The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 457-463

On-Line Analytical Processing and Research Problems for Statisticians

Jeong Yong Ahn[1], Kyung Soo Han[2]

Abstract

Recently, statistical analysis tools have been changed to the applications on the World Wide Web that access data stored in databases. On-line analytical processing(OLAP) is a class of technologies that give users statistical information with multidimensional views of data in databases. In this paper, we introduce the concept and requisites of OLAP system, and we propose some research issues.

Keywords : OLAP, Statistics, Research Issues

1. 머리말

최근 정보 기술의 급속한 발달은 교육 및 연구 분야를 포함한 우리 생활의 전반적인 환경에 많은 변화를 가져오고 있다. 교육에 있어 각종 소프트웨어의 활용은 이미 일반화되어 가고 있는 현상이며, 디지털(digital) 문명의 산물인 전자 교재(electronic text)와 전자 잡지(electronic journal), 전자 결재, 전자 상거래, 디지털 전화 등은 이제 더 이상 새로운 것이 아니다. 데이터를 분석하는 응용 프로그램들 역시 점차 네트워크 환경에서 이용될 수 있도록 변모하고 있다.

컴퓨터에 기반을 둔 이러한 디지털 기술의 발달은 인간의 활동무대를 가상 공간으로 확장시키고 있다. World Wide Web(이하 Web)과 같은 가상 공간을 통하여 엄청난 양의 데이터가 시시각각 데이터베이스에 누적되어 가고 있으며, 이러한 현상은 통계학 분야에서 취급하는 데이터 형태에도 많은 영향을 주고 있다.

일반적으로 과거의 통계학 분야에서 다루어지던 데이터는 행렬 형태로 표현되는 비교적 소규모의 데이터가 주를 이루어 왔으며, 기존의 데이터 분석 방법들은 소규모의 데이터에 적합하게 연구되어 왔다. 탐색적 자료 분석(EDA)은 그 대표적인 예이며, 이러한 데이터는 거의 대부분이 저차원적으로 구성되어 있고, 검정(testing)에 필요한 속성인 동질성(homogeneity)을 가지고 있는 특징이 있다.

그러나 현실 세계에서 나타나는 데이터의 대부분은 이와 같은 속성을 갖지는 않으며, 특히 최근 과학과 기술의 발달은 데이터 형태를 점차 대규모화 시켜가고 있다. 이들 데이터는 실험 데이터와

1) Assistant Professor, Division of Computer Science and Information Communications, Seonam University, Chonbuk, 590-170, Korea
E-mail: jyahn@tiger.seonam.ac.kr, jyahn@stat.chonbuk.ac.kr
2) Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University, Chonbuk, 561-756, Korea
E-mail: kshan@stat.chonbuk.ac.kr

는 달리 고차원적이며 동질성을 갖지 않는 경우가 대부분이다.

이러한 현상은 데이터 분석 및 활용 방법에 있어서 새로운 방식을 요구하고 있다. 수백 개의 변수(variables)와 수십억 개의 관측치(observations)로 이루어진 데이터를 분석하고 활용하는 것은 간단한 일이 아니며, Huber(1994)에서 지적하듯이 대량화되어 가고 있는 데이터를 기존의 방법으로 취급하는 것은 큰 의미가 없기 때문이다. Famili 등(1997)은 현실 세계의 데이터가 갖고 있는 문제점들을 지적하고 데이터 분석을 시작하기 전에 해결해야만 하는 과제들을 제시하고 있으며, "데이터의 양이 열 배의 속도로 증가되면 그것을 분석하는 방법을 완전히 다시 생각해야 한다"는 Friedman(1997)의 지적은 통계학자들에게 시사하는 바가 크다.

