多義動詞の意味分類における共起情報抽出手法の検討

横山 弘子　大町 真一郎　阿曽 弘具

東北大学大学院 工学研究科
〒980-8579 仙台市青葉区荒巻字青葉 05
TEL 022-217-7088　　E-mail yoko@aso.ecei.tohoku.ac.jp

あらまし　近年，テキストコーパスから共起情報を抽出し，語の統計的・表層的ふるまいにより意味知識を獲得する研究が行われている．本論文では，単言語コーパス内の単語の共起情報をもとに，クラスタリングにより日本語多義動詞の意味分類を行なう手法を提案する．これは，広範囲の文脈に着目し共起情報を取得し，また動詞の意味を決定する要因となる格として出現する名詞以外の名詞との共起情報も利用するという特徴を持つ．提案手法を既存の手法と比較し検討した結果，上記２つの点が意味分類に有効であることが示された．

キーワード　クラスタリング，多義動詞，共起情報

Evaluation of Methods to Extract Collocational Information from Corpus for Semantic Clustering of Japanese Polysemous Verbs

Hiroko Yokoyama　Shin'ichiro Omachi　Hirotomo Aso

Graduate School of Engineering,
Tohoku University,
Aoba 05, Aramaki, Aoba-ku, Sendai, 980-8579, Japan
TEL (+81)22-217-7088　　E-mail yoko@aso.ecei.tohoku.ac.jp

Abstract　Recently, many statistical approaches for acquiring knowledge of word meaning, especially using collocational information from text corpora, have been investigated. In this paper, we propose a semantic clustering method of Japanese polysemous verbs by collocational information obtained from Japanese monolingual corpus. This method has two features. One is to observe global contexts and the other is to utilize words not only arguments(case fillers) of a verb, but also utilize other nouns co-occur with the verb to get collocational information. This paper compares our method with other previous approaches and evaluates their results, and the results indicate our method is the most effective.

key words　clustering, polysemous verb, collocational information

1. はじめに

単語を意味 (語義) で分類し，体系化することは自然言語処理における多義語の曖昧性解消問題や，大量の文書の意味内容による分類の手がかりとなるものとして重要である．語を意味分類し体系化したいわゆる意味辞書といわれるものは，従来，人手により作成されたものが多かったが，この方法では労力やコストがかかること，作成者による主観が反映され均質なものができないという問題点が指摘されていた．この問題に対し，近年，テキストコーパスから共起情報を抽出し，語の統計的・表層的ふるまいにより意味知識を獲得する研究が行われている．筆者らは，既に，日本語多義動詞を対象として，コーパスにより自動的に語の意味分類を行う実験を試みている [7].

コーパスを用いた単語の意味分類研究としては，事例文内の動詞の項 (arguments) あるいは表層格に着目しそこに出現する名詞 (句) と動詞の共起情報によって動詞や名詞の意味を区別しようとする試みがある [1],[2],[3],[4]．これらはいずれも動詞の出現する事例文内のみでの共起情報を利用し，さらに，その中でも動詞の項 (arguments) あるいは表層格に出現する名詞との共起情報のみ取り出している．一方，意味や特徴を考慮した，高度な文書分類／検索を念頭においた研究においては前後100語という比較的広い文脈での単語間共起情報を表現するベクトルを用いて文書の構成要素となる単語の意味を表現することを考察した研究があるが [5]，主眼としているのは文書レベルの分類である．また，単語の多義解消問題を行うという，単語に主眼をおいた研究においても，対象単語の前後100語に関する共起情報から意味を同定する実験を行っているものがある [6]．この研究では92 %の精度で多義語の曖昧性解消が行なわれているが，ロジェシソーラスのカテゴリ情報という外部知識を利用している．

筆者らは，[5]，[6] の結果から，[1]～[4] のような事例文内のみでなく比較的広い文脈内での単語間共起傾向が意味の決定に重要であると考え，[5] で提案された単語間共起情報をもとに，単語の意味分類を行うことを目的とする．さらに，人手による作成のもたらす問題を避けるため，外部知識を使わず単言語コーパスのみを利用し

