

自己組織化による研究情報の構造化

高分子情報への応用

3B-1

門脇 道子*

藤原 譲**

* 筑波大学 理工学研究科

** 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の増大、計算速度の向上に伴い、情報の蓄積、管理、検索機能に加えて、人間が日常行なっている思考のような高度な情報処理が可能なシステムが要求されている。

情報ベースシステムは、複雑かつ大量の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[1] に関する研究論文の内容を、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 情報の構造化による情報ベースの構築

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、重なりなどである。情報の意味的構造とは、原情報の位置関係を示す物理構造、同値、上下関係などの概念構造、文脈や状況に応じた関係を指す論理構造の 3 つである。

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拡張した拡張ハイパーグラフ (Extended Hyper Graph)[2] を利用する。拡張ハイパーグラフは、基本のハイパーグラフに入れ子構造、再帰構造、ラベル付け、有向性を持たせることによりこれを拡張したものである。

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。本研究では今回、高分子研究に関する用語について特に PCMR-DB の材料表現を解析し、自己組織的に構造化することについて考察した。

3 PCMRDB の自己組織的構造化

PCMRDB は、高分子データベース研究会が編集、作成した高分子の C-13 核磁気共鳴データベースである。これは、C-13NMR に関する既発表論文および研究会で測定または収集したデータを編集し、この分野の研究者に利用しやすい形でデータや文献を編集したものである。この PCMRDB は文献検索、数値データの相互比較など総合的な利用を図ることを目的としている。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。レコードは大きな項目に分かれていて、ヘッダと呼ばれるレコードの最初の一行には、レコードの識別番号である IDNO、そのレコードの内容や化合物名を 8 文字以内の省略語で表現した REFCOD、その他検索や分類に必要な情報が入力されている。

本稿では、この PCMRDB の材料名及び REFCOD からもとの用語やその概念の判定を計算機によって実現し、動的に概念構造を構築する方法について述べる。

3.1 REFCOD

PCMRDB の REFCOD とは、前述のデータベース中で化合物名或いはレコードの内容を簡潔に表現した 8 文字以内の抽象語である。この抽象語は、検索を容易にするだけでなく、もとの用語の意味を反映している。PCMRDB では、特に共重合体等の語の断片である FRAGMENT と、その短縮形コードの FRAGCOD を定義し、これらを利用して REFCOD の作成に役立てている。REFCOD は物質別に付けられる事が主であるが、NMR の実験方法に重点があるものや総括的な内容によって付けられる場合もある。

Structuralization of Research Information by Self-Organization
On the Application to PCMR-Information

Michiko KADOWAKI*, Yuzuru FUJIWARA**

*Master's Degree Program in Sciences and Engineering,
University of Tsukuba

**Institute of Information Sciences and Electronics, Uni-
versity of Tsukuba

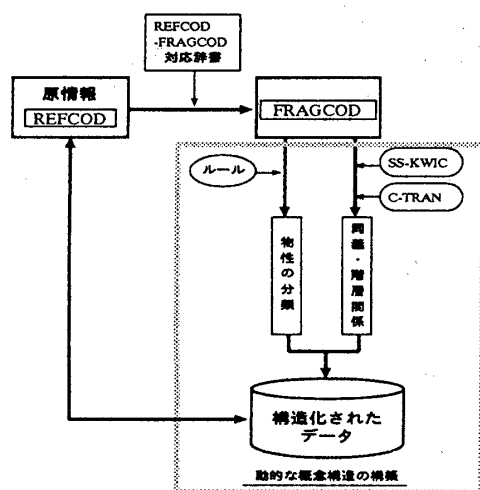
3.2 REFCOD の概念構造の構造化

構造化を行なう際、視点を何に置くかによって概念構造は動的に変化する。まず、大きく分けて以下のようなことに注目する。

- 物質名かそれ以外のものか。
- 化学的性質。
- 物理的性質。物質の構造。

レコードはまず REFCOD の頭文字により分類し、構造化する事ができる。頭文字は、P、C、M、R、N、B、G、L、-、*、の10種である。レコードの内容がある物質についてのものか、または、PCMRDBの利用法などについて述べているものなのかを REFCOD の頭文字によって分類し、構造化する。

さらに、REFCOD の中に含まれる FRAGCOD の情報により REFCOD の持つ概念を構造化することができる。REFCOD からの FRAGCOD の切り出しには、REFCOD-FRAGCOD 対応辞書を利用する。切り出された FRAGCOD と、対応している FRAGMENT の意味をもって REFCOD を分類していく。PCMRDB の REFCOD の自己組織的な概念構造の構築の概要を以下に示す。



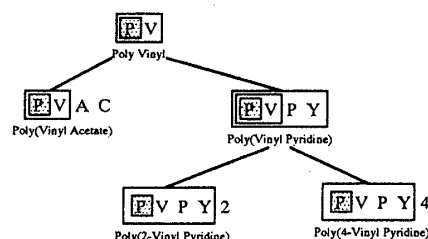
REFCOD の概念構造の構築

化学的な性質については、特定の FRAGCOD に注目することにより動的に概念構造を得ることが出来る。今回、物質の物理的性質、つまり物質の構造に注目し、これを自己組織的に構造化する。この分類を行なうためには幾つかのルールが存在する。物質の構造により、その物質の持つ化学的な性質の判断が可能になる。

物質の構造による分類と、頭文字やその他の FRAGCOD の概念構造と合わせて REFCOD を構造化していく。その際、C-TRAN、SS-KWIC という情報ベースにおける概念の同値、上下関係抽出システムを利用する。SS-KWIC は、ここでは、REFCOD に含まれる FRAGCOD の数と位置等により REFCOD 作成規則に基づき、同義、階層関係を構築する。SS-KWIC により注目する FRAGCOD に関して動的に概念の同値、上下関係を構築することが可能となる。以下に、

- FRAGCOD として V が抽出される。

というルールにより、Vinyl Polymer として分類された REFCOD の概念構造を示す。



Poly Vinyl に関する概念構造の例

この構造化により、注目する物質の構造による性質や化学的性質等の情報を得ることが出来る。

4 考察

今回、PCMRDB の REFCOD から自動的に概念構造を作成する方法を考察した。REFCOD は、一種の自然言語に基づく抽象語である。それ故に、より正確に REFCOD から FRAGCOD を抽出し、さらに概念構造を得るために自然言語的な処理を行なう必要がある。

現在、REFCOD を日本語の物質名と対応させた対訳辞書を作成中である。これと PCMRDB の化合物名項目の中に記載されている同義語を利用することによって、新たな概念の構造化を進めて行く。

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] N. Uda, Y. Fujiwara, Analysis of Information for Its Modeling, ASIS(To be publishe), 1993.