응용통계연구 제 17권 3호, 2004년, pp. 595-603

통계학과에서의 데이터베이스 교육 방안

안정용 [1] 한경수 [2] 최숙희 [3]

요 약

통계학과에서 데이터베이스 교육은 필요한가? 데이터베이스 교육이 필요하다면 그 교육 방안은 무엇인가? 본 연구에서는 통계학과에서 데이터베이스 교육의 필요성에 대해 논의하고 구체적인 교육 방안을 제안하고자 한다. 본 논문의 목적은 어떻게 통계학이라는 학문과 연관지어 데이터베이스를 교육할 수 있을 것인가에 대해 생각해보고자 하는데 있다.

주요용어: 데이터베이스 교육, 데이터 분석, 데이터베이스와 통계학, 정보 기술

1. 서론

컴퓨터와 정보통신 기술이 발전하면서 현대 조직 사회에서 온라인 업무처리(online transaction processing; OLTP)는 기본이 되어, 각종 업무와 관련된 다양한 종류의 데이터가 대량으로 양산되고 있다. 이들 데이터는 과거에 통계학 분야에서 다루어졌던 일반적인 통계 데이터와는 다른 특징을 가진다. 예를 들어, 모집단의 대표성도 의심스럽거니와 대부분의 데이터는 확률 표본(random sample)이 아니어서 통계적 기법들을 무분별하게 적용하면 문제가 발생할 수 있으며, Hand(1998)에서 지적하듯이 검정(testing)과 같은 통계적 분석 기법을 무의미하게 만들고 있다. 따라서 이러한 데이터를 활용하기 위한 새로운 기법이 요구되고 있으며, 데이터 마이닝(data mining), KDD(knowledge discovery in databases), CRM(customer relationship management), BI(business intelligence) 등에 대한 최근의 많은 관심은 이러한 흐름을 잘 반영하고 있다고 하겠다.

현대 사회에서 양산되는 데이터가 가지는 대용량적인 특성은 데이터 분석 기법뿐만 아니라 데이터 수집 및 관리에도 체계적인 방법을 요구하고 있다. 데이터의 수집과 체계적인 관리는 데이터 활용을 위한 선결요건이기 때문이다.

데이터베이스(database)는 이러한 대용량 데이터의 관리와 활용을 위한 적절한 기반을 제공하여 주는 응용 분야이며, 많은 분야에서 그 이용이 보편화되고 있다. 이러한 측면에서, 데이터베이스는 전산학 분야뿐만 아니라 통계학, 산업공학, 경영학, 의학 등 데이터와 관련된 많은 학문 분야에서 보다 적극적인 활용이 가능하다. 특히, 데이터베이스 이용 측면만 고려한다면 안정용과 한경수(2002)에서 언급한 바와 같이 전산학 분야보다는 오히려

1) (561-756) 전주시 덕진동 664-14, 전북대학교 수학 통계정보과학부, 조교수
E-mail: jyahn@chonbuk.ac.kr
2) (561-756) 전주시 덕진동 664-14, 전북대학교 수학 통계정보과학부, 교수
E-mail: kshan@chonbuk.ac.kr
3) (565-701) 전북 완주군 삼례읍 후정리 490, 우석대학교 전산정보학부, 교수
E-mail: shchoi@woosuk.ac.kr

여러 응용 분야에서 데이터베이스 활용에 대한 지식이 더 필요하다고 생각한다. 그 이유는 모든 사람이 한 데이터로부터 같은 정보를 추출하는 것이 아니듯이 전산학 분야와 응용 분야에서 데이터를 바라보는 시각이 분명 다를 것이기 때문이다. 그럼에도 불구하고 과거에 데이터베이스의 교육은 전산학 분야에서만 이루어져 왔으며, 위에서 언급한 응용 분야들에서는 거의 다루어지지 않은 것이 사실이다. 그 이유는 여러 가지가 있겠지만 다음과 같은 두 가지가 중요한 요인이 아니었나 생각된다. 첫째, 데이터베이스를 교육할 수 있는 인력과 실습 등의 주변 환경적인 여건(하드웨어와 소프트웨어 여건)이 충분하지 못했고, 두 번째로는 데이터베이스의 필요성에 대한 인식의 부족을 들 수 있을 것이다. 그러나 최근 컴퓨터 환경의 발전은 여러 응용 분야에서도 얼마든지 데이터베이스를 교육할 수 있도록 도와주고 있으며, 이 분야에 대한 관심이 높아지고 있다.

