特集解説

テキスト・マイニングの概要

非会員　保田　明夫*

Reviewing “ Text Mining ” : Textual Data Mining

Akio Yasuda*, Non-member

The objective of this paper is to give overviews of text mining or textual data mining in Japan from the practical aspects. Text mining is the technology utilized for analyzing large volumes of textual data applying various parameters for purpose of withdrawing useful knowledge and information. The essence of “Mining” is "the discovery of knowledge or information." And target of text mining is to objectively discover and extract knowledge, facts, and meaningful relationships from the text documents. This paper summarizes the related disciplines and application fields which are applied in text mining, and introduces features and application examples of text mining tools.

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，テキスト・マイニング・ツール
Keywords：text mining, textual data mining, data mining, text mining tool

1.　はじめに

インターネットやブロードバンドの時代を迎え，日本語の電子的処理環境や言語情報処理分野の諸研究が進展したことから，テキスト型データの取得方法や解析手法への関心が高まった。とくに，社会調査（意識調査・行動分析等）や市場調査等の自由回答型・自由記述型データ，あるいは，コールセンターやお客さま相談室などで収集した生活者や消費者の「生の声」や「本音」，さらには，営業報告書や議事録・会議録，製造部門での工程管理や品質管理の定性情報など，多種多様かつ膨大なテキスト型データを経営に活かすことを狙いとして，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）への期待が高まっている。

その一方で，TM は実のところ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）があり，DM は，関連書籍も数多く，コンピュータ・ソフトも多数登場しているが，TM 同様，分かったようで漠としたところがある。とくに，統計的データ解析の各種方法論と DM の違いは，いまひとつ明らかではなく，このことが TM をより捉えどころのないものにしていると思われる。

ここでは，TM の背景から関連する研究分野や方法論などを俯瞰すると同時に，TM や TM ソフトウェアの特徴や機能，技術的な諸要素，諸事項について総合的に要約する。

* (株)平和情報センター
〒112-0002　東京都文京区小石川 1-3-21
Heiwa Information Center Co., Ltd.
1-3-21,Koishikawa, Bunkyo-ku, Tokyo 112-0002

2.　テキスト・マイニングの背景

テキスト・マイニングの特徴は，膨大なテキスト型データを様々な観点から分析し，役に立つ知識や情報を見つけ出すことにある。一方，実務的な面からみれば，単なる技術ではなく，CRM（Customer Relationship Management）や KM（Knowledge Management）などと同様に，顧客の信頼を獲得し，維持し続けるためのマネージメントの一つとして捉えることもできる。

〈2･1〉　テキスト・マイニング（TM）とは

TM はデータ・マイニング（DM）から派生した方法論であるとの見方がある。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。いくつかの定義例を挙げると以下のようなことである（Neri, Nahm, Ye[1]（2003），Sullivan[2]（2001）他）。

（１）大量のテキスト，文書（ドキュメント）など，「生（原始）」のテキスト型データを直接操作し（データ処理），潜在する隠れた事実や関連性を発見する。

（２）大規模データベースに蓄積された膨大なテキスト（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。

（３）新たな知見・知識を得る一連の接近法であり，大量のテキスト型データを計量化し，探査の過程・推移や結果の要約・視覚化を通して，データ構造や内在する情報を理解可能な顕在情報に変換する。

（4）自然文（自然言語テキスト）から，規則性，典型，傾向などの特徴抽出や分類を行うことにより，有用な知識，知見を引き出す。

（5）データベースやデータウェアハウスから，DM 技法を活用し，顕著なパターンの発見を通して，新たな（これまでは未知の）有用な事実，知識を得る。

このようなことから， TM の特徴として，以下のようなことがあげられる。

・大量の「生（原始）」のテキスト処理
・大規模（ドキュメント）データベースの活用
・テキスト・コーパス（コーポラ）や辞書の利用
・パターンの探査，特徴抽出
・規則性，類似性，関連性の発見
・例外，変則，特異性の発見
・有用かつ新たな（予期しなかった）事実や知識の発見
・情報の視覚化，可視化
・情報の組織化，知識発見，知識獲得

計算機言語学の研究でも著名な Hearst[3](1999)によると，TM の目的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。また，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，「探索的アプローチ」の重要性を示唆している。

