自己組織化による研究情報の構造化 －PCMR への応用

6E－8　門脇 道子 *　藤原 譲 **
* 筑波大学　工学研究科　** 筑波大学　電子情報工学系

1 はじめに

計算機における記憶容量が増大し、データ、知識が複雑化し、計算速度の向上が実現するに伴い、情報の蓄積、管理、検索機能に加えて、高度な情報処理が可能なシステムが要求されている。

最近では、科学データベースが大きな研究対象となっており、そのサポートのための新しいデータベースの模索が始まっている。科学データベースとは、科学者が利用するためのデータベースであり、文献データベースのみでなく、科学の研究の対象として用いられる専門領域内の必要なデータを全て集めたものを意味している。

情報ベースシステムは、複雑かつ大量である情報を自己組織的に構造化し、研究開発支援の一環として類推、帰納、仮説推論などの高度な処理能力を持つシステムである。本研究では、高分子の C-13NMR に関する研究論文や測定データを、約２０年間に渡って蓄積してきたデータである PCMR-DB を情報ベースとして構築することを目標としている。

2 情報ベースシステム

2.1 情報ベースシステムの概要

情報ベースシステムは情報の特性に従い、情報の意味的構造を利用して意味処理を行なう。さらに、情報ベースシステムは大量かつ多様なデータを扱うため、自己組織的に情報を構造化する機能がある。情報の特性とは、概念の内部構造、相対性、重なり等である。情報の意味的構造は、次の３つである。

物理構造 所在情報や書誌情報に対応する。

概念構造 概念間の様々な関係に対応する。同値関係、上下関係、類似関係などである。

論理構造 論理関係など文脈や状況に応じた概念間の関係に対応する。

類推, 帰納推論などの高度な処理を行なうためには、概念構造は静的なものだけではなく、要求に応じた意味構造を利用時に自動生成しなければならない。

概念構造と論理構造を生成する際、従来のハイパーグラフを入れ子構造、ラベル付け、有向性を持たせて拡張した拡張ハイパーグラフ (Extended Hyper Graph) を利用する。これにより、概念の内部構造、重なり、相対性を扱うことが可能になる。

Structuralization of Research Information by Self-Organization
On the Application to PCMR-Information
Michiko KADOWAKI*, Yuzuru FUJIWARA**

3 情報ベースにおける概念の構造化

3.1 PCMRDB

PCMRDB は、高分子データベース研究会が編集、作成した主として高分子の C-13 核磁気共鳴に関するデータベースである。これは、C-13NMR[1] に関する既発表論文と本データベースの活用のために測定したデータ等を、この分野の研究者が利用しやすいデータや文献として編集したものである。それぞれの文献は一定のフォーマットに従ってレコードとして格納されている。レコードは大項目に分けられているが、特に HD と呼ばれるレコードの最初の一行に、レコードの一貫番号である IDNO、そのレコードの内容や化合物名を８文字以内の省略語として表現した REFCOD、その他、検索や分類に必要な情報が入力されている。本研究では、高分子研究に関する用語について特に PCMRDB の REFCOD を解析し、概念構造を構造化する。解析することにより自己組織化が可能になり、構造の再利用が出来るようになる。

3.2 REFCOD （レファレンスコード）

REFCOD は、化合物名或いはレコードの内容を簡潔に表現して見出しとして利用することで、検索を容易にすることを目的とした８文字以内の抽象語である。REFCOD は、特定の専門分野の視点から一般的事象を抽象化し、得られた概念を記号化したものと考えられる。よって、この記号には構文論的、意味論的、語用論的な規則があると考えられる。

REFCOD は、頭文字と名詞の組み合わせから成る。REFCOD は、頭文字によって大きく HOMOPOLYMER, COPOLYMER などに分類する事が出来る。REFCOD 中の名詞は、FRAGCOD という語の組み合わせから成っている。FRAGCOD は化学物質名の語の要素である FRAGMENT をコード化したものである。

3.3 REFCOD の構造化

REFCOD の概念を構造化するにはまずその中に含まれる名詞を抽出しなくてはならない。名詞を抽出するためにこれを構成する FRAGCOD を抽出する。FRAGCOD を抽出する際、FRAGCOD 辞書を利用する。これは、FRAGCOD とその品詞、FRAGMENT を対応させているものである。ただし、辞書を利用して抽出した FRAGCOD は、必ずしも正しいものではない場合がある。抽出された FRATCOD は単独のものであり、その組み合わせ、つまりその連続したものが REFCOD の構成要素である名詞としてふさわしくない場合があるからである [?]。

