OLAP 시스템에서 희박 데이터의 패턴 분류 및 성능 평가

강주영° 이봉재 송재주 신진호 용환승
한국전력 전력연구원, 이화여자대학교
{jykang°, bjlee, jjsong, jinho}@kepri.re.kr, hsyong@ewha.ac.kr

Korea Electric Power Research Institute, Ewha Womans University

Juyoung Kang° Bongjae Lee Jaeju Song Jinho Shin, Hwanseung Yong
Power Information Technology Group, KEPRI
Dept. of Computer Science and Engineering, EIST, Ewha Womans University

요 약

OLAP(On-Line Analytical Processing)은 데이터 웨어하우스 내의 방대한 양의 데이터에 대해 사용자와의 상호 작용이 가능하도록 질의에 대하여 빠른 응답성능을 보장해야 한다. 이를 위해 OLAP 시스템은 데이터에 대한 다량의 다차원 집계 연산을 수행해야 하기 때문에, 일반적으로 사전 연산 결과를 저장하여 직접적인 집계 연산을 줄임으로써 응답 성능을 높이는 방법을 사용하고 있다. OLAP 다차원 데이터의 희박성은 이러한 사전 연산 시 데이터 폭발 현상을 일으켜 도리어 성능을 저하시키는 요인으로 작용할 수 있다. 본 논문에서는 데이터의 희박성과 성능 문제에 대해 고찰하고 OLAP 응용에서 발생할 수 있는 다차원 데이터의 희박성 패턴에 대해 정의하였다. 또한 정의된 패턴에 따라 희박 데이터를 생성하는 데이터 생성기를 구현하고 이를 이용하여 생성된 데이터를 기반으로 MS SQL Server Analysis Services와 Pilot DSS의 두 OLAP 제품의 성능을 평가하고 결과를 비교하였다.

1. 서 론

OLAP은 전사적으로 통합된 데이터 웨어하우스에 저장된 다량의 데이터에 대한 임의의 질의에 대해 그 결과를 온라인으로 사용자에게 제공하는 기술을 말한다[1]. OLAP 시스템은 다루어야 하는 데이터 용량이 방대하고 OLTP 응용과는 달리 질의가 복잡하기 때문에 원초 데이터로부터 직접적으로 실행 될 경우 수 시간에서 수 일까지의 수행시간이 요구된다. 이러한 성능 병목 현상은 OLAP 응용이 많은 양의 다차원 집계 연산을 필요로 한다는 특성에서 기인하는데, 이를 해결하기 위해 분석 결과의 일부를 미리 계산하여 집계 테이블 내에 저장해 두는 방법을 주로 사용한다[2]. 이렇게 사전 연산 결과를 저장해 두어야 하는 경우, 데이터의 특성에 따라 저장해야 할 데이터의 분량이 분석하고자 하는 데이터에 비해 폭발적으로 증가할 수 있다. 실세계의 다차원 데이터는 대부분 매우 희박한 특성을 가지는데, 이러한 희박 데이터는 사전 집계 연산 수행시 데이터 폭발 현상을 일으키는 주된 요인이 된다. 데이터 폭발 현상은 디스크 공간의 낭비와 질의 응답 시간을 증가 시킨다는 점에서 OLAP의 성능을 저하시키는 가장 중요한 요인이라 할 수 있다[3].

기존에 제안된 대부분의 다차원 색인 및 저장 기법들은 일반적으로 데이터가 밀집한 경우를 가정하고 있으며, 실제 응용 데이터의 희박성을 고려한 색인 및 저장 기법에 대한 연구는 충분히 이루어지지 못한 상태이다[4]. 현재까지 연구된 희박성 관리 기법들로는 청크 개념을 이용한 배열 저장 구조, 복합 차원, 희박-밀집 분리 기법 등이 있다. 하지만 이러한 기법을 적용하고 있는 OLAP 시스템의 수가 극히 드물 뿐 아니라, 데이터가 특정 희박성 패턴을 가진다고 가정하여 처리하는 방법을 취하고 있기 때문에 복합적인 희박성을 띄는 실제 응용 데이터의 처리 기법에 대한 연구는 미흡하다고 할 수 있다.

