知 能 と 複 雑 系 118－3
（1999．10．15）

大量のテキストからの知識マイニング

那須川 哲哉　　長野 徹　　武田 浩一
日本アイ・ビー・エム株式会社 東京基礎研究所
〒242-8502 大和市下鶴間 1623-14
{nasukawa, tohru3, takedasu}@jp.ibm.com

あらまし： 大量のテキストデータは貴重な知識源となる可能性を持つ反面、個々の内容を解釈し全体的な傾向をつかんで有益な知識を獲得するためには大変な労力が必要となる。そのため、せっかくのテキストデータが有効に活用されていない場合が多い。本稿では、テキストマイニング技術の確立を目標として数十万件に及ぶ顧客からの問合せデータを実際に処理した試みを通じ、どのような処理を行えばどのような知識を獲得できるかについての知見を示す。
最初に、大量の文書データの複雑性と分析の困難性を示し、そこからいかにして有益な知識を自動的に抽出するかを考察した上で、具体的なデータへの適用例とその結果を示す。

キーワード： 自然言語処理　データマイニング　知識発見　テキストマイニング

Knowledge Mining from Huge Amounts of Text Data

Tetsuya Nasukawa　Tohru Nagano　Koichi Takeda
IBM Research, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502
{nasukawa, tohru3, takedasu}@jp.ibm.com

Abstract: A huge collection of important documents is naturally a valuable knowledge resource. However, because of the limitations of human document-handling capability, it is difficult to make good use of such documents. In this paper, we explore technology for discovering knowledge in vast amounts of documents, in particular, on customer claims that may directly lead to some various desirable results, such as fewer calls and faster detection of problems. After summarizing the requirements and our approach to the extraction of knowledge from text data, we illustrate actual procedures for mining knowledge from text data, and demonstrate the results.

Keywords: natural language processing, data mining, knowledge discovery, text mining

1 はじめに

新聞、雑誌、書籍、日記、各種報告書などに加え、近年ではメールやフォーラム、チャットなど，世の中には膨大な量のデータがテキスト形式で記述され蓄積されている。多くの人々が書物を読んで学習するように、このような文書データは知識の源泉となる可能性があり、計算機を用いて膨大な文書データから何らかの知識を抽出できないかという期待が生じる。特に、計算機科学の発展に伴い、ハード的にもソフト的にも膨大な量のデータを処理する環境が整ってきていることと、文書データの電子化が進み、計算機上に蓄積されている文書データの量が既に膨大な量に及んでいることが、この期待をますます強めている。

ところが、文書データを構成している自然言語による記述は非常に柔軟性が高く、多様な内容を自由に記述できる反面、その内容を解釈するには、膨大な知識が必要となる。したがって機械的な処理によって文書データから有益な知識を抽出するには様々な問題を解決する必要がある。すなわち、文書データを単なる文字列の集合と考えて処理しても、何らかの規則性に基づいた文字列のパタンしか抽出することができず、知識というイメージからはかけ離れた結果しか得ることができない。有益な知識を抽出することを考えると、文書データ中に表現されている意味内容を考慮する必要がある。ところが、意味を扱うためには知識が必要であり、大量の文章を処理するにあたって、予め機械的に処理できるような知識を十分に用意しておくことは現時点の技術レベルでは実質的に不可能である。したがって文章内容の解釈を自動化するのは非常に困難であり、文書データの処理はもっぱら手作業で行われているのが現状である。

ただし、人が実際に目を通せる文書の量には限度があり、膨大な文書の内容を全て把握するのは実質的に不可能であるため、これを支援するための技術が存在する。例えば、全データを網羅的に処理するのは不可能であることから、作業効率を上げるために処理対象を絞り込む、すなわち必要とする内容を含むデータに処理対象を限定するために検索の技術が用いられている。また、文書データ全体の中にどのような内容が含まれていて、その分布がどのようになっているかを調べる、いわば文書内容を分類・整理するための技術としてclusteringやclassificationなどの技術も利用されるようになってきている。しかし、これらは、あくまでも人間による知識発見の処理の効率化を図るための技術であり、知識の候補として着目すべき内容を提示してくれるものではない。