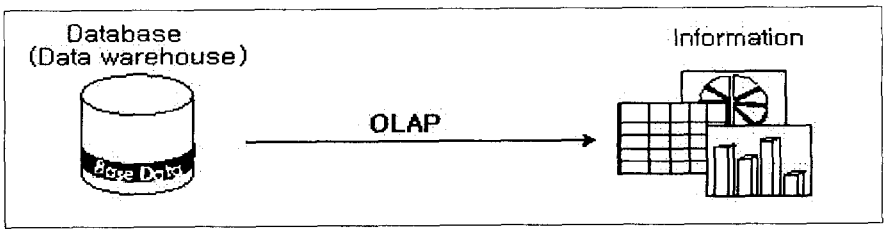
또한 이러한 데이터 형태의 변화는 데이터를 관리하는 방법에도 변화를 요구하고 있다. 데이터 관리 분야는 손건태, 허명회(1999)에서 지적하는 바와 같이 통계학자들에게 많은 관심을 받지 못한 분야이다. 그러나 대규모 데이터를 효율적으로 관리하고, 온라인 상을 통하여 실시간으로 이용하기 위해서는 데이터 관리에 대한 연구가 필수적이라 할 수 있다.

현대 사회의 정보 기술의 발달은 이러한 요구들을 해결하기 위한 적절한 환경을 조성해 주고 있다. 데이터 수집 및 관리, 컴퓨팅 능력 등으로 과거에 겪어야 했던 어려움의 많은 부분이 해결되어 가고 있다. 컴퓨터 네트워크와 데이터베이스 기술의 발달은 대용량의 데이터를 손쉽게 이용할 수 있는 환경을 제공하여 주고 있으며, 통계학자들이 그 동안 많은 관심을 기울이지 않았던 데이터 관리 및 대규모 데이터의 활용 분야에 대한 연구에 쉽게 접근할 수 있도록 도와주고 있다.

본 연구에서는 대규모화되어 가고 있는 데이터를 효과적으로 활용할 수 있는 방안에 대해 살펴보고자 한다. 그 방법의 일환으로 최근 데이터를 관리하고 활용할 수 있는 기술로 부각되고 있는 On-Line Analytical Processing(이하 OLAP)에 대해 살펴보고, 통계학자들에게 어떤 연구 과제들을 줄 수 있는지에 관해 생각해보고자 한다. 본 연구는 한경수와 안정용(1998, 1999)의 학술 발표 내용에 기초를 두고 있다.

2. OLAP

OLAP이라는 용어는 Codd 등(1993)에 의해 처음 사용되었으며, Lenz와 Shoshani(1997), Shoshani(1997)는 다차원 데이터베이스에 축적된 데이터로부터 통계적인 요약 정보를 제공할 수 있는 기술로 정의하고 있다. Carickhoff(1997)는 다차원으로 이루어진 데이터에 대한 이해뿐만 아니라, 데이터 요약 정보, 시계열 및 추세 분석과 같은 분석적인 내용을 사용자에게 제공하는 기업용 소프트웨어(business software)의 한 범주로 정의하고 있다. 다시 말하면, 데이터베이스로부터 온라인 상에서 정보를 추출하는 기술이 OLAP의 핵심이며 그 근간은 통계적 방법이라 할 수 있다.



<그림 1> 데이터베이스와 OLAP

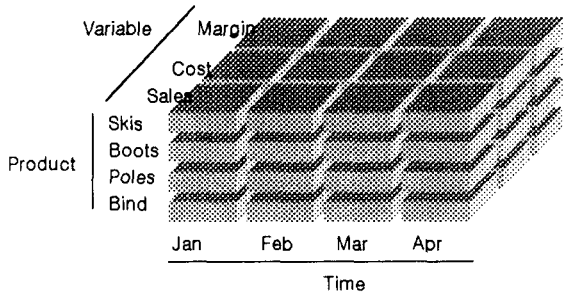
<그림 1>에서 보는 바와 같이 사용자들은 축적되어 있는 데이터로부터 어떤 정보를 요구한다. 요구되는 정보를 추출하기 위해 데이터의 통합과 분석 과정은 필수적이며, 이러한 과정에 있어 데이터베이스의 기능만을 이용하는 것은 한계가 있다. 데이터베이스는 Thomsen(1997), 조재희와 박성진(1999)에서 지적하는 바와 같이 데이터를 종합, 분석, 합병하는 목적으로 설계되지 않아 의사결정에 필요한 분석적이고 다양한 정보를 쉽게 제공하기 어려운 단점을 가지고 있기 때문이다. 이러한 데이터를 현재 범용으로 이용되고 있는 통계 패키지와 같은 프로그램으로 처리하는 것 또한 매우 어렵다. 그 이유는 대부분의 데이터는 매우 큰 용량이며, 온라인 상에서의 처리가 필수적이고, 다차원적인 구조를 다루어야 하기 때문이다.