- 動詞の出現する事例文の前後の文脈にも着目すること
- 格に出現する単語以外にも着目すること

により得られる単語の共起情報から，「1 つの意味を 1 つのクラスタに対応づける」ようなクラスタリングによって多義動詞を自動的に意味分類する手法を提案する．

文献 [7] では，いくつかの動詞について良好なクラスタリング結果が得られることを報告した．しかしコーパスの規模が小さかったという点から十分な共起情報が得られず，意味に応じたクラスタが形成されない動詞もあった．

そこで本稿では，より規模の大きいコーパスを用いて実験し，さらに，提案した手法の 2 つの特徴が意味分類にどの程度有効であるのか検証しその結果を考察する．

2. 意味分類手法

本章では，1 つの多義動詞に対して，共起情報をもとに動詞の意味に応じたクラスタリングを行なう手法を説明する．なお詳細は [7] に説明したためここでは概略を説明する．

2.1 クラスタリング用単語の選択

対象となる多義動詞 vの意味分類に用いる単語として名詞に着目し，vに対しそのクラスタリングに用いる名詞集合 N_vを以下の手順により選択する．

以下，多義動詞 vの意味を区別するのに重要と思われる格を着目格と呼ぶ．

(1) 与えられたコーパスにおいて，動詞 vが着目格を伴なって出現する文 (事例) を含む話題を複数収集して文書集合を作成する．これを S_vとする．

(2) S_vに出現する全名詞 (普通名詞・固有名詞・時相名詞・サ変名詞のいずれか) のうち，出現頻度の分散がしきい値θ_1より大きくかつ総出現回数がしきい値 c_{max}より少ない名詞の集合を候補名詞集合 Nとし，これに vと vの着目格に出現した名詞を加えた集合を共起情報抽出用集合 Wとする．

(3) S_v中の k番目の文脈 (以下文脈 k) を k番目の文とその前後 num(正の整数値) 個の文とからなる集合と定義する．ある文脈 kに対し共起ベクトル $\boldsymbol{c}_k$を

$$\boldsymbol{c}_k = (c_{k1}, \ldots, c_{ki}, \ldots, c_{km})$$

と定義する． mは Wの要素数，c_{ki}は W中の単語 iの文脈 kでの出現回数である．共起ベク

トルはS_vの全文の数(以下nとする)分作成される.

(4) 共起ベクトルより相関行列$C(=[\rho_{ij}], m\times m$行列) を作成する.$\rho_{ij}$は単語$i$と単語$j$の共起度を表わすものといえる.

(5) (4) の結果から,動詞vとの相関係数がθ_2以上かつ着目格に出現した名詞のうち少なくとも1つの名詞との相関係数の絶対値がθ_3以上の名詞をWから取り出す.着目格に出現した名詞は必ず取り出す.これをN_vとする.

2.2 クラスタリング

N_v中の名詞iと名詞jの距離d_{ij}を

$$d_{ij} = 1 - \rho_{ij}$$

と定義し,群平均法によりクラスタリングする.

3. 実験

3.1 実験条件

(1) 使用コーパス
コーパスとして [7] で用いたコーパス (朝日新聞社説'91年度分) よりも規模が大きい毎日新聞記事CD-ROM版'91年度分を用いた.このコーパスにおける1つの記事を,1つの話題に対応させた.このコーパスの記事フォーマットは表1のようなタグとそのうしろにタグに対応する内容が続いた形式になっている.