통계학과에서도 데이터베이스 교육이 증가되는 추세에 있으며, 몇몇 연구들을 통하여 그 필요성이 언급되어 왔다. 그러나 대부분의 연구에서, 통계학과에서 데이터베이스 교육이 필요하다는 사실만을 간단히 언급할 뿐 구체적으로 그것이 왜 필요하고, 또 교육은 어떻게 해야 되는지의 실천적 교육 방안에 대한 논의는 찾아보기 힘들다.

본 연구에서는 통계학과에서 데이터베이스 교육의 필요성에 대해 살펴보고, 그 교육 방안을 몇 가지 측면에서 제안해보고자 한다. 특히, 어떻게 통계학이라는 학문과 연관지어 데이터베이스를 교육할 수 있을 것인가에 대해 생각해보고자 하는데 본 논문의 목적이 있다.

2. 통계학과에서 데이터베이스 교육의 필요성

Friedman(1997)이 지적한 바와 같이 데이터베이스 분야에 대한 통계학자들의 관심은 최근까지도 미미한 실정이며, 통계학과의 정규 교육과정에서도 많이 다루어지지 않은 것이 사실이다. 물론 통계학과 교육과정의 개선에 관한 연구를 비롯한 몇몇 응용 논문을 통하여 통계학과에서 데이터베이스 교육이 필요하다는 사실은 간혹 제기되어 왔다. 박헌진 등(1998)은 통계계산 교과과정을 제언하면서 '전산통계 II' 과목에서 데이터베이스에 대한 내용을 다룰 수 있음을 제안하고 있으며, 조신섭 등(1999)은 정보 관련 통계학과의 교과과정에 '통계 데이터베이스' 과목의 추가를 제안하고 있다. 손건태와 허명회(1999), Bryce 등(2001) 또한 데이터베이스 교육의 필요성을 언급하고 있으며, Ritter 등(2001)은 통계학 전공자들의 취업을 하기 위해 요구되는 자질 중 하나로 데이터베이스의 활용능력을 언급하면서, 통계학의 단편적인 지식뿐만 아니라 종합적인 능력이 필요함을 강조하고 있다.

최근에 국내의 많은 통계학과들이 정보 관련 학과로 변신을 꾀하면서 데이터베이스에 대한 관심도 높아지고 있다. 이러한 변화는 각 학과에서 운영하고 있는 교육과정에 데이터베이스가 정규 과목으로 추가되고 있는 점을 살펴보면 쉽게 알 수 있다. 2002년도에 웹(Web) 상에서 조사 가능한 국내 45개 대학의 통계학과 학부과정의 교과과정을 필자들이 조사해 본 결과, 이 중 22개의 학과에서 데이터베이스 과목을 포함하고 있었다. 표 2.1은 국내 통계학과에서 개설되어 있는 데이터베이스 과목명을 정리한 것이다. 이러한 과목들의 강의가 어떻게 이루어지고 있는지, 또 실제로 설강이 되고 있는지 여부는 미지수이지만 과거에 거의 없었던 점을 감안하면 많이 증가된 것이 사실이고 앞으로 더욱 늘어날 것으로 생각된다.

표 2.1: 데이터베이스 과목

데이터베이스, 데이터베이스론, 데이터베이스 기초
데이터베이스 시스템, 데이터베이스 활용, 데이터베이스 실무
분산데이터베이스 실습, 응용 데이터베이스, 통계데이터베이스

통계학과에서 데이터베이스 교육이 필요한 이유는 여러 관점에서 논의될 수 있지만, 여기에서는 세 가지 측면에서 살펴보도록 한다. 첫째는 데이터 분석 능력의 향상이다. 이것은 데이터베이스를 활용할 수 있는 능력을 가지고 있으면, 데이터베이스에서 데이터를 가져와 통계 패키지와 같은 응용 프로그램을 이용하여 분석할 수 있는 능력을 갖게 된다는 것이다. 만약 데이터베이스를 모르면 어떤 데이터가 어떤 구조로 어디에 저장되어 있는지 모르기 때문에, 데이터를 분석할 수 없음은 물론 현실 세계에서 발생하는 여러 가지의 데이터에 대한 문제 인식조차 생길 수 없을 것이다. 예를 들어, Web에서 발생하는 대표적인 데이터 중의 하나가 로그(log) 데이터인데, 이 데이터를 활용하고자 하는 통계학자는 거의 없고 대부분 전산학 하는 사람들이 이 데이터를 분석에 활용하고 있다. 그 이유는 로그 데이터라는 것 자체와 이 데이터가 어디에 어떻게 존재하는지에 대해 모르기 때문일 것이다.

