同表記異義の処理とその情報検索への応用

阿部 賢司　片見 憲次　武田 和也　藤崎 博也

東京理科大学

1. はじめに

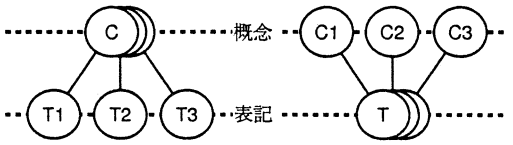
従来のキーワード検索では、語の表記のみに着目して検索するため、キーワードに異表記同義が存在する場合には検索洩れが生じ、また、同表記異義が存在する場合には不要な検索が生じる [1-3]。これらのうち、前者の異表記同義に関しては、表記概念対応辞書を参照して、同概念のキーワードを検索式に追加することにより検索洩れを軽減させることができる [4]。一方、後者の同表記異義に関しては、表記-概念対応辞書を用いることにより、表記に対応し得る概念を推定することはできるが、文書（データ）中で、そのキーワードがどの概念で用いられているかを明らかにしなければ、不要な検索を回避することはできない。したがって、同表記異義の関係にある語の概念を特定することの必要性は極めて高い。

このような観点から、我々は、同表記異義の現象を定量的に把握することを目的とし、特に、学術情報検索における同表記異義の現象に着目して、学術論文のキーワードから同表記異義の実例を収集した [5]。本報では、これらの同表記異義を処理するための一方法として、階層的クラスター分析の手法を用いて同表記異義のキーワードを含む論文をクラスタリングし、同表記異義の関係にあるキーワードの概念を特定する方法について検討した結果を述べる。また、この方法を実際の情報検索に応用し、その有効性を検証した結果について述べる。

2. 異表記同義・同表記異義の定義

本報では、1つの語は、1つの表記と1つの概念から構成されると考える。このとき、複数の語が概念のレベルで縮退する現象を異表記同義の現象と呼び (図1(a))、逆に、複数の語が表記のレベルで縮退する現象を同表記異義の現象と呼ぶ (図1(b))。

なお、語の表記とは、文字言語の場合には文字を、音声言語の場合には音声を意味するが、ここでは、文字言語の場合について議論する。



(a) 異表記同義　検索洩れをもたらす　　(b) 同表記異義　不要な検索をもたらす

図1. 異表記同義・同表記異義が存在する場合の表記と概念との関係

3. 同表記異義の収集・分類

同表記異義の現象を定量的に把握するため、学術情報センターから提供される情報検索システム評価用テストコレクション1 [6] に含まれる学術論文の日本語キーワード (英略語も含む) の中から同表記異義の実例を収集した。その結果、全キーワード 132,464 語のうち 0.36%のキーワードに同表記異義の現象が存在することを確認した。また、それらのキーワードを、表記に着目して以下の4種類に分類し、それらの出現頻度を調べた結果を表1に示す。

(1) 英略語：例．PC
　Personal Computer / Programmable Controller
(2) 平仮名：例．あそび
　遊戯 (play) / 機械のゆとり (clearance)
(3) 片仮名：例．アルバイト
　曹長石 (albite) / アルバイト (part-time work)
(4) 漢字：例．株
　切り株 (stump) / 株式 (stock)

表1 同表記異義の分類と出現率

	出現率 [%]
(1) 英略語	90.0
(2) 平仮名	1.2
(3) 片仮名	4.0
(4) 漢字	4.8

表1からも明らかなように、同表記異義の現象は英略語に最も多くみられる。このことから、学術論文に関しては英略語の同表記異義を処理する必要性が最も高いといえる。

4. 同表記異義の処理方法 [7, 8]

同じ概念の単語が用いられている論文同士は、その共起情報も類似していると考えられる。したがって、本研究では、論文中の単語に同表記異義の現象が存在する場合、共起情報を利用してその概念を推定する。

具体的には、まず、共起情報に基づいて、検索空間における論文の位置をベクトルで表す。次に、そのベクトルに基づいて、論文間の距離を求め、それを論文間の類似度を表す指標とする。さらに、階層クラスター分析の手法を用いて、距離が近い論文同士から階層的にリンクさせる。ここで、ある距離を閾値とし、その段階でのクラスタリングの結果を同表記異議の分類結果とする。なお、距離の求め方とリンクの方法にはいくつかの種類があるが、本研究では、距離には内積を、リンク法には最短距離法を用いた。

