

氏 名 宇 野 裕 之
 学位(専攻分野) 博 士 (工 学)
 学位記番号 工 博 第 1415 号
 学位授与の日付 平 成 7 年 3 月 23 日
 学位授与の要件 学 位 規 則 第 4 条 第 1 項 該 当
 研究科・専攻 工 学 研 究 科 応 用 シ ス テ ム 科 学 専 攻
 学位論文題目 Studies on the Optimization of Query Processing in Deductive Databases

(演繹データベースにおける質問処理の最適化に関する研究)

(主 査)
 論文調査委員 教 授 長 谷 川 利 治 教 授 茨 木 俊 秀 教 授 上 林 彌 彦

論 文 内 容 の 要 旨

データベースが処理するデータの量や複雑性は、時代とともに増大する一方なので、必要とする情報をできるだけすばやく抽出することを可能にするために、種々のアプローチが試みられている。その一つである演繹データベースは推論機能を付与することによって関係データベースを拡張したものであり、基本データとなる「事実」とそれらから推論可能なデータを導出するための「規則」を分離して保持している点に特徴がある。

演繹データベースにユーザが質問を発すると、この「事実」と「規則」を用いて必要な答を計算し出力することになるが、その際に、質問をいかに効率的に処理するかが重要であり、演繹データベースの実用性を決定する。与えられた質問に対して、データベースを処理してその答えを導出するには、いくつもの異なった方法があるのが普通であり、最も効率的な方法を見出すことが要求されるが、これを質問処理の最適化という。

本論文は、データベースを処理するための関係代数の各演算の平均処理時間を、広範なデータに対して評価できるように、データの分布を考慮した形で推定すると共に、より効率的な質問処理方法を定量的に明らかにすることを目的とする研究をまとめたものであり、全体で7章よりなる。その内容を以下に要約する。

第1章は序論であり、本研究に関連する研究の歴史、および本研究の背景と目的などについて述べるとともに、第2章以下の内容について概説している。

第2章では、データベースの基本モデルである関係データベースと演繹データベースの基本的概念や、その操作言語である関係代数あるいはDatalogについて説明している。とくに、これらの処理に用いられる各種の演算についてはそれらの一つ一つをとりあげ、詳しい定義と説明を、予備知識として与えている。

第3章では、まず、通常用いられている2項関係上の推移閉包を拡張して、一般化された推移閉包を定

義している。そして、データが一様分布にもとづいて生成されたという前提で、推移閉包を含む各演算の平均処理時間を近似的に導出している。さらに、その結果を用いて、具体的な演繹データベースと再帰的な質問の例を与えて、従来から提案されている方法の効率を定量的に評価している。

第4章では、データが一般の分布（とくに実世界に見られるような偏った分布）にもとづいて生成されたという前提のもとで、各演算の平均処理時間を近似的に導出している。さらに、Zipf 分布の場合について、その近似式の精度を実験的に評価している。その結果、分布の違いによって計算コストは大きく異なり、データの分布を考慮した評価が必要となることが判明した。

第5章では、演繹データベースを特徴づける重要な演算である推移閉包について、その大きさを有向ランダムグラフの立場から論じている。はじめに、有向グラフの推移閉包の平均的な大きさは、与えられた節点から他の任意の節点への到達可能確率から求められることを示し、その到達可能確率を厳密に計算するアルゴリズムを2通り示している。さらに、それら厳密な計算には多大の時間がかかることを考慮して、その理論的上界値が簡単に求められることを示し、その挙動を解析している。これらの結果に基づき、与えられた節点からの到達可能な節点の数や推移閉包の大きさの上界値についても論じている。

第6章では、もう一方の重要な演算である結合に着目し、その最適な処理順序を求める問題 OPTJOIN を定義する。その問題の計算の複雑さは NP 困難であり、最適化をはかる効率的なアルゴリズムが存在することは期待できないことを示すとともに、その中で、効率的なアルゴリズムが存在する部分クラスを発見し、それらに対するアルゴリズムを提案している。

第7章は、論文全体の成果、および今後の研究課題をまとめて結論としたものである。

論文審査の結果の要旨

データベースで処理されるデータの量や複雑性の増大に対処するため、演繹データベースでは、他のデータから推論可能なものは陽に持たず、必要なときに推論できるような機能を加えることで、処理能力の高度化をはかっている。この場合、ユーザが発した質問に対する答を効率的に導出する方法を見出すことが重要である。

本論文は、データベース処理に用いられる各演算の処理時間の評価法を提案するとともに、効率的な質問処理の方法を定量的に明らかにするための研究をまとめたもので、主な成果は以下の通りである。

1. データが一様分布にもとづいて生成されたという前提の下で、データベースの基本演算、とくに推移閉包演算の平均処理時間の近似導出式を提出し、その結果を用いて、従来から提案されている方法を定量的に評価・比較した。
2. データが一般の分布にもとづいて生成された場合についても同様の研究を行い、分布の違いによって計算コストは大きく異なることを明らかにした。
3. ランダムグラフの立場から、推移閉包の大きさの期待値の正確な評価を行うとともに、全体的な挙動を知るため、その上界値の理論的挙動についても論じた。
4. 複数の結合演算の最適な実行順序を求める問題 OPTJOIN を定義し、その計算の複雑さおよびアルゴリズムについて検討した。

以上、本論文は演繹データベースに用いられる諸演算、とくに推移閉包の処理時間の評価を可能にし、質問処理の効率化への道を拓いたものであり、学術上、実際上寄与するところが少なくない。よって、本論文は博士（工学）の学位論文として価値あるものと認める。また、平成7年1月12日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。