そこで、一度抽出されたFRAGCODを品詞的にとらえ、その組合せがREFCODの構成文法に適合しているかどうかを調べる。FRAGCODの品詞をを以下のように定める。

- 名詞の接辞部 (LN)
- 名詞の付属部 (AN)
- 修飾名詞 (MN)
- 一般名詞 (N)

FRAGCODはREFCOD中の名詞を構成するものであると考え、FRAGCODの組み合わせがREFCOD中の名詞として構成文法に適合しているものを、名詞部分としてとらえる。このように意味的に正しいFRAGCODが抽出されたかどうかを評価し、正しいと判断されたものについて概念の構造化を行う。これ以外の組み合わせからなるFRAGCOD候補からなるREFCODは、概念構造化の対象から削除する。

構造化には、SS-KWIC[3]という手法を利用し自己組織的にこれを行う。さらに、頭文字による分類と併せて視点に合わせた概念を抽出していく。

3.4 概念構造の抽出

全てREFCODの構成文法と一致する組み合わせから成るFRAGCODから構成されるREFCODについて概念の構造化を行う。概念の構造化を行う際、一つの分類法は頭文字を利用するものである。さらに、抽出されたREFCODを利用し、SS-KWICにより概念の同値、階層関係を構築する。SS-KWICでは、REFCODに含まれる名詞、さらにそれに含まれるFRAGCODの位置等とREFCODの構成文法に基づき、概念の同値、階層関係を構築する。

FRAGCODの中でも特に修飾名詞でないものは、概念として重要な役割を持つ。同じFRAGCODを含むREFCODは、同値な性質を持つと考えられるし、さらに含まれるREFCODの長さにより階層関係を抽出することが可能となる。基準としたものより長いFRAGCODを持つことで、より細分化された情報が付与することになるので概念もより詳細なものとなる。特に注目する名詞やFRAGCODに関して、動的に構造化をすることができる。

4 実験

現在、登録されているREFCODは７３５件であり、その中で頭文字がＰとＣのものだけで全体の７６％を占める。今回、特に頭文字がＰとＣのREFCODについて解析を行なった。結果を表に示す。

Initial	入力 *REFCOD* 数	解析された *REFCOD* 数
P	372	265
C	187	64

よって頭文字ＰのREFCODに関してはREFCODからFRAGCODの抽出率は約７０％となる。

評価されたFRAGCODからなる頭文字ＰのREFCOD２６５個をSS-KWICにより構造化する。一つのFRAGCODに注目して、その下位となるREFCODにリンクを張り、重なっているものを除くと７４５本となった。

さらに、**V(LN)** と **AMD(LN)** に注目してノードとリンクの数を示す。

FRAGCOD	ノード数	リンク数
P	117	745
V	32	62
AMD	8	13

よって概念がより一般的なものほど、辿らなければいけないリンクの数が多くなる複雑な多重継承をもつと考えられる。

5 まとめ

今回、情報ベースシステムにおけるPCMRデータのREFCODについて概念構造構築の方法について検討した。REFCODの構造化は、まずその意味的な要素であるFRAGCODを切り出し、その構成を文法のルールに合わせて限定した。さらにこのようにして切り出されたFRAGCODの概念構造の構造化をSS-KWICで行なった。この構造化により視点に応じた動的な概念構造を構築することが出来る。特に品詞によって、同じFRAGCODも持つ意味がことなるので、そのことに注目すれば概念の多義性や曖昧性を扱う時に有効に利用し得るものとなる。現在、REFCODを日本語の物質名と対応させた対訳辞書とPCMRDBの化合物名項目の中に記載されている同義語、またキーワード等の利用によって、新たに一般的な概念の構造化を行なっている。さらに、PCMRデータによる情報ベースに特に要求される状況に応じた推論、主に一般的ルール発見を行なう帰納推論とを併せて有効な研究開発支援システムの構築を行なっていく。

参考文献

[1] 石塚 英弘, 竹内 敬人. C-13 NMR 基礎と応用, 講談社,1976.

[2] 小川 泰嗣, 望月 雅子, 別所 礼子. “複合語キーワードの自動抽出法”, 情処研究報告 NL-97-15,pp.103-110(1993).

[3] Jingjuan Lai, Hanxiong Chen, Yuzuru Fujiwara Extraction of Semantic Relationships Among Terms-SS-KWIC Proc.of the 47th FID Conferene and Congress pp.155-159, 1994.