

関係代数操作による自然言語表現の変換

九州大学工学部 上林弥彦 (Yahiko Kambayashi)

九州大学工学部 天野浩文 (Hirofumi Amano)

1. まえがき

データベースの利用範囲の拡大に伴い、一般利用者の数も近年急速に増大しつつあるが、データベースを利用する場合には、通常、自分がどのような情報を要求しているかを記述した質問をデータベース管理システムに与える必要があり、利用者は質問の記述に用いられる言語に習熟していなければならない。本研究は、質問用の特殊な言語に不慣れな一般利用者の質問作成を支援するために、自然言語を利用する方法について考察したものである。

本稿では、単純な概念と強力かつ柔軟な検索・更新能力をもち、近年は商用システムの分野でも脚光を浴びている関係データベースを取り上げる。関係データベースは、利用者からは平坦な表の集まりに見えるという特色をもっている^[1]。

この関係データベースにおいて、質問の記述に自然言語を利用できれば、利用者が特殊な言語を習得する必要がなく、わかりやすいので、利用者から見た使い勝手の点では好都合である。しかし、自然言語(による質問)の処理が本質的に困難な問題を含んでいるため、データベースにおける自然言語インタフェースは、まだ必ずしも成功したとはいえない。たとえば、システムが複雑かつ膨大になることや、利用者の質問にあいまいさがある場合にフィードバックによってシステムの理解している質問を修正していく方法を取らなければならない^[2]ために生じるオーバーヘッドなどの問題点がある。

そこで、関係データベースにおける一般利用者のための使いやすいインタフェー

スとして、利用者からの自然言語による質問を処理するかわりに、利用者の質問言語による質問から、その質問の意味を表す文章を生成し、質問の内容が自分の意図どおりかどうか利用者が確かめることができるようなシステムを考え、このために必要な、利用者の与えた関係代数質問を自然言語表現に変換する方法を考察した。本稿で述べる方式は、あらかじめ、属性間の意味的なつながりを表す英語表現の集合を人間が用意しておき、利用者の質問に応じ、それらに簡単な変換を施して質問の意味に対応するような英文を出力する方法を用いる。ただし、簡単な削除・置換・連結等の操作で、質問の意味に対応する英文を生成するためには、もともになる表現の集合がある程度の制約を満たしていなければならない。

本稿では、データベーススキーマに対応する英文集合の作成規則、および、関係代数における選択・射影・結合質問に対応する、それらの英文の変換規則について述べる。

2. 基本的事項

2.1 関係モデル

データベースにおける関係モデルでは、データはすべて関係の形で記憶される。

関係は、直観的には、一枚の表に対応させて考えることができる。

表に見立てた関係の縦の列をそれぞれ区別するために、それらに属性あるいは属性名と呼ばれる名前をつける。関係の要素は組と呼ばれ、同じ表の横の列に対応する。

一つの関係 R を構成する属性の集合を関係スキーマと呼び、 R で表現する。

関係スキーマの集まりを (関係) データベーススキーマと呼び、対応する関係の現在の値を (関係) データベースと呼ぶ。

2.2 関係代数

質問言語の能力の基準として考案された関係代数は、特定の演算を関係に適用す

ることで質問を表現する。演算には次のようなものがある。

(1) 射影

関係を構成する属性のうちから、必要な属性だけを取り出して、新しい関係を作る操作を射影という。関係 R の要素である一つの組を t とする。属性集合 $X \subseteq R$ に対し、 t の X の部分のみを集めたものを t の X に対する射影と呼び、 $t[X]$ で表す。 R の X への射影は次の式で定義される。

$$R[X] = \{t[X] \mid t \in R\}$$

(2) 選択

関係の組のうちで、その組のある成分(成分の組合せでもよい)が与えられた条件を満たしているものだけを取り出す操作を選択という。

θ を比較演算子 $=$ 、 \neq 、 $<$ 、 \leq 、 $>$ 、 \geq のうちいずれかであるとする。関係 R の二つの属性 A と B の間の θ 選択を次のように定義する。

$$R[A \theta B] = \{r \mid r \in R \wedge (r[A] \theta r[B])\}$$

ただし、 A と B とは同じ定義域をもつ属性でなければならない。

また、右辺を $r[B]$ のかわりに定数 c として考えることもできる。

$$R[A \theta c] = \{r \mid r \in R \wedge (r[A] \theta c)\}$$

(3) 結合

二つの関係を、共通の定義域を手がかりとして結合し、新しい関係を作る操作を結合という。

θ を(2)と同様の比較演算子とする。関係 R の属性 A と関係 S の属性 B の間の θ 結合を次のように定義する。

$$R[A \theta B]S = \{(r, s) \mid r \in R \wedge s \in S \wedge (r[A] \theta s[B])\}$$

