4K-09 SuperSQLを用いたデータ分析システムの試作

田中 覚† 遠山 元道§

†慶應義塾大学 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

例えば、各地域毎のデータを分析する際には、各地域毎にクエリを発行し必要なデータを取得する必要がある。また、取得されたデータを集計するために表計算ソフトを用いて集計表を作成し、さらに必要に応じてグラフ化なども行う必要が生じる。これらの処理は独立しており、データ量が多い場合などには手作業で繰り返し同じ処理を行わなくてはならない。

関係データベースからの出力結果として、直接クロス集計表やグラフを得ることができれば、データ分析が行いやすくなるので、そのメリットは大きい。これはSuperSQL[1, 2] を拡張した質問文中に計算の指示を埋め込むことで実現できる。この「RDBからのデータの抽出」と「データの分析処理」の2つの処理過程を統合する方法を論じる。

2 データ分析システムの試作

本システムは、1) 集計表の作成 2) グラフ描画 を実現するためのExcel上で動作するVBA(Visual Basic for Application)[3] マクロを自動生成する。

2.1 Pivot関数

Excelの代表的な機能としてピボットテーブルを利用した集計がある。標準SQLでは、集計表の形で出力結果を得ることはできない。また、地域毎の集計表を得たい場合には、ピボットテーブルを多数作成する必要がある。このような場合には、SuperSQLのグルーピング ([ ]) と新たに定義したPivot関数を利用してこの問題を解決することができる。以下に具体例を示す。

```
GENERATE excel
 [verb(地域),          <=    索引部          (1)
  [所属 %              <=    索引部          (2)
   [所属,              <= フラットテーブル (3)
    pivot([            <= ピボットテーブル (4)
     月@{name=月,position=columnfields},
     t.名前@{name=講師,position=rowfields},
     s.名前@{name=生徒},
     s.性別@{name=性別,position=pagefields},
     s.科目@{name=科目,position=pagefields},
     単価@{name=時給},
     時間@{name=コマ数},
     cell@{name=日報,formula=時給*コマ数,
           position=pivotfields,calc=sum}
    ]!@{type=積み上げ縦棒,series=Rows,   (5)
        title=月別集計グラフ,            (6)
        category=月,value=給料(円)})     (7)
   ]! (3)
  ]!  (2)
 ]!   (1)
from  ...
where ...
```

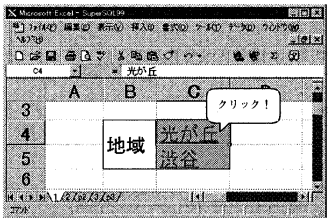

図 1: 実行結果(索引部)

はじめに地域に関する索引部を生成する ((1),(2), 図1)。次に地域(所属)毎にフラットテーブルを生成する ((3)-(5), 図2)。二項演算子%(2) により索引部の各所属(2) から各所属のフラットテーブル ((3)-(5)) へのリンクを自動生成する(図2)。その後、Pivot関数によりフラットテーブルを元にピボットテーブルを生成する。このピボットテーブルは新規のワークシート上に生成され、フラットテーブルからピボットテーブルへのリンクが自動生成される ((4)-(5), 図3)。最後にPivot関数のオプションとして、ピボットテーブルからグラフを自動生成することもできる ((5)-(7), 図3)。

A Prototype of data-analysis tool using SuperSQL

TANAKA Satoru†, 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

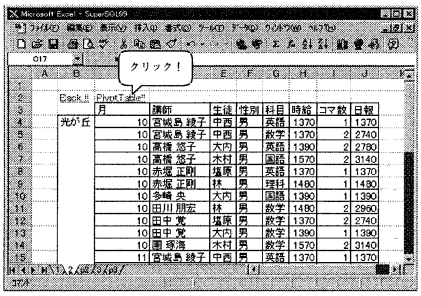

図 2: 実行結果 (フラットテーブル)

