응용통계연구 제 13권 2호, 2000년, pp. 287-296

웹을 이용한 데이터 수집 및 관리에 관한 연구 : 강의평가 시스템 구현

안정용[1] 최승현[2] 한경수[3]

요 약

데이터 수집과 관리, 그리고 분석을 통한 정보의 제공은 통계학의 기본적 사안이며, 다량의 데이터가 양산되는 현대 사회에서 그 중요성은 더욱 확대되고 있다. 그러나 이러한 분야에 대한 연구는 분석분야에 비하여 상대적으로 많이 이루어지지 못한 것이 사실이다. 본 연구에서는 웹을 활용한 데이터 수집과 관리 방법에 대한 일반적인 사항들에 대해 살펴보고, 데이터를 수집함과 동시에 분석하여 사용자에게 정보를 전달해 줄 수 있는 강의평가 시스템을 사례로 제시한다.

주요용어: 데이터 수집 및 관리, 웹 어플리케이션, 강의평가 시스템.

1. 서론

현대 사회에서 정보의 활용은 가장 관심 있는 사항이며, 대부분의 정보는 데이터를 통하여 얻어진다. 따라서 통계학의 기본적 영역인 데이터의 수집과 관리(collection and management), 정보를 추출하기 위한 분석(analysis), 추출된 정보를 사용자에게 전달하는 기술 등은 매우 중요하다. 그러나 손건태와 허명회(1999), Short와 Pigeon(1998)에서 지적하듯이 데이터 수집 및 관리는 분석 분야에 비하여 상대적으로 소홀하게 취급되어 온 것이 사실이며, 데이터 수집에 대한 연구는 Schwarz(1997)에서 지적하는 바와 같이 대부분 표본 조사에 국한되어 왔다.

이러한 현상은 통계학의 학문적 영역을 스스로 제한하는 결과를 가져왔을 뿐만 아니라, 학생들이 데이터를 다룰 수 있는 전반적인 능력보다는 데이터의 계산과 분석에 치중할 수 밖에 없는 교육 환경을 조성하는 결과를 초래하였다. Higgins(1999)는 이러한 교육 환경에 대한 문제점을 지적하면서 데이터를 수집하고 관리할 수 있는 데이터 전문가(data specialist)를 양성할 것을 주장하고 있으며, 손건태와 허명회(1999), 허명회(1999)는 인터넷을 활용한 데이터 수집 또는 정보탐색 등의 교과 개발과 더불어 데이터 관리 분야를 발전시킬 수 있는 연구의 필요성을 강조하고 있다.

1) (590-711) 남원시 광치동 720, 서남대학교 컴퓨터정보통신학부 조교수
E-mail: jyahn@tiger.seonam.ac.kr
2) (561-756) 전주시 덕진동 664-14, 전북대학교 전산통계학과
E-mail: shchoi@stat.chonbuk.ac.kr
3) (561-756) 전주시 덕진동 664-14, 전북대학교 수학 통계정보과학부 교수
E-mail: kshan@stat.chonbuk.ac.kr

한편, Chambers(1993)는 'Greater or Lesser Statistics: A Choice for Future Research'에서 통계학이라는 학문을 좀 더 넓은 관점에서 바라볼 것을 제안한다. 그는 통계학의 영역이 데이터 분석이라는 생각을 버려야 하며, 데이터를 포괄적으로 다룰 수 있는 어플리케이션 개발에 대해 학생들에게 교육하고, 그러한 연구에 통계학자들이 참여할 것을 요구한다.