본 논문에서는 OLAP 데이터의 희박성 문제와 기존의 희박성 처리 기법에 대해 살펴보고, OLAP 데이터의 희박성 패턴에 관해 분석하고 정의한다. 또한 정의한 패턴에 따라 희박 데이터를 생성하는 데이터 생성기를 설계 및 구현하고, 이를 기반으로 실험 데이터를 생성하여 Microsoft SQL Server 2000 Analysis Services(MS AS)와 Pilot Decision Support Suite(Pilot DSS)의 두 가지 OLAP 제품의 데이터 희박성에 따른 성능을 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 희박성 처리 기법들에 대한 관련연구를 소개하고 3장에서는 OLAP 시스템의 희박 데이터의 패턴을 분류하고 정의하며 이를 기반으로 하여 설계, 구현한 희박 데이터 생성기에 대해 기술한다. 4장에서는 두 가지 OLAP 제품들을 기반으로 성능 평가를 수행하고 결과를 분석한다. 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

기존의 OLAP 다차원 색인 및 저장 기법들은 대부분 데이터의 희박성이 0%인 밀집 데이터를 가정하고 있지만, 몇몇 상용 제품에서 특정 형태의 희박 데이터를 처리하는 기법이 사용되고 있다. ROLAP은 희박한 셀은 튜플로 저장하지 않기 때문에 희박성이 큰 문제가 되지 않지만, MOLAP의 경우 다차원 배열구조를 사용하기 때문에 희박 셀에 대한 처리가 매우 중요하다. 이러한 MOLAP 배열 기반의 희박성 처리 기법으로 청크 기반 기법이 제안된 바 있다[5]. 이 방법은 배열을 이용하여 작은 단위의 청크를 구성하고 이를 밀집청크와 희박청크로 구분하여 밀집청크는 그대로, 희박청크는 각 유효 셀의 색인 값으로 데이터를 저장하는 방법이다.

현재 EssBase 제품에서 사용되고 있는 희박-밀집 분리(Sparse-Dense Split) 기법은 희박 셀의 분포적 특성을 고려하여 두 차원간의 관계를 밀집차원과 희박차원으로 구분한다. 밀집차원인 경우 데이터를 배열 형태의 밀집블록으로 저장하고, 희박차원인 경우는 각 밀집블록들에 대해 인덱스의 역할을 하도록 블록 포인터 배열이나 이진 트리 인덱스 구조로 구성한다. 이렇게 차원간의 관계를 두 종류로 구분함으로써 유효 셀들을 포함한 밀집블록만을 저

장하여 저장공간을 효율적으로 사용할 수 있다[6].

Oracle Express의 복합 차원(Composite Dimension) 기법은 실제 OLAP 데이터에서는 차원 항목 전체에 걸쳐 희박성이 나타나는 것이 아니라, 차원의 특정 항목 집합간의 관계에서 희박성이 나타난다고 가정하고 있다. 이 기법은 기본 차원의 항목들을 조합하여 복합 차원을 구성하고 실제로 유효 데이터를 포함한 셀만을 저장하도록 하여 희박성을 처리하고 있다[7].

이 외에도 비트 연산 방법을 통해 데이터에 접근할 수 있도록 인덱싱 하여 청크로 나뉘어진 유효 셀들만을 저장하는 BESS(Bit-Encoded Sparse Structure)가 제안된 바 있다[8].

지금까지 여러 가지 희박 데이터 처리 기법들이 제안된 바 있으나, 대부분 데이터가 특정 희박성 형태를 띤다고 가정하고 있다. 따라서 복합 희박성 형태의 실세계 데이터를 다루기 위한 일반화된 희박성 제어 기법에 대해서는 더 많은 연구가 필요하다.