〈2･2〉 データ・マイニング（DM）とテキスト・マイニング（TM） DM と TM は，「鉱脈探し」（mining）という共通語があり，TM のある部分，とくにデータ処理や解析エンジン（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。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が，DM については多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できない。ここでは DM を概観し，それをとおして TM の位置づけを探る。

近年，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。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，知識発見の方法論として DM が提唱されてきた。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）がある。狭義の 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を把握し，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 に知識発見の道具立てとしての DM が加わり，今の新たな 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（アルゴリズム）など，予測，解析，発見，検証などの関連技法の集合体であり，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としてとらえることができる（Fayyad，Piatetsky-Shapiro 他[4)-6)]（1996））。

ここで，従来からの統計的手法，統計的データ解析（とくに探索的方法論）と KDD の考え方がどう異なるのかがひとつの疑問となる。DM は，統計学・統計的データ解析の関連手法を適用するという意味で密接な関係がある一方，多くの書冊によると，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」「大規模データベースやデータウェアハウスを活用する」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。

膨大なデータの中から「金の鉱脈」（有用かつ新たな事実や知識）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もある。DM に関連する多くの書に「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」「どうすればゴミではないデータをあつめられるのか」という極めて素朴な疑問にたどり着く。しかし，DM の多くの方法論にはこれに対する明解な答えはなく，「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ているように思われる。

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，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議論にとどまってしまうこともある。

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」が重要であり，それを前提とした「データ主導型」の解析過程が必要である。これについて，統計的データ解析の領域では，「データ科学」（data science）という発展的なパラダイムがある（林知己夫[7]（2001））。「データ科学」では，現象解析の基本は「データ」にあると考え，「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。その要点は，

（1）データをどう計画的に取得するか
（2）データを具体的にどう集めるのか
（3）現象解明に適した解析法はどうあるべきか

の3つにあり，これらを探索的に繰り返し行う過程にある。

要は，DM にせよ，TM にせよ，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。

近年，国内では，社会調査や市場調査，あるいはコールセンターやお客さま相談室での「生の声」の分析活用などの分野で，TM の活用に高い期待がもたれている。これに応えるには，自由回答設問や記述方法の設計はもとより，データ収集環境やデータベース構築技法など，幅広い視野に立ったデータ取得機構の設計指針が重要となる。

3. テキスト・マイニングの関連研究分野と方法論

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，考え方も様々である。また実際に TM は，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。ここでは，TM と関連する研究分野，利用される方法論，さらには適用範囲の面から TM を概観する。

〈3･1〉 関連する研究分野 TM に関連する研究分野は，自然言語処理・計算機言語学，エキスパートシステム，知識獲得・知識工学，機械学習などの人工知能（AI），情報検索（IR）・情報処理，計量文献学，計量言語学，さらには，言語学，社会学，行動科学，記号論，カテゴリー論，などなど実に多彩である。さらにそれぞれの分野の諸要素が相互に絡み合っている（図 1 参照）。

TM はテキスト型データを対象とすることから当然ではあるが，自然言語処理や計算機言語学と密接に関連しており，形態素解析，統語解析（syntactic analysis, parsing），構文解析，意味解析などの技法が TM には必要となる。

「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前のことだが，それ以前も様々な研究が行われてきた。とくに文書情報管理・検索機能は重要で，例えば KWIC（keyword in context）やコンコーダンス（concordance）により，語句の文章内での使い方や共起の関係を調べ，またあわせて共起語，コーパス頻度，共起頻度の閲覧や統計的指標なども観察する。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[8]（2003），Popping[9]（2000），