따라서 OLAP, 데이터 마이닝(data mining)과 같은 기술의 필요성이 부각되고 있으며 이 기술들은 관계형 데이터베이스의 단점을 보완하고, 통계 패키지의 한계를 극복할 수 있는 기술로 인식되고 있다.

2.1 OLAP 데이터 모델

OLAP은 사용자들에게 데이터의 흐름을 다차원적인 구조로 빠르게 보여주면서 몇몇 미리 계산된 값들을 제공함으로서 데이터에 대한 정보를 쉽게 제공해 주는 것을 목적으로 한다.

OLAP 데이터 모델에서 정보는 <그림 2>와 같이 차원(dimension)과 관측된 데이터(measures)로 구성되어 있는 데이터 큐브(data cube)의 형태로 표현할 수 있으며, 차원 항목은 다시 범주형의 속성을 가지는 차원과 실제 데이터 요소로 분석을 요하는 변수 차원 항목으로 분류할 수 있다. 다시 말해, 큐브는 분석이 필요한 데이터를 다차원적으로 조직화해 정리해 놓은 것이라 할 수 있다.



<그림 2> 3차원 큐브

<그림 2>와 같이 모델링된 다차원 데이터는 주로 스타 스키마(star-schema)라는 관계형 데이터베이스 설계 기법을 이용하여 표현한다. 스타 스키마는 다차원 데이터를 사실 테이블(fact table)과 차원 테이블(dimensional table)로 분류하여 표현한다. 사실 테이블은 분석을 요하는 변수 차원의 항목들을 포함하고 있는 테이블이며, 차원 테이블은 사실 테이블의 변수들을 살펴보기 위한 범주형 속성의 계층적(또는 비계층적) 정보를 포함한다. 이러한 스키마가 완성되면 데이터베이스에 테이블과 데이터 큐브를 생성하고 OLAP 서비스를 제공하기 위한 응용 프로그램을 제작한다. Chaudhuri와 Dayal(1997)은 OLAP시스템을 개발할 때 이용할 수 있는 몇몇 도구들을 제시하고 있다.

2.2 OLAP 시스템의 요건

OLAP 시스템은 데이터가 가지고 있는 정보를 직관적이고 빠르게 제공해야 하고, 다양한 관점에서 작성된 레포트를 통하여 데이터를 살필 수 있어야 하며, Gray와 Watson(1996)에서도 제시하듯이 다음과 같은 사항들을 지원할 수 있어야 한다.

- 데이터 요약(통계 분석)과 다차원적인 뷰(view)
- 데이터 조작과 레포팅 기능
- 네트워크 환경에서 다중 사용자 환경
- 클라이언트-서버(client-server) 구조
- 시계열 데이터 분석
- 대용량의 데이터 처리

3. 연구 과제

OLAP 분야는 기업용 응용프로그램(business applications)에 의해 출발되어 발전하였으며, 통계의 다른 분야들과는 달리 통계학자들의 관심을 많이 받지 못해 왔다. 최근 들어 통계학자들이 OLAP을 포함한 데이터 웨어하우징, 다차원 데이터베이스, 데이터 마이닝 등의 분야에 많은 관심을 표명하고 있으나, 아직까지 연구와 교육에서 그 실적은 매우 미미한 실정이다. 이러한 분야들에 대한 연구할 만한 가치가 있는 사항들 몇 가지를 제시해 보면 다음과 같다.

