テキスト型データのマイニング

－ 最近の動向とそれが目指すもの －

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

１．まえがき

社会調査，市場調査を始め，文章や文字で記述された，いわゆる**テキスト型データ**（textual data）の利用場面が増えている．また，その分析に関する方法論を求める声が高まっている．

とくにマーケティングや市場調査の分野では，ワン・トゥ・ワン・マーケティングの時代にあり，顧客や消費者との関係を的確かつ総合的に把握するための CRM/eCRM 等がキーワードといわれている．こうした中で，CRM/eCRM 等を支援する有力な方法のひとつとして，定性情報の有効活用が注目されている．たとえば，顧客満足度（CS）の評価，そして，コール・センターやコンタクト・センターなどの実装化レベルでのシステム構築の過程で，「顧客の生の声」「消費者の本音を知る」とのキャッチコピーのもとに，いわゆる**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）を活用するという考え方が拡がりつつある．

同じく，社会調査に携わるあるいはそれを利用してきた研究者からも，定性的情報（qualitative information）の分析を行うための方法，**質的研究**（qualitative research）の一環として，内容分析やコーディング処理法，あるいは文章型データの分析法の提供が求められている．

こうした要請に応えるかのように，ここ数年の間に，ソフトウェア市場にはテキスト・マイニング対応をうたった製品が次々と登場している（後述）．

こう記すと実体の明らかな TM，TDM という「何か役に立つ道具」があるように見える．しかし改めて考えると，いわゆる TM とは実に曖昧であり，いろいろに解釈できる概念である．似たような言葉に**データ・マイニング**（DM: data mining）がある．これも流行り言葉の一つであるが，TM と同様に，分かったようで実は漠としたものである．また，これを表題とする論文や書籍は無数に発刊されており，併せて沢山のコンピュータ・ソフトが現れ，正に百花斉放の感がある．しかし子細に眺めると，DM を適用して画期的な成果があったという話しもそう多くは見られない．また，統計学あるいは統計的データ解析で利用されてきた方法論とどう異なるかというと，これもいまひとつ明らかではない．これは TM についても似たような事情にあると思われる．そして，TM と DM の言葉の類似から類推されるように，この両者の差異や類似がどこにあるのかも，さほど明確ではない．

ここではまず，TM についてその特徴を俯瞰すると同時に，これに関連する技術的な諸要素，諸事項について，“総合的に”概観する．そもそも TM とは何を言うのか，またそれが対象とするものは何か，関連する研究分野は，とくに DM とはどう関わるのか，といったことを総合的に要約してみたい．要約であるから，個々の要素についての知識を深めることには困難があり，ここはあくまでも全体を見渡すという姿勢で記してみたい．

１．１　データ・マイニングとテキスト・マイニング

ときとして，TM は DM からの派生した方法論であるとの記述が見られる．「**鉱脈探し**」（mining）という共通語からの類推であろう．しかし，これは必ずしも正しくはない．後述するように，TM の発祥あるいはルーツは，実は様々なところにある．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（いわゆる解析手法やそのアルゴリズム）については，DM にかなり類似したものがある．

では，どう異なるのか，それを知るには，そもそも DM とは何かを知る必要がある．これについては，前述のように雲霞のごとく無数の研究報告や書冊がある．しかし，DM に関連する個々の技法や方法論にまで言及するスペースもない．ここでは，一般的に考えられている DM の概念を眺め，これに続いて TM とは何かをみることにする．

最近は，DM を**知識発見**（KD: Knowledge Discovery）にリンクして議論することが多い．しかし，元々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．いわゆる**人工知能研究の一つの支流**として，80 年代広

範から 90 年代に入って登場した狭義の **KDD**（Knowledge Discovery in Databases）である．ここで KDD とは「データに潜在的に内在する，確かな，しかし予期しないような特徴の把握，また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．さらに，この狭義の KDD にデータ・マイニング（DM: Data Mining）が加わって，今の新たな KDD（Knowledge Discovery and Data Mining）がある．つまりここで始めて，知識発見の道具立てとしての DM の役割が明らかになる．すなわち 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（アルゴリズムなど）に相当する処理過程，また，検証，発見，予測，記述などに関連の諸要素の集合体が DM の基本的な役割という見方である（実際に実現可能かは別のこと）．

たとえば，DM 分野で先行的な研究を続け，またエバンジェリスト的な役割を果たしている Fayyad，Piatetsky-Shapiro 他（1996）やその周辺の研究報告によると，KDD と DM の関係を以下のように説明している．まず，DM の中心的な方法論の多くは，統計学の支流として登場した．これに加えて，1989 年頃に登場した KDD に対して，パターン認識，機械学習，大規模データベース環境を前提とするデータベース技法を背景に，この研究分野を特徴付けるために DM と名付けたものである．

ここで，従来からの統計的手法や**統計的データ解析**（とくに**探索的方法論**）の知識が多少なりともある者にとっては，図 1 に見るような考え方がデータ解析とどう異なるのか，俄には分かりかねるであろう（筆者もその一人だが）．また，Fayyadも指摘するように，統計学・統計的データ解析の関連手法の支援も受けるという意味で，不可分の関係にある．しかしDMに関連する多くの書では，その違いは「統計的な分布の仮定がない，母集団概念などが不要」「そもそも扱うデータの規模・ボリュームが異なる」そして「（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やデータ・ウェアハウスを用いる」等の言葉が返ってくる．しかし，最近の統計的方法論では，こうした指摘に対する解決策は提供されており，その違いはどこにあるかは依然として曖昧であり，こうした主張だけではDMを特徴付けるための説得力がない．つまりは，DMという耳に心地よい言葉に惹かれた一種の流行のように見えるのである．

確かに，膨大なデータセットを目前にしたとき，それがデータベース上にあるかどうかは別として，その中から“金の鉱脈，ダイヤモンド”を探し当てる方法があるなら，それに越したことはない．しかし，いま考えられている DM あるいは KDD 過程には重大な落とし穴がある．DM の多くの書に**「ゴミを入れればゴミが出る」**（GIGO: garbage in garbage out）との言葉が頻繁に登場する．しかし冷静に考えると，「ではゴミではないデータはどこにあるのか」という極めて素朴な疑問に行き着く（ニワトリと卵の論法である）．しかしながら，DM の多くの方法論はこれには答えてはくれない．**十分な量の適切で良質なデータがあれば，という前提**で議論が展開されるのである．つまり，あてがわれた（整った）データから何かを探るというきわめて面白みのない考え方であり，はたして真の意味の現象解析がこのアプローチで可能かという疑問に突き当たるのである．

では，統計学ではここはどう考えるのであろうか．古典的な統計学では母集団（目標母集団，枠母集団）を想定し，そこで実験計画なり調査計画を厳密に構築し，サンプリングという操作をもって分析対象（標本）を用意する．この厳密さがあるがゆえに，現実の現象解析に適したデータ取得環境が作れず，結果として，数理の枠の中での些末な議論となって，事が矮小化されてしまうこともある．しかし「ゴミは，所詮はゴミ」であり，やはりそこには問題とする現象解明にとって必要とされる**“目的に合ったデータの取得方法”**が必要であり，またそれを前提とした**“データ主導型”**の解析過程を必要とするのである．この点では，統計学の方が，明確な枠組みを示していると言える．これを発展的に考える概念がいわゆる**「データ科学」**（data science）である．データ科学では，現象解析の基本は「データ」にあると考え，つまり「データによる現象理解」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探索的データ解析（EDA）

の発展型と考えてよい．その要点は，

①データをどう計画的に取得するか（experimental design）
②データを具体的にどう集めるのか（data collection mode）
③そして，問題とする現象解明に適した解析法はどうあるべきか（analyzing）