ただし、 A と B とは同じ定義域をもつ属性でなければならない。

その他に、除算や和集合、共通集合などの演算がある。除算は他の演算で表現できることが知られているので、代表的な関係代数演算は選択・射影・結合の三種類である。この三つの演算のみで表現できる質問を SPJ (selection/projection/

join) 質問という。この SPJ 質問は関係代数質問の重要なクラスとなっている。ただし、本稿ではこの SPJ 質問を、自己結合 (self join) を含まないものに制限して考察する。

2.3 従属性と正規形

一般にデータ値として自由な値の組が許されているわけではなく、実世界のデータには、何らかの制約が課せられている場合が多い。このような意味的制約の中で、データベースにおいて特に重要なのは、従属性^[1]と呼ばれるものである。

一般に、属性集合 X の値の組をひとつ決めると、属性集合 Y のすべての値が一意に決まる場合に、 X から Y への関数従属性 (functional dependency, 以下 FD と略す) が成立する、といい $X \rightarrow Y$ で表す。

FD: $X \rightarrow Y$ が成立していれば、定義より、 Y の任意の要素 A に対し、FD: $X \rightarrow A$ が成立する。また、任意の属性集合 Z に対し FD: $XZ \rightarrow A$ が成立する。したがって、FD: $X \rightarrow A$ において、左辺の X をより小さな集合で置き換えることのできる可能性がある。FD: $X \rightarrow A$ が、どのような X の真部分集合 X' に対しても $X' \rightarrow A$ という FD がないという性質を満足するとき、これを全関数従属性 (full functional dependency, FFD) と呼び、 $X \Rightarrow A$ で表す。 Y のすべての要素 A に対し $X \Rightarrow A$ であるとき、FFD: $X \Rightarrow Y$ が成立するという。

関係 R において、 $K \Rightarrow R$ となるような極小の $K (\subseteq R)$ を R のキーと呼ぶ。

データベース内の関係は、現実世界の意味制約を反映した従属性集合を満足していると考えられる。この従属性集合の保持を容易にするために、関係の正規形を定義する。正規形はいくつかの種類が考えられているが^[1]、ここでは、本稿で述べる方式において重要な第三正規形のみについて触れる。

非キー属性が " \rightarrow " の右辺に現れるすべての自明でない FD の左辺が、キーもしくはキーを含む属性集合になっている関係を、第三正規形という。

3. 関係スキーマに対する標準表現群

本節では、データベース管理者が用意すべき英語表現の作成規則について述べる。ただし、本稿では、各々の関係は第三正規形に分解されているものとする。

各々の関係に対して、それに含まれる属性を意味的なまとまりに分け、それらのつながりを表す英語表現を作り、これを標準表現と呼ぶことにする。この標準表現は以下のような規則に基づいて作成する。

- (1) 各属性名のうち、ある単語(あるいは語句)の省略形を用いてあるものや、何か他の語句を用いた方がわかりやすいものは、属性名のほかにそれらの単語や語句を加えて一つの要素とする。
- (2) 非キー属性の一つ一つについて、それとキーとを一つの属性グループとする。
- (3) 標準表現はひとつの属性グループについて一つの文であるものとし、カンマを用いずに表現する。
- (4) キーが単一の属性である場合は、そのキー属性を主語にする。キーが複数の属性からなる場合は、その中から一つ主語を選ぶ。この主語はその関係について属性グループが複数個あっても共通とする。
- (5) 各属性グループの主語でない属性が単一の述語動詞の目的語、補語、あるいは、修飾語句の一部となるように適当に動詞を選び、標準表現とする。
- (6) 関係詞節の使用はなるべく避ける。

(1)の規則は、属性名が何かの語句の省略形である場合にそれが何を指すのかをはっきりさせるためのものである。(2)は、第三正規形の関係のキー属性から個々の非キー属性へのFDの一つ一つに、意味表現を対応させることに相当する。