3. OLAP 시스템의 희박 데이터 패턴 분류 및 희박 데이터 생성기

본 논문에서는 OLAP 응용 시스템에서의 희박 데이터의 패턴에 대해 분석하고 이를 기반으로 사용자의 정의에 따라 희박 데이터를 생성하는 희박 데이터 생성기를 설계 및 구현하였다. 또한 실험 희박 데이터를 생성하여 이를 기반으로 다차원 모델을 설계하고 두 가지 상용 OLAP 시스템에 적용하여 각 시스템의 성능을 비교 분석하였다. 전체적인 시스템 구성도는 다음과 같다.

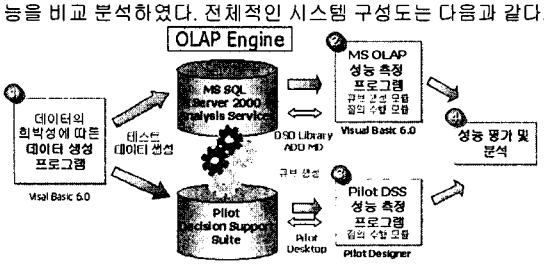


그림 1 시스템 전체 구성도

3.1 OLAP 시스템의 희박 데이터 패턴 정의

희박성(Sparsity)이란 다차원 모델의 데이터 저장을 위한 기본 공간인 셀에 해당되는 값이 없이 비어있는 경우를 말한다[9]. 즉 희박 셀은 관계형 데이터베이스에서 특정 속성의 값이 비어있는 Null 값과 유사하다고 할 수 있다. 다차원 모델에서 희박성은 특정 차원에 독립적인 것이 아니라 두 개 이상의 차원이 서로 연계되어 희박성이 나타나게 된다. 예를 들어 그림 2(a)와 같이 제품, 매장으로 이루어진 2차원 판매 데이터의 경우 판매된 제품과 매장 항목들 사이에 특정 상관 관계가 없다면, 그림 2(b)와 같이 희박 셀들이 흩어져 있는 무작위(Random) 패턴을 나타내게 된다. 또한 특정 일자에 휴무로 전 매장의 판매 데이터가 전혀 없거나, 특정 매장에서 특정 제품군을 판매하지 않은 경우 그림 2(c)과 같은 줄(Stripe) 형태의 희박성 패턴이 나타나게 된다. 마지막으로 그림 2(d)의 클러스터(Cluster) 패턴은 특정 매장의 매출 특성상 해당 일자에 특정 제품들만이 판매 된 경우 발생할 수 있다. 즉 이와 같이 2차원 모델에서 나타날 수 있는 희박 데이터 패턴은 무작위, 줄, 클러스터의 세 가지 패턴이라 분석할 수 있다.

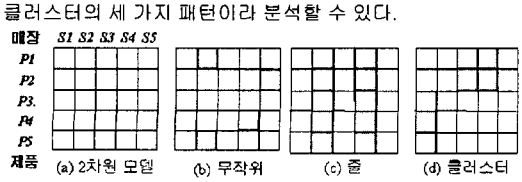


그림 1 2차원 데이터의 희박성 패턴

2차원 모델에 기간 차원을 추가한 3차원 데이터 모델을 가정해 보자. 각 차원들 모두 서로 희박한 관계를 가진다고 할 때, 무작위, 줄, 클러스터 희박성 패턴은 2차원 모델과 유사한 조건에서 발생한다. 3차원 모델의 경우는 이 패턴들 외에 슬라이스(Slice) 패턴이 발생할 수 있다. 예를 들어, 전 기간에 걸쳐 전 매장에서 특정 제품의 판매가 지속된 경우, 그림 3(e)와 같이 슬라이스 형태의 패턴이 발생한다. 즉, 3차원 모델의 희박성 패턴은 그림 3과 같이 무작위, 줄, 클러스터, 슬라이스의 네 가지로 구분할 수 있다.

