

論理式による内包情報のモデル化とその操作

6 A A - 5

掛下 哲郎

村田 美友紀

佐賀大学 知能情報システム学科

八代高専 情報電子工学科

kake@is.saga-u.ac.jp

mmurata@polaris.yatsushiro-nct.ac.jp

1 まえがき

データベースの高度化に伴い、その複雑さは増大の一途をたどっている。このようなデータベースを活用するためには内包情報の活用が不可欠である。ここで、内包情報とは実体集合（外延情報）の各要素に共通な性質を表現した情報を指す。その例として、知識発見によって得た情報^[1]、クエリー、一貫性制約などがある。内包情報の適切な活用によりデータベース理解を促進できる。また、再利用された内包情報は、より高度なデータベース活用のための構成要素にできる。

内包情報は数が多いため、これを活用するためにはデータベース化が必要である。本稿では内包情報をモデル化するために命題論理式を用いる。また、命題論理式を用いて表現された内包情報を検索する手法及び、命題論理式の編集を通じて対応する外延情報を一括更新する手法を提案する。

2 論理式による内包情報のモデル化

内包情報は複数の実体（複合実体を含む）を集約して得られる情報である。これを表現するためには論理的な枠組を用いることが自然である。内包情報のうちSQL/OQLクエリー及びビューのようにSPJ質問に対応しているものは、選択/結合条件を命題論理式で表現できる。なお、射影属性は外延的に定義された集合であるが、これに対する操作（検索や編集）は命題論理式に対する操作より容易なので、本稿では議論しない。また、知識発見によって得られた知識にもSPJ質問に対応するものが多く含まれる。その例としては、実体を分類するための決定木や相関ルール（支持度及び確信度は属性として別に扱う）などがある^[2]。

これに対して、全称/存在限量子を含んだ情報を表現するためには述語論理式が必要とされる。また、再帰的データ構造（木、リスト、グラフ等）の集合を内包的に表現するための手法も提案されている^[3]。

上述したように、内包情報を表現するためのモデル化手法は3種類に大別される。本稿ではこの中で最も基本的な命題論理式（以下、論理式と略記する）を用い

て表現される内包情報を操作対象とする。しかし、以下で提案する論理式の検索・編集手法は命題論理式とは独立性が高いため、他のモデル化手法にも適用できると予想している。

3 サンプルを用いた論理式検索

論理式を用いたデータベース検索を行うと、論理式中の自由変数の組に対応する実体が検索される。ここで、論理式を検索するために論理式を用いることはできない。これは充足可能性判定（NP完全問題）に対応するためである。これを避けるために、本稿ではデータベース中の実体を用いて論理式を検索する手法を提案する。

実体¹ e は、論理式 L を満足する時に L の正例、そうでない時に負例となる。サンプル S は実体の集合であり、各要素について正例/負例のいずれかが明示されている。 L が S 中の正例を全て満足し、負例を全て満足しない時、 S は L を検索すると定義する。

サンプルを用いた論理式検索は、以下の手順で行われる。(1) 単一の実体からなる初期サンプルを構成する。(2) サンプルによって検索された論理式が唯一に決定されるまで、サンプルに実体を追加する。この作業は利用者が行う必要があるが、それを支援するために以下の機構を用意している。

テンプレート機能

各論理式はスキーマの部分グラフを検索範囲としている。この部分グラフ（テンプレート）の属性値を設定することにより、リンク値の設定なしにサンプルを構成する実体を作成できる。テンプレートは論理式をデータベース化する際に自動的に作成できる。また、テンプレート自体をデータベース化した場合、目的とするテンプレートの検索はスキーマ中での部分グラフ検索に帰着できる。

論理式による実体の検索

論理式を用いた論理式検索は実際的でないが、論理式を用いてサンプルに属する実体を検索する手法は実用性がある。検索対象論理式を L とすると、 $L' \rightarrow L$ なる L' によって検索された実体は L の正例になる。ま

¹データベース中にない仮想的な実体でも良い。

た、 $L \rightarrow L'$ なる L' によって検索されない実体は L の負例になる。

一般に L' によって検索した(されなかった)実体は複数個存在する。これらのうち適切なものをサンプルに追加することによって論理式検索の手間を減らすことができる。そのための手法として以下の2つが挙げられる。(1) サンプル中の正例(負例)とのハミング距離が最小になる負例(正例)を追加する。(2) サンプル中の正例(負例)とのハミング距離が最大になる正例(負例)を追加する。実体間のハミング距離は、一方の実体にのみ出現する属性の数と双方の実体に出現するが属性値が異なるものの数の和によって定義する。予備的評価では(1)の手法が効果的であるとの結果を得ている^[4]。