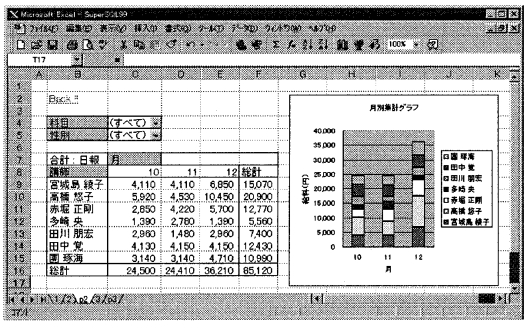

図 3: 実行結果 (ピボットテーブルとグラフ)

Pivot 関数には、1)name により属性の名前を指定する　2)position により集計表の位置を指定する　3)calc により集計対象となる属性の集計の方法を指定する　等の主なるオプションを備えている。

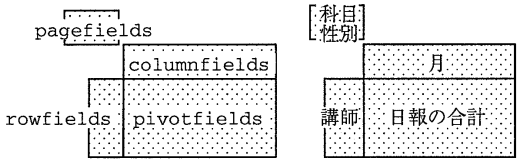

図 4: ピボットテーブル

この様にひとつの質問文の記述で、各地域毎にピボットテーブルとグラフを自動生成することができる。これにより手作業で行う繰り返しの処理をする必要がなくなる。さらに、View 表 ↔ ピボットテーブル ↔ グラフ の相互にハイパーリンクが自動的に張られるので、ワークシート上での出力結果の閲覧も容易である。

2.2 属性間の計算機能

先に示した例で、Pivot 関数の中で属性間の計算により “日報” を計算していた。日報は “時給 * コマ数” により計算される (*calc* を *sum* と指定することでピボットテーブルで集計する際に “日報” を合計 (*sum*) する様に指定している)。このように Pivot 関数中には、属性間の計算機能が備わっており、DB には存在しない新しい属性の値を導出することができる。また Pivot 関数中に記述されている *cell* という属性は属性間の計算結果を格納するためのスペースを表計算のワークシート上に確保する役割がある。

計算処理自体は Excel のワークシート関数を用いて行っている。したがって SuperSQL の質問文中に記述された計算式をワークシート関数に変換する処理が必要である。(尚、Pivot 関数の中で記述できる計算式または条件式は Excel のワークシート関数であれば大抵可能)

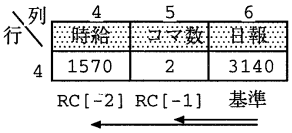

図 5: 式の変換処理

この変換処理は、formula で指定した属性の名前の部分を計算結果が表示されるセル位置からの相対座標で置換することで達成できる。

```
Call Set_Formula(_
  MyCell:=Cells(4,6),_
  MyFormula:="時給*コマ数",_
  MyElement:=Array("時給","コマ数"))
  日報="時給*コマ数",
=> "RC[-2]+コマ数" => "RC[-2]+RC[-1]"
=> "="+"RC[-2]+RC[-1]"
```

3 評価・検討

SuperSQL 質問文の反復連結を利用し、複数のピボットテーブルを一括生成できることは他の *4GL* ツールにはない特徴である。さらに、出力先でデータの更新があった場合に、その変更を DB にフィードバックする機構が備わっていれば、DB の更新に役立てることができるだろう。

4 おわりに

本稿では、「データの抽出」と「データの分析」を統合したシステムを提案した。これにより多くの分析者に対して DB から常に最新のデータを提供できる上に作業効率の向上が望める。今後としては、DB とワークシート上のデータの一貫性が保てるシステムの構築や、分析結果を分析するといった多段階の分析をサポートできる機能の実装を検討している。

参考文献

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

[2] SuperSQL,http://www.db.ics.keio.ac.jp/ssql

[3] 井上 俊宏 : 『Excel97 VBA の応用 70 例』、ソフトバンク、1998