둘째는 데이터 분석을 고려한 데이터베이스 설계를 할 수 있다는 점이다. 전산학 전공자들은 일반적으로 효율적인 업무처리, 즉 OLTP에만 관심을 갖고 데이터베이스를 설계하므로 의사 결정에 필요한 데이터 저장에는 무관심할 수밖에 없다. 따라서 데이터에서 의사결정에 필요한 유용한 정보를 끄집어 낼 수 있는 통계인이 데이터베이스 설계에 반드시 참여해야 한다. 그렇지 않으면 쓸모 없는 데이터가 잔뜩 쌓여 있는 곳에서 유용한 정보를 찾아야 하는 어려움에 당면하게 되고, 이는 원천적으로 매우 힘든 일이기 때문이다.

셋째는 현대 사회에서 데이터베이스를 배제한 데이터 활용은 생각할 수 없는 환경이 되어가고 있다는 점이다. 현대 사회에서 발생하는 많은 데이터는 데이터베이스에 저장되며, 데이터의 특성에 따라 여러 개의 테이블들에 나뉘어져 보관된다. 따라서 이러한 데이터를 활용하기 위해서 데이터베이스에 대한 이해는 필수적이다.

3. 교육 방안

앞 절에서 우리는 통계학과에서 데이터베이스 교육의 필요성에 대해 살펴보았다. 그러면, 교육은 어떻게 해야 되는가? 이 절에서는 데이터베이스를 교육할 때의 구체적인 교육 방안에 대해 살펴보고자 한다.

통계학과에서 데이터베이스를 교육할 때 가장 우선적으로 고려해야 할 점은 데이터베이스의 일반적이고 이론적인 내용에 치중하기보다는 통계학이라는 학문과의 관련성을 높일 수 있도록 해야 한다는 것이다. 이것은 데이터베이스를 교육할 때 항상 데이터 처리/분석 및 활용을 염두에 두어야 한다는 의미이며, 실생활에서 데이터베이스를 활용할 수 있는 교육이 되어야 한다는 의미이기도 하다. 여기에서는 교육 방안을 데이터베이스 내용적인

측면, 통계학과의 관련성 고려 측면, 교육효과 개선과 실습 위주 교육을 지원하기 위한 정보기술 활용 측면 등으로 구분하여 살펴보기로 한다.

3.1. 데이터베이스 내용적인 측면

데이터베이스를 다루는 여러 교재들을 살펴보면, 구성된 내용을 다음 표 3.1과 같이 데이터베이스에 관한 기초적인 내용과 고급/응용적인 내용으로 분류할 수 있다.

표 3.1: 데이터베이스 내용

구분	내용	비고
기초내용	데이터베이스 개념	*
	관계형 데이터베이스	*
	관계형 데이터베이스 설계	*
	SQL	*
고급내용	질의 처리와 최적화	
	트랜잭션 처리	*
	동시성 제어	
	보안 및 권한 관리 등	

앞에서 언급한 바와 같이, 통계학과에서는 데이터베이스를 이론적으로 접근하기보다는 실제적인 활용 측면이 강조되어야 할 것이다. 따라서 데이터베이스 기초 내용은 실습을 병행하면서 강의하고, 고급 내용은 트랜잭션 처리 부분 정도가 우선 필요할 것으로 생각된다(표에서 * 표시 부분).

트랜잭션(transaction) 처리 부분은 현재 구축되어 있는 대부분의 데이터베이스 시스템이 트랜잭션 처리를 위하여 구축되어 있기 때문에 간단히 살펴볼 필요가 있다. 트랜잭션 처리 위주의 데이터베이스는 데이터 분석적인 측면보다는 데이터의 운영측면에 관심이 있기 때문에 전산학적인 관점에서는 문제될 것이 없겠으나 데이터 분석적인 관점에서는 많은 문제점이 노출되고 있다. 이 부분에 대한 이해를 통하여 데이터 분석을 고려한 데이터베이스 설계 방향을 설정하는 데 도움이 될 수 있다.

