웹 로그 데이터의 OLAP 연산을 위한 희박성 분석

김지현, 용환승
이화여자대학교, 컴퓨터 학과
{pshike, hsyong}@ewha.ac.kr

Web Log Data Sparsity Analysis for OLAP

Ji-Hyun Kim, Hwan-Seung Yong
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

하루에도 수십 수백 메가 바이트까지 증가하는 웹 로그 데이터를 이용하여 실시간에 다차원분석을 가능 하게 하기 위해서는 OLAP의 적용이 필요하다. 하지만 OLAP을 적용하는데 있어서 빠른 응답시간을 얻기 위해 사전처리(Precomputation)를 수행 할 시 심각한 데이터의 희박성으로 인해 데이터 폭발 현상이 발생된다. 본 논문에서는 실제 웹 로그 데이터를 사용하여 OLAP적용 시 희박성을 일으키는 원인들을 밝히고, 2,3 차원에서의 희박성 형태를 분석함으로써 웹 로그 데이터의 희박성 처리 방식 및 성능평가에 기반이 되게 한다.

1. 서 론

최근의 경쟁적 비즈니스 환경에서 CRM(Customer Relationship Management)를 위한 IT(Information Technology)기술들이 급속도로 발전하고 있다. 이러한 CRM을 위해 최근 개발되고 있는 분야는 간단하게는 통계적 수단을 통해 웹 사이트를 분석하는 방법과 좀더 정교하게는 다차원 분석이 가능한 OLAP(On-Line Analysis Processing)기술을 적용한 방법, 그리고 Data Mining을 이용한 방법들이 있다[1].이러한 분석을 위해 고객, 구매, 공급업체, 경쟁사 등의 데이터들이 사용되고 있다. 특히 고객 데이터 중에서 각 기업이나 단체의 웹 서버에 저장되어 있는 웹 로그 데이터는 가장 쉽게 얻을 수 있고, 유용하게 사용될 수 있다.

매일 수십 수백 메가 바이트까지 증가 하는 이러한 데이터를 실시간에 다차원 분석을 효율적으로 하기 위해서는 OLAP을 사용해야 한다. 하지만 OLAP을 적용하는데 있어, 웹 로그 데이터 자체가 가지고 있는 특성에 의해 희박성이 발생되고, 사전집계 연산 수행 시 데이터의 폭발(Data Explosion)현상이 일어난다.

본 논문에서는 실제 데이터를 사용하여 Microsoft SQL Server 2000 Analysis Services와 DBMiner의 3D Cube Explore 로 OLAP적용 시 발생 되는 희박성 원인과 희박성 형태에 대해 분석 및 도식화 한다.

2. 관련 연구 및 연구 동향

OLAP제품은 저장 구조에 따라 크게 ROLAP(Relational OLAP), MOLAP(Multidimensional OLAP), HOLAP(Hybrid OLAP)으로 나누어 진다. ROLAP의 경우 데이터를 테이블 형태로 저장함으로 큐브 생성 및 연산 시간이 느리고, 데이터 저장에 있어 오버헤드가 발생되게 된다. MOLAP은 데이터 저장 시 다차원 배열형태로 저장하여 연산 시간에 있어서 좋은 성능을 얻을 수 있다. 마지막으로 HOLAP의 경우는 ROLAP과 MOLAP 두 가지 방식을 혼용하여 사용하고 있다[2]. MOLAP 방식의 경우 ROLAP과는 달리 데이터의 조작이나 분석 시 빠른 성능을 보이나, 다차원 모델 형성 시 모든 차원의 멤버들은 각 멤버들의 가능한 위치가 정해져 있게 되어 무수한 무효 셀들이 생성되게 된다.

따라서 대부분의 데이터가 희박한 OLAP어플리케이션 데이터와 사전연산 결과들을 저장하기 위해서는 데이터 폭발 현상이 발생되게 된다[3,4]. 이러한 현상은 데이터 모델의 차원이 증가할수록 더욱 심해지는데, 대부분 7,8차원 이상의 데이터 모델을 필요로 하고 데이터가 매우 희박한 OLAP어플리케이션이 대부분이라는 점을 고려할 때, 이러한 현상은 일반적인 것이라 할 수 있다.