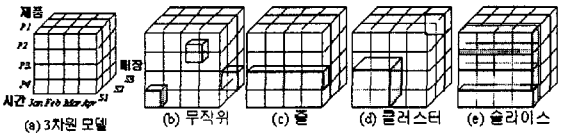


그림 3 3차원 데이터의 희박성 패턴

2, 3차원 모델에 관한 희박성 패턴은 크게 위와 같이 분류할 수 있는데, 이러한 패턴들은 서로 완전히 독립적인 것이 아니기 때문에, 물리적인 데이터 저장 시 차원 항목들의 위치를 조정하여 다른 희박성 형태로 전환할 수 있다. 이러한 특성은 기존에 제안된 희박성 처리 기법을 또 다른 희박성 형태의 처리에 적용시킬 수 있다는 점에서 중요하다.

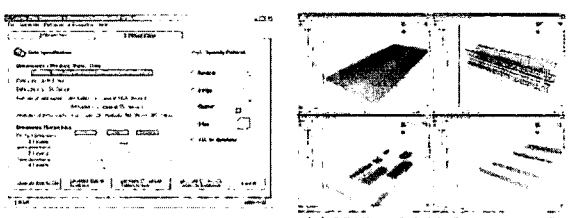
3.2 희박 데이터 생성기

본 논문에서 설계 및 구현한 희박 데이터 생성기는 사용자가 지정한 희박성 패턴에 따라 텍스트 파일이나 MS Access 포맷으로 차원 및 사실 테이블 데이터를 생성한다. 본 논문에서는 구현된 프로그램을 통해 실험 데이터 생성하고 이를 기반으로 OLAP 시스템의 성능을 평가하였다. 이를 위하여 다음과 같이 다차원 모델을 설정하고 이에 따라 실험 데이터를 생성하여 OLAP 시스템에 적재 하였다. 구현된 데이터 생성기는 그림 4(a)와 같다.

OLAP 데이터는 일반적으로 5차원을 넘는 다차원 모델이 대부분이지만 본 논문에서는 앞서 분석된 2, 3차원 모델에 대해서 다음과 같이 다차원 모델을 설계하였다. 2차원 모델은 매장, 기간의 두 차원으로 이루어지고, 3차원 모델은 제품 차원을 더한 세 개 차원으로 이루어진다. 2, 3차원 모델 모두 사실 테이블의 변수 항목으로 판매량과 판매액을 설정하였다. 설정된 다차원 모델의 계층 구조 및 상세 내용은 표1과 같다

표 1 2, 3차원 모델의 상세 정보

차원	2차원			3차원	
각 차원 항목 수	매장차원	9600 (S01 – S9600)		제품차원	120 (P01 – P120)
	기간차원	730 (1999, 2000년)		매장차원	500 (S01 – S500)
				기간차원	365 (1999년)
차원의 계층구조	매장차원	2계층	Store -> Retailer	제품차원	3계층 : Division -> Group -> Product Code
	기간차원	4계층	Year -> Quarter -> Month -> Date	매장차원	2계층 Store -> Retailer
				기간차원	4계층 : Year -> Quarter -> Month -> Date
튜플 수	7008000 개 (9600 × 730)			21900000 개 (120 × 500 × 365)	
희박성	95% (유효 셀수 : 360400 개)			95% (유효 셀 수 : 1096000 개)	



(a) 데이터 생성기 인터페이스 (b) DBMiner 2.0 3D Cube Explore로 가시화한 테스트 데이터

그림 4 구현된 데이터 생성기와 생성된 실험 데이터

생성된 데이터의 희박성 패턴을 확인하기 위해 DBMiner의 3D Cube Explore를 이용하여 데이터 가시화하여 보았다(그림 4(b)). 이 툴은 다차원 데이터를 사용자가 알아보기 쉽도록 3차원 큐브 모양으로 보여준다. 앞서 정의한 무작위, 줄, 클러스터, 슬라이스 희박성 패턴에 맞도록 데이터가 생성된 것을 볼 수 있다.

