

5J-03 SuperSQL を利用した XML データの格納と利用

赤堀 正剛† 有澤 達也‡ 遠山 元道§

†慶應義塾大学大学院 理工学研究科 管理工学専攻 ‡慶應義塾大学大学院 理工学研究科 計算機科学専攻

§慶應義塾大学 理工学部 情報工学科, JST さきがけ研究 21

1 はじめに

近年 XML で表現されたデータが増大し、データベース内のデータとの統合利用の必要性が高まりつつある。管理の為に XML 文書を関係データベースへ格納する [5] 以外に、運営中のデータとの連携を狙って XML データ*を単一リレーションへ格納し、SQL による既存のデータと統合利用する方法がある [6] が、単一リレーションへの格納ではデータの冗長性やデータの持つ意味をスキーマに反映しない、既存のリレーションへの追加・統合が行いにくい等の問題がある。そこで、データベース出版技術 SuperSQL [1, 2, 3] により XML のデータを関係データベース内の複数のリレーションへ正規化・分割して格納し、既存のデータと連携して統合的な情報の利用を行う†。

2 XML データの格納

2.1 SuperSQL

SuperSQL の質問文は SQL の SELECT 句を GENERATE < medium > < TFE > の構文を持つ GENERATE 句で置き換えたものである。ここで < medium > は出力媒体の指定で、< TFE > は構造とレイアウトを指定できる一種の式である。

2.2 XML データの分解における問題点

図 2-B の構造の XML データを図 2-A の複数リレーションへ格納する場合、キーと参照関係の生成が重要である。又、同じ著者が複数の論文を書いている場合などでは同一視できるオブジェクトが複数作られ、同一オブジェクトの同定が重要である。関係

データベース内に既存のデータと重複する場合もあり、XML データ内もしくは XML データと既存データ間で同一オブジェクトの統合・同一視、もしくは区別するかの選択が必要である。区別して格納する場合は既に使用されていないキーの生成が必要である。

2.3 格納と利用の流れ

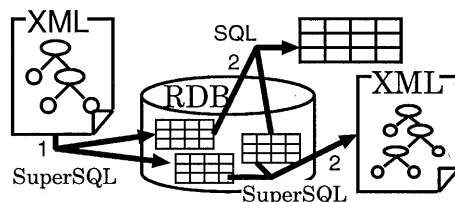


図 1: XML データの格納

1. SuperSQL により XML データを複数リレーションへ分解する。
2. 格納したデータと既存のデータを統合し、SQL や SuperSQL 等で出力を得る。

2.4 格納の為の指定

格納を指示する質問文には以下が必要である。

1. XML-関係データベーススキーマのマッピング
2. 重複処理指定
3. キーの生成に関する条件
4. 関係データベーススキーマに関する指定 (参照条件等)
5. 格納する XML データの指定
6. 格納先関係データベースのリレーションの指定
7. XML に関する条件

1 には SuperSQL の TFE の反復子と結合子を利用し 2 ~ 4 は TFE の修飾子 (@{ オプション }) によるオプション指定で指定する。5 は FROM 節で、6 は INTO 節で、7 は WHERE 節で指定する。例として、格納する XML と格納先のリレーションを図 2 に、格納するための SuperSQL 質問文を図 3 に示す。

2.5 質問文の構文

GENERATE 句の TFE では、生成するリレーションのスキーマを表す。入れ子の内側の反復子 ([]) でリレーションを表し、その名前は relation オプション

SuperSQL based XML repository on a relational database
AKAHORI Masatake†, ARISAWA Tatsuya‡ and TOYAMA Motomichi§

†Department of Administration Engineering, Keio University.

‡Department of Computer Science, Keio University.

§Department of Information and Computer Science, Keio University. PRESTO, JST.