文書データを完全に活用するためには究極的には、膨大な文書データの内容を全て理解し、知りたい内容を対話的に教えてくれるようなシステムが理想かと考えられるが、そこまでの道のりは遠い。本稿では、これらの中間に位置付けられる技術、すなわち、検索や分類よりも深いレベルで処理を行い、着目すべき内容を自動的に認識・抽出する技術を対象とする。テキストマイニングとして、このような技術を対象としようという考え方は、次第に出てきているものの、まだ決定的な技術のメドは立っていないのが現状である。[1]

本稿で対象としている知識発見技術と、検索、分類技術との位置付けや特徴を表１にまとめる。どれも文書データを対象とし、何らかのレベルで言語処理を行い、文書データの内容を処理しやすい形式に変換することでは一致しているが、処理のレベルが異なっている。検索においては、全文検索方式の場合、文書データを単なる文字列の集合として表現するが、キーワード検索方式の場合は、文書データから形態素解析等を用いて単語を切り出し、文書データを単語の集合体として扱うことになる。分類においては、ベクター・スペース・モデルのような単語の分布状況が文書データの内容を表現するものとして処理される。それに対し、知識発見を考えると、意味的な処理がより重要となってくる。すなわち何が着目すべき有益な内容かを認識する上では文書データ中にどのような内容が含まれているかを把握する必要があり、それはベクター・スペース・モデルのよう

表 1：膨大な文書データの活用技術

処理の深さ	処理の概要	機能	技術的要素	処理対象データ内容の表現形式	自然言語処理内容
レベル1	検索 Search	目を通す対象を絞り込む	Information Retrieval	文字列 単語列	・単語の抽出（語の基本型へ置換など）
レベル2	分類・整理 Organize	全体的にどのような内容が含まれているか把握	Clustering Classification	単語の集まり (Vector Space Model)	・単語の分布状況の分析
レベル3	知識発見 Discover	面白い内容を自動的に抽出	NLP Data Mining Visualization	意味的概念の集まり	・意味の分析 ・依存関係の分析

な単語の分布状況のみで処理することはできない。

例えば多義語の問題があり、同じ単語でも全く違う内容を表現していれば区別する必要がある。また、何についての記述であるかを把握するためであれば、文書中にどのような単語が含まれているかを参照するだけですむが、何がどうしたかまで知ろうとすれば、単語と単語の関係まで参照する必要性が出てくる。検索や分類では自立語、特に体言のみが用いられるのが一般的であるが、何がどうしたか、良いのか悪いのかといった内容を扱うためには用言の情報も重要であり、さらに、否定や疑問、要望といった区別を行うためには、付属語、特にモーダル表現が重要な役割を果たす。

本稿では、知識発見のためのテキストマイニング技術の確立を目標として数十万件に及ぶ顧客からの問合せデータを実際に処理した試みを通じ、どのような処理（特に、どの程度の意味処理）を行えばどのような知識の獲得を実現できるかを示す。

まず、第2節で、大量の文書データの複雑性と分析の困難性を示し、そこからいかにして有益な知識を抽出するかを第3節で考察した上で、具体的なデータへの適用例とその結果を第4節で示す。

2 テキストデータの複雑性

文書データは、記述内容の自由度が高く、多様な内容の表現が可能であるために、分析という観点からは、その多様性が処理の困難度を高めることになる。

マイニングを行って何らかのパタンを見つけるためには、対象データの中に同じパタンが繰り返し現れていなければならないが、表現の多様性が大きいほど、同じパタンが繰り返される割合は低くなる。従って、出現する言葉の種類が少なく、表現している内容が限定されているデータほどマイニングに適していると考えることが出来る。そのような観点から、現在対象としている顧客問合せデータの特徴を示す。

顧客問合せデータの特徴

対象データは，日本ＩＢＭ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い合わせに関する報告書である．このヘルプセンターでは，ＩＢＭのＰＣ関連製品に関する様々な問い合わ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
- 問い合わせの種別（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共に，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，その総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.