최근 컴퓨터와 네트워크(network) 기술의 발전은 이러한 요구들을 해결하기 위한 적절한 환경을 조성해 주고 있다. 특히 World Wide Web(WWW, 이하 웹)으로 대표되는 네트워크의 발전은 전 세계 컴퓨터들간의 정보 공유를 가능토록 해 주고 있으며, West 등(1998)에서 주장하듯이 통계학의 교육 및 연구 분야에 많은 기회를 제공하고 있다. 학생들의 표본조사 교육을 위하여 설계한 'StatVillage'(Schwarz, 1997)는 데이터 수집 교육에 웹을 활용하는 좋은 예이며, 웹을 이용하면 데이터 수집과 관리는 물론 데이터를 분석하여 추출한 정보를 손쉽게 전달할 수 있다.

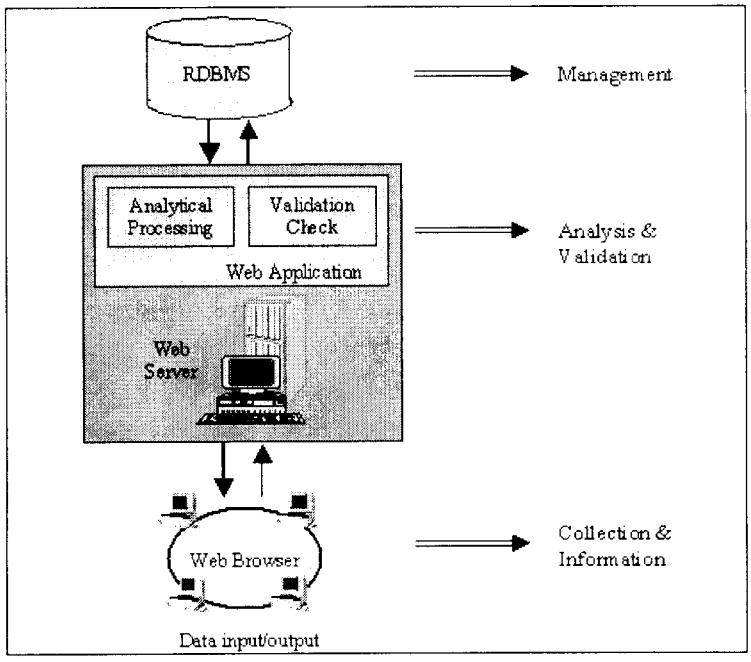

그림 1.1: 웹 상에서의 데이터 활용

그림 1.1은 웹 상에서 데이터 수집, 관리, 분석, 그리고 분석 정보를 전달할 수 있는 시스템의 구조를 간단히 표현한 것이다. 웹 브라우저를 통해 생성된 데이터는 타당성이 검토된 후에 데이터베이스에 저장되어 관리 된다. 어플리케이션에 데이터의 타당성을 검토하는 과정의 추가는 과거의 설문조사 방식에서 겪었던, 타당성이 의문시 되는 데이터 처리의 어려움을 줄일 수 있으며, 데이터를 온라인 상에서 입력받음으로써 비표본 오차 중 하나인 데이터 입력자의 실수에 따른 입력오차를 염려하지 않아도 된다. 또한 이렇게 수집된 데이터는 분석 처리 과정을 거쳐 사용자에게 실시간으로 정보를 전달할 수 있는 장점이 있다.

본 연구에서는 웹을 활용한 데이터 수집에 관한 방법론적인 측면과 데이터 관리에 대

한 일반적인 사항들에 대해 살펴보고, 데이터를 수집함과 동시에 분석하여 사용자에게 정보를 전달해 줄 수 있는 강의평가 시스템을 사례로 제시한다.

2. 데이터 수집 및 관리

2.1. 데이터 수집

웹의 발전은 사용자들이 수 많은 정보를 손쉽게 이용할 수 있고 다양한 데이터를 수집할 수 있는 환경을 제공하고 있다.

이 논문에서 우리는 웹을 이용하여 수집할 수 있는 데이터를 크게 (1)탐색가능한 데이터(웹상에 산재되어 있거나 웹을 통해 제공되는 이미 존재하는 데이터)와 (2)생성 데이터(로그인(login), 설문조사, bar-code, 거래 처리, 공정 관리 등으로부터 생성되는 데이터)로 나누어 이러한 데이터 수집에 관한 방법론적인 측면과 확장할 수 있는 연구 분야를 살펴보고자 한다.

