化学データベースにおける名称検索の適合率の向上（３）

７Ｅ－９　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、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。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てきた。

今回は、以前の発表で指摘を受けた、片仮名をローマ字に変換して照合を行う方法を新たに試し、従来の手法との比較を行った結果を報告する。

2　類似度を用いた検索

前報[1]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。

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。[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

Improvement of the Relevancy of Search in Chemical Databases

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

3　辞書照合による適合率の向上[2]

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　ローマ字変換による検索

表１．同一物質と異物質

	名称	違っている箇所
同一物質	クロロメタン クロルメタン	ｒｏ ｒｕ
異物質	メタノール ブタノール	ｍｅ ｂｕ

表１に示すように、違っている箇所をローマ字に変換し、子音が一致しているかどうかで同一物質であるかの判定をする。

今回はこの手法を２つの目的で試す。

ローマ字１：ＬＭのかわりにローマ字照合で検索を行う。

ローマ字２：ＬＭによる検索結果に対してローマ字照合によるふるい落としを行う。

5 比較実験

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[4] 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 2．実験結果

	検索率	適合率
*LM**	81.7%	47.9%
LM(辞書付き) *	81.7%	63.7%
ローマ字 1	58.3%	92.1%
ローマ字 2	81.7%	56.3%

※注：長音とマイナス記号のミスタイプを考慮したため前報と多少数値が異なっている。

5.1 ローマ字 1 について

従来の手法と比べて、適合率はかなりの向上が見られたが、検索率の点で不十分であると言える。

検索率が悪くなった理由として、次の 2 つが挙げられる。

- 文字列長の違うものが救済できない。
- ミスタイプの救済ができない。

特にミスタイプの救済は本研究で当初から目的としたものではないが、実際の検索失敗例の約 11.2 %を占めており、重要であることがわかる。

なお、ミスタイプの救済ができない、というのはユーザーが仮名入力した場合であり、ローマ字入力した場合、母音におけるミスタイプは救済される。

5.2 ローマ字 2 について

辞書の代わりにこの手法を用いた結果、適合率において効果がみられた。ただ、辞書を用いたものよりは若干適合率がよくない。

6 まとめと今後の方向

今回は、従来の方法との比較として、違っている箇所をローマ字に変換して同一物質であるかを判定する方法を試した。その結果、総合的にみて従来のＬＭによる方法の方がよい結果が得られた。

これらの結果は、本研究の対象が化学データベースであり、入力する文字列が日常ふれることの少ない化学物質名称であるということによるとみられ、他の日本語検索における表記のゆらぎへの対処（例えば飯田ら[5]の方法）とは異なっているとみるべきであろう。

従って今後は、ミスタイプの救済ができるという点で、従来のＬＭを用いる方法を採用しようと考えている。

なお、本研究ではこれまで検索速度やメモリ負荷の問題は度外視してきたが、今後はこれらの問題についても検討する必要がある。

参考文献

[1] 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

[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（２）, 情報処理学会第 50 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.

[5] 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.