流行ことば・流行コンセプト予測手法

―単語間距離による流行コンセプト評価法の提案―

池田　定博・大橋　正和・金田　重郎

あらまし

社会の流行がどのような方向に向かうのかを予測することは、商品販売、CM制作、ドラマ制作等において、極めて重要である。しかし、従来、どのことばが「あたる」かは、人間の直感や、利用者からのヒアリングにより直感的に判断されてきた。このような、感覚的な方法は、数値的な裏付けが無いため、説得力に乏しく、また、誤りも入りやすいと思われる。そこで、この問題を解決するために、本論文では、現代用語辞書と自然言語処理の意味ベクトル法による距離計算を用いた、流行予測手法を提案する。具体的には、コピーライターは、複数の流行語、流行コンセプトの候補を作成する。そして、候補毎に作成された説明文と、現代用語辞書との各見出し語との距離計算を行い、距離が小さいものを意味的に近いとする。

実際に、自由国民社発行の『現代用語の基礎知識』を利用してプロトタイプシステムを構築した。『現代用語の基礎知識』が持つ、約10種類の分野区分について、相関係数、金融工学的予測手法により流行分野の推定を行った。その結果、「政治」「経済」等の分野では、統計学的に有意な検定結果を得た。この結果は、従来から言われれている「流行の背景には社会現象がある」との見方を数値的に裏付ける。

１．はじめに

「24時間戦エマスカ」。これは、バブル華やかなりし頃、流行したCMのキャッチフレーズである。良く知られているように、これは、三共の「リゲイン」というブランドのCMである。それから10年数後、同じブランドの商品がヒットさせたCMのキーワードは「癒し」となっている。坂本龍一のCM音楽と共にその年の流行語にもなった[4]。企業の担当者は、このようなヒットするキャッチフレーズやコンセプト作りに力を費やしている。そして、そのコンセプトも時代につれて変化する。

具体的なフレーズ作りは人間のみがなしえる創作行為である。多くのコンセプトやフレーズの中から、政策の柱となる候補を絞り込むことは決して容易ではない。現実には、段階を経るいくつかの部署の決定者によって経験的・曖昧な選択基準で決定される[27]。その結果「ありふれたもの」や成功事例の「二番煎じ」に終わってしまうことが多い。

本論文では、これら現状の問題を解決するための「流行を予測する」システムを提案する。そして、その手法を「流行ことば予測」と呼ぶ。「流行ことば予測」では、過去・現在・将来における流行語の背景となる社会的要因を「ことば」として表現する。そして、今後流行の可能性がある新しい「ことば」と、これら社会的要因との距離を計算し、「近い」と算出された「ことば」から、流行の可能性の高いコンセプト・キーワードを開発する。

具体的には、まず作成した流行語候補の各々に説明文を作成する。一方、数値化のための比較対象物として、複数の「見出し語」を持つ辞書（テキストベース）を準備する。そして、作成した流行語候補と「見出し語」との距離を、自然言語処理技術の「ベクトル空間法」で求める。最終的に近いとされた辞書中の「見出し語」と当該流行語候補への距離に基づいて、流行語候補がどのよ

うな要因により人々にアピールするかを分析する。この際、比較対象辞書として、一般の国語辞典は必ずしも十分ではない。流行を支配するような時代背景を表す単語が、見出し語として、必ずしも豊富ではないからである。そこで、提案手法では、辞書として、自由国民社発行の『現代用語の基礎知識』を利用している。

以下、第2章では、流行語と社会的背景との関係について触れる。第3章では、現代用語辞書を用いた、「流行ことば予測」システムを提案し、そのプロトタイプ実験の結果を述べる。第4章は、得られた結果を統計的に検証する。第5章では、「流行ことば予測」への金融工学的手法の適用可能性について論じる。第6章において若干の議論を行い、第7章で本論文の結論を述べる。

2．流行語とマーケティング

2.1 流行と社会的背景

「ショムニ」ということばがある。1998年ヒットしたフジテレビ系ドラマのタイトルで、98年度の流行語大賞にランクインした当時の新語である。このドラマは。ある商社を舞台に。庶務二課（通称ショムニ）とは名ばかりの、備品管理室に追いやられたOLたちが痛快に活躍する、漫画原作のストーリーである。女優・江角マキコ扮するキャラクターは共感を呼び、高視聴率を得た[3]。

ここで、このドラマが流行した理由を考えてみたい。むろん、江角マキコら人気女優や人気漫画をもとに制作した点は無視できない。しかし、それだけでは、高視聴率の要因としては不十分である。ヒットの理由を思いつくままに列挙すると、以下の**要因があるように思われる。**

1．長期にわたる不況で、「いつ閑職に追いやられるか分からない」不安を抱えて日を過ごし、一方で、幹部の無策に不満を抱きながらも声にはできない、多くの「悩める声なき社員」達からの共感。
2．主人公達は、ことばはキツイが、一人一人が個性を持ち、周囲への配慮や思いやりがある。この「古き良き男社会」へのノスタルジー。
3．男女雇用機会均等法の施行による、女性の職場への進出。

いずれの理由ももっともらしい。この例からも、流行の背後に社会的・時代的な要因が隠されていることは疑い得ない。流行語が時代時代の社会・経済的な背景の影響を受けていることは、特に、1990年代に入ってから、マスコミ論の世界では、一つの通説になっている。社会的・経済的背景と流行語に強い相関が推定されることは本論文の前提である。

しかし、単に、感覚的に思いつくままに、社会的要因を想起するだけでは、流行語と社会・経済的背景との関連を証明したことにはならない。上記強調部分のように、「要因があるように思われる」に過ぎない。必要なことは、何らかの科学的・定量的な手法によって、流行語と社会・経済的背景との関係を証明することである。そして、もし、そのような関係が定量的に証明できれば、将来の流行語を「予測する」道が開ける可能性がある。

以上の分析から、流行を生むためのコンセプト作り、コピー作成等において、コンピュータに求められる機能として、本論文では、以下の2点を設定した。

1．来年の「流行語」を支える社会的要因が何であるかをある程度定量的に予測したい。この場合、社会的要因を表わすものは「ことば」であるが、種々の粒度（抽象度）が考えられる。例えば、「経済」と言った大きな概念かもしれないし、「サプライチェーン」といった、細かいレベルのことばの可能性もある。
2．コピーライター等により創作された新しいコンセプトやことばが、どんな社会的要因を人々の心の中に想起させるかを定量的に推定したい。そのためには、上記の社会的要因を表わすことばと、創生された新語との間の距離を測定したい。

2.2 現代語研究と言語によるマーケティング支援

システムの提案の前段として、既存の研究に

ついていくつか触れておく。「流行ことば予測」はコンテンポラリに社会的要因を示すことばとして現代語を扱う研究である。そのことから、まず現在、現代語の周りで行われている研究を言語学、自然言語処理、そして本研究のようにマーケティングの支援ツールとして行われているものを検証する。

2.2.1 現代語研究と流行語

まず、現代語の研究に関して、代表的な文献がある。一つは梅花大学の米川明彦のもので、中でも著書『現代若者ことば考』・『新語と流行語』・『若者語を科学する』は現代の若者の会話を生きたことばとして集録し、整理を行ったものである[5]。

研究の方法は主に調査に基づくもので、実際の学生のよく使うことばを集めている。これは会話の収集とともにアンケート調査を行い、使用頻度を 5 段階で評価してもらい傾向を出す一方、そのことば群を生成の過程を含め的確に分類している。また、明治時代からの学生語とも比較して、若者語の生まれる背景を心理的要因・社会的要因・歴史的要因の観点から見ている[6]。特に社会的要因という観点は本研究を進めていく中で比較検証できるものである。

また、成城大短期大学部の小林千草の研究で、源氏物語から現代の若者ことばまでの歴史的変化をたどっているものがある[7]。現代語に関しては、こちらも学生に対する調査に基づいている。ただ、歴史的関わりとの中で、現代日本語の乱れとしてよく指摘されている「ら抜きことば」を源氏ことばや江戸ことばとして特徴づけられたように、現代のことばの「ゆれ」として特徴づけている。

どちらの研究もそうであるが、現代語をテーマにする際、どうしても若者語が脚光を浴びる。若者のことばが、変化しやすく社会的要因の影響を受けやすいからである。しかし「流行ことば予測」では逆に若者ことばだけでなく、あらゆるジャンルからの現代語を見る必要がある。加えて調査によって進められているこれまでの研究とは別に、ことばの数値化による一部特定の人が対象でないデータによって流行の変化を見る。

もうひとつのアプローチは、言語の統計である[15]。語彙の調査によるもので、特に国立国語研究所が行っているものである。同研究所の『現代新聞用語の一例』[8]、『婦人雑誌の用語－現代語の語彙調査 -(4) 』[9]、『総合雑誌の用語－現代語の語彙調査前編後編』[10]、『現代雑誌九十種の用語用事』[11]、『テレビ放送の語彙調査』[12] は本論文の研究にも参考となる研究である。しかし、これらは、テレビ放送の語彙調査を除くと、いずれも古い調査である。これら研究の収集の基準をヒントに新しいデータを独自に収集する必要性が感じられる。

2.2.2 言語を用いたマーケティング

人文科学分野で、言語の収集研究が進む一方、コンピュータを使った科学的な言語の分析技術、「自然言語処理技術」も進化している。コンピュータによる言語の解析は特に、1950 年前後から米国で発展してきた。なかでも 1957 年チョムスキー（N.Chomsky ）の構文構造理論は現在も応用されている形態素解析の手法に用いられ、その後の研究にも大きな影響を与えている。形態素解析は後述する提案手法でも重要となる技術である。

その後、研究ではオートマンと呼ばれる手法による情報検索研究、あるいは人工知能の一環としての研究を経て、現在ではより人間に近い言語理解の研究が進んでいる。コンピュータによる人にとっての自然な文章の生成も急速に発展している[17, 18] 。

