうした構造化されたテキスト・データ集合を対象とした方法論が多いので，非構造的なテキストである自由回答・自由記述文の解析では様々な問題が生じる．

一つは，同義語・類語（シソーラス）である．表記や表意の違いがあって同じことを意味する語句をどう扱うかは難題である．例として筆者の関与した Web 調査で取得の自由回答データを考える．設問をどんなに工夫しても回答の内容は様々である．「友人」を「友達」「友」「ともだち」「とも達」「だち公」「仲間」「と・も・だ・ち」…と書き，「夫」「ダンナ」「旦那」「旦那さま」「パパ」…と記すという具合である．状況によっては広義語や関連語等の整理や関連付けの検討も必要である．「家族」を，「ファミリー」「身内」に括り，さらに「親類」「親族」「血族」「縁者」「父母兄弟」…とあるから，どこまでを類似語句としてまとめるか判断に迷う．さらにケータイ語，電子メール語，チャット語と様々で，同義語・類語の処理を厳密に考えること自体に無理があるし，はたして厳密な操作が必要かも考えねばならない．

TM ソフトウェアの側から考えると，ユーザがこの問題をどう理解し，要求内容がどの水準にあるかを知らねばならない．シソーラス辞書が整備できるか否か，ユーザがどこまで辞書編集を行うのか，同義語・類語・関連語の処理を考えなくても解析が可能か，解析のどの段階で利用できるかをユーザは知るべきであるし，ソフト提供者はそれらの情報を明示すべきである．

別の課題として語彙・コーパスをどう考えるかがある．語彙とはある一定の範囲で使用される単語，語句の集合体をいう．一定の範囲とは，ある作家の作品，個人の利用範囲等をいう．もっとも大きな括りは「日本語語彙」があり，小さなものでは個人の日記などがある．また，言語生活を営むうえで必要な基本的な語彙を基本語彙という．類似テーマを同一パネルに繰り返し調査し自由回答データを取得するとき，同一テーマで異なる調査対象に意見を聞くとき等の場面では，得られた回答には同じような使い回しの語句や単語が登場する．このようなとき，整備されたコーパスや，それを目的別に複数集めたコーポラがあると便利である．TM ソフトウェアはこうした機能を備えるべきであろう．また，CD 化されたシソーラスやコーパス，辞典・事典類を補助的に使う工夫も必要である（デジタル類語辞典 2003，日本語語彙大系，類語大辞典，類語検索大辞典日本語大シソーラスなど）．

3.3 適用の範囲からみたテキスト型データの分析の難易度

筆者の少ない体験からも，TM が“汎用的に”様々なテキスト型データに適用できるとは考えていない．つまり多くの TM ツールがうたっているように，多目的に利用できるものではないということであり，多くの場合，分析目的に応じたカスタマイズが必要であることを意味している．

理由の一つは，日本語分析の困難性にあると考えられる．換言すると，TM を有効に活用するには，それなりの「データ取得法」を考えるべきということである．また，既に集積化されたテキスト型データの分析を行う場合には，以下に示すように，その対象データが，ど

のような段階，様相にあるかを見極めたうえで対処すべきである．

［テキスト型データの多様な様相］

(1) 単に集めただけのテキスト・データ

サンプル・調査対象の背景やデータ取得状況や素性，取得目的があまり明らかでないデータ，多くの場合，分析が厄介で，有益な知見も期待しにくい．

(2) 元来が文字情報であるとき

これには，文学書・文芸書，新聞・雑誌，各種の記録文書などがある．コーパスなどの利用も比較的可能であり，TM の対象としては扱いやすい．

ただし，分析目標は，全文検索，文書分類，要約化処理，表現法の比較，記事分類，ドキュメント・マイニングなどである．

(3) 過去の蓄積データの見直し・再評価

“再発掘”等の過程を経て取得したデータ，例えば蓄積されていたアーカイブなどに付帯情報，データ取得履歴を付加し整理可能なデータ，蓄積した定性情報データベース，メタ・アナリシス，複数のデータベース情報の併合利用などがある．

(4) 調査データ，とくに選択肢型設問との併用

調査データに限って考えると，選択肢型設問等と併せて用いる自由回答設問がある．マーケティング，市場調査などで，もっとも多いと思われるタイプである．

(5) 計画的に設計された取得環境から収集のデータ

テキスト型データの取得を主目的として調査設計された中で取得のデータ．自由回答取得を主目的として設計された調査や特定の商品ユーザのモニター形式の継続的調査など．

このように，扱うデータの様相が様々であることが，数値型データを扱う通常のデータ解析とは根本的に異なることである．しかしながら一方では，現状の TM ツールを用いる限り，テキスト型データという本来の特性を，ある形で“計量化・数量化した”うえで，(従来型の) データ解析方法論を適用することが多いということを忘れてはならない（後述のように，この意味で真の TM とは何かを考えるべきである)．

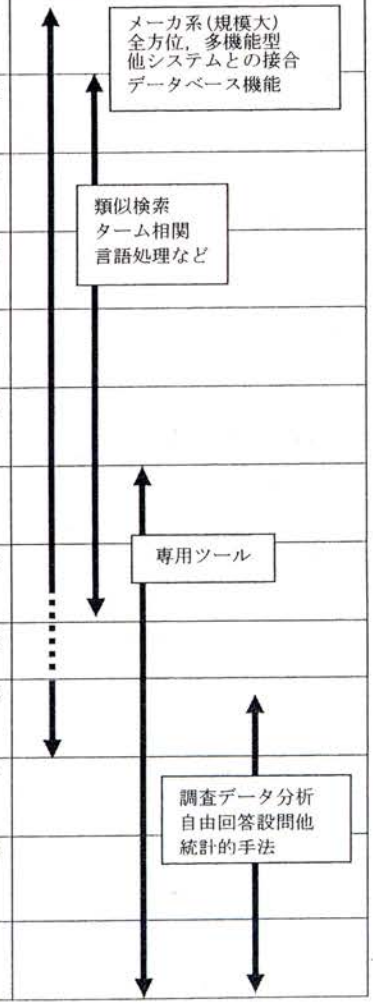
4. テキスト・マイニングのソフトウェア

TM ソフトウェアが備えるべき要件を知ることは重要である．例えば大項目としては，拡張可能性（スケーラビィティ)，分析対象資源やテキストの適用可能範囲，既存システムとの互換性，更新サービスの充実度，テキストの要約化・視覚化機能，解析機能の充実度，辞書機能，多言語対応の可能性，価格と処理機能の関係（コスト・パーフォーマンス）等がある．個別の詳細機能についてはここでは省略する．