論理式間の差演算

論理式 L によって検索される実体集合を $E(L)$ とする。論理式集合 $\{L_1, \dots, L_n\}$ において L_i を特定する実体の集合 $uniq(L_i)$ は以下のように定義される。

$$uniq(L_i) = E(L_i) - \bigcup_{i \neq j} E(L_j)$$

各 L_i について $uniq(L_i)$ に属する実体を一つずつ求めることにより、サンプルに追加する正例や負例の候補を列挙できる。これによって検索対象論理式を絞り込める。なお、全ての L_i について $uniq(L_i) = \phi$ となる場合がある。その際には、論理式集合内の全ての2つの論理式間の差 $dif(L_{ij}) = E(L_i) - E(L_j)$ を求めることで正例や負例を列挙できる²。

差演算を用いる手法は、利用者が実体を選択する必要があるため、検索対象論理式数が多い場合には適用が難しい。しかし、論理式等を用いて検索対象の絞り込みを行った後に適用すると効率的な論理式検索が行える。

4 論理式を通じた外延情報の間接編集

クエリーやビューを表現する論理式は、それによって検索される実体集合に対応している。この性質を利用して、論理式に対する編集操作に対応する実体集合に対する編集操作に拡張できる。これを外延情報の間接編集と呼ぶ。間接編集により実体集合を一括して編集(変更、コピー、削除、メソッド起動)できるため、データベース編集が容易になる。

実体に対する編集操作は、論理式中に出現する自由変数(単一実体、属性、リンク等に対応する)に対する編集操作の集合として定義される。曖昧さを避けるために、同一の自由変数に対する編集は高々一つしか定義

しない。論理式によって検索された各(複合)実体について、自由変数に対応する実体等に対して、定義されている編集操作を実行することで、論理式に対する編集操作を実体に対する編集操作に拡張する。

検索された実体が共通要素(単一実体またはその集合)を持たない場合には、編集操作の結果は唯一に決定される。複数の実体が共通要素を持つ場合には、共通要素に対する編集操作は重ねては施さない。この際、メソッド起動時には、同一の実体に複数のメッセージ(少なくとも一つは副作用を持つ)が送られる場合がある。間接編集においてはメッセージの到着順が定義されないため、変更結果が唯一に決定されない。これを防ぐために、編集操作の確定操作は、同一実体が複数メッセージを受信していない場合に限りで行う。なお、外延情報の部分集合に対するメッセージ送信を定義することで、上記の実体に対するメッセージの到着順を利用者が指定することもできる。

外延情報の間接編集を定義する場合、ORを含む論理式については注意が必要である。OR結合された複数の項を同時に満足する実体に対しては矛盾した編集操作が定義される場合がある。これを防ぐために、論理式を主加法標準形に変形し、各AND項毎に編集操作を定義する。異なった編集操作が定義された項の間には優先度を定義することで矛盾を防ぐことができる。

5 むすび

データベース発掘に代表される知識発見技術は、内包情報を生成する技術とも考えられる。しかし、これをデータベース化して共有するための研究は、ほとんど行われていない。現在、サンプルを用いた内包情報の検索手法を評価中である。今後の課題としては、(1) 集合以外の内包情報(リスト、木、集合とリストの組み合わせ等)に対する操作手法、(2) 実体と内包情報が混在したデータベースに対する条件式とサンプルを組み合わせた検索手法等が挙げられる。

謝辞 本研究の一部は文部省科学研究費(重点領域研究 #08244105)の援助を受けている。

参考文献

- [1] U. Fayyad, R. Uthurusamy, ed., "Data mining and knowledge discovery in databases", *Comm. ACM*, Vol. 39, No. 11, pp. 24-68, 1996.
- [2] M-S., Chen, et al., "Data mining: an overview from a database perspective", *IEEE TKDE*, Vol. 8, No. 6, pp.866-883, 1996.
- [3] T. Miura, "A logical framework for deductive object", *Information Systems*, Vol. 17, No. 5, pp.395-414, 1992.
- [4] 村田、掛下、"サンプルを用いた論理式検索機構の評価", 情報研報 DBS-113-3, 平成9年7月。

² $dif(L_{ij}) = dif(L_{ji}) = \phi$ ならば、現状のデータベース中で L_i と L_j を同一視できる。