OLAP 큐브의 다중 차원계층구조에 대한 분석*

박 영 선°, 김 지 현, 임 윤 선, 김 명
이화여자대학교 컴퓨터학과
{ys_park°, jhrosa, lys96, mkim}@ewha.ac.kr

Analysis of Multiple Dimension Hierarchies of OLAP Cubes

Youngsun Park°, Jihyun Kim, Yoonsun Lim, Myung Kim
Department of Computer Science and Engineering, Ewha Womans University

요 약

롤업과 드릴다운은 다차원 데이터 분석을 위한 주요 연산으로, 각 차원에 정의된 계층구조를 통해 상세 데이터로부터 점차적으로 요약되는 정보를 분석가에게 제공한다. 이러한 연산 속도를 고속화하기 위해 OLAP 시스템은 사전에 집계 테이블들을 생성해 놓는다. 각 차원은 다중 계층구조를 가질 수도 있으며, 이런 경우 집계 테이블들을 모두 생성하게 되면 데이터 폭발 현상이 발생하게 된다. 본 연구에서는 다중 계층 구조를 분류하고, 집계 테이블과 데이터 큐브의 크기를 계산하는 모델을 정립하였다. 이를 통해 분석가는 다중 계층구조에 따른 큐브 크기를 미리 예측할 수 있으며 계층 구조의 모양과 개수를 변경하여 데이터의 양을 조절할 수 있다.

1. 서 론

OLAP(On-Line Analytical Processing) 시스템은 사용자가 요구하는 데이터를 다차원적으로 신속·정확하게 제공하기 위해 분석결과의 일부분을 미리 계산하여 큐브(cube)에 저장해 놓는다. 큐브의 크기는 각 차원에 정의되어 있는 계층구조의 모양에 따라 원본 데이터의 수 천 배가 될 수도 있으며, 특히 차원에 여러 개의 계층구조가 정의되어 있을 때, 차원에 "다중 계층 구조"가 있다고 하며, 이런 경우 데이터 폭발 현상은 더욱 심각하다 [1].

실제 응용에서 사용되는 차원 계층구조들은 정형화되어 있지 않은 경우가 많으며, 각 차원에 여러 개의 계층구조가 정의되어 있는 경우 이러한 구조가 큐브 크기에 미치는 영향을 미리 예측하기는 힘들다. 계층 구조가 조화되지 않은 경우도 흔하고, 특정 차원에 정의된 여러 계층 구조가 서로 차원 레벨을 공유하는 경우도 있고, 복잡하고 체계화되지 않은 연결 구조를 가지는 경우도 있다 [2, 3].

다중 계층 구조에 대한 연구로는 신속한 연산 처리를 위해 이들을 구현하는 방법들이 제시되어 왔다. 차원 생성 방법[4]과, 선형 방법[5]이 대표적 연구 결과이다. 그러나 차원의 다중 계층구조의 모양에 대한 집중적인 분류 작업은 아직 제안되지 않았으며, 본 연구는 차원 계층 구조를 분석하고 이를 통해 생성될 큐브의 크기를 사전에 계산해 볼 수 있도록 하여 분석가가 계층 구조를 큐브 생성 전에 조절할 수 있도록 하는 방안을 제시하는 것을 목표로 한다.

구체적으로, 본 연구에서는 특정 차원이 여러 개의 계층 구조를 가질 때 어떤 모양을 하는지, 여러 계층 구조가 서로 어떤 영향을 주는지를 분류하였다. 이러한 모양에 따라 OLAP 큐브에 속할 집계 테이블의 개수와 큐브의 예상 크기를 계산하는 모델을 제시하였다. 이를 통해 분석가는 데이터 증가를 예상할 수 있고, 자신이 사용할 환경에 적절한 사전 연산을 할 수 있도록 하였다.

본 논문의 구성은 다음과 같다. 2절에서 다중 계층구조의 분류 체계를 제안하고, 3절에서는 본 연구에서 제안한 다중계층구조 분류에 따른 집계 테이블과 큐브 크기를 나타내는 모델을 제시한다. 4절에서 결론을 맺는다.

2. 다중 계층구조의 분류

본 연구에서는 다중 계층 구조를 표 1에서와 같이 4가지 기준으로 분류하였다. 한 차원에 다중 계층 구조가 정의되는 경우 다중 트리가 존재한다고 보지 않고, 루트는 공유하는 것으로 간주하였다. 분류 체계의 첫째 기준은 계층구조를 나타내는 트리가 완전 트리인가에 대한 것이다. 계층이 완전트리 구조를 이루는 경우를 '조화 계층 구조'라고 하고, 그렇지 않은 경우를 '부조화 계층 구조'[5, 6]라고 한다.