TM ソフトウェアは国内外ともに無数にある．とくに国内ではここ数年の間に次々と登場した．例えば表１，２に我々の知る範囲を要約した．また，備える機能，分析対象の守備範囲を表中に書き入れた．

表1　主要なテキスト・マイニング・ソフトウェアの一覧（国内）

No.	製品名 (†)	開発元・販売元	特　徴	守備範囲
1	Symfoware Text Mining Server テキストマイニングソフトウェア	富士通(株)	キーワード間の関連性をビジュアルに表示する「コンセプトマッパー」。OLAP 製品と組み合わせ使用可能。 http://software.fujitsu.com/jp/symfoware/products/textmining/	
2	DocumentBroker 文書管理基盤	(株)日立製作所	ターム（単語・語句）の共起関係による相関分析・分類、自然文検索、概念検索など、統合的文書管理システム http://www.hitachi.co.jp/Prod/comp/soft1/docbro/	
3	TAKMI テキストマイニングシステム	日本アイ・ビー・エム(株)	概念（キーワードとなる文字列とそのカテゴリー）を抽出し、定型情報と共に統計量を計算・結果表示。 http://www.trl.ibm.com/projects/s7710/tm/takmi/takmi.htm	
4	Knowledge Meister ナレッジマネジメントシステム	(株)東芝	キーワードの出現頻度・関連度によるクラスタリング、依存・品詞分析によるテキスト・マイニング（要因分析） http://cn.toshiba.co.jp/prod/km2/function/mining.htm	
5	Knowledgeocean(ナレッジオーシャン) ナレッジマイニング支援システム	(株)NTT データナレッジ	コンセプト（主要語、概念）の抽出によるコンセプトの共起分析、クラスタリング、類似文書検索 http://www.knowhowbank.com/html/sol/sol_kso_1.html	
6	MiningPro21 文書マイニングシステム	日本ユニシス(株)	単語の相関度による文書分類、連語抽出・判別関数による文書判別、日本語文章による類似文書検索 http://www.unisys.co.jp/MP21/bunsho/	
7	CB Market Intelligence テキストマイニング・ソリューション	(株)ジャストシステム	意味認識手法（自然言語処理技術がベースのテキスト分析技術）による主題・評価・感性・機能要求分析 http://www.justsystem.co.jp/cbmi/	
8	VextSearch テキストマイニングツール	クオリカ(株) （旧コマツソフト）	コンテキストベクタ（似た文脈の中で用いられる単語のベクトルは似た方向を持つ）方式による知識モデル生成 http://www.qualica.co.jp/develp/vextminer/	
9	DE-FACTO	電通リサーチ	発想支援ソフト，テキスト型データから単語・語句の関連性を重要度に応じて類型化し，視覚化する.	
10	Survey Analyzer(サーベイアナライザー) ［Topic Scope として改編された］	日本電気(株)	確率的コンプレキシティ（統計尺度）に基づき、分析対象と結びつく固有の言葉や語句を抽出・発見 http://www.nec.co.jp/press/ja/0110/0502.html	
11	Text Mining for Clementine（LexiQuest） テキストマイニングツール	エス・ピー・エス・エス(株)	コンセプト（意味ある言葉の組み合わせ）の抽出。データ・マイニングツール Clementine のプラグインツール http://www.spss.co.jp/product/cle_text/text.html	
12	TRUE TELLER(トゥルーテラー) 統合型テキスト・マイニング分析システム	(株)野村総合研究所	係り受け(主語－述語)構文解析、話題・因果関係マッピング、文書スコアリング、分析結果の EXCEL 出力 http://www.trueteller.net/	
13	WordMiner(ワードマイナー) (‡) テキスト型データ解析ソフトウェア	日本電子計算(株)	構成要素（語や語句）抽出による多次元データ解析（対応分析、クラスター化）、コンコーダンス（用語検索） http://www.jip.co.jp/bs/products/Shohin/52/kihon/kihon.html	



（†）製品名等は、各社の登録商標もしくは商標　　（‡）WordMiner は筆者等のグループが開発したソフトウェア

表２　欧米のテキスト・マイニング・ソフトウェアの例

No.	製品・サービス名	開発元・販売元	特　徴
1	Sphinx Survey Plus2 & Lexica	Le Sphinx Développement SCOLARI http://www.scolari.com http://www.pugh.co.uk/Products/scolari/surveyplus.htm http://www.lesphinx-developpement.fr/	・調査データの集計・分析を主とする ・内容分析，文脈分析を行う ・多変量解析（主成分分析．対応分析など）
2	SPAD.T (Système Portable pour l'Analyse des Données-Donnée Textuelles)	L. Lebart（ENST）とそのグループ	・記述的・探索的ツール ・調査データ（自由回答など）の解析を重視 ・選択肢型設問とのクロス分析 ・多変量解析（対応分析，クラスター化） ・単語・語句の有意性テストによる特徴抽出 ・コンコーダンスによる単語・語句の利用パターン観察 ・WordMiner™の元となったソフト
3	WORDSTAT (V4.0)	Provalis Research Inc. http://www.simstat.com/home.html	・内容分析を主とする ・統計ソフト SIMSTAT ，CodeMiner にリンク (*)CodeMiner: Qualitative Data Analysis Tool
4	STATISTICA Text Miner	StatSoft Inc. http://www.StatSoft.com/ http://www.StatSoft.com/textminer.html	・統計ソフト STATISTICA と併用（add-on），統計処理機能の利用（PCA, k-means クラスター化，その他のデータマイニング） ・STATISTICA に渡す前の事前処理 ・種々のテキスト・フォーマットに対応 ・削除機能とそのルール，stub-list の生成 ・stemming algorithm の適用 ・多言語対応（オランダ，ドイツ，英語，フランス，イタリア，ポルトガル，スペイン，スウェーデンなど） ・文章要約化の機能 ・SVD（特異値分解）による特徴抽出
5	Text Analysis	MEGAPUTER Inc. http://www.megaputer.com/	・セマンティック・テキスト・マイニング：キー概念と非構造的テキスト型ノードとの関係から意味論的（セマンティック）分析を行う ・Link Analysis を使って，意思決定に役立つような視覚化を行う
6	WEBSOM	Helsinki University of Technology http://websom.hut.fi/websom/	・ドキュメント探査ツール，視覚化ツール ・Self-Organizing Maps（SOM）を使う ・Kohonen が主催するグループの研究公開

