2 テキストマイニング

―膨大な文書データの自動分析による知識発見―

那須川哲哉　日本アイ・ビー・エム（株）東京基礎研究所
諸橋　正幸　多摩大学経営情報学部
長野　徹　日本アイ・ビー・エム（株）東京基礎研究所

膨大な文書，集めてはみたものの...

Information Overload （情報過多）という言葉を実感する機会が増えてきている．WWW上に限らずさまざまな場所で日々多様な情報が蓄積され，アクセス可能な情報の量が急速に増加している．中でも文章形式で蓄積されたデータの増加が目立つ．

なんらかの情報を記録しようとする場合，現時点では，記述形式の柔軟性や計算機に蓄積する際のデータ量という観点から，文章にするのが最も一般的な手法であろう．一説によれば，世の中に蓄積されているデータの8割以上が文章形式であるとも言われている．ところが文章にした場合，後からその情報を引き出すには，基本的には人が読むしかないのが現状である．

たとえば大量の営業報告書の分析について考えてみよう．この場合，目的に応じて必要なデータだけを形式的に記述する（たとえば表の中の項目を埋めていく）形で報告書が作成されていれば，それを集計するのは比較的容易である．ところが営業活動のノウハウ（たとえば顧客との応対内容や営業上の工夫）をうまく捉えるような項目をあらかじめ用意しておくのは困難であるため，必然的に文章で記述する場合が多くなる．その上，作成した報告書からどのような情報・知識を活用するかは利用者次第であり，利用時点にならないと分からないのが普通である．したがって報告書の作成時に，その後の利用を意識して十分な記述を行うのは不可能である．その結果，すでに作成された大量の文書から目的とする情報・知識を読み取るには大変な労力が必要となる．また対象文書の量が膨大な場合は，すべてを人手で読むことは不可能である．

別の例としてアンケートで顧客の声を集めることを考えよう．その場合，顧客の意見にどのような内容がありそうかを最初から予想し，選択肢を用意したり表を埋めるような形式にしておけば，集計が楽になる．反面，最初から仮定できる内容であれば，新たな知見は得難いというジレンマが生じる．そのためアンケートには，選択肢形式の回答欄だけでなく自由回答形式の記述欄が含まれている場合が多い．そして，回答結果の中で貴重なのが，この自由回答部分と考えられる．ところが実際には集計の困難さから，アンケートの規模が大きくなるほど自由回答部分の活用度が低くなる傾向が見られる．

このように文章は多様な内容の記述を可能にする反面，その内容を機械的に集計することができないという処理の困難さから，せっかく文章にしたデータが十分に活用されていないというのが現状である．

本稿ではこのような文章形式のデータを分析し，その中から有効な知識を獲得する技術を紹介する．

実際にこの技術を用いて，お客様相談センターに電話で寄せられる顧客からの問合せ内容を記録した大量文書を分析した結果，問合せ傾向をつかむことで，顧客の求める情報を積極的に公開することにより問合せを減らしたり，製品の障害を早い段階で発見したりする効果をあげられることが確認されている．本技術の適用により，従来は手作業で大量文書のごく一部をサンプリングして行っていた分析の効率が飛躍的に向上し，大量の文書全体にわたってのより深い分析が効果的に行えるようになる．

膨大な文書データの分析技術

テキストマイニングという言葉は，比較的曖昧な形で用いられているのが現状である．大量の文書データを対象とする点では共通している場合が多いが，マイニングの持つ「掘り出す」という意味が，自分の欲しい内容を含む文書を引き出してくるという検索の意味で使われることも多い．本稿では，「大量の文書データに記述されている多種多様な内容を対象として，その相関関係や出現傾向などから新たな知識を発見する.」という，よりデータマイニングに近い意味でテキストマイニングという言葉を用いる．

データマイニングの技術を用いてテキストの情報をマイニングできないかという発想は，決して目新しいものではなく，データマイニングが注目を集め始めた当初から存在する．研究用に無償で公開されているJUMAN[7]や茶筅[8]のような形態素解析ツールや市販のキーワード抽出ツールを用いれば，名詞句を中心としたキーワードを文書から抽出することが比較的容易に実現できる．そして，抽出されたキーワード集合における語と語の相関関係を分析したり，語のクラスタリングを行ったり，文書に含まれるキーワード集合を利用して文書をクラスタリングするなどの処理を，既成のデータマイニングツールの利用により容易に試すことができる．

