# ファジィマッチング関数とその情報検索への応用

4K－1

三宅輝久、宮本定明、中山和彦
（筑波大学）

## 1．はじめに

情報検索システムに対してファジィ集合理論を応用する試みは種々行われてきた。そこにおいては、従来の2値論理に基づいた情報検索システムでは問題とならなかった新しい要素がいくつか存在する。その中の一つとして、キーワードと索引語のマッチングの問題がある。従来、マッチングについては、完全一致を基本とする方法で行われてきたが、ファジィ情報検索システムにおいては、ファジィ集合間の演算として定義し、マッチングの機能を拡張することができる。本稿では拡張されたマッチング関数（ファジィマッチング関数）として部分列のマッチングを用いた方法を提案し、それをファジィ情報検索に応用した例について報告する。

## 2．情報検索におけるマッチング

従来の情報検索システムにおいては、キーワードと索引語の二つの文字列のマッチングを考える場合、その長さ、文字種共に完全に一致するかどうか、即ち完全一致するかどうかを基準としてきた。索引語が統制されている場合には完全一致でも問題点は少ないが、自然語索引の場合には同一の意味の索引語でも語尾変化等で多くの語形をとり，完全一致ではその一部にしかマッチングしないという難点があった。そのため、従来から完全一致だけではなく部分一致（トランケーション）が用いられてきた。（表1）

表1　部分一致の種類

| | |
|---|---|
| 前方一致 | 文字列@ |
| 後方一致 | @文字列 |
| 両端一致 | 文字列@文字列 |
| 中間一致 | @文字列@ |

（@：不定文字記号）

これによりマッチングの機能はかなり拡大されたが、部分一致をうまく使用するためには、対象となる索引語の形についての知識を利用者が持っており、明示的に検索質問中のキーワードを組み立てることが必要である。これは、全ての利用者が容易に行なえるものではない。利用者に意識させることなく、キーワードと形態的に最も似た索引語に自動的にマッチングさせるための機能拡張が必要である。マッチング関数のファジィ化により、これを実現することが可能である。

## 3．マッチング関数のファジィ化

完全一致のマッチング関数は次のように定義される。マッチング関数をm、キーワードをv、索引語をwとすると、

$$m(v,w)=\begin{cases}1 & v=w\\ 0 & v\neq w\end{cases}$$ となる。

ここでmをファジィ化された関数に拡張し、vとwの一致の程度を〔0，1〕の値で表現することを考える。そのためvとwに何らかの集合を対応させ、それらの間の演算によりmを定義する。文字列は（1）含まれる文字、（2）文字の順序、という二つの情報を持っている。文字列を、単に含まれる文字の集合に対応させるだけでは、順序情報が失われ、望ましくない一致が多くなることは明らかである。そこで、文字列をそれを構成する全ての部分列の集合に対応させることを考えた。即ち長さLの文字列は、それ自身から個々の文字に到るL＊（L＋1）／2個の部分列の集合に対応させることになる。mは二つの文字列に対応する部分列の集合間の完全一致している要素の数を用いて定義することとする。

マッチング関数としてよく用いられるものを表2に示す。

表2　マッチング関数

| | |
|---|---|
| Diceの係数 | $\dfrac{2\,\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$ |
| Jaccardの係数 | $\dfrac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$ |
| Cosine係数 | $\dfrac{\lvert A\cap B\rvert}{\lvert A\rvert^{1/2}\cdot\lvert B\rvert^{1/2}}$ |
| Overlap係数 | $\dfrac{\lvert A\cap B\rvert}{\min\{\lvert A\rvert,\lvert B\rvert\}}$ |

（ただし、A，Bは集合，|A| は集合Aの要素の数）

これらは、部分列間のマッチングに適用した場合、図1に示すように特性に差が認められる。

情報検索に用いる場合、一致している部分の長さの変化に対応して値の変化が大きい点から、Jaccardの係数を採用し、次の様にファジィマッチング関数m'を定義した。

$$m'(v,w)=\frac{\lvert sub(v)\cap sub(w)\rvert}{\lvert sub(v)\rvert+\lvert sub(w)\rvert-\lvert sub(v)\cap sub(w)\rvert}$$