(1) 탐색가능한 데이터 수집

웹을 통하여 탐색가능한 데이터를 수집할 수 있는 첫번째 방법은 웹 페이지에 산재하는 데이터를 수집하고 취합하는 것이다. 각종 신문이나 언론 매체들에서 제공하는 데이터를 예로 들 수 있다.

필요한 데이터를 수집하기 위해서는 웹 페이지에 대한 검색이 필요하며, 검색은 주로 검색 디렉토리와 검색 엔진(예를 들어, 심마니, 까치네, yahoo, altavista, lycos 등)을 통해 이루어진다. 그러나 이 방법은 웹상에서 데이터를 제공하는 사이트 수가 기하급수적으로 증가하기 때문에 매우 지루한 작업이 될 수 있으며, 원하는 정보를 포함하고 있는 웹 페이지를 찾기가 쉽지 않다. 이러한 문제는 웹상에서 정보를 찾고자 하는 사용자 대부분이 느끼고 있으며, 이 문제를 해결하기 위한 많은 연구(Carriere와 Kazman 1997, Fred 등 1996)들이 진행되고 있으나 현재까지는 특별한 해결책이 제시되지 못하고 있는 실정이다. 따라서 이 방법을 이용하여 원하는 데이터를 검색, 수집, 취합하는 것은 많은 시간과 노력을 필요로 하는 단점이 있다. 그러나 실 생활과 밀접한 관계를 가지고 있는 데이터를 수집할 수 있다는 점에서 데이터가 수집되면 교육적인 측면이나 사회 상황의 추이를 살펴보는데 활용도가 높다고 생각 되며, 데이터 검색시 HyperLink기능, 검색엔진, Bookmark의 적절한 활용 등을 추천한다.

두번째 방법은 파일(file) 형태나 데이터베이스(database)로부터 제공되는 데이터를 수집하는 것이다. 이 방법은 현재 가장 흔하게 접할 수 있는 데이터 제공 방법이며, JSE(Journal of Statistics Education), StatLib Archive, 통계청 데이터 정보 등을 예로 들 수 있다.

웹의 수많은 사이트 중에 자신이 원하는 데이터를 제공하고 있는 사이트를 어떻게 찾을 수 있을까? 단순히 검색엔진에서 키워드 검색 결과만 가지고는 그 사이트가 적합한 내용을 담고 있는지를 평가할 수가 없다. 이러한 유형의 데이터 수집시 학술정보 resource page(The WWW Virtual Library (http://vlib.org)) 또는 Digital Librarian(http://www.servtech.com/

public/mvail/home.html), Carnegie Mellon 대학의 StatLib(http://stat.cmu.edu/)와 같은 사이트를 활용할 것을 추천한다. 이 방법으로 제공되는 데이터는 쉽게 수집할 수 있고 일반적으로 다량의 데이터인 경우가 많기 때문에 연구 분야를 확장할 수 있다. 대용량 데이터로부터 통계 정보를 효율적으로 추출하기 위하여 Olken(1993), Haas(1997)등이 연구한 '데이터베이스에서 표본을 추출하는 기법'에 관한 연구는 좋은 예이다.

(2) 생성 데이터 수집

이 방법은 설문조사나 사용자의 로그인(login), bar-code, 거래 처리, 공정 관리 등을 통해 직접적으로 데이터를 수집하는 방법이다. 현재 많은 웹 사이트들에서 사용자의 의견을 설문 조사하여 현실성 있게 활용하고 있으며, 어떤 시스템에 접속할 때 로그인 정보를 입력하는 사이트를 흔히 볼 수 있다. 이 방법은 온라인(online)상에서 데이터를 수집 함으로써 비용 절감은 물론 온라인 또는 실시간 환경에서 데이터를 동적으로 처리할 수 있어 정보를 빠르게 이용할 수 있는 장점이 있다. 데이터의 온라인 동적처리는 급변하는 사회여건을 감안할 때, 여러 분야에서 아주 유용하게 이용될 수 있다. 특히, 제조업, 금융 분야를 포함한 많은 기업 환경에서 중요한 사안이며, 앞으로 많은 연구가 필요할 것으로 생각된다.