自然言語処理は、近年、より一般化し、コンピュータ能力の向上や、デジタル化された辞書の存在により、容易に適用できるものになった。特に、マーケティング方面への応用は注目されている。日本でも KJ 法やインスピレーションといったソフトウェアがすでに市販されている。これらのツールも本研究と同じように発想支援を目的としたものであるが、これはあくまでもカードやブレーンストーミングによって行われる作業をパソコンによって整理していくものである。自然言語処理機能を有している訳ではない。本論文の予測機能とは異なるが、安価で誰でも使えるという点で先駆的である。

一方、プロがマーケティングに使用するツールとしては、電通と富士通が開発したＤＥ-

FACTO（DENTSU FLEXBLE ANALYSER OF CONTEX AND OPINION）がある[21]。このシステムは開発後関連会社の電通リサーチに置かれマーケティングの専用ツールとして調査・分析に活躍している。このようなツールは博報堂やアサツー・ディー・ケイなど広告代理店がシステムメーカーと共同で開発しているケースが多い[1]。

DE-FACTOでは、ある2種類の単語が、一つの文の中に共起する確率が高いほど、2単語の意味的に近いとする考え方に基づいている。そして、この距離に基づいて、画面上に単語相互の関係を図示して、クリエーターの発想・分析を支援する。すなわち、DE-FACTOは、流行語の候補からあたりそうなものを選択するためのツールではない。詳細は、本論文の主旨からずれるので、付録Aにおいて紹介する。

3．用語辞書を用いた流行語予測

本章では、「流行ことば予測手法」を提案する[29]。提案手法の「流行予測」は、**「あたることばは社会的背景を有する」**との仮説に基づく。

3.1 流行ことば予測システムの提案

本節では、流行ことば予測システムを提案する。提案手法では、流行語の背景となる社会的要因を「ことば」で表現し、流行語とそれらのことばとの距離を計測し、社会的要因と「近い」と判定されたものから流行コンセプトを開発する。

3.1.1 『現代用語の基礎知識』

社会的要因をことばとして表現するには、まず、ことばの辞書（テキストベース）が必要になる。著者らは、この辞書として、自由国民社発行の『現代用語の基礎知識』[2, 3]を用いた。『現代用語の基礎知識』は、その名の通り、現代用語である、新語・その年に注目されるようになった単語（見出し語）について解説した辞書である。『現代用語の基礎知識』は、広辞苑等の国語辞書にはない以下の特徴を持つ。

- 見出し語（1998年版で約9,000語）として、現代的なことばのみが選ばれている。
- 毎年（1984年開始）、この見出し語の中から、「流行語大賞」が選定されており、各年毎に流行したことばとして利用できる。
- 見出し語は、その上位概念として3レベルのグループを持ち、見出し語のレベルよりも、より高い概念レベルで、社会的背景を分析できる。具体的には、収録語を**【経済・経営】、【情報・産業】、【国際情勢】、【政治】、【社会、生活】、【健康・医療】、【科学・技術】、【風俗・流行】、【文化・芸術】、【今年の人物・今年の発言】**とジャンル別にまとめている。1年の流行語がどのジャンルから派生する傾向が多いかなど比較が行える。以下、この10個の概念を、**「最上位概念」**と呼ぶ。本提案手法では、この最上位概念を評価に多用している。

『現代用語の基礎知識』は、1948年（昭和23年）の創刊以来、現代語を取り上げた実績があり、1984年（昭和59年）から「日本新語・流行語大賞」の解説文も収録し各界の評価も高い。

3.1.2 システム基本構成

著者らの流行ことば予測システムの概要を図1に示す[29, 30]。システムは大きく分けて、2つの部分から構成される。一つは、『現代用語の基礎知識』を格納した、辞書M_i（$i = 1, 2, 3 \ldots n - 1$）である。『現代用語の基礎知識』の見出し語は、毎年、そのかなりの割合が入れ替えられており、各年毎の世相を反映したことばがこれらの辞書には格納されている。図1において添え字は、年度を表す。M_{n-1}が最新の『現代用語の基礎知識』を格納した辞書である。

一方、毎年の**流行語**R_i ($i = 1, 2, 3 \ldots n$) に格納しておく。各R_jは流行語の集合である。今回の実験では、この部分には、「流行語大賞」とその説明文を各年毎に格納してある。個数は、各年度毎に、およそ10個から20個である。目的が流行語予測なので、未来の流行語が選定・評価できな

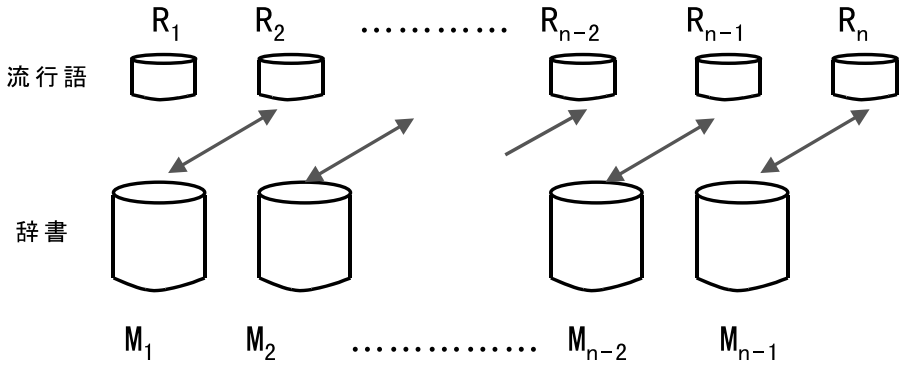

図1：システム構成概要

いと意味がない。そこで、来年の流行語の候補と思われることば（複数）を、流行語R_nとして、コピーライター自身が作成した流行語候補の説明文とともに格納する[1]。

流行語が影響を受けるとすると直近・直前の年度の社会的状況と考えられるので、R_i中の各流行語からの距離を、辞書M_{i-1}との間で計算する[2]。R_i中に含まれる各流行語は新語であるので、原則として、M_{i-1}には含まれていない[3]。

3.1.3 意味的距離の計算

【流行語候補に対する距離計算法】

本提案の手法では、基本的に、意味ベクトル法により単語の間（正確には、各年の流行語と、その前年の辞書の見出し語の間）の距離を計算している。まず最初に、意味ベクトル法[14]について説明する。

まず、辞書中の見出し語を$Midashi_i$とする。一般に、ある年の辞書には、この見出し語は1万程度ある。見出し語$Midashi_i$は説明文を持つが、その説明文が含む単語を$T_i0, T_{i,1}, T_{i,2} \ldots T_{i,ni-1}$とする。$n_i$は、見出し語$Midashi_i$に現れる単語の個数である。ただし、ここで、単語は重複を許すものとする。すなわち、異なる添え字を持つTが一致することがある。

これに対して流行語（候補）をr_jとする[4]。その説明文が含む単語を、$T_j0, T_{j,1}, T_{j,2} \ldots T_{j,nj-1}$とする。$n_j$は、流行語候補$r_j$に現れる単語の個数である。見出し語$Midashi_i$と流行語$r_j$で、一致する単語の個数を$Match_{i,j}$とする。意味ベクトル法では、$Match_{i,j}$の数が多いほど、より意味的に$Midashi_i$および$r_j$が近いとする[5]。

また、意味ベクトル法では、単一の説明文中に当該単語が出現した回数を考慮したり（TF:Term Frequencyと呼ぶ）、あるいは、どの説明文にでもよく出てくる単語か否かを評価（IDF:Inverted Document Frequencyを利用する。詳細は本論文では省略する。）している。今回の実験では、TFは

[1] 今回の実験では、実際には、流行語大賞として最新のものをR_nに登録して、そのひとつ前までの流行語大賞から、この最新の流行語大賞を評価する実験を行っている。

[2] ある年の流行語と、その前年の見出し語との間で距離を計算する。一般に、流行語はその年の新語である。

[3] 実際には、稀に含まれていることもあったが、そのような場合には、M_{i-1}から当該見出し語は削除した。

[4] 流行語候補集合であるR_iと区別するために、小文字のrを用いた。rはR_iに含まれる個々の流行語候補である。

[5] 実際には、(1) 単語の中には、多数の見出し語に出現する頻出単語と、極めて稀にしか出現しない単語がある。(2) 見出し語に対する説明文は、見出し語によって異なるため、どうしても、長い説明文を持つ見出し語の方が、距離計算で有利になりやすい、といった問題がある。そこで、上記(1)に対しては、「ある程度、稀にしか現れない単語」に注目して、例えば、助詞や接続詞のような意味のない頻出単語については、計算から除外している。単語の取り出しには、形態素解析を用いている。詳細は、付録Cを参照されたい。また、上記(2)に対しては、各見出し語のもつ単語数が等価的に一致するように、各単語に重みを与えて、正規化を行っている。

特に考慮していない[6]。また、IDF は適用しているが、極端に出現頻度が少ない単語は無視している[7]。

なお、システムの作成にあたっては、処理を高速化して、メモリ量を削減するために、見出し語、および、説明文中の単語は、すべて数値化を行っている。詳細は、付録 B, 付録 C を参照されたい。

上記の意味ベクトル法により、与えられた流行語候補r_jに対して、その前年の辞書から、近いものから順々に見出し語が得られる。また、この結果を更に加工して、以下の距離計算を行った。

【上位概念に対する距離計算法】

上記の方法により、流行語候補r_jについて、近いとされた見出し語の上位q個が得られる[8]。この時、当該年度の流行語候補 $r_j(j= 0, 1, 2 \ldots, n_j - 2, n_j - 1)$(ここで、$n_j$は当該年度の流行語候補の個数) のすべてから$q$個の見出し語を取り出して、その$q \times n_j$個の見出し語が持つ最上概念を調べる。『現代用語の基礎知識』が持つ「経済」「風俗」等の上位概念である。そして、この$q \times n_j$個の上位概念の中に占める「経済」等の各上位概念の割合を調べる。ここで、もし、「経済」の割合が最も高ければ、流行語候補r_jは、「経済」の影響を最も強く背後に持っているものと解釈する。

