研究論文　3-6

自然言語情報の分析手法と経営学的諸問題への応用

後藤正幸[1]

1. はじめに

情報技術の高度な発展によって、膨大な情報が蓄積される時代となった。インターネット上には刻々と増加する情報が氾濫しており、データや情報の取り扱い方のスタンスも変化している。同様に、データウェアハウスやデータベースの恩恵により、企業は膨大な顧客データ、購買履歴データ、アンケートデータ、苦情データなど、ビジネスに活用しうるデータを電子媒体で手に入れることが可能となっている。しかしながら、これらの溢れる情報の量は、人手による処理能力の限界を超えており、計算機能力を上手く活用して有用な情報をいち早く活用することに成功した企業が、他社をリードすることが可能となる時代となった。

情報がある程度大量に蓄積されると、それらの中から必要な情報を効率よく検索するための技術が必要となる。このような情報検索（Information Retrieval）の概念は1950年代後半に生まれ、計算機科学の分野で高度に発達した［1］,［2］。情報検索は、図書館データベースの検索といった場面で応用されてきたが、近年ではインターネットの Web ページの検索エンジンで、一般のユーザに広く利用されるに至っている［3］。情報検索はもともと、文章や画像、音声などの非数値データの内容分析を行い、その結果を蓄積することで、後の検索要求に効率的に応えようとするものである。情報検索における文書データの内容分析は、計算量を抑えつつ、検索の精度を表す指標である適合率と再現率を高めるための準備ともいえる。これに対し、自由な形式で記述された大量の文書データから意味のある情報を抽出したり、長文からなるテキストに書かれている内容の趣旨を要約文として提示するというように、分析そのもの（自然言語処理）に興味の対象がある場合も多い［4］。これらの分野は、もともと書物の作者推定やテキストの分類問題、エッセイやレポートの自動採点［5］、文書の自動要約や自動翻訳など多くの応用を持ち、言語モデル研究の分野で議論がなされてきた。最近では、大量のテキストデータから価値のある情報を掘り出すという意味を込めてテキストマイニングといった新しい言葉も定着している［6］－［8］。実務的に利用可能なソフトウェアの登場によってさらに注目度は高まっており、今後さらなる応用上の研究が進むものと考えられる。

本稿では、情報検索やテキストマイニングに見られる自然言語処理技法のうち、企業の経営問題、経営工学的問題に有用となる要素技術に焦点を絞って議論を行う。これらの自然言語情報の分析手法について、経営工学的諸問題への応用という点で意義のある方法論について概要をまとめ、具体的応用例を与えることでその有用性を示す。

2. テキストマイニング

テキストマイニングは膨大な自由記述テキストファイルに対する分析手法の総称であり、自然言語処理とデータマイニング技術の双方をあわせ持った技術として注目を集めている。テキストマイニングの要素技術には、テキストクラスタリング、テキスト分類、相関ルール抽出、自動スコアリング、情報抽出、情報縮約などがあげられる［6］。

①テキストクラスタリング：複数の文書データを分析し、それらの間の類似性を考慮して、似たもの同士を集めた文書ファイル集合に仕分ける処理をいう。例えば、多くの学生レポートについて、こ

1　武蔵工業大学環境情報学部助教授

れらを似た内容を扱ったレポート同士をグルーピングして、いくつかのクラスタに仕分けるような処理である。

②テキスト分類：与えられた文書データに対し、すでに与えられているクラス（トピック）のどれに属するかを判定する処理である。例えば、ニュース記事は、“社会”、“経済”、“スポーツ”など、既存のトピックのどれかに属すものと仮定し、新しく得られたブログ記事がどのトピックについて論じたものであるのかを自動で判別するといった処理である。

③相関ルール抽出：文章中に同時に表れる単語の共起関係を分析し、単語同士の関係性を可視化する技法である。例えば、多くの論文において、“ベイズ”という単語が使われれば、同時に“損失”という単語も使われている頻度が高い場合、これらには共起関係があるとみなし、意味的な関係性があると関連付けるといった処理である。

④自動スコアリング：エッセイやレポートなどの質を評価し、専門家に代わって採点をするための技法である。例えば、優れたレポートの特徴を予め定義しておき、新たに与えられたレポートについて、この特徴群とマッチングすることで採点を行うといった処理である。

⑤情報抽出：文書データ中に含まれる重要な情報を抽出する処理技法である。例えば、文書データの中から、ある話題について論じられた箇所のみを取り出すといった処理である。

⑥情報縮約：長文の文書データに対し、意味を損なわずに自動で要約し、短いエッセンスの形で提示する技法である。例えば、長い Web テキストに対して“要するに何が書かれているのか”を短い文に集約する処理である。