4.1 共起情報からのベクトルの生成

まず、ある一つの単語T_0に着目して、検索対象となる全論文中の共起単語を$T_1, T_2, \cdots, T_i, \cdots, T_m$とする。ここで、ある論文$A$の位置を表すベクトル$\vec{a}$を次式(1)のように表す。

$$\vec{a} = (w_1 x_1, w_2 x_2, \cdots, w_i x_i, \cdots, w_m x_m) \qquad (1)$$

式(1)のx_iは共起単語T_iの有無を表す値であり、T_iが存在すればx_iの値を1とし、存在しなければ0とする。また、w_iはT_iの重要度を表す重み係数であり、重要な共起単語ほどw_iの値を大きく設定すべきである。

本研究では、共起単語として、学術論文におけるキーワード(KW)、著者名(AU)、学会名(SO)、アブストラクト中の名詞(AB1)・未知語(AB2)、タイトル中の名詞(TI1)・未知語(TI2)の7種類を用いた。ここで、アブストラクトおよびタイトル中の単語については、日本語形態素解析システム「茶筌」[9]を用いて形態素解析を行い、そのうち名詞および未知語と表記してあるもののみを利用した。

4.2 重みの設定

クラスタリングの結果は、重み係数w_iの設定により変化する。ここで、w_iの適切な値を調べるため、共起単語の種類に着目して、w_iを表2の設定にしたがって変化させ、それぞれの場合について、同表記異議の単語を含む論文をクラスタリングする実験を行った(同表記異議の単語としては、10種類の英略語を用いた)。その結果、重み係数を共起単語の種類にしたがって、KW:1, AU:1, SO:0.1, AB1:0.1, AB2:0.9, TI1:0.1, TI2:0.9と定めた場合の分類精度が最も高かったため、以下の実験では、重み係数としてのこれらの値を用いた。

表2 重み係数の設定

共起単語の種類	重み
KW	1
AU	1
SO	0, 0.1, $\cdots$, 1
AB1	0, 0.1, $\cdots$, 1
AB2	0, 0.1, $\cdots$, 1
TI1	0, 0.1, $\cdots$, 1
TI2	0, 0.1, $\cdots$, 1

4.3 情報検索への応用方法

クラスタリングの結果は図2に示すようなデンドログラムを用いて表すことができる。この図の①～⑧は同表記異義の単語を含む論文の番号を表し、また、その同表記異義の単語は、論文①, ②においては概念1、論文③～⑤においては概念2、論文⑥～⑧においては概念3の意味で用いられている。また、図の縦軸は距離を表し、リンクしたときの距離が小さいほど論文間の類似度は高いとみなすことができる。以下、この図を用いて、クラスタリングの結果を情報検索に応用する方法(方法1、方法2)を説明する。

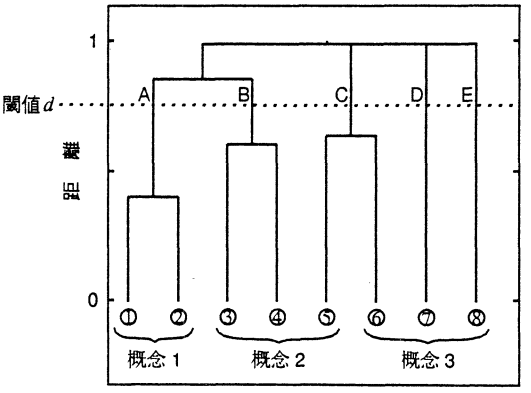


図2. デンドログラムの例

[方法1]

図2の例において、閾値dにおける分類結果を利用する場合、システムは、論文①～⑧がA～Eの5つのクラスに分類され、各論文に含まれる同表記異義の単語の概念はクラス毎に異なることをユーザに示す。このとき、ユーザは、その単語を自分の要求する