3.2. 통계학과의 관련성 고려 측면

통계학과에서 데이터베이스를 교육할 때 발생하는 가장 큰 문제는 통계학 분야와의 연결이 미흡하다는 점이다. 이 문제는 통계학과에서 데이터베이스를 교육하면서 전산학 전공자들을 활용한다는 데에서 기인된다. 전산학 전공자들은 데이터베이스를 통계학 분야에서 어떻게 연결하여 응용할 수 있는가에 대해 생각하지 않는다. 이것은 매우 중요한 문제이다. 그 이유는 통계학과에서 데이터베이스를 교육하면서 그것이 왜 그리고 어디에 필요

한지에 대한 것을 생각하지 않고 교육이 이루어진다는 것을 의미하는 것이며, 당연한 결과로 교육 성과를 기대하기는 대단히 어렵기 때문이다.

이러한 맥락에서 통계학과에서 데이터베이스를 교육할 때는 데이터베이스 그 자체가 목적이 되어서는 곤란하며, 통계학 분야와 항상 연결하여 생각하여야 한다. 따라서 다음과 같은 사항들이 포함되도록 교육 체계를 구성할 필요가 있다.

- 데이터 분석 응용 프로그램과의 연결

  데이터를 분석할 때 SAS, SPSS, EXCEL 등과 같은 응용 프로그램이 자주 이용되며, 많은 통계학과 및 관련학과에서 정규 수업 시간을 통하여 다루어지고 있다. 이러한 응용 프로그램을 이용하여 데이터를 분석하는 일반적인 방법은 첫째, 적당한 형식에 맞추어 데이터를 입력하고, 둘째, 원하는 분석 방법을 선택하여 결과를 산출하는 과정으로 이루어진다. 물론 데이터를 꼭 응용 프로그램 안에서 입력할 필요는 없으며, 외부에서 텍스트 파일 형태로 만들어 이용해도 된다. 중요한 것은 첫 번째 단계에서 누군가는 어떤 형태로든지 데이터를 입력해야 된다는 사실이고, 이것이 지금까지 데이터를 분석해온 전형적인 방법이다.

  그러나 이제는 이러한 방법으로 데이터를 분석하는 것은 학교에서 수업을 목적으로 하는 경우에는 상관이 없을지 몰라도 사회에서 다양하게 나오는 요구를 충족하기에는 부족하다. 위에서 언급했듯이 현대 사회에서 발생하는 많은 데이터는 데이터베이스에 저장되고 있으며, 대용량인 경우가 많기 때문이다. 따라서 데이터베이스에 저장된 데이터를 SAS, EXCEL 등에서 분석하는 방법이 다루어져야 될 필요성이 있다.

- 대용량 데이터라는 특성 감안

  대용량 데이터인 경우 요구된 질의를 수행하고, 그 결과를 반환하는 시간이 매우 많이 소요될 수 있기 때문에 이러한 데이터를 분석하기 위한 다양한 방법이 연구되고 있다. 통계학 분야와 친숙한 대표적인 방법은 표본을 추출하여 이용하는 기법과 충분통계량(sufficient statistics)을 이용하는 기법 등을 들 수 있다. 표본 추출 기법은 데이터를 검색하는 비용과 통계적 분석에 필요한 계산의 양 등을 줄일 수 있는 이점을 제공해 줄 수 있으며, Olken(1993), Vitter(1987), Chaudhuri 등(1999)에서 표본 추출 알고리즘에 대한 연구가 진행되었다.

  한편, Moore 등(1998)은 대용량 데이터의 분석을 위하여 충분 통계량을 이용하는 기법을 연구하였다. 이 기법은 충분 통계량을 미리 계산해 놓고 분석에 이용한다는 점에서 요약 테이블을 이용하는 방법과 개념적으로 비슷하다고 할 수 있다. 또 Tendick(2002)은 대용량 데이터에서 EDA(exploratory data analysis) 기법의 활용에 대해 연구하였다. 이러한 기법들을 데이터베이스를 교육할 때 소개하면 데이터 분석 능력의 향상은 물론 통계학의 개념과 연결시킬 수 있다는 장점이 있다.

- 데이터 분석을 고려한 데이터베이스 설계

  데이터베이스를 어떻게 설계하는가에 따라 데이터로부터 정보를 추출하는 효율성이 크게 좌우될 수 있다. 그러나 데이터베이스 설계는 꽤 어려운 문제이기 때문에 학부

과정에서는 기본적인 내용만을 다룰 수밖에 없다. 가능하다면, 통계적인 관점에서 결측치(missing values) 처리, 시계열 데이터의 효과적 활용을 위한 데이터베이스 설계 문제 등을 생각해보면 좋을 것이다.