テキストマイニングでは、これらの要素技術を援用し、様々な課題について応用研究が進んでいる［6］。例えば、多くの営業スタッフの営業日報を分析したり、Web の検索結果を自動分類したり、あるいはコールセンターへの問い合わせの分類といった様々な応用例が報告されている。日頃の営業スタッフが足で稼ぎ、蓄積した膨大なデータを分析し、必要な時に過去の事例が容易に参照できたり、顧客の要望を自動で集計できたりすれば、営業スタッフの生産性は格段に向上するであろう。コールセンターへの問い合わせの問題においても、顧客の問い合わせ内容を自動分類し、適切な対応や回答をテンプレートとしてオペレータに提供できれば、迅速な対応が可能となり、顧客満足度の向上が期待できる。また、過去の問い合わせと対応データから、今後の適切な対応に向けて新たな知見を発見することも可能である。このように、テキストマイニングの技術は、企業活動においても多大なベネフィットを与える可能性が高く、今後の応用研究に期待が持たれている。

2.1　形態素解析

文章を構成する最小の要素を形態素（morpheme）という［4］。形態素は文章を分解した時の最小の言語要素といえる。自由記述文章は、人間が書いたり、読んだりすれば比較的容易に意味が取れるものであっても、これをコンピュータに理解させようとすると非常に難しいものを多々含んでいる。そこで、自然言語処理の伝統的なアプローチとして、文章を形態素に分解し、それらをもとにして文章の特徴量を規定することで、コンピュータが処理し易い形に成形する方法が取られている。すなわち、自然言語処理の分野では、まず文章を構成要素である形態素に分解し、語形変化分析、品詞のタグ付けなどを行う必要があり、これを形態素解析（morphological analysis）という。例えば、日本語ワープロにおける“かな漢字変換”においては、入力されたひらがな文章を単語に分割し、その後、文脈を考慮してそれぞれの単語の漢字変換を行うという処理が必要になる。情報検索分野においても、ロボット型のサーチエンジンでは、Crawler　や Robot といったプログラムがひたすらインターネット上を徘徊し、新規更新された Web ページの情報を自動収集して、検

索キーワードを抽出し蓄積するという処理が行われる。すなわち、分析対象である文章を単語に切り分ける操作(単語分割)がどこかで必要となってくる。ここでは、以上のような形態素への分割という重要な役割を担う形態素解析について概要を述べる。

2.1.1 形態素解析の概要

例えば、「野球がやりたい。」という文章を形態素解析により分析すると、

野球　(名詞―一般)
が　(助詞―格助詞―一般)
やり　(動詞―「やる」の連用形)
たい　(助動詞―「たい」の基本形)
。　(記号―句点)

という結果が得られる。このような分析を行うフリーのツールとして“茶筅”　などが公開されている［9］。

実はこのような形態素への分解は、コンピュータにとっては容易な処理ではない。例えば、「各自営業に出る」では、

各　(接続詞―名詞接続)
自営業　(名詞―一般)
に　(助詞―格助詞―一般)
出る　(動詞―「出る」の基本形)
。　(記号―句点)

という結果が得られたりするが、人間が見れば、“各自”(で)“営業”“に”“出る”“。”という分割を容易に思いつくはずである。このように、特に日本語では、英語のように単語間にスペースを挿入する“分かち書き”の習慣がないため、単語への区切り方が一意に定まらず、常に曖昧性が残る。

計算機能力がまだ貧弱であった時代には、経験則をルール化して形態素に分割する方法や複数の区切り方の候補についてコスト関数を定義してコスト最適解を求める方法などが使われていたが、そのチューニングには人間が介在しており、精度向上には壁が存在した。しかしながら、最近ではコンピュータパワーにものを言わせて、大量のテキストデータから、各形態素の出現頻度情報を蓄積して利用することが可能となり、これを受けて確率モデルを用いた確率的言語モデル[10]の研究が進んだ結果、95%以上というかなりの高精度が達成されている。

長さ n のテキストデータを x^n とし、長さ m の単語系列を w^m とする。このとき、確率モデルに基づく形態素解析の問題は、テキストデータ x^n が与えられたときの単語列 w^m の条件付確率 $P(w^m|x^n)$ を最大化する問題として定式化できる。すなわち、最適な推定量 $\hat{w}^m$ は、

$$\hat{w}^m = \arg\max_{w^m} P(w^m|x^n) \qquad (1)$$

で与えられる。これは、推定の誤り確率を最小化する解、すなわち0-1損失関数を用いた場合のベイズ最適解である[11]。単語列 w^m の条件付確率 $P(w^m|x^n)$ の計算には、マルコフモデルや隠れマルコフモデル[10]などの確率モデルが用いられる。また、それらの確率モデルのパラメータ推定には、予め人手によって形態素に分類された教師データをコンピュータによって処理し、頻度情報を統計値として用いることができる。

実際に確率モデルが与えられたもとで、条件付確率 $P(w^m|x^n)$ を最大化する単語列 w^m を探索する問題については、動的計画法の一種であるビタビアルゴリズムによって実用的な計算量で、最適な単語分割を得ることが可能である［4］。