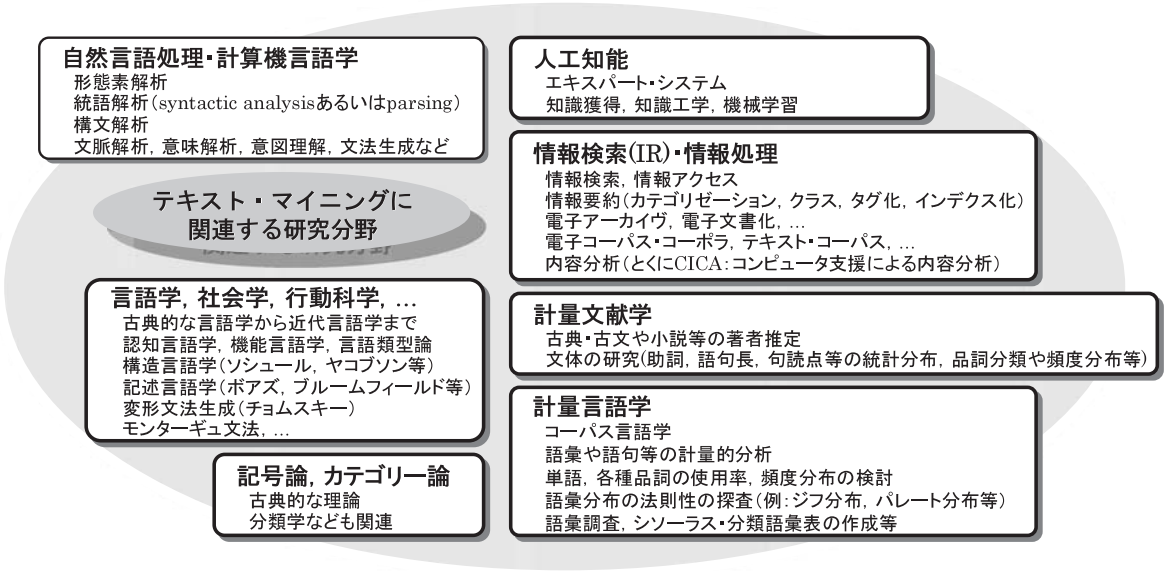

図 1 テキスト・マイニングに関連する研究分野
Fig. 1. the related disciplines of textmining.

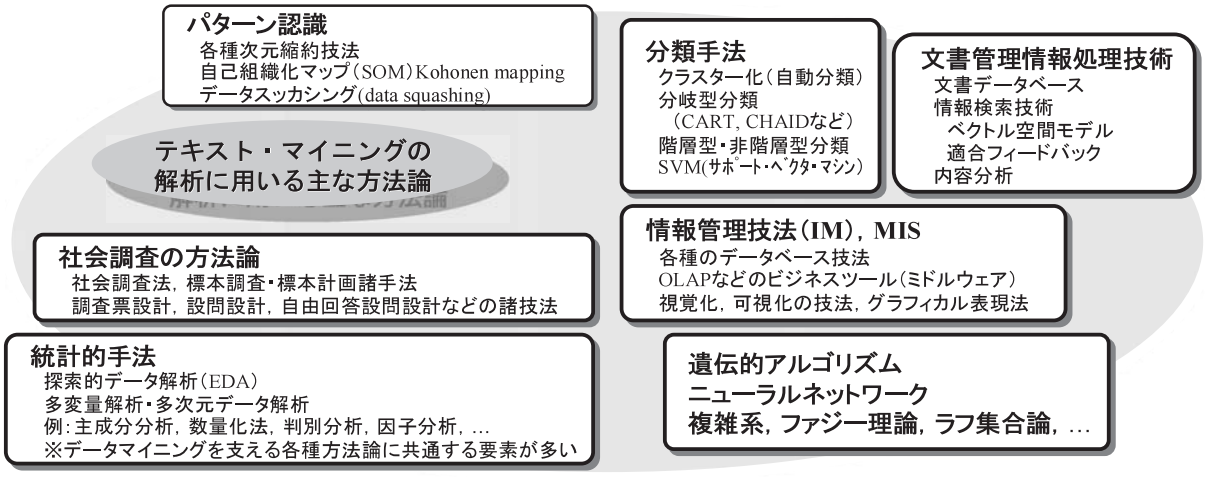

図 2 テキスト・マイニングの解析に用いる主な方法論
Fig. 2. the methodology of textmining.

Neuendorf[10]（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関係などの機能がTMに活用されている。

〈3・2〉 利用される方法論 TMの解析に用いる方法論は，関連分野と不可分の関係にあり，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，社会調査の各種調査技法や自由回答設問設計等，各種統計的手法（とくに，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），統計的言語処理の精度向上を狙いとした機械学習法や分類学習法としてのサポート・ベクタ・マシン（SVM:support vector machine）がある。また，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），膨大なテキストから必要な情報を検索する方法としてのベクトル空間モデル（vector space model）や適合フィードバック（relevance feedback），情報管理技法（IM），情報管理システム（MIS），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。

この他にも，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TMの解析に用いられている（図2参照）。

このように，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と同じである。TMというなにか特定な方法論があるように思われるが，実はそうではなく，それぞれの分野の利用技術の特色を活かし，また各方法論の利点と課題を把握し，「どう使いこなすか」がTM活用の鍵となる。つまり「どんな方法を使うか」ではなく，分析目的に応じて「何故，何をするのか」そして「どう使いこなすか」が重要である。

