

現代用語辞書に基づく流行コンセプト予測手法

Concept Generation Support System using a Contemporary Word Dictionary

池田 定博 大橋 正和 金田 重郎

同志社大学大学院総合政策科学研究科

Sadahiro IKEDA Masakazu OHASHI Shigeo KANEDA

Graduate School of Policy and Management, Doshisha University

Generating a keyword/concept is one of the most important tasks in marketing activities. It is very difficult, however, to choose a best concept from many generated candidates written by copy-writers. To solve this problem, we are developing a new approach to choose best candidates by using a contemporary word dictionary and natural language processing technology. This paper clarifies, numerically and statistically, whether a chosen candidate reflects the social background or not. The statistical result shows this approach has a sense in social and economical fields. But it also shows this approach hardly has a sense in cultural fields.

1 研究の背景

流行語・流行コンセプト(以下、「流行語」と呼ぶ)は、CF(コマーシャルフィルム)の中で利用されたり、スローガンとして利用される。流行語の出来・不出来は、製品そのものの販売額・商品イメージに多大な影響を与えることがある。適切な流行語・流行コンセプトの選定は、企業にとって、今日、重要な課題である。

しかし、コピーライターが創生した流行語の中で、どれが「当たる」のかを知るのは容易ではない。コピーライターは自分自身の直感によって「当たりそうな」キャッチコピーを選ぶしかない。キャッチコピー(流行語候補)の善し悪しを客観的に判断する方法はなかった。また、CFを発注する側の企業にとっても状況は同一であり、どのコピーに社運をかけるかは主観的に判断するしかなかった。

上記問題点を解決するため、著者らは、時代を反映するテキストベースとして現代用語辞書を利用し、流行語候補群と用語辞書との意味的な距離を利用することにより、流行コンセプト作成を支援する『流行ことば予測手法』の研究を進めている[11, 12]。これまでの予備的実験によって、流行予測がある程度可能との結果を得ている。しかし、その統計的な検定は十分とは言えなかった。本稿では、著者らの『流行ことば予測手法』を紹介するとともに、結果の統計的分析を行い、課題を明らかにする。

2 用語辞書を用いた流行語予測

2.1 流行語と社会的背景との関係

本節では、著者らの『流行ことば予測手法』の概略を説明する。詳細は文献[11]を参照されたい。

2.2 「現代用語の基礎知識」と「流行語大賞」

「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。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、ひとつの定説になっている。しかし、この理論を、流行語の評価・生成に利用した試みは、著者らの知る限り存在しない。

本手法では、流行語の背景となる社会的要因を「ことば」として表現し、流行語とそのことばとの距離をベクトル空間法によって計測し、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。ここで、流行語の背景を表すことばの辞書(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」[1, 2]を用いた。

現代用語の基礎知識は、国語辞書にはない以下の特徴を持つ。1) 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。2) 毎

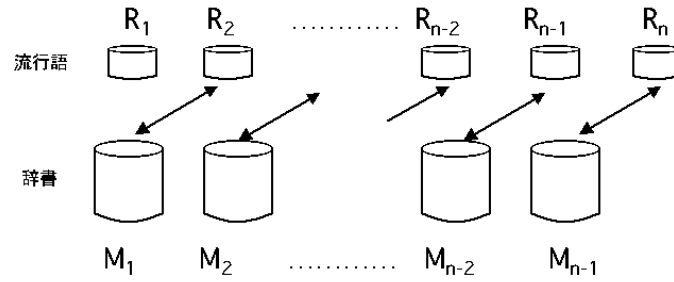


図 1: 流行ことば予測システムの構成

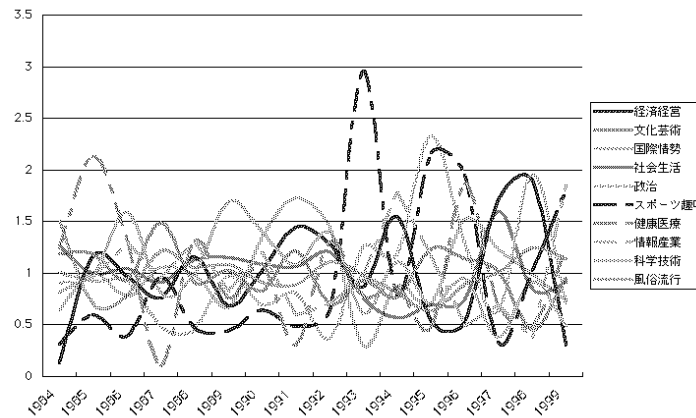


図 2: 最上位概念の流行語との平均距離年度推移（分野毎に平均値に対して正規化）

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。3) 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。

現代用語の基礎知識は、普段慣れ親しんでいる新しいことばに対応できることと、見出し語が多層の上位概念によりグループ化されているので、背景となる「ことば=社会的要因」の特定に有効と考えた。