이 방법으로 데이터를 수집하여 활용하기 위해서는 적절한 인터페이스 개발과, 수집된 데이터를 관리하기 위한 관리 기술의 이용이 필수적이며, 통계데이터베이스 분야를 비롯한 여러 가지 연구분야(예를 들면, 결측 데이터 처리 문제, 데이터 타당성 문제, data visualization 문제, sparse matrix handling 문제 등)를 제공한다.

2.2. 데이터 관리

각종 계획 수립을 위한 기초자료는 물론 합리적 의사결정에 필요한 정보 취득에 이르기까지 데이터의 역할은 매우 중요하며 차후 여타 분야에서 중요한 자료로 이용될 수 있는 데이터 재사용 측면에서, 효율적인 관리와 보급은 매우 중요한 의미를 갖는다고 할 수 있다.

일반적으로 과거의 통계학 분야에서 다루어지던 데이터는 행렬 형태로 표현되는 비교적 소규모의 데이터가 주를 이루어 왔으며, 이러한 데이터는 거의 대부분이 저차원적으로 구성되어 있고, 검정(testing)에 필요한 속성인 동질성(homogeneity)을 가지고 있는 특징이 있다. 그러나 최근의 컴퓨터에 기반한 디지털 기술의 발달은 웹과 같은 가상 공간을 통하여 엄청난 양의 데이터를 생산해 내고 있다. 이들 데이터는 실험 데이터와는 달리 고차원적이며 동질성을 갖지 않는 경우가 대부분이다.

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.

데이터의 효율적인 관리는 여러 측면에서 살펴볼 수 있겠으나 여기에서는 일반적인 측면과 통계적인 측면으로 구분하여 생각하기로 한다.

일반적인 측면에서, 컴퓨터를 이용하여 데이터를 관리하는 전통적인 방법은 파일 단위

관리 방식이다. 이 방법은 하나의 데이터 파일이 있을 경우 다른 파일이 갖고 있는 내용을 참조하기가 어렵고, 중복되는 데이터가 많이 발생할 수 있으며, 다량의 데이터를 취급하는 것은 사실상 불가능하다는 단점을 가지고 있다. 데이터베이스 기술은 이러한 문제를 해결하기 위한 개념으로 출발하여 현재 대부분의 데이터 관리를 위하여 이용되어지고 있으며, 방대한 양의 데이터를 관리하기 위해서 데이터베이스의 이용은 필수적이라 할 수 있다. 통계분야에서도 데이터베이스에 대한 몇몇 연구가 진행되었으며(Shoshani 1997, Gillman 등 1996, Lenz와 Shoshani 1997) 앞으로 많은 발전이 예상되는 분야이다.

통계적인 측면에서 데이터 관리는 파일이나 데이터베이스에 단순히 데이터를 축적하는 것만을 의미하지는 않는다. 효율적으로 데이터를 이용하기 위해서는 데이터의 입력, 수정 및 검색 등이 용이하고 통계 메타데이터(metadata) 질의와 필요한 통계 정보를 쉽게 추출해 낼 수 있도록 관리되어야 한다. 또한 다른 응용프로그램에서 쉽게 이용할 수 있는 형태로 제공될 수 있어야 한다. 따라서 데이터베이스 테이블(table)의 효율적인 물리적 설계, 집계(aggregation) 데이터를 어떻게 구성하고 활용할 것인가에 대한 문제, 데이터 제공 형태 등의 고려는 중요한 요소라 할 수 있다.