概念で用いている論文(クラス)を1つ選択できるものとする。例えば、ユーザが概念1を要求している場合はAを選択し、概念2を要求している場合はB、Cのいずれかを選択し、概念3を要求している場合はC、D、Eのいずれかを選択する。ここで、簡単のために、ユーザが要求するものが概念1～概念3である確率は等しいものとする（この場合$\frac{1}{3}$）。また、概念2を要求している場合にBあるいはCを選択する確率は共に等しく$\frac{1}{2}$とし、同様に、概念3を要求している場合にC、D、Eのいずれかを選択する確率はいづれも$\frac{1}{3}$とする。

ここで、ユーザが概念3を要求し、かつ、Cを選択した場合には、同表記異義の単語を概念3の意味で用いている論文⑥,⑦,⑧のうち、⑦,⑧の2つが検索されていないため、検索洩れ率は$\frac{2}{3}$となる。また、検索した論文⑤,⑥のうち、⑤は不要なものであるため、不要検索率は$\frac{1}{2}$となる。このように、ユーザの要求するものが概念1～概念3であると想定した各々の場合において求めた検索洩れ率・不要検索率を平均することにより、閾値dにおける分類を情報検索に利用したときの平均的な検索洩れ率・不要検索率を求めることができる。したがって、検索洩れ率・不要検索率を最も小さくするような閾値dを求め、そのときの分類結果をユーザに提示することにより、検索洩れ率・不要検索率を軽減することができる。

[方法2]

方法1では、分類された論文における同表記異義の単語の概念は、互いに異なるとみなす。しかし、閾値dにおいて他のいずれともリンクしていないもの(図2の⑦,⑧)は、実際に共起単語が他のいずれとも異なるのか、あるいは、共起単語の不足によりリンクできなかったのかを判定することが難しい。したがって、検索洩れを軽減することを最優先とし、閾値dにおいて他のいずれともリンクしていない論文は必ず検索する。例えば、図2の閾値dにおいてユーザがAを選択した場合には、Aの他に、DとEも検索する。

5. 提案手法の評価実験

前節の方法に従い、同表記異義の現象が存在する単語をキーワードとした場合の検索洩れ率・不要検索率を求める実験を行なった。実験では、同表記異義の現象が存在する英略語11種類を用意し、また、閾値dは、0から1まで0.05刻みで変化させた。

実験の結果、方法1を用いた場合は閾値dを0.95とした場合の平均検索洩れ率・平均不要検索率がもっとも低く、平均検索洩れ率は28.2%、平均不要検索率は0%となった。また、方法2を用いた場合は閾値dを0.90とした場合の平均検索洩れ率・平均不要検索率がもっとも低く、平均検索洩れ率は12.0%、平均不要検索率は34.5%となった。ここで、これらの値と、同表記異義を処理しない場合、すなわち、キーワードの表記のみに着目して検索した場合の値(平均検索洩れ率・平均不要検索率)とを比較した結果を表3に示す。

表3 検索洩れ率と不要検索率の比較

	処理しない場合	方法1 ($d=0.95$)	方法2 ($d=0.90$)
検索洩れ率	0	28.2%	12.0%
不要検索率	57.6	0%	34.5%

この表からも明らかなように、方法1、方法2を用いた場合の方が、同表記異義を処理しない場合よりも平均不要検索率が軽減している。また、方法1と方法2とを比較すると、平均検索洩れ率に関しては方法2を用いた場合の方が低く、逆に、平均不要検索率に関しては方法1を用いた場合の方が低い。

ここで、実験結果の例として、同表記異義の英略語：SFC（Sequential Function Chart / Space Filling Curves）を含む論文をクラスタリングした結果を図3に示す。また、方法1、方法2に基づいてクラスタリングの結果を情報検索に利用したときの、閾値dと検索洩れ率・不要検索率の関係を、図4, 5に示す。