(3)～(6)は変換を容易にするための規則である。

以上のような規則に基づいてどのような表現を作成すべきか、次のようなデータベーススキーマを例にとりながら述べる。属性名に下線が引かれているのは、それがキー属性であることを示している。

(a) STUDENT (SNAME, SECT, ADDRESS, AGE)

(b) PROFESSOR (PNAME, SECT, TEL, ADDRESS, SAL, AGE)

(c) SECRETARY (SECNAME, DIV, ADDRESS, SAL)

(d) DEPARTMENT (SECT, DEPT)

(e) EXAM (SNAME, TITLE, MARKS)

(f) LECTURE (SNAME, SUBJECT, PNAME)

これに対する標準表現群は、たとえば、次のようになる。

(a') Student {SNAME} : belongs to the section {SECT}

: lives in {ADDRESS}

: is {AGE} years old

(b') Prof. {PNAME} : belongs to the section {SECT}

: has a telephone {TEL}

: lives in {ADDRESS}

: earns a salary of {SAL}

: is {AGE} years old

(c') Secretary {SECNAME}: works in the division {DIV}

: lives in {ADDRESS}

: earns a salary of {SAL}

(d') The section {SECT} : is in the department {DEPT}

(e') Student {SNAME} : got {MARKS} marks for an exam {TITLE}

(f') Student {SNAME} : takes the lecture {SUBJECT}

given by Prof. {PNAME}

これらの英語表現は、次節で述べる変換規則や出力文の生成規則による変換・連結の操作を受ける土台となるものである。操作によっては、必ずしも適合しないものもあり、データベース管理者があらかじめ数種用意しておかなければならないこともある。

4. 標準表現群の変換

本節では、質問に応じてこれらの表現に変換を加え、質問に対応する英文を生成する際の規則について述べる。ここでは、質問言語として関係代数を考え、SPJ質問(ただし、自己結合を含まないものに制限する)に含まれる選択・射影・結合の操作について、それらに応じて標準表現をどのように変換すべきかについて考察する。

4.1 選択

選択の条件を表す式を $r[A] \theta r[B]$ (A, B は属性名、 θ は比較演算子)とする。選択に対応する変換は、語句の挿入と置換が基本である。変換規則は次のようになる。

(i) 右辺が定数でない場合、 A に関する記述の代わりに、 θ の種類に応じて以下の語句を挿入する。ただし、 $\langle A \rangle$ は A に関する記述の部分を示す。

A, B が数値データのときは、

= $\langle A \rangle$ equal to $\langle B \rangle$

> $\langle A \rangle$ greater than $\langle B \rangle$

< $\langle A \rangle$ less than $\langle B \rangle$

\geq $\langle A \rangle$ not less than $\langle B \rangle$

\leq $\langle A \rangle$ not greater $\langle B \rangle$

\neq $\langle A \rangle$ greater or less than $\langle B \rangle$

非数値データのときは、

= the same $\langle A \rangle$ as $\langle B \rangle$

\neq $\langle A \rangle$ different from $\langle B \rangle$

(ii) 右辺が定数の場合、 θ に応じて以下の変換を行う。

= 標準表現の中の A に関する記述をその定数で置換する。

それ以外 (i)のときの変換を行いき $\langle B \rangle$ の部分とその定数で置換する。

属性については、元の関係の組を識別する重要な要素であるので削除しない。文法的にみて削除できるかどうかは、標準表現作成の際にデータベース管理者が指定する。

(iii)は、ある関係から一つのキー属性しか射影されない場合、それがその関係から射影されたものであることしかいえないであろうという理由による。

例4.3 LECTURE[SNAME,SUBJECT]

射影されないPNAMEは、非キー属性であるので、上記の規則にもとづき、次のように変換する。

Student {SNAME} : takes the lecture of {SUBJECT}

例4.4 STUDENT[SNAME,SECT]

属性 ADDRESS、AGEが他の操作に無関係ならば、SNAMEとSECTに関する部分のみを残す。(もし、他の操作でADDRESS、AGEを用いるならば、射影に対応する変換規則は適用しない)

Student {SNAME} : belongs to the section {SECT}

例4.5 EXAM[SNAME,MARKS]

射影されない属性TITLEはキー属性であるので、

Student {SNAME} : got {MARKS} marks for some exam.