2.1.2 単語単位の情報源

1948年に Shannon が提唱した情報理論は、近年の高度情報化社会における基盤理論として、現在もその価値を失っていない[12]，[13]。この理論は、通信や符号化の問題を美しい理論体系で説明したものであり、情報圧縮限界としてのエントロピーの性質や通信路符号化の限界（通信路容量）について論じている。情報理論では、もともと意味的な情報(セマンティック情報)の解釈を諦め、記号的な情報(シンタックス情報）の統計的性質を用いて通信のモデルを構築したことによって成功を収めた。一方で、

自然言語の分析技術との関連性についても興味深いテーマとなっている［14］。ここでは、形態素解析と類似した問題を情報理論的に考察した例について述べる。

いま、情報源が確率的にテキストデータを出力することを想定する。もし、情報源から出力されるデータが1文字ずつで規定された確率に従っているとすれば、これは従来の Shannon 理論で扱っているモデルとなる。いま、情報源からのデータは1文字ずつに与えられた確率ではなく、単語単位に与えられた確率に従って出力するモデルを考える［15］。これは、自然言語が形態素を単位として構成されていると仮定する限り、妥当な仮説であろう。すなわち、情報源は単語単位でデータを出力するが、それを受け取るユーザにとっては、それは単語を連接して得られるデータ列としてみなされることになる。

テキストデータの x^n エントロピーは、

$$H(X) = \lim_{n\to\infty} \frac{1}{n} \sum_{x^n} P(x^n) \log P(x^n) \qquad (2)$$

で与えられ、これはこの情報源から出力されるデータの平均的な情報圧縮限界を意味している。一方、情報源からのデータは単語単位で出力されるため、単語の確率構造に依存したエントロピーが存在し、

$$H(W) = \lim_{m\to\infty} \frac{1}{m} \sum_{w^m} P(w^m) \log P(w^m) \qquad (3)$$

で与えられる。このとき、もしテキストデータ x^n が与えられたときの単語列 w^m への分解の仕方が一意であれば、

$$H(W) = \frac{H(W)}{E[|W|]} \qquad (4)$$

という関係が成り立つ。ここで、$E[|W|]$は平均単語長（単語の長さの平均値）である。データ圧縮の問題に限れば、この情報源について再起時間定理を証明することができ、これを用いた Ziv と Lempel による符号化法が漸近最適であることも示すことができる［15］。この事実は、現在主流となっている Ziv-Lempel の符号化が、実は単語単位でデータを出力する情報源についても良好な圧縮技法となっており、そのために強力な圧縮ツールとして現在もなお利用されていることを裏付けるものといえる。

一方、日本語の文章に見られるように、自然言語においては、単語への分割が一意に定まらない方が一般的である。このような情報源モデルについては、残念ながら、明示的なエントロピーの数式は導かれていない。そこで石田らは、単語への分割が一意に定まらない場合について、情報源のエントロピーの上限と下限を導出している［16］。さらに、実際にこのような情報源から得られるデータをベイズ決定に基づいて単語に分割するアルゴリズムを構成している。この方法は、実際には形態素解析で用いられるビタビアルゴリズムによる探索法と同等である。

2.2 ベクトル空間モデル

本節では、形態素解析により、各文書 $d \in \Delta$ について単語への切り出しが行われた後、情報検索やテキスト分類の問題を取り扱い易い問題に落とし込んだモデルであるベクトル空間モデルについて述べる［17］。ただし、Δ は分析対象である文書集合を表す。ベクトル空間モデルでは、文書中の出現単語の頻度に基づき、文書の特徴量を1つのベクトルで表現することで、文書を空間上の点として表す。出現単語に基づくベクトル空間を構成し、文書を空間上の点として表現することで、文書同士の類似性を距離の概念によって数学的に取り扱うことが可能である。このモデルは、計算機で実装する際に強力な枠組みを与えるものである。

2.2.1 ベクトル空間と文書－単語行列

分析対象である文書集合を $\Delta = \{d_1, d_2, \cdots, d_D\}$とする。$\Delta$ 内の全ての文書 d_i について、文書内に含まれる単語を抽出する。この単語抽出には、通常、文書の分類や検索のために有効となる単語（有効語）を選定して抽出する。すなわち、助詞や句読点など、文書の内容にあまり関係なく出現する語は分類や検索には意味をなさないため除外する。通常は、有効語として名詞や動詞の語幹の中から全文書中での頻

度を考慮して選定される。

全文書から抽出された有効語の集合を $\sum = \{w_1, w_2, \cdots, w_{|W|}\}$ とすれば、各文書の特徴ベクトルを各特長語の出現頻度に応じて、W 次元ベクトルで表現することができる。すなわち、文書集合 Δ から得られる全有効語によってベクトル空間が構成され、文書 d_i を次式で表現することができる。

$$d_i = (v^i{}_1, v^i{}_2, \cdots, v^i{}_W)^T \quad (5)$$

ただし、T は転置を表す。ここで、この文書ベクトルを集めた行列

$$A = (d_1, d_2, \cdots, d_D)^T \quad (6)$$

を文書―単語行列（document word matrix）と呼ぶ。本稿では触れないが、この特徴ベクトルに含まれるノイズを除去し、意味のある空間において分析を行うための方法として Latent Semantic Indexing という方法が提案されており、この方法ではこの文書―単語行列を特異値分解することによって得られる主成分空間上でベクトル空間を構成する。