タグ	内容
ＩＤ	ＩＤ番号(ユニーク)
ＡＤＤ	記事番号(ユニーク)
ＡＥ	掲載日付,ページ
ＡＦ	索引記事番号
ＤＴ	掲載日付
ＰＧ	掲載ページ
ＺＵ	写真,図の有無
ＭＥ	掲載面種別コード
ＴＩＮ	記事見出し
ＴＩ,ＱＴＩ	記事見出しキーワード(カナヨミ)
ＴＩＫ,ＱＴＩＫ	記事見出しキーワード(漢字表記)
ＨＯＮ	記事本文
ＨＯ１,ＱＨＯ１	本文キーワード(カナヨミ)
ＨＫ１,ＱＨＫ１	本文キーワード(漢字表記)
ＣＦ	著作権フラグ

表 1. 毎日新聞記事 CD-ROM 版'91 年度分フォーマット

この中で,1つのＨＯＮタグには1文以上の文が対応している.本実験では,6個以上のＨＯＮ部分を持つ記事を用いることにした.これを満たす記事は20114記事あった.これは,ある程度の文数を持つ記事でないと十分な共起情報が得られないと考えたためである.

(2) 分類する意味の数
今回の実験においては,2つの意味の分割を考える.多義動詞が3つ以上の意味を持ち,かつ3つ以上の意味に対応する事例が出現した場合,特に出現回数の多い2つの意味に対応する事例が出現する記事のみを選び,文書集合を作成する.

(3) 文書集合中の記事数
使用するコーパスの先頭から,2.1の (1) に述べた条件に適合する記事を,先頭から最大100記事取り出し文書集合とする.

(4) 実験対象動詞及び着目格
多義動詞「積む」を実験対象とし,着目格は「を」格とする.辞書 [8] によるとその意味 (語義) は3つあり,その全ての意味に対応する事例が使用コーパスに出現したが,既述したようにそのうち主として出現した2つの意味の分類実験を行う.表2に,辞書 [8] に示される「積む」の意味のうちクラスタリングに用いた2つの意味と,各々の意味に対応する「を」格に出現した名詞の例を示す.

	意味	格に出現した名詞例
積む1	搭載する (荷を積む等)	テント,寝袋, 灯油等 (計60種類)
積む2	繰り返す (訓練を積む等)	経験,実績, 修業等 (計16種類)

表 2. クラスタリングに用いた意味

(5) パラメータの値
本実験での集合N_vを求める過程における各パラメータは予備実験により以下のように定めた.
$\theta_1 = 0.02$, $\theta_2 = 0.1$, $\theta_3 = 0.4$,
$c_{max} = 50$, $num = 5$.
なお, $m = 2659$, $n = 3339$ であった.

3.2 結果の評価方法

結果の評価は,作成されたクラスタが辞書 [8] で挙げられている語義 (表2) にどの程度対応しているか見ることで行なう.具体的には,クラスタリン

グによって2つのクラスタに分割した場合に，表2での「積む1」と「積む2」の意味に対応付けられる格に出現した名詞(「積む1」ならテント等，「積む2」なら経験等)が各々のクラスタの要素にどの程度排他的に振り分けられているかで評価する．定量的に評価するため，便宜上「積む1」に対応するクラスタをC_1,「積む2」に対応するクラスタをC_2とし,「積む1」に対するクラスタ誤り率$err1$,「積む2」に対するクラスタ誤り率$err2$を下式のように定義する．この値が小さい程，意味に応じたクラスタが形成されたと評価することにする．

$$err1 = \frac{C_2\text{に含まれた「積む1」に対応する名詞の数}}{\text{「積む1」に対応する名詞の数}}$$

$$err2 = \frac{C_1\text{に含まれた「積む2」に対応する名詞の数}}{\text{「積む2」に対応する名詞の数}}$$

4. 比較のための対照実験

提案手法との比較のため，提案手法による実験に加えて以下の4種類の実験を行う．

4.1 着目する文脈を出現動詞の事例文内に限定したクラスタリング

提案手法の特徴の1つである「動詞の出現する事例文の前後の文脈にも着目すること」に対する比較のための対照実験として，文脈を出現動詞の事例文内に限定したクラスタリング実験を以下の2通り行う．これは，提案手法での文脈の範囲$num = 0$として実験したものである．それ以外のパラメータは提案手法で用いたものと同じ値とした．

(1) 着目格を「を」格のみにした場合(対照実験1)

(2) 着目格を複数の格にした場合(対照実験2)
動詞の持つ複数の格に出現する名詞に着目して意味分類を行う研究との比較も考慮し，着目する格として「を」格以外に以下の6つの格も加えた計7種類の格を着目格とした実験を行った．