4. 성능 평가

본 논문에서는 앞서 정의한 생성한 희박 데이터를 기반으로 MS AS와 Pilot DSS의 두 가지 OLAP 시스템에 대해 성능 평가를 수행하였다. 성능평가를 위해 OLAP 응용에서 가장 일반적으로 사용되는 다차원 질의들을 기반으로 질의 모델을 설계하였다. OLAP 다차원 질의의 기본은 사용자가 큐브의 어떤 부분을 볼 것인지 정의하는 것으로[9] 다차원에 걸친 데이터를 요약, 통합 정리(consolidate)하거나, 사용자가 원하는 특정 뷰를 보여주고 계산식을 적용하는 작업들로 이루어진다[3]. 성능 평가는 OLAP 응용에서 가장 일반적으로 사용되는 Exact Match, Range, Slice, Dice, Pivot, Drill Down , Roll Up의 7가지 질의를 기반으로 수행하였다

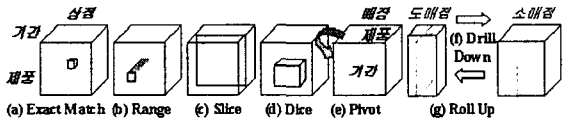


그림5 OLAP의 일반적인 7가지 질의

본 논문에서는 MS AS의 성능 평가 프로그램을 구현하기 위해 DSO(Decision Support Object) 라이브러리를 이용해 다차원 모델을 설정하고, ADOMD (ActiveX Data Objects Multidimensional) API와 MDX(Multidimensional Expression)를 이용하여 다차원 큐브에 대한 질의를 수행하였다. Pilot DSS의 성능 측정 프로그램은 Pilot DSS의 Desktop과 Designer라는 객체지향 개발 툴을 이용하여 다차원 모델을 구성하고, 질의 수행 부분을 구현하였다.

성능 평가를 위해 각 질의에 대한 응답 시간을 ms 단위로 측정하였으며, 각 질의를 5번 반복 실행하여 평균 성능을 관찰하였다. 그림 6은 MS AS 시스템에 대해 수행한 Exact Match, Drill Down, Slice 질의에 대한 성능 결과를 보여준다. 실제로는 모든 질의에 대해 성능을 평가하였으나, 나머지 네 가지 질의의 경우 희박성 패턴 별 응답 성능이 평균 질의 응답 시간의 20% 이내로 거의 유사한 성능을 보여 희박성 패턴에 따라 성능에 유의할만한 차이가 있다고 판단하기는 어려웠다. 그림에서 볼 수 있듯이 Exact Match 질의의 경우 클러스터 패턴에서 가장 우수한 성능을 보이며 Drill Down 질의의 경우 3차원 무작위 패턴에서 성능이 저하됨을 알 수 있다. 또한 Y축의 수행 시간을 살펴 보면 알 수 있듯이 Slice 질의의 경우 다른 질의에 비해 100배 정도 느린 응답 성능을 보였다.

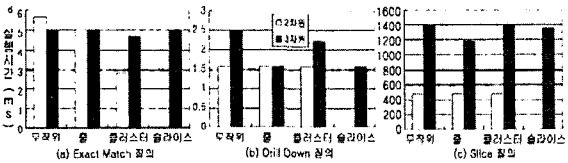


그림 6 MS AS의 성능 평가 결과

그림 는 Pilot DSS의 질의 수행 결과를 보여준다. Pilot DSS의 경우 각 성능이 데이터의 희박성 패턴에 상관없이 거의 유사하였는데 이는 Pilot 시스템이 데이터의 희박성 특징과 무관하게 균일한 성능을 제공한다는 것을 보여준다. 그림 7(a)의 Drill Down과 Roll Up의 경우 2차원 클러스터 희박성 패턴에서 성능이 가장 저하됨을 알 수 있다. Slice 질의의 경우 3차원 무작위 패턴에서 최고 성능을, 3차원 줄 패턴에서 최저 성능을 보였다