欧米の TM の評価や比較検証については，多数の報告がある．とくに，U. Nahm には TM に関する総合的な紹介サイト，24 のテキスト・マイニング・プロダクツのサイトへのリンクがある．

また「内容分析」の歴史は古く，コンピュータ利用もかなり早くから始まっているので多数のソフトがある．Roel Popping (2000)には，38 のソフトの紹介（かなり詳しい説明，評価）がある．Kimberly A. Neuendorf and Paul D. Skalski (2002)では一つの章を割いて，「Paul D. Skalski , Computer Content Analysis Software, pp325-239」とし，20 のソフトの紹介，評価説明を行っている．また，Robert P. Weber (1990)にもソフトウェアと利用可能データアーカイヴの簡単な紹介がある．

5. 簡単な分析例

5.1 用いる調査データの概要

我々が行った実験調査のうち，調査方式としてインターネット調査（Web 調査）を用いた調査データがある［調査の詳細は大隅他（2002，2004）を参照］．これの一部を用いた分析例を示す．

【調査概要】
調査実施時期：2002 年 5 月 16 日～5 月 23 日
調査方式：インターネット調査（Web 調査）
調査実施機関：電通リサーチ（DENTSU_R-net を利用）
計画標本数：1,542（サンプル），有効調査依頼発信数：1,512（サンプル）
有効回収標本数：894（サンプル）（有効回答率：59.1%）

この調査でいくつか自由回答設問を用意した．この一つを分析対象として取り上げる．

［用いた設問］

Q3. 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、どのようなことがらに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、どのようなことがらに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。

この 2 つの設問は，回答者と「インターネット」とのかかわりについて尋ねた設問である．Q3-1 は，「回答者自身」にとってインターネットがどのように活用できるかを，また，Q3-2 は，「一般的に」インターネットがどのように活用できるか，他にはないような活用法の提案を求める設問となっている．

ここでは２つの設問のうち，Q3-1 を用いた分析のいくつかを例示する．なお，この２設問が意図するところは「２つの設問の差異を表す特徴的な語句，単語は何か」「属性，とくに性別，年齢区分などの違いが回答に現れるのか」（設問文の違いが自由回答にどう現れるか）を知ることにある．ただここでは，TM ツールがどのようなことを行うのか，あるいは結果の情報要約がどのように得られるかといった技法的な側面から例を紹介する．なお用いるデータの自由回答書き込み状況は，自由回答書き込み数「回答記入あり」が 878 サンプル（98.2% 書き込みあり），うち分析に有効なサンプル数は 848 サンプル（94.9%）である．

5.2 分かち書き機能の比較

既述のように，用いるソフトによって分かち書き処理の結果は異なる．これは自明のことでありながら，ほとんど関心が持たれることはない．また，ソフト間の相互比較を行う機会も少ないのではなかろうか．ここでは，同じ自由回答文について４種のツールを適用し，分かち書きで得た単語群の頻度集計の一部について比較する．

表３　分かち書きソフトウェアの比較例（４種のツールの比較）

頻度順位	各ソフトウェアの単語・語句の頻度順位（上位 20 位）			
	WordMiner	ソフト D	ソフト V (†)	ソフト S (‡)
1	情報	情報	情報	情報
2	情報収集-情報集め	収集	収集	収集
3	事-こと	調べる	でき	仕事
4	できる	仕事	調べ	趣味
5	する	検索	検索	検索
6	趣味	入手	入手	調べる
7	電子メール-メール	ショッピング	趣味	メール
8	検索	メール	仕事	ショッピング
9	仕事	趣味	メール	入手
10	して	友人	知り	友人
11	友達-友人	連絡	ショッピング	連絡
12	もの	旅行	連絡	できる
13	調べる	得る	友人	する
14	入手	手段	旅行	上
15	等	コミュニケーション	手段	知る
16	いる	買う	なり	得る
17	コミュニケーション	自分	コミュニケーション	コミュニケーション
18	収集	活用	あり	買い物
19	買い物	ニュース	ニュース	旅行
20	連絡	予約	活用	ニュース

(†) ソフト V は分かち書きに茶筌を利用
(‡) ソフト S は分かち書き結果の単語集計機能がないので WordMiner で再集計した

表の内容は，分かち書き後の単語・語句の頻度集計の結果，出現頻度が高いものからそれぞれ上位 20 位を一覧としたものである．同頻度もあったが，ここは自動的に上から 20 位で切った．また，具体的に正確な頻度数が算出できない（その機能がない）ソフト S については分かち書き相当の出力情報を用いて WordMiner™で順位付けした（WordMiner™は我々が開発した市販ソフト）．ソフト V は分かち書きに「茶筌」を用いているとあったが，別途茶筌を用いて同じデータセットを分かち書きしたが，同じ結果とはならなかった．何らかの処理（再編集など）を行っているものと思われる．

また，WordMiner™とソフトDの結果は，分かち書き処理の後にごくわずかの編集作業を行った．また，ソフトV，ソフトSは，実はどのような編集機能があるかがあまり判然としないので，ほぼ分かち書きのままとした．分かち書き処理後の編集機能の内容は日本語テキスト型データを扱ううえでの重要な要素であるが，一般に不透明でありこのことが問題でもある．