*本論文では XML 文書で表現されたデータを XML データと呼ぶ。

†ユーザは XML・関係データベース双方の構造が既知である必要がある。ただし XML データに対応するリレーションを自動生成する場合はこの限りでない

で指定する。構成するデータベース属性は変数(\$で始まる)とパス式によるXMLデータの要素(属性は->でアクセス)のリストで表現する。リレーションに対するxmldistinct オプションをon とすると、XMLデータから同値のオブジェクトが生成された場合には一つしか格納しないことを示す。また、dbdistinct オプションをon とすると挿入先リレーションに存在しないデータしか格納しない(両者共デフォルト値はoff)。キーは関数KEY_GEN()で生成し、これはSkolem関数である。引数となる値が格納先のリレーションで使用されている場合にはそのキーの値を、そうでなければ新しいキーの値を返す。ただし、newkey オプションをon と指定すると常に新しいキーを生成する。データベース属性名には基本的にはXMLの要素名を流用するが、変更したい場合はname オプションを修飾子で指定する。value オプションにより変数でバインドすることも可能である。

```

GENERATE relation
[ [KEY_GEN($XP. 題名, $XP. 論文誌)
  @{name=id, value=$PID},
  $XP. 題名, $XP. 論文誌
]@{relation=$P,
  xmldistinct=on, dbdistinct=on},
[KEY_GEN($XPA. 著者名, $XPA. 所属)
  @{name=id, value=$AID},
  $XPA. 著者名, $XPA. 所属
]@{relation=$A,
  xmldistinct=on, dbdistinct=on},
[$PID, $AID
]@{relation=$PA,
  xmldistinct=on, dbdistinct=on}
]@{database=$PPP}
FROM Papers.xml $X, $X. 論文誌. 論文 $XP,
  $XP. 著者リスト. 著者 $XPA
INTO jdbc:postgresql://postgres/PAPERS $PP,
  $PP. 論文 $P, $PP. 著者 $A,
  $PP. 論文-著者 $PA
WHERE $X. 論文誌 = VLDB and
  $X. 論文誌->年度 = 2000

```

図 3: SuperSQL 質問文例

A. 格納先関係データベーススキーマ

論文

ID	論文タイトル	掲載論文誌
----	--------	-------

著者

ID	著者名	所属
----	-----	----

論文-著者

論文ID	著者ID
------	------

B. 格納するXMLデータのDTD

```

<!ELEMENT 論文リスト (論文)+>
<!ELEMENT 論文 (題目, 論文誌, 著者リスト)>
<!ELEMENT 題目 (#PCDATA)>
<!ELEMENT 論文誌 (#PCDATA)>
<!ELEMENT 著者リスト (著者名, 所属)>
<!ELEMENT 著者名 (#PCDATA)>
<!ELEMENT 所属 (#PCDATA)>
<!ATTLIST 論文誌 年度 CDATA #REQUIRED>

```

図 2: 格納するXMLデータと格納先のスキーマ

3 応用

3.1 格納データの統合利用

格納したデータと既存のデータの統合利用を行う。SQL等で使用する他、SuperSQLでXMLデータとして出力する事も出来る[4]。

3.2 TFEの自動生成

格納するXMLデータと同構造のXMLデータを出力するSuperSQL質問文がある場合はそれを元に、無ければ[7]等の手順で自動生成する。DTDで「A|B」と表記される構造に関してはA,B双方の属性を含むリレーションを用意し、要素が存在すればその値を格納し存在しない方の値はNULL値とする。

4 おわりに

本論文ではXMLデータと関係データベース内の既存データとの連携・統合を実現する為、XMLデータを複数リレーションへ分解し、既存のリレーションへ追加する方法を提案した。今後の課題として、半構造化への対応やユーザ関数による自由なキー生成、データ更新への対応等を検討している。尚、本研究の一部は情報処理振興事業協会(IPA)の高度情報化支援ソフトウェアシーズ育成事業の補助による。

参考文献

- [1] SuperSQL: <http://ssql.db.ics.keio.ac.jp>
- [2] 遠山元道 他: レイアウト式 TFE の拡張, 情報処理学会研究会報告, 95-DBS-104 pp.217-224
- [3] M.Toyama, SuperSQL: An Extended SQL for Database Publishing and Presentation. in *Proc.SIGMOD'98*, ACM(1998), 584-586.
- [4] 赤堀正剛 他: SuperSQL による XML データドキュメントの自動生成, 情報処理学会研究会報告, 2000-DBS-122 pp.455-462
- [5] 石川博: XML とデータベース, 情報処理学会誌, Vol.41, No.1, pp.68-73, 2000.
- [6] 春日史朗 他: 異種情報源環境における情報検索・統合検索方式, 電子情報通信学会 DEWS2000 論文, 7A-5,2000
- [7] Jayavel Shanmugasundaram et al.: Relational Databases for Querying XML Documents: Limitations and Opportunities. VLDB 1999: 302-314