〈3・3〉 適用範囲・応用範囲 TMの適用範囲は幅広い。報告書，研究論文に登場するキーワードを列記すると，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，知的検索・類似文書検索などがあり，さらに，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）など，TMの応用分野は実に様々な分野に拡がっている。

とくに最近では，Webページ間の相互の引用関係に着目することによって，膨大なWeb空間を整理体系化・俯瞰するWebマイニングや，Webサービスプロバイダ，電話会社やコールセンターの通信記録，あるいは，金融や流通分野の取引記録に代表される大量の電子化データの流れであるデータストリームから有用な情報を少ない資源で取り出そうとするデータストリーム・マイニング（data stream minig）が着目されてきた。

既に欧米では様々な分野でのTMの活用がはじまっている。とくに，「構造化された（structured）」膨大な文書データベースやドキュメント・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のTMがある。（Ye[1]（2003），Sullivan[2]（2001））。

一方，日本国内では，社会調査や市場調査における調査データ（自由回答設問）の分析やコールセンターやお客さま相談室等で収集した「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，今後の期待・発展分野であろう。文献[11]（大隅，Lebart（2000））には，自由回答データの取得方法や日本語文章・テキストの解析方法の方向性，統計的データ解析との関係など，様々な研究成果や活用事例の詳細な報告がある。

4. テキスト・マイニングの機能

TMは，前述したような様々な研究分野や方法論を組み合わせた複合的な技術であり，その実装方法や適用範囲も多岐にわたる。ここでは，国内でTMの活用が期待されている社会調査や市場調査，あるいはコールセンター等で収集された「非構造的なデータ」を扱う場合について，TMの機能をその処理の流れから考察する。

テキスト・マイニングの処理の流れは，テキスト型データから情報や概念を抽出するステップと抽出された情報や概念を解析する（マイニング）２つのステップに分けることができる。この２つに，データの基本情報(背景)や解析過程・結果を分かり易く提示するための情報の分析・可視化を加えた３つの機能がテキスト・マイニングの主要な要素である。

〈4・1〉 情報や概念の抽出 分析対象のテキスト型データを形態素解析や構文解析などを用いて，その内容をあらわす情報や概念を抽出する。構造化・形式化されていないテキスト型データを次の解析ステップ（マイニング）で扱えるようにするための数値変換処理であり，テキスト・マイニングの特徴的（象徴的）な機能である。

単に類似文書（テキスト）の検索や分類を行うのであれば，形態素解析による単語分割を行わなくても，N-gram法（N文字が隣接して生じる文字の共起関係）や字種切り法（句読点などの区切り符号，漢字，カタカナ，ひらがな，英字，数字などの辞書の切れ目を利用する）などにより切り出された文字列を利用する方法もあるが，情報や概念の抽出にあたっては，単なるキーワードの抽出だけでは不十分であり，単語の同義性や多義性を考慮に入れた抽出が必要になる。例えば，「米」「米国」「合衆国」などを「アメリカ」という一定の表現に置換することで同義性を吸収したり，「米ー食物」または「米ー国名」というように意味属性を加えて「米」という語（文字）の持つ多義性を解消したりする。

また，「性能と価格は良いが，デザインが悪い」といった文章の場合，何が「良い」のか，何が「悪い」のか，語句

間の関係を表す係り受け情報や，文節や文章間の複合概念まで抽出することも必要になる。さらに，概念や情報を端的に表現するという意味では，文書要約や自動タイトル付けなどの技術を活用する場合もある。

〈4・2〉　情報や概念の解析（マイニング）　ここでは，抽出された情報や概念から，今までに知られていない新しい事実や知識を得るためのマイニングを行う。

相関分析，クラスタリング，クラス分類，時系列分析など，解析の目的とデータの質や量，抽出した情報や概念の特徴に合わせて，データマイニング・アルゴリズムや多変量解析手法，あるいは統計的手法を適用する。

ここでクラスタリングとは類似のパターンを教師なし学習アルゴリズムによりグループ化すること(clustering)をいい，クラス分類とは教師付きアルゴリズムにより入力パターンを識別する処理(判別：discrimination)をいう。例えば，前者は文書群を情報や概念の類似性により仕分けるような場合のことであり，後者は既存の分類体系に振り分けるような場合に相当する。いずれも欠かせない機能である。