이러한 희박성을 처리하기 위한 기법들로 Oracle Express에서 사용하고 있는 복합 차원기법[5], Hyperion EssBase의 Sparse-Dense Split 기법[6], Chunk 기법[7]들이 연구 되어 왔다.

3. 웹 로그 데이터의 OLAP 연산을 위한 희박성 분석

본 장에서는 분석을 위한 원시 로그 데이터에 대한 설명과, 희박성이 발생되는 원인 및 그 형태에 대해 살펴본다.

3.1 웹 로그 데이터의 희박성의 원인

웹 로그 데이터를 사용한 다차원 모델에서 희박성은 여러 가지 이유로 만들어 질 수 있다.

첫째는 각 차원들이 가지고 있는 데이터의 고유한 특성으로 인해 밀집 항목과 희박 항목으로 분리 됨으로써 발생된다. 예를 들어 시간차원의 경우, 쇼핑장 사이트는 시즌별로 늦가을과 겨울철에 성수기를 이루어 방문하는 사람의 수가 많아지고 나머지 시즌에는 방문자의 수가 희박하게 나타나게 된다. 둘째로, 데이터베이스 설계에 있어서 로그 데이터 형식의 변화로 새로운 값이 추가됨으로써 발생하게 된다. 세 번째로는 분석하고자 하는 웹 서버의 장애와 같은, 시스템상의 오류로 인해 희박성이 발생 할 수 있다. 마지막으로는 특정 기간 동안만 존재하고 없어져 버리는 데이터들에 의해 발생된다. 이 중 희박성을 일으키는 주 원인이 되는 첫번째 경우를 실제 세가지 웹 로그 데이터 사용하여 알아본다.

다음 절에서는 분석에 사용된 웹 로그 데이터에 대한 설명과 데이터 모델 구조에 대해 설명한다.

3.1.1 실제 웹 로그 데이터와 데이터 모델

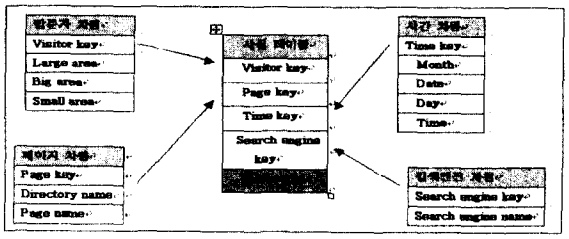
분석에 사용된 웹 로그 데이터의 형식은 W3C 확장 형식을 기본으로 하며, 일주일 분량의 웹 로그 데이터를 가지고 사전처리-데이터 정제 과정 및 사용자 확인 모듈, 세션 확인 모듈 - 를 적용한 후 OLAP 서비스에 적용한다[8]. 데이터 모델은 4개의 차원 -방문자, 페이지, 검색엔진, 시간 -을 가지며, 방문횟수를 측정값으로 한다.

첫번째 사례에 사용된 사이트는 세계각국의 스포츠 소식을 전하는 사이트로 회원가입을 통해 주요정보를 서비스 받는다. 분석에서 사용한 이 사이트의 웹 로그 데이터는 표 1로 나타내었다.

[표 1 사례 1의 웹 로그 데이터의 설명]

출 처	스포츠 정보 회사	
기 간	2000년 10월 28일 ~ 11월 4일	
데이터 용량	실제 로그 데이터 :	2513975records (163M)
	사전 처리된 로그 데이터 :	35635records (1.83M)

사전처리가 끝난 로그 데이터와 고객 데이터 정보와 결합하여 만든 데이터 모델은 그림 1과 같다.



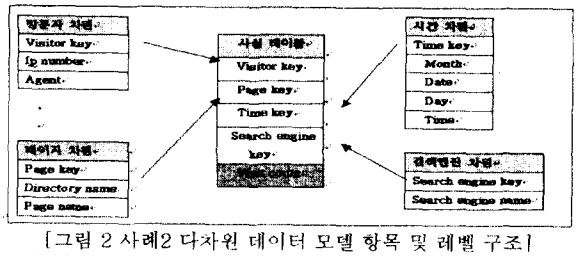
[그림 1 사례1 다차원 데이터 모델 항목 및 레벨 구조]

두 번째는 본 연구실의 웹 서버 로그 데이터를 사용하였다. 회원 가입을 요구하지 않으며, 15일 간의 데이터를 사용하였다. 로그 데이터의 설명은 표2에 제시하였다.