3.3. 정보기술 활용 측면

최근의 정보 기술의 발전은 여러 학문 분야의 교육적 활용에 많은 도움을 주고 있다. 데이터베이스 교육에서도 정보 기술, 특히 웹을 이용하면 교육의 효과를 높일 수 있을 것으로 생각된다.

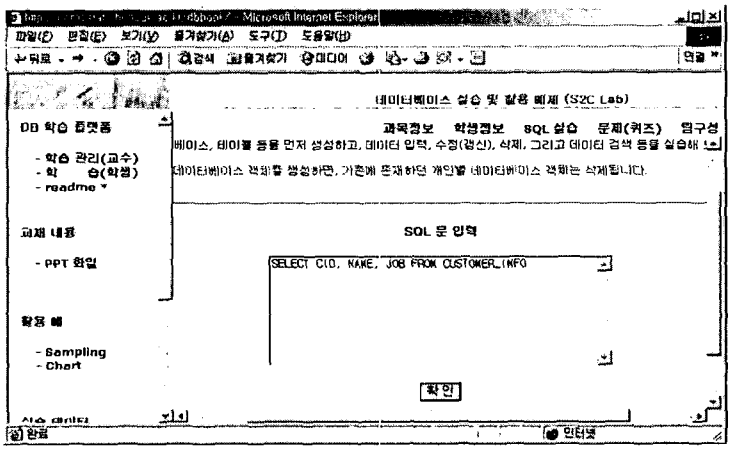

그림 3.1: SQL 문장 입력

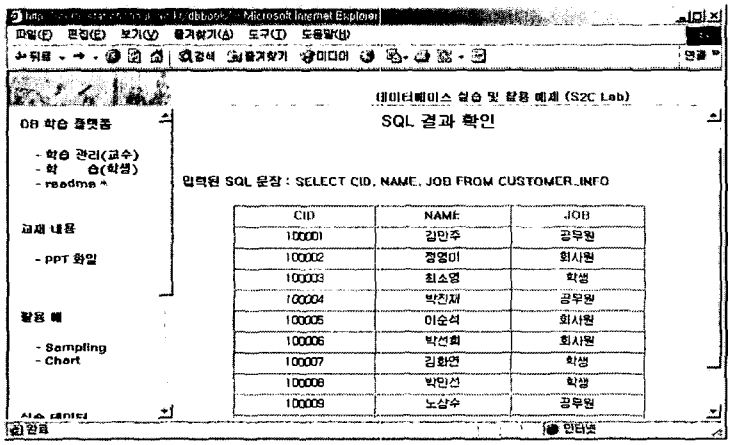

그림 3.2: 질의 결과

데이터베이스와 같은 응용 과목은 실습이 뒷받침되어야 그 효과를 배가시킬 수 있기 때문이다. 또한 실생활에 활용할 수 있는 예제를 구축해봄으로서 데이터베이스의 실제적인 활용에 대한 이해를 높일 수 있을 것으로 생각된다. 웹을 데이터베이스 교육에 이용할 때의 장점은 첫째, 교재에서 다루기 어려운 다양한 데이터나 문제를 제공할 수 있고, 둘째, 그림 3.1과 같은 실습 도구를 제공함으로서 학생들이 데이터베이스 설계 및 SQL을 직접 실습할 수 있으며, 셋째, 교수 입장에서는 학생들의 학습 현황 모니터링 및 적절한 평가를 할 수 있다는 점을 들 수 있다. 그림 3.1과 그림 3.2는 웹 상에서 SQL을 실습하는 예제이다. 학생들에게 데이터베이스에서 질의하는 실습 문제를 제시하고, 학생들이 질의를 수행하면 그 질의 내용과 결과를 제공한다. 이 결과는 학생과 교수 모두에게 제공될 것이고, 교수는 이러한 내용을 평가에 활용할 수 있을 것이다.

3.4. 교재 개발

통계학과에서 데이터베이스를 교육할 때의 어려운 문제점 중의 하나는 이용할 수 있는 교재가 거의 없다는 점이다. 기존의 데이터베이스 교재는 전산학 전공자가 전산학과 학생들의 교육에 적합하게 작성한 것이 대부분이어서 그 내용이 다분히 이론적일 뿐만 아니라 통계학과와 같은 응용 분야의 학생들에게는 부적합하다고 생각된다. 따라서 위에서 언급한 내용들을 고려하여, 통계학과에서 활용할 수 있는 좋은 교재의 개발에도 많은 관심이 요구된다. 다양한 교재의 개발 경험을 통해서 보다 효과적인 교육으로 연결시킬 수 있을 것이기 때문이다.

