# 構造化文書とデータベースの統合利用の研究 — ランキングを含む問合せの記述とその処理方式 —

5F－6

品川 徳秀　　森嶋 厚行　　北川 博之

筑波大学 工学研究科　　筑波大学 工学研究科　　筑波大学 電子・情報工学系

## 1 はじめに

近年、コンピュータ技術の発展やネットワークの普及によって様々な情報資源の利用が容易になり、それらの異種情報資源の統合利用が重要な課題となっている。各種の情報資源の中でも構造化文書はデータベースと共に重要な情報資源となってきている。我々の研究グループでは構造化文書とデータベースの統合環境を提供すべく、NR/SD+ の開発を行なっている [1], [2], [3]。

**NR/SD+** では、入れ子型リレーショナルモデルに抽象データ型として**構造化文書型 (SD 型)** を導入している。また、入れ子型リレーションに対する入れ子型リレーショナル代数とテキストに対するリージョン代数という 2 つのデータ操作体系に、それらの間を相互変換するコンバータと呼ばれる演算子を核とする演算子群を提供する事により、構造化文書と入れ子型リレーションの統合利用が可能な環境を提供している。

文書の検索に関しては、一般にユーザの検索要求を問合せ式として明確に記述する事が難しい場面が多くある事が知られている。このため、情報検索においては、対象データの検索要求に対する適合度を導入した問合せ処理がしばしば行なわれている。構造化文書の検索においても同様の問題が発生する。そこで、本稿では NR/SD+ に適合度の概念を導入した問合せ記述とその処理機構の導入を検討する。まず、NR/SD+ について簡単に述べる。次に、その拡張として適合度を持つリレーションと、適合度を与える演算子、適合度によりタプルを絞込むための演算子について述べる。また、適合度がどのように結果へ伝播されるかを述べ、例を示す。最後に、まとめと今後の課題を示す。

## 2 NR/SD+

NR/SD+ で導入された SD 型の値は **DTD** と**タグ付きテキスト**の組である。テキスト中の同名の**タグ**で区切られた部分を**要素**と呼び、各要素の構造の定義は DTD の要素型定義で行なう。

入れ子型リレーション構造と文書構造とを相互変換する**コンバータ**には **Unpack** と **Pack** がある。Unpack は SD 値中の要素群を値とする副リレーション構造を作成し、Pack は副リレーション構造中の要素を持つ SD 値を作成する。Unpack(**U**) と Pack(**P**) の適用例を次に示す。$s_2$ の属性 C の各副リレーションには、要素型 b を持つ要素のうち、任意の要素の第一副要素として現れるものを持つ SD 値が含まれる。$s_2$ に (2) 式を適用すると、各副リレーションを $s_2$ の属性 B の SD 値をテンプレートとして構造化文書に変換し、$s_1$ を得る。

$$s_2 = \mathbf{U}_{B \to C(O, D[b \cap 1] \text{ as } x)}(s_1) \quad (1)$$

$$s_1 = \mathbf{P}_{C(O, D \text{ as } x) \to B}(s_2) \quad (2)$$

Integration of Structured Documents and Databases
-- Specification and Processing of Queries Incorporationg Ranking --
Norihide Shinagawa †, Atsuyuki Morishima †, Hiroyuki Kitagawa ‡
† Doctoral Degree Program in Eng., Univ. of Tsukuba,
‡ Institute of info. Sci. and Elec., Univerety of Tsukuba

| A | B |
|---|---|
| 1 | ⟨a:seq(b,c:rep(b))," ⟨a⟩⟨b⟩d1⟨/b⟩ ⟨c⟩⟨b⟩d2⟨/b⟩⟨b⟩d3 ⟨/b⟩⟨/c⟩⟨/a⟩" ⟩ |

| A | B | C | |
|---|---|---|---|
| | | O | D |
| 1 | ⟨ a:seq(b,c:rep(b))," ⟨a⟩&x.1; ⟨c⟩&x.2;⟨b⟩d3⟨/b⟩⟨/c⟩⟨/a⟩" ⟩ | 1 | ⟨ b," ⟨b⟩d1⟨/b⟩" ⟩ |
| | | 2 | ⟨ b," ⟨b⟩d2⟨/b⟩" ⟩ |

コンバータ **U/P** の適用例 $s_1$ (上) と $s_2$ (下)

## 3 NR/SD+ の拡張

### 3.1 曖昧な条件に対するタプルの適合度の導入

NR/SD+ で曖昧な検索条件を扱うために、ある問合せ条件に対するタプルの**適合度**を表現する枠組を用意する。具体的にはリレーションの特別な属性として $LB, UB$ を導入する。$UB$ と $LB$ はそれぞれタプルの適合度の上限と下限を表し、0 以上 1 以下の値である。適合度が未評価のタプルに関しては、未定義を表す値である [1,0] であるものとする。このような、区間を真偽値として持つデータモデルとして [4], [5] などがあり、リレーションに対する基本演算子群もこれを演算結果に伝播させるよう拡張されている。ここでは、基本的に [5] に基いた拡張を考える。尚、簡単化のため、タプルの最上位の SD 型の属性 のみを適合度評価の対象とし、タプルの適合度とするものとする。また、副リレーション中のタプルに関する適合度は考えない。

| | Paper | Names O | Names Name | LB | UB |
|---|---|---|---|---|---|
| $r_1$ | $paper_1$ | $o_1$ | 山田一郎 | 0.8 | 0.9 |
| | | $o_2$ | 田中花子 | | |
| | $paper_2$ | $o_1$ | 中村二郎 | 0.1 | 0.4 |
| | | | | | |