2.3 流行ことば予測システムの概要

2.3.1 システム構成

著者らの流行ことば予測システムの概要を、図 1 に示す [11, 12]。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 ($i = 1, 2, 3, \dots, n-1$) 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、毎年の流行語を流行語 R_i ($i = 1, 2, 3, \dots, n$) に格納しておく。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10個から20個程度である。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。

流行語が影響をうけるとすると直前の年度の社会的状況と考えられるので、各流行語 R_i は、辞書 M_{i-1} との間で距離を計算される。 R_i 中の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。距離は、流行語の各々が持つ説明文と、辞書の見出し語の説明文との間で、意味ベクトル法 [3] を用いて計算する。

	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	1.000									
文化芸術	-0.593	1.000								
国際情勢	-0.080	-0.270	1.000							
社会生活	0.109	0.038	0.033	1.000						
政治	-0.097	-0.129	-0.257	-0.602	1.000					
スポーツ趣味	-0.277	0.066	-0.439	-0.657	0.571	1.000				
健康医療	-0.205	-0.086	0.236	0.060	-0.189	-0.103	1.000			
情報産業	-0.052	-0.023	0.184	-0.314	-0.081	-0.055	0.059	1.000		
科学技術	0.362	0.015	0.110	0.232	-0.363	-0.338	0.079	-0.384	1.000	
風俗流行	-0.142	-0.094	-0.053	0.393	-0.253	-0.268	-0.214	-0.402	-0.110	1.000

図 3: 10 種類の最上位概念相互の相関係数

辞書の見出し語は、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。最上位の階層は、「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」である。そこで、各流行語に近い見出し語を、近い見出し語から 20 個ずつ選択して、その最近隣 20 見出し語の属する上位概念が何であるかで、社会的な背景を探る手法を用いている¹。

2.4 システムを用いた実験

2.4.1 年度毎の推移

以上のシステムを用いて、1984 年から 1999 年までの間で、最上位概念の出現割合の推移を調査した。利用した辞書 M_i は全て 99 年版 [2] である²。結果を、図 2 に示す。推移に何となく周期があることは分かるが、明確な傾向は分からない。

本システムの目的は、各最上位概念が、今後伸びるか、落ちるかを予測することである。しかし、その妥当性の検定は難しいので、ここでは、相関係数に着目する。

2.4.2 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

¹例えば、ある年の流行語 R_i は 12 個あるとする。その各流行語から見て最近隣 20 見出し語は、240 見出し語ある。この 240 見出し語が有する最上位概念 240 個の中で、上記の最上位概念がどのような割合で含まれているかを調べ (10 概念分の合計で 100%)、割合が多い最上位概念ほど、 R_i 年の流行を支配していると判断する。

²本来、各年の現代用語の辞書を設定すべきであるが、全ての現代用語の基礎知識を準備できないためである。その意味では、図 2 は予備実験の結果である。

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。

3 距離計算 (意味ベクトル法) の評価

3.1 共通要素単語の分析

図 3 の最上位概念 (分野) 毎の上がり下がりを行推定に利用するためには、少なくとも、意味ベクトル法で距離が近くなる原因である、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) が人間から見ても妥当である必要がある。そこで、以下の分析を行った。

- 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 4、図 5 に示す。

3.2 実験結果分析

「経済経営」分野の結果 (図 4) では 1990 年から 1993 年にかけて、単語「バブル」が登場している。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。

	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 6: 無相関検定結果

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。即ち、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、常にそのように上手く行くとは限らない。「国際情勢」分野の結果（図 5）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。

しかし、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。また、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。「国際情勢」の分野に関しては、機械的に判定されたことばと人間の想像の範囲とは一致しない可能性がある。

総じて言えば、「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、人間の直感にあわない単語が、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言葉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、更に統計的に確認する必要がある。

4 実験結果の統計的分析

本システムでは、流行を数値化して捉え、各分野（最上位概念）の動向を、予測に試みようとしてい

1990株	年	バブル	経済	状況
・	9	6	5	5
1991会社	証券	株	投資家	株価
・	13	12	11	10
1992経済	バブル	崩壊	年	不況
・	15	12	11	7
1993価格	バブル	会長	行使	手
・	4	3	3	3
1994価格	物価	下落	商品	制度
・	13	9	7	5
1995年	3			
・				
1996				
・				
1997証券	改革	バン	ビッグ	売買
・	11	10	9	9
1998金融機関	債権	金融	経営	年
・	25	23	18	12
1999				11
・				

図 4: 「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
1991年	派	革命	月	事件
・	11	5	4	4
1992大統領	年	勝利	選	国民
・	17	13	9	9
1993日本				8
・	3			
1994思想	社会	主義	人	党
・	5	4	4	4
1995政党	選挙	候補		
・	6	5	3	
1996首相	月	政治	社会	英語
・	5	4	4	3
1997				2
・				
1998年	共産党	首相	選	委員長
・	6	5	5	4
1999なら	選挙			3
・	3	3		

