化学データベースにおける名称検索の適合率の向上

5V－3

伊東靖史　吉川雅修　片谷教孝
（山梨大学）

1　はじめに

本研究で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。データベースに全ての別名称を登録すれば問題はないが、現実にはデータ量が膨大となり、データ作成のコスト増加の問題が起こる。

全く異なる文字列の別名称に関してはデータベースに登録する他はないが、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すれば、類似した名称については照合処理で対応する方法が考えられる。

本研究では、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　前回のシステム[1]とその問題点

前回は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、1文字の違いを容認する照合に絞って手法の効果を確認した。容認する事により目的物質以外の多数の物質名称が検索結果に紛れ込む事を抑制する為に、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を判定する規則を照合システムに導入した。以下、1文字違い容認に規則を加えたものを単に『1文字違い容認』と表記する。

規則については、判定に有効でかつ同一物質をふるい落としてしまう問題がない完全な規則集合を作成する事は難かしい事がわかった。

昨年の考え方では対応できないパターンとして、以下の2つが挙げられる。

・2文字以上異なるもの
例：クロロベンジラート と クロルベンジレート
・文字列長の異なるもの
例：ナトリウム と ナトリュウム

1文字違い容認に加えて上記の場合にも対応する為、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。

3　類似度を用いた検索

3.1　Likeness Measure

LM の定義は以下のとおりである。[2]

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

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

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、決められた類似度の許容値を上回るもの全てを照合結果として出力する。

3.2　類似度の許容値の改良

最初、類似度の許容値は一率65%とした。この場合、文字列長によって2文字以上の違いを容認する事になり、検索結果に含まれる目的物質以外の物質名の数が1文字違い容認に比べて著しく多い。

従って、類似度の許容値を文字列長によって変える事が望ましい。

今回の実験に用いたデータについて、何文字か容認する事によって検索に成功した例を抜きだし、それらから文字列長別に許容値の検討を行なった。

Improvement of the Relevancy of Search in Chemical Databases
Yasushi ITO, Masanobu YOSHIKAWA, and Noritaka KATATANI
Yamanashi University.

類似した名称の同一物質名の多くはアルファベットからカタカナに直すときの表記のゆれであり、1箇所につき2文字違いまでである。また、3文字以上違ったものは、文字列長が短い名称の場合には類似したものとは言えないであろう。

以上の点を踏まえて、許容値の決め方を次のように提案する。

$s = max(|A|, |B|)$ 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》$= 75\%$

このような決め方によって、3文字以下では完全一致のみ、6文字以下では1文字容認、それ以上は2文字容認ということになる。

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。

4 検索実験

1文字違い容認とLMとによる検索をおこない、検索の性能を比較した。検索実験に際し、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、その中でも出現回数が15以上のもの113件の検索実験を行った。ただし、別名称も含めデータベースに存在しない物質については対象外としている。

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。

検索効率の良否を判定する基準として、次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。

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

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

表1．検索率と適合率

	1文字違い	LM	LM'	LM''
検索率	44.8%	86.2%	86.2%	86.2%
適合率	100%	10.7%	24.3%	61.2%

表1で、『LM』は類似度の許容値を一率65%とした検索結果である。『LM'』は3.2節で述べた類似度の許容値を変えた場合である。

さらに、入力文字列と完全に一致したものが存在する場合には他に類似度の高いものがあっても出力しないように改良を加えて検索をおこなった。（『LM''』）

- 表1から、類似度を用いた今回の方法では1文字違い容認と比べて検索率において効果があった事がわかる。しかし、適合率は著しく低下している。
- 『LM』と『LM'』とを比べると検索率は落さずに適合率がやや向上することがわかる。
- 『LM''』を見ると、適合率においてかなりの向上がみられる事がわかる。

5 まとめ

前回の『1文字違い容認』では検索できなかった2文字違っているもの、文字列長の異なるものに対応する為、類似度を用いる方法を試みた。結果として、検索率において向上がみられた。また、この試みによって低下した適合率についても、類似度の許容値を文字列長によって変える、といった改良により多少向上した。さらなる適合率の向上の為には化学物質名称の特徴をより有効に利用する手法が必要となる。

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46回全国大会

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

[3] 富士通 FIP: 神奈川県化学物質安全情報システム, 1992.