このような評価方法を取ったのは、今回の実験が、あくまで、社会的背景との関係を探るのが目的であったのと、流行語大賞に対する説明文が極端に短く、個々の流行語候補に近い見出し語を分析しても、あまり意味がないと判断したためである。しかしながら、コピーライターに直接的に興味があるのは、むしろ、流行語候補に近いとされた具体的な単語であるとも想定される。一つの流行語候補に近い見出し語を参考とするか、あるいは、最上位概念の割合を分析の対象とするかは、本システムの利用者が判断すべきことと考える。

3.2 予備的評価実験

3.2.1 データの準備

以上のシステム構成により、実験を行った。過去の流行語の集合 R_j, $(j = 1, 2, 3, , , n - 1)$には、自由国民社が発表している「流行語大賞」に選ばれた単語と、その『現代用語の基礎知識』に収録された当該流行語大賞の説明文を用いた。来年度の流行語候補の集合をR_nとする。R_nには、本来、コピーライター等の作成者自身が、自分で説明文を付加することとなる。ただし、今回の実験では、最新の流行語対象をこのR_nに格納して、より過去のデータからこの判明している流行語が予測できるかどうかを検証した。

一方、各年度に利用されていたことばの辞書 M j としては、過去数年分の辞書が必要である。しかし、このデータの収集は容易ではなかった。現在、『現代用語の基礎知識』CD-ROM から抽出されているのは、1998 年判と 1999 年版の『現代用語の基礎知識』のみである[9]。そこで、今回の実験では、以下のように同様の実験を 2 回繰り返した。

- 【実験 1】1998 年の流行語は、1998 年版の『現代用語の基礎知識』をすべての辞書M_jに導入。
- 【実験 2】1999 年の流行語は、1999 年版の『現代用語の基礎知識』をすべての辞書M_jに導入。

ここで、1998 年版の辞書を用いて、1998 年の流行語を予測していることに奇異な印象を持たれるかもしれない。しかし、1998 年版辞書は、1997 年秋に発行される[10]。1998 年版の CD-ROM

[6] これにはあまり深い意味はないが、異なるベクトル同士の共有単語数が少ないので、単語が一致したことを最重要視したかったためである。

[7] 出現頻度 3 以下は無視

[8] qは、任意に設定される取り出す見出し語の個数である。今回の実験では、20 としている。

[9] CD-ROM からのデータ取り出しには著作権の問題があり、すべて自由国民社の了解のもとに実験を進めている。しかし、CD{RM は特殊なフォーマットになっており、データの取り出しは容易ではなかった。CD-ROM から集出されたテキストは、SGML 化を行って実験に利用している。SGML の詳細については、付録 A を参照されたい。

[10] 正確には、秋に発行されるのは、紙印刷の『現代用語の基礎知識』である。データを取得した対象である 1998 年版 CD-ROM は、1998 年の 1 月頃に発行される。しかし、この CD-ROM 版の内容は、紙印刷のものに準拠しており、時期的にも、1998 年のことばを網羅する余裕はない。

の内容は、1997年版の現代語しか掲載されていない。流行語大賞は、毎年、夏が終わった頃に発表される。1998年版CD-ROMには、1998年の流行語大賞に属することばは掲載されていない。1998年の流行語大賞は、次年度版である1999年版に掲載される[11]。

上記の２つのケースは、97年初頭段階、および、98年初頭段階で、前年の現代語辞書から、流行語大賞を推定していることになる[12]。

3.2.2 実験１の結果（1998 年版 CD-ROM）

図２は、1984年から1997年までの間で、流行語大賞から近いとされた見出し語（各流行語大賞毎に20個とした。20×流行語大賞の数が見出し語個数である。）が持つ最上位概念の各最上位概念毎の割合である[13]。図２では、周期があることはある程度判明するが、明確な傾向は分からない。ただし、1984年から1997年までを考えると、大きな社会経済的な変化がある。バブルの崩壊である。実際、1989年くらいまで、どんどん増加してきた「経済・経営」が、この時期に大きく落ち込み、一方で、それまでは落ち込む一方であった「文化・芸術」が大きく跳ね上がっている。この一例からも、流行コンセプトが、社会的・経済的な背景と密接に関連していることがわかる。

【相関係数】

図３は、最上位概念の出現割合の年次推移について、各最上位概念毎の相関係数である。年次変化だけを見れば、バラバラのように見えるが、最上位10概念相互の相関係数を計算してみると、一部、強い相関を見せる。「経済経営」と「文化芸術」は逆の相関であり、「スポーツ趣味」と「政治」が意外にも正の相関である。後述するように、説明文が短いので、意味的距離計算をするにはかなり厳しい環境であることを考えると、意外に明確な相関である。

- 「文化・芸術」は「経済・経営」と逆の相関が強い。これは、前述のバブル前後の動きとも合わせ、一つの傾向と考えられる。また、これだけではなく、「風俗・流行」や「文化・芸術」のような概念は、「国際情勢」「経済・経営」といった概念と、逆の相関を持つ。

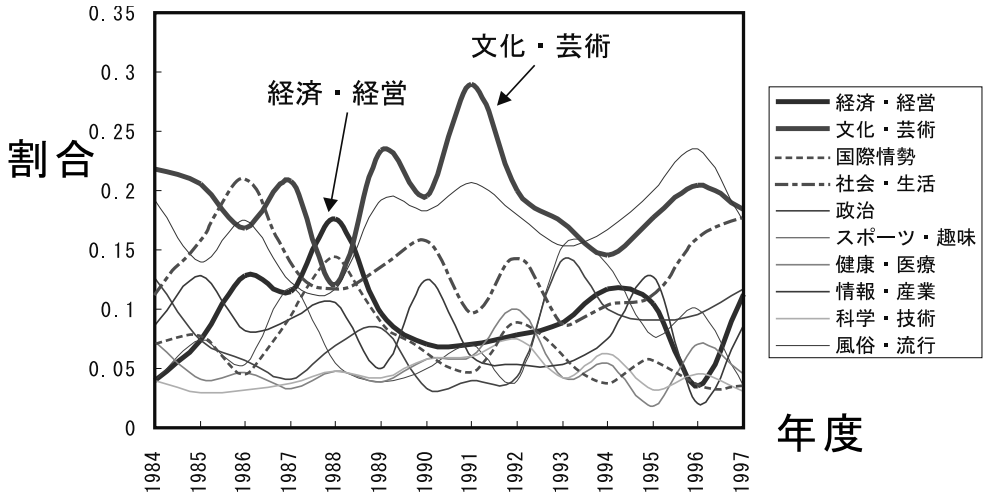

図２：年度による距離の変化（1998 年版 CD-ROM による）

[11] 1999年版の『現代用語の基礎知識』が発売されるのが、1998年の秋である。

[12] 流行語が創生される前に、具体的に流行語自体をあてることはできない。本提案のシステムに可能なことは、あくまでも、コピーライター等が流行語の候補となりそうなものを創生した際において、その評価を行うことである。

[13] グラフはExcelで平滑化した。

	経済・経営	文化・芸術	国際情勢	社会・生活	政治	スポーツ・趣味	健康・医療	情報・産業	科学・技術	風俗・流行
経済・経営	1.00									
文化・芸術	−0.78	1.00								
国際情勢	0.55	−0.39	1.00							
社会・生活	−0.25	0.23	−0.34	1.00						
政治	0.38	−0.60	0.09	−0.54	1.00					
スポーツ・趣味	0.11	−0.26	−0.15	−0.46	0.30	1.00				
健康・医療	−0.60	0.42	−0.19	0.26	−0.44	−0.31	1.00			
情報・産業	0.09	−0.05	0.16	0.22	−0.35	−0.23	−0.28	1.00		
科学・技術	−0.37	0.36	−0.10	−0.03	−0.57	−0.05	0.77	−0.16	1.00	
風俗・流行	−0.64	0.50	−0.66	0.14	−0.07	−0.23	0.36	−0.39	0.23	1.00

図３：10 種類の最上位概念相互の相関係数（1998 年版 CD-ROM）

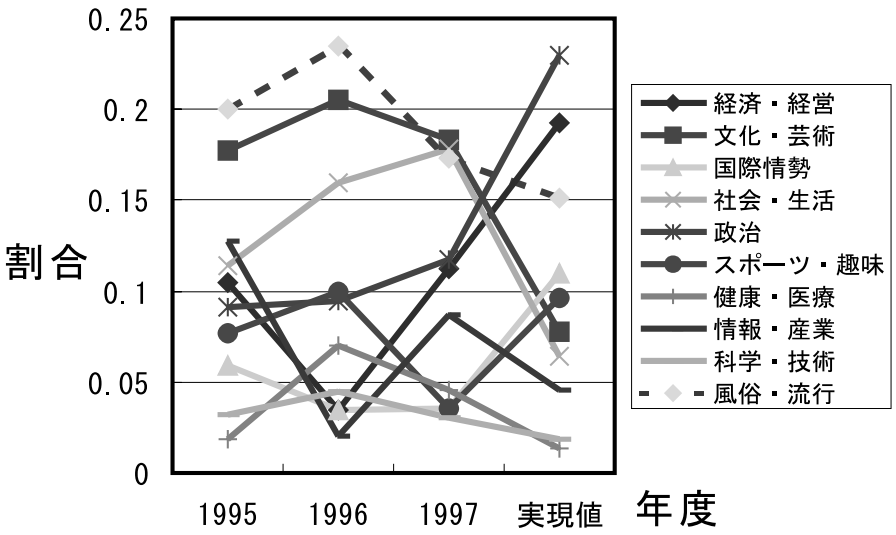

図４：流行語との距離（1995,1996,1997,1998）

- 「健康・医療」が「科学技術」と強い相関を示す傾向がある。これは、理由が分からないが、相関係数の絶対値が大きいので無視できない。

- 多少、意外なのは、「スポーツ」が、必ずしも「健康・医療」と相関しないことである。「スポーツ」「情報・産業」は、他の分野とはあまり強い相関は見せず、独立した傾向を伺わせる。