表の内容の観察結果は説明するまでもないだろう．類似の単語・語句は登場するが，かなり異なる結果となっている（頻度数までを示すと，さらに違いが顕著である）．分かち書きはさらなる分析，とくに多次元データ解析などの高度な分析の出発点であるから，このように出発点ですでに異なる結果を得るということを知って対応することが重要であるが，多くの場合ここが明確に呈示されないことも事実なのである．

5.3 情報の視覚化の例

同じ自由回答設問への回答データを用いて分析を進め情報を視覚化した例をみよう．図1は，ある大手広告会社（電通と電通リサーチ）が独自開発した DE-FACTO™による出力情報の一部である．なお，開発社（者）によると，DE-FACTO は TM ツールというよりも「発想支援ソフト」の一つと位置づけている．しかし TM ツールとしての基本的な機能は十分に備えており，マーケティング情報として有効に活用できるよう独自の設計思想に基づいて作られていることが特徴である．図1は，そうした独自機能の一つで，抽出単語・語句間の関係を単語マップという張り木（spanning tree）で表した図である．ここには 49 語の単語をもっとも重要なキー「情報」を中心に位置づけ，各単語の関係を張り木で表してある．また，結線の太さが関連の強度を表している．精密な分析よりも，回答の中にある主要な単語・語句をとりあえず見やすい構造に表すという発想である．

次に示す例は，多次元データの特徴を2次元イメージとして描画する SOM（自己組織化マッピング）である．この種の方法論は，初期データ（入力データ）として何を用いるか（適用可能か）がいささか曖昧なこともあって，いい加減な使い方をされているようにみえる．また多くの指摘があるように，SOM の基本原理は，分類手法として知られる k-平均法や ISODATA（Iterative Self-Organizing Data Analysis Technique-A）に類似した手法である．従って，原則としては量的データに適用することが妥当と思われる．

ここでは，各回答者の発言単語のパターンの分析に注目して，つまり「(回答者) × (単語・語句群)」のクロス表に対応分析を適用して得られる数量化スコア（成分スコア）を使って単語・語句の分類マップを作る．具体的には，単語の出現頻度が 10 以上の単語 157 語を採用し，それに有効な回答数が 844 サンプルから得た大きさ「(844 サンプル) × (157 語)」の2元表の対応分析を行い，個々の単語に付与された数量化スコアの初めの 15 成分を用いて SOM を行った（Viscovery を利用）．SOM のパラメータの細かい設定条件は省略し，およそ 10 前後のクラスター数で学習を行った結果が図2である（もちろんパラメータ設定条件によって解は様々である）．図の左上に「情報授受」に関係する語群が，また検索の機能の使い方の関連する多くの語句が，それぞれある位置関係をもって布置されているようにみえる．