2.2.2 TF･IDF

前節において、各文書 d_i のベクトル表現を与えたが、各要素の値を如何に決めるかという問題が残っている。最も簡単な方法として、各単語の出現頻度とする方法があるが、しばしば検索や分類性能が、多くの文書に出現する単語に大きく影響されてしまうという問題がある。通常、全ての文書にまんべんなく表れる単語は、文書の特徴を規定するためにはあまり意味がない。むしろ、少数の文書において集中的に表れる単語は分類や検索に有効である。そこで、各単語の出現頻度だけでなく、全文章中でその単語が現れる割合を考慮した特長量の算出が必要であり、そのための方法が TF･IDF 法である［4］。TF は Term Frequency の略であり、文字通り単語の出現頻度を表す。一方、IDF は Inverse Document Frequency の略であり、全文書中の単語の出現割合の減少関数を表す。文書 d_i における単語の出現頻度 TF を $tf(d_i, w_j)$ とおく。IDF は単語 w_j を含む文書の数を $df(w_j)$ とすると、

$$idf(w_j) = \log \frac{D}{df(w_j)} \quad (7)$$

のような関数で定義される。このとき、文書 d_i における単語 w_j の特徴量 $v^i{}_j$ は、

$$v^i{}_j = tf(d_i, w_j) \cdot idf(w_j) \quad (8)$$

で与えられる。最近では、TF･IDF の情報理論的な解釈についても研究が行われている［14］。

2.2.3 文書間の類似度判定

各文書の特徴量がベクトル表現されれば、文書 d_i と文書 d_k の類似度（内容的近さ）は、これらの距離を使って測ることができる。この距離には、ユークリッド距離や内積を用いることも可能であるが、これらの距離は原点付近の 2 点が近いものであると判定する。ほとんどの単語の特徴量が 0 に近い文書同士は内容的に類似しているとは言えないが、ユークリッド距離や内積によれば類似していると判定してしまう。そこで、文書ベクトル d_i と文書ベクトル d_k の余弦をとって類似度とする方法が一般的である。

$$sim(d_i, d_k) = \frac{d_i{}^T d_k}{|d_i||d_k|} \quad (9)$$

文書検索の問題においては、検索語を特徴ベクトル（クエリベクトルと呼ぶ）で表現し、このクエリベクトルと類似度の高い文書を検索結果として提示する。類似度の高いものからリスト表示することにより、検索結果のランキング機能も有していることになる。文書の類似性評価については、様々な問題に対して、問題の特性を考慮した方法が研究されている［18］。

3 経営工学的諸問題への適用例

本章では、自然言語データのソースを用いて解析を行う問題として、企業経営に関わる諸問題への適用可能性について議論する。ここでは、マーケティ

ングリサーチにおける顧客ロイヤリティ構造化の問題［19］、自由記述アンケートデータからの重要意見のランキング問題［20］、ビジネスモデル成功事例検索の問題［21］を取り上げ、先に述べた自然言語処理やテキストマイニングの技法の適用可能性について議論を行う。

3.1 自由記述文書データに基づく顧客ロイヤリティの構造分析［19］

本節では、インターネットの評判サイトに掲載された自由記述コメントから、顧客ロイヤルティの構造を視覚化するための構造図の作成方法について紹介する。ここでは、形態素解析で単語分割した抽出した後、人手によって単語間の類似性による構造化を行い、この構造を用いて各ユーザコメントを分類して特長を把握するという方法をとる。本来は、形態素解析の結果から自動的に、その構造を構築する方法が望まれるが、本研究ではそのような自動分析技術構築のための基礎研究として行った。

3.1.1 分析の手順

分析手順:

［手順1］評判サイト（後述）からの自由記述意見の抽出，収集

［手順2］抽出した自由記述意見に対する形態素解析による単語分割及び不要語の除去

［手順3］分割した単語を類似度でKJ法により分類

［手順4］分類した単語群をまとめ，先行研究［22］で得られている知見を参考として顧客ロイヤルティ構造図作成

［手順5］顧客ロイヤルティ構造図に対して，自由記述文章の内容がどの程度含まれるかを集計

［手順6］数量化Ⅲ類を用いた成分の抽出と特徴の把握

3.1.2 顧客ロイヤリティ構造図

評判サイトのユーザコメント群から得られた単語集合を仕分けし、構造化した結果、表1に示すような構造図が得られた。この構造図の構築にあたっては、従来の顧客ロイヤリティやブランドロイヤリティに関する研究で得られている知見［21］を盛り込んでいる。実際には、それぞれのカテゴリ内に仕分けされた単語が分類整理されている。

