化学データベースにおける名称検索の適合率の向上

1C－2

吉川雅修†　定盛浩之‡　片谷教孝†
（† 山梨大学　‡ ソニー幸田株式会社）

1　はじめに

化学データベース検索にはさまざまな検索項目があるが，物質名称からの検索（名称検索）の利用が最も頻度が高い．化学物質には類似する複数の日本語別名が存在するものが多数あり，物質の正式な主名称だけを検索の対象とするだけでは検索に不便である．

名称検索を補助する方法の１つとして文字列の部分一致検索が挙げられるが，これを有効に使うにはある程度の化学の知識が必要である．より広い範囲の利用者を対象とするために，物質の別名称による検索を柔軟に行なう手法について考える．

2　化学物質の別名称での検索

化学物質の名称については国際名に加えて慣用名が広く使われている．さらにその日本語名については字訳基準がある [1] が，やはり慣用を認めるなどの例外を含んでいる．従って，複数の別名称を持つ物質が多く存在する．化学物質の別名称は文字列が1ヶ所だけ違っているものから，全く違っているものまでさまざまである．

このような別名称での検索において検索洩れを防ぐ方法には次のものが考えられる．

1. 別名称をすべてデータベースに登録する
   - ○ 登録した名称を持つ物質を確実に検索することができる．
   - ● データ作成コストが増加する．
2. 名称が類似しているものは容認する
   - ○ データ作成コストを増加させない．
   - ● 文字列間の関係，他の物質名称との関係を考える必要がある．

全く異なる文字列の別名称に関しては登録する方法を採用する他はない．しかし，化学物質名称では，化合物の系統によって共通の類似性を有する多数の別名称パターンが存在するので，すべて登録データとして用意する方法ではデータ追加のコスト増加が問題となる．本研究では，入力文字列に関して類似性の高い名称を照合結果として出力し，データを個々に登録することなく検索システム側で対応する方法を考える．

2.1　１文字違いを容認する照合システム

１文字違いの同一化学物質の名称が他数存在することに着目し，第一段階として，１文字の違いを容認する場合に絞って照合プログラムを作成した：

> 照合の時に，文字列が同一であるか、またはどこか１文字のみ異なる場合に一致と判定する．

これを物質名称の完全一致，部分一致の名称検索に適用する．

2.2　同一物質 / 異物質の判定規則

単純に類似文字列の容認を行う照合では，目的物質以外の多数の物質名称を検索結果として出力してしまうので，実用的でない．本研究の対象が化学物質データベースであることから，化学物質名称の特徴を利用して，類似文字列が同一物質であるか異物質であるかを判定する規則を照合システムに導入する．これによって，目的以外の物質名が検索される数を小さくする．

今回の１文字違い容認の照合に対応した判定規則は，多数存在する１文字違いの同一物質名称と１文字違いの異物質名称の組を分析し，容認を抑制するべきであるパターンを抽出し，照合時の規則に組み込んだものである．

Improvement of the Relevancy of Search in Chemical Databases

Masanobu YOSHIKAWA, Hiroyuki SADAMORI, and Noritaka KATATANI
†Yamanashi University, ‡ Sony Kouda Co.,Ltd.

容認を抑制するパターンの一部

パターン1: 前1文字、後2文字が一致する

(例) エステル ⟷ エーテル
サリチル ⟷ サ○チル

パターン2: 末尾の4文字が一致し、末尾から5文字目が異なる

(例) エタノール ⟷ ○タノール
ピクリンサン ⟷ ピ○リンサン

パターン3: 1文字目が異なり、残り2文字が一致する

(例) ヨード ⟷ ○ード
メタン ⟷ エタン(ブタン)

3 検索の実験: 性能評価

それぞれの機能について検索をおこない、検索の性能を調べた。この実験では神奈川県の環境化学物質データベース[3]に登録されている名称の中から約4500件(1800物質)をデータとして用いた。被験者は理科系(非化学系)の大学生6人である。各被験者が20-30個の名称を記憶に頼って記述した文字列について検索を行なった。

検索効率の良否を判定する基準として、下記のように定義する目的物質の検索率、出力結果の適合率を用いる。

検索率 = (結果に目的物質名が含まれた検索回数)/(全検索回数)

適合率 = (ユーザーが目的とするデータの数)/(検索されたデータの数)

表1: 目的物質の検索率(%)

検索方法	検索率
完全一致検索	65.6
完全一致(1文字容認)	74.8
部分一致検索	83.3
部分一致(1文字容認)	88.9

表2: 出力結果の適合率(%)

検索方法	規則適用	規則なし
完全一致検索(1文字容認)	79.8	76.1
部分一致検索(1文字容認)	12.9	1.4

実験結果において、「完全一致(1文字容認)」、「部分一致(1文字容認)」とある項が前節で述べた1文字違いの容認と判定規則の適用とを行なう方式による場合である。

この実験では、今回実現した方法により完全一致、部分一致における検索率と適合率とがともに向上した。

表1から、今回実現した方法が検索率に関して効果があることがわかる。慣用名や日本語化の際の表記のゆれにより生じる別名称での検索が、名称登録の追加なしで実現することができる。なお、副次的な効果ではあるが、一部分が不正確な入力による検索を可能にする効果もある。

1文字違いを容認する場合において異物質か同一物質かを判定する規則を与えることの効果を確認した実験結果を表2に示す。完全一致、部分一致の両方の場合において適合率が向上している。

現段階での容認の手法を部分一致検索に適用することの検索率における効果は小さく、文字列によっては大量の出力結果を出すので適合率が低くなって実用的でない。しかし表2を見ると、判定規則を検討することによって適合率の向上が可能であることがわかる。

4 まとめ

化学物質データベースの検索の効率化の1つとして別名称での柔軟な検索を登録データを増やさずに実現することを試みた。第一段階として文字列の1文字以内の違いを容認する照合と、物質名称の性質から導出した判別規則との組合せを作成し、その有効性を確認した。

参考文献

[1] 小川雅弥・村井真二: 有機化合物・命名の手引, 化学同人, 1990.

[2] 化学式辞典編集委員会: 化学辞典, コロナ社, 1990.

[3] 富士通 FIP: 神奈川県化学物質安全情報システム, 1992.

[4] 千原秀昭・時実象一: 化学情報, 東京化学同人, 1991.