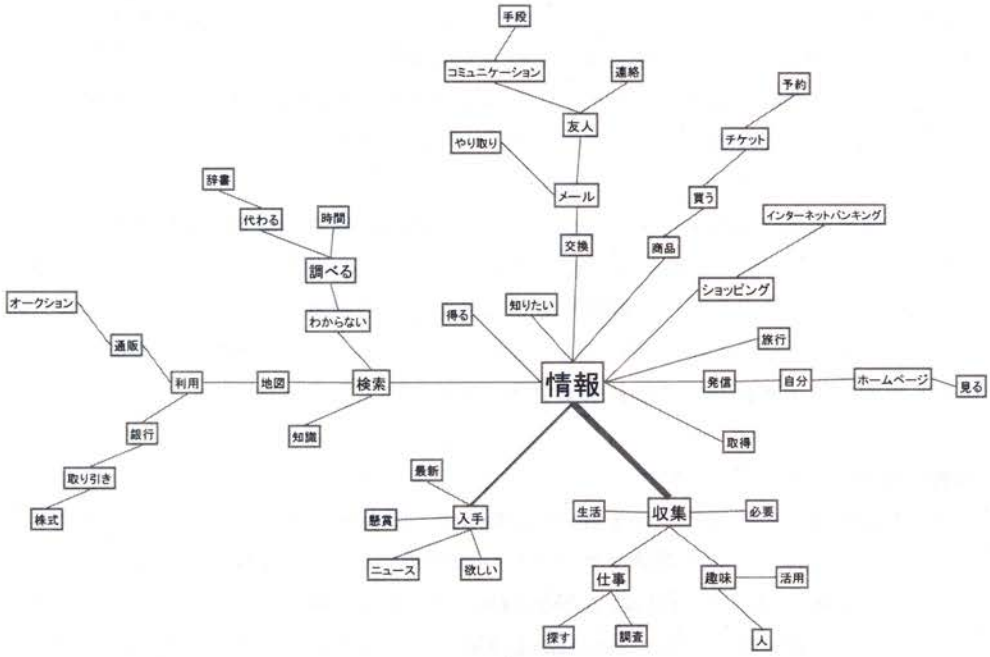


図１　DE-FACTO の出力例（49 語を「情報」をコアとしてリンク）
（Q3-1. 自分自身にとって「インターネット」はどのようなことに活用できるか）

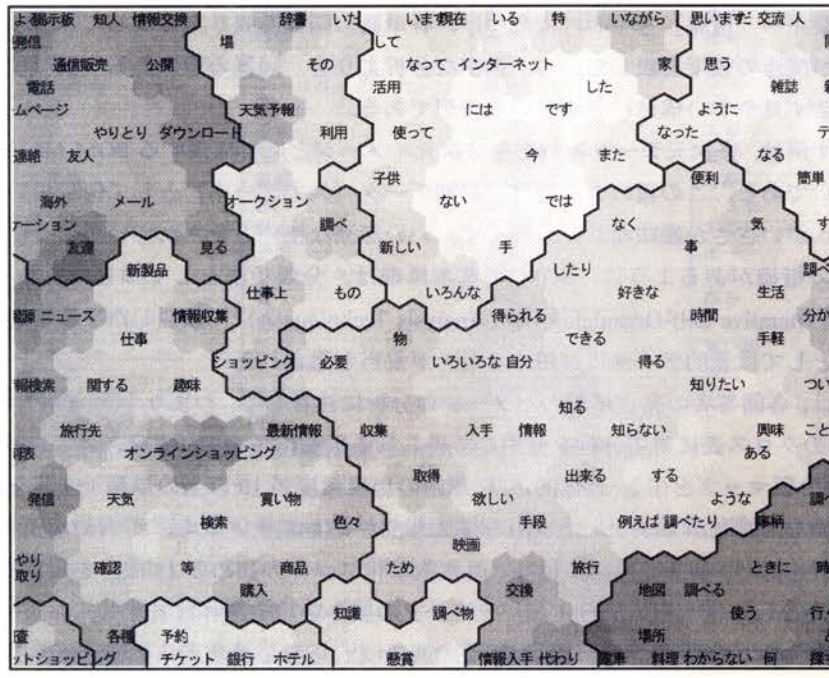


図２　SOM による分析例（Viscovery を利用して 155 語を布置）
（Q3-1. 自分自身にとって「インターネット」はどのようなことに活用できるか）

しかし，図１，２ともに，結果解釈は恣意的となり，こうした視覚化からはたしてどこまで客観的な情報が得られるか難しいように思われる．３．２節の（５）で指摘したように，統計ソフトウェアを用いて出力されるグラフとそこに付与された統計値を併用観察することとは，使い方がかなり異なり，この種の視覚化情報は慎重に利用すべきである．また，こうした視覚化で利用できる単語数は“せいぜい数百語程度”であって，大量の単語群から有効な情報，知見を得るというTMの目標に対して満足できる解とはなっていない．換言すると，ここには，大量の語句から知識発見に必要な少数語句をいかに合理的に抽出するかという別の課題がある（残念ながら，我々はその解を知らない）．

5.4 出現単語の特徴抽出—統計手法を用いた客観的な考察—

図と数値統計情報とを併用することの利点を示す例を挙げよう．繰り返しになるが，TMの主たる目標の一つは，大量の単語・語句群の中から意味ある知見を得ることである．このための視覚化操作とは，単語群と回答（者）の関係，あるいは単語群が自由回答以外のどの設問群とどの程度の近い関係にあるかを知る手段が必要である．もっとも簡単な操作として，前述の「(回答者）×（単語・語句群)」のクロス表，さらに「(選択肢型設問）×（単語・語句群)」のクロス表に注目し，これの数量化（対応分析など）を行うことである．我々が開発したWordMiner™はこうした機能を備えている．数量化スコアの布置図を描くことは通常のソフトと同様である．加えて，分析に用いた個々の単語・語句が，果たしてどの設問や属性と有意に関係し，またその設問，属性の各選択肢でどう単語の利用パターンの類似，差異が見られるかを出現単語･語句の有意性検定を用いて客観的に評価する．これの例を一つ示そう．

ここでは，いままでと同じ設問（Q3-1：自分自身にとって「インターネット」はどのようなことに活用できるか）を用いる．分析処理条件としては有効回答数：894 サンプル中，対象サンプル：848 サンプル，出現頻度が 3 以上の単語 312 語とした．またここでは分かち書きの単位そのものではなく，特徴的な語句群として選んだ「キーワード」を用いる（注：WordMiner™では，分かち書き結果そのものと，そこから主として名詞・固有語として抽出したキーワードと，２種の単語群が利用できる)．属性として「性年齢区分」（表 4 のように，男女別に，年齢区分 24 才以下，25 才～29 才，30 才以上 60 才未満は 10 才区分，60 才以上とあわせて 6 段階）を，すなわち「(性年齢区分：12 カテゴリー）×（312 単語)」のクロス表を分析対象データ表とする．通常の分析にならえば，語句と性年齢区分との数量化スコアを求め，これの同時布置図を描き観察することで，性年齢区分と単語との関係を読み取る．例えば，それぞれの性年齢区分の近くにある語句から，その性別・年齢層に関連する語句を観察する．しかしこの程度の単語数であっても布置図は煩雑で観察が困難となりしかも解釈も主観的である．

そこで，布置図の観察に加えて，各性年齢区分内に現れる単語群が，全単語の出現頻度分布に対してどう偏るかを統計的な有意性テストを用いて検証する．この結果が表４である．ここで上位とは，ある区分内で用いられた語句の頻度が，全体の出現頻度に比べて相対的に多いことを意味し，下位とはその逆に，その区分であまり重要ではない語句を表している．例えば，「女性 24 才以下」の層では，「レポート」「就職活動」「情報検索」以下に続く語句が

表4　性年齢区分別にみた特徴的な単語群

有意の順位	男性-24才以下 サンプル数：39 異なり構成要素数：106	男性-25才～29才 サンプル数：58 異なり構成要素数：132	男性-30才～39才 サンプル数：201 異なり構成要素数：223	男性-40才～49才 サンプル数：119 異なり構成要素数：189	男性-50才～59才 サンプル数：51 異なり構成要素数：120	男性-60才以上 サンプル数：27 異なり構成要素数：81
上位 1	現在	コミュニケーション	様々	安価	情報入手	ビジネス
上位 2	雑誌	ファイル	仕事	相手	自分自身	株式投資
上位 3	やり取り	取得	情報収集-情報集め	仕事	親戚	旅-旅行
上位 4	学術論文-論文	情報収集-情報集め	趣味	個人	仕事上	学校
上位 5	データ	ツール	仕事上	物品購入	懸賞応募	取引
上位 6	学校	知識	バンキング	余暇-レジャー	製品	交通情報
上位 7	就職活動	通信販売	共有	手続き	百科事典	情報交換
上位 8	友達-友人	手紙	調べ物	製品	収集	いろいろ
上位 9	仲間	人	テレビ	他人	会社	株価
上位 10	勉強	やり取り	情報検索	飛行機	仲間	事象
上位 11						宿泊先
上位 12						宿泊予約
上位 13						出張時
上位 14						地図検索
上位 15						土地
下位 12						インターネット-ネット
下位 11	手段			テレビ	事柄	情報入手
下位 10	調べ物	情報交換	検索	家	手	必要
下位 9	仕事上	旅-旅行	やり取り	料理	商品	仕事上
下位 8	色々な	興味	天気予報	新聞	ショッピング	予約
下位 7	情報収集-情報集め	手	場所	人	買い物	時
下位 6	等	購入	生活	地図	電子メール-メール	人
下位 5	予約	ショッピング	海外	場所	手段	自分
下位 4	旅-旅行	予約	料理	生活	調べ物	ショッピング
下位 3	趣味	事-こと	旅-旅行	活用	コミュニケーション	買い物
下位 2	買い物	ニュース	情報入手	海外	色々な	仕事
下位 1	仕事	活用	友達-友人	自分	趣味	情報収集-情報集め