4.3 結合

4.1節、4.2節で述べた変換は、単一の関係内で考えることのできる操作に対応するものであった。これに対し、結合操作は複数の関係にまたがる操作であるから、結合によって得られる新しい関係に対応する意味表現を求めるためには、単一の関係に対応する表現どうしを何らかの形で連結すればよい。この連結の方法には、次の2つを考えることができる。

- (a) 結合される2つの関係に対応する標準表現群はまとめて配置し、それらの間に結合の条件を表す文を置く。
- (b) 一方の関係の標準表現の中の、結合に関与する属性の後ろに、他方の関係の

中の結合に関与する属性が持っている意味を表す表現を挿入する。

本稿では、(a)を疎連結法、(b)を密連結法と呼ぶことにする。これらの方法で生成される表現の模式図を図4.1, 4.2に示す。

疎連結法では、それぞれの関係に対応してまとめられた標準表現群とは独立した文が設けられるので、標準表現の構文にほとんど依存しないが、結合の経路が鎖状になっていない場合に、結合されるすべての関係の対に対応する標準表現どうしが隣接する位置に配置できないので、必ずしもわかりやすい表現にならない場合がある。

これに対して、密連結法では、鎖状質問だけでなく、木構造の質問にも対処できる。しかし、標準表現の最後部にない属性に関係代名詞節などの形で修飾語句を付けると問題が生じる可能性があるので、括弧でくくるなどの処置が必要である。また、この方法では修飾語句(非常に長くなる場合がある)を挿入された文が長くなるのは避けられない。

本稿では、上記の2つの方法のうち、もとの質問との対応を考えやすく変換操作も簡単な疎連結法を用いることにする。したがって、以下では、結合される各関係

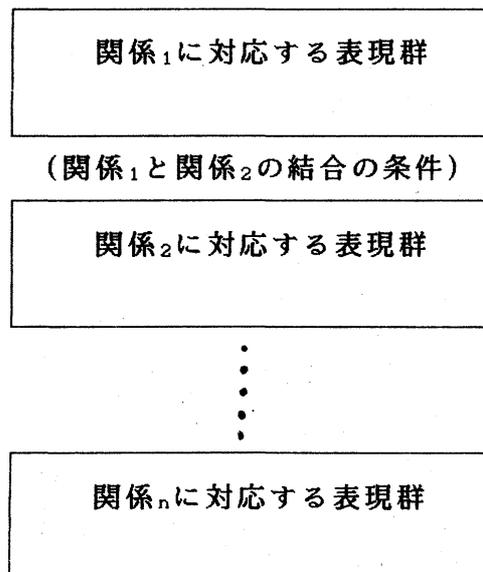
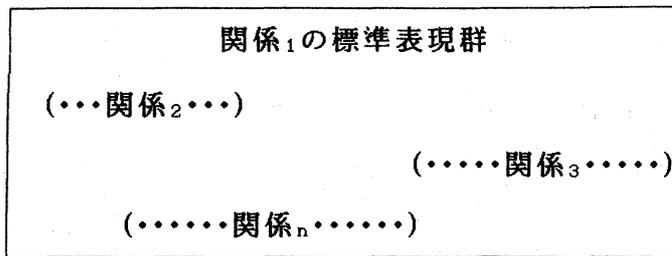


図4.1 疎連結法



()内には、結合される各関係の標準表現が埋め込まれている

図4.2 密連結法

に対応する標準表現群どうしの中間に置かれる文の生成法について述べる。

関係RとSとの結合の条件を表す式を $r[A]\theta s[B]$ (A, B :属性名、 θ :比較演算子)とする。以下に示すように、結合に対応する変換の基本は、文の接続である。 $\{A\}$ は属性名、 $\langle B \rangle$ はBに関する記述を示す。

(i) 次の語句を用意する。

$\{A\}$ is

(ii) AとBの属性名が異なるときは、 θ に応じて上の語句のあとに次の語句を加える。

A、Bが数値データのときは、

= equal to $\langle B \rangle$ that

> greater than $\langle B \rangle$ that

< less than $\langle B \rangle$ that

\geq not less than $\langle B \rangle$ that

\leq not greater than $\langle B \rangle$ that

\neq greater or less than $\langle B \rangle$ that

A、Bが非数値データのときは、

= the same one as $\langle B \rangle$ that

\neq different from $\langle B \rangle$ that

(iii) AとBの属性名が等しいときには、 θ に応じて (i) の語句のあとに次の語

句を加える。

A、Bが数値データのときは、

= equal to one that

> greater than

< less than

≥ not less than

≤ not greater than

≠ greater or less than

非数値データのときは、

= the same one that

≠ different from one that

- (iv) Bが主語でない場合、上記のような語句のあとに、Bを含む標準表現から(文法的にみて可能ならば)Bを取り除いたものを加える。
- (v) Bが主語のときは、Bの標準表現の中に、他の射影・選択・結合に関与する属性があれば、そのうちの一つの標準表現から(文法的にみて可能なら)Bを除いたものを上記(ii)、(iii)のような語句のあとに加える。なければ、次の語句を加える。