表1　顧客ロイヤリティ構造図

顧客ロイヤルティ向上	便益上昇	品質全般	機能	機能性
				特性
			デザイン	色
				形
				種類
			独自性	ブランド価値
				顧客へのサービス
		個人特性	意思決定	購入意思
				比較
				願望達成
				批評
			製品へのイメージ	イメージ
			使用用途	使用場所
				使用方法
			利便性	便利さ
	コスト低下	費用		外部費用
				価格
				コストパフォーマンス
	信頼性	信頼性向上		製品評価低下
				イメージとの差異
				（プラス）
		信頼性低下		製品評価低下
				イメージとの差異
				（マイナス）

3.1.3 顧客ロイヤリティ構造図を用いたユーザコメントの関係性評価

表1で得られたロイヤリティ構造図の最も細分化された右側の各項目に属する単語が含まれるか否かを0と1でデータ化した各ユーザコメントを数量化Ⅲ類で分析した結果を以下に示す。図1は分析の結果得られた主成分1と2に対して、各ユーザコメントの主成分得点の散布図を描いたものである。この結果、一般に知られている商品特性、とくに最寄品、買回品、専門品によって特徴に差が出ることがあき

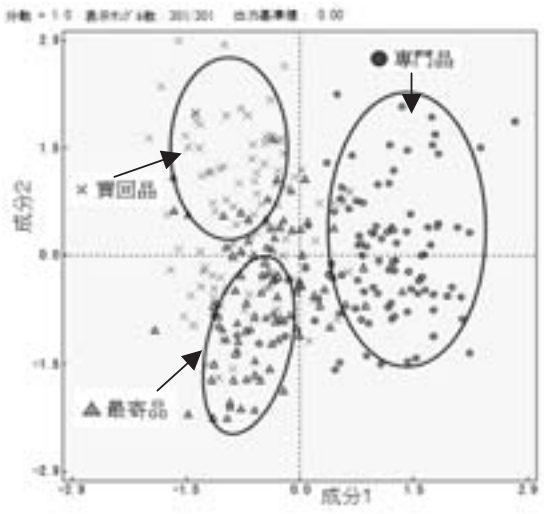

図1　数量化Ⅲ類結果（成分1,2）

らかとなった。

このように、インターネットの評判サイトから得られたユーザレビューの自由記述文書データをもとに、各ユーザコメントの内容が商品特性などによって特徴付けられることが明らかとなった。今後は、ユーザコメントから切り出した特徴語を、顧客ロイヤリティ構造図などのメタ情報によって自動仕分けするなど、自動分析のシステムを構築できると考えられる。

3.2　自由記述アンケートデータからの重要意見ランキング手法［20］

本節では、自由記述式のアンケートデータから、重要な意見をランキングして提示する方法について述べる。自由意見は、例えば企業の問い合わせ先メールアドレスや Web 記入欄などに大量に投稿され、それらの自由意見を総合して判断を下すための方法が有効である。

3.2.1　モデル化

一人の被験者の意見を文書ベクトル d_i とみなし、他の意見との類似度を式（9）によって測ることを考える。このとき、全体意見を代表する重要な意見は、他の全ての意見からの類似度が平均的に高いものであると考えられる。そこで、重要度指標として、

$$imp(d_i) = \frac{1}{D-1}\sum_{j\neq i} sim(d_i, d_j) \qquad (10)$$

で測ることができる。この重要度の大きいものからランキングして表示すればよい［23］。しかし、このまま順位付けをすると、上位の抽出意見としてお互いに似たような意見ばかりが抽出されてしまうという傾向がある［24］。また、アンケートの分析では、類似する意見が全体のどれくらいを占めているのかという情報が重要であることが多い。そこで、次のように、重要度上位の意見を取り出すごとに、これと類似する意見数を数え、リストから外して再度ランキング計算をするという処理により、得られる重要意見リストの網羅性を高め、類似意見数という情報を付加する方法［20］を考える。

【重要意見抽出アルゴリズム】

Step 1　要約したい文書集合の形態素解析を行って単語を抽出し，各文書を単語ベクトルで表す。

Step 2　文書集合中の各文書間の類似度を式（9）より計算する。

Step 3　全ての文書について，文書の重要度を式（10）より計算する。

Step 4　未抽出文のうち重要度の最も大きい 1 文書（意見）を抽出する。

Step 5　Step 4 で抽出した文書との類似度がパラメータ γ 以上の文書を抽出候補リストから除外し，それを類似文書として扱う．類似文書数を数えて抽出文書の付随情報として付加する。

