SuperSQLを利用した構造化文書生成の試み

5P－4

赤堀 正剛† 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 理工学部 情報工学科, さきがけ研究 21/JST

1 はじめに

SuperSQL[1, 2] は、関係データベースからの問い合わせに対する結果を階層的に構造化し、種々のメディアに変換して出力する。

データベースに格納されているデータは表の中だけでなく文章中にも使用されることがある。一般に文書中には深さが一定ではない章や節があり、各々に複数の文章が存在する。また、段落などが存在し、図や表なども付加される事もある。このような構造の文書は通常の SuperSQL の階層構造では扱えず、生成できない場合がある。

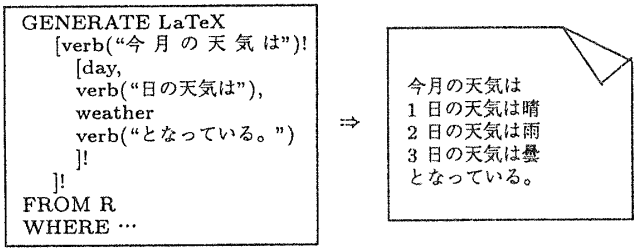
そこで本研究では文章中でのデータ利用の方法を検討し、構造化文書を生成する事を目的とする。

2 従来の SuperSQL による文章生成

2.1 従来の手法

従来の SuperSQL では、verb 関数*によって質問文中に書かれた文字列を出力に反映させていた。この手法により、データ部分のみが変化する繰り返しなどの単純な構造的な文章は簡単に生成できる。

表 1: 従来の手法による構造的な文章生成



Generate Structured Documents by SuperSQL

AKAHORI Masatake†, TOYAMA Motomichi§

†Department of Administration Engineering, Faculty of Science and Technology, Keio University.

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

*verbatim の略:引数をそのまま出力する関数

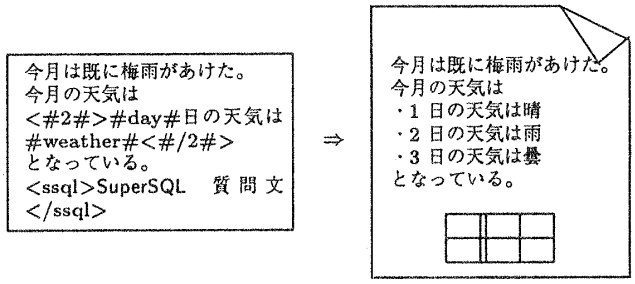
2.2 問題点

ところが、文章が大量に存在する長大な文書を出力する場合データの配置 (属性の並べ方) や関係等がわかりにくくなり、全体の流れがつかみにくい。

3 提案する手法

質問文の中に文章を記述するのではなく、質問文とは別に文章を記述する手法を提案する。別に記述する文章中には SuperSQL に似た構造を表現するタグと属性変数を埋め込み SuperSQL と同等の表現力を持たせる。また 4 次元以下の SuperSQL 質問文を付加することにより表を挿入する。

表 2: 提案する手法による構造的な文章生成



4 章・節構造を持つ文書の生成

章や節などからなり、それぞれに複数の文章が存在する文書 D について考える。文書中には表や図などの要素も存在する。章・節・文章・表・図などの要素は場所により存在しないこともあると仮定する例えば図 1 のような構造の文書などである。本論文

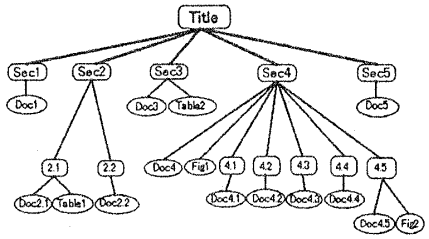


図 1: 章・節構造を持つ文書例 (本論文の構成)

では、文書 D のように場所により任意の深さをとる構造の文書を木構造†文書と呼ぶ。

4.1 文章生成の拡張

3章で提案した手法において元々章・節だてされた文書を記述しておくと、先に仮定した文書 D を生成できる。これは通常の文書を記述することと似ているが、応用データである文書を直接記述することと比べ、使用しているデータの変更があった際のデータの整合性が取りやすい等の利点がある。

4.2 等しい深さの階層構造の文書の生成

SuperSQL で扱える次元数に上限は無いのだが、今まで使用されている次元は4次元までである。そのため、データの出力形式としては表形式が主たるものであり、他の形式の試みは余りなされていない。

そこで、構造化文書を生成するように或次元を対応させることにより章・節構造を持つ文書を生成する (表3参照)。

表 3: 等しい深さの階層構造の文書の生成

```
GENERATE LaTeX
    [
      sec!sentence_s![tab_s]![fig_s]! !
      [ssec!sentence_ss![tab_ss]![fig_ss]!]!
    ]!
FROM …
WHERE …
```

ただし、通常の SuperSQL では反復している各々の要素は同質で階層構造の深さはどの場所でも等しく、バランスが取れている。このため、異なる深さをもつ要素を反復出来ない。

4.3 木構造文書の生成 I

SuperSQL では外部結合を用いると木構造データ‡を扱うことは可能である [3]。外部結合による木構造データの生成により木構造文書が生成できる。この手法を取る場合、最大の深さとなる階層までを記述する。

4.4 木構造文書の生成 II

木構造文書の場合、構造をデータベースに格納しておくとしても深さが不定であるため、この構造をそのままスキーマに反映して格納しておくことは好ましくない。変更等があった場合に対処しにくいからである。そのため、親子関係を基準として格納しておく方法が考えられる。

†深さが均一でない階層構造
‡[3] では半構造データと呼んでいる

SuperSQL では親子関係を基準としたバランスのとれていない木構造の出力を生成可能である [4]。この特性を文書構造に適用する。

4.5 参照を持つ文書

他の文書や同じ文書内にあるオブジェクトへの参照を持つ文書も存在する。同じ文書へ自己参照する場合、その文書の全文を出力すれば問題はないが、一部だけを出力すると必要な参照先が欠如することもある。この為、文章中の参照の情報を保存しておき、部分文書を出力する際に一緒に参照先を出力する。

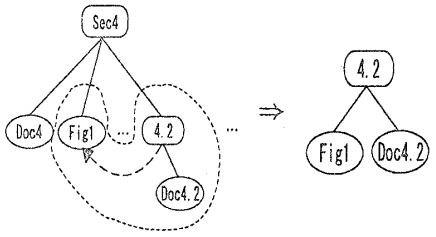


図 2: 参照のある文書例 (Sec.4.2 のみ出力する)

また、複数箇所から同じオブジェクトへの参照がある場合には重複しないように処理する。

5 おわりに

本研究では、データを含む文章の生成と SuperSQL の持つデータの階層構造化を利用した構造化文書の生成を行った。その際、従来の SuperSQL では質問文だけで出力を得る方式を採用していたが、SuperSQL 質問文だけでなく別個に文章の構造を指定する手法を提案した。今後の課題として、提案手法の改善や有効性の検討などが挙げられる。

参考文献

[1] SuperSQL: http://ssql.db.ics.keio.ac.jp

[2] M.Toyama, SuperSQL: An Extended SQL for Database Publishing and Presentation, in *Proc.SIGMOD'98*, ACM(1998), 584-586

[3] 遠山元道: 関係データベースに基づく半構造データの実現と管理, 情報処理学会データベースシステム研究会資料, 98-DBS-114-15 pp.105-112

[4] 小原 彰、遠山元道: TFE を用いた再帰的問い合わせ表現とその処理, 第 55 回全国大会, 2X-02 pp.3-306～3-307