ところが，単に名詞句を中心としたキーワードの集合に対し一般的なデータマイニングの技術を適用しても，新たな知識を発見するという点ではあまり有効な結果は得られない．

これは，多種多様な概念を表現する言葉の種類が膨大な数に及び，そのすべてを混合して処理しても，混沌とした結果しか得られないためである．たとえば，語と語の相関関係を分析しようとすれば膨大な数の相関ルールが抽出され，その中から価値のある相関ルールを抜き出すのは至難の技である．その上，処理の単位が名詞を中心としたキーワードのレベルであるため，相関ルールとして取り出される情報も，『「データ」と「分析」は共起しやすい』といったものであり，実用的に価値のある情報はなかなか取り出せない．

また，クラスタリングするには膨大な次元数のベクトルを扱うことになり，そのような多次元空間においてキーワード間や文書間の距離を適切に評価し有効な類似度を求めるための工夫が必要となる．

さらにその処理の前提となるキーワードであるが，自然言語の特性ともいえる多義性（同じ言葉が複数の意味をとり得る性質．たとえば「山口」は人名にも地名にもなり得る）や同義性（同じ概念が複数の言葉で表現される性質．たとえばパーソナル・コンピュータという概念は「パーソナル・コンピュータ」のみでなく「パソコン」や「PC」などでも表現される）を備えている．そのため，これらの性質を無視して処理すると，語の意味を取り違えた結果を出力する危険性があり，すべてのキーワードを同等に扱ってマイニングを行う処理には無理がある．

すなわち，文書中から単なる文字列としてのキーワードを抽出し，その結果に対して一般的なデータマイニングの手法を適用しても有効な結果はなかなか得られない．

したがって，文書データを本格的にマイニングするには，さまざまな工夫が必要となることが分かってきている．まず自然言語処理の部分では，単なるキーワード抽出よりも深い解析を行い，曖昧性を処理して意味のあるデータを抽出した上で，自然言語の性質を活かしたマイニングを行う必要がある．また，膨大なデータから抽出した情報の全体像を把握しやすいように視覚化し，得られた結果を適切に解釈できるよう，インタラクティブに多種多様な観点から分析できるようなインタフェースも用意する必要がある．以下では，この流れに沿って概念抽出，マイニング，視覚化および分析において必要となる処理の内容を解説する．

概念抽出

1つの文書にはさまざまな話題が含まれているため，テキストマイニングにおいては，文書そのものではなく，文書中から抽出した概念を単位とした分析が必要である．また，文書中に示されている内容を表現する上では，単なる文字列としてのキーワードは曖昧性が高すぎる．そこで，多義性や同義性を考慮した上で，同義表現を1つの表現にまとめたり，多義表現は別々の内容として扱えるようにする必要がある．たとえば，「山口」に対し「山口〔人名〕」「山口〔地名〕」というようにカテゴリーもしくは意味素性を付加することで多義性を解消したり，「パソコン」や「PC」などを「パーソナル・コンピュータ」のような一定の表現に置き換えることで同義性を吸収すればよい．以下，本稿では，単なる文字列としてのキーワードでなく，同義性を吸収した上で，「山口〔地名〕」のようにカテゴリーもしくは意味素性を与えたものを概念と

呼ぶことにする．

さらに，文書中の概念を表現しようとすると，名詞句のみでは不十分であり，述語やその付属語の情報なども重要である．

例として以下の4つの文の処理を考えてみよう．

(a) Win98を入れたら起動が早い．
(b) Win98を入れても起動が早くない．
(c) Win98を入れたら起動が早いか．
(d) Win98を消したら起動が早い．

通常のキーワード抽出では名詞句のみを抽出するため，いずれの文からも，キーワードとして「Win98」と「起動」のみを抽出する．したがって，このようなキーワードレベルの情報を使う処理であれば，いずれの文もこの2つの語によってラベル付けされることになる．つまり，このレベルの情報を使った処理では，いずれの文も『「Win98」と「起動」に関する文』という意味で解釈され，まったく同じ扱いを受けることになる．

しかし，一歩踏み込んで，

- 各文の意図するところは何か（苦情か要望か質問か）
- どのような行為が関係しているのか

といった分析をしようとすると，「Win98」と「起動」というキーワードのみでなく，述語やその付属語などの情報を利用することで，要望か質問か苦情か，また否定か肯定かなどの判別をしたり，「○が△した」「▽を□する」といった複合概念を抽出することも必要になってくる．