テキスト部分の特徴としては、

- オペレータは文筆のプロではなく、限られた短い時間で概要を記述するため、省略や文法的誤り、誤字脱字などが多い
- 複数（数百）人のオペレータが記述しているため、表現方法に個人差があり、同じ内容を異なる表現で示している場合が多い。例えばThinkPad という機種名を人によっては TP や T/P と表現している
- 内容的にはPC製品に関連した問合せという非常に限定された世界を対象としているため、使われている語は比較的限られている

という性質をあげることができる。

一例として、1997年8月のデータ約4万件において、個々の文書データのサイズや、語の分布などを見てみると、データ一件あたりの応答内容のテキスト部分は１５０文字程度であり、キーワード数は３５語程度となっている。使われている語はかなり限られており、約4万件のデータ中の

総キーワード数:　1473444語

に対し、

キーワード異なり語数:　101987語

であり、同じ語が何度も繰り返し出現していることが分かる。具体的には、

2回以上出現 95.6%（異なり語数36811語）
5回以上出現 91.9%
10回以上出現 89.1%
20回以上出現 85.8%
100回以上出現 75.1%
760回以上出現 50.1%
1000回以上出現 45.0%

というように、20 回以上出現している語だけでデータ全体の 85%以上をカバーしている。すなわち、データ中に20回以上出現する語を辞書登録し、その語に意味的属性を与えておけば、データ中の 85%以上の語に対して辞書から情報を取ってこられるということになる。

従って、このデータを処理する上では、オペレータによる表現の個人差を吸収するための同義語辞書と、高頻度で出現する語に対して意味的属性を与える辞書を構築することが有効であるという見通しが得られる。

分析内容の多様性

テキストデータから有益な知識を抽出するといっても、何が有益かは利用者の価値観によって異なってくる。例えば、ヘルプセンターの責任者であれば、業務効率の向上やコストダウン、顧客満足度の向上に関心があるため、顧客問合せデータを分析することによって、どのような問合せに対して処理時間がかかっているか、オペレータは適切な応対をしているかといった知識が得られれば良い。それに対し、開発部門の責任者にとっては、製品の不具合や使い勝手の良し悪しに関する知識が有益である。また、マーケティング部門の責任者であれば顧客の行動形態に関する知識を得て販売戦略につなげようと考えるであろう。

従って、多様な観点からの分析が必要であり、このような観点をいかにして取り込むかも重要な課題となる。

3. テキストデータからの知識発見

以上の観点から、我々は知識発見を目的としたテキストマイニングのプロトタイプシステムTAKMIを構築し、膨大な文書データの処理を通して知識発見技術の開発に取り組んでいる。[2] TAKMIは大きく分けて3つのコンポーネントから成り立っており、各コンポーネントの処理は以下のような特徴を持つ。

自然言語処理を中心とした概念抽出部[3]

- 名詞句などの体言を中心に抽出したキーワードのみでは、表現できる内容に限界があるため、動詞や形容詞などの用言の情報も抽出する
- 抽出結果は単なる文字列として扱うのではなく、カテゴリ分けという形で意味付けする
- 述語に関しては、モーダルを分析することで意図を解釈し、「疑問」「要望」「不満」などの区別を行う
- 体言を中心とする概念と用言を中心とする概念とを組み合わせ、「何がどうした」「何をどうする」といった概念を抽出する

マイニング部

- 概念抽出部でテキスト部分から抽出した概

念と、同じデータの定型部分から抽出した情報とを組み合わせて分析する
- 増減傾向など時間的な変化の特徴を捉える
- データ全体における概念の分布と、特定の概念を含むデータ集合内の概念の分布とを比較することにより、特定の概念の特徴を捉える
- 概念抽出部でカテゴリという形で付加された各概念の性質を考慮して処理を行う

視覚化及びインタラクティブな分析部
- 多様な観点を考慮してカテゴリという側面からデータ内容を視覚化する
- ユーザーが興味を覚えた対象をインタラクティブに指定し、その対象に関して掘り下げた分析を可能にする

このような枠組みによって、実現される分析内容に関しては、[2]で概要を述べてあるので、以下では特に、相関ルールの導出を中心に考察する。

相関ルールの導出

大規模データからの知識発見としては、データマイングにおける相関ルールの導出[4]が良く知られている。文書データに関しても同様の技術を適用し、言葉と言葉の相関ルールを導出する事で何らかの知識が得られるのではないかとう期待が生じる。ところが実際には、文書データではアイテムとなる言葉の種類が膨大な数に及ぶため、導出される相関ルールの数も膨大になり、その上、雑多な内容が混合されているため、この中から有用なルールを見出すのは非常に困難である。

実際に、顧客問い合わせデーター月分（約４万件）からの相関ルールの導出実験を行った結果を以下に示す。

一つの問合せデータを一つのバスケットとみなし、その中に含まれる概念をアイテムとして、相関ルールを導出した。一つの問合せデータ中には、顧客からの問合せの内容とそれに対する応対内容とが、「Q：～」と「A：～」というように書き分けられており、ここでは、顧客の問合せ内容を対象とするために、文書データ中の「Q：～」の部分のみを対象とした。