〈4・3〉　解析結果の分析・視覚化　マイニングで得た結果を様々な視点から分析し可視化（視覚化）することにより，自明で興味のない事実と，価値ある新たな知識との違いが明らかになる。また，解析結果を導いた過程やルールの説明機能（例えば，解析に用いた単語群やサンプル，抽出した情報や概念，適用ルールや評価指標，根拠となる元のテキスト型データの提示など）も重要である。

とくに，解析で得た類型の典型や多数派の傾向だけではなく，少数派の特徴を知ることも大切であり，単なる文章要約や分類結果の表示だけではなく，個々の元のテキスト型データやそこから抽出した情報や概念に容易に立ち戻れるような構造の視覚化が必要となる。

また，テキスト型データはテキスト情報として独立して存在しているのではなく，アンケートであれば属性や選択肢型設問などで得た数値型データ，文献であれば書誌・所蔵情報などから数値化されたデータなど，定量的なデータとともに存在している。従って，間接的ではあっても定量的分析（データ・マイニング）の理論や知識の援用を得た

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）
Table 1.　Textminig software (in Japan).

No.	製品・サービス名	開発元・販売元	特　　徴
1	Symfoware Text Mining Server ﾃｷｽﾄﾏｲﾆﾝｸﾞｿﾌﾄｳｪｱ	富士通㈱	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	㈱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 ﾃｷｽﾄﾏｲﾆﾝｸﾞｼｽﾃﾑ	日本ｱｲ・ﾋﾞｰ・ｴﾑ㈱	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 ﾅﾚｯｼﾞﾏﾈｼﾞﾒﾝﾄｼｽﾃﾑ	㈱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://pf.toshiba-sol.co.jp/prod/km2/index_j.htm		
5	Knowledgeocean(ﾅﾚｯｼﾞｵｰｼｬﾝ) ﾅﾚｯｼﾞﾏｲﾆﾝｸﾞ支援ｼｽﾃﾑ	㈱NTT ﾃﾞｰﾀﾅﾚｯｼﾞ	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/knowledgeocean/		
6	IVMap 情報検索ｼｽﾃﾑ	三菱電機㈱	文書の類似性に基づく大量文献の２次元マップ(自動分類)表示(kohonen の SOM に類似したアルゴリズム)
7	MiningPro21 文書ﾏｲﾆﾝｸﾞｼｽﾃﾑ	日本ﾕﾆｼｽ㈱	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/		
8	CB Market Intelligence ﾃｷｽﾄ・ﾏｲﾆﾝｸﾞ・ｿﾘｭｰｼｮﾝ	㈱ｼﾞｬｽﾄｼｽﾃﾑ	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/		
9	VextMiner ﾃｷｽﾄﾏｲﾆﾝｸﾞﾂｰﾙ	ｸｵﾘｶ㈱	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/products/txt/marketing/index.html		
10	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．
11	Survey Analyzer(Topic Scope) 自由記述ｱﾝｹｰﾄ分析ｼｽﾃﾑ	日本電気㈱	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.labs.nec.co.jp/DTmining/products/s_analyzer/		
12	Text Mining for Clementine ﾃｷｽﾄﾏｲﾆﾝｸﾞﾂｰﾙ	ｴｽ・ﾋﾟｰ・ｴｽ・ｴｽ㈱	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		
13	TRUE TELLER(ﾄｩﾙｰﾃﾗｰ) 統合型ﾃｷｽﾄ・ﾏｲﾆﾝｸﾞ分析ｼｽﾃﾑ	㈱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/		
14	WordMiner(ﾜｰﾄﾞﾏｲﾅｰ) ﾃｷｽﾄ型ﾃﾞｰﾀ解析ｿﾌﾄｳｪｱ	日本電子計算㈱	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://wordminer.comquest.co.jp/		

※会社名，製品名等は，各社の登録商標もしくは商標（順不同）

表2　欧米のテキスト・マイニング・ソフトウェアの例
Table 2.　Textminig software (in Europe and America).