「が」,「に」,「へ」,「から」,「まで」,「で」

4.2 共起名詞集合を格に出現する名詞に限定したクラスタリング

提案手法のもう1つの特徴である「格に出現する単語以外にも着目すること」に対する比較のための対照実験として格に出現する名詞のみ用いたクラスタリング実験を以下の2通り行う．これは提案手法での共起情報抽出用集合Wを，動詞vと着目格に出現する名詞のみから成る集合として実験したものである．文脈の範囲は提案手法と同じく$num = 5$とした．(これ以外のθ_1，θ_2，θ_3，c_{max}は条件から不用である)．

(1) 着目格を「を」格のみにした場合(対照実験3)

(2) 着目格を複数の格にした場合(対照実験4)
4.1と同様の理由で，複数の格を着目格とした実験も行う．

5. 実験結果

図1～図5に，提案手法及び4種類の対照実験によるクラスタリング結果をデンドログラムにより示す．各図のクラスタ木は左から右へ成長している．図の上方の数値は，クラスタ併合の順番を表すもので，クラスタ間の実際の距離を表わすものではない．すなわち，i番目の併合を表す節点を位置iに置くように描いている．クラスタ併合の順序を見やすくするためにこの方法をとった．また,「積む1」の意味に対応する「を」格に出現した名詞は白楕円で囲い,「積む2」の意味に対応する「を」格に出現した名詞には斜線の楕円を重ねた．さらに，2つのクラスタに分割した場合に「積む1」に対応するクラスタと判断したものにC_1のラベルを,「積む2」に対応するクラスタと判断したものにC_2のラベルを付けた．

表3に各実験に用いたN_vの要素数を示す．

手法	N_vの要素数
提案手法	158
対照実験1	80
対照実験2	111
対照実験3	76
対照実験4	103

表 3. 各クラスタリング実験でのN_vの要素数

6. 考察

各クラスタリング結果のクラスタ誤り率を表4に示す．

クラスタリング結果	クラスタ誤り率	
	$err1$	$err2$
提案手法	0.05	0.19
対照実験1	0.00	0.75
対照実験2	0.00	0.75
対照実験3	0.15	0.50
対照実験4	0.57	0.44

表 4. クラスタリングに用いた意味

提案手法と各対照実験での誤り率から結果を評価する．

- 対照実験1及び対照実験2との比較
  文脈範囲を事例文内に限定したクラスタリング結果である対照実験1及び2と提案手法において，$err1$(「積む1」に対応するクラスタの誤り率) は提案手法が0.05，対照実験1及び2では0.00である．一方$err2$(「積む2」に対応するクラスタの誤り率) は提案手法の値が0.19であるのに対して対照実験1及び2の値は0.75であり対照実験1及び2の値が極端に大きい．

  対照実験1及び2では文脈範囲が狭い($num = 0$)ため，今回使用したコーパスでは，特に，提案手法で多数抽出された「積む1」の意味に対応する格に出現した名詞を1つにまとめる名詞群 (格に出現した名詞以外の名詞) が対照実験1及び2では得られなかった (実際，表3に示したように，対照実験1及び2でクラスタリングに用いた名詞の数が提案手法よりも少ない)．また「積む2」の意味に対応する格に出現した名詞群も文脈範囲が狭いため共起情報が十分に得られず，多数が「積む1」クラスタに含まれてしまい，2つの意味に対応するクラスタが分割できなかった．結果として$err2$の値が非常に大きくなり，一方「積む1」の意味に対応する格に出現した名詞群は1つにまとまったため$err1$の誤りがゼロになったと考えられる．作成されたクラスタ全体としてみると，提案手法の結果の方が良好であると言える．

- 対照実験3及び対照実験4との比較
  クラスタリングに用いる集合を，格に出現した名詞に限定したクラスタリング結果である対照実験3及び4と提案手法においては，$err1$, $err2$とも提案手法での値が小さく，提案手法の方がより良い結果が得られた．