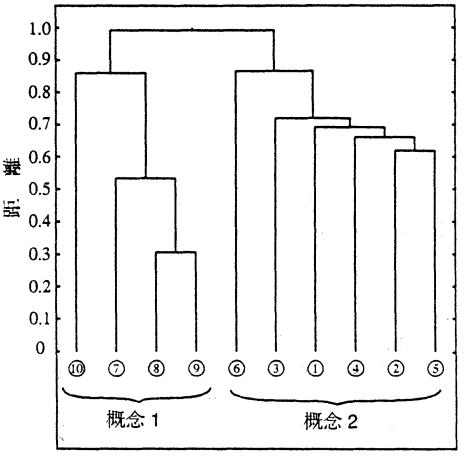


図3. SFCを含む論文をクラスタリングした結果

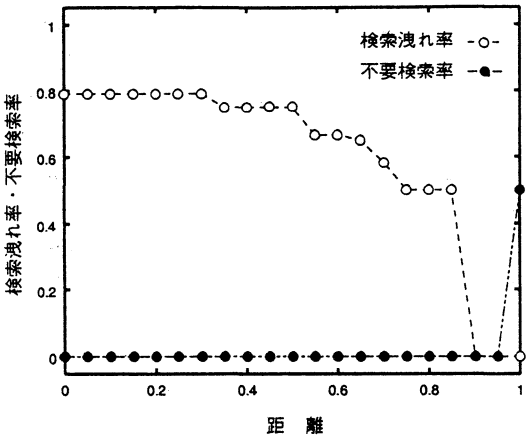


図 4. 方法 1 を使用した場合の英略語 SFC の検索洩れ率・不要検索率

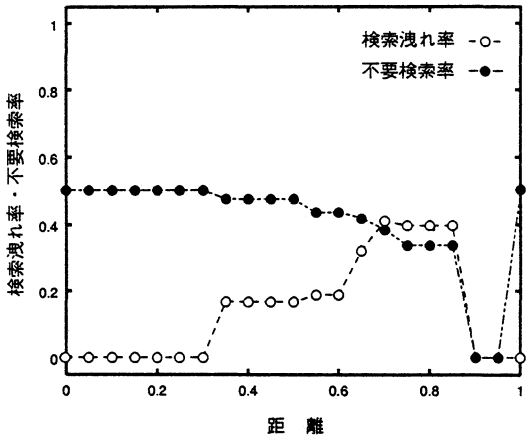


図 5. 方法 2 を使用した場合の英略語 SFC の検索洩れ率・不要検索率

6. おわりに

本報では、同表記異義の関係にあるキーワードの概念を共起情報を用いて推定し、情報検索に利用する方法を提案した。

参考文献

[1] H. Fujisaki, H. Kameda, S. Ohno, T. Ito, K. Tajima and K. Abe : "An intelligent system for information retrieval over the Internet through spoken dialogue," *PROCEEDINGS of EUROSPEECH'97*, vol. 3, pp. 1675-1678 (1997).

[2] 劉軼, 戸井田和重, 八杉大輔, 阿部賢司, 大野澄雄, 藤崎博也, 久保村千明, 亀田弘之 : "学術情報検索における異表記同義・同表記異義の分類・分析および処理," 言語処理学会第 4 回年次大会発表論文集, pp. 108-111 (1998).

[3] 藤崎博也, 大野澄雄, 阿部賢司, 片見憲次, 飯島岐勇, 鈴木匡芳 : "キー概念に基づく情報検索方式の高度化 (2) -キーワードの同表記異義の処理-," 情報処理学会第 57 回全国大会講演論文集, vol. 3, pp. 239-240, (1998).

[4] 藤崎博也, 大野澄雄, 阿部賢司, 飯島岐勇, 片見憲次, 鈴木匡芳 : "定量的基準に基づく検索結果の順位付けの検討," 情報処理学会第 58 回全国大会講演論文集, vol. 3, pp. 125-126, (1999).

[5] 藤崎博也, 阿部賢司, 片見憲次, 武田和也, 白井克彦 : "共起情報を用いた同表記異義の処理," 情報処理学会第 59 回全国大会講演論文集, vol.2 , pp.333-334 , (1999).

[6] http://www.nacsis.ac.jp/nacsis.index.html

[7] http://www.gls.co.jp/glsoft/chemomet/chemometo/hca/hca.htm

[8] http://aoki2.si.gunma-u.ac.jp/lecture/misc/clustan.html

[9] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明 : "日本語形態素解析システム「茶筌」version 1.5 使用説明書," Technical Report NAIST-IS-TR97007(1997).