Step 6　残り全ての文書が除外された場合は終了。

Step 7　Step 4 に戻り，残った文書の中から再び重要度が最も高いものを抽出する。

3.2.2　適用実験例

ここでは、某大学に導入された履修登録管理システムの利用感想について、ユーザである大学生に質

1．簡単にできて、良かった。類似意見数419件
2．時間も早く満足のいく結果だった。類似意見数25件
3．大学まで登録に行くのが面倒だし、パソコンの方が選択ミスも気付きやすいと思うからよい。類似意見数38件
4．これまでの実験よりも、スムーズに行うことができた。類似意見数４件
5．12：30ちょうどにアクセスし始めたのでスムーズに登録できました。実際の科目登録でもこのくらいの軽さで出来るなら、ぜひＷｅｂ登録をしたいです。類似意見数21件
6．前回、前々回に比べて一連の作業がとてもスムーズであったと思います。類似意見数14件
7．特にトラブルなく、スムーズに登録できたのでよかったと思う。類似意見数０件
8．マークシートより格段に楽　類似意見数２件
9．3月に行った科目登録と同じようにスムーズに行えたと思います。　類似意見数８件
10. 大変良いシステムだと思います！来年からこの方法で登録できたら本当に夢のようです。便利で楽しくて明確で面倒な手間がなくて文句なしです。類似意見数３件
11. 各学部ごと、曜日ごとなどに分類されていて、科目検索が非常に分かりやすかったです。類似意見数８件
12. 大学まで足を運ばずにすむから。類似意見数２件
13. レスポンスが早かったので、特に問題は感じなかった。類似意見数１件
14. 今回は正常に作動したが、今までの経緯から使用したいとは思わない。類似意見数７件
15. 実用化されればすごく便利だと思う。類似意見数０件
16. 保健体育科目の場所が多少分かり難かったです。類似意見数４件
17. おもってたよりスムーズにできた。登録も難なくできたが、実験実証の手順の説明がわかりにくくて時間がかかった。ＰＤＦをひとつにまとめてほしい。類似意見数４件
18. 時間と交通費の節約になる。24時間申請可能なシステムならば時間がとても有効に使えるようになる。類似意見数０件

図２　重要意見の抽出結果

問した自由記述式アンケートの分析結果について概要を示す（図２）。サンプル数は605である。

このように、全体で600件以上の自由記述意見を集約し、全体意見として多いと考えられる重要なコメントをランキングして表示することが可能である。この方法は本質的には情報検索で使われているベーシックな方法と同じであり、この方法であれば完全自動化が可能である。しかしながら、アンケートデータの分析においては、”意見間の類似性”をどのように定義するかについてさらなる検討を加えることにより、アンケート分析手法独自のさらに有効性の高い方法が構築できると考えられる。とくに、アンケートデータでは、少数意見の中に改善のアイデアが含まれるケースもあり、この点をどのように分析に取り込むかが課題である。

3.3　中小企業経営のための戦略事例検索モデル［21］

本節では、形態素解析などの自然言語処理手法を用いず、人手によって過去の事例を分類し、事例‐特徴量空間を構成することにより、中小企業の経営者が自社の現状に最も参考となる他社の成功事例を検索するモデルについて述べる。企業戦略の過去の成功事例は、いくつかのビジネス雑誌などの情報誌に特集で掲載されており、かなりの分量となっている。例えば、日経ビジネス誌では“小さなトップ企業”という特集で、毎週個性的な市場で高シェアを達成している中小企業の事例が紹介されている。このような事例は、今後も電子的にテキストファイルとして蓄積されていくことが期待され、企業戦略立案のための有用なデータベースとして活用できる。

この応用例では、経営戦略事例ベース構築のための基礎研究として、自然言語処理の技法を駆使せず

に、企業戦略論の方法論を駆使してベクトル空間モデルを構築した。すなわち、企業事例のアナロジーという概念［25］を、情報検索技術に類似する考え方によって導入する。これにより、中小企業の経営者が、自社の状況と最も類似する事例を抽出するための方法を示す。

3.1.1　方法論

［Step 1 ］戦略事例の収集

戦略事例として『日経ビジネス』の記事、”小さなトップ企業”［26］に掲載されている計44社の事例を収集した。この記事は、国内あるいは世界でトップシェアを持つ中小企業だけを紹介しており、企業がどのようにトップシェアを握るようになってきたかを時間経過とともに説明しているため、戦略立案のための参考事例としては最適である。

［Step 2 ］戦略事例の分析

戦略事例にアナロジーという概念を導入するにあたり、まず、多くの成功している事例に共通点があるということを仮定する。そして、共通点を探すために、戦略事例を、企業戦略立案のための重要かつ有効な手法である SWOT 分析の視点を取り入れ分析する［27］。具体的には、戦略事例に対し、トップシェアを持つに至るまでの S（強み）、W（弱み）、O（機会）、T（脅威）、イベントなどに着目し、それぞれを説明している部分を抜き出す。

［Step 3 ］事例の特徴を説明する要素項目の整理

Step 2 で抽出された事例の分析結果である、S、W、O、T などそれぞれの要素を分類整理する。全ての事例の要素を列挙し、似たような単語を持つという基準で分類する。そして、各分類群に所属する要素がすべて同じ意味であるかを確認し、違う意味を持つ要素はその群から排除する。

［Step 4 ］アナロジー評価指標の構造化

アナロジー評価指標をより分かりやすいものとするため構造化する必要がある。そこで、S･W についてはマッキンゼーの 7 S や、マイケル　E. ポーターのバリューチェーン［27］を利用しアナロジー評価指標の上部に当たる中分類項目を作成し、指標をその各項目に当てはめる（表 2 の左から2, 3、 4 列目の部分）。O･T についてはマイケル　E. ポーターの 5 つの競争要因や、PEST 分析［27］を利用し、中分類項目を作成し指標を各項目に当てはめる。