の３つにあり，この①～③を探索的に“行きつ戻りつ”する過程をいう．またここでは，いわゆる探索的か，確証的か，(仮説）検証的かはあまり本質的ではない．重要なことは，モデリングが先にあって，それにデータが追従する（当てはめる）のではないということ，現象を見て理解するためのデータのありようは何か，その取得方法は，…と仮説発見的に考えることにある（図１）．これは，最近の（統計）科学研究の目指す方向とは，ある意味で微妙に異なる概念である．とくに，最近の統計解析は，①と②が軽視される傾向にあり，ともすると統計的モデリングに偏っていると考えられる．
　とくに調査における自由回答・自由記述データの分析に際しては，（第 II 部で）後述するように，データ取得をいかに適切に行うか，行える環境が作れるかが生命線とも言える．データ科学を意識しつつ分析に臨むことが期待される．

１．２　TM の発祥と関連分野

　いささか横道にそれた感がある．しかし実は，後述するように TM を考えるうえで，ここに述べたことが重要な「まえおき」となる．前述のように，日本国内では，とくにマーケティングや市場調査の世界では，電子テキスト化されたデータ，たとえば調査で取得した自由回答・自由記述設問の回答，あるいはコール・センターなどで収集した顧客情報データなど，いわゆる定性情報から，有益な情報を得る手段として TM を適用することが，高い期待感を持って迎えられている．しかし，TM とはそもそもどのような考え方であり，多数の関連ソフトツールは何を行ない，どのような情報を提供してくれるのだろうか．
　ここではまず，現状の TM とは何か，どのような方法論に支えられているのかを，筆者の観点から俯瞰的に概観することにしよう．そもそも，TM とはどのような分野の成果物であり，何が行われるのか，ある側面だけに注目が集まり，TM の本質が正しく理解されていない面があるようだ．ちなみに，検索エンジン（Google）を使って，TM に関連する幾つかの用語を検索し，どのくらいのアイテムがヒットするかを調べてみた．ここでは英語と日本語について，以下の用語を確認した．
（注）これらは，これを執筆した 2003 年～2004 年あたりの情報である．

テキスト・マイニング（9,650 件），text mining（961,000 件）
テキスト型データ・マイニング（51 件），textual data mining（26,300 件）
テキスト・データ解析（7,690 件），
テキスト型データ（144 件），textual data（559,000 件）

さらに，言語処理あるいは言語分析で，登場する幾つかの語句も併せて検索してみよう．

context analysis（文脈解析）（3,460,000 件）
content analysis（内容分析）（4,090,000 件）
lexical（630,000 件），lexical analysis（語彙分析）（185,000 件）
semantic analysis（意味分析）（394,000 件）

　もちろんヒットした内容を網羅的に精査したわけではない．重複もあり，あまり関係のないものも含まれる．しかし，それぞれがかなりの数にのぼることがわかる．同時に利用

頻度の高い用語も見えてくる．ここで，この検索結果を「どう読み解くか」である．ここで早くも情報検索結果のマイニング（mining）を行う必要に迫られる．しかも，これだけのデータ量となると，さほどうまい方法がないことにも気付くのである（このようなとき，流行りの Web マイニング・ツールが役立つのだろうか）．手作業によるマイニングの「主観的な要約」で，以下のような特徴があるように見える．

① 用語によって出現頻度に偏りがあること，とくに英語，日本語の差異が大きい
② 無数の White Paper（簡易報告書）が出ていること，つまりこの話題はホットであること
③ 関連語，類語，同義語が多数あり，つまりジャーゴンに溢れ，しかもかなり曖昧に使われているらしいこと
④ 同時に，既にかなり長い研究履歴があるテーマもあること（ヒットした件数とその情報のアップロードの時期から類推），たとえば内容分析，文脈解析など
⑤ どのような関連分野があるかが，わずかだが透けて見えること

２．テキスト・マイニングの背景

２．１　なぜ，いま，テキスト・マイニング」なのか

ではここで，なぜ，いま，TM なのかを考えてみよう．言うまでもなく，現用の TM ツールやその設計指針の基礎となる考え方には，様々な研究分野での成果が反映されている．たとえば，一般的な言語学研究（日本語，欧米語共に），計量言語学，コンピュータ利用を前提とした自然言語処理あるいは計算機言語学，言語情報学，あるいは内容分析などがある．さらには人工知能研究や機械学習と，多くの知識要素を取り入れた「融合体」として，テキスト・マイニングやそのツールが登場している．

こうした従来からの多様な研究・応用分野で進展を見た諸要素に加えて，コンピュータ・ネットワーク，とくに WWW 環境に基づくインターネットの普及による，データ収集あるいは取得機構の急速な変容がある．かつて，電子化されたドキュメント（文書）やテキスト，あるいは調査における自由回答・自由記述データの電子的データの収集や整備にはかなりの労力を要した．しかし現在は，以下に挙げるような環境となってきた．

- コンピュータ環境の改善により，膨大な文書，テキストが扱えること
- インターネット時代にあってデータ収集過程に大きな変容があること
- 大量の電子化されたデータが容易に取得できる環境があること
- とくに，ハードウェアの性能向上で，非数値型データ（テキスト，映像，音声など）の扱いが容易となったこと
- データベースの機能向上により，データベース上の数値データや構造化されたデータ（structured data）だけでなく，非構造化データ（unstructured data）の扱いの自由度が高まったこと
- インターネット調査などの調査方式の普及で，簡単かつ平易にデータ取得が可能と思われていること（例：Web 調査による自由回答取得など）
- コール・センター，コンタクト・センターなどの普及で，データベース上に定性的情報の蓄積が増えたこと

しかし，こうした現象は，利点だけでなく問題点も多々併せ持つのである．たとえば，多くの場合は「**ボリュームは大きいがゴミだらけ**」である（ゴミとはなにかが既に検討課題である）．しかも電子的データ取得が容易となったことが「何か意味ある情報が取れること，取れそうにみえること」との期待を抱かせる．多くの場合，ここらが正しく理解されてはいないことがある．

同時に，蓄積された膨大なテキスト型データから，何らかの操作で有用情報が抽出され

るなら，そのようなうまい方法論があれば，確かにありがたいことではある．ここに TM への過剰な期待が生まれる素地がある．

２．２　テキスト・マイニングとは？

では，TM（テキスト・マイニング）とはなにをいうのであろうか．もっとも安易な言い方は，DM（データ・マイニング）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．つまり，これらの諸研究の融合体，各種の技術要素，種々の方法論の集合体と考えてよい．確かに，人工知能研究の派生であること，また膨大な情報から意味ある関係や特徴を抽出する，という意味では，つまり炭鉱採掘（マイニング）という点では共通性がある．

ステロタイプな言い方が多いのだが，そのいくつかを書き上げてみる．まさに通俗的な言い方であるが，もっとも一般的でもある定義，概念は以下のようなことであろう（F. Neri, U. Nahm 他）.

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，それらの間に潜在的にある関連性を分析し，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・知識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が最大の特徴である．

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，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こと，などを行う一連の操作をいう．

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，傾向を探査することである．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．
- たとえば，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．

これらを見ると，幾つかの共通項があることに気付く．たとえば，以下のような項目である．

- 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（コーポラ）

- 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

計算機言語学の研究者である 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かつ誘導し，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．彼女はさらにその本質は，単に，情報検索の技術，あるいは発話解析，語義曖昧性の解決（解除），辞書作成などの，単なる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**と言っている．つまり，ここでも，事の本質が探索的アプローチにあることの重要性を指摘しているのである．

注：テキストとは？
ここでは，主に電子化された文字情報あるいはそれに替わる類似情報をいう．ドキュメント（文書），E-mail や Web ページ上の文字情報，電子アーカイブ，そして，定性調査で取得の情報，たとえば，自由回答設問，グループ・インタビュー，フォーカス・グループなどで取得のテキスト型データの集合体．これらが原則として**「電子化された」情報**をいう．換言すると電子コード化された文字情報，**テキスト型データ**（textual data）である．

２．３　TM と関連する分野，方法論，そして適用の範囲

ところで，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，考え方は様々である．しかも TM は，学際的かつ広範な分野にまたがっており，考え方としてこれといった厳密な制約や境界もない．たとえばここで，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てみよう．