[표2 사례 2의 웹 로그 데이터의 설명]

출 처	이화여자 대학교 데이터 베이스 연구실	
기 간	2001년 03월 01일 ~ 03월 15일	
데이터 용량	실제 로그 데이터 :	198591records (120M)
	사전 처리된 로그 데이터 :	72309records (25.1M)

본 사이트는 고객 데이터가 존재 하지 않으므로 순수 웹 로그 데이터를 사용하여 그림 2 데이터 모델로 구성하였다.



[그림 2 사례2 다차원 데이터 모델 항목 및 레벨 구조]

세 번째 사례에 사용된 웹사이트는 자연 다큐멘터리 인터넷 방송국 사이트로서 회원 가입을 통해 사이트의 정보들을 볼 수 있다. 로그 데이터에 대한 설명은 표 3에 제시하였다. 여기서 사용된 데이터 모델은 그림 1의 것과 같은 구조를 갖는다.

3.1.2 각 차원별 데이터를 이용한 희박성 원인 분석

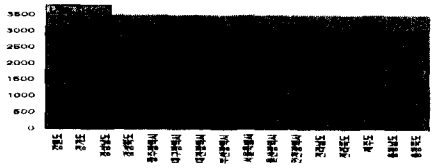
앞에서 설명한 것처럼 웹 로그 데이터에서 희박성이 발생하는 주 원인은 각 차원의 데이터 항목이 밀집 항목과 희박 항목으로 나타나고, 희박 항목이 밀집 항목에 비해 넓은 분포로 존재할 경우에 나타나게 된다. 여기서는 실제 로그 데이터를 사용해 각 차

[표3 사례 3의 웹 로그 데이터의 설명]

출 처	자연 다큐멘터리 인터넷 방송 사이트	
기 간	2001년 03월 28일 ~ 04월 03일	
데이터 용량	실제 로그 데이터 :	21575records (13.1M)
	사전 처리된 로그 데이터 :	19739 records (4.72M)

원별로 밀집 항목과 희박 항목이 나타나게 되는 웹 로그 데이터의 고유한 특성에 대해 고찰해 본다.

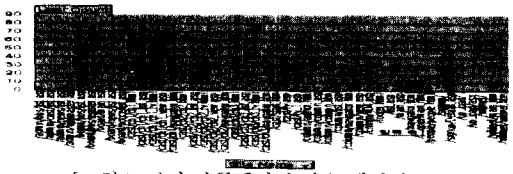
우선 첫번째로 방문자 차원을 살펴 보면, 스포츠 정보 회사의 경우 계층 구조 중 대권역 수준에서 그림 3로 보여 주었다.



[그림 3 스포츠 정보 사이트의 방문자 차원]

방문자 차원의 경우 방문 정도에 따라 자주 방문하는 사람과 희박하게 방문하는 사람으로 나눠 볼 수 있다. 이러한 분류가 발생되는 원인들은 위의 그림처럼 주요 도시에서는 접속 횟수가 많고 나머지 지역에서는 넓게 분산되어 있는 지역적 특징에 의해 발생 될 수도 있고, 사이트의 정보에 관심이 많아 회원 가입을 한 회원과 비 회원의 여부, 북 마킹의 유무, 이벤트나 그룹 활동 등의 참여 여부, 그리고 개인의 취향 및 관심 분야, 직업 ,성별 등의 요소들로 인해 발생한다.

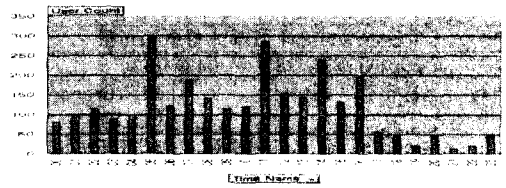
두 번째는 페이지 차원을 기준으로, 각 페이지에 방문한 방문자의 횟수를 알아봄으로써 인기 항목 와 비인기 항목을 구별해 볼 수 있다. 그림 4는 자연 다큐멘터리 인터넷방송 사이트의 페이지 차원별 방문횟수를 보여준 그림이다.