このように意図を解釈したり，複合概念を抽出する技術の具体例としては，文献1)，6）を参照されたい．

マイニング

文書中から適切な概念が抽出できれば，次の処理内容は，この概念の集合からどのような情報を得たいかによって決まってくる．

たとえば，報告書のようなデータを処理して，最近の動向を探りたい場合には，時間的な経過に応じて出現する概念の増減傾向を調べればよい．また，誰が（あるいは何が）どう評価されているか知りたければ，人名（あるいは事物）の概念と評価の概念との相関を調べればよい．

このように，文書中の概念をマイニングする上では，すべての概念を混合して扱うのではなく，カテゴリーもしくは意味的性質を区別して扱う工夫が重要である．特にテキストマイニングにおけるマイニングでは，一般的なデータマイニングと異なり文書データの特徴を活かした処理が必要となる．文書データの特徴の1つは，その中に含まれる個々の概念が多様な性質を持っていることである．たとえば，「パソコン」は，計算の道具であると同時に最近ではコミュニケーションの道具でもあり，さらに物質としての性質も備えていて，空間的な場所を占め重量が備わっている．また，データマイニングでは，通常個々のデータの値自体は大きな意味を持たないことが多いのに対し，文書中の概念には，目的に応じた重要度が存在し，出現頻度は低くても重要視しなければならない概念も存在する．このような重要視すべき概念をいかにして認識すべきかも大きな課題である．

このように文書から抽出した概念を分析する技術の具体例としては文献2)，3）を参照されたい．

視覚化および分析

通常，1つの文書の中には多種多様な概念が含まれており，それが蓄積された膨大な文書集合を分析するためには，そこに含まれる膨大な種類の概念を何らかの観点で整理しながら分析する仕組みが必要である．ここでは，全体像を視覚化することに加え，視覚化された内容を直接操作できる対話的分析の機能が分析能力向上の大きな鍵となる．すなわち，文書データの情報を多様な観点から視覚化し，インタラクティブに観点を変えながら分析する仕組みを取り入れることで，全体像が把握しやすくなる．

さらに全体像を捉えた上で，着目する部分に対象を絞り込み，絞り込んだ集合の情報を同じように多様な観点から提示する機能を実現することで分析を掘り下げていくことが可能になる．

また，現段階で自動的に行える言語処理の精度には限度がある．したがって，自動分析結果の誤りや不明な部分をユーザが容易に元の文書で確認できるようにする機能も実用的なシステムの構築には重要である．

このように，インタラクティブに視点を変えながら分析する技術の具体例としては文献5）を参照されたい．

コールセンターにおける問合せ記録の分析事例

以上解説してきたテキストマイニングの機能を

実現した例として，日本アイ・ビー・エム（株）東京基礎研究所において構築したプロトタイプシステム TAKMI（Text Analysis and Knowledge MIning）を用いて実際に大量のデータを分析した例を紹介する．対象としたデータは，日本アイ・ビー・エム（株）のPCヘルプセンターにおいて受け付けている顧客からの問合せに関する報告書である．このPCヘルプセンターでは，IBMのPC関連製品に関するさまざまな問合せを電話で受け付けており，その応答内容をデータベースに蓄積している．この報告書データには，

- 機種名
- 問合せの種別（技術的QA，購入相談，要望など）
- 処理に要した時間

などの定型的な情報とともに，オペレータが具体的な応答内容を自由な形式の文章でワープロ入力した内容が含まれている．このデータは毎月約4万件ずつ蓄積されていっており，そのすべてに人が目を通して分析するのは不可能な量となっている．

処理の概要

TAKMIの基本的な処理の流れを図-1に示す．まず入力データの文章部分から概念（キーワードとなる文字列とそのカテゴリー）を抽出し，定型的な情報とともにさまざまな統計量を計算した上で，その結果をさまざまな観点からユーザに提示する．

◎概念抽出

カテゴリーとしては，名詞概念に対する「ハードウェア」「ソフトウェア」「専門用語」などとともに，述語概念に対して「動作」「質問」「問題」などのカテゴリーを定義した．また，述語概念と名詞概念の同文内での組合せを複合概念として抽出している．