有意の順位	女性-24才以下 サンプル数：37 異なり構成要素数：82	女性-25才～29才 サンプル数：55 異なり構成要素数：119	女性-30才～39才 サンプル数：155 異なり構成要素数：212	女性-40才～49才 サンプル数：68 異なり構成要素数：154	女性-50才～59才 サンプル数：29 異なり構成要素数：89	女性-60才以上 サンプル数：9 異なり構成要素数：28
上位 1	レポート	場所	子供	病気	健康	行き先
上位 2	就職活動	場	天気	事-こと	旅行先	旅-旅行
上位 3	情報検索	店	地域	買い物	宿	宿泊先
上位 4	手軽	電話番号	電話	色々	新聞	宿泊予約
上位 5	手配	音楽	ため-為	応募	利用	方
上位 6	人	レストラン	ネット銀行	にて	株取引	料金
上位 7	活用	ひまつぶし	買い物	主	時刻表	予約
上位 8	自分	飲食店	映画	出来	天気予報	重宝
上位 9	ネット上	ダウンロード	レシピ	地図検索	旅-旅行	電子メール-メール
上位 10	最近	注文	レストラン	ショッピング	メディア	売買
上位 11	作成				意見交換	
上位 12	日記				解決	
上位 13	問題				娯楽	
上位 14	ときに				交通機関	
上位 15	発信				時間帯	
上位 16					時刻	
上位 17					新幹線	
上位 18					値段	
上位 19					判断	
上位 20					目的地	
上位 21					話題	
下位 15						友達-友人
下位 14						友達探し-友達づくり
下位 13						有効
下位 12						予定
下位 11						余暇-レジャー
下位 10	購入	事-こと	趣味	インターネットショッピング	商品	容易
下位 9	手段	知識	予約	情報発信	購入	様々
下位 8	調べ物	コミュニケーション	コミュニケーション	連絡	インターネット-ネット	利用
下位 7	必要	オークション	情報源	自分	情報入手	旅館
下位 6	やり取り	ニュース	情報検索	情報入手	仕事上	旅行情報
下位 5	仕事上	情報交換	やり取り	必要	予約	旅先
下位 4	電子メール-メール	仕事	仕事上	仕事上	活用	連絡手段
下位 3	情報収集-情報集め	便利	収集	趣味	収集	話題
下位 2	等	簡単	各種	仕事	仕事	仕事
下位 1	仕事	ショッピング	仕事	情報収集-情報集め	情報収集-情報集め	趣味

表5　テキスト・マイニングの位置づけ

データの型（種類）		対　象	目　標　と　対　応		
			対応方法 適用の方法論	単純なパターンの発見	意味ある複雑な情報の発見
数値型データ	質的データ （名義，順序） 量的データ （区間，比例）	・テキスト型データを計量化・数量化し，数値型データとみなして処理 ・数値型データとテキスト型データの併用	典型的なデータマイニング 統計解析手法 （*）特徴，傾向，規則性の探査・発見 （*）モデリングの支援	・データベース問い合わせ （*）単純な検索，情報アクセス，参照など ・タグ化，コード化，カテゴリー化など ・情報検索，情報抽出	・形式的 TM の実行 （*）テキストの計量化を通じて探査 （*）現状の TM の主流 ◆これで「意味ある複雑な情報の発見」は本当に可能か
非数値型データ	テキスト型データ	小説，自由記述文，自由回答など （非構造的） 一般文書類 （構造的）	自然言語処理・計算機言語学 （*）構文解析，意味解析，文脈解析，… （*）共起，係り受け等	文書要約 文書分類 内容分析 全文検索	◆真の TM とは？ さらに有効な TM の方法はあるのだろうか？
	非テキスト型データ	画像，音声など	計算機言語学 言語学 音声学	自動翻訳技術 多言語間翻訳	自動通訳機能 意味理解 音声認識 画像認識

いくつかの例示でみたように，現状のTMツールの盲点は，とくに入り口（本当に大量のデータセットの処理が可能か）と出口（解析結果，その解釈に科学性があり客観的か）にある．TMが本当に「大量のテキスト型データから知識発見し知識組織化を目指す」方法論であるなら，これに適切な解を与えるべきである．TMやDMが最終目標とする「知識発見，価値ある知見の発見」とは何をいうのか，また，今のTMの利用環境でこの目標が本当に達成されるのだろうか．そして真のTMの目指すべき道はどこにあるのだろうか．一つのつたない試みとして，表5を作ってみた．

ここでは，TMが扱う「データの型（種類）」「対象」そして「TMが目標とする内容と対応（用いる方法論，考え方）」の関係を示している．明らかなことは，既述のように，所与のテキスト型データを，その生の情報のまま扱うのではなく，一度「数量化・計量化の手続き」を経て，従来型DMの方法論が適用可能な形に情報を変換することがある（情報の量と質の両面での変換操作がある）．この意味ではKDDプロセスと変わることはない．

単純な操作としては，語句・単語を抽出しコード化，カテゴリー化，タグ化などを通じてテキスト情報を数値として扱い易い形とし情報検索や情報抽出を行う．またテキスト情報を多変量解析や多次元データ解析手法により数量化を行い，情報要約・次元縮約を図って，テキスト型データの定性情報を扱い易い“量的データ”として処理する．こうした接近法は“単純なパターンの発見”や“節約の原理”の達成には有効である．

一方，もっとも関心のある非構造的な自由記述文（自由回答を始め，多くの自然語文書体）のTMを行うには，まったく異なる視点からのアプローチが必要と思われる．しかし，残念ながら，筆者等にはいまこれへの的確な解を即答できるだけの情報がない（表5のセル「真のTMとは？」に相当）．ここでは，テキスト型データの数量化・計量化を通じて知識発見を行う現状の方法論を越えて，あるいはこれに加えて，“何か別の道”があるだろうとしか言えない．