（ただし、sub(v)は、vの部分列の集合とする。）

---

Fuzzy matching functions and its application for information retrieval

Teruhisa Miyake, Sadaaki Miyamoto and Kazuhiko Nakayama
University of Tsukuba

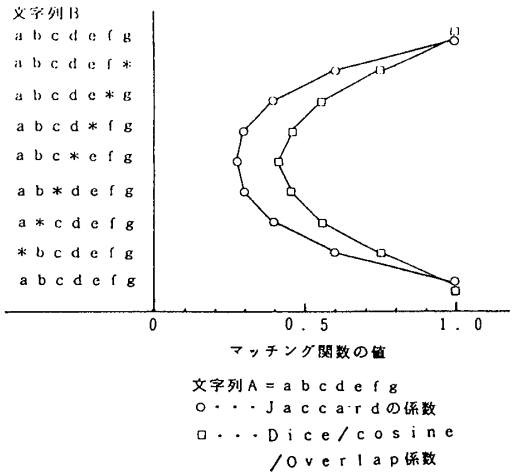


図1　部分列間のマッチングにおける特性の差

## 4．情報検索システムへの応用

ここで定義したファジィマッチング関数を、図2に示す文献データベースを対象としたファジィ情報検索システムに適用した。このシステムにおいては、検索質問中のキーワードはファジィシソーラス中の用語と比較され、一致した用語の関連語のファジィ集合に展開され、ファジィ索引により検索されて出力文献のファジィ集合を出力する。

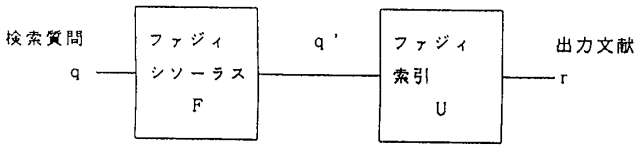


図2　ファジィ情報検索システム

ファジィマッチング関数を、（1）検索質問中のキーワードとファジィシソーラスの用語、（2）ファジィシソーラスで展開されたキーワードの関連語とファジィ索引の二つのマッチングで用いることができるようファジィ情報検索システムの拡張を行った。ファジィマッチング関数を用いるかどうかは利用者がコマンドで指定するようにしている。ファジィマッチング関数を用いた検索の例を図3に示す。同一の検索環境において、完全一致によるマッチングでは出力が得られない場合であっても、ファジィマッチング関数を用いた場合は、最も似た索引語を持った文献を得ることができる。

## 5．ＤＮＡ塩基配列のホモロジーサーチへの応用

部分列の集合を用いたファジィマッチング関数のもう一つの応用として、ＤＮＡの塩基配列間の類似性を求めるホモロジーサーチが考えられる。ホモロジーサーチでは対象とする塩基配列間の長さは必ずしも同程度ではないため、マッチング関数としては、Ｊａｃｃａｒｄの係数ではなくＯｖｅｒｌａｐ係数の方が適している。また、ＤＮＡの場合は、配列を単独の塩基とアミノ酸に対応したコドンのうちいずれのレベルで取り扱うか、マッチング関数の値に加えて最適な一致配列を求める必要がある等、文献データの場合とはいくつかの違いがある。しかしながら、実験的に求められた塩基配列に対して、ＤＮＡデータベースから最も類似したＤＮＡやその部分配列を探すといった用途には適している。

## 6．おわりに

ファジィマッチング関数を用いたファジィ情報検索システムでは、対象となる索引語等の全てとマッチングを行っており、完全一致の場合に比べて膨大な処理時間を必要とする。しかしながら、従来の部分一致と比べて利用者が意識せずに使用できるという利点がある。また、核酸配列のホモロジーサーチの場合は、従来から総当たりで処理を行っているため同様の処理時間で実行可能である。ファジィマッチングに適した索引やデータ構造の検討等、いくつかの問題点が残されているが、ファジィマッチング関数は、情報検索における有効なツールとなるものと考える。

〔参考文献〕

1．三宅、宮本、中山、ファジィ情報検索システムのためのマッチング関数、第5回ファジィシステムシンポジウム講演論文集(1989)、pp.439-440

2．L. Kohout, E. Keravnou, W. Bandler, Automatic documentary information retrieval by means of fuzzy relational products, TIMS/Studies in the Management Sciences, 20(1984), pp.383-404

3．S. Needleman, C. Wunsch, A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins, J. molec. Biol., 48, pp.444-453

```
) FSEA FUZZY ORDERING RELATION
 1 .FUZZY ORDERING RELATION(1.0000000 )
*** RESULT  : TERM(S) =   1
FSEARCH RESULT EMPTY
```

従来の検索の場合

（出力が得られない。）

```
) FSEA FUZZY ORDERING RELATION
 1 .FUZZY ORDERING RELATION(1.0000000 )
 2 .FUZZY RELATIONS(0.22222221 )
 3 .BOOLEAN STRATEGIES(0.22222221 )
 4 .SIMILARITY MEASURES(0.22222221 )
*** RESULT  : TERM(S) =   4
           : LAYER 0      0 DOCUMENT(S)
           : LAYER 1      5 DOCUMENT(S)
           : LAYER 2      0 DOCUMENT(S)
           : LAYER 3      0 DOCUMENT(S)
           : TOTAL        5 DOCUMENT(S)
```

ファジィマッチングの場合

図3　ファジィ情報検索システムへの応用例