その結果概念抽出で例にあげた4つの文

(a) Win98を入れたら起動が早い．
(b) Win98を入れても起動が早くない．
(c) Win98を入れたら起動が早いか．
(d) Win98を消したら起動が早い．

から抽出される概念は以下のようになる．

(a)「Windows98（ソフトウェア）」「入れる（動作）」「Windows98…入れる（ソフトウェア…動作）」「起動する（動作）」「早い（好評）」「起動する…早い（動作…好評）」

(b)「Windows98（ソフトウェア）」「入れる（動作）」「Windows98…入れる（ソフトウェア…動作）」「起動する（動作）」「遅い（不満）」「起動する…遅い（動作…不満）」

(c)「Windows98（ソフトウェア）」「入れる（動作）」「Windows98…入れる（ソフトウェア…動作）」「起動する（動作）」「早い？（疑問）」「起動する…早い？（動作…疑問）」

(d)「Windows98（ソフトウェア）」「消す（動作）」「Windows98…消す（ソフトウェア…動作）」「起動する（動作）」「早い（好評）」「起動する…早い（動作…好評）」

ここで表現の多様性を吸収するため，「Win98」を「Windows98」に，「早くない」を「遅い」に変換している．

このようにカテゴリーを付与したり同義語の表現を統一するための知識源としてカテゴリー辞書が必要になる．ここでは，対象としたデータが，

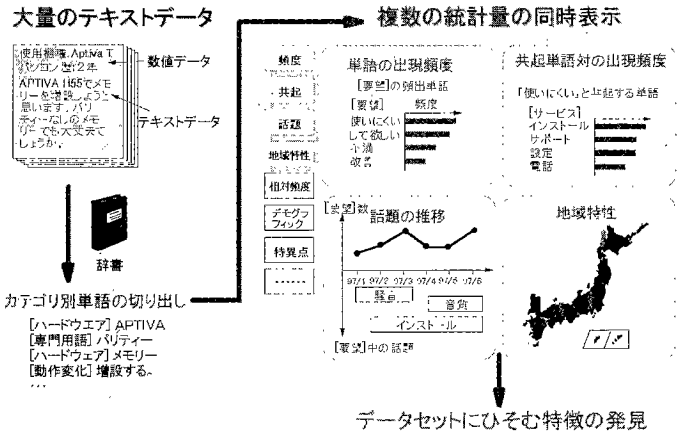

図-1 TAKMIによるテキストマイニング処理の概要

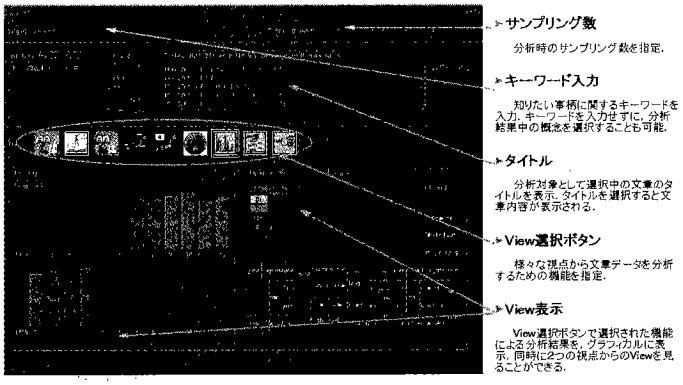

図-2 TAKMIの分析ツール

図-3　ViaVoiceに関する問合せの分析

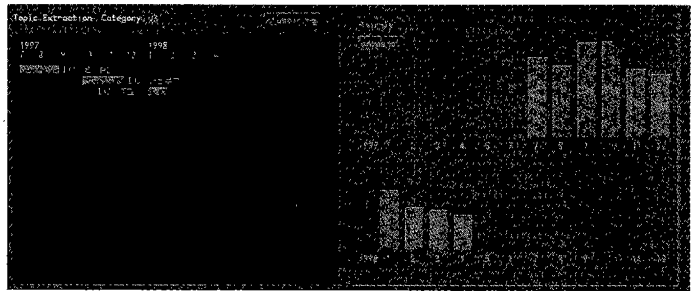

図-4　VoiceTypeに関する問合せの分析

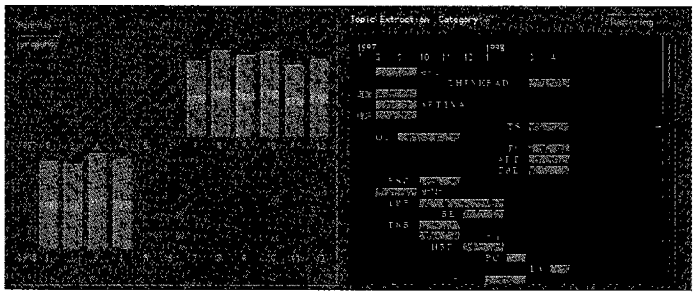

図-5　インターネットに関する問合せの分析

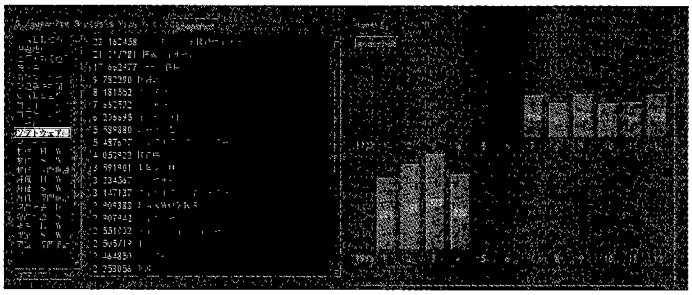

図-6　インターネットおよびThinkPadに関する問合せの分析

PC関連製品に関する問合せという比較的限定された世界を対象とした文書であるため，語彙が限定されており，カテゴリー辞書に登録した語数は1万数千語と，比較的小さい作業量で辞書が構築できた．

具体的な数字を示すと，応答内容のテキスト部分は平均して150文字程度であり，1カ月分のデータにおける自立語の異なり数は10万語程度であったが，同じ語が何度も出現しているため，出現頻度の高い1万語程度の語を登録すれば，文書データ全体に出現する語の7～8割程度をカバーできるという結果が得られている．