No.	製品・サービス名 / 開発元・販売元	特　　徴
1	Sphinx Survey Plus2 & Lexica Le Sphinx Développement SCOLAR http://www.sphinxdevelopment.co.uk/functions1.htm	・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 http://www.enst.fr/egsh/lebart/	・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
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/products/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

※会社名，製品名等は，各社の登録商標もしくは商標（順不同）

いこと，あるいは比較可能となっていること，つまり，定性的（テキスト型）データと定量的データの相互関連性の分析が重要になる。

5.　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。ここでは，個別の詳細機能については省略し，TM ソフトウェアの特徴や課題について概観する。

〈5･1〉　TM ソフトウェア　　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（表 1，表 2 参照）。

TM のソフトウェアは，日々急速な進歩を遂げており，適用効果も数多く報告されているが，現状は「隠された事実の発見」を期待するというよりは，分析者の判断や思考を支援することが主たる利用目的となっている。そのため，理論の明快さや技術の優秀性，あるいは，コンセプトの先進性よりは，日常業務での活用性と実務性に重きが置かれている。

一方，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，Nahm[12]には TM に関する総合的な紹介サイトやテキスト・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 [9](2000)には，多数のソフトの紹介（かなり詳しい説明，評価）がある。Kimberly A. Neuendorf and Paul D. Skalski[10](2002) では一つの章を割いて，「Paul D. Skalski , Computer Content Analysis Software」とし，ソフトの紹介，評価説明を行っている。また，Robert P. Weber [13](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･2〉　TM ソフトウェアの課題　　元々，得られるデータや情報が曖昧かつ多様な表現を持ち，とくに E-mail や携帯電話などの世界では独特な記号や造語を容認している。例えば，記号を利用して表情や感情を表現するスマイリー（Smiley，例えば，大笑い縦型「(^0^)」横型「8D」，泣いている縦型「(;_;)」横型「;-(」），発音のもじり（Hakspec，例えば「F2F」face to face，「IC」I see），頭文字をつなげたアクロニム（Acronym，例えば「FYI」For Your Information），さらには，組合せ文字（例えば「ナ」と「＝」で「ナ＝」で「た」を表現）や絵文字等，きりがない。

また，表記上の問題は比較的発生しないが，学術文献情報は，元来，新しい事実や知見を報告するものであり，記載したとき（書き手）と読むとき（読み手）の概念や意識の差は大きく，これを自動的に認識し理解するのは難しい。すなわち，これまでと類似の概念であっても，書き手は無意識のうちに（あるいは「あえて」）これまでとは異なる新

しい概念（知見）として表現する傾向がある。

結局のところ，分析者には対象データや業務に十分精通していると同時に，マイニング・ソフトウェアの特徴を把握し，目的に応じて分析結果を評価する能力が必要とされる。

6. テキスト型データの多様性と分析の難易度

筆者の少ない体験ではあるが，これまでに経験したテキスト型データの分析事例（TM の活用事例）について，分析の難易度を主観的に要約した（表 3 参照）。経験をとおして言えることは，「万能の」あるいは「汎用的な」TM ツールや技法はなく，解析の目的や対象とするデータにより，様々な創意と工夫が必要となることである。つまり，現在の TM ツールや技法が多目的にそのまま利用できるものではないということであり，とくに TM ツールには，分析目的に応じたデータの編集・加工・抽出，あるいは用語や概念の抽出など，カスタマイズが重要であることを意味している。

テキスト型データの分析（TM）の難しさは，そもそも自然語の分析の困難性にあると考えられる。言語情報としての表現や描写が困難な抽象概念を取り扱わなければならず，これの計量化が容易ではないことが TM の難しさのひとつの理由である。比較的容易とされるアンケート調査の自由回答・記述回答においても，調査者の意図を超えて回答は多様となり，回答者により表現は様々であり，回答内容や概念の組み合わせは爆発的となる。例えば，「良い点か，悪い点か」といった評判分類，「褒めているのか，不満なのか」といった発言の性格的分類，「私が思うのか，誰かが言っていたのか」といった主観性分類など，分類の観点も様々である。

従って，TM を有効に活用するためには，まずは，解析の目的にあったデータ取得法を考えるべきということである。また，既に収集し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上で対処すべきである。

（1）サンプルや調査対象の背景やデータ取得状況，データの素性，取得目的があまり明らかでない「単に集めた

表 3　テキスト・マイニングの適用場面の例

Table 3.　The application fields of textminig (for examples).

