化学グラフデータベースシステムの設計と構築

4Q-4

°霜山　友肖、欒　玉琴、北川　博之、大保　信夫、藤原　譲（筑波大学）

1．はじめに

近年、図形、数学的グラフ等の非数値データを対象とするデータベース技術への要求が高まっている。しかし、既存のＤＢＭＳにおいて、これらの構造を持つ非数値データを表現することは困難である。

そこで、本稿では、関数型データモデルに抽象データ型の概念を導入したデータベースが、この様な構造を持つデータに有効であることを、化合物構造情報を対象とした化学グラフ構造データベースシステムを例にして述べる。

2．化学グラフ構造データベースシステム

化学の分野において、グラフとして表現される化合物構造情報を対象とした、グラフデータベースシステムに対するニーズが高まってきているが、従来のＤＢＭＳでは、この様な構造情報は、複数のレコードにわたるデータ構造として表現されるために、一つのグラフをアクセス単位として認識しようとすると、非常に複雑なデータ表現、操作を強いられることになる。

この様な問題を解決するための方策として、現在様々なアプローチが研究されているが、現在主流となっている有効なアプローチとして、オブジェクト指向型データベースが挙げられる。また、これを実現する手段として、関係型データベースの拡張や関数型データモデルの研究が盛んである。関数型データモデルでは、実在する関連を柔軟に辿ることはできるが、データベース中に存在するデータ、あるいは、複数のレコードにわたるデータ構造に付随する応用世界の意味を、表現操作することが困難である。そこで、複数のレコードにまたがるデータ表現を簡単に操作できる機構を作り、化学グラフ表現を一つのデータの単位として扱いたいという要求から、関数型データモデル上に、データのセマンティック表現を可能にするために、抽象データ型の概念を導入し、ひとつの化学グラフ構造をひとまとまりのオブジェクトとして認識し、それによって、効率の良いデータ操作が行えるような、化学グラフ構造データベースシステムを構築することにした。

3．化学グラフ構造の認識モデル

化学グラフ構造は、データベース内のデータとして表現する場合、全体のグラフとその部分構造の階層関係として表現することができる。化学グラフの部分構造を一意に決定するために、スーパブロック分割法という手法を用いる。化学グラフ構造を部分グラフ構造に分割して、その部分グラフ構造をデータベースにおけるアクセス対象とすることは、化学グラフの同型性の判定のコストを著しく減少することになり、化学グラフ構造データベースに対するアクセスの効率化が期待できる。

（１）スーパブロック分割法

スーパブロック分割とは、化学グラフを環状構造部分と非環状部分に一意に分割する分割法のことである。分割されたスーパブロック（ＳＢ）間の対応を表現するために、スーパブロック・コネクショントリー（sc tree）を用いる。図１に化学グラフをＳＢ分割した例を示す。

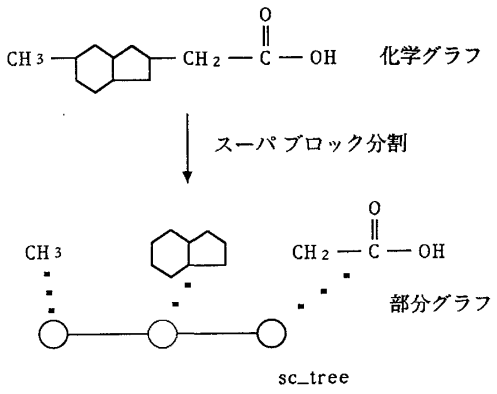

図1 スーパブロック分割の例

（２）概念モデル

（１）で定義された部分構造を一つのノードとみなすと、化学グラフは、部分構造に対応するノードの集合（Ｎ）と部分構造同志のつながりを表すエッジの集合（Ｅ）として表現される。化学グラフを一つのデータ型（compound）として表すと、compound はＮとＥの組合せで表現される sc_tree というデータ型に一意にマッピングされる。また、sc_tree の個々のノードは、

Design and Implementation of Chemical Graph Database System
Tomonori SHIMOYAMA, Yuquin LUAN, Hiroyuki KITAGAWA, Nobuo OHBO, Yuzuru FUJIWARA
University of Tukuba

部分構造となるNの中の一つの要素（super_nodes）に展開され、super_nodes に関数を適用することによって、グラフ表現（chem_graph）を表すことができる。

これによって、化学グラフ構造を一つのデータ単位として容易に扱うことができる。図２に、化学グラフ構造の認識モデルを表すデータ型と、それに付随する関数群の一部を示す。＊印がついたデータ型は、そのデータ型のインスタンスを要素とするリストの形式を表すことにする。例をとって説明すると、compound 型に付随する関数 include は、compound 型のインスタンスを引数にとり、結果としてその化合物に含まれるsuper_nodesのリストを返すことを表す。