【流行しそうな分野の推定】

前節の分析から、10 個の最上位概念を、予測に利用できそうとの感触は得た。しかし、「あたる」キーワードを選別できるか否かには疑問が残る。そこで、1998 年度の流行語大賞[14]を、図２に追加して距離を計算した[15]結果を図４に示す。

本システムに要求されることは、最近数年（1995年から1997年とした）のスコアで、その次

[14] 利用したのは、「ショムニ」「ハマの大魔人」「だっちゅーの」「貸し渋り」「老人力」「モラルハザード」「凡人・軍人・変人」「冷めたピザ」「日本列島総不況」「スマイリング・コミュニスト」「ボキャ貧」である。「環境ホルモン」は今回の実験では、スコアが低かったので、除外。

[15] 図４は、図２の右端部を拡大したものである。

の1998年（「実現値」として図４中には記載）が予測できるか否かである。ここでは、直感的な判断であるが、以下の前提で考える。

- 各最上位概念が基本的には、周期を持ってはやったり、はやらなかったりしているとする。

- ピークを過ぎた最上位概念は下降に転じ、一方、下がりきった最上位概念（分野）は、上昇に転じるとする。

この判断で97年での推移で判定すると、図４からは、(1)「風俗・流行」「文化・芸術」は値は高いが基本的に下降線を辿っている。(2)「社会・生活」は、ピークに近づいているが、絶対値としては高い。(3)「政治」「経済・経営」「情報・産業」「国際情勢」は上昇に転じる。(3)「科学・技術」「科学・技術」「スポーツ・趣味」は低迷か不明。との様子が観察できる。

結果として、98年度の流行語大賞の最上位概念をみると、「スポーツ・趣味」が意外に跳ね上がり、「情報・産業」が落ちたのを除くと、測される方向で推移している。即ち、10概念中、８概念は的中である。図４を見る限り、ある程度「将来の流行語が関係する分野（最上位概念）を予測できる」と思われる。尚、「スポーツ・趣味」及び「情報・産業」については、意外な動きをしているが、これら２つは、前記の相関のところで述べたように、他分野と無相関であるとすると、予測不能はやむをえないと思われる。

【流行しそうなことば】

では、実際に、個々の流行語（候補）に関係が近いとされたことばについて確認しておく。流行語候補を「ショムニ」として、1998年度版『現代用語の基礎知識』との間で計算した距離を表1に示す。『現代用語の基礎知識』の見出し語の中で、「ショムニ」との距離が近い上位20個を示した。この場合、「経済・経営」の個数が多い（7/20）。ショムニは、風俗・流行に近いことばであると同時に経済・経営に近い。これは、相関係数の表で見たように、もともとTVドラマという風俗・流行のことばであるが、図２のグラフで上昇傾向にある逆相関の経済・経営の影響を受けていると見なし得る。

ただし、この結果は、あくまでも、ある流行語候補に関係したことばをこのシステムが自動的に検出できるかどうかの確認である。もとより、図１に示したように、完全に適切なことばのみを抽出することは不可能である。むしろ、これは、コピーライターが自分の作成した流行語候補（正確には、流行語候補の説明文）に含まれた内容が、どんな社会的なことばを想起させるかの確認に利用できる可能性を示すものである。

【流行しそうなことばの選択】

次に、複数の流行語候補があった場合につい

表１：「ショムニ」における上位20の見出し語とスコア

レベル１	レベル２	レベル３	見出し語	スコア
風俗・流行	ワードウォッチング用語	失楽園する	クサナギ君	0.523988
科学・技術	エネルギー用語	エネルギー一般	エネルギー	0.350158
文化・芸術	現代映画用語	映画作品のジャンル	スペース・オペラ	0.325636
経済・経営	労働運動用語	労働組合の種類と形態	企業別	0.317828
情報・産業	放送・映像用語	放送番組関連	チャイドル	0.315954
風俗・流行	子ども文化用語	マスコミ文化	「金田一少年の事件簿」	0.294557
風俗・流行	広告批評用語	９８年のキーワード	フジテレビの移転絶対反対	0.277011
科学・技術	原子力用語	最新キーワード	廃炉／原子炉廃止措置	0.256982
情報・産業	広告宣伝用語	広告媒体と広告表現	インタラクティブ・メディア	0.24621
情報・産業	マーケティング用語	環境変化とマーケティング	従業員満足	0.24032
経済・経営	経営問題用語	組織と人事・人材	エンパワーメント	0.235825
国際情勢	中国問題用語	最近の中国の情勢	工業企業法	0.225067
経済・経営	雇用・就職用語	雇用と労働市場	内部労働市場	0.223636
経済・経営	経営問題用語	環境変化と企業	人本主義企業	0.222368
風俗・流行	ワードウォッチング用語	マルタイ	透明な人材	0.221491
経済・経営	経営問題用語	環境変化と企業	ＣＩ	0.219497
経済・経営	労働運動用語	労働組合の種類と形態	本部／支部／分会	0.21308
風俗・流行	社会風俗用語	９７年の流行現象から	チャイドル	0.21126
情報・産業	放送・映像用語	放送事業	ＡＴＰ	0.210865
経済・経営	雇用・就職用語	雇用と労働市場	終身雇用	0.209556

表２：45個の流行語に対する評価結果

	政治	風俗・流行	経済・経営	文化・芸術	総合点
冷めたピザ	15	3	0	0	18
自主廃業	1	3	14	0	18
ショムニ	0	17	0	1	18
ビジュアル系	0	2	1	15	18
フィスコ	1	1	12	1	15
不良債権	0	1	8	6	15
新潮実名報道	0	14	0	0	14
ＭＯＦ担	4	1	8	1	14
チリワイン・ブーム	0	8	1	5	14
レオ様	0	4	0	10	14
サロン	0	3	0	11	14
スーパー高校生	0	12	1	0	13
盗聴社会	2	1	2	7	12
ダディ	0	2	0	10	12
丸のみ	5	0	6	0	11
棒の手紙	0	7	2	2	11
電車でＧＯ！	0	5	0	6	11
感情かっこ	0	4	0	7	11
世間	1	1	1	8	11
ノーパンしゃぶしゃぶ	4	1	4	1	10

て考える。具体的には、次のようにして実験を行った。1999年版『現代用語の基礎知識』には、前年に流行した（すなわち、1998年版『現代用語の基礎知識』に入っていない）新語47個が収集されている。例えば、「ビジュアル系」「小顔」等である。ただし、この中には、1998年版の流行語大賞に選ばれたものが３個（「冷めたピザ」「ショムニ」「だっちゅーの」が含まれている。そこで、この47個を1998年段階での流行語の候補と見なし、実際に、ベクトル空間法によって、大賞を取った３個が浮かび上がるか否かを検証した。

ただし、評価にあたっては、「政治」「文化・芸術」「風俗・流行」「経済・経営」に注目した。これにはあまり深い意味はないが、図４の中で、絶対値が大きいものであって、かつ、図３では、他分野との相関が比較的よく見られるものを選択した（「社会・生活」は図４では、値が大きいが、図３では、他分野と、あまり強い相関を示していない。）。結果を表２に示す。

表２に示すように、流行語大賞の対象となった３候補中で２候補が、１，３位として、入っている。スコアの高い「自主廃業」は、流行語大賞に選択されていないが、もともと、流行語大賞では暗い話題は選択されない傾向にあり、その意味では、この２位の候補はもともと流行語対象外と考えてよい。なお、「だっちゅーの」はスコアが低かった。ひとつの大きな原因は、この流行語47個の説明文が極めて短いことがあげられる。事実、「ショムニ」については、表１とは異なり、「風俗・流行」がほとんどを占めている。これは、説明文が短い場合に、書き手に依存して、近いとされる分野が異なることを明確に示している。これは、本システムの予想される問題点である。

3.2.3 実験２の結果：1999年版CD-ROMによる評価

同様の実験を、1999年版のCD-ROMを辞書データとして利用した。流行語としては、2000年版CD-ROMに掲載された流行語大賞（1984～1999）の説明文を利用した。これらの説明文は、2000年度版から大幅に書き換えられた。説明文は、1998年版のCD-ROMの評価で利用した流行語大賞の説明文に比較して短い。図５にその例を示す[16]。

[16] 後述するように、このような短い説明文でも、流行がある程度予測できることが本研究の意外性の一つである。

「ショムニ」とは庶務第二課のことで、“役にたたない”社員の島流しのような部署である。シュンとする男性陣に対し、女性陣はとにかく元気。ある意味では、極めて今日的なテーマをマンガチックに描いている。元気な女性が大活躍というストーリーがヒットの理由という。

図５：今回利用した流行語の説明文の例

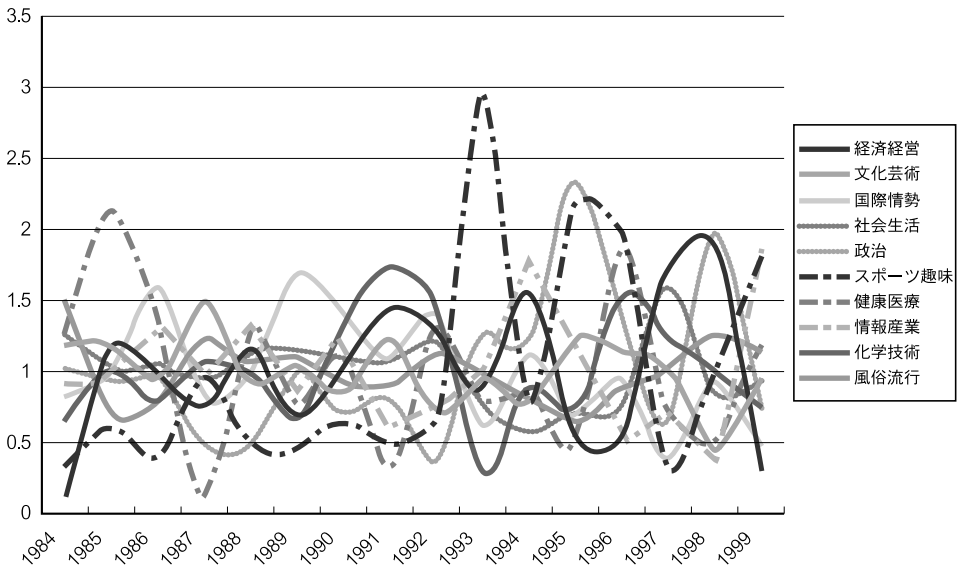

図６：年度による距離の変化（1999 年度版 CD-ROM による）

図６は、1999年版CD-ROMによる年度経過である。ただし、図２とは異なり、各分野を、その平均値に対して、正規化している。これは、図２の分析から、各分野の割合の絶対値より、分野の割合値の増減のほうが重要と推定されたからである。