利用/適用の場面	分析の難易度	分析の課題/留意点
一般的なアンケート調査における自由回答データ（消費者行動調査分析，自由回答方式の研究など）	比較的容易	・調査票や設問の設計 ・選択型設問による定量分析との併用
インターネット調査など電子調査で取得したデータ（Web 調査，電子メール調査など）	比較的容易	・調査設計（とくにデータの取得法や取得環境，サンプリング，回答比率，再現性や客観性など）
Web ページ上での(特定)製品ユーザの意見聴取	比較的容易	・計画的に設計された取得環境からのデータ収集
電子メールによるモニターからの回答収集分析	やや面倒	・設問設計や回答方法（複合的な意見の選別）
製品に添付の意見葉書の自由記述分析	比較的容易	・選択型設問による定量分析との併用
コールセンター，コンタクトセンターでの収集情報（お客さまのご意見・ご要望，オペレータの応対，両者の対話履歴，FAQ 作成など）	やや面倒	・教師あり(既存分類体系)分類と教師なし(自由仕分け)分類の併用 ・お客さまとオペレータとの対話分析
グループインタビュー，フォーカスグループにより取得したデータ	難しい	・発話の反応，心境の変化など，対話分析
インターネット上でのチャット，対話データ	難しい	・同上
面接法による録音データ（逐語録），ヒアリングや聞き取り調査により取得したデータ	難しい	・記録の書き起こし方法と分析用データの作成 ・(コンピュータ支援の)分析目的に依存
グループ作業に対する意識調査，品質評価分析（医療チームと患者の意識比較，メンバ別評価）	やや面倒	・設問設計，データ取得法
発想法，KJ 法等の文字データ解析	ほぼ適用可	・意見の類型化(クラスター化)処理を適用
行政/自治体「市民の声」の活用分析[14]（苦情・要望，政策評価・提言など）	やや面倒	・取得環境構築（電話，投書，電子メール，Web，面談など，様々な収集ルートからのデータ取得法）
論述形式，記述問題の解答や評価(感想)の分析[15]（先生/生徒評価，事前/事後評価，選択式との比較など）	ほぼ適用可	・分析の目的に依存（解答や評価の特徴抽出や比較・傾向分析は比較的容易だが，正誤判定・得点付けは難しい）
新聞記事，特許公報，Web コンテンツ（動向・傾向分析，競合他社・技術分析など）	ほぼ適用可	・分析の目的に依存（時系列変化，企業や機関・人物・技術等の属性による特徴分析などは比較的容易）
日記形式の記述文の解析（日記による心理分析，行動分析など）	かなり難しい	・分析者のスキルに依存
書籍，小説・文芸作品などの文章解析	やや面倒	・分析の目的に依存（著者やテーマなどの属性による特徴分析は比較的容易，要約・主題分析などは難しい）
議事録，会議録(速記録)，事故記録など記録データ（議会議事録，アクシデント/インシデント分析など）	やや面倒	・分析の目的に依存（傾向分析や特徴分析は比較的容易，因果関係分析や対策の絞込みなどは難しい）
TV ドラマ，映画，スポーツ番組の発話分析（あらすじ分析，シナリオ分析，臨場感分析など）	かなり難しい	・分析者のスキル，分析の目的，さらには，分析データの作成法にも依存
情報検索（類似文書検索，自動仕分け，要約・タイトル付けなど）	適用可	・従来の自然言語処理技術の活用
Web マイニング（Web コンテンツ評価，Web 空間の整理・体系化など）	適用可	・膨大な Web ページからの特徴抽出，および Web ページ間の引用関係の集約
コーパスの生成，シソーラス辞書の構築（語彙の蓄積と意味づけ，典型文例の作成など）	かなり難しい	・意味解析や文脈解析などの高度な言語処理が必要

だけ」のテキスト型データは，多くの場合，分析が煩雑となるばかりか，有用な知見を得ることができない。

（2）新聞記事，特許公報，雑誌，文芸書，あるいは各種の記録文書など「元来が文字情報」のテキスト型データの分析には，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，著者や表記法の比較，ドキュメント・マイニング（マッピング）などである。

（3）蓄積されていたアーカイブなどに付帯情報，データ取得履歴などを付加し，データを整理し，蓄積した定性情報データベース，あるいはメタ・アナリシス（複数のデータベース情報の統合・併合利用）など，過去の蓄積データの見直し・再評価を行う。

