3U-05 # キー概念検索のための異表記同義・同表記異義の処理

藤崎博也　武田和也　阿部賢司　堀越修平　猪股尚典　山崎潤

東京理科大学

1. はじめに

従来のキーワード検索方式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や同表記異義が存在すると、検索洩れや不要な検索が生じる。これらを回避するためには、キーワードの概念にまで遡って検索するキー概念検索方式が有効である [1]。

本報では、キー概念検索を具体化するための自然言語処理として、特に、学術情報検索における同表記異義・異表記同義の現象に着目し、その処理方法について検討した結果を述べる。

2. 異表記同義の収集・分類

学術情報検索における異表記同義の現象を定量的に把握するため、情報検索システム評価用テストコレクション 1[2] に収録されている 339,483 件の論文情報の中から、1,000 件分の情報を任意に抽出し、その中のキーワード 3,906 語から異表記同義の実例を収集した。なお、この際には、キーワードの表記と概念を対応付けた表記-概念対応辞書 [3] を参照し、着目したキーワードと同じ概念を持つ語が辞書に複数存在する場合には、それを異表記同義の現象とみなした。その結果、3,906 語のうち、830 語 (21.0%) に異表記同義の現象が存在することを確認した。

ここで、収集した異表記同義を、表記に着目して分類した結果を以下に示す。また、分類した各要素の出現率を表 1に示す。

(1) 表記の多様性によるもの
- (a) 漢字仮名混じり表記： 例. 捻りモーメント / ねじりモーメント
- (b) 片仮名表記： 例. コンピュータ / コンピューター
- (c) 数字表記： 例. 3 次元モデル / 三次元モデル
- (d) ローマ字表記： 例. fiber / fibre

(2) 略語によるもの
- (e) 片仮名表記： 例. ミリメートル波 / ミリ波
- (f) ローマ字表記： 例. ESB / Electron Spin Resonance

(3) 辞書上の概念が一致したことによるもの
- (g) 同言語間： 例. 同定 / 識別
- (h) 多言語間： 例. 誤り率 / error rate

表 1 異表記同義の種類と出現率

異表記同義の種類	出現率 [%]
(1) 表記の多様性	(小計：24.8)
(a) 漢字仮名混じり表記	3.8
(b) 片仮名表記	20.0
(c) 数字表記	0.1
(d) ローマ字表記	0.9
(2) 略語	(小計：1.4)
(e) 片仮名表記	0.4
(f) 英語表記	1.0
(3) 辞書上の概念の一致	(小計：73.8)
(g) 同言語間	43.2
(h) 多言語間	30.6

3. 異表記同義の処理方法と処理結果

本報では、ユーザが提示したキーワードと異表記同義の関係にある全てのキーワードを自動的に抽出して検索式を拡張することによって、検索洩れを回避する手法を提案する。具体的には、まず、表記-概念対応辞書にもとづいて着目するキーワードに異表記同義の現象が存在するか否かを判断し、存在する場合には、異表記同義の関係にある全てのキーワードを検索式に追加して検索する。

この方法にしたがい、異表記同義の現象が存在する語をキーワードとした場合の情報検索実験を行なった。なお、この際には、情報検索システム評価用テストコレクション 1 に収録されている 339,483 件の論文情報を検索対象として用いた。

異表記同義の現象が存在するキーワード 30 語を用意し、(1) 異表記同義を処理しない場合の平均ヒット件数、(2) 異表記同義を処理した場合の平均ヒット件数、(3) 異表記同義を処理した場合にヒットしたものの中で、誤って検索されたものの平均件数、を調べた結果を表 2に示す。

この表からも明らかなように、異表記同義を処理した場合の方が、処理しない場合よりヒット件数が増加しており、提案手法が検索洩れを軽減するのに有効であることが確認できる。なお、(3) の誤り検索は、拡

表 2 異表記同義の現象が存在する語をキーワードとした場合の情報検索実験の結果

	件数 [件]
(1) 処理しない場合の平均件数	228.9
(2) 処理した場合の平均件数	278.2
(3) 処理によって誤って検索されたものの平均件数	1.6

Processing of synonymy and polysemy/homonymy for concept-based search
Hiroya Fujisaki, Kazunari Taketa, Kenji Abe, Syuhei Horikoshi, Takanori Inomata, Jun Yamazaki
Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510