［Step 5 ］各事例データの登録

Step 4 で構築された各アナロジー評価指標に対し、各事例が該当する箇所を 1 、該当しない箇所を 0 として事例を登録し、各戦略事例をベクトル空間モデルで表現する。

表 2　アナロジー評価モデル

	中分類項目		アナロジー評価指標	自社の状況	ケージーエス	飯塚製作所	…	プリンス電機
					1	2	…	44
強み	マッキンゼーの7S	戦略	高付加価値戦略	0	0	1	…	0
			差別化戦略	0	0	0	…	0
			廉価品と高級品の両方を手がける	1	0	1	…	0
			多品種少量生産	0	0	0	…	0
			市場特化	0	0	0	…	0
			ブランド化	0	0	0	…	0
			新規事業参入	0	0	0	…	0
			価格競争力	0	0	0	…	0
			多くの新製品を投入	1	1	0	…	0
			OEM を手がける	0	0	0	…	0
			技術のブラックボックス化	0	0	0	…	0
			海外進出	0	0	0	…	0
		スタッフ	特殊な人材を持つ	0	1	0	…	0
			意識の高い社員を持つ	0	0	0	…	0
			社員が経営意識を持つ	1	0	1	…	0
			多くの技術者を持つ	1	0	0	…	0
		システム	成果主義、自己評価制度の導入	1	0	1	…	0
	バリューチェーン	製造	新商品、新技術の開発	0	0	0	…	1
			幅広い製品ラインナップ	0	0	1	…	0
		出荷物流	スピード納品	0	0	0	…	0
		販売マーケティング	高い提案力	0	0	1	…	0
			直販体制	0	0	1	…	0
			需要動向の先読み	0	0	0	…	0
		サービス	アフターサービスの充実	1	0	0	…	0
			顧客要望、クレームの吸い上げ	0	0	0	…	0
			顧客要望にきめ細かく対応	1	0	0	…	1
		研究開発	研究に注力できる環境を持つ	0	0	0	…	0
			技術開発に注力	1	0	1	…	0
			製造装置の自社開発	0	0	0	…	0
			高い生産能力	0	0	0	…	0
			生産の自動化	0	0	0	…	0
			最新技術の導入	1	0	0	…	0
			共同開発	0	0	0	…	0
			経験・勘の機械化	0	0	0	…	0
			高い技術力を持つ	0	0	0	…	0
			独自ノウハウを持つ	0	0	0	…	1
			特許を多数取得	1	0	0	…	0
			商品の高性能化	0	0	1	…	1
			IT 技術によるシステム構築	0	0	0	…	0
		調達活動	中国に生産拠点を持つ	0	0	0	…	0
	その他		環境に配慮	0	0	0	…	0
				類似度結果	1	4	…	0

以上のステップによって、各戦略事例を構造化して保存することにより、中小企業の経営者が、自社の強み、弱み、機会、脅威の分析から、参考とすべき成功事例をランキングして表示することが可能となる。

3.1.2 分析結果と考察

実際に、連載記事”小さなトップ企業”［26］に掲載されている計44社の事例について分析した結果の一部を表2に示す。結果として、いくつかのサンプルケースについて、有用な過去の戦略事例が提示できることが明らかとなった。しかしながら、この戦略事例検索モデルは、検索部分に関しては情報検索モデルと同等であるが、現在では戦略事例のテキストデータを自動で構造化する方法を構築するまでには至っていない。情報検索と同じ評価基準で、戦略事例の類似性を定義できれば、情報検索の方法論を適用可能であるが、実際には戦略事例の類似性には経営学の知見を十分加味する必要がある。経営学の知見を集約した独自のコーパスを構築するなど、今後の成果が期待される。

4 まとめと今後の展望

本稿では、いくつかの事例を通じて、自然言語処理の技法を経営学、経営工学の分野で活用する方法について考察した。取り上げた事例の他、テキスト分類の応用も極めて実用性が高い。テキスト分類ではナイーブベイズ法［28］やサポートベクターマシン（SVM）［29］といった様々な学習理論による分類法が研究されている。どのような分類器が有効であるかは問題の特性にも依存するが、今後さらに様々な技術が研究されていくであろう［30］－［32］。企業経営の分野では、独自のナレッジが蓄積されており、テキスト形式で蓄えられているデータを分析するためにも、経営学や経営工学の知見を加味した分析の方法論を構築する必要がある。現在では、研究が活発に行われ、いくつかの製品が登場し始めている段階であるが［33］、今後も企業経営に関わる諸問題に対して実用的な方法論の整備が順次なされていくものと考えられる。

参考文献

［1］伊藤哲郎．情報検索．昭晃堂，1986．

［2］G. Salton and M.J.McGill.Introduction to Modern Information Retrieval. McGraw-Hill. 1983.