is in << relation (関係名) >>

結合される属性を含む標準表現を接続する場合には関係代名詞を用いる。(本稿では、標準表現から適当な語句を削除することで関係代名詞節を得られると仮定しているが、標準表現の構文によっては、このような簡単な方法では、求める表現が得られないことがある。この場合には、データベース管理者があらかじめ検査しておき、不都合があれば、代わりの表現を準備しておかなければならない。)

AとBの属性名が異なる場合、どちらの属性名も残すことが望ましいので、(ii)では、Bを含む記述を添えてある。等しい場合には、重複を取り除いて(iii)のような表現を用いる。(iv)、(v)の操作により、関係代名詞節が完成する。(v)で" is in ..."の表現を用いるのは、ある関係RとSを結合する際に、Rの組のAと同じ

値を B の値として持つ組が S の中にあるかどうかだけを問題とすることがありうるためである。

例4.6 STUDENT[SECT=SECT]PROFESSOR

これは、所属講座が等しいという条件のもとで、STUDENTとPROFESSORを結合するものである。2つの標準表現群の間に置かれる文は次のようになる。

{SECT} is the same one that Prof. {PNAME} belongs to

4.4 選択・射影・結合質問に対する出力文の合成

4.4節では、これまでに述べた変換規則にもとづいて変換された標準表現を連結し、出力とするための規則について述べる。

- (i) 文頭は大文字、標準表現の主語の頭の大文字は出力文中では原則として小文字となるようにする。
- (ii) 各関係の標準表現はひとまとまりにして、カンマや " and " で連結するものとする。結合に関与する二つの関係は隣接して置かれるものとし、結合の条件となった属性を含む記述はその間に置く。そのあとに、関係代名詞を含む記述を添える。関係代名詞を含む文のあとはピリオドとし、さらに文が続く場合は、その先頭に、"Moreover," を置く。
- (iii) 出力属性は、[] でくくるものとする。

以上の規則による出力文を模式的に表すと、たとえば次のようになる。()内は結合操作が含まれる場合である。

主語₁ [] , ..., [] .

(Here, { A₁ } is the same one that ...,

.....,

and { A_n } is greater than

Moreover, 主語₂ [] , ..., [] .)

本節までに述べた規則にしたがって生成される英文の例を以下に示す。

例4.7 ((STUDENT[AGE=22])[SECT=SECT]PROFESSOR)[SNAME,SECT,PNAME]

これは、22才の学生と同じ講座に所属するような教授を選んで結合し、(SNAME, SECT,PNAME)を答として求める質問である。

Student [SNAME] is 22 years old, and belongs to the section [SECT].

Here, {SECT} is the same one that Prof.[PNAME] belongs to.

例4.8 (((SECRETARY [DIV='WELFARE']) [SAL=SAL] PROFESSOR) [SECT=SECT] STUDENT)
[SNAME]

厚生課に勤務する秘書と同じ給与を得ているような教授と同じ講座に所属する学生の名前を求める質問である。

Secretary {SECNAME} works in the WELFARE division, and earns a salary of {SAL}.

Here, {SAL} is equal to one that Prof.{PNAME} earns.

Moreover, Prof. {PNAME} belongs to the section {SECT}.

Here, {SECT} is the same one that student [SNAME] belongs to.