（１）関連研究分野からの観察

いうまでもなく，「ことば」や「文字」は，人と人とのコミュニケーションの道具でありこれを用いることは人間行動に関わる重要な行為である．従って関連する研究分野は当然多様なものとなり，

- 自然言語処理（NLP：natural language processing）あるいは計算機言語学（CL：computational linguistics）
- 人工知能（AI），エキスパート・システム，知識獲得，知識工学
- 認知科学や認知モデリング（cognitive modeling）
- 情報検索（IR：information retrieval），情報処理
- 機械学習理論（ML：machine learning）
- 計量言語学
- コーパス言語学
- 計量文献学
- 言語学，社会学，行動科学など
- 記号論，テクスト論，カテゴリー論，意味論など
- 内容分析（content analysis）あるいはテキスト分析（text analysis）

等がある．これらに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．これを要約すると図２のようになる．

また当然のことながら，自然言語処理あるいは計算機言語学との関連が強い．すなわち，**形態素解析，統語解析**（syntactic analysis，parsing）あるいは**構文解析（統語解析），文脈解析，意味解析・意味理解，文法生成**などの技法も，TMには必要とされるだろう．また，係り受け，n-gramなどの技法がTMのツールの利用されることもある．

研究の長い歴史がある**内容分析**（content analysis，text analysis）についても同様のことが言える．コンピュータ利用の内容分析（CACA： Computer-assisted content analysis）が登場したのは既に半世紀近くも前のことであり，またそれ以前から様々な内容分析の研究が行われてきた．中でも，文書情報管理・検索機能は重要で，インデクスやカテゴリーをサーチし，さらにある語句（キーワード）とその語句の使用前後の文脈を調べるKWIC（keyword in context），さらにはコンコーダンス（concordance）機能で，ある語句の文章内での使い方や共起の関係を調べる．併せて共起語，コーパス頻度，共起頻度の閲覧や統計的指標の出力なども得られる．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)］．こうしたCACAの成果も，TMを考えるうえで無視できない．

（注）内容分析については，「内容分析とコーディング」（樋口耕一，2004）などが参考になる．

（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 を考えてみよう．ここでは，以下のような手法が登場する．

- 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，自由回答設問設計など
- 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，情報検索技術など）
- 各種の視覚化，可視化の技法，グラフィカル表現法

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，実に多様である．これを要約すると図３となる．

このように多様な分野の“技術要素の集合体”であることが TM の特徴であり，この点では DM に同様である．従って，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，それぞれの分野の利用技術の特色を活かし，また方法論の利点を，目的に応じてどう使いこなすかという「使い方」が TM をうまく活用するためのキーとなる．つまり，分析対象に応じて「何を（どんな方法を）使うか」ではなく，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

当然のことであるが，TM が関与する適用範囲は実に多彩である．様々の報告書，研究論文にあるアイテムを要約すると，次のような言葉が登場する．

- 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）
・ テキストの分割（text segmentation）
・ テキスト，文書の要約化と収集分析（summarization analysis of text collection）
・ 知識取得と理解（knowledge capture & understanding）
・ テキスト・ナビゲーション，視覚化のためのユーザー・インターフェース（text navigation, visualization and user interface）
・ 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（text search），全文検索（full text documents search），文書検索（document retrieval）
・ 情報抽出（information extraction）

等々，枚挙にいとまがない．ここで注目すべきことは，欧米諸国では，ここに列記したようにTMの応用分野が実にバラエティに富み，様々な分野に拡がっていることである．とくに，整備された（構造化された：structured）**膨大な文書データベースやコーパスを用いた，知識発見ツール**としてのTMがある．

一方，日本国内では，ビジネス面での適用，とくにマーケティングや市場調査分野におけるTMの適用場面は，調査データ（自由回答）の分析やコール・センターやコンタクト・センターで収集の非構造的なデータ（unstructured data）への適用例が多い．本来のデータ・マイニング的な利用法であるドキュメント分類，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，研究としては散見されても，ビジネスでの利用は少ないように思われる．あるいは，そうした機能を利用した実用分析例の報告を目にすることは少ない（そこまでTMの活用度が高まっていないか，あるいは成功事例は紹介がなされないのだろうか）．つまり，TM の応用の範囲や浸透の方向・拡がりにかなりの差異がある．ある一面だけが強調され，しかも研究の深化が極めて浅いといえる．

３．テキスト・マイニングはどう活用すべきか

３．１　マーケティングにおける適用可能性

前述のように，本来のTMが目標とする対象は，**大量文書・テキストからの“有用な情報・知識発掘”**にある．しかし，TMのマーケティングにおける適用可能性や利用範囲を考えたとき，やはり前提としては，CRM に関連した顧客対応の場面で，いかに活用できるかにあるだろう．砕いて言えば，調査における自由回答・自由記述データ，グループ・インタビューやフォーカス・グループなどの定性型データから有効な知見を得る方法としてのTMの役割をどう考えるかだろう．しかしこれは，TMの広範な適用分野の**ごく一部に過ぎない**ということを知っておこう．

TMの対象をこうした分野に限定したときに，そしてTMやその関連ツール（ソフト）いかに有効に利用するかを考えるとき，どのような視点で取り組めばよいのだろうか．闇に鉄砲ではなく，そこにはある指針や前提が必要である．そこで，次のことを考えてみたい．

・使い方のコツは，利用上の留意事項は？
・調査における利用法，活用法は？
・とくに，調査における自由回答設問の考え方は？

これに対する答えとして，また目標を絞るという意味で，以下のことを挙げておこう．

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，あくまでも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．
③ 日本語の品詞分類特定の正確性，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達成が難しい．
④ テキストの意味のニュアンスの違いなどへの拘りはあまりしない，つまり意味論的，内容分析的なアプローチには限界があるし，いま必要かをコスト面からも考慮すべきである．
⑤ 有用な知見や情報を得るためには，解析結果に適切な解釈（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）．たとえば，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．

３．２　テキスト・マイニングが行うこと，何ができるのか

繰り返しになるが，TM が目標とする方向は多方位的であり，また適用対象も多岐にわたる．したがって「TM が何を行うのか」を考えたとき，まともに正面から取り組むと，その分析の手順は膨大な組み合わせとなる．しかし一般論としては，あるいは基本的な枠組みとしては，既述の KDD プロセスのアナロジーを想起すればよい．とくに，マーケティングや市場調査などの関心対象である自由回答設問あるいはそれに類した**テキスト型データの探索的分析**に的を絞って議論することが容易である．

TM の処理過程は，図 1 で示した KDD プロセスにおいて，次のように考える．まず，数値型のデータだけではなく，いわゆる「テキスト型データ（textual data）」までを対象とすることがある．次に，解析エンジン部つまり DM 技法に相当の部分に，TM の関連手法ならびに DM 手法を含めた多様な方法論を適用する．とくに，計算機言語・自然言語処理系の技法，たとえば形態素解析，構文解析，係り受けなどの技法，あるいは言語学系の類語・同義語辞書（シソーラス）の機能，さらには分析対象に応じた語彙群の準備（コーパス・コーポラ，テキスト・コーポラ）などへの配慮も必要だろう．これらの関係は，図 4 のように考えればよい．そしてこれが現状の TM の考え方でもある．

とくにここでは，TM を“日本語の”テキスト型データの解析に適用するうえで考慮すべき幾つかの要素について，述べることにする．また既述のように，TM で最重要なことは，**対象とする事象の解明に適したデータ取得法**の設計にある．これを前提として，実際の 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

TM に限らずデータ解析すべてに共通することであるが，集めたデータセットについての事前処理や初動探査を必要とする．データランドリ，論理チェック，単純集計による探査，等々の処理が必要である．また，必要に応じて，大量データセットから分析対象とする一部を抽出するサンプリング操作を用いる．

この段階で, 既存の統計ソフトウェア･統計システムを利用することは必須要件である．なお，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．しかし，**例外的なもの，はずれ値的なもの，変則的なもの見抜くことが不得手**である．TM の課題の一つとして，ここをどう処理できるかがある．

