★――論文

テキスト・マイニングが目指すもの
～最近の動向，そしていま何を必要とするか～

大隅　昇
●文部科学省統計数理研究所

横原　東
●株式会社電通リサーチ

❶――まえがき

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がキーワードとされる。こうした中で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」とのキャッチコピーのもとにテキスト・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）の活用が拡がりつつある。しかし実は，TMとは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りTM同様に分かったようで漠としたものであるが，関連書籍は無数にあり，併せて沢山のコンピュータ・ソフトが現れ百花斉放の感がある。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の差異も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についてそ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”要約する。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が目標ではない。また紙幅の都合もあるので全体を大まかに概観する。

1．データ・マイニングとテキスト・マイニング

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。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，あるいは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。これについては前述のように無数の研究報告や書冊があるが、ここで個々の技法や方法論に言及することは難しい。単に一般的なDMの概念を眺め，これに続いてTMとは何かをみる。

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。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。このとき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）］。

従来からの統計手法，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」等にあると言う。しかし最近の統計的方法論では，これらに対する解決策は提供されており，この主張だけでDMを特徴付けることは説得力がない。DMという耳に心地よい言葉，流行のように見える。

膨大なデータセットを目前にしたとき，その中から“金の鉱脈”を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の多くの書には「ゴミを入れればゴミが出る」（GIGO：garbage in garbage out）とあるが，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」との素朴な疑問に至るが（ニワトリと卵の論法），DMの多くの方法論にはこれへの答えはない。“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。しかしこれで真の現象解析が可能かという疑問に突き当たる。

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，サンプリング操作により分析対象（標本）を用意する。この厳密さがあるがゆえ，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。つまり「ゴミは所詮はゴミ」であり，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，それを前提に“データ主導型”の解析過程を必要とする。この点，統計学は明確な枠組みを示している。「データ科学」（data science）はこれを発展的に考える［林知己夫（2001）］。ここでは，現象解析の基本は「データ」にあり，「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。

❷――テキスト・マイニングの背景

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，いくつかの定義を挙げると以下のようなことであろう［Nahm，Ye（2003），Sullivan（2001）］。

【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，それらの間に潜在する関連性を分析，類型化する。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。

■ 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操作して（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。

【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。

■ 例えば，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，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。

ここには表現はやや異なるが以下に見るような幾つかの“共通項”がある。

・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）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。しかしその本質は，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，事の本質が探索的アプローチにあると主張している。

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。

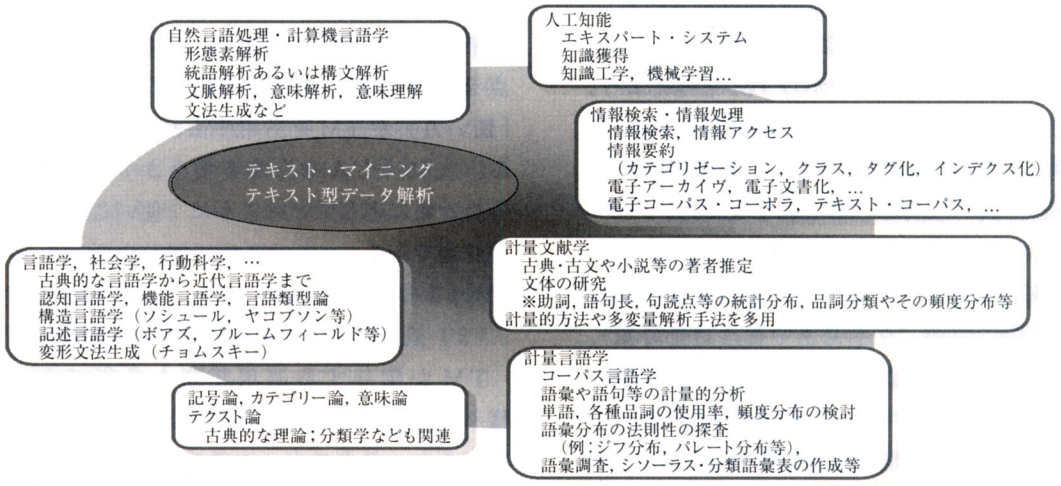
（1）関連研究分野からの観察

これは**図－1**のように要約される。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，実に多彩である。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