3장에서 사례로 제시되는 강의평가 시스템은 이러한 문제들을 적용하고자 하는 실험적인 프로토타입이라 할 수 있다. 설계된 테이블은 설문 문항의 변경에 영향을 받지 않으며(이것은 데이터베이스 테이블(table)의 구조를 변경할 필요가 없다는 뜻이다), 데이터가 수집 또는 삭제됨과 동시에 온라인 상에서 동적으로 필요한 집계 데이터를 미리 산출하여 사용자의 요구에 응답하도록 설계되어 있다.

데이터 관리에 대한 연구 분야는 매우 포괄적이라 할 수 있으나, 현재까지 사실상 미개척 분야이다. 최근에 많이 연구되어지고 있는 데이터 웨어하우징(data warehousing), OLAP(online analytical processing), 데이터 마이닝(data mining), 다차원 데이터베이스(multidimensional database) 등은 모두 데이터 관리 및 이용과 밀접하게 연관되는 분야라 할 수 있다.

3. 강의평가 시스템

3.1. 개발도구

본 연구에서 구현 사례로 제시하는 강의평가 시스템은 데이터 수집, 관리와 분석, 그리고 분석 정보의 전달 등을 포함한다.

웹 어플리케이션 개발시 이용할 수 있는 도구들은 매우 많으며, 각각 장단점이 있다. 표 3.1은 이번 연구에서 이용한 개발도구를 정리해 놓은 것이다.

3.2. 강의평가 시스템 내용

데이터 수집은 강의평가 설문지를 통하여 이루어지며, 필요에 따라 결측 데이터 처리 방안이나 데이터의 타당성을 검토할 수 있다. 수집된 데이터는 웹 어플리케이션을 통하여 데이터베이스 테이블에 저장 관리 된다. 분석 정보는 데이터의 민감성을 고려하여 일반 사

표 3.1: 개발 도구

운영체제	Windows NT Server
웹 서버	IIS(Internet Information Server)
데이터베이스 관리 시스템	SQL Server
웹 어플리케이션 개발도구	MS Visual InterDev, ASP(Active Server Page), Dynamic HTML 등
Chart 도구	ChartFX

용자와 관리자를 구분하여 제공되며, 관리자에게는 전체적인 정보를, 일반 사용자에게는 자신에 해당되는 정보만를 전달하도록 설계되었다.

강의평가를 하기 위하여 login 하면 강의평가를 이미 실시한 과목과 실시하지 않은 과목을 구분하여 수강 신청한 과목들의 명단을 볼 수 있다.

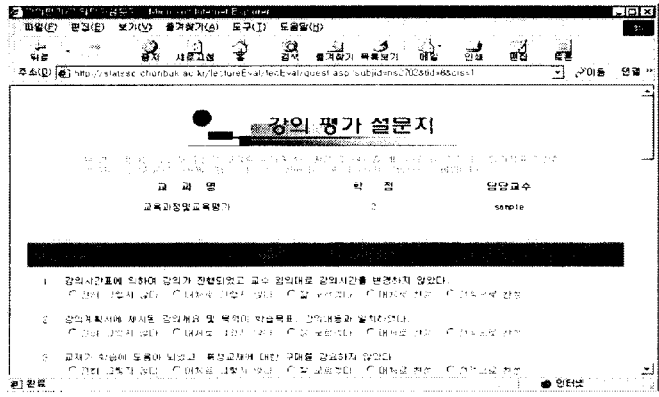

그림 3.1: 강의평가 설문지

평가 하고자 하는 과목을 선택하면 그림 3.1과 같은 강의평가 설문 항목들을 볼 수 있다. 학생들은 이 설문에 응답한 후 '완료'버튼을 누르면 해당과목의 강의평가가 완료되며 다른 과목의 평가를 계속할 수 있다.