(1) 데이터 분석 기법에 관한 연구

대규모 데이터를 처리하고자 할 때 기존의 분석방법을 적용하는 것은 의미가 없다. 예를 들어, 대규모 데이터에서 검정(testing)의 결과가 무슨 의미가 있겠는가? Hand(1997)에서 주장하듯이 대규모 데이터에 대한 분석은 전통적인 방법 이상의 기술을 요하며, 컴퓨팅 기술 등과 연관되어 개발되어야 한다.

(2) 데이터 시각화(visualization) 기법에 관한 연구

그래프를 이용한 데이터 표현은 오랫동안 이용되어진 데이터 탐색 방법이다. 그러나 대규모의 다차원 데이터를 표현하는 것은 간단한 일이 아니며, 이 분야 역시 애니메이션, 렌더링(rendering) 등 컴퓨터 그래픽 기술과 연관되어 개발되어야 효과적으로 데이터를 표현할 수 있다.

(3) 결측 데이터의 처리를 포함한 데이터 정제(data cleaning) 방법에 관한 연구

결측 데이터 처리와 데이터 정제 방법은 주로 이질적인 데이터를 통합하는 과정에서 이용되는 기법들이며, 데이터 정제는 데이터 마이닝 분야와 밀접한 관계가 있다.

(4) 효율적인 데이터베이스 설계 문제에 관한 연구

데이터베이스를 어떻게 설계하는가에 따라 데이터로부터 정보를 추출하는 효율성이 크

게 좌우될 수 있다. 특히, 대규모 데이터를 처리하고자 할 경우 데이터베이스 설계는 매우 중요한 문제이다. 어떤 질의에 대한 반응 시간(response time)이 매우 오래 걸릴 수 있기 때문이다. 따라서 질의의 효율성과 활용할 수 있는 컴퓨터 기억용량을 고려하여 적절한 설계가 이루어져야 한다.

(5) 희소 행렬 처리(sparse matrix handling) 문제에 관한 연구

Shoshani(1982)에서 제시한 'World trade' 데이터는 널 값(null value)을 갖는 비율이 아주 많다. 이러한 데이터를 효율적으로 저장할 수 있는 기법과 이용 방안에 대한 연구 또한 필요하다

(6) 특정 환경에 맞는(customized) 효율적인 시스템 개발에 관한 연구

통계 패키지와 같은 응용 프로그램들은 데이터를 분석할 수 있는 많은 기법을 포함하고 있다. 그러나 일반 사용자들이 이용하기에는 어려운 점이 많으며, 실제로 특정한 분석 기법만을 필요로 하는 경우가 많다. 또한 기업과 같은 환경에서는 필요한 분석 결과가 정형화되어 있는 경우가 많기 때문에 사용자들이 편리하게 이용할 수 있는 시스템의 개발이 요구된다.

(7) 메타데이터(metadata) 구축 및 활용방안에 관한 연구

정보 기술의 발전은 전 세계 컴퓨터들간의 데이터를 공유할 수 있는 환경을 제공하여 주고 있다. 데이터 공유 문제에 있어 메타 데이터의 역할은 매우 중요하며, 메타데이터의 표준 설정, 구축 및 활용방안 등에 관한 연구가 요구된다.

(8) 데이터베이스에서 표본 추출 기법에 관한 연구

대규모 데이터인 경우 요구된 질의를 수행하고, 그 결과를 반환하는 시간이 매우 많이 소요될 수 있다. 통계적인 관점에서, 필요로 하는 데이터가 근사치로 충분한 경우 데이터베이스로부터 표본을 추출하여 활용하는 것은 매우 좋은 방법이다. 이 방법은 데이터를 검색하는 비용과 통계적 분석에 필요한 계산의 양, 물리적인 검사 비용 등을 줄일 수 있는 이점을 제공한다. 따라서 표본 추출은 데이터베이스 내에 축적된 대 용량 데이터를 분석하기 위한 기본적인 수단이라 할 수 있으며, Olken(1993), Vitter(1987), Chaudhuri 등(1999)에서 표본 추출 알고리즘에 대한 연구가 진행되었다.