  対照実験3及び4は，文脈範囲は広く取ったが，クラスタリングに用いる名詞は格に出現する名詞に限定したものであった．このため，表3に示したように，対照実験3及び4においても提案手法に比してクラスタリングに用いた名詞の数が少ない．しかし提案手法では格に出現した名詞に限らず，それらと共起度の大きい名詞群もクラスタリングに用いるようにしている．従って，対照実験3及び4では「積む1」「積む2」のクラスタの形成に十分な名詞を用いていないため，誤り率が大きくなったと考えられる．

全体として，提案手法では，その特徴である

◇ 動詞の出現する事例文の前後の文脈にも着目すること

◇ 格に出現する単語以外にも着目すること

によって「積む1」,「積む2」の意味に対応する格に出現する名詞 (今回の実験では「を」格) をクラスタとしてまとめる働きをする名詞群をクラスタリング用の名詞として取り出すことができ，結果として提案手法による分類が多義動詞「積む」の2つの意味を良く分割し，対照実験より良好な結果となったと言える．

7. 結論及び今後の課題

単言語テキストコーパスから抽出した共起情報を利用して多義動詞の意味分類をクラスタリングにより行う手法を提案した．

本報告では，単語の意味分類に対し，提案手法の有効性を確認するため，多義動詞「積む」を対象として既存の方法との比較実験を行った．結果として提案手法の特徴である2点が意味分類に有効であることが確認された．

今回の実験では比較のため，共起情報としては相関係数を用いる方法で統一したが，既存のアプローチでは相互情報量を用いるなどの手段がとられるものもある．これを考慮し，今後，提案手法でも，どのような共起情報を用いるのがより有効であるか検討する予定である．

また，今回は1つの多義動詞の意味分類実験に限定したが，この手法を発展させ複数の動詞に対して行うことで，動詞間の意味の類似関係を体系化したシソーラス的なデータを構築することも考えている．

参考文献

[1] Hindle.D,"Noun Classification from predicate-argument structures," 28th Annual Meeting of the Association for Computational Linguistics, Morris Town,New Jersey, pp.268–275, 1990.

[2] 福本文代，辻井潤一，"コーパスに基づく動詞の多義解消"，電子情報通信学会技術研究報告,NLC94-24,pp15–22,1994.

[3] 内元清貴，宇津呂武仁，長尾真，“動詞の語彙的知識獲得における類義語の用例を用いた多義性の類別”，情報処理学会研究報告,NL101-14,pp105–112,1994.

[4] 平岡冠二，松本裕治，“コーパスからの動詞の格フレーム獲得と名詞のクラスタリング”, 情報処理学会研究報告,NL104-11,pp79–86,1994.

[5] Schütze.H,“Dimensions of meaning,”, Proc. of supercomputing, pp.787–796,1992.

[6] David Yarowsky,“Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained On Large Corpora”,In Proceedings of COLING-92, pp454–460,NANTES,France,1992.

[7] 横山弘子，大町真一郎，阿曽弘具，“コーパスからの共起情報を用いた日本語動詞の意味分類”，電子情報通信学会技術研究報告,NLC-97-55,pp1–8,1998.

[8] 金田一京助他，“新明解国語辞典 第5版”, 三省堂,1997.