最後の４年間について、結果を図７に示す。ただし、上記分析から、各分野（最上位概念）毎の出現割合よりも、むしろ、増減のほうが問題に思われたので、ここでは、各最上位概念毎に、およそ 10 年間の割合の値の平均を求めて、これを「１」とする正規化を行った。この図からは、この15年間において、ある年に平均値の値が1.5を越えたものは、ほぼすべてがその翌年に下降することが伺える。

そこで、96 年以降の様子を図７に示す。この図７の 98 年までのの部分を見る限りでは、以下の傾向がある。

- 1998 年段階では、「経済・経営」「社会・生活」のスコアが高いが、すでにピークであり、99 年は下降すると思われる（事実、そうなっている）。

- 上昇機運に乗っているのは、「スポーツ・趣味」「国際情勢」「風俗・流行」である。

ただし、「風俗・流行」は変化が小さい。つまり、急に注目されている訳ではない。

- 「文化・芸術」「健康・医療」は底を打っている。急速に上昇すると思われる。

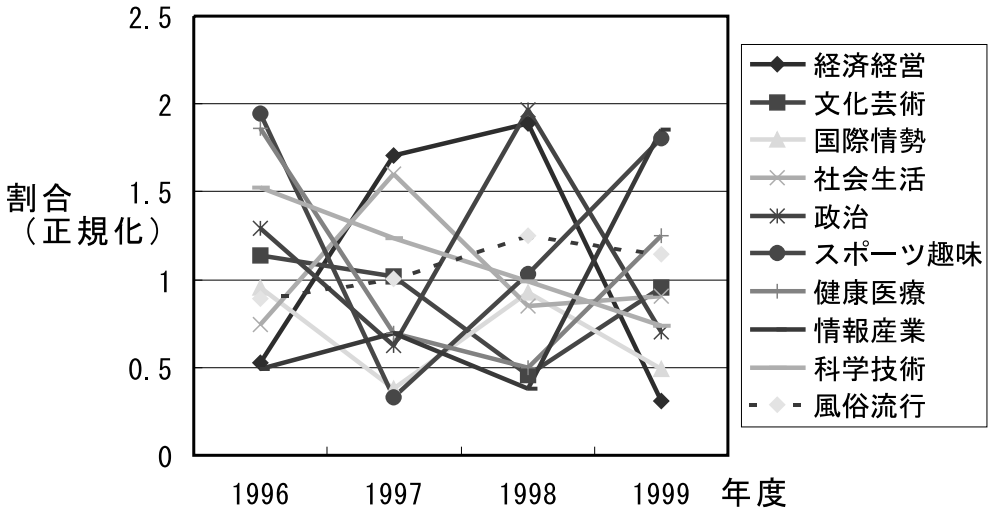

図7：1999年版CD-ROMによる実験結果（細部）

2000年の流行には「文化・芸術」「健康・医療」「スポーツ・医療」分野が大きく関わってきそうな気配がここからは読みとれる。2000年4月からの「介護保険制度」の発足が、ここにあることは不思議な一致である。

実際、「文化・芸術」「健康・医療」「スポーツ・医療」「国際情勢」への注目を予測させるようなデータがある。2000年3月まで放送されていたテレビドラマに、TBS系の「ビューティフルライフ」と、CX系の「二千年の恋」とがある。この2つは売れっ子の俳優や脚本家などを使い、どちらのドラマも高い視聴率をとってもおかしくないものだった。「ビューティフルライフ」は、カリスマ美容師と障害を持つ女性の物語である。一方、「二千年の恋」は、国際テロリストとその恋人を題材とする。

「ビューティフルライフ」は、上記注目4分野中の3分野に関係している。これに対して、「二千年の恋」は国際関係に対してのみ関係する。「ビューティフルライフ」はビデオリサーチが調査を開始して以来ドラマ部門で歴代2位の高視聴率[17]をとったのに対し、「二千年の恋」は14%程度であり視聴率ベストテンの中には出てこなかった。やはり、「あたりそうな」複数の分野を組み合わせて作ってゆくことがドラマの場合は効果的なように、このデータからは伺える。

3.3 中間的結論

以上の実験から、以下の課題が明らかとなった。

1. 今回の実験では、流行語説明文から生成されたベクトルと、辞書説明文から生成されたベクトルが共有する要素の数は必ずしも多くなかった。見出し語の説明文があまり長くないことも一因である。従って、スコアが、個々のベクトル毎に大きく揺らいでいる可能性は残る。しかし、それでもなお、人間の直感と合致した結果がでている点は興味深い。
2. 上記の議論では、グラフが「上がっている」といった、主観的な尺度で、どの分野が流行するかを論じている。しかし、これは、明らかに学問的に不十分であり、これらのグラフが統計的にどのような性質を有しているかを示し、統計的に翌年度の予測値を出す必要がある。
3. 以上の実験では、ベクトル空間上で距離が最近傍のものを利用することなく、近傍

[17] 最終回の瞬間最高視聴率41.3%

の 20 個を選び、その全体的傾向を見ている。結果が良かった一つの原因と思われる。しかし、この 20 個の個数に技術的必然性はない。今後は、どの程度の個数を利用するのが効果的であるかも検討を要する。

4．結果の統計的検証

本章以降では、前章までの結果を統計理論的に検証する。但し、本章では、評価をより厳しく出すために、流行語の説明文が極めて短い、1999 年版 CD-ROM を用いた実験（図６）等に基づくこととする。以下に相関係数の検定があるが、相関係数は、図３に示したものとは異なる。相関係数そのものがより低めに出た厳しいデータとなっている。

4.1　意味ベクトル法の評価

本節では、まず、実際に、意味ベクトル距離計算において合致した単語そのものについて、分析する。

4.1.1　共通要素単語の分析

図６の最上位概念（分野）毎の上がり下がりを流行推定に利用するためには、意味ベクトル距離の原因である、共通要素の単語（意味ベクトルで双方の説明文で共通に現れた単語）が人間から見て妥当である必要がある。そこで、以下の集計を行った。

- 各年度・各分野ごとに、意味ベクトル法の共通単語を抜き出し、単語毎に度数を集計
- 度数が２回以下になったものを除き、上位５位までの単語を抽出

但し、上位５位までに入るものが５個以上あったらすべて抜き出し、上位５位までに登場回数２回以下のものが含まれる場合はそれは抜き出していない。結果の一部を図８，図９に示す。

4.1.2　実験結果分析

「経済経営」分野の結果（図８）では 1990 年から 1993 年にかけて、単語「バブル」が登場して

1990	株	年	バブル	経済	状況
・	9	6	5	5	4
1991	会社	証券	株	投資家	株価
・	13	12	11	10	9
1992	経済	バブル	崩壊	年	不況
・	15	12	11	7	7
1993	価格	バブル	会長	行使	手
・	4	3	3	3	3
1994	価格	物価	下落	商品	制度
・	13	9	7	5	5
1195	年				
・	3				
1996					
・					
1997	証券	改革	バン	ビッグ	売買
・	11	10	9	9	9
1998	金融機関	債権	金融	経営	年
・	25	23	18	12	11
1999					
・					

図８：「経済経営」の上位５位までの要素

1990	大統領	イラク	年	軍	世界
・	10	9	8	7	7
1991	年	派	革命	月	事件
・	11	5	4	4	4
1992	大統領	年	勝利	選	国民
・	17	13	9	9	8
1993	日本				
・	3				
1994	思想	社会	主義	人	党
・	5	4	4	4	4
1195	政党	選挙	候補		
・	6	5	3		
1996	首相	月	政治	社会	英語
・	5	4	4	3	2
1997					
・					
1998	年	共産党	首相	選	委員長
・	6	5	5	4	3
1999	なら	選挙			
・	3	3			

図９：「国際情勢」の上位５位までの要素

いる。このころは、「バブル景気」が終わり、日本が不況に向かってまっしぐらに落ちていった時期である。また、1997年の「ビッグ」「バン」や、1998年の「金融機関」「破綻」といった単語も、金融ビッグバンや、相次ぐ金融機関の破綻があった時期と一致する。

この例を見ると、「経済経営」の分野に関しては、流行語に近いとされた見出し語の説明の中には、確かに、当時の社会的影響として重要なものが含まれている。提案手法の距離計算には、一定の根拠があると考えられる。すなわち、機械的に判定された近い「ことば」と人間の感覚とは、ある程度、一致している。

しかし、どのような最上位概念についても、そのようになっている訳ではない。「国際情勢」（図９）では、1990年の「イラク」、1992年の「大統領」「選」など、当時の国際情勢を的確に反映した単語も見受けられる。

また、1991年では、スコアは高いが、それらから何も推測できないような単語が根拠になっている。1998年の「共産党」「委員長」のように、日本共産党の不破委員長を指す「スマイリング・コミュニスト」の説明文から抜き出されたと思われる、的外れな単語もある。