まずは、一般的なキーワード抽出を行った結果を用いて相関ルールを抽出した場合にどうなるかの実験を行った。

概念として名詞句のみを利用した場合
- アイテム数（名詞句の種類）：55716
- バスケットあたりの平均アイテム数：10.3
- サポート：120 以上
- コンフィデンス：1%　以上
- 出力ルール数：2973
- 導出ルール例：

英語版　==>　英語
VOICE　==>　TYPE
アプリケーション　==>　CD
ハードディスク AND ROM　==>　CD
ROM AND インストール　==>　CD
FOLLOW　==>　LOG
SUPER　==>　OFFICE

この場合、アイテム数が非常に多くなるのが特徴で、導出されるルールは、例に示したように「同じ複合語を構成する要素の語は共起し易い」という程度の情報を示すルールが大半を占めている。

そこで、名詞句を中心とした従来のキーワードベースの処理による知識獲得は困難という判断から、述語概念も利用すると共に、名詞概念としては、約１万６千語登録されているカテゴリ辞書中に定義されている主要概念のみを対象として相関ルールの導出を行った。その結果は以下の通りである。

概念としてカテゴリ辞書で定義された語（述語類を含む）を利用した場合
- アイテム数：8183
- バスケットあたりの平均アイテム数：11.7
- サポート：130 以上
- コンフィデンス：1%以上
- 出力ルール数：31278
- 導出ルール例：

投入する ==>　電源
王様　==>　翻訳
増設　AND　初期化==>　BIOS
不正　==>　処理する
黒い　==>　画面
接続するAND応答する ==>モデム

この場合は、高頻度語のみが対象となったため、

アイテム数が減少すると共に導出されるルールの数が増大するという結果が得られた。ルールの例を見ると、名詞句のみの場合には、「何と何が共起し易い」というルールが多かったのに対し、述語も対象としたために、「何がどうなる」「何をどうする」といった内容を示すルールが多く見られた。しかし数万件に及ぶルールから有益なルールを選択するのは困難であり、有益なルールを判定する基準が必要となる。その際、サポートやコンフィデンスの値のみでは、なかなか有効な判断ができない。例えば、ある製品に対して「操作が難しい」というコメントが高い割合で存在したとしても、全ての製品に関して同じようなコメントが同じような割合で存在するなら、「その製品の操作が特別に難しいというわけではなく、注目に値しない」という判断ができる。

特異性の検出

そこで、『類似概念に関して、同じ性質の出現度合いが極端に異なる個所に注目すれば、そこに意味のある情報が含まれている可能性が高い』という判断基準でルールを選択する仕組みを取り入れた。ここで、類似概念とは、「ハードウェアに属する概念」や「ソフトウェアに属する概念」、「同じタイプの製品」など、何らかの観点から同じ性質を共有する概念を示す。例えば、全ての製品に対して「操作が難しい」というコメントが多少なりとも存在する中で、その比率が他と比べて極端に高い製品が存在すれば、その製品には何らかの問題がある可能性が高いという判断を行う。

このような処理を行うためには、文書中の各表現が基本的にどのような性質の概念を示しているかを考慮して文書中から情報抽出を行う必要があり、単なる文字列としてのキーワードを抽出する処理では不十分である。また、正確な判断を行うためには、異なる表現であっても同じ内容を示している場合には、それを同じものとして処理する同義語の処理が重要である。

4 顧客問合せデータへの適用結果

前節で述べた手法を、第2節で示した顧客問合わせデータに適用した結果を示す。まず、概念抽出部においては、形態素解析と係り受け解析を行い、その過程でカテゴリ辞書と同義語辞書を参照することにより、カテゴリ付けした概念を抽出すると共に同義性を吸収している。カテゴリとしては、名詞概念に対しては、「ソフトウェア」、「ハードウェア」、「それ以外の専門用語」といった程度の大まかな分類を行い、述語類に対しては、「質問」、「要望」、「問題」、「好評」といった分類を行った。さらに、これらの名詞概念と述語概念とが係り受け関係にある場合は、その組み合わせを複合概念として抽出した。例えば、ある月のデータ約4万件の中で、「ソフトウェア」に分類された概念と「問題」に分類された概念との複合概念を見ると、「ファイル…見つからない」という概念が55件のデータから抽出されていた。そのうち「ファイルが見つからない」という表現（文字列）が含まれているデータは13件のみであり、残る42件のデータでは、「ファイルがみつかりません」「ファイルがみつからない」「ファイルが見つかりません」という表現から、「ファイル…見つからない」という複合概念が抽出されていた。このような述語概念や複合概念は文章内容の分類に有効であることが確認されている。[5]