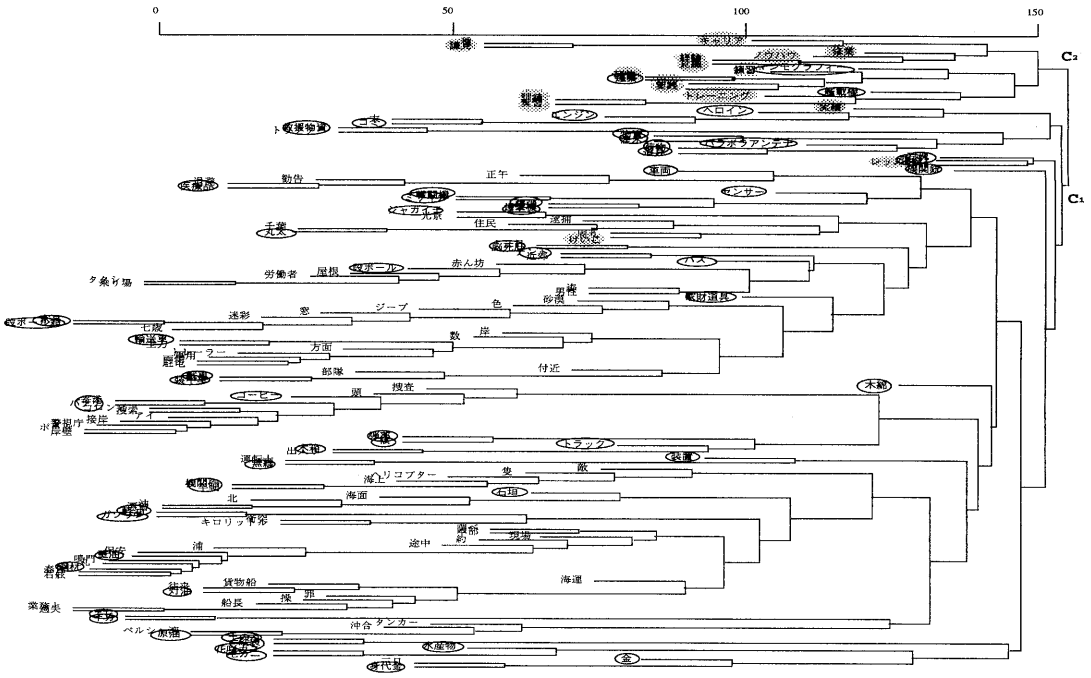

図 1. 提案手法によるクラスタリング結果

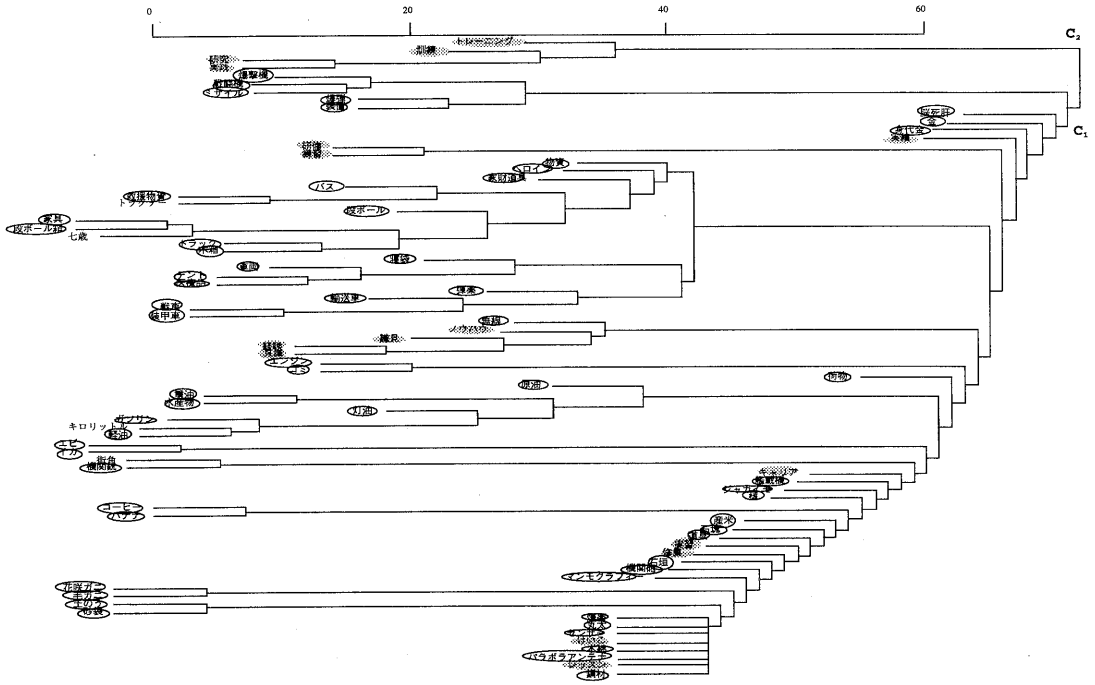