4. 결론

통계학은 데이터를 다루는 학문으로 흔히 정의되며, 데이터베이스는 데이터를 활용하기 위한 기반 기술이다. 그러한 맥락에서 데이터베이스는 통계학 분야에서 활용할 수 있는 매우 유용한 도구이지만, 현재까지도 활용범위가 넓지 않다. 그러나 최근에 그 필요성에 대한 인식이 많이 달라지고 있으며 교육에 대한 관심도 높아지고 있다.

본 연구에서는 통계학과에서 데이터베이스 교육의 필요성에 대해 살펴보고, 교육 방안을 몇 가지 측면에서 제안해 보았다. 물론 여기에서 살펴본 내용이 전부일 수는 없겠지만, 통계학과에서 데이터베이스를 교육할 때 다소간 도움이 되지 않을까 기대해보며, 통계학과에서 데이터베이스를 왜 교육해야 되는지에 대한 당위성을 찾을 수 있는 하나의 동기(motive)를 제공할 수 있지 않을까 생각한다.

참고문헌

박헌진, 신봉섭, 심송용, 유종영, 이승천, 이정진 (1998). 변화하는 정보화 사회에 대응되는 통계계산 교과과정의 제언, <한국통계학회 춘계 학술논문발표회 논문집>, 75-79.

손건태, 허명회 (1999). 토론 : 통계학 학부전공 프로그램의 비전과 전략에 비추어, <응용통계연구>, **12**, 705-709.

안정용, 한경수 (2002). 통계학과에서의 데이터베이스 교육 방안, <한국통계학회 학술논문발표회 논문집>, 231-234.

조신섭, 신봉섭, 이상복, 한경수 (1999). 정보관련 통계학과의 교과과정, <응용통계연구>, **12**, 683-703.

Bryce, G. R., Gould, R., Notz, W. I. and Peck, R. L. (2001). Curriculum guidelines for bachelor of science degrees in statistical science, *The American Statistician*, **55**, 7-13.

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.

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

Hand, D. J. (1998). Intelligent data analysis : issues and opportunities, *Intelligent Data Analysis*, **2**, 1-14.

Moore, A. and Lee, M. S. (1998). Cached sufficient statistics for efficient machine learning with large datasets, *Journal of Artificial Intelligence Research*, **8**, 69-91.

Olken, F. (1993). Random sampling from databases, *Ph.D. Dissertation*, University of Califonia at Berkeley.

Ritter, M. A., Starbuck, R. R. and Hogg, R. V. (2001). Advice from prospective employers on training BS statisticians, *The American Statistician*, **55**, 14-18.

Tendick, P. (2002). EDA in the age of very large databases, *Proceedings of the Joint Statistical Meetings - Business and Economic Statistics Section*, 3436-3441.

Vitter, J.S. (1987). An efficient algorithm for sequential random sampling, *ACM Transactions on Mathematical Software*, **13**, 58-67.

[2003년 9월 접수, 2004년 6월 채택]

A Note on Database Education in Statistics Undergraduate Course

Jeong Yong Ahn [1] Kyung Soo Han [2] Sook Hee Choi [3]

ABSTRACT

Does database education need in statistics undergraduate course? Then, how must we do education? In this article we examine the necessity of database education in statistics department and propose some concrete plans for instruction. The goal of this article is to explore how to educate database in connection with statistics.

Keywords: Database education; Data analysis; Database and statistics; Information technology.

1) Assistant Professor, Div. of Mathematics and Statistical Informatics, Chonbuk National University, 664-14 duckjin-dong, duckjin-gu, Chonju, Chonbuk, 561-756, Korea
E-mail: jyahn@chonbuk.ac.kr

2) Professor, Div. of Mathematics and Statistical Informatics, Chonbuk National University, 664-14 duckjin-dong, duckjin-gu, Chonju, Chonbuk, 561-756, Korea
E-mail: kshan@chonbuk.ac.kr

3) Professor, Div. of Computer and Information Science, Woosuk University, 490 hujeong-ri, samrye-eup, Chonju, Chonbuk, 565-701, Korea
E-mail: shchoi@woosuk.ac.kr