しかし新たなTMが見つかるまでの“代替策”は，発話者・発言者（回答者）の“言いたいこと，述べたいこと”を的確に拾い上げる「仕組み作り」（データ取得機構）を考えることであろう．一例として，ある自治体の「市民の声」分析を挙げよう（2003 SURF研究報告（2003））．収集ルートは電話，投書，電子メール，市庁来訪と様々であり，収集情報から確かに悪臭対策，騒音対策，地下鉄問題，介護問題と多様な特徴や傾向が見える．この膨大な意見データから政策決定，意思決定に有効な意見が即座に集約されるかというとそう簡単ではない．真の知識発見とは何かという根本的な問題「ただ集めてみても適切な意見が出るとは限らない」という現実に直面する．市民の意見の述べ方，提案方法の指導を始め，的確な情報をリアルタイムに汲み取る仕組み作りを構築することが求められる．これは既述のデータ科学の精神であり，現在のKDD，TM，DMに欠けている部分である．「生の声，本音」をTMで知るというもっともらしい言葉に惑わされることなく，真のTMとは何かを再考すべき時期にある．過去にも様々な方法論が高い期待をもって登場したが大半はいつの間にか忘れ去られた．TMが同じ轍を踏むことなく育つことを期待したい．

【謝辞】

2 名の査読者から詳細なご指摘と有益なコメントをいただいたことに謝意を表します．また，例とした Web 調査データの取得と分析に際して，㈱電通リサーチ，横原東氏のご協力ならびに情報開示のご了解をいただいたことに，この誌面をお借りし御礼申し上げます．

【注】

1) コーパスとは「ある言語の言葉（話し言葉，書き言葉等）や語彙の集積で，主にコンピュータ処理が可能な集合体」のこと．「言語学的分析のために収集された一群のデータ」のこと．コーパスがどう利用されるかについては，例えば中村（2003）を参照．

2) 品詞とは，語を分類し，いくつかのグループに分けたとき，その同類となったグループの名称をいう．いわゆる，名詞，動詞，形容詞，助詞，助動詞などのこと．このグループ分けの操作を品詞分類という．

3) 形態素解析ツールに，茶筌（奈良先端科学技術大学院大学），JUMAN（京都大学），ALTJAWS（NTT コミュニケーションズ科学基礎研究所），Breakfast（富士通），すもも（NTT コミュニケーションズ科学基礎研究所），QJP（リコー），SuperMorpho-J（オムロン）などがある．

【文献】

飽戸弘編著. 1994. 『食文化の国際比較』日本経済新聞社.

Baayen, R. H. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers.

Dhillon, I. S. Mallera, S. and R. Kumar. 2002. "Enhanced Word Clustering for Hierarchical Text Classification." in *KDD-2002: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, SIGKDD.: 191-216.

Fayyad, U. and R. Uthurusamy. 1995. "Preface." in Fayyad, U., M., and R. Uthurusamy.(eds.) *The Proceedings of the First International Conference on Knowledge Discovery and Data Mining.* AAAI Press.

Fayyad, U., G. Piatetsky-Shapiro,and P. Smyth. 1996. "Knowledge Discovery and Data Mining: Towards a Unifying Framework." in Simoudis., E., J. Han, and U. Fayyad (eds.) *Proceedings Second International Conference on Knowledge Discovery & Data Mining*. AAAI Press.: 82-88.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. "The KDD Process for Extracting Useful Knowledge from Volumes of Data." *Communications of the ACM* 39(11): 27-34.

Feldman, R. 2003. "Mining Text Data." in Fayyad, U., M., and R. Uthurusamy.(eds.) *The Handbook of Data Mining*, edited by N. Ye. Lawrence Erlbaum Associates.: 481-518.

Feldman, R. and I. Dagan. 1995. "Knowledge discovery in textual databases (KDT)." in *the Proceedings of the First International Conference on Knowledge Discovery and Data Mining.* AAAI Press.: 112-17.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

言語学研究所. 2003. 『類語・シソーラス辞典ソフト「デジタル類語辞典 2003」』.

林知己夫. 2001. 『データの科学 シリーズ<データの科学＞1』朝倉書店.

Hearst, M. A. 1995. "TileBars: Visualization of Term Distribution Information in Full Text Information Access., in *the Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.*: 59-66
[http://www.sims.berkeley.edu/~hearst/papers/tilebars-chi95/chi95.html]

Hearst, M. A. 1998. "Current Topics in Information Access." in *SIAM Academic Course* 296a-5-3.

Hearst, M. A. 1999. "Untangling Text Data Mining,." in *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*. (invited paper).

今井浩. 2001.「データマイニングとは？−情報システムとしての温故知新」『ESTRELA』8 月号:2-9.