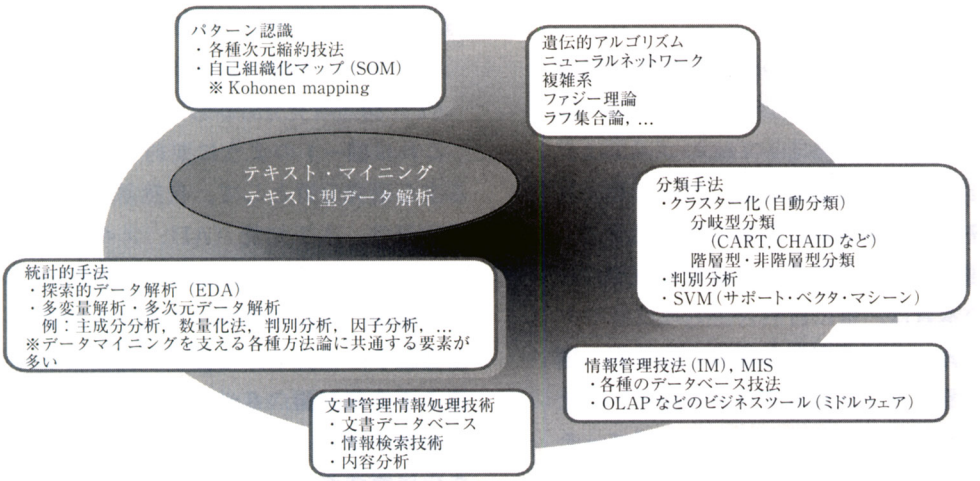
■図―1

テキスト・マイニングに関連する研究分野



■図—— 2

テキスト・マイニングに利用される主な方法論



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見る。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［Popping（2000），Neuendorf 他（2002）］。この CACA の成果は 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。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。しかし解析部の主体となる方法・手法としてこれを見ると，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，実に多様である（**図—2**）。

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 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，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 を活用するためのキーである。先に「どんな方法を使うか」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。様々な報告書，研究論文にあ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，ルール発見，概念抽出，関係の発見，情報の有機的統合化，特定なトピックスの検出，テキストの分割，テキスト・文書の要約化と収集分析，知識取得と理解，テキスト・ナビゲーション，視覚化，ユーザー・インターフェース，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，全文検索，文書検索，情報抽出とテキスト型データ，文書情報に関する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がある［Sullivan（2001），Ye（2003）］。

しかし日本国内では，とくにマーケティングや市場調査分野では，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等で収集の非構造的なデータ（unstructured data）など，限定された範囲の利用が多い。本来の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，研究として散見されてもビジネスでの利用は少ない。TMの応用の範囲や浸透の方向・拡がりにだいぶ差異があり，ある一面だけが強調され研究の深化が極めて浅い。

❸――テキスト・マイニングはどう活用すべきか

1. マーケティングにおける適用可能性

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”にあるが，的を絞ってCRMに関連した顧客対応の場面でいかに活用できるかを考える。具体的には，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いかに有効利用できるかである。対象をこの分野に限定し，TMやその関連ソフトの活用法を考えたとき，どのような視点で取り組めばよいだろうか。「使い方のコツ，利用上の留意事項」「調査における利用法，活用法」「とくに，調査における自由回答設問の考え方」に対するガイドとして以下を挙げておこう。

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，本当に必要かをコスト面からも考慮すべきこと
⑤ 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

2．テキスト・マイニングが行うこと，何ができるのか

“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，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，はずれ値的なもの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。言語類型論により形態的特徴で区分すると日本語は膠着語とされる。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。しかしコンピュータにはこの「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ツールがそのまま日本語処理に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。形態素とは「意味をもつ最小の言語単位」をいい，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。また分かち書き処理で得た要素単位がそのまま形態素とはかぎらない。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。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答が得られるとは限らないことに留意せねばならない。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。形態素解析だけでなく，自然言語処理系では言語的知識（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。

参考：形態素解析ツールに，茶筅（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と同様に，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，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，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に入る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），STATISTICA-Text Minerなどをみれば明らかである。[**表－1**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。そもそも定性的情報であるテキスト型データに潜在的にある漠然とした特徴，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で視覚化することは有効である。一方，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）。ここをどう解決するかが今後の課題である。

コホーネン（Kohonen）の提案したSOMマップ（自己組織化マップ）も良く利用される。SOM（Self-Organizing Maps）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），Murtaugh（2000），Sullivan（2001）］。これら視覚化過程での検討課題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
- ここで，要約や縮約化に伴う情報損失をどう評価するか，あるいは客観的に知るか，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，視覚化の意義や意味解釈の方法の説明が総じて明らかではない。TMの目標のひとつであるのに設計指針が曖昧であり提供情報の意味解釈を与える客観情報に乏しい。

（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。形態素解析や分かち書き処理を行うために，大抵の場合は辞書を備えている。しかしここで分析対象が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。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（とくに欧米），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も同じことを意味する語句をどう扱うかはかなり難題である。例としてWeb調査で取得の自由回答データを考えればよい。設問を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」「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。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」「ファミリー」「身内」から「親