그림 3.2는 강의평가 중 또는 완료된 후 강의평가 종합 관리자에게 강의평가 결과의 전반적인 상황을 주는 화면이다. 관리자는 일반 사용자(일반 교수)들과는 달리 어떤 의사결정을 하기 위한 종합적인 정보를 필요로 한다. 학생들의 참여율, 총괄적 결과, 과목별 또는 학년별 결과 등을 점검하고 관리할 수 있도록 구성되어 있다.

그림 3.3과 그림 3.4는 교과목 담당 교수가 볼 수 있는 결과 화면이다. 자신이 강의한 과목에 대한 평가 결과의 데이터는 물론 요약된 정보를 차트등과 함께 살펴볼 수 있으며 다른 과목들과 비교하여 볼 수 있다.

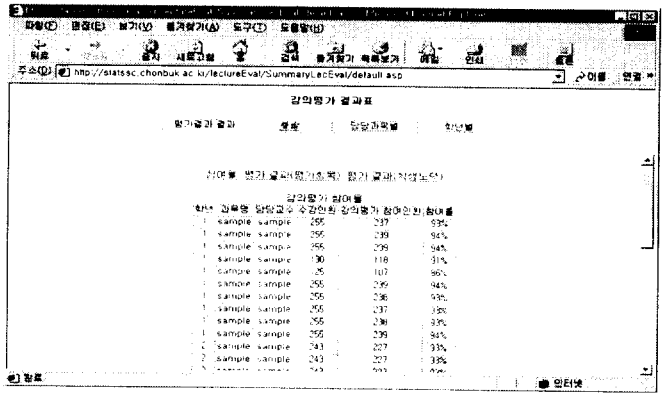

그림 3.2: 전체적인 결과표

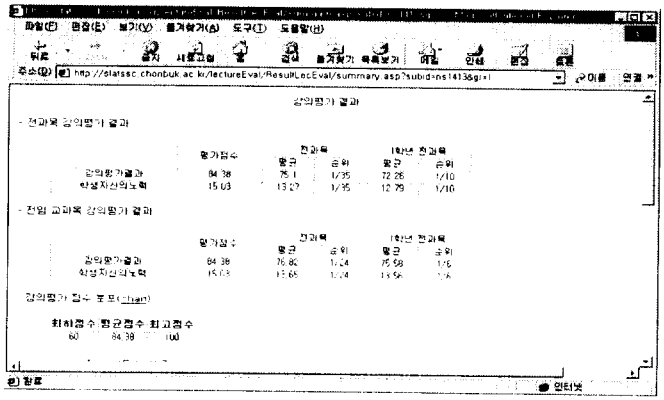

그림 3.3: 과목별 결과표

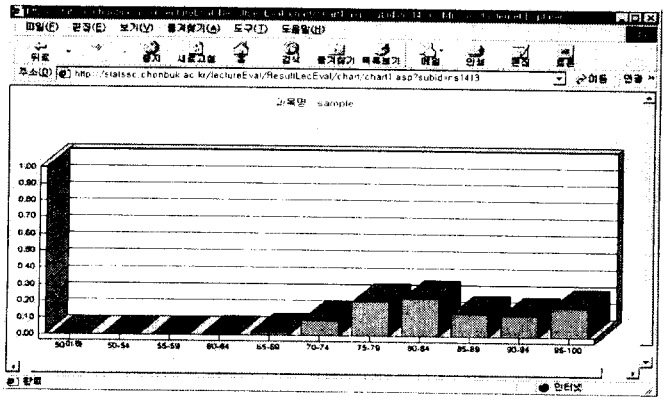

그림 3.4: 과목별 결과 Chart

4. 결론

사회가 발전함에 따라 정보(또는 데이터) 활용에 대한 중요성은 더욱 확대되고 있으며, 데이터 수집과 관리 분야 역시 중요한 문제로 부각되고 있다. 컴퓨터와 네트워크, 그리고 데이터베이스 관리 시스템 등은 이러한 분야를 더욱 발전시킬 수 있는 효율적인 도구이며, 이러한 기술의 효과적인 활용은 통계 이론 개발과 더불어 중요한 요소이다.