総じていえば、「経済経営」「政治」分野では機械の判定と人間の想像がおおむね一致しているが、その他の分野では、必ずしも、人間の直感に合わない単語が、意味ベクトル法の共通単語となっている場合がある。真面目で堅いイメージのある分野、どちらかといえば流行語の持つイメージを書きことばで表現できるような分野は機械による判定に向き、お遊びのイメージのある分野、どちらかといえば流行語の持つイメージを話しことばで表現するような分野は機械による判定に向かないことになる。このような分析は直感的なものであるので、さらに統計的に確認する必要がある。

4.2 相関係数の統計的検定

まず最初に、前章で示した相関係数を統計学的に検定する。ただし、相関係数は、両分野の推移カーブが意味を持っていないと有意ではない。その意味では「きつい」検定である。

4.2.1 相関係数の統計的検証

相関係数を求めた際の各ベクトルの要素数は

表３：相関係数の区間推定結果

相関の対象分野	上限	下限
経済経営×文化芸術	-0.848	-0.116
政治×社会生活	-0.852	-0.130
スポーツ趣味×社会生活	-0.875	-0.218
スポーツ趣味×政治	0.838	0.083

	経済経営	文化芸術	国際情勢	社会生活	政治	スポーツ趣味	健康医療	情報産業	科学技術	風俗流行
経済経営	***									
文化芸術	0.020	***								
国際情勢	0.777	0.330	***							
社会生活	0.699	0.893	0.907	***						
政治	0.731	0.647	0.355	0.018	***					
スポーツ趣味	0.318	0.815	0.102	0.008	0.026	***				
健康医療	0.464	0.761	0.397	0.832	0.500	0.715	***			
情報産業	0.854	0.935	0.512	0.254	0.774	0.846	0.835	***		
科学技術	0.185	0.958	0.696	0.405	0.184	0.218	0.780	0.158	***	
風俗流行	0.614	0.739	0.851	0.147	0.363	0.334	0.444	0.137	0.696	***

図 10：無相関検定結果

15 である。これらを、母集団からのサンプル要素と見なし[18]、「無相関検定[36] 」を行った。更に、「無相関ではない」と判定されたものに限って、その母集団の真の相関係数の区間を推定した。

4.2.2 相関の検定

具体的な方法は以下の通りである。

- 帰無仮説を「母集団の相関係数は０である」とする。
- 標本数と標本相関係数から、検定統計量を計算し、それを元にして *P* 値を算出する。
- *P* 値が有意水準（α とする）以下であれば、最初に立てた仮説が棄却されるので、「母集団の相関係数は０ではない」といえる。

これらの計算は群馬大学の青木繁伸教授のウェブサイト[36, 37] を利用している。有意水準は１％と５％とした。結果を図 10 に示す。

有意水準１％のかなりきつい条件で検定した結果、有意性が認められるのは「スポーツ趣味」＝「社会生活」だけであった。有意水準５％では、４個が有意となる。これらは厳しい結果である。しかし、相関が有意となるためには、本来、それを構成する双方の最上位概念に手法が有効でなければならない。その意味では、10 個中、約半分が、統計的に意味がある最上位概念といえなくもない。

4.2.3 相関係数の区間推定

標本数と標本相関係数をもとに、母集団の相関係数が 95% の確率で取り得る範囲（信頼区間 95%）を求めた。結果を表３に示す。注目されるのは、「スポーツ趣味」×「社会生活」である。この組み合わせは有意水準１％で有意性が認められた組み合わせであり、その相関係数は最大で -0.875 と強い負相関である（図３とは異なる CD-ROM データであるため、結果が異なっている。）。しかし、検定結果が強い支持をするのは、この組み合わせに限定されており、相関係数については、一つの目安にはなるが、統計的に強い

[18] つまり、相関係数は、２分野の時系列的な関係を無視し、単年度での相互の関係を見ている。

根拠とするのには無理があるように感じられる。

5．金融工学的手法の適用

5.1 金融工学的手法

「未来の予測」は、流行の世界だけではなく、他の世界でも行われている。その代表として、株や為替などを予測する金融界がある。金融商品の運用において、最も重視されるのは、利益を極大化することと同時に、リスクの分散である。これに関して、1952年にマーコビッツによって発表された「ポートフォリオ理論」以来、さまざまな理論が生まれた。

このように、金融界では約50年前から「利益の極大化」と「リスクの分散」に関して多くの理論が生まれている。これをベースとして、流行予測にも同様の考え方を導入できないか、と考えたのが金融工学的手法の適用である。

今回は、「利益の極大化」＝「最もあたる分野の予測」というのは今までの手法で一応の目的を達しているので、金融工学的手法では、「リスクの分散」＝「外れそうな分野の回避」を目的とする。

5.2 適用手法とその結果

今回は、株の値動きの予測に使われる指標のうちの「RSI」をベースとした[40]。RSIは、テクニカル分析の指標一つであり、単純に株価の上昇・下降を見る「サイコロジカルライン」に、株価の波の動きを足したものである。RSIの計算式は以下の通りである。なお、基本的にRSIは2週間分（10日から14日の間）程度のデータを用いて計測される。

$$RSI = \frac{\text{上昇した日の値幅の合計}\times 100}{\text{上昇した日の値幅の合計+ 下降した日の値幅の合計}} \ (\%)$$

株の世界では、RSIが30%を下回れば売られ過ぎ、70%を越えれば買われ過ぎと解釈される。

RSIは、計算式そのものは非常に単純であるが、波の「振れ」を計測して流れを分析する点や、買われ過ぎ・売られ過ぎといった、人の心を数値化するという考え方では、人の心を読み取りたい流行予測の考え方と非常に近い。そこで、これを流行予測向けに改良した。計算式は以下の通りである。但し、「ポイント」とは図6の（最上位概念毎に正規化された数値の）上げ幅、あるいは、下げ幅の絶対値である。

$$\text{疑似}RSI = \frac{\text{上昇年度のポイント幅の合計}\times 100}{\text{上昇年度のポイント幅の合計＋下降年度のポイント幅の合計}} \ (\%)$$

この「疑似RSI」では、本来のRSIを考えると12年分程度のデータが欲しいところであるが、サンプル数が少ないため、6年分のデータを用いて計測した。そして、その結果を10%ごと（1の位切り捨て）に区切って、その数とその値を取った翌年に本当に上昇したかを調べた。その結果を表4に示す。なお、この結果は、分野ごとの振り分けを行っていない。また、疑似RSIを算出したのは1998年度までである。（1998年までの6年分のデータから計算したRSIを用いて、1999年を予測するため）

表4：RSI値と翌年流行上昇数

RSI値	事例数	翌年上昇数	割合（%）
0.1〜0.19	1	1	100
0.2〜0.29	2	2	100
0.3〜0.39	16	10	62.5
0.4〜0.49	30	13	43.3
0.5〜0.59	35	10	26.6
0.6〜0.69	14	4	28.6
0.7〜0.79	2	0	0
0.8〜0.89	0	0	0

5.3 結果の信憑性

RSIが10〜30%の範囲では、サンプルが少ないものの、確率＝1でその翌年は上昇している。また、70%台も同様にサンプル数が少ないが、確率1で下降している。

この結果から、RSIの値の大きいほど、翌年には下がる可能性が高いと推測できる。しかし、このデータがどの程度の統計的信頼性を有するかが明らかではない。そこで、サンプル数の少なさを統計的に補うために、この結果に対してブートストラップによる検定を行った。

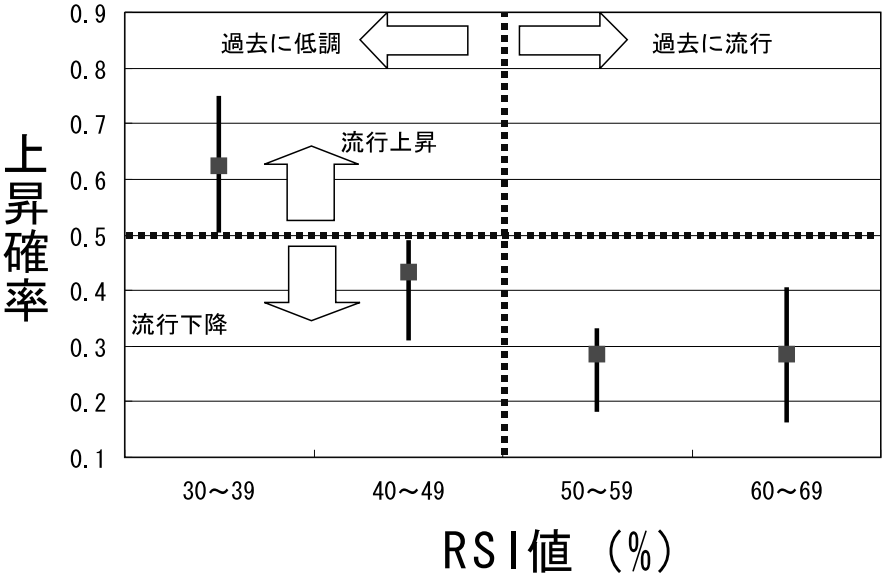

図 11：RSI 値と上昇確率（Bootstrap 併用）

5.4 ブートストラップ検定の適用

5.4.1 ブートストラップ

ブートストラップは、統計値の分布を得るための汎用的な手法であり、Efron により提案された[38]。基本的には、n 個得られたサンプルから、重複を許してn 個をリサンプリングすることを何度も繰り返す。そして、このリサンプリング毎に目標となる統計値を計算する。統計値はある分布を持つこととなるが、この分布は、サンプルではなく、母集団からリサンプリングした場合の統計的分布に等しいことを利用する。

ブートストラップの優れた点は、特に、事前に誤差分布を仮定しない、すなわち、ノンパラメトリックな統計量に対しても、当該統計量の誤差分布を求めることができる点にある。今回のRSI は、ノンパラメトリックな統計量であるため、ブートストラップの適用に好適である。

5.4.2 予測への適用

ブートストラップによる検証を行う上で、次の点に留意しなければならない[39]。

- サンプル数が 9 を下回ると正確に検証できない。30 以上あれば申し分ない。
- 繰り返し回数は、標準偏差を取るには、1000 回。正確な区間推定には、10000 回が推奨される。

今回のデータは、RSI が 30% ～ 70% のサンプルは 10 個以上ある。そのため、ブートストラップによる検証は、上記の最初の留意点をクリアした 4 種類について行った。また、2 番目の留意点に関しては、推奨の通り 10000 回繰り返しを行った。結果を図 11 に示す。エラーバーは、上と下に各々σである。

RSI 値が低い 30 ～ 40% に対しては、統計的なバラツキを考慮しても、明らかに有意に流行する方向である。それに対して、40 ～ 50% の多少ははやっていたものを含めて、過去に流行していた最上位概念は飽きられている。とりわけ、この結果を見る限り「過去に特にはやらなかったもののみが当たる」傾向があるように思われる。

5.5 リスクヘッジ

ブートストラップ実験によって、一部の例外はあったがおおむね、「RSI の値が低いほうが翌年に上昇する可能性が大きい」という結果が得られた。ただし、仮に 60% 台の値を取ったとしても、上昇する可能性は 25% 程度はあるのも事実である。そこで、この結果を逆に考え、「翌年上昇する可能性の低い分野は、RSI の値が高い」と考えることで、あたるはずもないドラマを仕立てるような危険を、回避できることになる。

株の世界でも、上昇の見込める銘柄の選定というのは難しい。しかしながら、下降する可能性のある銘柄を購入対象から除外するということは、上昇銘柄の予測よりは比較的簡単である。

50 年程度研究され続けている金融工学にはまだまだ及ばないが、今まで統計的な検証が行われてこなかった流行予測の分野において「リスクヘッジ」の手法を提案できるようになったことは、このシステムの予測に対して統計的な裏付けが出来たということは、コピーライター・クライアントの双方にとって非常に有意義と考える。

6．実験結果の考察

今回利用している流行語の説明文は極めて短いものであり、これに基づいて意味ベクトル法による距離計算を行っても、どこまで、それが意味があるかに疑念を有する部分もあった。しかし、相関係数の統計的検定（最上位概念の相互関係の単年度限定の評価）、RSI 検定（最上位概念単独での、時間的変化）の何れも、全てとはいえないが、統計的にも有意の結果が得られている。

本システムで評価対象としているのは、最上位概念であり、かなり広い分野をカバーしている。その意味では、これが、直接に流行語選定に寄与できるか否かは今後の課題である。また、今回の実験では、本来、各年度の辞書を利用すべきところを、単年度の辞書で代用している。今後、各年度の『現代用語の基礎知識』をデータとして整備し、その状態で、予測が可能であるか否かを検証する必要がある。しかし、相関係数等が、たまたま得られた数値ではなく、統計的に有意であることが今回の実験で確認できた。今後は、これらデータの充実を図りながら、より、精度の高い研究へと発展させたい。

7．まとめ

本稿の主要な結論を以下にまとめる。

- 意味ベクトル法の共通要素単語の分析からも、相関係数の統計的検定結果から見ても、「経済・経営」「政治」等の社会経済的な分野では、マスコミ論で論じられてきた「流行語の背景には社会経済的影響がある」との見解が定量的に裏付けられた。これらの分野では、提案手法が、流行語候補評価手法として有効である可能性を示唆する。

- 一方、その他の分野、特に「風俗」等の分野では、相関係数が出ても統計的に信頼度が低い。また、意味ベクトル法の共通要素となっている単語自体を見ても、意味的に妥当な単語とはいい難い面がある。風俗的な流行については、統計的に「あてる」ことが難しいと示唆される。これらは社会学者の見解と一致するものかもしれないが、工学的に流行を予測するシステムの立場からいえば、別のアプローチが必要と思われる。

- 流行予測は、株価のように、本来的には困難なタスクである。従って、株価のアナロジーで考えるなら、予測はできなくても、「あたりそうにないものを外す」「複数のあたりそうな物を組み合わせる」等の対処法も必要と思われる。

- 尚、今回の分析は、あくまで、辞書 M_i として単一年度の辞書を利用し、しかも、統計的検定は、分野（最上位概念）相互の相関係数について論じている。相関係数は、双方の分野が意味を持たない限り、有意とはならず、厳しい評価である。辞書の充実を図り、単一の分野（最上位概念）の動きが有意か否かを、継続して検討する必要がある。

謝辞

本研究に際して、『現代用語の基礎知識』の利用を許諾頂いた、自由国民社に深謝いたします。本研究に関してご支援を頂いた、北寿郎さん、松沢和光さん、大山芳史さん、金杉友子さんを始めとするNTT コミュニケーション科学基礎研究所関係者各位に深く感謝の意を表します。尚、本論文の内容は、平成10, 11, 12年度のNTT コミュニケーション基礎研究所よりの受託研究の内容を含みます。また、本研究の一部は、文部省補助に基づく同志社大学・学術フロンティア研究プロジェクト「知能情報処理科学とその応用」として行ったものです。

参考文献

[1] 堀浩一・鏡明・小林健一・鈴木宏衛『ことばの「連鎖」が生む新しい発想法（対談）：月刊アドバタイジング４月号より』電通、1998、pp.52-57
[2] 「1998 年版現代用語の基礎知識」自由国民社
[3] 「1999 年版現代用語の基礎知識」自由国民社
[4] 「2000 年版現代用語の基礎知識」自由国民社
[5] 米川明彦、『現代若者ことば考』丸善ライブラリー、1996年、pp.85-172
[6] 米川明彦、「若者語を科学する」明治書院、1998年、pp.85-148
[7] 小林千草、「ことばの歴史学源氏物語から現代若者ことばまで」丸善ライブラリー、1998、pp.200-227
[8] 「語彙調査 - 現代新聞用語の一例―フォネーム研究序説」国立国所研究所、1978
[9] 「婦人雑誌の用語―現代語の語彙調査(4)」国立国所研究所、1953
[10] 「総合雑誌の用語―現代語の語彙調査―（前編・後編）(12,13)」国立国所研究所、1957, 1958
[11] 「現代雑誌九十種の用字用語」国立国所研究所、1962, 1963
[12] 「テレビ放送の語彙調査」国立国所研究所、1995, 1997, 1999
[13] 松本裕治、影山太郎、永田昌明、斎藤洋典、徳永健伸「岩波講座言葉の科学３単語と辞書」岩波書店、1997、pp.157-159
[14] 長尾真、黒橋禎夫、佐藤理史、池原悟、中野洋「岩波講座・言語の科学９・言語情報処理」岩波書店、1998
[15] ibid、第４章、言語の統計
[16] ibid、pp.14-15
[17] ibid、pp.151-194
[18] 長尾真編、『岩波講座ソフトウェア科学15　自然言語処理』岩波書店、1996 、pp.1-12
[19] ibid 、pp.117-130
[20] 日本ユニテックSGML サロン「はじめてのSGML」技術評論社、1995、pp.12-28
[21] 電通・自然言語解析システム DE-FACTO, 株式会社電通・電通リサーチ、1999.
[22] 金杉友子 , 松澤和光 , 笠原要「アバウト推論の「ことば遊び」への適用」信学技報 NLC96-31,1996.
[23] 松澤和光、湯川高志、笠原要、藤本和則「アバウト推論技術」NT R&D Vol.45 No.11, 1996
[24] 笠原要、松澤和光、石川勉「国語辞書を利用した日常語の類似性判別」情報処理学会論文誌、第 38 巻、第７号、pp.1272-1282、1997.
[25] 阿部明典「広告などにおける感性つきことばの概念ベースによる生成の可能性」, 第59回情処全大, Vol.2, 5N-04, pp.397-398, 1999.
[26] 阿部明典「ことば工学の地平線―あとがきにかえて―」人工知能学会研究会資料, SIG-LSE-9901-12, 1999.
[27] 奥野貴司「広告効果のメカニズムとその現状」, 情報処理学会 , 情報メディア研究会資料 , 14-3, pp.13-18, 1994.
[28] 鈴木賢一郎、稲積宏誠、楠本和也「広告におけるメディア選択支援方式の研究―DEA を用いた予測システムの検討―」, 情報処理学会第58回（平成11 年前期）全国大会 2K-4, 291-292, 1999.
[29] 池田定博、金田重郎、金杉友子、加藤恒昭「現代用語辞書を用いた流行コンセプト作成支援」電子情報通信学会・言語理解とコミュニケーション研究会 NLC99-93, 2000 (パターン認識・メディア理解研究会と合同、PRMU99-276, pp.113-120, 2000).
[30] 大橋正和、池田定博、金田重郎、金杉友子「自然言語処理を用いた流行コンセプト予測支援」経営情報学会・2000 年春季全国研究発表大会 1C-1-4, pp.66-69, 2000.
[31] 大谷實、太田進一、山達志編著「総合政策科学入門」、成文堂、1998、pp.4-5、pp.12-13
[32] 全日本シーエム連盟「ACC 年鑑 2000」、宣伝会議
[33] 東京コピーライターズクラブ「TCC 年鑑 2000」、宣伝会議
[34] オリジナルコンフィデンス「オリコン年鑑」、オリジナルコンフィデンス、1999
[35] 松澤和光、堀浩一、金杉友子、安部明典「ことば工学入門」人工知能学会誌 , 15 巻３号 , pp.446-455, 2000.
[36] http://aoki2.si.gunma-u.ac.jp/lecture/Corr/corr.html
[37] http://aoki2.si.gunma-u.ac.jp/lecture/Corr/corr3.html
[38] B. Efron and R. J. Tibshirani, “An Introduction to the bootstrap”, Chapman & Hall, 1993.
[39] M. R. Chernick, “Bootstrap Methods - A Practical Guide”, John Wiley & Sons, Inc. 1999. Chap. 9. Too Small Sample Size.

[40] 林康史、「株価が読めるチャート分析入門」かんき出版、2000.

付録 A：DE-FACTO の機能と仕組み

本付録では、DE-FACTOについて、簡単に紹介する。そして、本論文のアプローチとの差異を述べる。

(1) 関係抽出

DE-FACTO では、まず入力されたテキスト情報を「テキスト単位」（文、段落など意味的なまとまりを持ったことばの集まり）に分割する。次に、各テキスト単位から「単語」を切り出し「同じような単語が現れるテキスト単位同士は近い」「よくいっしょに現れる単語同士は近い」といった性質を用いて、テキスト単位や単語の関連度（連想の強さ）を統計的手法により計算する。

(2) 連想検索

(1) で計算された関連度の情報を用いることにより、単語から関連単語、単語から関連テキスト単位、テキスト単位から関連単語、テキスト単位から関連テキスト単位という四種類の連想検索ができる。またこれらの基本機能に加え、組み合わせ検索、分野を限定した検索、連続検索、連想パス検索といった応用機能も用意されている。

(3) 図解化

(1) で求められるテキスト単位・単語間の関係を、人間が直感的に把握しやすい「図解」として表示できる。図解化に利用される自動レイアウト機能には複数のタイプのものが用意されており、似た内容の要素をグループ化して階層的な図解（KJ 法的な図解）を作成や近い内容の要素が近くに配置されるような図解（平面マップ）を作成ができる。また、図解を編集する機能、編集結果をレイアウトしなおす機能、部分拡大機能、図解の変化をアニメーションで見せる機能などもあり、パソコン上で簡単に図解を操作できる。

DE-FACTO もコンセプト開発のための支援ツールである。しかし語群としては一般辞書を使用しているため、本論文の提示する新語・造語を語群に持った本研究とは異なる。ただし、リサーチ会社のツールとして消費者へのグループインタビューをかなり長時間にかけて行いその内容をテキスト化し検証している点で特定の商品コンセプトには優れている。唯一の問題点はそのテキスト化や微修正などの時間の浪費とコストである。「流行ことば予測」のシステムはマーケティングの担当者が使用することを前提に不要な部分を省き、より簡易に結果を出せるものとして開発されている。

以上

付録 B：SGML 化について

本付録では、実験に際して行った『現代用語の基礎知識』のSGML化と、処理に際しての見出し語の内部表現法について、簡単に紹介する。

(1) 現代用語の基礎知識の SGML 化

SGML とは、Standard Generalized Markup Language の略で、テキストの論理情報を記述するための国際的な標準規格（ISO8879 ）である。1986 年に定められ、その後、アメリカの公的機関や業界団体などの積極的支持によって普及した。日本では 1992 年に JISX4151 として規格化された。われわれは、辞書の文面を見ると、どれが「見出し語」で、どれが「解説文」かを容易に理解する。これに対して、コンピュータでは、このような意味の理解はできない。そこで、SGML では、「見出し語」には＜見出し語＞、解説文には＜解説文＞という目印「タグ」を付けておく。この際、テキストの論理的構造は、DTD（Document Type De nitioh 文章定義）と呼ばれる別のテキストで正確に定義される [20] 。

本研究では、先に紹介した「現代用語の基礎知識 1998, 1999」の全文を SGML 化した。「現代用語の基礎知識」のデータは全部で 4 階層のレベルを持ち、最下位のレベル 4 が見出し語を形成する。データはすべて SGML 化している。例を以下に示す。1998 年版で CD-ROM から収録されている 1998 年度の「日本新語・流行語大賞」（1997 年の語対象）で話題になった「失楽園（する）」を見てみる。

原文は次の通りである。

「失楽園（する）」

作家・渡辺淳一が日経新聞に連載した小説「失楽園」がベストセラーになり映画化、ついでＴＶドラマ化された。小説も、お堅い経済新聞が「ここまで描いて」低迷する財界を励ましたと言われたが、映画になってみると高年齢層を中心に大ブームを巻き起こし、主演女優黒木瞳、果ては「失楽園したい女ベストテン」が男性週刊誌の定番企画になった。いっぽう、ＴＶウォッチャーナンシー関によれば「川島なお美ひとりが屍の上に立っている」状態になりはしないかとドラマ開始前から危惧し、見事彼女の危惧は的中した。なんの脈絡もなくシャワーシーンが挿入される「前回までのあらすじ」、日活ロマンポルノの手法である「画面中央の花瓶」など、一般視聴者を辟易させたが、おやじ連中は大喜び。瞬間最大視聴率は 40% を記録した。

以下は SGML 化された「現代用語の基礎知識」の例である。

< LEVEL 1 name=“ 風俗・流行 ” >
< LEVEL 2 name=“ 社会風俗用語 ” >
< LEVEL 3 name=“ 97 年の流行現象から ” >
. . . . . （中略）. . . . . .
< LEVEL 4 name=“ 失楽園（する）” >
作家・渡辺淳一が日経新聞に連載した小説「失楽園」がベ

ストセラーになり映画化、ついでＴＶドラマ化された。小説も、お堅い経済新聞が「ここまで描いて」低迷する財界を励ましたと言われたが、映画になってみると高年齢層を中心に大ブームを巻き起こし、主演女優黒木瞳、果ては「失楽園したい女ベストテン」が男性週刊誌の定番企画になった。いっぽう、ＴＶウォッチャーナンシー関によれば「川島なお美ひとりが屍の上に立っている」状態になりはしないかとドラマ開始前から危惧し、見事彼女の危惧は的中した。なんの脈絡もなくシャワーシーンが挿入される「前回までのあらすじ」、日活ロマンポルノの手法である「画面中央の花瓶」など、一般視聴者を辟易させたが、おやじ連中は大喜び。瞬間最大視聴率は 40 % を記録した。

< /LEVEL 4 >
< /LEVEL 3 >
< /LEVEL 2 >
< /LEVEL 1 >

(2) 実際の内部処理

SGML 化した各＜見出し語＞は、そのままコンピュータの内部で処理はしていない。これは、ストリングのままであると、内部のメモリを大幅に食うからである。そこで、一種のハッシュを行い、各見出し語は、番号にして表現した。これにより、検索の高速化と、メモリの削減を図っている。例えば、1998 年版で生成された見出し語は 8901 個であった。この結果、例えば、以下のように、ハッシュ値と見出し語が対応する。

1 地域連携
2 東南アジアの通貨不安
3 アムステルダム条約
4 ツチ族の逆襲

. . .（中略）. . .

6361 アダルトチルドレン
6362 複雑系
6363 失楽園
6364 酒鬼薔薇
6365 アジアアロワナ
6366 東電ＯＬ殺人事件

. . .（以下省略）. . .

以上

付録 C：形態素解析

本付録では、実際の距離計算に利用する単語の抽出方法、ならびに、内部表現法について簡単に紹介する。

(1) 形態素解析の実行

各＜見出し語＞は＜説明文＞を持ち、ここには、「社会的要因を示す語」が多数含まれている。そこで、この説明文中で、実験の妨げとなりがちな助詞・助動詞などは省き名詞、サ変動詞を残すため、品詞別に分類する形態素解析プログラムにかけ分類した。

形態素解析とは、自然言語処理のひとつである。具体的にいえば、日本語は英語等と違って分かち書きされていない。そこでコンピュータで単語単位に区切る必要がある。このため文法的な規則等により文字列に自動的に区切りを入れ、各単語の品詞等を決定するツールとして形態素解析がある[13, 19]。

具体的に先ほど SGML 化した「現代用語の基礎知識 1998」の中から「失楽園」を形態素解析にかけた例を示す。

次に、社会的要因を表す語としては意味のない、「てにをは」などを省く。実際には、今回の実験では、名詞、サ変名詞、形容詞、動詞を抽出した。

失楽園　作家　渡辺　淳一　日経　新聞　連載　小説　失楽園　ベストセラー　映画　ドラマ　小説　堅　経済　新聞　低迷　財界　映画　年齢　層　中心　ブーム　主演　女優　黒木　瞳　果て　失　楽園　女　ベストテン　男性　週刊誌　定番　企画　関　川島　美　ひとり　屍　状態　ドラマ　開始　彼女　危惧　的中　なんの　脈絡　シャワー　シーン　挿入　あらすじ　日　活　ロマン　ポルノ　手法　画面　中央　花瓶　一般　視聴　辟易　おやじ　連中　喜び　最大　視聴率　記録

(2) 内部処理について

形態素解析の結果得られた名詞等は、前述の＜見出し語＞と同様にして、内部ではハッシュして扱っている。メモリ容量を削減しつつ、処理の高速化を図る多面である。番号は辞書に現れてきた順とした。下記はその一部である。

1 地域
2 主義
3 一体化
4 統合
5 指向
6 とも
7 協力
8 協定
9 経済
10 技術
11 文化
. . .（以下省略）. . .

この番号となった名詞等は、同じく数値化された＜見出し語＞と併せて、数値に置き換えられた「現代用語の基礎知識 1998 」を構成する。「失楽園」は次のように数値ベクトル化される。

このように処理することで初めて、現実的な処理時間となっている。今回の実験では、プログラムは、Windows 上の Perl5 で書いた。一回の予測実験に、Pentium 500MHz 程

度のパソコンでは、数時間以上を要する。しかし、上記のようなハッシュ化を導入しなければ、さらに実験時間が増大したと推定される。なお、Perl ではなくて、C++ 等でのプログラム作成を行えば、より高速になると思われる。しかしながら、パタンマッチの強力さや、プログラム容易性から、Perl で実験を行うこととしている。

6363
(9,69,161,253,253,266,476,503,742,811,918,930,1052,1338,1544,1707,2098,2312,3211,3350,4268,4538,4850,5443,5445,6396,6396,6534,7944,8006,8006,8224,8239,8239,8264,8341,9000,10480,11443,11443,11462,11523,12602,13574,14796,15131,15529,15544,15651,15677,15723,16026,16329,17337,17622,19472,19553,19554,19554,19555,19556,19557,19558,19559,19560,19561,19562,19563,19564,19565)

以上