図 2. 対照実験 1 によるクラスタリング結果

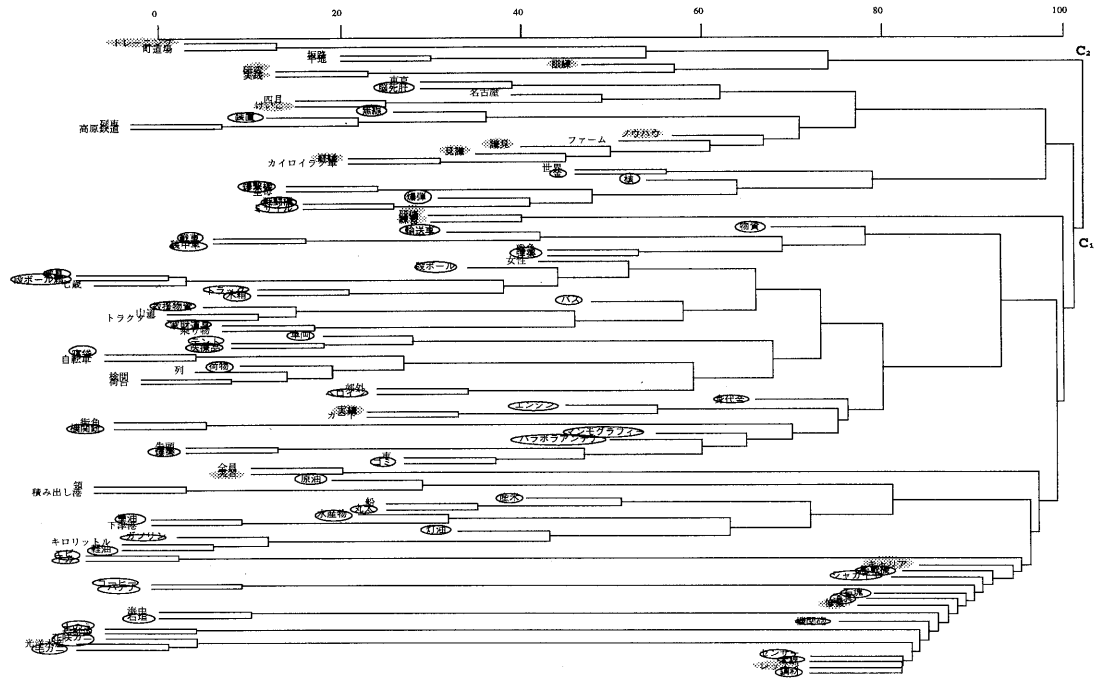

図 3. 対照実験 2 によるクラスタリング結果

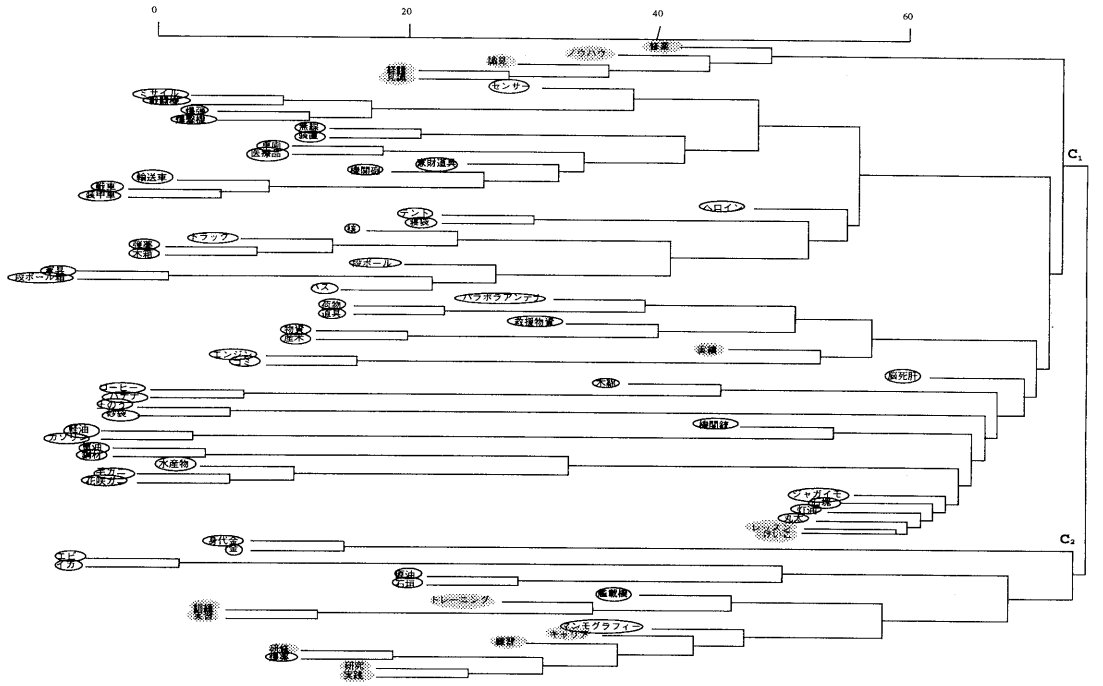