このように語彙の限定されたデータは，辞書的知識を構築しやすく，テキストマイニングの効果を出しやすい．

◎マイニング内容の視覚化

図-2にTAKMIの分析ツールを示す．多様な観点から対象を分析するため，多様な情報の表示機能として，

- カテゴリー別の概念一覧を表示する機能
- 時系列に沿って増減傾向を表示する機能
- 時系列に沿って話題性の高い概念を表示する機能

などが組み込まれており，ユーザはインタラクティブに各表示機能を呼び出しながら，対象としているデータの特性を調べたり，対象を絞り込むことができるようになっている．

マイニング事例

1997年7月から1998年4月までの10カ月分，約40万件のデータを処理して得られた分析結果の一部を紹介する．

図-3は，IBMの音声認識ソフトウェアViaVoice™に関する問合せのデータ（文章中に「ViaVoice」という概念を含む報告書）に対象を絞り込み分析している画面である．左下に月別の件数が表示されており，1997年（上段）後半から1998年（下段）前半にかけて問合せ件数が増加していることが分かる．また，問合せ内容に関しては，右下に表示されている概念一覧において，上位に表示されている「教える…稼働環境」「確認する…稼働環境」という内容から，稼働環境を知りたいという問合せが多いことが分かる．

図-4には，ViaVoiceの前身のソフトウェアであるVoiceType™に関する問合せデータの分析画面の一部が示されている．右側の月別件数グラフで

示されている通り，ViaVoiceと対象的にVoice-Typeに関する問合せは減少しており，その問合せの種類は，最初の頃は購入相談であったのが，次第に質問（総合案内）に変化し，1998年に入ってからは要望が目立っている．このように，いわば製品のライフサイクルが自動的に取り出されている．

図-5で示しているのは，インターネットに関する問合せのデータである．インターネット関係の問合せは，件数的には変動が少ないが，内容的には変化しており，1997年の8月から9月頃にはAptiva関連の問合せの比率が高かったのに対し，1998年の3月から4月頃にはThinkPad関連の問合せの比率が高くなっていることが分かる．

図-6では，インターネットに関する問合せのデータのうち，ThinkPadに関連するものに対象を絞り込んだ結果が示されている．実際に件数が1998年に入って増加していることが確認できると同時に，この問合せに特に目立っているソフトウェアとして「インターネット接続ウィザード」があげられている．