한편, 데이터를 다루는 어플리케이션 설계에 있어 통계적인 접근으로 이루어졌는가 아니면 전산학적인 관점으로 이루어졌는가는 많은 차이가 있으며, Chambers(1993)는 이러한 시스템 개발의 중요성을 설명하면서 통계학자들이 현재와 같이 이러한 연구에 계속 무관심하면 다른 분야의 학자들이 연구를 수행할 것이고, 그러면 통계학의 위치가 흔들릴 것임을 경고한다.

본 연구에서는 웹을 활용한 데이터 수집 및 관리 방법에 대해 살펴보고, 강의평가 시스템을 소개하였다. 여기에서 다룬 평가 결과와 같은 데이터는 설문 구성 항목에 따라 민감한 사안이 될 수도 있기 때문에 이용에 신중을 기해야 하며, 일반적인 설문조사나 표본조사에서 발생할 수 있는 결측치 처리 문제, 데이터 타당성 평가 문제, 응답자 비밀 보장에 대한 문제 등이 고려되어져야 할 것으로 생각된다. 데이터 수집 및 관리 방법에 대해서는 데이터 분석 및 분석 정보전달 기술등과 함께 체계적이고 구체적인 연구가 앞으로 많이 이루어질 것으로 생각된다. 이 분야는 현재 많은 관심을 받고 있는 OLAP, 데이터 마이닝, 다차원 데이터베이스 등에 대한 연구와 밀접한 관련이 있기 때문이다.

이번 연구는 웹을 이용하여 데이터 수집 및 관리와 동시에 그 결과를 분석하여, 사용자에게 정보를 제공하는 실제적인 데이터 활용에 관한 문제를 다루고 있으며, http://statssc.chonbuk.ac.kr/lectureEval에서 이용할 수 있다. 이러한 연구는 사회 환경 변화에 따른 데이터 분석의 온라인화, 실시간화의 요구에 부응하고, 사용자들에게 데이터 활용에 대한 새로운 인식을 심어주는데 도움이 될 것으로 기대된다.

참고문헌

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

[2] 허명회 (1999). 토론: 통계학, 새로운 모습의 탐색, 〈응용통계연구〉, 제12권 1호, 309-313.

[3] Chambers, J.M. (1993). Greater or Lesser Statistics: A Choice for Future Research, *Statistics and Computing*, Vol. 3, No. 4, 182-184.

[4] Carriere, S.J. and Kazman, R. (1997). WebQuery: searching and visualizing the Web through connectivity, *Computer Networks and ISDN Systems*, Vol. 29, 1257-1267.

[5] Fred, D., Thomas, B., Yih-Farn, C. and Eleftherios, K. (1996). WebGuide: Querying and navigating changes in Web repositories, *Computer Networks and ISDN Systems*,

Vol. 28, 1335-1344.

[6] Gillman, D.W., Appel, M.V. and LaPlant, W.P. (1996). Design Principles for a Unified Statistical Data/Metadata System, *Proceedings of the International Conference on Scientific and Statistical Database Management*, 150-155.

[7] Haas, P.J. (1997). Some Sampling and Estimation Methods for SQL Database, *Summer Research Institute on Data Mining*, http://www.research.microsoft.com/uwmsrdmi/abstracts.htm

[8] Higgins, J.J. (1999). Nonmathematical Statistics: A New Direction for the Undergraduate Discipline, *The American Statistician*, Vol. 53, No. 1, 1-6.

[9] Lenz, H.J. and Shoshani, A. (1997). Summarizability in OLAP and statistical databases, *Proceedings of the international conference on scientific and statistical database management*, http://www.lbl.gov/ arie/papers.

[10] