データベースシステム 119−52
（１９９９．７．２２）

複合階層索引の質問記述能力

河野弘史　　古川哲也

九州大学経済学部

データの流通量の増大と多様化によって，収集した大量のデータを整理し，必要に応じて利用できるようなシステムの開発が重要となっている。データを性質ごとに分類した階層を組み合せることで，そのような要求に柔軟に対応することが可能である。本稿では，分類階層を合成して得る複合階層を用いて必要なデータを指定する方法について述べ，その質問記述能力が十分であることを示す。検索システムを構築する際には，利用者の利便性とシステムの効率性を考える必要がある。利用者に負担をかけずにデータを検索するためのシステムの機能についても議論している。

Query Expression Power of Complex Hierarchical Indices

Hirofumi Kawano　　Tetsuya Furukawa

Dept. Economic Eng., Kyushu Univ.

It becomes more important to construct the systems which arrange and serve a large amount of collected data. Complex hierarchies, which are combinations of classification hierarchies, cope with these requirements flexibly. In this paper, we discuss how to select data on complex hierarchies and show that the query expression power is sufficient. Convenience of users and efficiency of systems have to be considered when a retrieval system is constructed. The paper also discusses the system facilities to retrieve data without users' burden.

1. はじめに

近年の高性能化と小型化に伴い，計算機は様々な分野で利用されている。またネットワーク技術の発展は，流通するデータの多様化をもたらしている。そのため，様々な分野で大量の情報を蓄え，それを利用するシステムの開発が必要となっている。情報検索の分野では，そのような利用を実現するための様々な研究が行われている。

必要なデータを取り出すために，データを整理し，分類しておくことは有用である。収集したデータを整理し，利用できるようにするために，複合階層を用いてデータの分類をするものがある[1]。複合階層は，データを性質ごとに分類した複数の階層を合成することにより構成される。本稿では，複合階層を用いて必要なデータを指定する際の記述能力について検討する。

データを様々な視点から分類した分類の階層を，すべて準備しておくことは現実的ではない。性質ごとに分類して，それを組み合せることにより対応する。作成した複数の分類階層を用いてデータの検索を行うときに，必要なデータを記述する能力が十分かという問題がある。[2] では，分類階層を独立に使用し，必要なデータを記述することを検討しているが，論理和や否定の記述を想定しておらず，十分ではない。分類階層を合成した複合階層を用いることによって，十分な記述能力が得られるかどうかについて検討する。

大量の多様なデータを利用するとき，必要なデータを記述するためには，質問記述のためのデータの構造が必要である。[3, 4] は半構造データに対して，スキーマと同様の構造を適用することを議論している。また半構造データに対する問い合わ

せの研究も行われている[5, 6]。本研究はこのような分野にも応用が可能である。データの分類については，データベースのアクセスを高速化するためのクラスタリングに関する研究がある[7]が，データの物理的な構成を議論するものがほとんどである。

実際に検索システムを構築するときには，利用者にとって使いやすいかということや，システムにとって効率がよいかということを検討しなければならない。利用者に負担をかけず，求めるデータを的確に検索でき，それをシステム側で効率よく支援することが望まれる。

本稿では，複合階層索引を用いた検索システムの記述能力について検討するとともに，利用者の使いやすさや検索の処理効率の点から検索システムの構成について考察する。2節では，複合階層を構成する手法について述べる。3節では，必要なデータを指定する際の複合階層の記述能力は十分であることを示す。4節では，複合階層を用いた検索システムの効率性について検討する。5節では，利用者を支援する検索システムはどうあるべきかについて議論する。6節は全体のまとめと今後の研究課題である。

2. 複合階層の構成

検索の対象となるデータをオブジェクトとし，分類によって生成されたオブジェクトの集合をグループ g とする。またグループ g を構成するオブジェクトの集合を $m(g)$ で表す。分類によってできたグループをさらに分類すると，分類の階層ができる。

［例 1］ 図1 (a) はオブジェクト集合を地域という性質で分類した分類階層である。(a) の階層中のグループをさらに異なる性質で分類するのではなく，性質ごとに分類階層を作成する。(b) は同じオブジェクト集合を，産業という性質で分類した階層である。 □

複合階層の一部分として得ることができる階層をグループ階層とする。

［定義 1］ グループ g とその子グループ集合 $\boldsymbol{g}$ に対し，$(g, \boldsymbol{g})$ を原子階層，g を原子階層の親，$\boldsymbol{g}$ の要素を原子階層の子という。原子階層の集合 $H = \{(g_i, \boldsymbol{g}_i) \mid 1 \leq i \leq k\}$ は，その親がすべて異なり $(g_i \neq g_j\ (i \neq j))$，$H$ 中の他の原子階層の子とはならない親が唯一であるとき，グループ階層であるという。

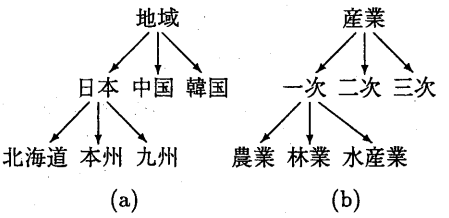


図1 分類階層

H 中の原子階層の親グループとはならない子グループを H の子グループといい，H の子グループの集合を $leaf(H)$ で表す。他の原子階層の子グループとはならない親グループを H の根グループといい，$root(H)$ で表す。 □

［例 2］ 図1 (a) の分類階層では，原子階層は h_1(地域, { 日本, 中国, 韓国 }), h_2(日本, { 北海道, 本州, 九州 }) である。グループ階層は $H_{11} = \{h_1\}$, $H_{12} = \{h_2\}$, $H_{10} = \{h_1, h_2\}$ となる。また図1 (b) では，原子階層は h_4(産業, { 一次, 二次, 三次 }), h_5(一次, { 農業, 林業, 水産業 }) であり，グループ階層は $H_{21} = \{h_4\}$, $H_{22} = \{h_5\}$, $H_{20} = \{h_4, h_5\}$ となる。 □

複合階層は，ある順序で複数個のグループ階層を組み合せたものである。グループ階層の合成を次のように定義する。

［定義 2］ グループ g_0 によるグループ g の制限 $g_0 * g$ は，$m(g') = m(g_0) \cap m(g)$ となるグループ g' である。グループ集合 $\boldsymbol{g}$ に対し，$g_0 * \boldsymbol{g} = \{g_0 * g \mid g \in \boldsymbol{g}\}$ とする。原子階層 $(g, \boldsymbol{g})$ の制限 $g_0 * (g, \boldsymbol{g})$ は，階層 $(g', \boldsymbol{g}')$ $(g' = g_0 * g, \boldsymbol{g}' = g_0 * \boldsymbol{g})$ である。

グループ g_0 によるグループ階層 H の制限 $g_0 * H$ は，グループ階層 $\{(g, \boldsymbol{g}) \mid (g, \boldsymbol{g}) = g_0 * (g', \boldsymbol{g}')(g' \neq root(H)), (g, \boldsymbol{g}) = (g_0, g_0 * \boldsymbol{g}')(g' = root(H)), (g', \boldsymbol{g}') \in H\}$ である。

グループ階層 H_i, H_j の合成 $H_i * H_j$ は，$H_i \cup \bigcup_{g \in leaf(H_i)} g * H_j$ である。 □

複合階層は，複数個のグループ階層の合成により構成される。定義より明からなように，グループ階層の合成演算は結合則が成り立つ，すなわち $(H_i * H_j) * H_k = H_i * (H_j * H_k)$ である。したがって，複合階層の定義では，グループ階層の順序のみが必要であり，合成演算の適用順序は問題とな

らない。

［定義 3］ 複合階層は，階層を制限するグループ g_0 とグループ階層 H_i $(1 \leq i \leq n)$ の列 $(g_0, H_1, H_2, \cdots, H_n)$ で定義される制限された原子階層の集合 $\{(g,g) \mid (g,g) \in g_0 * H_1 * H_2 * \cdots * H_n\}$ である。 □

複合階層は仮想的なグループ階層なので，すでに定義された他の複合階層を用いた再帰的な定義も可能となる。また，複合階層の定義で，階層を制限するグループを全オブジェクト集合とすれば，複合階層は，H_1 の親グループのオブジェクト集合に対する階層構造となる。