活用事例

ここでは，このような技術が実際の業務にどう役立てられるかを紹介する．

本技術の実用化にあたり，実際の業務で分析を行うユーザの意見を取り入れて開発した新しいTAKMIのGUIを図-7に示す．図-7に示されているのは1998年1月から1999年2月にかけてのデータであり，毎月4万件程度のデータが蓄積されている．図-8には傾向分析機能を用いた結果として6月後半から7月初めにかけて増加率の高いソフトウェアが表示されている．この結果から，この期間にはWindows98の増加率が高かったことが分かる．

そこで分析対象をWindows98に絞り込んだ結果が図-9であり，下段に示されている月別頻度分布より，6月から8月にかけてWindows98に関する問合せが急激に増加していることが分かる．また，上段右側には，対象期間を7月にした際の「ソフトウェア…質問」というカテゴリーに属する概念が頻度順に表示されており，問合せの多くが「Windows98を（自分のPCに）導入可能かどうか」という内容であったことが分かる．このように問合せの多い内容とそれに対する回答を整理して，ホームページ上で公開（図-10）することで問合せ件数を減らしたり，また，問合せに対しても即答できるようにすることで応対の時間を短縮させる

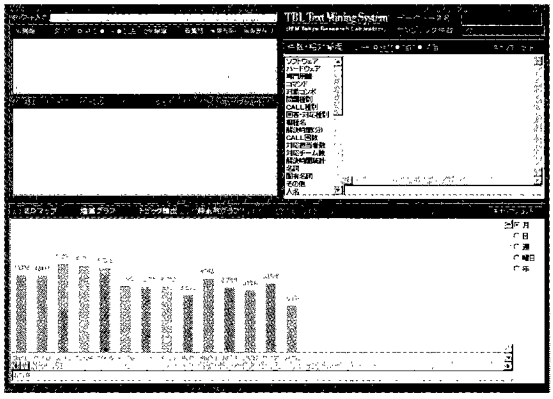

図-7　新しいGUIによる1999年2月までのデータの分析

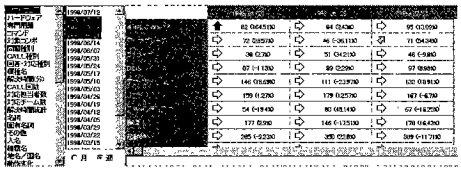

図-8　問合せが増加しているソフトウェアの分析

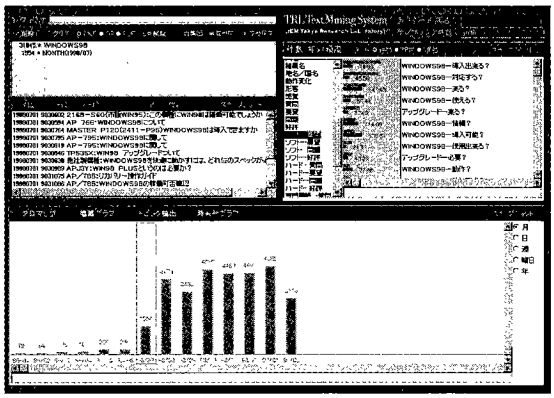

図-9　Windows98に関する問合せの分析

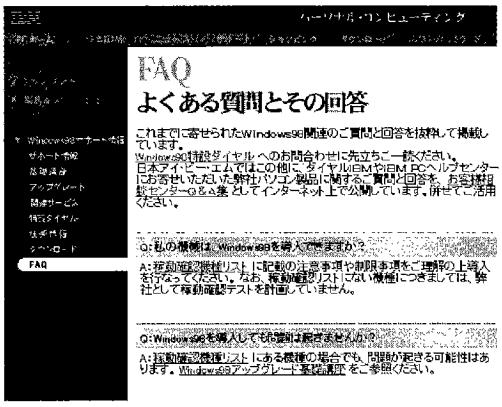

図-10　日本IBMのホームページ

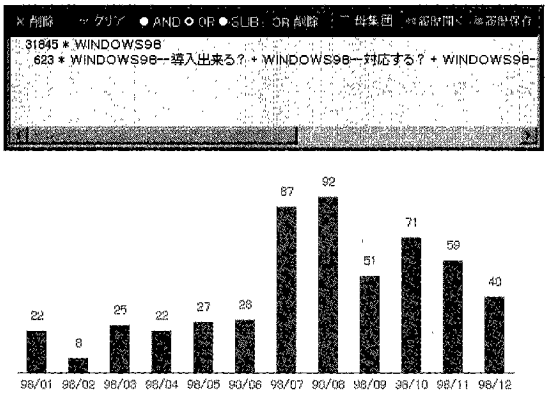

図-11　Windows98の導入に関する問合せの時系列分布

ことが可能になる．

さらには，このような対処の結果として，この内容の問合せが実際に減っているかどうかの確認も本システムを利用することで簡単に実現できる（**図-11**）．その上，製品障害に関する問合せが目立っている機種があれば，それを開発部門に報告することで，製品障害の初期発見を行えるという効果も確認されている．

従来の手作業による分析では，適当に選択したデータ（たいていの場合，先頭から200ないし300件程度）を抽出し，全体の分布をまとめた上で，問題になりそうな内容に関しては，全データを検索し，似たようなレポートを集めてきてさらに掘り下げ，分析結果をまとめるという手順がとられている．しかしこの場合，抜け（最初に目に付かない問題）が存在する危険性が高いことや，実質的に対象とするレポートが数百件程度に限られてしまい，過去のデータとの比較や時間的経緯の傾向などまで手が回らないことなどが問題になる．

このように人間が実際に文書を読み，分析結果として内容の内訳などを出そうとする場合，1時間あたりせいぜい300件程度の文書を処理するのが限度である．それに対し，テキストマイニングの技術を用いて処理することにより，全データを対象にできるとともに，さらに深い分析を行う余裕ができてくる．