현재 이용되고 있는 데이터베이스 관리 시스템에서는 표본 추출에 대한 질의 연산을 제공하지 않는다. 따라서 효율적인 표본 추출 기법에 관한 연구가 요구되고 있다.

(9) 통계학과 학생들에 대한 OLAP 관련 기술의 효과적인 교육 방안에 관한 연구

최근 통계학과의 교과과정 운영 및 통계학과의 현실에 관한 논의가 활발하게 이루어지고 있다. 과거의 이론 지향적이던 방식에서 탈피하여 응용 분야에 대한 연구와 교육이 필요하다는 주장이 제기되고 있으며, 특히 현실적인 데이터 처리에 대한 교육의 필요성이 강조되고 있다. 따라서 OLAP, 데이터베이스, 데이터 마이닝, 패턴 인식, 데이터 수집 및 관리, Web 관련 기술 등의 분야를 효과적으로 교육할 수 있는 방안에 관한 연구도 매우 중요하다.

4. 결론

통계학은 종종 데이터를 다루는 학문으로 정의된다. 그러나 데이터에 관련된 응용 분야에 대한 통계학자들의 기여는 매우 적었다고 Friedman(1997)은 지적한다. OLAP, 패턴 인식, 데이터베이스 관리, 그래픽 모델, 데이터 시각화 기법, 데이터 마이닝 등과 같은 분야는 통계학자들에게 거의 무시되어진 분야이다. 그 이유는 무엇인가? Friedman의 지적대로 통계학을 데이터에 관한 학문으로 볼 때 "이들 분야는 통계학이 아니다"라는 이유는 설득력이 없다. Friedman은 또한 통계학자들이 현재와 같이 이론적인 연구를 계속 지향한다면 정보사회에서 통계학의 역할은 차츰 감소할 것임을 경고하면서, 데이터에 관련된 여러 기술들과 데이터 마이닝 분야에 통계학자들의 연구와 관심의 필요성을 주장한다. Kettenring(1997)은 정보사회에서 통계학이 나아갈 방향을 제시하면서, 데이터에 관련된 응용 분야들이 통계학의 중요한 연구 분야로 떠오를 것으로 예측하고 있다.

West 등(1998)은 통계학자들이 침체되어 가고 있는 통계학 분야를 계속 유지하기 위해서는 새롭고 흥미 있는 데이터가 필요함을 강조하면서, Web 기술의 활용을 적극 주장한다. Selfridge 등(1996)은 실세계에서 발생되는 데이터의 특징을 설명하고, 적절한 분석 방법의 개발 및 활용을 주장한다. Hand(1997)는 현대사회에서의 통계학은 과거에 통계학 분야에서 다루어지던 소규모의 데이터로부터 탈피해야 하며, 데이터 분석을 위한 새로운 기법의 개발은 실세계에서 나타나는 문제들을 해결하기 위한 것이어야 한다고 주장한다.

본 연구에서는 이러한 요구에 부응할 수 있는 한 방법으로서 OLAP 기술에 대해 살펴보았으며, 관심을 가질만한 몇몇 연구 과제를 제시해 보았다. OLAP, 데이터 마이닝과 같은 기술의 응용 분야는 실세계의 문제를 해결하고 사용자들에게 데이터 이용의 중요성에 대한 인식을 더욱 새롭게 심어줄 수 있을 것으로 생각된다. 또한 실세계에 이용할 수 있는 응용학문으로서 통계학을 연구하는 연구자에게 많은 기회를 줄 것으로 생각된다.