### 3.2 演算子の追加

**$\gamma$ 演算子**　問合せに対する対象文書の適合度は、検索要求が与えられた時に初めて決定可能である。そこで、$\gamma$ 演算子を導入する事で、SD 値に対する問合せ条件に基いてその適合度を評価し、当該 SD 値を含むタプルの $LB, UB$ の値として格納する。

$$\gamma_{A_i:\{parameters\}}(r) := \{(a_1, ..., a_n, l, u) \mid t \in r \wedge t = (a_1, ..., a_n) \wedge [l, u] = g_{\{parameters\}}(a_i)\}$$

ここで $parameters$ には、例えば対象文書に対する問合せ条件を記述するためのキーワードなどが与えられ、それをパラメータとして用いる得点付け関数 $g$ によって文書の適合度が測られる。$g$ による適合度は母集団となる情報資源の特性とは無関係に決定され、適合度の上下限であるものとする。$g$ が単価関数である場合は、$l = u = g_{\{parameters\}}(a_i)$ とする。

**$\rho$ 演算子**　問合せに応じた演算結果の適合度を用いて絞込みを行なうための演算子である。これは $\sigma$ 演算子とは異なり、適合度属性の値を明示的に参照可能である。

$$\rho_{condition}(r) := \{t \in r \mid condition(t) \text{ が成立}\}$$

条件式に次の $t \in r$ に対する順位付け関数を与える事により、上位 n 位までのタプルに絞込むなどもできる。

$$ord_{r:expr}(t) := 1 + |\{t' \in r \mid expr(t') > expr(t)\}|$$

例えば $s' = \rho_{ord(UB)\leqq 3}(s)$ は、$s$ に含まれるタプルを $UB$ の大きい順に 1, 2, ... と順位付けを行ない、1 位から 3 位までのタプルのみを結果 $s'$ とする。

**$\kappa$ 演算子** 適合度をタプルに付加する事により、データ部が同値であるにも関らず適合度の異なるタプルが出現する事があるが、それらは異なるものとして扱われる。そのようなタプルを 1 タプルに統合するための $\kappa$ (companction) 演算子 が [5] で与えられている。

### 3.3 適合度の伝播

$\gamma$ 演算子によって付加された適合度は、各演算子においてその結果へと伝播される。その伝播方式は [5] に従う。演算子を大きく分けると、元の適合度をそのまま伝播すればよいもの ( $\sigma, \pi, \mathbf{U}, ...$ ) と、元の適合度から新たな適合度を算出する必要のあるもの ( $\cap, \cup, \bowtie, \times, \kappa, \mathbf{P}, ...$ ) とに分かれる。新たな適合度の算出方法には元の適合度間の関係に依存して次の 3 通りがあり、演算子と共にユーザによって指定される。(1) 2 つのタプルの適合度間に負の相関を持つ時 "nc" (inconsistent)、(2) 正の相関を持つ時に "pc" (positive corelation)、(3) 相関を持ち合わせていない時 "ig" (ignorance) 。正の相関を持つというのは、例えばキーワード「インターネット」への適合度の高いものは同「コンピュータ」への適合度も高い、と考えられるような事を指す。以下に、各場合における新たな適合度の算出式を示す。但し、$\cup$ には $\oplus$ が、それ以外では$\otimes$ の定義が利用される。

$[\alpha_1, \beta_1] \otimes_{nc} [\alpha_2, \beta_2] := [0, 0]$
$[\alpha_1, \beta_1] \otimes_{ig} [\alpha_2, \beta_2] := [max(0, \alpha_1 + \alpha_2 - 1), min(\beta_1, \beta_2)]$
$[\alpha_1, \beta_1] \otimes_{pc} [\alpha_2, \beta_2] := [min(\alpha_1, \alpha_2), min(\beta_1, \beta_2)]$
$[\alpha_1, \beta_1] \oplus_{nc} [\alpha_2, \beta_2] := [min(1, \alpha_1 + \alpha_2), min(1, \beta_1 + \beta_2)]$
$[\alpha_1, \beta_1] \oplus_{ig} [\alpha_2, \beta_2] := [max(\alpha_1, \alpha_2), min(1, \beta_1 + \beta_2)]$
$[\alpha_1, \beta_1] \oplus_{pc} [\alpha_2, \beta_2] := [max(\alpha_1, \alpha_2), max(\beta_1, \beta_2)]$