둘째 분류 기준은 차원 계층 구조가 한 개의 트리를 이루는지 루트를 공유하는 여러 개의 트리로 나뉘어 지는가에 대한 분류이다. 단일 트리로 된 경우를 '단일 계층구조' 그렇지 않은 경우를 '다중 계층 구조'로 분류하였다.

셋째 분류 기준은 다중 계층 구조인 경우, 여러 계층 구조가 레벨을 어떤 방식으로 공유하는가에 대한 분류이다. 그림 2(a)와 같이 요약되는 과정에서 상위 레벨에서 만나는 구조를 '교차 다중 계층 구조'라고 하고, 그림 2(b)와 같이 하위 레벨 (상세 데이터 방향)에서 공유하다가 상위 레벨로 올라가면서 분기하는 경우를 '분기 다중 계층 구조'라고 한다. 공유하는 레벨이 많은 경우 OLAP 큐브 생성시 집계 테이블의 공유가 증가하여 큐브 크기를 줄일 수 있다.

* 본 연구는 2003년 한국과학재단 우수여성과학자 도약지원연구사업(R04-2001-000-00191-0(2003))지원에 의해 수행되었음.

분류의 마지막 기준은 해당 레벨의 멤버의 순서가 정렬된 채로 다중 계층구조를 이루는가에 대한 분류이다. 이와 같은 분류 체계에서 멤버들의 순서 정렬 또는 비정렬 구조이면서, 하나의 계층 경로와 또 다른 계층 경로가 만나 공유레벨이 발생하는 교차 다중계층구조인 경우 집계 테이블이나 큐브 데이터 크기에 가장 큰 영향을 미치는 구조가 된다.

표. 1. 계층 구조의 분류.

	Balancing	Number of Path	Intersection of shared level	Member ordering
Hierarchy	Balanced	Single Path		
		Multi-Path	Bipolar (Intersection)	ordered
				not ordered
			Polar (Branch)	Ordered
				not ordered
	Not Balanced	Single Path		
		Multi-Path	Bipolar (Intersection)	ordered
				not ordered
			Polar (Branch)	ordered
				not ordered

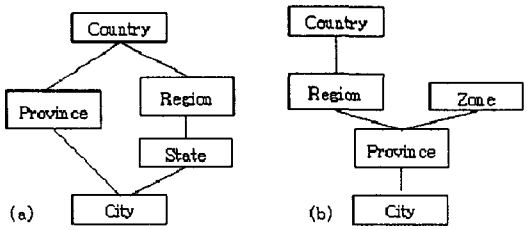

그림 2. (a) 교차다중계층구조 (b) 분기다중계층구조.

3. OLAP 큐브 크기 계산 모델

이 절에서는 계층 구조와 데이터 크기의 연관성을 분석하였다. 표 1의 분류체계에서 조화 교차 다중 계층 구조를 중심으로 모델을 세웠다. 부조화 계층 구조의 경우는 큐브 크기 계산에서는 조화 계층 구조, 단일 계층 구조와 분기 다중 계층 구조는 조화 교차 다중 계층 구조의 특별한 형태로 취급될 수 있기 때문이다. 구체적으로 본 연구에서는 단일계층구조를 갖는 경우, 교차 다중계층구조에서 멤버들의 순서를 고려하여 집계 테이블의 수와 큐브 셀 개수 계산 모델을 세웠다.

(1) 단일 계층구조를 갖는 경우

OLAP 큐브는 여러 집계 테이블(summary table)로 구성되고, 집계 테이블은 데이터 셀로 구성된다. 원본 데이터가 D_1, D_2, D_3, $\cdots$, D_n 차원으로 구성되어 있을 때, 이로부터 생성되는 집계 테이블의 수는 2^n 이 된다. 또한 큐브 전체의 크기는 이런 집계 테이블의 셀 개수를 합한 값이다. 차원 D_i의 멤버의 수를 d_i라고 했을 때, 큐브 셀 개수는 $\prod_{i=1}^{n}(d_i + 1)$이다.