類」「親族」「血族」「縁者」「父母兄弟」…とあって，どこまでを類似語句として括るか判断に迷う。さらにケータイ語，電子メール語，チャット語とあっては同義語・類語の処理を厳密に考えること自体に無理がある。

TMソフトの側からは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，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，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・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。例えば、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自由回答データを取得する，同一テーマで異なる調査対象に意見を聞く等の場面では，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。この場合に，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。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，類語大辞典など）。

■表――1

主要なテキスト・マイニング・ソフトウェアの一覧

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/	メーカ系(規模大) 全方位，多機能型 他システムとの接合 データベース機能
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	類似検索 ターム相関 言語処理など
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/	専用ツール
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) 自由記述アンケート分析システム [Topic Scopeとして改編されたもよう]	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	調査データ分析 自由回答設問他 統計的手法
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	

※ 会社名、製品名等は、各社の登録商標もしくは商標

3．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（スケーラビィリティ），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）等があるだろうが，個別の詳細機能については紙幅の都合もあるので省略する。

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**に，保田（2003）によるサーベイを元に一覧とし，備える機能，分析対象の守備範囲を図中に書き入れた。

❹———むすび：真のテキスト・マイニングの目指す方向とは？

日本語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の表現，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。自由回答設問に限っても，調査者の意図を越えて回答内容は多様であり，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。多変量的かつ高次元性があり膨大な特徴の組み合わせの可能性の中で，知識発見や組織化をどう行うかがある。個別的に優れた技術要素があってもそれらを有機的に融合し使いこなすにはかなりの習熟度を要する。

現状は（とくに国内），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。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。その対極に何事も精密かつ厳密にモデル化されるべきとの考え方がある。しかしいずれも極端，どちらも説得力があるとはいえない。利用者・分析者の要求に応じて“的確に”，“信頼できる”客観的情報を提供できることが望ましいのだが，現状のTMはいかにも中途半端である。理由の第一は，利用者側の方法論への正しい理解が十分ではないことがある。次に，解析ツール提供者側に，TMが満たすべき要件を十分に吟味した設計指針があってソフト開発に取り組む姿勢が不足していること，そして「ノウハウ」という都合のよい隠れ蓑に保護され，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がある。

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，その解釈は科学性があり客観的か）にある。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。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。一つの試みとして，**表－2**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

■表――2

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（名義尺度，順序尺度） 量的データ（区間尺度，比例尺度）	・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　音声認識 意味理解　画像認識

る。ここで明らかにしたいことは，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。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”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が，いまこれへの的確な解を即答できる段階にはない（表－2のセル「真のTMとは？」に相当）。ここで指摘できることは，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だけではなく，“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」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。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は出てこない」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，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で知るという美味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。今まで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。過剰な期待も困るが，一過性の流行りものに終わらせてはならない。

参考文献

Ah-Hwee Tan (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]

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

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

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

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.

Kimberly A. Neuendorf and Paul D. Skalski (2002), *The Content Analysis Guidebook*, Sage Publications.

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

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

Marti A. Hearst (1999), Untangling Text Data Mining,; 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).

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

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

Ronen Feldman and Ido Dagan (1995) : Knowledge discovery in textual databases (KDT), in the *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (KDD-95), Montreal, Canada, August, AAAI Press, 112-117.

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

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/]

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.

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.

Usama Fayyad and Ramasamy Uthurusamy (1996), Preface for "KDD-95: Proceedings First International Conference on Knowledge Discovery & Data Mining," AAAI Press.

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.

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.

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（特集：使いやすくなった自然言語処理のフリーソフト－知っておきたいツールの中味），41，11，1203-1207.

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.

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」.

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.

柴田武，山田進編（2002），類語大辞典，講談社.

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.

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．
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．
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/］

大隅　昇（おおすみ のぼる）

現職は文部科学省統計数理研究所調査実験解析研究系パターン解析研究部門・教授（1981年から）：専門分野はデータ科学，多次元データ解析，社会調査法など：理学博士（1979年）：著書として「統計的データ解析とソフトウェア」（日本放送出版協会），「記述的多変量解析法」（日科技連出版社）他．