特徴的相関ルールの導出例を図1から図3に示す。図1には、98年1月から5月までの期間内で総合案内に分類されたデータのうち、Aptiva、ThinkPad、PS/Vという3つのブランドの機器に関する問合せ約1万件（9356件）のデータを対象とした分析の様子が示されている。この図の下半分に特徴的相関ルールを分析する画面が表示されており、左側のリストで「専門用語…要望」の複合概念と「ハードウェア」の概念との組み合わせが指定されている。その右側に対象データ中の各「ハードウェア」概念に対する「専門用語…要望」の複合概念の分布状況が示されている。このうち、ブランド別の要望の特徴が示されている部分を切り出して拡大したのが図2である。各セルにおいて左側の数値が問合せの件数を示し、括弧内の数値は左端に表示されているブランドに関する問合せの中での比率を示している。問合せ総数の比は、Aptiva：ThinkPad：PS/Vがおよそ

10：3：1 となっている。そのため、例えば「メモリを増設したい」という要望を見ると、絶対数ではAptivaに関する問合せが46件と一番多いが、構成比ではPS/Vに関する問合せにおける比率が2.15%と他より特に高い値になっている。したがって、PS/V に関する「メモリを増設したい」という問合せのセルがハイライトされている。

図3には、ある月のデータに本手法を適用し、機種別の問題の分布を分析した結果が示されている。いくつかの機種で特定の問題がハイライトされる結果が得られているが、ここでは、ある機種に対して「遅い」という問題を指摘される割合が他の機種に比べて特に高いという結果に着目し、その問合せに関連性の高いハードウェアを右上のリストに表示している。図に示されているとおり、この問題に関しては、ハードディスクの関連性が高いという結果が得られている。実際にこの機種においては一部のハードディスクでシールディングの問題が発見されており、本手法が製品の問題の発見に役立つ事が検証されている。このように、テキストマイニングにおいては、単に、（ある機種が特に遅いというような）相関ルールを発見するだけではなく、その内容を掘り下げて、その現象に関連する文脈の情報を分析することにより、（何が原因かといった）さらに詳細な分析が可能になるところに一つの特長がある。

5 おわりに

以上、膨大な文書データから有益な知識を発見する試みを示した。今回対象としたデータは対象分野が限定されていたことから、意味情報を与えるためのカテゴリ辞書を比較的小さな労力で構築することができ、その結果、単なる文字列としてのキーワードよりも、ある程度意味を持ったレベルの概念を扱うことの有効性を確認することができた。

今後の課題としては、カテゴリの定義、カテゴリ付け、同義性の判断を行うための知識の自動獲得があげられる。特に、「何がどうした」「何をどうする」といった複合概念は、元の文書データに目を通すことなく文書データの内容を把握するのに有効であるが、組み合わせとしての数が非常に大きくなるため、統計的な処理を行う上では、データがスパースになってしまうという問題が発生する。そこで、複合概念の同義性をうまく認識し、同じ意味の代表的な表現におきかえることが、有効なマイニングを行う上では重要である。

参考文献

[1] Hearst, M: “Untangling Text Data Mining,” In Proceedings of ACL-99, pp. 3-10 (1999)

[2] 那須川，諸橋，長野：「テキストマイニング—膨大な文書データの自動分析における知識発見—」情報処理，pp. 358-364 (1999)

[3] 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 —意図の認識—」情報処理学会第57回全国大会講演論文集，3-76 (1998)

[4] Agrawal, R., Imielinski, T., and Swami, A.: “Mining Association Rules between Sets of Items in Large Databases,” In Proceedings of the ACM SIGMOD ‘93, pp. 207-216 (1993)

[5] 長野，那須川，諸橋：「テキストマイニングのための情報抽出手法」人工知能学会全国大会論文集，pp. 411-412 (1999)

[6] 鈴木「データベースからの特徴的ルール発見のための一般性と正確性の信頼性同時評価手法」人工知能学会誌，p. 139-147 (1999)