図 4. 対照実験 3 によるクラスタリング結果

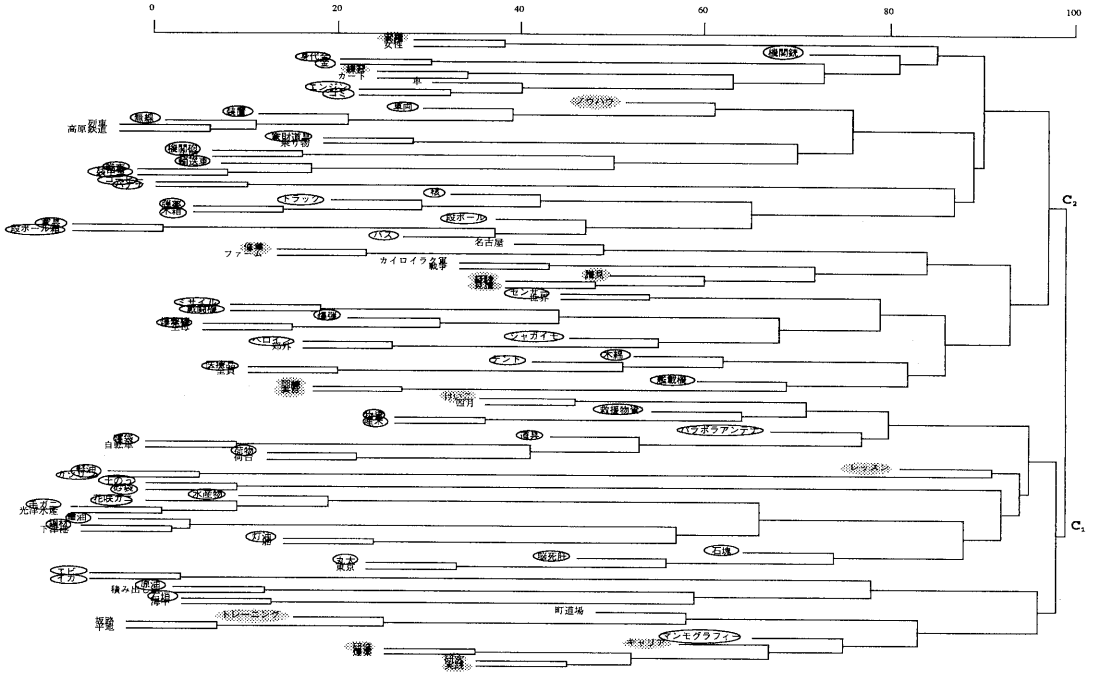

図 5. 対照実験 4 によるクラスタリング結果