伊藤雅光. 2002. 『計量言語学入門』大修館書店.
川端亮・樋口耕一. 2003.「インターネットに対する人々の意識-自由回答の分析から-」『大阪大学 大学院人間科学研究科紀要』29:163-181.
北原保雄（監修），斎藤倫明（編）. 2002. 『語彙・意味 朝倉日本語講座 4』朝倉書店.
北原保雄（監修）. 2003. 『日本語の使い方，考え方辞典』岩波書店.
小池清治，小林賢次他（編） 1997. 『日本語学キーワード事典』朝倉書店.
国立国語研究所. 1964. 『分類語彙表 国立国語研究所資料集』大日本図書.
国立国語研究所. 2004. 『分類語彙表 増補改訂版（CD-ROM 付）』大日本図書.
Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology*(second edition). Sage Publications.
Lagus, K., T. Honkela., S. Kaski, and T. Kohonen. 1996. "Self-Organizing Maps of Document Collection: A New Approach to Interactive Exploration." in E. Simoudis., J. Han, and U. Fayyad (eds). *Proceedings Second International Conference on Knowledge Discovery & Data Mining*. AAAI Press: 238-43
Lebart, L.,A. Salem, and L. Berry. 1998. *Exploring Textual Data*. Kluwer Academic Publishers.
町田健. 2003. 『コトバの謎解きソシュール入門』光文社新書.
松本祐治・今井邦彦他. 1997. 『言語の科学入門 岩波講座言語の科学 1』岩波書店.
Miles, M. B. and M. Huberman. 1994. *An Expanded Sourcebook: Qualitative Data Analysis*(second edition). Sage Publications.
Murtagh, F. 1999. "Data Mining, Statistics and Data Science." in *Proceedings of ISM Symposium: Data Mining and Knowledge Discovery in Data Science.*:1-12.
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（CD-ROM 版）』 岩波書店. [http://www.kecl.ntt.co.jp/icl/mtg/resources/GoiTaikei/]
長尾真・黒橋禎夫他. 1998. 『言語情報処理 岩波講座言語の科学 9』岩波書店.
長尾真（編）. 1996. 『自然言語処理 岩波講座ソフトウェア科学第 15 巻』岩波書店.
Nahm, U. A Roadmap to Text Mining and Web Mining, Department of Computer Sciences, The University of Texas at Austin. [http://www.cs.utexas.edu/users/pebronia/text-mining/]
中村純作. 2003.「コーパス言語学」 山梨正明・有馬道子（編）『現代言語学の潮流』 勁草書房:233-245.
Neuendorf, K. A. and P. D. Skalski. 2002. *The Content Analysis Guidebook*. Sage Publications.
大隅昇. 2004.「調査環境の変化に対応した新たな調査法の研究」『インターネット調査産学協同研究報告』CD-ROM.
———. 2002. 「インターネット調査の適用可能性と限界—データ科学の視点からの考察—」『行動計量学』 29(1):20-44.
———. 2002. 「テキスト型データの多次元データ解析—Web 調査自由回答データの解析事例」 柳井晴夫・岡太彬訓・繁桝算男・高木廣文・岩崎学（編）『多変量解析実例ハンドブック』朝倉書店：757-83.
———. 2000. 「定性情報のマイニング-自由回答データの解析-」『ESTRELA』5 月号:14-26.
大隅昇・L. Lebart. 2000. 「調査における自由回答データの解析-InfoMiner による探索的テキスト型データ解析-」『統計数理』 48(2):339-376.
大隅昇・丸岡吉人他. 1997. 「自由回答データの解析法についての提案-実験調査におけるいくつかの試み-」『第 25 回日本行動計量学会大会報告要旨集』:176-79.
奥村学. 2000.「自然言語処理関連ツールあれこれ-使えるフリーソフト-」『情報処理』41(11):1203-07.
Popping, R. 2000. *Computer-assisted Text Analysis.* Sage Publications.
Renz, I. and J. Franke. 2003. "Text Mining." in Franke, J., G. Nakhaeizadeh, and I. Renz(eds.). *Text Mining: Theoretical Aspects and Applications*. Physica-Verlag: 1-19.
仙台都市総合研究機構. 2003.『「市民の声」の活用法に関する調査研究』2003 SURF 研究報告.
柴田武・山田進（編）. 2002. 『類語大辞典』講談社.

Software, E. 1999. Viscovery SOMine version 3.0 Scientific and Enterprise Edition, User's Manual. Stone Analytic, Inc. Evaluating Text Mining Applications.
[http://www.secondmoment.org/atats-column/stats-textmining.php]

Sullivan, D. 2001. *Document Warehousing and Text Mining*. John Wiley.

Tan, A.H. 1999. Text Mining: The state of the art and the challenges, in *Proceedings: PAKDD'99 Workshop on Knowledge discovery from Advanced Databases* (KDAD' 99), Beijing.
[http://www.ntu.edu.sg/home/asahtan/publications.htm]

Thuraisingham, B. 1999. *Data Mining Technologies, Techniques, Tools, and Trends.* CRC Press.

Weber, R. P. 1990. *Basic Content Analysis* (second edition) Series: Quantitative Applications in the Social Sciences 49. Sage Publications.

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.

山梨正明・有馬道子（編）. 2003. 『現代言語学の潮流』勁草書房.

Ye, N. (ed.). 2003. *The Handbook of Data Mining*. Lawrence Erlbaum Associates, Publishers.

（受稿 2004 年 5 月 6 日／掲載決定 2004 年 8 月 10 日）

Reviewing Textual Data Mining in Japan

Noboru OHSUMI
The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku,
Tokyo 106-8569, Japan

Akio YASUDA
Heiwa Information Center Inc.
Nihon Seimei Kasugacho-Dai-ni Building
1-3-21 Koishikawa Bunkyou-ku,
Tokyo 112-0002, Japan

The objective of this paper is to give overviews of text mining or textual data mining in Japan from the practical aspects. Firstly, we explain that text mining is defined as a branch of data mining (in particular, KDD: knowledge discovery and data mining) which is applied to large amount of text datasets. And target of text mining is to objectively discover and extract knowledge, facts, and meaningful relationships from the text documents. We will also briefly outline the related disciplines and application fields which are applied in text mining.

In addition, we discuss the applicability of text mining in the field of qualitative research and also examine about how to solve some problems faced in using text mining techniques. Moreover, the computer programs for conducting text mining are given as the summarized tables.

As concrete examples, using a data set of some open-ended questions obtained by Web-based survey, we illustrate several analyses of segmentation of Japanese responses to the open-ended questions, visualization of mining results, and statistically significant test based on the frequencies of characteristic words and the corresponding statistical test-values obtained from the aggregated lexical table for "words by gender-age variable" with 12 categories, generated by cross-tabulating gender (two categories) and age (6 categories).

Finally, we propose a perspective of text mining that we expect, that is, about how to solve questions which knowledge is needed and how to be able to suitably gather the text data sets required for understanding the target phenomena. At any rate, from the point of view of data science, question about how a sort of "acquisition system" for obtaining the appropriate data sets can be integrated will have to be examined in future.

Keywords and phrases: text mining, textual data mining, data mining, knowledge discovery and data mining (KDD), qualitative research, data science, open-ended questions, Web-based survey, correspondence analysis of lexical tables

(Received May 6, 2004/Accepted August 10, 2004)