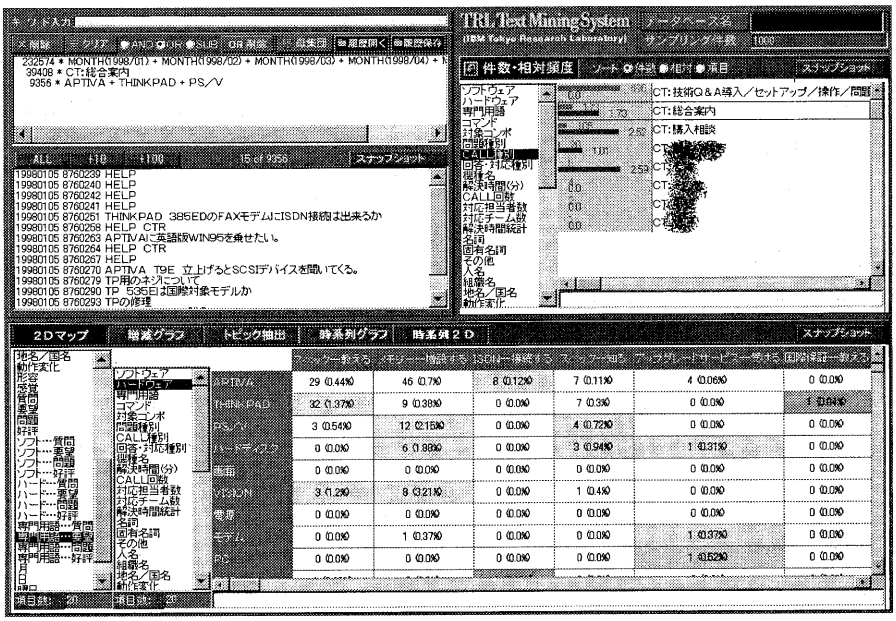

図１：ブランド別問合せ内容の特徴分析（１）

	アベック→勧める	メモリー→増設する	ISDN→接続する	アベック→知る	アップグレードサービス→受ける	国際保証→教える
APTIVA	29 (0.44%)	46 (0.7%)	8 (0.12%)	7 (0.11%)	4 (0.06%)	0 (0.0%)
THINKPAD	32 (1.37%)	9 (0.38%)	0 (0.0%)	7 (0.3%)	0 (0.0%)	1 (0.04%)
PS／V	3 (0.54%)	12 (2.15%)	0 (0.0%)	4 (0.72%)	0 (0.0%)	0 (0.0%)

図２：ブランド別問合せ内容の特徴分析（２）

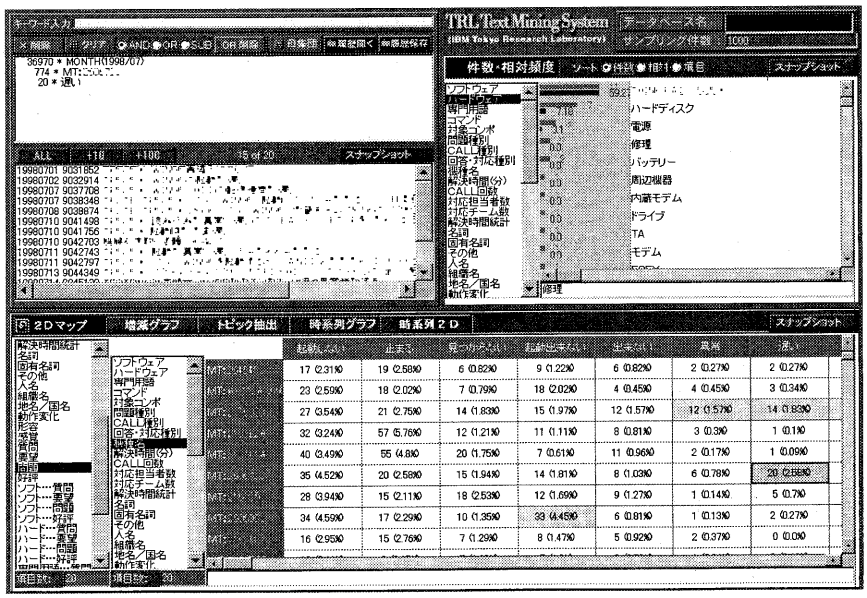

図３：機種別問題の特徴分析