［3］河野浩之，川原　稔．Web 検索におけるテキストマイニング．人工知能学会誌　Vo.16, No.2, pp.212-218, 2001.

［4］金明哲，村上征勝，永田昌明，大津起夫，山西健司．言語と心理の統計．岩波書店，2003．

［5］石岡恒憲．記述式テストにおける自動採点システムの最新動向．行動計量学，Vol.31, No.2, pp.67-87, 2004.

［6］市村由美，長谷川隆明，渡部勇，佐藤光弘．テキストマイニング－事例紹介．人工知能学会誌　Vol.16, No.2, pp.192-200, 2001.

［7］那須川哲哉，河野浩之，有村博紀．テキストマイニング基盤技術．人工知能学会誌　Vol.16, No.2, pp.201-211, 2001.

［8］山西健司．データ・テキストマイニングの最新動向－外れ値検出と評判分析を例に．応用数理，Vol.12, No.4, pp.7-22, 2002.

［9］日本語形態素解析システム「茶筌」
http：//chasen.aist-nara.ac.jp/hiki/ChaSen/

［10］北　研二．確率的言語モデル．東京大学出版会，1999．

［11］繁枡算男．ベイズ統計入門．東京大学出版会，1985．

［12］平澤茂一．情報理論．培風館，1997．

［13］植松友彦．現代シャノン理論．培風館，1998．

［14］A.Aizawa. The Feature Quality： An Information Theoretic Perspective of Tfidf-like Measure, 23rd Annual International ACM SIGIR Conference, SIGIR 2000, pp.104-111, 2000.

［15］Masayuki Goto, Toshiyasu Matsushima, and Shigeichi Hirasawa. A source model with probability distribution over word set and recurrence time theorem. IEICE Trans. on Fundamentals, Vol.E86-A, pp.2517-2525, 2003.

［16］石田崇，後藤正幸，松嶋敏泰，平澤茂一．語頭条件を満たさない単語集合をもつ Word-Valued　Source の性質について．電子情報通信学会技術研究報告 IT 2003-5, pp.23-28, 2003.

［17］大谷紀子．情報検索におけるベクトル空間モデルの応用．武蔵工業大学環境情報学部紀要，No.5, pp.99-109, 2004.
［18］深谷　他．単語の頻度統計を用いた文章の類似性の定量化．電子情報通信学会論文誌　D-II, Vol.J87-D-II, No.2, pp.661-672, 2004.
［19］三川健太，高橋勉，後藤正幸．継続購買につながる顧客ロイヤリティの構造分析手法に関する一考察．日本経営工学会春季研究大会予稿集，pp.44-45，2006.
［20］渡辺智幸，後藤正幸，石田　崇，平澤茂一．情報検索技術を用いたアンケートデータの分析手法に関する研究．日本経営工学会春季研究大会予稿集，pp.126-127，2006.
［21］斉藤靖夫．中小企業における戦略事例のアナロジー評価手法の提案．武蔵工業大学環境情報学部卒業論文，2005.
［22］加藤雄一郎，川村法征．顧客接点の役割～ブランド評価メカニズムと顧客接点の関係～，　経営情報学会2005年秋季全国研究発表会予稿集，pp.144-147，2005.
［23］亀田雅之．擬似キーワードによる重要キーワードと重要文の抽出，言語処理学会第2回年次大会発表論文集，pp.97-100，1996.
［24］伊藤潤，石田崇，後藤正幸，平澤茂一．文間の単語共起類似度を用いた重要文抽出法．FIT 論文集，pp.83-84，　2002.
［25］ジョバンニ・ガベッティ，ジョン W．リブキン．アナロジカル・シンキング．ハーバード・ビジネス・レビュー，pp.48-61，2005年7月号.
［26］小さなトップ企業，日経ビジネス，2004年2月2日～2005年2月21日.
［27］大石達也：　最新『経営戦略』とケース分析，秀和システム，2004年3月.
［28］ H. Langseth, T. D. Nielsen. Classification using Hierarchical Naive Bayes models. Mach Learn, Vol.63, pp.135-159, Springer, 2006.
［29］津田宏治．サポートベクターマシンとは何か．電子情報通信学会誌，Vol.83, No.6, pp.460-466, 2000.
［30］平澤　茂一，石田　崇，足立　鉱史，後藤　正幸，酒井　哲也．文書分類技法とそのアンケート分析への応用．経営情報学会　2005年春季全国研究発表大会，pp.54-57，2005.
［31］酒井　哲也，石田　崇，後藤　正幸，平澤　茂一．自然言語表現に基づく学生アンケート分析システム．第3回情報科学技術フォーラム講演論文集　pp.325-328, 2004.
［32］ Masayuki Goto, Takashi Ishida, Shigeichi Hirasawa. Representation method for a set of documents from the viewpoint of Bayesian statistics. 2003 IEEE International Conference on System, Man and Cybernetics, pp.4637-4642, 2003.
［33］日刊工業新聞，記事“NEC　必要な経営情報可視化して活用するソリューションを体系化”，2006年6月9日第9面，2006.