４．化学グラフの表現例

上述した化学構造グラフの認識モデルを用いた、化合物の表現例を図３に示す。ここで、○印は関数の合成を表し、関数名の前につく＊印は、その引数リストの個々の要素に関数が適用されることを表す。ここで、c#1,sn#1,sn#2,sn#3,sc#1はそれぞれの型に対応するインスタンスを表し、それにその型に付随する関数を適用することによって、化学グラフ構造を表現することができる。

また、想定されるデータベースにおける問い合わせとして、部分構造検索、検索結果の選択、検索結果の集合に対する演算などが挙げられるが、図４にその一例を示す。この例において、IN 述語は、IN の後に現れる集合の中に、IN の前に現れるものが存在するかどうかを判定する述語である。PRINT 述語は、chem_graph 型のインスタンスの並びに対して適用され、結果としてそのグラフ表現を部分グラフ同志のつながりも含めて出力する。

５．おわりに

提案した化学グラフ構造認識モデルにより、化学グラフ構造の効率的な表現、操作が可能であることが分かった。このモデルは、一般的に、階層構造を持つデータの表現に適用可能であると考えられる。

現在、ＵＮＩＸマシン上の関係データベースをベースとし、その上に関数型データモデルの関数を取り扱う機構と、抽象データ型を表現できる機構を持つシステムを構築中である。今後、本システムがこの様な構造を持つデータを操作するために有効であるかどうかを検証していく。

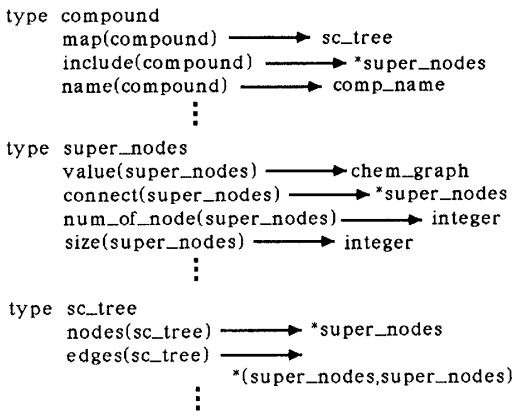

図2 化学グラフ構造認識モデルにおけるデータ型

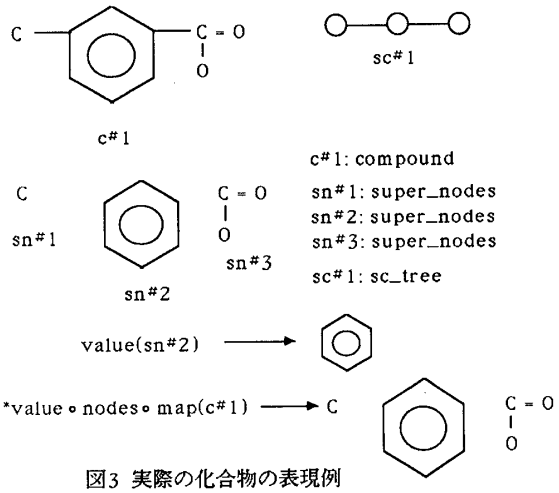

図3 実際の化合物の表現例

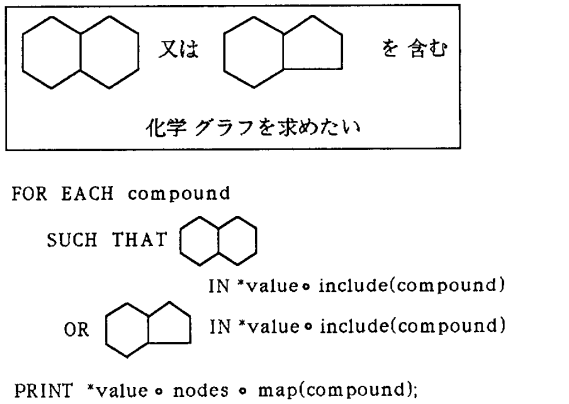

図4 データベースに対する問い合わせ例

参考文献

（１）黒沢他：化学グラフのデータベース，情報処理学会第２５会全国大会 6P-2,6P-3 論文集 705-708

（２）DAVID W.SHIPMAN：The Functional Data Model and the Data Language DAPLEX,ACM TODS(1981)

（３）P.BUNEMAN,R.E.FRANKEL,R.NIKHIL：An Inplementation Technique for Database Query Languages, ACM TODS(1982)