グループ階層 H_1 に対し，$leaf(H_1)$ に含まれるすべてのグループに他のグループ階層 H_2 を適用するのではなく，その一部のグループのみを H_2 の構造としたい場合がある。H_2 を適用するグループを特定するため，どのグループに H_2 を適用するのかを記述するように複合階層の定義を拡張するのは，軽微な定義の変更である。

［例 3］ 図 2 は図 1 の分類階層を合成した複合階層である。この複合階層の階層構造は，H_{10}(北海道$*H_{22}$, 本州$*H_{22}$, 九州$*H_{22}$) として定義できる。 □

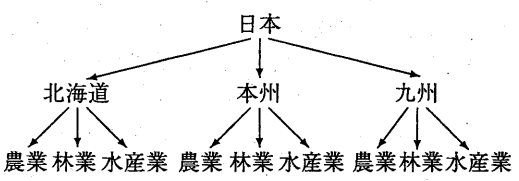


図 2 複合階層

このように，グループ階層はオブジェクトが実際に分類された階層構造であり，複合階層は分類の仮想的な構造である。

3. 複合階層の記述能力

複合階層の特徴は，主となる階層のグループを他の階層によって分類することである。本節では，合成してできた複合階層で，必要なデータを指定する方法の記述能力について検討する。

階層の数を n 個とする。複合階層を用いずに，n 個の階層を独立に使うことでデータの指定ができる[2]。各階層に対して 1 つの性質を指定する。それらをすべて満たすオブジェクト集合は，i 番目の階層での性質を示すグループを g_i とすると，

$$m(g_1) \cap m(g_2) \cap \cdots \cap m(g_n) \qquad (1)$$

で表すことができる。各階層において複数の性質を指定できるようにすると，データ集合は，

$$\begin{aligned} & (m(g_{11}) \cup m(g_{12}) \cup \cdots \cup m(g_{1m_1})) \\ \cap\ & (m(g_{21}) \cup m(g_{22}) \cup \cdots \cup m(g_{2m_2})) \\ \cap\ & \cdots \\ \cap\ & (m(g_{n1}) \cup m(g_{n2}) \cup \cdots \cup m(g_{nm_n})) \end{aligned} \qquad (2)$$

となる。ここで g_{ij} は i 番目の階層で必要なデータのグループであり，m_i は i 番目の階層で指定する性質の数である。この記述では，1 つの階層のグループは他の階層で同じ性質を持つことになる。例えば $(m(g_{11}) \cup m(g_{12})) \cap (m(g_{21}) \cup m(g_{22}))$ は，1 番目の階層のグループ g_{11} と g_{12} のどちらに対しても，2 番目の階層の性質は g_{21} または g_{22} であり，異なる性質を記述できない。これは性質ごとに論理和を適用しているためで，論理式における和積標準形の特殊な形に対応する。

任意のデータを指定できる記述は，論理式の和積標準形または積和標準形に対応する形式となる。グループ g_i に含まれないという否定の記述は，全体の集合が定まっているので，g_i を除いた他のグループの和集合 $g_1 \cup g_2 \cup \cdots \cup g_{i-1} \cup g_{i+1} \cup \cdots \cup g_n$ で表すことができる。積和標準形に対応するデータ指定の一般形は，

$$\begin{aligned} & (m(g_{11}) \cap m(g_{21}) \cap \cdots \cap m(g_{n1})) \\ \cup\ & (m(g_{12}) \cap m(g_{22}) \cap \cdots \cap m(g_{n2})) \\ \cup\ & \cdots \\ \cup\ & (m(g_{1m}) \cap m(g_{2m}) \cap \cdots \cap m(g_{nm})) \end{aligned} \qquad (3)$$

となる。この形式で表されるデータ集合を指定できれば，十分な記述能力がある。

複合階層は階層の合成を行うもので，積項 $m(g_{1j}) \cap m(g_{2j}) \cap \cdots \cap m(g_{nj})$ は，$H_1 * H_2 * \cdots * H_n$ における経路で指定することができる。求めるデータはその和集合となるので，合成の拡張を利用して，$H_1(g_{1j} * H_2(g_{2j} * \cdots * H_n))$ における g_{nj} として指定する。

このような階層の合成は必ずしも直接できるわけではない。g_{ij_1} と g_{ij_2} に包含関係がある，すな

わち g_{ij_1} が g_{ij_2} の祖先のとき，g_{ij_1} は g_{ij_2} への階層と他の分類による階層が必要なため，2つの異なる性質の子孫を持つこととなる。すなわち (3) 式の全ての積項を直接指定できるような複合階層を構成することはできない。逆に包含関係がないときには，すべてのグループ g_{ij} は子孫のグループを持たないので，他の階層を直接合成できる。したがって包含関係があるときには，包含関係がない形に変換すればよい。$m(g_{1j_1}) \cap \cdots \cap m(g_{ij_1}) \cap \cdots \cap m(g_{nj_1})$ で，$m(g_{ij_1})$ を g_{ij_2} のレベルのグループ集合 $\{g_{ik_1}, g_{ik_2}, \cdots\}(\ni g_{ij_2})$ に対して，$m(g_{ik_1}) \cup m(g_{ik_2}) \cup \cdots$ で置き換える。置き換えることで新たな包含関係は生じないので，包含関係に対して，繰り返しこの操作を適用することにより，積和標準形に対応する形式を，すべての階層でグループの包含関係のない形式に変換することができる。

［例 4］　九州の農林業と農業については日本全体のデータを得る場合を考える。(3) の形式による記述は，(九州 ∩ 農林業) ∪ (日本 ∩ 農業) である。九州は日本の子孫であり，日本は産業への分類も必要であるから，2つの異なる性質の子孫が存在することになる。そのままでは1つの階層にできないので，日本を (北海道 ∪ 本州 ∪ 九州) で置き換えて (北海道 ∩ 農業) ∪ (本州 ∩ 農業) ∪ (九州 ∩ 農業) ∪ (九州 ∩ 林業) とする。これにより1つの複合階層を得ることができる。図2は結果の複合階層である。質問の記述は，この階層で各積項に対応するグループを選択することになる。　□

以上のように，階層を合成した複合階層で任意の必要なデータを指定することができる。すなわち，複合階層はデータの指定に対して十分な記述能力を持つ。

複雑な指定が必要な検索を行う際に，どのように複合階層を合成すればデータの指定が可能になるかを考えることは，利用者にとって大きな負担である。利用者が必要なデータを指定する方法として，次の場合が考えられる。例えば，九州の農林業と日本の農業に関するデータの検索では，図1の分類階層において，(a) の階層で九州，(b) の階層で農業と林業を指定した検索結果と，(a) の階層で日本，(b) の階層で農業を指定した検索結果の和集合をとる。これは (2) 式を複数回適用することに対応する。この場合，利用者にとってデータ指定の記述は容易であるが，利用者またはシステムが結果の和集合をとる必要がある。(2) 式を複数回適用することは，複合階層を複数回利用すること，すなわち (3) 式を複数回適用することで代替できる。

4. コストの比較

積和標準形に対応した形式で任意のデータ指定を行う方法は，包含関係のない形に変換し，1つの複合階層を合成して指定する方法と，複合階層を複数回用いてデータを指定する方法の2通りがある。これは，複数の複合階層で指定したデータをシステムで処理できるかどうかで考えることができる。(3) 式と同等の記述能力を持つ方法は次のようになる。

(1) 単指定：1つの複合階層で指定する
- 単検索：1回の検索で必要なデータを得る
- 複数検索：複数回の検索に分けて必要なデータを得る

(2) 複数指定：複数の複合階層で指定する

単指定は，1回の検索で利用できる複合階層が1つのときである。このとき (3) 式に対応する指定を行う方法として，次の2つがある。単検索は，すべての必要なデータを1つの複合階層で指定する場合で，複数検索は，複合階層でデータが指定された検索を複数回行い，利用者が結果の和集合をとる場合である。複数指定は，同時に利用できる複合階層が複数であり，単指定複数検索で利用者が行う和集合の処理を，システムが行うものである。特別な形として，複数の分類階層で必要なデータを複数回指定する場合も含まれる。

システムにとってどの方法の効率がよいかを検討するために，単指定と複数指定のコストを比較する。比較の際に，グループのオブジェクトを取り出すことにコストはかからないと仮定する。共通集合と和集合の計算に関するコストのみを考え，それぞれについて次の記号を用いる。

- 分類階層のグループ g_i と g_j の共通集合を求めるコストを $I(g_i, g_j)$ とする。
- 和集合を求めるために重複を除くコストを U とする。

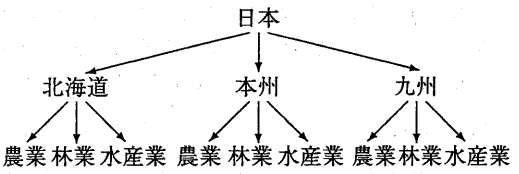
分類階層に対して，九州の農林業と日本の農業に関するデータを取り出す場合についてコストを比較する。

［複数指定］　複数指定の場合には，1つの階層で日本の農業，もう1つの階層で九州の農林業に関

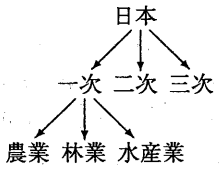
するデータを指定する。図3はそのための2つの複合階層である。まず図3 (a) の階層で九州の子の農業と林業を，次に (b) の階層で日本の子孫の農業を選択して和集合をとる。このときのコストは，

$$C_1 = I(\text{九州}, \text{農業}) + I(\text{九州}, \text{林業}) + I(\text{日本}, \text{農業}) + U$$

となる。



(a)



(b)

図3　複数の複合階層

[単指定単検索]　図2の複合階層に対して，例4で示したようにデータを指定する場合である。この場合図2の複合階層で北海道の子の農業，本州の子の農業，そして九州の子の農業と林業を選択する。このときのコストは，

$$C_2 = I(\text{北海道}, \text{農業}) + I(\text{本州}, \text{農業}) + I(\text{九州}, \text{農業}) + I(\text{九州}, \text{林業})$$

となる。

図4も図1の分類階層を合成して得た複合階層である。これは地域を主とするのではなく，産業を主とし，どの産業に対してどの地域のデータが欲しいかということを考えた場合である。このとき農業の子の日本，林業の子孫の九州を選択する。コストは，

$$C_3 = I(\text{農業}, \text{日本}) + I(\text{林業}, \text{九州})$$

となる。

複数指定と単指定単検索のコストを比較する。共通集合を求めるコストは，集合の表現方法，計

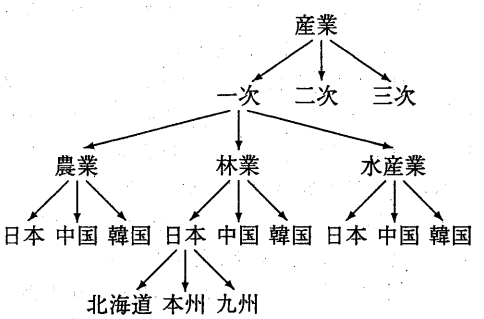


図4　順序を変えた複合階層

算方法やディスクのアクセス速度などに依存するため，グループ x, a, b で $m(a) \cap m(b) = \phi$ のとき，$I(x,a) + I(x,b) = I(x, a \cup b)$ であると仮定する。

$C_1 - C_2$ は，$(I(\text{日本}, \text{農業}) + U) - (I(\text{北海道}, \text{農業}) + I(\text{本州}, \text{農業}))$ であり，仮定より $I(\text{日本}, \text{農業}) \geq (I(\text{北海道}, \text{農業}) + I(\text{本州}, \text{農業}))$ であるから，$C_1 - C_2 > U$ となり，C_1 は C_2 よりも大きい。また $C_1 - C_3$ は $I(\text{九州}, \text{農業}) + U$ なので，C_1 は C_3 よりも大きい。

したがって，どちらの単指定単検索についても複数指定よりコストが小さい。C_2 と C_3 のどちらが小さいかについては，実現方法の違いによって異なる。

[定理1]　単指定単検索のコストは複数指定のコストより大きくなることはない。　□

(証明)　$I(x,a) + I(x,b) = I(x, a \cup b)$ という仮定により，1つの複合階層索引でのデータ指定では，階層の合成の順序が異なっていても，コストは全て同じである。複数指定で，各複合階層での検索結果に重複したデータがあると，そのデータを余分に求めるコスト及び，和集合により重複を除くコストが余分にかかる。　(証明終)