もちろん機械的な処理には限界があり，自動的に完全な分析ができるわけではないが，膨大な量を対象とした場合，全体的な傾向は顕著に現れるため，ここで紹介した事例においては，多少の分析エラーが混在していても，それはノイズとして無視できる範囲に収まっている．

テキストマイニングが可能にすること

以上，宝の山としての可能性を秘めていながらも活用の難しい膨大な文書データから有効な知識を発掘する技術を紹介した．

テキストマイニングは，比較的新しい分野であり，自然言語処理とデータマイニングや視覚化の技術が密接に結び付くことにより可能になってきている．

このようなテキストマイニングの対象となるのは，アンケートやコールセンタの報告書による顧客の声の分析に限らない．たとえば，特許文書を対象とすることにより，ある技術をどのような組織がいつ頃開発しているかといった業界動向の分析や，ある企業がどのような技術をいつ頃開発しているかといった企業の特許戦略の分析などが容易に実現できる[4]．また，売上げと営業活動内容を記述した営業報告書を対象にすれば，どのような活動が売上げに結び付くかといったノウハウを獲得できる可能性がある．

近年ナレッジ・マネージメントに対する取り組みが盛んになっている中で，実際に行われている活動は，さまざまな知識を文書にして共有できるようにする仕組みの構築であることが多い．このような取り組みの中でも，電子化された文書がますます増えており，それが次第に手に負えない量になるにつれて，テキストマイニングの必要性がますます高くなっていくものと考えられる．

謝辞　データを利用させていただくとともに，日頃貴重なアドバイスをいただいている日本アイ・ビー・エム（株）PCヘルプセンターの皆様に感謝いたします．

参考文献

1) 乾, 内元, 村田, 井佐原: 文末表現に着目した自由回答アンケートの分類, 情報処理学会研究報告, 98-NL-128, pp.181-188 (1998).
2) 大隅, Lebart, 片山, 竹前: テキスト型データの統計解析システムーSPAD. T/Jー, 第11回日本計算機統計学会シンポジウム, pp.19-22 (1997).
3) 那須川, 諸橋, 長野: テキストマイニング：膨大な文書データからの知識獲得 ー概要ー, 情報処理学会第57回全国大会講演論文集, 3-77 (1998).
4) 那須川: テキストマイニング: 大量文書データの戦略的利用技術, 人工知能学会全国大会第12回論文集, pp.640-641 (1997).
5) 野美山, 武田: 情報獲得のパラダイムとしての Information Outlining, 情報処理学会, 情報学基礎シンポジウム (1996).
6) 諸橋, 那須川, 長野: テキストマイニング：膨大な文書データからの知識獲得 ー意図の認識ー, 情報処理学会第57回全国大会講演論文集, 3-76 (1998).
7) http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html
8) http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

（平成11年3月11日受付）