張したキーワードに同表記異義の現象が存在する場合に生じるものであるが、これに関しては、同表記異義の処理を加味すれば軽減できるものと考えられる。

4. 同表記異義の収集・分類

学術情報検索におけるの同表記異義の現象を定量的に把握するため、情報検索システム評価用テストコレクション 1 に収録されている 339,483 件の論文情報の中から、1,000 件分の情報を任意に抽出し、その中のキーワード 3,906 語から同表記異義の実例を収集した。なお、この際には、表記-概念対応辞書を参照し、着目したキーワード (表記) に複数の概念が対応している場合には、それを同表記異義の現象とみなした。その結果、3,906 語のうち、221 語 (5.7%) に同表記異義の現象が存在した。

ここで、収集した同表記異義を、表記に着目して分類した結果を以下に示す。また、分類した各要素の出現率を表 3に示す。

(1) 漢字表記： 例. 分散
　概念 1： variance (of a variable)
　概念 2： dispersion (of light, matter)
(2) 平仮名表記： 例. ひずみ
　概念 1： deformation of a material
　概念 2： distortion of a waveform
(3) 片仮名表記： 例. チャネル
　概念 1： communication channel
　概念 2： a band of radio waves
(4) ローマ字表記 (英略語)： 例. IR
　概念 1： information retrieval
　概念 2： infrared

表 3 同表記異義の種類と出現率

同表記異義の種類	出現率 [%]
(1) 漢字表記	56.6
(2) 平仮名表記	0.4
(3) 片仮名表記	39.8
(4) ローマ字表記 (英略語)	3.2

5. 同表記異義の処理方法と処理結果

本報では、同表記異義の関係にある語を、それらの共起情報にもとづいて、それぞれの概念ごとに分類し、その中からユーザの意図と合致するものを特定することによって、不要な検索を回避する手法を提案する [1]。

提案手法の概要は次の通りである。まず、着目したキーワードを含む論文の検索空間における位置を、共起情報にもとづいてベクトルで表す。次に、そのベクトルにもとづいて、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、階層的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 d を閾値としたときのクラスタリングの結果を参照し、クラスターの数が n 個 ($n \geq 2$) の場合には、キーワードは n 種類の意味で用いられている、すなわち、キーワードには同表記異義の現象が存在すると予測する。このとき、各クラスターの共起情報とユーザの検索意図との合致度を考慮して、以下の 2 つの方法にもとづいて検索する。

[方法 1]：不要な検索の少なさを重視する方法

キーワードの概念はクラスター毎に異なる意味で用いられているとみなし、ユーザの検索意図に合致するもののみを検索する。(本研究では、ユーザの検索意図に合致する 1 つのクラスターのみを検索することとする。)

[方法 2]：検索洩れの少なさを重視する方法

閾値 d において他のいずれともリンクしていないものは、実際に共起単語が他のいずれとも異なるのか、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ものは必ず検索する。

この手法にしたがい、同表記異義の現象が存在するキーワード 10 種類を用意し、それらを情報検索に用いたときの平均検索洩れ率・平均不要検索率を求めた。その結果、方法 1 を用いた場合には、閾値 d を 0.95 とした場合の平均検索洩れ率・平均不要検索率が最も低く、方法 2 を用いた場合には、閾値 d を 0.90 とした場合の平均検索洩れ率・平均不要検索率が最も低かった。方法 1、方法 2 のこれらの値と、同表記異義を処理しない場合の値を表 4に示す。

表 4 平均検索洩れ率と平均不要検索率の比較

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

この表からも明らかなように、同表記異義を処理した場合の方が、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、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。

6. おわりに

本報では、キー概念検索を具体化するために、異表記同義、同表記異義の処理方法について検討した。また、その方法を情報検索に応用することにより、その有効性を検証した。

参考文献

[1] K. Abe, M. Iijima, K. Katami, M. Suzuki, K. Kurokawa, K. Taketa, S. Ohno and H. Fujisaki: "Concept-Based Search and User Modeling in Information Retrieval Based on Human-Machine Dialogue," *Proceedings of RIAO'2000*, vol.2,pp.1728-1743 (2000).

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

[3] 日本電子化辞書研究所:EDR 電子化辞書仕様説明書 (第 2 版),1995.