尚、一方の適合度が未定義を表す $[1, 0]$ の時は、もう一方の値をそのまま返すものとする。

## 4 問合せ例

次のような問合せに対するデータ操作を考える。

「情報資源として、論文の構造化文書リポジトリと、大学教員に関するデータベースがある。この時、その執筆論文の中に SGML に関する論文と SQL 問合せ最適化に関する論文がある T 大学の教官のうち、適合度上限が上位 3 位までの人の氏名、肩書、担当科目、専門の一覧を求めよ」

Faculty

| FID | Name | Dep | Title |
|---|---|---|---|
| | | | |

Papers

| Paper |
|---|
| |

SD 型属性 Paper の値の DTD の例

```
paper    = seq(title, authors, abstract, p-body, reference)
authors  = rep(author)
author   = seq(name, affiliation)
p-doby   = rep(section) + fig
section  = seq(sectitle, rep(para), opt(rep(section)))
...
```

この問合せは次のように行なえる。まず、SGML に関する論文の適合度を付与した一覧を作成し、その著者名を Unpack によって取り出す。結果は前出の $r_1$ を経由して $r_2$ になるとする。同様に、SQL 問合せ最適化に関する論文については $r_4$ となったとする。

$r_1 = \mathbf{U}_{Paper \to Names(O,Name)[name]}(\gamma_{Paper:\{SGML\}}(Papers))$
$r_2 = \kappa_{pc}(\pi_{Name}(\mu_{Names}(r_1)))$
$r_3 = \mathbf{U}_{Paper \to Names(O,Name)[name]}(\gamma_{Paper:\{SQL,\ 問合せ最適化\}}(Papers))$
$r_4 = \kappa_{pc}(\pi_{Name}(\mu_{Names}(r_3)))$

$r_2$

| Name | LB | UB |
|---|---|---|
| 山田一郎 | 0.8 | 0.9 |
| 田中花子 | 0.8 | 0.9 |
| 中村二郎 | 0.1 | 0.4 |

$r_4$

| Name | LB | UB |
|---|---|---|
| 山田一郎 | 0.8 | 0.9 |
| 田中花子 | 0.4 | 0.5 |
| 高橋健太 | 0.4 | 0.6 |

次に、これらの共通部分を取る。ここでは $r_2$ と $r_4$ の相関が特に認められないであろう事から $\otimes_{ig}$ の演算を適用する。

$r_5 = r_2 \cap_{ig} r_4$

$r_5$

| Name | LB | UB |
|---|---|---|
| 山田一郎 | 0.6 | 0.9 |
| 田中花子 | 0.2 | 0.5 |

最終的に、$UB$ の大きな順に 3 位までを結果とする。

$r_6 = \rho_{ord(UB)\leqq 3}(\pi_{Name,Title,Cource,Speciality}(\sigma_{Dep='T大学'}(Faculty \bowtie_{ig} (r_5))))$

$r_6$

| Name | Title | Course | Speciality | LB | UB |
|---|---|---|---|---|---|
| 山田一郎 | 助教授 | DB | データ工学 | 0.6 | 0.9 |
| 石山純一 | 教授 | OS | OS | 0.7 | 0.8 |
| 木村雅人 | 助教授 | 情報処理概論 | データ工学 | 0.6 | 0.8 |

## 5 まとめと今後の課題

本稿では、不確定性を扱うリレーショナルデータモデルに関する研究をベースに、適合度評価演算子 $\gamma$ と絞込み演算子 $\rho$ を導入する事によって、NR/SD+ において問合せ条件への適合度を考慮した演算を行なうための拡張を検討した。今後の課題として、ここで行なった様々な仮定の解除、ネスト構造を考慮した適合度の導入などが残されている。また、[6] に見られるような問合せ処理の最適化などについても考察を進めている。

## 謝辞

本研究の一部は文部省科学研究費補助金重点研究「高度データベース」の助成による。

## 参考文献

[1] A. Morishima and H. Kitagawa, *A Data Modeling Approach to the Seamless Information Exchange among Structured Documents and Database*, Proc. ACM SAC'97, pp.78-87, Feb. 1997.

[2] A. Morishima and H. Kitagawa, *A Data Modeling and Query Processing Scheme for Integration of Document Repositories and Relational Databases*, Proc. DASFAA'97, pp.145-154, April 1997.

[3] 森嶋厚行, 北川博之, 参照の導入による構造化文書とデータベースの統合操作の検討, 第 113 回情報処理学会データベースシステム研究会, 1997 年 7 月.

[4] V. S. Lakshmanan and F. Sadri, *Probabilistic Deductive Databases*, Proc. of the 1994 International Logic Programming Symposium, ILPS'94, Ithaca, NY, November 1994, pp.254-268.

[5] V. S. Subrahmanian, *Uncertainty in Databases and Knowledge Bases, Part V of Advances Database Systems*, pp.315-410, Morgan Kaufmann Publishers.

[6] M. J. Caret and D. Kossmann, *On Saying"Enough Already!" in SQL*, Proc. ACM SIGMOD International Conference on Management of Data, May 1997, pp.219-230.