차원 D_i, $1 \le i \le n$, 가 h_i 개의 계층을 갖는다고 하면, OLAP 큐브의 집계 테이블 개수는 $(h_1+1)\times(h_2+1)\times\cdots\times(h_n+1)$이 되고, 이를 달리 표현하면, $\prod_{i=1}^{n}(h_i+1)$로 표시될 수 있다. 집계 테이블의 셀 개수는 집계 테이블을 구성하는 각 차원의 멤버의 수의 곱이므로, 차원의 계층 개수를 쓰는 대신에, 각 계층의 멤버의 수를 쓰면 된다. 이와 같이 계산하면, OLAP 큐브의 크기는 $\prod_{i=1}^{h_i}(d_{i,1} + d_{i,2} + \cdots + d_{i,h_i} + 1)$ 이다.

(2) 순서 비정렬 다중계층구조를 갖는 경우

교차 · 분기 다중계층구조 중에서도 순서 비정렬로 다중계층구조를 이루고 있는 경우는 공유 레벨이 존재해도 공유레벨을 추가하여 집계 테이블의 개수를 계산해야 한다. 공유레벨의 멤버들이 순서가 바뀌면서 다중계층구조를 이루게 되면 순서가 바뀐 멤버들에 해당하는 레벨의 집계 테이블들을 모두 다시 생성해야 하기 때문이다. 그러므로 순서 비정렬 다중계층구조를 갖는 큐브는 공유레벨의 의미가 없어진다. 따라서 모든 다중 계층구조는 공유하는 집계테이블이 존재하지 않게 되어 독립적인 집계테이블의 수를 계산해야 한다.

순서 비정렬 다중 계층구조를 갖는 n차원 큐브의 집계 테이블의 수와 셀 개수는 다음과 같이 계산된다. 우선 큐브는 $D_1, D_2, D_3, \cdots, D_n$ 차원으로 구성되고, 차원 D_i는 m개의 계층을 갖는다고 하자. D_i차원의 j번째 계층의 총 레벨 수를 l_{ij}라고 하면 큐브의 집계 테이블의 개수는 $\prod_{i=1}^{n}(\sum_{j=1}^{m} l_{ij} + 1)$이다. 즉 집계 테이블은 n개 차원의 조합으로 구성되고, 집계 테이블에 포함되는 차원은 해당 차원의 특정 레벨이기 때문이다.

$d_i(j,k)$를 차원 D_i의 j번째 계층의 레벨 k의 멤버 수라고 하면, 큐브의 셀 개수는 다음 식으로 계산된다.

$$\prod_{i=1}^{n}\left(\sum_{j=1}^{m}\sum_{k=1}^{l_{ij}} d_i(j,k) + 1\right).$$

(3) 순서 정렬 다중계층구조를 갖는 경우

멤버의 순서가 정렬되어 한 레벨에서 공유된 후 상위 레벨에서 다시 공유되는 구조의 다중계층구조는 공유 레벨의 멤버들이 다시 새로운 레벨로 그룹핑 될 때 멤버들의 순서의 변함없이 그룹핑 되는 다중계층구조를 말한다. 이러한 다중계층구조에서는 공유레벨에서 생기는 집

계 이블들을 서로 공유할 수 있다. 그림 3은 시간 차원에서 두 개의 계층구조 C와 F를 갖는 순서 정렬 다중계층구조의 예이다. 그림 3의 시간 차원은 Month 레벨에서 분기한 후 상위레벨인 Year에서 공유되는 구조이다. 이 시간 차원의 집계 테이블은 분기되는 C_2와 F_2만 각각 독립적으로 생기고 공유레벨에서는 어느 계층구조이든 한쪽의 계층구조 레벨만 포함시켜 집계테이블을 생성하면 된다. 즉 그림 3의 예제에서는 C_1, C_2, C_3, C_4, C_5, F_2 레벨에 대한 집계 테이블을 생성하면 된다.

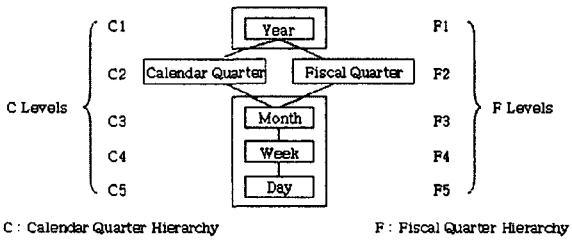

그림 3. 시간 차원의 순서정렬 교차 다중계층구조

순서 정렬 다중계층구조를 갖는 n차원 D_1, D_2, D_3, $\cdots$, D_n 큐브의 집계 테이블 수는 다음과 같이 계산된다. 우선 S_i를 다음과 같이 정의하자.