（4）調査データ，とくに選択肢型設問と併せて用いる自由回答設問がある。社会調査や市場調査などで，最も多いタイプである。

（5）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したデータ，例えば，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で，解析目的が明確で，その成果を検証できる。

このように，TMは，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異なる。しかし一方では，現状の主たるTM技術・技法は，膨大な生のテキスト型データを対象とするという本来の目的を，ある形で圧縮し，計量化・数量化した上で，（従来型の）データ解析の方法論を適用することが多いということも事実である。

7. おわりに

今後，テキスト型データの活用の声が高まるとともにテキスト・マイニングへの期待もますます高まるものと思われるが，現状は，いかにも安易な発想でTMが「役に立つ」と考える風潮があり，課題評価の上，期待倒れになる懸念をもつ。TMツールや技術を利用する立場では，ともすれば「方法論はともかく（ブラックボックスのまま）簡単な操作がよい」，「主観的であれ，視覚的に分かりやすければよい」，というような安易な考えが蔓延しつつあり，これがTMの健全な発展を阻害するのではないかという危惧がある。

ゴミのデータをいくら分析しても，重要な分析結果は得られない。膨大なデータが容易に収集できるときにこそ，集まったデータの解析法だけではなく，データの収集・取得法もあわせて設計する必要がある。そもそも自由回答・自由記述とは，単に自由に書いて（発言して）貰うだけでは不十分であり，周到に実験計画された環境下で「いかにデータを取得するか」といったデータ取得方法の研究とも密接に結びついている。適切なデータ取得法があって，はじめて解析が意味を持つ。即ち，現象解析の基本はデータにあるというデータ科学（data science）のアプローチを尊重しつつ，テキスト・マイニングを使いこなすための知恵と工夫がこれからますます重要となる。

謝　辞

日頃より御指導をいただくともに，本解説記事の報告にあたり，貴重なご意見ならびに情報や資料のご提供をいただいた大学共同利用機関法人，情報システム研究機構，統計数理研究所（旧，文部科学省統計数理研究所）の大隅昇名誉教授に深謝いたします。

（平成17年2月25日受付，平成17年2月25日再受付）

文　献

（1） Ye, N. (ed.)：“The Handbook of Data Mining”, Lawrence Erlbaum Associates, Publishers (2003)
（2） D. Sullivan：“Document Warehousing and Text Mining”, John Wiley (2001)
（3） M. A. Hearst :“Untangling Text Data Mining, in the Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics” ,University of Maryland, June pp.20-26 (1999)
（4） U. Fayyad and R. Uthurusamy：“Preface for KDD-95: Proceedings First International Conference on Knowledge Discovery & Data Mining”, AAAI Press (1996)
（5） U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth：“Knowledge Discovery and Data Mining: Towards a Unifying Framework, in Proceedings Second International Conference on Knowledge Discovery & Data Mining”, AAAI Press, pp.82-88 (1996)
（6） U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth：“The KDD Process for Extracting Useful Knowledge from Volumes of Data”, Communications of the ACM, 39, 11, pp.27-34 (1996)
（7） 林知己夫：データの科学，シリーズ<データの科学>1，朝倉書店. (2001)
（8） 中村純作：現代言語学の潮流，コーパス言語学（山梨正明，有馬道子編），勁草書房，pp.233-245 (2003)
（9） R. Popping：“Computer-assisted Text Analysis”, Sage Publications (2000)
（10） K. A. Neuendorf and P. D. Skalski：“The Content Analyisis Guidebook”, Sage Publications (2002)
（11） 大隅　昇・L. Lebart.：「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，Vol.48, No.2, pp.339-376 (2000)
（12） 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/]
（13） R. P. Weber：“Basic Content Analysis (second edition)”, Series: Quantitative Applications in the Social Sciences 49, Sage University Paper (1990)
（14） 仙台都市総合研究機構：「市民の声」の活用法に関する調査研究，2003 SURF 研究報告 (2003)
（15） 須永恭子・上野栄一・保田明夫：「内容分析を用いた臨地実習における学習達成の自己評価と指導者評価の分析」，Quality Nursing, Vol.10, No.3, pp.57-65 (2004)
（16） 大隅　昇・保田明夫.：「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，Vol.19, No.2, pp.135-159 (2004)

保　田　明　夫

（非会員）　1978年3月東京理科大学卒業。現在，(株)平和情報センター勤務。テキスト・マイニングをはじめ，知的情報処理システムの研究・開発に従事。