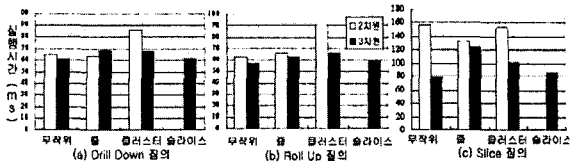


그림 7 Pilot DSS의 성능 평가 결과

이러한 성능 차이는 비록 두 시스템 모두 MOLAP 구조를 사용하고 있으나 내부적으로 상이한 희박성 처리 기법을 사용한다는 점에서 기인한다고 할 수 있다. MS AS의 경우 불록 멀티큐브 형태로 데이터를 저장하는 반면 Pilot 시스템은 시리즈 멀티큐브 형태로 저장한다. 블럭 멀티큐브 방식은 동일한 차원을 공유하는 변수 항목들의 집합에 대해 각각 차원을 설정해 주는 반면 시리즈 멀티큐브의 경우 모든 변수 항목들을 별개의 큐브로 다룬다. 즉 Pilot DSS는 각 변수 항목들을 별개의 큐브로 저장하여 주어진 질의에 대하여 다수개의 큐브를 조회하는 방식을 취하기 때문에 위와 같은 응답 성능의 차이를 보인다고 생각할 수 있다.

5. 결론

OLAP 응용 시스템은 사용자에게 빠르고 인터액티브한 질의 결과를 제공하기 위해서 다량의 집계 연산에 대한 사전 연산을 수행 한다. 그러나 실제 OLAP 환경에서 다루어야 하는 데이터는 매우 희박한 특성을 가지기 때문에 이러한 연산 작업 시 데이터 폭발현상을 일으키게 된다. 본 논문에서는 2, 3차원 OLAP 데이터의 희박성 패턴에 대해 정의하고, 이를 기반으로 희박 데이터를 생성하는 데이터 생성기를 설계 및 구현하였다. 또한 생성한 실험 데이터를 이용하여 다차원 모델을 구성하고 MS SQL Server Analysis Service와 Pilot DSS의 두 시스템에 대해 성능 평가를 수행하였다.

본 논문의 결과를 기반으로 Oracle과 EssBase와 같은 다양한 OLAP 제품의 성능에 대한 비교 평가나 일반화 된 희박 형태 처리 기법에 대한 연구가 가능할 것이다. 또한 복합 희박 패턴 데이터 생성기, 희박 데이터의 효율적인 다차원 저장 및 처리 기법, 희박성 패턴의 자동 분석 등에 대한 지속적인 연구가 필요할 것이다.

6. 참고문헌

[1] MicroStrategy, White Paper, The case for relational OLAP, 1995
[2] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh, Data cube: A relational aggregation operator generalizing group-by, cross-tabs and sub-totals". In Proc. of the 12th Int'l Conference on Data Engineering, pages 152-159, 1996
[3] White Paper, OLAP Report, Data Explosion http://www.olapreport.com/DatabaseExplosion.htm
[4] Y. Zhao, P.M. Deshpande, and J.F. Naughton, An Array-Based Algorithm for Simultaneous Multidimensional Aggregates, In Proc. Of ACM SIGMOD '97, pages 159-170, Tucson, 1997
[5] Erik Thomsen, OLAP Solutions, Building Multidimensional Systems, 1997
[6] Robert J. Earle. Arbor Software corporation, Method and Apparatus for Storing and Retrieving Multi-dimensional Data in Computer Memory", U.S. patent #5359724, Oct. 1994
[7] Oracle Corp., Sparsity Management System for Multi-dimensional Databases, U.S. patent #5943677, Aug, 1999
[8] Sanjay Goil and Alok Choudhary, Sparse Data Storage of Multi-Dimensional Data for OLAP and Data Mining, Technical Report CPDC-TR-9801-005, Center for Parallel and Distributed Computing, Northwestern University, 1997
[9] 조재희, 박성진, 시그마 컨설팅, "데이터 웨어하우스의 효과적 활용 기법 OLAP 테크놀러지", 1999