システムの内部では，1つの複合階層におけるデータ指定，またはそれに対応する (3) 式の形式での表現を用いるのが良い。

5. システム構成

検索システムの利用について考える。どのような情報を示すことができれば，利用者は必要なデータを的確に得ることができるか，またその際に，システム側でどのように支援すれば良いかというこ

とが重要である。本節では，そのような検索システムの構成について議論する。

単指定は利用者にとって負担が大きい。複数検索は和集合をとるための重複を除くコストがかかる。単検索は利用者が1つの複合階層を構成するために，階層の組み合せ方やその順序について検討する必要があり，現実的ではない。したがって利用者にとっては複数指定が適している。

またシステムについては，複数指定を直接実現するのは重複を除くコストがかかり，効率的ではない。単指定単検索の方が効率が良い。

検索システムにおいては，利用者が複数の複合階層でデータを指定することで，システムは(2)式の形式の複数の情報を得る。その情報を(3)式に変換する。変換した(3)式に包含関係がないかを調べ，グループ間に包含関係がある場合には，3節で示したように，そのグループを子孫グループ集合で置き換えて，包含関係のない形に変換する。

検索の結果を利用者に分かりやすい形で示すために，結果の形を整える必要がある。例えば，結果が(九州, 農業), (九州, 林業), (九州, 水産業)になった場合には，(九州, 一次)にまとめたり，結果が(日本, 農業), (九州, 農業)となった場合には，(九州, 農業)を削除するといったことを行う。例4は日本を(北海道∪本州∪九州)で置き換えた結果，(北海道, 農業)∪(本州, 農業)∪(九州, 農業)∪(九州, 林業)となっているが，置き換えによって生じた(九州, 農業)の重複を除いている。

利用者が複数の階層で指定して，システムが階層を合成した結果，例えば，産業の階層を主として地域の階層を合成した階層を示し，次は順序を逆にした階層を示す。このようにシステムが更なる条件の変更を支援することができれば，より柔軟な指定が可能になる。

6. むすび

オブジェクト集合を複数の性質ごとに分類し，それらの階層を合成して得られる複合階層の記述能力について検討した。1つの複合階層でのデータ指定で十分な記述能力があるが，利用者の利便性を考えると，1回の検索で複数個の複合階層を用いて指定できる方が良い。またそのような利用者インタフェースに対するシステム構成をコストの面から検討し，システムが利用者に対して提供することができる機能について議論した。

分類階層をどのように物理的に実現するか，すなわち階層の情報をどのように記憶するかによって，単指定単検索のコストが変わる。したがって，階層情報の記憶方法に合わせて階層の適用順序を決定する必要がある。

データの絞り込みをするためには，どのような条件を適用すればよいか，また検索結果を増やすためには，どのように条件を緩和すればよいかということを利用者は理解できなければならない。そのように利用者の立場で検索システムを考えた場合，様々な視点から検索条件を見ることができるように，システムで支援する必要がある。

参考文献

[1] 古川哲也, “データの多重分類階層の構成”, 情報処理学会研究報告, 98-DBS-116-35, 平成10年7月.

[2] 河野弘史, 古川哲也, “複合階層索引を用いたデータ検索システムの利用者インタフェース”, 情報処理学会研究報告, 98-DBS-116-79, 平成10年7月.

[3] 宝珍輝尚, 都司達夫, “半構造データの構造表現のための動的スキーマの生成法について”, 情報処理学会研究報告, 99-DBS-118-5, 平成11年5月.

[4] Soe, D.-Y. et al, “Schemaless Representation of Semistructured Data and Schema Construction,” *Proc. Int'l Conf. and Workshop on Database and Expert Syst. Applications*, pp.387-396, 1997.

[5] Abiteboul, S., “Querying Semi-Structured Data,” *Proc. Int'l Conf. on Database Theory*, pp.1-18, 1997.

[6] Buneman, P. et al, “A Query Language and Optimization Techniques for Unstructured Data,” *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pp. 505-516, 1996.

[7] Zhang, T. et al, “BRICH: An Efficient Data Clustering Method for Very Large Databases,” *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pp. 103-114, 1996.