化学データベースにおける名称検索の適合率の向上（２）

3F－10

伊東靖史　吉川雅修　片谷教孝
（山梨大学）

1　はじめに

本研究は化学データベース検索の中で最も利用頻度が高い物質名（日本語名称）による検索を対象とする。化学物質には別名称をもつものが多数存在するため、一つの名称のみでしか検索できないとなると不便である。

本研究では、化学物質名称には類似した名称の同一物質が多数存在している所に着目し、データベースに登録されている名称の中で入力文字列と類似度の高い名称を照合結果として出力する検索システムを考える。

2　類似度を用いた検索

前回は、文字列間の類似度を計る尺度である、likeness measure（$LM(A,B)$）を用いて検索を行うことを考えた。

検索に際しては、入力文字列とデータベースに登録されている全ての名称との間の類似度を計り、後述する類似度の許容値を上回るもの全てを照合結果とする。

2.1　Likeness Measure

LM の定義は以下のとおりである。[3]

$$LM(A,B)=\frac{LLCS(A,B)}{max(|A|,|B|)}$$

ただし、

$LLCS(A,B)$ ： 文字列 A と B の最長の共通部分列

$|A|$ ： 文字列 A の文字列長

2.2　類似度の許容値

類似度の許容値は、大きすぎると検索結果に含まれる目的物質以外の物質名の数が著しく多くなる。

従って、類似度の許容値は文字列長によって変える事にし、以下のように定義した。

$s=max(|A|,|B|)$ とするとき、

（1）$s\geqq 7$ のとき、《類似度》$=\frac{(s-2)}{s}\times 100\%$

（2）$s\leqq 6$ のとき、《類似度》$=75\%$

文字列長の長い複合名称では容認すべき箇所が複数個ある可能性があり単純ではない。しかし本研究で今回扱った入力データでの最長のものは14文字であり、この範囲で考えた。

なお、検索結果に入力文字列と完全に一致するものがある場合はそれのみ出力する。

3　適合率の向上

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

$$\text{検索率}=\frac{\text{(結果に目的物質が含まれた検索回数)}}{\text{(全検索回数)}}$$

$$\text{適合率}=\frac{\text{(目的とするデータの数)}}{\text{(出力されたデータ数)}}$$

先に述べたようにLMを用いて検索を行なう事により、検索率については効果があったが、適合率についてはまだ満足のいく結果ではない。

検索実験から、互いに類似度の高い基名等を含む物質名が適合率低下の主な要因であることがわかった。

そこで、次に示す単語を登録した辞書を用意し、LMによって類似度が高く同一物質である可能性があるとみなされた検索結果のうち、明らかに異物質であるものをふるい落とす事を考えた。

Improvement of the Relevancy of Search in Chemical Databases

Yasushi ITO, Masanobu YOSHIKAWA, and Noritaka KATATANI
Yamanashi University.

～ 辞書に登録するもの ～

○主要基名（メチル、エチル等）

○官能種類名（フッカ、エンカ等）

○元素名（タンソ、サンソ等）

○炭化水素（メタン、エタン等）

上記のものの内、辞書内の他のどの単語とも2文字以上異るものは登録しない。なお、今回登録した単語数は243件であり、平均文字列長は4.43文字、分散は1.88である。

3.1 ふるい落としの手順

§2で述べたようにLMで検索を行ない、一致とみなされた場合、その文字列（出力文字列）と入力文字列との間で以下の手順を実行する。

（1）部分一致検索によって、前述の辞書に登録されている部分列が入力文字列、出力文字列に存在するかを調べ、存在する場合その位置を記憶しておく。

（2）入力文字列と出力文字列のそれぞれ何文字目が不一致であるかを調べる。（類似度の許容値の定義から、不一致はそれぞれ最大で2箇所である）

（3）入力文字列、出力文字列ともに不一致箇所が全て（1）の部分列によるものである時、異物質であるとみなす。

3.2 文字列間の不一致箇所

2つの文字列A [1..n]、B [1..m] の不一致箇所を探す為に、(n+1)(m+1) の行列を用いて2つの文字列のLLCSの計算を行なうWAGNER-FISCHERアルゴリズムに以下の変更を加えて利用する。

○位置 (i,j) で文字が一致するとき、従来は無条件で (i-1,j-1) の値に1を足したものを (i,j) の値としたが、以下の制約を加える。

1. 文字列Aにおいて一致とみなされた最新の位置を記憶しておき、iがその値より小さい場合は一致とみなさない。
2. 同じ行で一致する箇所が複数ある場合、既に同じ行に一致とみなした箇所がある時、及びそこを一致とみなすと最終的に2つの文字列のLLCSの値に満たない時は一致とみなさない。

以上のようにして計算された行列の第m行がAの一致状況を示し、左から順に値が1ずつ増えている所は一致、前の値と同じ所は不一致となる。同様にして第n行はBの一致状況を示す。

4 検索実験

検索実験には、神奈川県の環境化学データベースの検索ログファイルを用いた。ファイル中から検索に失敗したデータを取り出し、中でも出現回数が15以上のもの113件の検索実験を行った。

検索実験の対象は、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4812の化学物質名称である。なお、平均文字列長は9.76文字、分散は39.8である。

表1．検索率と適合率

	LM	*LM*（辞書付き）
検索率	86.2%	86.2%
適合率	61.2%	80.0%

表1より、今回の手法によって適合率において効果が見られたことがわかる。

5 まとめ

基名などを登録した辞書を用意し、類似度は高いが異物質であるものを排除することにより適合率を上げることに成功した。

参考文献

[1] 吉川雅修・定盛浩之・片谷教孝: 化学データベースにおける名称検索の適合率の向上, 情報処理学会第47回全国大会

[2] 伊東靖史・吉川雅修・片谷教孝: 化学データベースにおける名称検索の適合率の向上, 情報処理学会第49回全国大会

[3] Shufen Kuo,George R.Cross: A TWO-STEP STRING-MATCHING PROCEDURE,Pattern recognition,Vol.24,No.7,pp.711-716,1991.

[4] 富士通 FIP: 神奈川県化学物質安全情報システム, 1992.