図 5: 「国際情勢」の上位 5 位までの要素

上限	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	-0.116	***								
国際情勢			***							
社会生活				***						
政治				-0.130	***					
スポーツ趣味				-0.218	0.838	***				
健康医療							***			
情報産業								***		
科学技術									***	
風俗流行										***

図 7: 真の相関係数の区間の上限

下限	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	-0.648	***								
国際情勢			***							
社会生活				***						
政治				-0.852	***					
スポーツ趣味				-0.875	0.083	***				
健康医療							***			
情報産業								***		
科学技術									***	
風俗流行										***

図 8: 真の相関係数の区間の下限

る。前述のように、本来、各分野毎の動きについて予測可能性の検定を行うべきであるが、サンプル数が少なく、技術的に難しい。そこで、相関係数を統計学的に検証する。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考えられるからである。

4.1 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 15 である。各最上位概念を無限の母集団とみなし、その母集団の相関関係を推定した。

最初に、「無相関検定 [13]」を行った。即ち、「無相関ではない」と判定されたものに限り、その母集団の真の相関係数の区間を推定した。「真の相関係数の区間推定」とは、ある有限個数のデータを元にして割り出した相関係数から、その集団が無制限個数であった場合に、どれだけの範囲の相関係数を取りうるか、を推定することである。

4.2 区間推定手法

具体的な方法は以下の通りである。

- まず、「母集団の相関係数は 0 である」という仮説を立てる。(当然、その逆は「母集団の相関係数は 0 ではない」である。)

- つぎに、標本数と標本の相関係数から、検定統計量を計算し、それを元にして有意確率 P を算出する。
- P 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。

なお、検定統計量や有意確率の計算式はここでは省略する。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 [13, 14] を利用している。有意水準は 1% と 5% である。結果を図 6 に示す³。

有意水準 1% で検定した結果、有意性が認められるのは 45 個中わずかに「スポーツ趣味」=「社会生活」の 1 個だけであった。有意水準 5% にゆるめると、4 個が有意となる。これらは厳しい結果である。

4.3 真の相関係数の区間推定

では、相関係数に有意性が確認された組み合わせにおいて、真の相関係数の値はどれくらいになるのだろうか。実験によって、相関係数が 0 になることはほとんどないので、その相関が強い弱いを検証する。ここでは、元の集団の標本数とそこから計算された相関係数をもとに、母集団の相関係数

³ この値が、有意水準 5% の時には 0.05、有意水準 1% の時には 0.01 を下回れば、「無相関でない」と結論づけられる。

が95%の確率で取りうる範囲（信頼区間 95%）と、80%の確率で取りうる範囲（信頼区間 80%）を求めた。計算式は、かなり複雑なのでここでは割愛する。信頼区間 95%における結果を図 7 と図 8 に示す。

図 7 と図 8 の「スポーツ趣味」=「社会生活」の組み合わせを参照されたい。この組み合わせは図 2 唯一有意水準 99% 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-0.657 と、強い負の相関である。しかし、相関係数の範囲は信頼区間 95%だと-0.218 から-0.875 である。これをどう解釈するかは難しいが、少なくとも、この分野に限っていえば、本システムの手法が、統計的にも裏付けられたと言ってもよいように思われる。

5 まとめ

上記分析結果から以下のことが言える。

- 意味ベクトル法の共通要素単語の分析からも、相関係数検定から見ても、「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来た「流行語の背景には社会経済的影響がある」との見解を定量的に裏付けている。従って、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えない面がある。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言えば、別のアプローチが必要と思われる。今後は、個々の分野毎の統計的妥当性についても更に検討してゆくとともに、予測はできなくても、「あたりそうにないものを外す」方法の開発も必要と思われる。
- 但し、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、きつめの評価である。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今後は確認してゆく必要がある。

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。本研究に関して支援を頂いた、北寿郎部長、大山芳史主幹研究員を始めとする NTT コミュニケーション科学基礎研究所各位に深く感謝の意を表します。本研究の一部は、学術フロンティア共同研究「知能情報処理科学とその応用」による。

参考文献

- [1] 1998 年版現代用語の基礎知識、自由国民社
- [2] 1999 年版現代用語の基礎知識、自由国民社
- [3] 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
- [4] 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.
- [5] 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.
- [6] 松澤和光、湯川高志、笠原要、藤本和則「アバウト推論技術」NTT R&D Vol.45 No.11,1996
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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(パターンの認識・メディア理解研究会と合同、PRMU99-276,pp.113-120,2000).
- [12] 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.
- [13] <http://aoki2.si.gunma-u.ac.jp/lecture/Corr/corr.html>
- [14] <http://aoki2.si.gunma-u.ac.jp/lecture/Corr/corr3.html>