[그림4 자연 다큐멘터리 방송 페이지 차원]

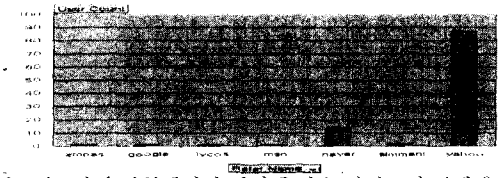
사이트의 페이지를 방문할 수 있는 방식은 크게 브라우저 내의 북 마크를 이용한 히스토리컬 기법과 검색 엔진 이나 참조 사이트들을 사용하는 방식 그리고 직접 방문으로 구분 할 수 있다. 히스토리컬 기법의 경우 대부분 흥미 페이지나 사이트 주된 페이지에 바로 방문하게 됨으로, 주제가 되는 페이지들의 방문횟수가 많아진다. 직접 방문의 경우 시작할 때의 초기 페이지와 회원 가입페이지, 로그인 페이지등과 주제가 되는 페이지에 방문 횟수가 많아진다. 따라서 나머지 페이지에서는 접속 분포가 넓게 분산되어 희박한 항목이 발생됨을 볼 수 있다. 하지만 주제 페이지의 수가 안내를 위한 보조 페이지 수 보다 많을 경우 보조 페이지의 접속 횟수가 많이 질 수 있다. 세 번째로는 시간 차원의 각 시간대별 사용자들의 방문 횟수를 사용하여 바쁜 시간대와 한가한 시간대를 알아본다. 그림 5은 데이터 베이스 연구실 사이트의 시간대별 방문횟수를 그래프화 한 것이다. 이 사이트에서 각 시간대별 접속자수의 접속현황을 보면 저녁시간을 제외한 나머지 시간대가 바쁜 시간대 임을 볼 수 있다. 사이트의 성격이 시간 차원에 매우 민감한 스키장이나, 수영장 사이트와 같은 경우 접속자 수가 확연하게 시즌별 월별 요일별 희박항목이 현저하게 나타나게 된다.

마지막으로 검색엔진 차원의 경우 각 검색엔진에 의해 자신의 사이트에 오게 된 방문자 수를 알아봄으로써 인기 검색



[그림5 데이터베이스 연구실 사이트의 시간차원]

엔진과 비 인기 검색 엔진을 알 수 있다. 그림 6은 자연 다큐멘터리 방송사이트의 검색 엔진의 참조 횟수를 그래프화 한 것 이다.



[그림6 자연 다큐멘터리 인터넷 방송 사이트의 검색엔진 차원]

검색 엔진차원이 밀집 항목과 희박 항목으로 뚜렷한 구분이 일어나는 원인은 각 검색엔진의 검색어를 다양하고 적절하게 제시 함으로써 사이트를 찾는 사람들이 여러 루트로 본 사이트를 발견할 수 있게 도와주기 때문이다.

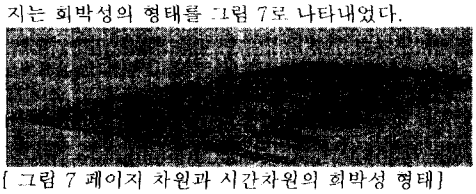
위 4가지 차원들에 대해 각 차원마다 희박 항목과 밀집 항목이 존재 하고, 이를 발생시키는 로그 데이터의 특징들에 대해 알아 보았다. 다음 장에서는 이러한 차원들에 의해 2,3 차원에서 만들어 지는 희박성 형태에 대해 알아본다

3.2 웹 로그 데이터의 희박성 형태

본 장에서는 실제 웹 로그데이터를 OLAP적용 시 만들어지는 희박성 형태에 대해 알아본다. 다차원 모델에서 희박성 형태는 각 차원이 희박 항목과 밀집 항목이 존재하고 이들 사이의 관계에 의해 만들어 진다. 다음 절에서는 2, 3차원 데이터 모델 형태에서 나타날 수 있는 희박성 형태에 대해 알아 본다.