참고문헌

[1] 손건태, 허명회 (1999), 토론 : 통계학 학부전공 프로그램의 비전과 전략에 비추어, 「응용통계연구」, 제12권 2호, 705-709.

[2] 조재희, 박성진 (1999), 「OLAP 테크놀로지」, 시그마컨설팅그룹.

[3] 한경수, 안정용 (1998), 데이터베이스에서 통계정보를 제공하는 OLAP의 소개, 「한국통계학회 추계학술발표회 논문집」, 87-91.

[4] 한경수, 안정용 (1999), OLAP 시스템의 구조 및 활용, 「한국통계학회 추계학술발표회 논문집」, 73-76.

[5] Carickhoff, R. (1997), A New Face for OLAP, *Internet Systems*, http://www.dbmsmag.com/9701i08.html

[6] Chaudhuri, S. and Dayal, U. (1997), An Overview of Data Warehouses and OLAP Technology, *ACM SIGMOD Record*, Vol. 26, No. 1, 65-74. http://www.acm.org/sigmod/record

[7] Chaudhuri, S., Motwani, R. and Narasayya, V. (1999), On Random Sampling over Joins, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 263-274.

[8] Codd, E.F., Codd, S.B. and Salley, C.T. (1993), Providing OLAP to User-Analysts : An IT Mandate, *White Paper*.

[9] Famili, A., Shen, W. M., Weber, R. and Simoudis, E. (1997), Data Preprocessing and Intelligent Data Analysis, *Intelligent Data Analysis*, Vol. 1, No. 1, http://www-east.elsevier.com/ida/browse/vol1.htm

[10] Friedman, J.H. (1997), Data Mining and Statistics : What's the Connection?, *Proceedings of the International Conference on the Interface : Computing Science and Statistics*, http://www.stat.rice.edu/interface97.html

[11] Gray, P. and Watson, H. J. (1996), The New DSS : Data Warehouses, OLAP, MDD, and KDD, *Proceedings of the Second Americas Conference on Information Systems*, http://hsb.baylor.edu/ramsower/ais.ac.96/papers/graywats.htm

[12] Hand, D. J. (1997), Intelligent Data Analysis : Issues and Opportunities, *Intelligent Data Analysis*, Vol. 2, No. 2, 1-14.

[13] Huber, P.J. (1994), Huge Data Sets, *COMPSTAT(Proceedings in Computational Statistics)*, 3-13.

[14] Lenz, H.J. and Shoshani, A. (1997), Summarizability in OLAP and Statistical Data Bases, *Proceedings of International Conference on Statistical and Scientific Database Management*, http://www.lbl.gov/~arie/papers/

[15] Kettenring, J.R. (1997), Shaping Statistics For Success in the 21st Century, *Journal of the American Statistical Association*, Vol. 92, No. 440, 1229-1234.

[16] Olken, F. (1993), Random Sampling from Databases, Ph.D. Dissertation, University of Califonia at Berkeley.

[17] Selfridge, P. G., Srivastava, D. and Wilson, L. O. (1996), IDEA : Interactive Data Exploration and Analysis, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 24-34.

[18] Shoshani, A. (1982), Statistical Databases: Characteristics, Problems, and Some Solutions, *Proceedings of the International Conference on Very Large Databases*, 208-213.

[19] Shoshani, A. (1997), OLAP and Statistical Databases : Similarities and Differences, *Proceedings of the ACM Symposium on Principles of Database Systems(PODS)*, 185-196. http://www.lbl.gov/~arie/papers/

[20] Thomsen, E. (1997), *OLAP Solutions*, John Wiley & Sons.

[21] Vitter, J.S. (1987), An Efficient Algorithm for Sequential Random Sampling, *ACM Transactions on Mathematical Software*, Vol. 13, No. 1, 58-67.

[22] West, R. W., Ogden, R. T. and Rossini, A. J. (1998), Statistical Tools on the World Wide Web, *The American Statistician*, Vol. 52, No. 3, 257-262.