横原　東（よこはら ひがし）

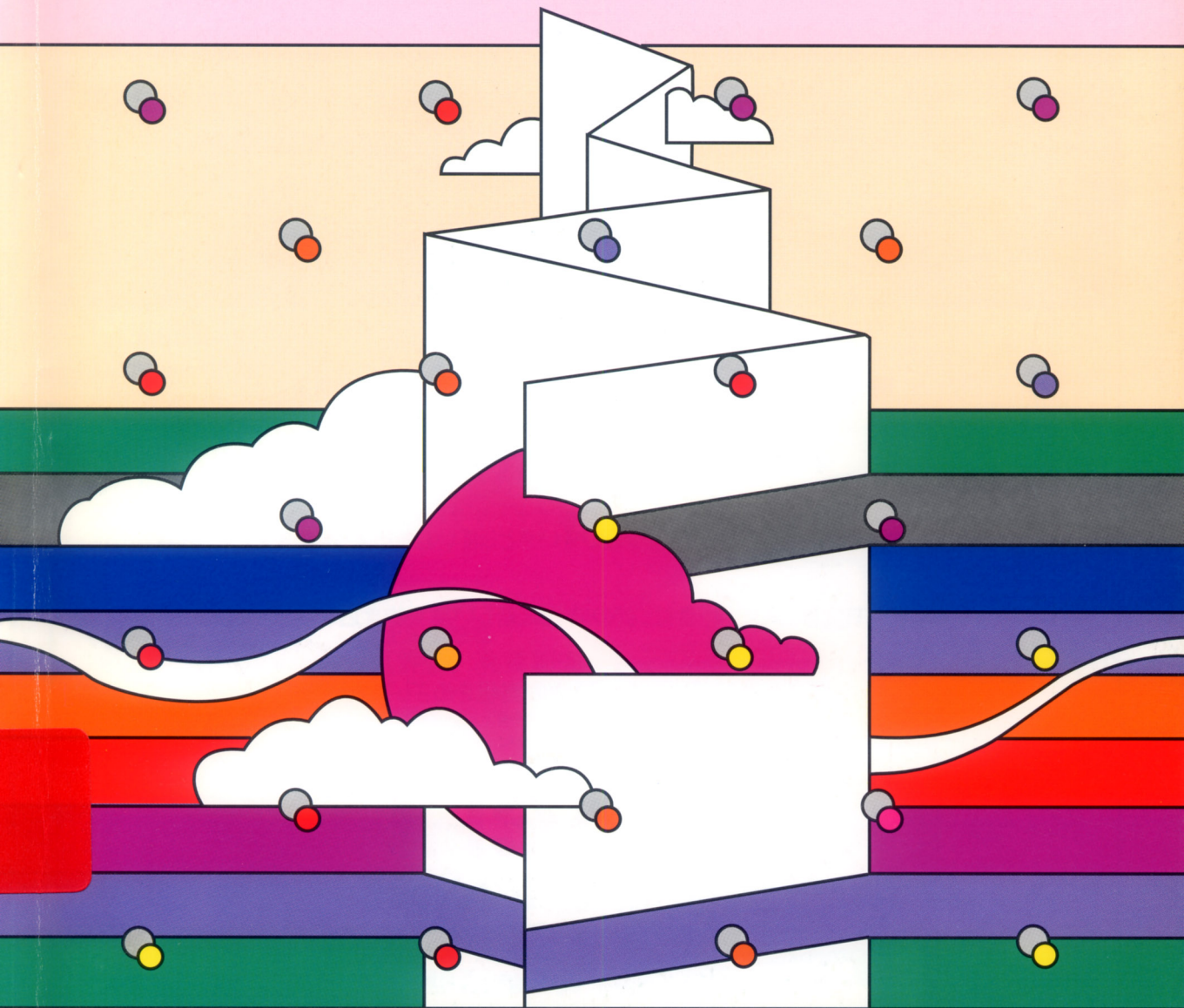
1972年　立教大学社会学部卒業。同年株式会社電通リサーチ入社。現職、研究開発部主席部長。

季刊
マーケティング ジャーナル
JAPAN MARKETING JOURNAL
ISSN 0389-7265
2004 91

●——本号執筆者

丸岡吉人〈巻頭言〉　大隅　昇　横原　東　池田謙一　小林哲郎　繁桝江里
中野幹久　福冨　言　上野　博　棚橋　豪〈訳〉　須永　努　恩蔵直人
村下　訓　高橋郁夫

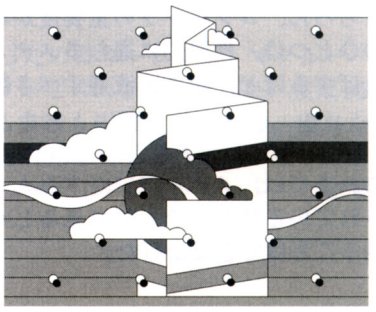
マーケティング・エクセレンスを求めて・第53回／パイオニア株式会社
テーマ書評シリーズ・第46回／広告研究の新しい視点



社団法人 日本マーケティング協会

季刊 マーケティングジャーナル ◉目次

91 2004
VOL.23—No.3



Design—山口英之