化学データベースにおける名称検索の適合率の向上（４）

２Ｑ－１　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日本語物質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、データベースに登録されていない名称からでは検索できない事が多々ある。通常の日本語の表記のゆらぎへの対処としては飯田ら[1]などの例があるが、本研究で対象とする化学物質名称は基本的にゆらぎの構造が異なるため、別のアプローチが必要と考えられる。

そこで本研究では、別名称には正式名称と類似した名称のもの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てきた。[2][3][4]

今回は、これまでの研究のまとめと新たに改良を加えた部分について報告する。

2　類似度を用いた検索

本研究では、文字列間の類似度を計る尺度である、likeness measure（$LM(A,B)$）を用いる方法を提案した。この方法を用いることにより、データベースに登録されていない別名称のうち主名称に類似したものについては検索漏れを免れることができる。また、ミスタイプ等による検索漏れも同時に防ぐことが可能となる。

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、許容値を上回るもの全てを照合結果とする。

ただし、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。

LMの定義は以下のとおりである。[5]

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

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

3　辞書照合による適合率の向上

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。

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

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

LMを用いて検索を行う事により、検索率については効果があったが、適合率は必ずしも高くなく、その原因の多くは互いに類似度の高い基名等を含む物質名であることがわかった。そこで、主要基名、元素名等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よう改良した。

4　改良点

今回新たに以下の改良を加えたので報告する。

4.1　特定のミスタイプへの対応

実際の化学データベースの検索ログを見ると、明らかなミスタイプとして、長音「ー」をマイナス記号「－」としているケースが多く見られる。このために検索もれとなってしまったり、辞書によるふるい落としがなされないことも起こり得ると考えられる。

従って、このミスタイプについては検索前に入力文字列をチェックし修正するように改良した。

表１．検索率と適合率

	改良前	改良後
検索率	81.7%	81.7%
適合率	63.7%	67.1%

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

表1に改良前と改良後の検索率と適合率を示す。これより、この実験においては適合率においてのみ向上がみられたことがわかる。さらに入力文字列によっては検索率においても効果があることが予想される。

4.2 位置番号の入った物質名の検索補助

化学物質名の中には「1,1,1-トリクロロエタン」のように位置番号を含むものがある。これらの名称は数字部分の誤り、カンマやハイフンの不統一等により検索漏れとなる例が極めて多い。また、たとえ入力が正しくてもデータベース側には総称（上の例では「トリクロロエタン」）で登録されている可能性もある。

これらの名称の場合、不一致箇所が多く従来のLMによるシステムでは救済されないケースが多い。従って、従来のシステムで検索を行った結果、類似した文字列が1つもない場合、位置番号の部分を取り除いた文字列で部分一致検索にかけるという方法を考えた。

実際の化学データベースにおいて検索に失敗した例の内、位置番号を含むもの100件について検索実験を行った（表2）。なお、救済ができなかったものの中には、データベースに存在しない為に救済できないものも含まれている。

表2．位置番号を含む名称の検索結果

	件数
新しい方法で救済されるもの	13
従来のLMで救済されるもの	43
総称が得られたもの	3
救済できなかったもの	41

表2から、従来のLMでは救済できなかったもの57件の内13件が新たに救済された。また、救済されなかったものの中でも、位置番号のみ違う物質が出力される事がよくあり、これもユーザにとって何らかの参考となるであろう。

この方法の採用に伴って生じるであろう問題点は以下の2つが考えられる。

- ○ 部分一致を用いる為、名称によっては出力される件数が多数になってしまう。
- ○ 計算速度の低下

4.3 検索速度に関する改良

これまでは、検索率、適合率を向上されることに専念してきたが、今回、検索速度の面で効率化をはかるために、以下の変更を加えた。

- ○ LLCSの計算において、計算途中で既に類似度が許容値を下回ることが確定した時点で計算を打ち切る。
- ○ 検索途中で入力文字列と完全一致のものが見つかったらそこで検索を打ち切る。

この他にも、データベースを文字列長でソートしておき、入力文字列長に応じてサーチする範囲を限定する、といったことが考えられる。

5 まとめ

今回は、これまでの研究のまとめと新たに加えた改良点について報告を行った。

今回新たに追加した、位置番号を含む名称への対応は効果は見られたが、今後は適合率の面も考慮する必要がある。

また、検索速度の問題は今回2つ改良を加えたが、さらなる効率化を目指すとともに、速度効率の評価をする必要がある。

参考文献

[1] 飯田敏幸・中村行宏：変形ルールと禁則ルールを用いた片仮名の表記ゆらぎの解消法, 情報処理学会論文誌,Vol35,No11,pp.2276-2282, 1994.

[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 4-169

[3] 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（2）, 情報処理学会第50回全国大会 4-37

[4] 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（3）, 情報処理学会第51回全国大会 4-249

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

[6] 富士通FIP: 神奈川県化学物質安全情報システム開発報告書, 1995.