$S_i = \{h_i(j,k) \mid 1 \leq j \leq m, 1 \leq k \leq l_{ij},$
$\forall j_1, j_2, k_1, k_2$ such that, $h_i(j_1,k_1) \neq h_i(j_2,k_2)\}$

S_i는 차원 D_i를 위해 정의된 m개의 계층 구조에 포함된 모든 레벨들 $h_i(j,k)$, $1 \leq j \leq m, 1 \leq k \leq l_{ij}$, 중에서 중복되는 레벨들을 제외한 레벨들의 집합이다. $L_i = |S_i|$로써, S_i의 원소들이 개수이다. 이렇게 정의할 때 집계 테이블의 개수는 $\prod_{i=1}^{n}(L_i+1)$로 계산될 수 있다.

순서 정렬 다중계층구조를 갖는 n차원 D_1, D_2, D_3, $\cdots$, D_n 큐브의 셀 개수는 다음과 같이 계산된다. 우선 Q_i를 다음과 같이 정의하자.

$Q_i = \{d_i(j,k) \mid 1 \leq j \leq m, 1 \leq k \leq l_{ij},$
$\forall j_1, j_2, k_1, k_2$ such that, $h_i(j_1,k_1) \neq h_i(j_2,k_2)\}$

Q_i의 원소인 $d_i(j,k)$는 S_i에 속한 $h_i(j,k)$의 셀 개수를 뜻한다. 이렇게 정의할 때 큐브의 셀 개수는 $\prod_{i=1}^{n}(\sum_{j=1}^{m} Q_i+1)$로 계산될 수 있다.

예를 들어, 시간, 상점, 상품 차원을 갖는 3차원 데이터 큐브에서 상품과 상점 차원이 각각 P_1, P_2, P_3와 S_1, S_2, S_3의 레벨 3인 단일 계층구조이고, 시간 차원이 그림 3과 같이 이중 계층구조를 갖는 경우 생성할 수 있는 집계테이블은 $(P_1+P_2+P_3+1)\times(S_1+S_2+S_3+1)\times(C_1+C_2+C_3+C_4+C_5+F_2+1)$이다. 따라서 112개의 집계테이블이 생성 된다.

4. 결 론

본 연구에서는 OLAP에서 가질 수 있는 모든 계층구조들을 분류하고, 각 계층구조에 따라 발생하는 집계테이블의 개수와 셀의 개수를 구하는 모델을 제시하였다.

본 연구에서 분류한 계층구조 중 다중계층구조는 계층의 경로에서 다시 만나는 공유 레벨이 있는지 없는지에 따라 교차 다중계층구조와 분기 다중계층구조로 1차 분류하였으며, 다중계층구조의 공유하는 레벨에서 멤버들이 순서대로 다시 그룹핑 되어 경로를 가지는지 아니면 순서가 바뀌어 그룹핑 되어 경로를 가지는지에 따라 2차 분류하였다. 이 2차 분류는 다중계층구조를 가지는 차원의 집계 테이블이나 데이터 큐브 크기 구하는 식이 달라지는 중요한 부분이다. 이와 같은 구조 중 다중계층구조가 순서 정렬로 구성되어 있는 차원일수록 중복된 데이터의 감소로 인한 데이터 큐브 크기를 줄일 수 있다는 것을 증명하였다.

계층구조 분류 체계와 데이터 크기를 예상할 수 있는 모델을 제안함으로써, 분석가들이 데이터 증가를 예상하여 데이터 폭발 및 계층의 혼돈이 발생하지 않도록 자신이 사용할 환경에 적절한 사전 연산을 할 수 있게 하였다는데 의의를 가질 수 있다.

5. 참고 문헌

[1] White Paper, http://www.olapreport.com /DatabaseExplosion.htm.

[2] E.Pourabbas, M.Raffanelli, "Characterization of hierarchies and some operators in olap environment," ACM Press, pp. 54-59, 1999, NewYork, NY, USA.

[3] R.Agrawal, A.Gupta, S.Sarawagi, "Modeling Multidimensional Databases," ICDE, pp. 232-243, 1997.

[4] http://msdn.microsoft.com/library/default.asp?URL=/library/techart/dimmsdn.htm

[5] R.Pieringer, V.Markl, "HINTA:A Linearization Algorithm for Physical Clustering of Complex OLAP Hierarchies," DMDW(11.1-11.11), 2001.

[6] http://www.paristech.com/PowerOLAP_help_files/powerolap