5. 変換後の表現の検査と修正

これまでに述べた方法では標準表現の構文を限定しており、また、変換も機械的で簡単なものに限っているため、利用者にとって理解しやすい自然な英文が得られない場合がある。しかし、前述の作成規則にもとづく標準表現群は、共通の主語をもつ表現をカンマや "and" で連結するだけで文法上の不都合を生じることはない。さらに、標準表現に加えられる可能性のある変換は

- (1) ある属性の削除あるいは置換
- (2) ある属性に対する修飾語句 "greater than ……" などの後置
- (3) ある属性の削除による関係代名詞節の生成

の3つだけである。したがって、もとの標準表現が妥当でわかりやすい表現であったのに変換によって不自然な表現が生成されたとすると、それは上記の3点のいずれかに問題があったためである。そこで、データベース管理者は標準表現群を作成

する際(あるいはこの利用者インタフェースの運用が開始される前)に、上記の三点についてのみ変換後の個々の表現を検査すればよい。不適切な表現が生成される場合には、その変換後の表現に代わる表現(代替表現)を準備する。

ただし、この代替表現は、他の(変換された)標準表現と同格にあつかえることが望ましいので、次のような規則にもとづいて作成する。

- (1') ある属性が削除あるいは置換された表現が不適切である場合、作成規則の(1)、(3)、(5)、(6)にしたがう。主語はもとの標準表現のそれと異なってもよい。
- (2') ある属性の後ろに修飾語句を置くことで不都合が生じる場合、その属性に関する記述の部分のみに、代わりの表現を準備する。
- (3') ある属性を削除して作成した関係代名詞節に問題がある場合、"{属性名} is"の後に置くことができ、しかも、その比較演算子の意味に対応する表現を準備する。

例5.1 関係EXAMの属性MARKSが質問に無関係な場合の表現は、

Student {SNAME} : got some marks for an exam {TITLE}

である。この表現は必ずしも自然でないので、上の(1')にもとづき、次のような表現を作成するとよい。

Student {SNAME} : took an exam {TITLE}

例5.2 関係STUDENTの属性AGEに対する修飾語の後置を行うと比較演算"AGE>**"に対しては

Student {SNAME} : is {AGE} greater than ** years old

となる(下線部は挿入された語句)。これは文法的に不適当なのでデータベース管理者はたとえば次のような表現をこれに代わるものとして作成する。

Student {SNAME} : is over ** years old ({AGE})

うしろの({AGE})は、この表現に対応する属性名を明らかにするためのものである。比較演算子が≠、<、≤、≥のときはそれぞれ

: is over or under ** years old ({AGE})

: is under ** years old ({AGE})

: is ** years old or under ({AGE})

: is ** years old or over ({AGE})

6. むすび

本稿で述べた方法により、自然言語処理にともなう困難な問題を避けたうえで、関係代数における鎖状のSPJ質問の意味に対応する表現を生成できる。データベース管理者が事前に検査しなければならない項目はかなり多いが、これも自然言語インタフェースにおける辞書作成/管理その他の作業に比べれば、はるかに容易であろう。

また、関係データベースにおける質問言語の中で最も著名なもののひとつであるSQL言語^[11]の記述の単位は、関係代数のSPJ質問とほぼ対応するので、本方式をSQL質問に拡張することも容易である。

ただし、本稿で述べたような、結合の経路に沿って単純な記述を順序よく重ねていく方法では、

(a) 結合の経路が鎖状でないような木質問

(b) 結合の経路に閉路がある巡回質問

に対して、わかりやすい表現が生成できない場合があることも明らかになった。

(a)の場合については、4.3節で述べた密連結法を併用すれば、もとの質問との対応のわかりやすさをあまり損なわずに、ある程度対処できるであろう。

しかし、巡回質問に対しては、結合の経路をたどりながら表現を連ねていくかぎり、閉路の出発点にもどったとき、自然言語表現の上では、出発点からもっとも離れてしまっており(自然言語表現では出発点は文頭にあり、もどってきたときは文末になっている)、かなり人間的な(あるいは技巧的な)表現を用いないかぎり、わかりやすい表現が得られない場合が起こりうる。

このことは、巡回質問に関し、自然言語が(記述不能ではないにせよ)十分な記述

能力をもたないことを示唆していると考えられる。

謝辞

本研究に関して、御検討いただいた高嶋裕一氏(現在YHP)、ならびに、熱心に御討論・御助言いただく最所圭三助手、古川哲也氏、Mohamed Elsharkawi氏をはじめとする研究室のみなさまに深謝いたします。

参考文献

- [1] Ullman, J. D., (國井・大保訳) "データベース・システムの原理"、日本コンピュータ協会。
- [2] Codd, E. F., "How About Recently? (English dialog with relational databases using RENDEZVOUS Version 1)", DATABASES: IMPROVING USABILITY AND RESPONSIVENESS (Shneiderman, B. eds.), Academic Press, 1978.