（２）　形態素解析と統計処理

日本語の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．言語類型論（linguistic typology）により言語の形態的特徴で区分すると，日本語は膠着語（agglutinating language）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」という意味ではない（言語の分類を，単語がどのような活用をするかという基準で類型化した，ということである）．単に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．

（注）日本語処理を行う際に，形態素解析が必須の操作であるとは限らない．日本語の全文検索では，形態素解析を用いない文字ベースの全文検索手法もある（インデックスを文字に対して適用する）．

また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．実はこうした混用は「くぎり」を示す役割を果たしているので，視認により意味の誤解が避けられる．しかし，コンピュータにとってはこの「くぎり」つまり分かち書きが難問となる．

語句・単語が連なった**「べた書き」**ということは，欧米語とは異なり，解析時の処理単位が明らかでなく，そのままでは扱うことはできない．欧米で開発された TM ツールがそのまま日本語処理に適用できない理由の一つがここにある．そこで，少なくとも，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．たとえば日本語学キーワード事典によると「単語をさらに細かく分析して得られる意味上の最小の言語単位」とある．従って，分かち書き処理で得た単位要素がそのまま形態素とはかぎらない．一般には，ここらの用語の使い分けは曖昧である．通常，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が行われる．また，分かち書き処理を含む形態素解析のツールは多数登場しており，またその処理方式も様々である．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こと）ができるとは限らない．また分かち書きの約束も一通りではない．文節単位で扱うのか，単語を単位とするか，あるいは助詞などを付けたままとするか否か，と様々である．データ解析上はこのことに十分に留意せねばならない．**出発点が異なるデータセットを用いた解析から同じ解答が得られるとは限らない**のである．多くの場合，TM の分析結果に，こうした基礎情報の説明がなされないことは，結果解釈や信頼性をひどく損なうものであり，分析者は報告に際してこれら基礎情報を明らかにする必要がある．

また，自然言語処理系では，形態素解析を始め，いわゆる言語的知識（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（あるいは構文解析），文脈解析などを行う．TM はいわばこうした技法体系の一部を利用している．

参考：

形態素解析を行うツールとしては，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．

（３）　多変量解析，多次元データ解析

TM は，DM 同様に，解析部の方法論として，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）が多用される．しかし TM ツール（ソフト）の内容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 類など），回帰分析型手法，多次元尺度構成法（MDS）などが利用される．
一般に，TM で扱うデータセットのサイズや項目数，語句数などは膨大であり，高次元であることから，**次元縮約**（reduction，reducing）や**節約原理**（principle of parsimony）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），いわゆる教師なし分類をはじめ，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）などが利用されている．非階層的分類では k-平均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である CART（二進木解析）や CHAID なども頻用される．

こうした多変量解析や分類手法では，モデリングや最適化に関連してニューラル・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．ここらは統計ソフトウェアの開発社やベンダーにとっては，従来からの技術資源を核として，これにデータベース機能や機械学習型機能を付加することによって, あらたな DM ツールとして提供できる素地がある．実際，Enterprise（SAS 社）や Clementine（SPSS 社），あるいは IntelligentMiner（IBM 社）などをみれば，このことは明らかである．［後述の TM ソフトウェアの一覧も参照，表１，表２］

また多くの手法は，数値化されたデータを扱うことから，実は**テキスト型データそのものを直接扱うわけではなく，いったん数量化された情報から推論を行う**．つまり，「情報の質の変換」という重要な操作が背景にあることに注意しよう．［６節も参照］

（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. そもそも定性的な情報であるテキスト型データに，潜在的にあるであろう，そこはかとない特徴，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で視覚化することは有効な手段である．一方，この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, TM の目標である知識抽出に有効な指針を示すことが実際にどこまで可能かを常に問うべきである．

たとえば，これを統計ソフトウェアの視覚化情報と比べてみるとよい．多くの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，それらの統計指標の意味解釈の助けとする．一方，TM では膨大な文字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 (はたしてどのように計量化されたか，である)．ここらをどう解決するかが今後の課題である．

少し違った視点から視覚化を考えるのがコホーネン（Kohonen）の提案した SOM マップ（自己組織化マップ，ニューラル・ネットワークの応用）であろう．もちろん，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 マイニングなどと関連してテキスト型データの分析に SOM マップ（Self-Organizing Maps）を適用する例が増えている［Lagus 他（1996），川端・樋口（2003），Murtaugh（2000）］．このような視覚化過程での課題は次のようなことだろう．

- 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グラフィカル表現すること
- 本来は数値化されない仮想的あるいは概念的な情報を可視化すること

- とくに，膨大なテキスト情報があるとき，はたして適切な視覚化が可能か，たとえば無数の単語を布置した図を観察しても簡単には解釈できない(知識取得にならない)
- つまり，ある種の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こと
- そのとき，要約や縮約化に伴う情報の損失をどう評価するか，あるいは客観的に知るか，ここでリダクションの方法を誤ると，誤った解釈を与えることになること

たとえば，このような例を考える．何かの方法で抽出した単語群について，

- 多変量解析手法，たとえば主成分分析や対応分析を使って求めた単語のスコアの布置図を見て解釈するとき
- 布置された単語群の図柄にはとくに意味がなく，単にグラフィカル表現を行ってみたとき
- 単語群を描画するための何らかのアルゴリズムがあるのだが，結果として示されたグラフ表示の意味解釈は分析者の主観に委ねられる場合

と視覚化の内容は様々である．それぞれを比較すること自体にあまり意味がなく（視覚化表現の設計指針が異なる），ここはその分析ツールの開発設計指針の問題となる．

これらについて，現状の市販の多くの TM ツールは，それぞれ視覚化の意義や意味解釈の方法を説明しているものもあるが，総じて明らかではない．ここらの設計指針が曖昧であり，また提供される情報の意味解釈を与える客観情報に乏しい（これを行うことが TM の真の目標であるのに，である）．ここでは，視覚化を考える際の検討課題として指摘するに留める．

（6）　辞書の機能

これも TM にとって重要な要素でありながら，扱いがきわめて厄介な事の一つである．形態素解析や分かち書き処理を行うために，大抵の TM ツールは辞書を備えている．しかし，もっとも問題とされることは，多くの場合，分析対象が非構造的なテキストが多いということである．

一方，高度なコンピュータ化が進んだコーパスが利用できるような場合は，あるいは構造化された文書データベースを利用するような場合は，かなり的確な分析結果が期待できる．TM の本来の対象はこうした整った（構造化された）コーパスや文書データベースを前提とした方法論が主流であるから，非構造的なテキストが多い，調査における自由回答・自由記述文の解析には，さまざまな問題が生じる．

その一つは，いわゆる**同義語・類語の扱い**である．表記や表意の違いがあっても同じことを意味する表現語句をどう扱うかは，かなりの難題である．典型的な例として，Web 調査で取得する自由回答データを考えればよい．どんなに設問を工夫しても，回答の内容はバラエティに富み，様々である．たとえば，「友人」を「友達」「友」「ともだち」「だち公」「仲間」…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．さらに，状況によっては**広義語や関連語**等を，どう整理し関連付けるかも求められる．「家族」「ファミリー」「身内」から「親類」「親族」「血族」「縁者」「父母兄弟」…となってくると，どこまでを類似の語句として括るか，悩ましいことになる．加えて，携帯語，電子メール語やチャット語とあっては，同義語・類語の扱いを精密に考えること自体に無理がある．

TM ツールの側からみれば，この問題をユーザがどう理解し，要求内容がどの水準にあるかを知らねばならない．シソーラス辞書がどのようの整備できるのか，あるいはユーザがどこまで辞書編集を行うのか，さらには**同義語・類語・関連語の扱い**はほとんど考えな

いまま，解析を行うことが可能なのか，…どのレベルで利用できるかを，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，**語彙やコーパス**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，などをいう．であるから，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い，たとえば，国語研究所が発刊している「分類語彙表」などもこの一つであろう．

調査などで，類似テーマで同一パネルに繰り返し自由回答データを取得する，あるいは同一のテーマで，異なる調査対象に意見を聞くなどを考えたとき，当然，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．こうした場合に，コンピュータ処理が可能なコーパスや，それを目的別にいくつか集めたコーポラがあると便利である．あるいはCDとして製品化されたシソーラスやコーパスを補助的に使うこともよいだろう（例：デジタル類語辞典 2003，日本語語彙大系など）．

注：品詞と品詞分類

品詞とは，語を分類し，いくつかのグループの分けたとき，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．

注：コーパス

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．
例：コーパスがどう利用されるかについては，たとえば「コーパス言語学」［中村純作，「現代言語学の潮流」，（山梨正明，有馬道子編），勁草書房（2003），pp233-245.］を参照.

３．３　適用の範囲からみたテキスト型データの様相

筆者のそう多くはない体験でも，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はいない．理由は多々あろうが，大元は日本語分析の困難性にあると考える．換言すると（繰り返しになるが），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］

⑴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，つまり，分析が厄介で，有益な知見も期待しにくい場合.

⑵元来が文字情報であるとき

これには，文学書・文芸書など，新聞・雑誌など，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，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，全文検索などである.

⑶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，たとえば蓄積されていたアーカイヴなどに付帯する情報，データ取得履歴が整理可能なデータ，蓄積した定性情報データベースやそのメタ・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.

⑷調査データ，とくに選択肢型調査との併用

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，市場調査などではもっとも多いと思われるタイプである.

⑸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（Web 調査など）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根本的に異なることである．しかし一方では，現状の TM を用いる限り，テキスト型データであるという本来の特徴を，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も見逃してはならない（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

４．１　テキスト・マイニング・ツールの採用，判断に際しての考慮項目

ここで，TM 向けのソフトウェアの備えるべき要件を考えよう．見方は様々あるであろうが，およそ次にあげるような項目を検討時の目安とすればよい．

⑴拡張可能性（スケーラビィリティ）
- どの程度のデータサイズが扱えるのか
- 処理速度
- 既存アプリケーションとの接合性の自由度
- ネットワーク機能との整合性
- データベース機能の水準
- ドキュメント・ウェアハウスの利用可能性

⑵分析対象とする資源，テキストの範囲
- 扱い可能なテキスト型データの種類
  - 例：テキスト（txt，csv），pdf，SQL/ODBC, html/xml/SGML，等々
    - （*）ODBC：open database connectivity　（データベース接続規格）
    - （*）SGML：standard generalized markup language

⑶既存システムへの互換性
- 特定なプラットフォーム，コンピュータ向けに設計されていないか
- 汎用の PC 上で利用可能か
- とくに，既存の統計ソフトウェアとの接合性，中間作業データ授受
- OLAP などを行うミドルウェアとの接合性

⑷更新サービス
- アップデート情報を常に更新する機能
- 最新情報が常に利用可能なこと

⑸テキストの要約化，視覚化機能
- ドキュメント分類，マップ表示など
- テキスト・ランドスケープ（俯瞰機能）
- クラスター化結果の要約化・視覚化
- 一般的なビジネス・グラフ他

⑹解析機能の充実度

・初動探査の標準機能（データランドリ，集計機能など）
・自然言語処理系の各種機能（とくに形態素解析）
・パターン認識，多変量解析，多次元データ解析などの解析機能
・解析手法の精密さ・正確さ，理論的背景・記述の正確性と透明性
・解析結果の解釈支援の機能
・知識組織化のための支援機能

⑺辞書機能
・辞書の利用可能性，その範囲
・コーパスが利用できるか，あるいはコーパス作成の補助機能
・シソーラスへの対応，シソーラス利用可能性
・辞書機能のユーザー・インターフェース

⑻多言語対応
・一つの言語だけでなく複数の言語の解析に対応
・多言語の比較分析の可能性

⑼価格と処理機能の関係（コスト・パーフォーマンス）
・機能操作性と価格の関係
・解析結果の有効性

４．２　テキスト・マイニング・プロダクツ

TM のソフトウェアは国内，国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．表１，表２は，保田（2003）によるサーベイを元に，それに欧米のソフトも加えて一覧とした表である．ここで，備える機能ないしは得意とする分析対象で分類すると以下のようになる．

- ソフトウェアの規模が大きい統合化システム
  （文書データベース，データウェアハウスなどの利用）
- 機能が自然言語処理系に中心がある
- 調査データの分析向き，統計処理機能を含む
- 価格帯の幅が広い（非常に高価から廉価なものまで）
- 欧米に比べてシェアウェアが少ない

なお，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（たとえば，［1］,［8］,［14］,［23］）．とくに， U. Nahm［23］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，評価など）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」に，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．これらの内容分析の研究や関連ソフトは，TM を考えるうえで無視できない領域である．

６．むすび

６．１　真のテキスト・マイニングの目指す方向とは？

テキスト型データの分析は，なぜ厄介で手に負えないのであろうか．理由の一つは，そ

もそも定性情報として表現，描写が困難な抽象概念が多いこと，つまり**計量化がそう容易ではないこと**がある．表記された内容，概念間の微妙な捉えがたい情報を表す“無数の”組み合わせが考えられることがある．

たとえば，自由回答設問を考えても，調査者の意図に反して，回答内容，表現方法は実に多様であり，しかも同じことを述べるにも類似概念を表わす多数の表現方法がある．多変量的な言い方をすれば，高次元性があり数万～数十万もの特徴の組み合わせの可能性がある中で，知識発見やその結果の知識組織化をどう行うかがある．これはたとえ，個別的には優れた技術要素があっても，それらを有機的に融合化して使いこなすにはかなりの困難性を伴うということである．

しかし，現状を見ると（とくに国内の），いかにも安易な発想で TM が“役に立つ”と考える風潮がないとはいえない．事が単純に分かればよい（分かった気持ちになる），簡単な事がよい，主観的であれ分かり易いことがよい，という発想がなくはないか．一方，その対極として，何事も精密かつ厳密であるべき，との考え方もある．しかしいずれもが極端，どちらも説得力があるとはいえない．要は利用者・分析者の要求に応じて“的確に”，“信頼できる”情報を提供できることが望ましいのであるが，現状の TM は（とくに国内の多くのソフトが行う TM），いかにも中途半端である．

理由は，第一に，利用者側の方法論への正しい理解が十分ではないと考えられること，次に，解析ツール提供者側にも，本来 TM が満たすべき要件を十分に消化した設計指針があってソフト開発に取り組む姿勢が今ひとつであること，そして，「ノウハウ」という都合のよい言葉に保護されて，ソフトの中味が暗箱化され**「何を分析したかが」明示的に見えない**こと，などがある．

現状の多くの TM ツールの盲点は，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，その解釈は客観的か）に問題がある．TM が本当に「テキスト型データから知識発見，そして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．

そもそも TM あるいは DM の最終目標とされる「知識発見，価値ある知見の探査」とは，何をいうのであろうか，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であろうか．あるいは真の TM の目指すべき道はどこにあるのだろうか．一つの試みとして，表３を作ってみた．

ここでは，TM が扱うであろう「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．

まずここで明らかにしたいことは（既述のように），既存の TM ツールの多くは，所与のテキスト型データを，その生の情報を扱うことを行うのではなく，一度「数量化・計量化の手続き」を経て，従来型の DM などの方法論が適用可能な形に情報を変換して扱うということがある（つまり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．

たとえば，単純な操作としては，語句・単語の抽出でコード化，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ことである．ここでは，伝統的な自然言語処理の技法や情報検索技術が利用される．別の方向として，テキスト情報を多変量解析や多次元データ解析手法を用いて，数量化を行い，同時に情報縮約・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計量化されたデータとして処理するというものである．確かにこうしたアプローチは，類似性や関連性の“単純なパターンや規則性の発見”には有効である．とくにデータベースやコーパスとして高度に構造化されたデータセットについては効力を発揮するであろう．

一方，我々が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，これに対する的確な解をここで即座に提供できるものではない（表３のセル「真の TM と

は？」に相当）．いま指摘できることは，現状の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方法だけではなく，“何か別の道”があるだろうとしか言えない．

ただ，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”を拾い上げるような**「仕組み作り」を考えること**ではなかろうか．

一例として，最近の体験を示そう．ある自治体で，様々なルートを通じて「市民の声」を集めてきた．電話，投書，電子メール，市庁来訪，…と様々である．集まった情報を眺めると，少なくともある特徴，傾向が見える．たとえば，悪臭対策，騒音対策，地下鉄問題，とアイテムとしては多種多様である．しかし，分析を深めて，ではこうした膨大な意見データから，政策決定に本当に有効な意見が集約されるかというとそう簡単ではない．つまり真の知識発見とは何かという基本的な問題に突き当たるのである．換言すると「ただ待っていても適切な意見は出てこない」という常識的な答えが出るだけである．発言者である市民の意見の述べ方，提案の仕方の根気の要る指導に始まり，それをリアルタイムにうまく汲み取る仕組み作り，という基本的な課題**「データ取得をどう行うか，その仕組みの設計は」**をクリアすることから始めるということである．

つまるところ，これは始めに述べた，**データ科学**の精神であり，現在の KDD，TM，DM に抜けている部分である．「顧客の声」「生の声」を TM で知るといった美味しい言葉に惑わされることなく，真のテキスト・マイニングとは何かを再考すべき時期にある．今までも多くの方法論が高い期待をもって登場したが，その大半はいつの間にか忘れられている．TM が同じ轍を踏むことなく，しっかりと育つことを期待したい．また，過剰な期待も困るが，粗雑に扱って一過性の流行りものに終わらせてはならないのである．

【参考文献】

[1] Ah-Hwee Tan, Text Mining: The state of the art and the challenges.

[2] Bhavani Thuraisingham (1999), *Data Mining Tecnologies, Techniques, Tools, and Trends*, CRC Press.

[3] Christine Fellbaum (ed.) (1998), *WordNet: An Electronic Lexical Database*, The MIT Press.

[4] Feldman, R. and Dagan, I. (1995): Knowledge discovery in textual darabeses (KDT), in *the Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada, August , AAAI Press, 112-117.

[5] Dan Sullivan (2001), *Document Warehousing and Text Mining*, John Wiley.

[6] Fionn Murtagh (1999), Data Mining, Statistics and Data Science, in the *Proceedings of ISM Syposium: Data Mining and Knowledge Discovery in Data Science*, organized by the Institute of Statistical Mathematics, Tokyo, 1-12.

[7] Inderjit S. Dhillon, Subramanyam Mallera, and Rahul 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, pp191-216.

[8] Ingrid Renz and Jurgen Franke (2003), Text Mining, in *Text Mining: Theoretical Aspects and Applications*, 1-19, Physica-Verlag.

[9] Kimberly A. Neuendorf and Paul D. Skalski (2002), *The Content Analyisis Guidebook*, Sage Publications.

[10] Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen (1996), Self-Organizing Maps of Document Collection: A New Approach to Interactive Exploration, in *Proceedings Second International Conference on Knowledge Discovery & Data Mining*, AAAI Press, pp238-243.

[11] Ludovic Lebart, André Salem, and Lisette Berry (1998), *Exploring Textual Data*, Kluwer Academic Publishers.

[12] Marti A. Hearst (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*, pp59-66, Denver, CO, May 1995. [http://www.sims.berkeley.edu/~hearst/papers/tilebars-chi95/chi95.html]

[13] Marti A. Hearst (1998), Current Topics in Information Access, in SIAM Academic Course 296a-5-3, Fall 1998.

[14] Marti A. Hearst (1999), Untangling Text Data Mining; This paper appears in the *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 20-26, 1999 (invited paper).

[15] Michael P. Oakes (1998), *Statistics for Corpus Linguistics*, Edinburgh University Press.

[16] Nong Ye (ed.) (2003), *The Handbook of Data Mining*, Lawrence Erlbaum Associates, Publishers.

[17] R. Harald Baayen (2001), *Word Frequency Distributions*, Kluwer Academic Publishers.

[18] Robert P. Weber (1990), *Basic Content Analysis* (second edition), Series: Quantitative Applications in the Social Sciences 49, Sage University Paper.

[19] Roel Popping (2000), *Computer-assisted Text Analysis*, Sage Publications.

[20] Roren Feldman (2003), Mining Text Data, in *The Handbook of Data Mining*, Nong Ye (ed.) ,Lawrence Erlbaum Associates, Publishers, 481-518.

[21] Stone Analytic, Inc., Evaluating Text Mining Applications [http://www.secondmoment.org/atats-column/stats-textmining.php]

[22] Tony McEnery and Andrew Wilson (1997), *Corpus Linguistics*, Edinburgh University Press.

[23] U. Nahm, 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/]

[24] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996), Knowledge Discovery and Data Mining: Towards a Unifying Framework, in *Proceedings Second International Conference on Knowledge Discovery & Data Mining*, AAAI Press, pp82-88.

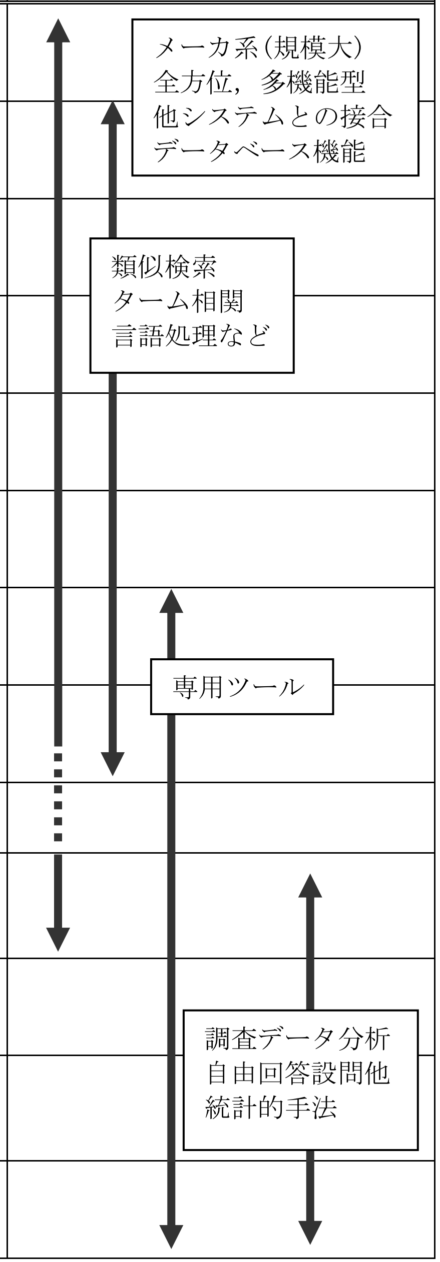
[25] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996), The KDD Process

for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, **39**, 11, 27-34.
[26] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasary Uthurusamy (1996), *Advanced Knowledge Discovery and Data Mining*, AAAI Press/MIT Press.
[27] Usama Fayyad and Ramasamy Uthurusamy (1996), Preface for "*KDD-95: Proceedings First International Conference on Knowledge Discovery & Data Mining*," AAAI Press.
[28] 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/］
[29] 伊藤雅光（2002），計量言語学入門，大修館書店．
[30] 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」．
[31] 国立国語研究所（1964），分類語彙表，国立国語研究所資料集．
[32] 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，2001，㈶統計情報研究開発センター，2-9．
[33] 山梨正明，有馬道子編（2003），現代言語学の潮流，勁草書房．
[34] 柴田武，山田進編（2002），類語大辞典，講談社．
[35] 小池清治，小林賢次他編集（1997），日本語学キーワード事典，朝倉書店．
[36] 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．
[37] 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 巻，3 月，163-181．
[38] 樋口耕一（2003），内容分析とコーディング，統計数理研究所公開講座資料「テキスト型データのマイニングとその応用」，115-153．
[39] 大隅昇，Ludovic 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．
[40] 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．
[41] 大隅昇（2000），定性情報のマイニング−自由回答データの解析−，ESTRELA，74 号，2000 年，5 月号，14-26．
[42] 町田健（2003），コトバの謎解き　ソシュール入門，光文社新書．
[43] 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．
[44] 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．
[45] 飽戸弘編著（1994），食分化の国際比較，日本経済新聞社．
[46] 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．
[47] 北原保雄（監修）（2003），日本語の使い方，考え方辞典，岩波書店．
[48] 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．

※ 下記のテキスト・マイニング研究会のホームページから，多数の関連文献，ホワイト・ペーパー，ソフトウェア・リンクサイトなどの情報が取得できる．
http://wordminer.comquest.co.jp/

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・サービス名	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，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	SurveyAnalyzer(サーベイアナライザー) 自由記述アンケート分析システム ［TopicScope として改編］	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（単語や語句）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	



※ 会社名、製品名等は、各社の登録商標もしくは商標

◆表 1 に記載の情報の確認・整理

1） 2003 年～2004 年あたりの情報であり，やや内容が古い.
2） 以下に新たに確認した情報を分かる範囲で整理しておく.
3） 多くの場合変更や更新が激しいので，検索エンジンなどで再確認するとよい.

表 1 補足　主要なテキスト･マイニング･ソフトウェアの一覧（国内）

No.	変更／追加	旧・製品・サービス名	製品・サービス名	開発元・販売元
1	変更	Symfoware Text Mining Server テキスト・マイニング・ソフトウェア	Interstage Navigator Explorer Server （*）この名称で，統合化システムとして機能 （*) この中にデータマイニング, テキストマイニングが含まれる.	富士通(株)
2	変更	TAKMI テキスト・マイニング・システム	IBM TAKMI（Text analysis and Knowledge Mining） テキストマイニングシステム	日本アイ・ビー・エム(株)，東京基礎研究所（TRL）
3	追加		テキストマイニングシステム DIAMining®EX	三菱電機インフォメーションシステムズ(株)
4	変更	VextSearch テキスト・マイニング・ツール	VextMiner テキストマイニング・マーケティング・ツール （*）VextSearch，VextMiner，VextContact として適用.	クオリカ(株)
5	追加		Text Mining Studio (for Windows)	（株）数理システム
6	変更	SurveyAnalyzer(サーベイアナライザー) 自由記述アンケート分析システム ［TopicScope として改編］	CSVIEW シリーズ （*）このソリューション・ツールの一つとして「CSVIEW/テキストマイニング」がある.	日本電気(株)

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 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
7	SAS Text Miner	SAS 社が提供するテキスト・マイニング・ソフト （*）SAS Enterprise Miner の部分機能として用いる	・ File processing ・ Text parsing ・ Dimension reduction (transformation) ・ Document analysis

表３　テキスト・マイニングの位置づけ

<table>
<tr><th colspan="2" rowspan="2">データの型
（種類）</th><th rowspan="2">対　象</th><th colspan="3">目　標　と　対　応</th></tr>
<tr><th>対応方法
適用の方法論</th><th>単純なパターンの発見</th><th>意味ある複雑な情報の発見</th></tr>
<tr><td>数値型データ</td><td>質的データ
（名義，順序）
量的データ
（区間，比例）</td><td>・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</td><td>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</td><td>・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</td><td>・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</td></tr>
<tr><td rowspan="3">非数値型データ</td><td rowspan="2">テキスト型データ</td><td>小説，自由記述文，自由回答など
（非構造的）</td><td rowspan="2">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</td><td rowspan="2">文書要約
文書分類
内容分析
全文検索</td><td rowspan="2">◆真の TM とは？
さらに有効な TM の方法はあるのだろうか？</td></tr>
<tr><td>一般文書類
（構造的）</td></tr>
<tr><td>非テキスト型データ</td><td>画像，音声など</td><td>計算機言語学
言語学
音声学</td><td>自動翻訳技術
多言語間翻訳</td><td>自動通訳機能
意味理解
音声認識
画像認識</td></tr>
</table>

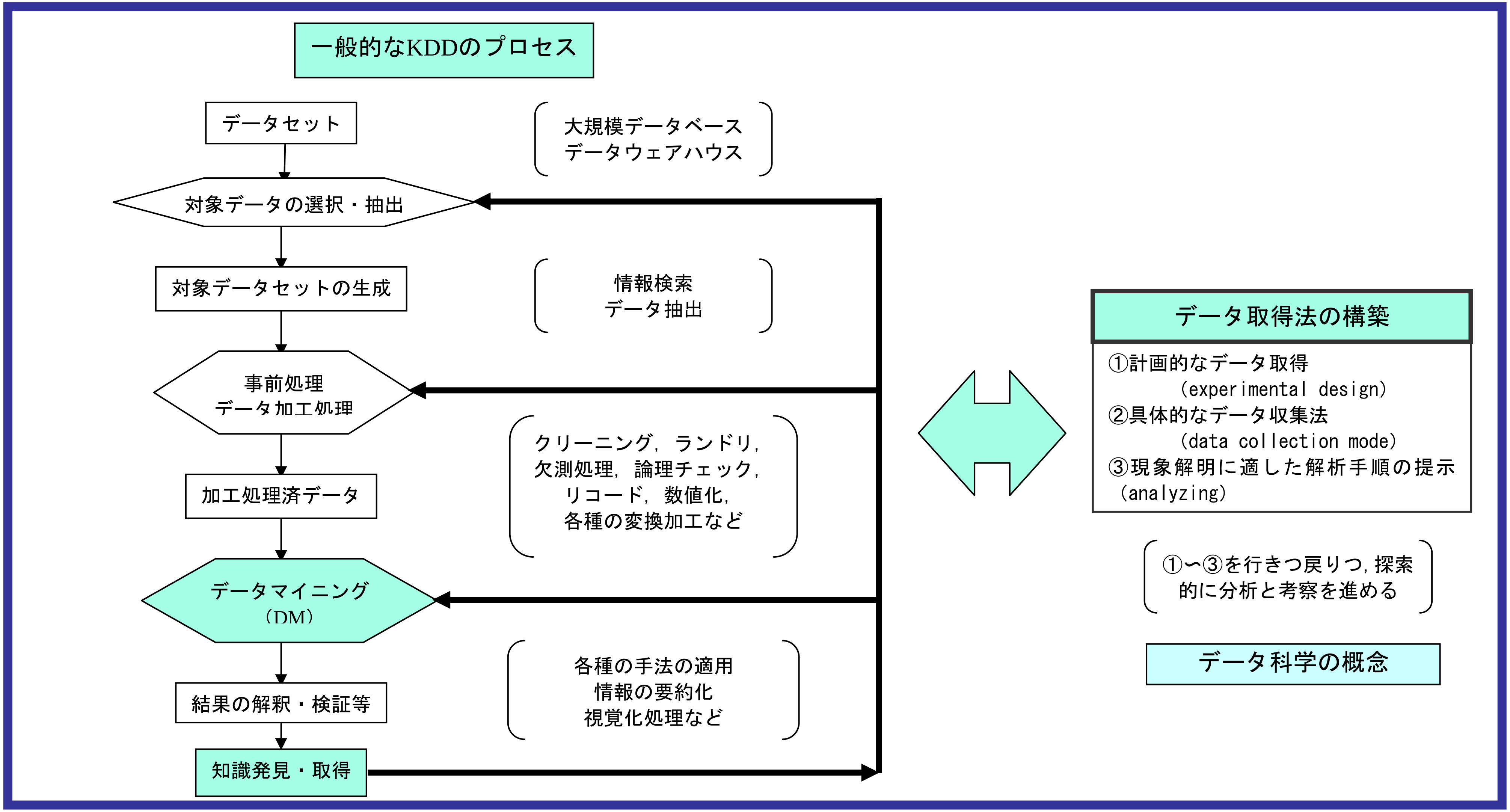


図１　KDD における DM の位置づけとデータ科学の関係

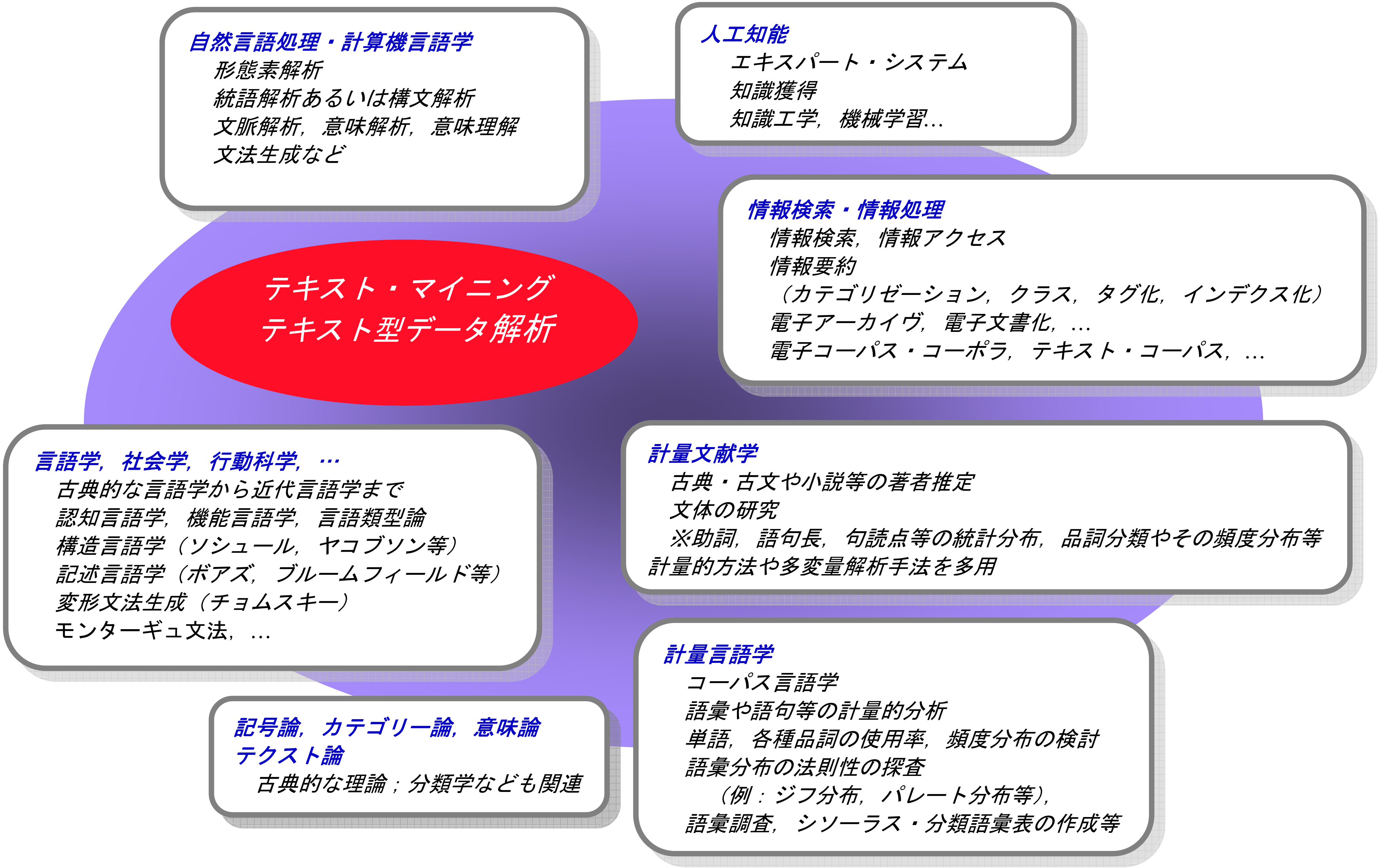


図２　テキスト・マイニングに関連する研究分野

図３　テキスト・マイニングに利用される主な方法論

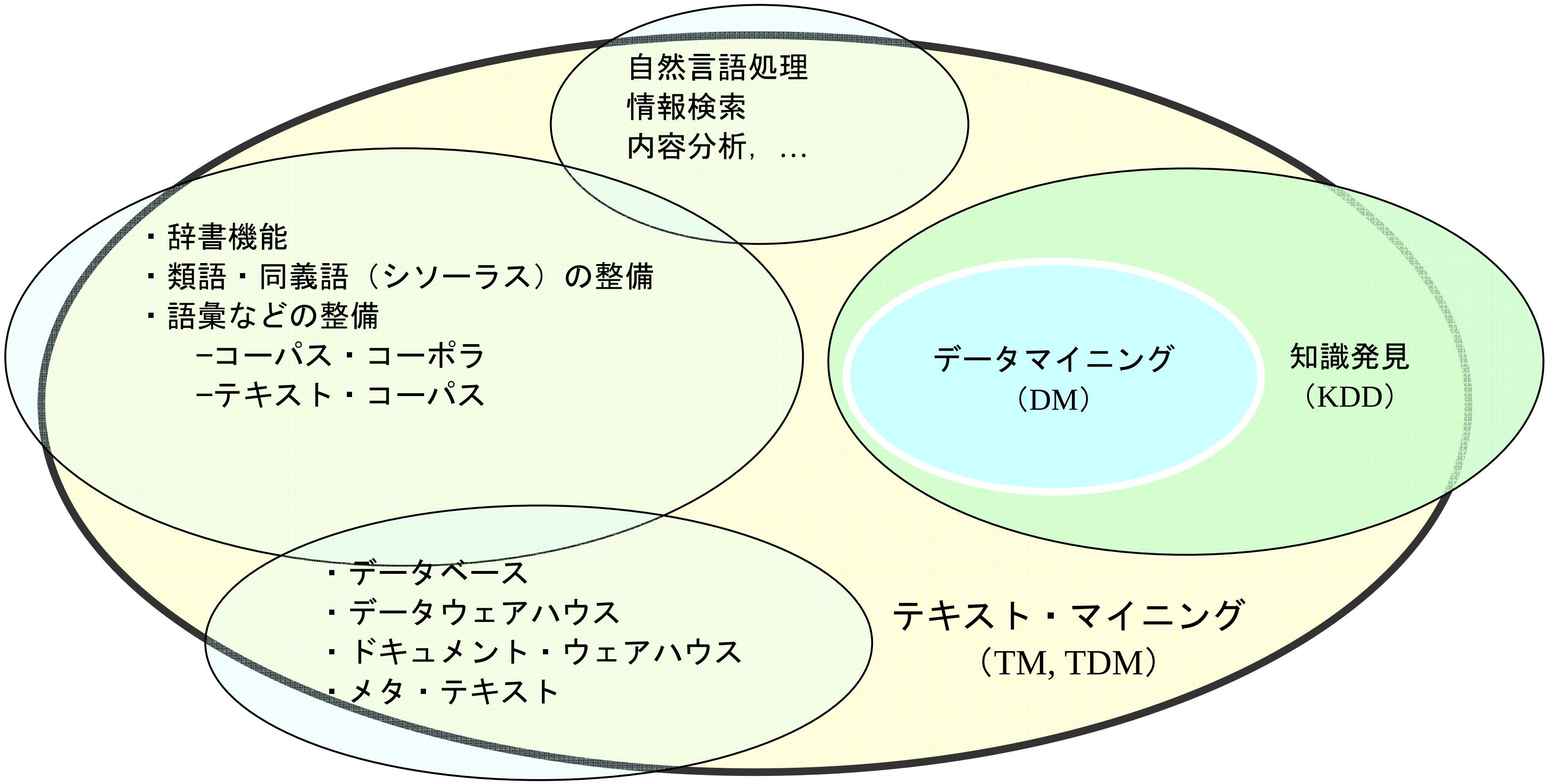


図４　KDD，DM，TM の関係