3.2.1 2차원 모델에서의 희박성 형태

앞에서 제시한 로그 데이터들을 사용하여 OLAP적용을 위한 다차원 모델 생성시, 2차원 모델에서 만들어 지는 희박성 형태에 대해 알아 보도록 하겠다. 페이지 차원과 시간 차원을 이용하여 만들어 지는 희박성의 형태를 그림 7로 나타내었다.



[그림 7 페이지 차원과 시간차원의 희박성 형태]

위의 그림에서 볼 수 있듯이 2차원 모델에서는 GRID모양의 희박성 형태를 볼 수 있다. 위의 경우 정보가 많은 페이지의 경우 바쁜 시간대와 관계없이, 북 마킹이나 검색 엔진을 통해 해당 페이지로 바로 오는 사람들과 직접 방문한 사람들에 의해 방문 횟수가 많고, 사람들이 많이 들어 오는 시간대에는 정보가 많은 페이지뿐 아니라 그 외의 페이지에도 방문횟수가 증가 한다. 즉 이는 각 차원이 희박 항목과 밀집 항목의 구분이 뚜렷하고, 이들 차원 항목들 사이의 상호 의존성이 약할 때 발생하게 된다. 하지만 각 차원 항목의 나열 순서가 바뀌게 되면 CLUSTER 형태를 나타내게 된다.

3.2.2 3차원 모델에서의 희박성 형태

이번 절에서는 3차원 내에서 발생되는 희박성의 형태에 대해 알아본다. 그림 8은 데이터 베이스 연구실 사이트의 페이지 차원과 시간차원 그리고 검색엔진차원 사이의 희박성 형태를 보여주는 그림이다. 3차원 데이터 모델에서도 또한 GRID형태와

[그림 8 페이지 검색엔진 시간차원사이의 희박성 패턴]

CLUSTER형태의 희박성 형태가 발견됨을 볼 수 있다.

사용자 차원이 들어간 3차원 모델의 경우 사용자 항목이 많은 관계로 도식화 하기 어려우나, 사용자 항목이 들어가는 모델의 형태 또한 사용자 차원 자체가 밀집 항목과 희박항목으로 구분이 뚜렷하게 나타남으로 위와 같은 형태의 희박형태가 나타나게 됨을 알 수 있다. 하지만 사용자 항목이 들어가는 모델의 경우 다른 차원의 항목들과 달리 가변성을 가장 많이 가지고 있어 RANDOM형태가 넓은 영역으로 분포된다.

4.결론 및 향후 과제

본 논문에서는 웹 로그 데이터의 희박성 분석을 하기위해 사전처리 과정을 거친 데이터를 사용하여 MS-SQL 2000 server의 analysis service와 DBMiner의 3D Cube Explore를 통하여 웹 로그 데이터의 희박성이 발생되는 원인 중 주 요인인 각 차원별로 데이터 항목을 밀집항목과 희박항목으로 구분하는 웹 로그 데이터 자체의 고유의 특성에 대해 알아보았다. 그리고 2,3차원의 데이터 모델을 설계하여 희박성 형태를 발견 할 수 있었다.

본 논문에서 파악한 웹 로그 데이터의 희박성 형태를 기반으로 웹 로그 데이터를 다차원 분석하는데 있어 빠르고 효율적인 저장 기법 및 희박성 처리 색인 기법을 개발하고, 웹 로그 데이터 분석을 위한 제품들의 성능평가에 사용하는 것을 향후 과제로 제시 한다.

5. 참고 문헌

[1] MaxScan Corp, White Paper. *An Accelerator for Click Stream Data Analysis Applications*

[2] Pilot Software, White Paper: *An Introduction to OLAP: Multidimensional Terminology and Technology*

[3] 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

[4] White paper :
http://www.olapreport.com/DatabaseExplosion.htm

[5] Oracle Corp, *Sparsity Management System for Multidimension al Databases*, U.S. patent #5943677, Aug, 1999

[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] Y.Zhao, P.M. Deshpande, and J.F.Naughton, *An Array-Based Algorithm for Simultaneous Multidimensional Aggregates*, In Proc.of ACM SIGMOD, pp 159-170, 1997

[8] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava: *Data Preparation for mining world wide web browsing patterns*, the Journal of Knowledge and Information System, Vol. 1, No. 1, 1999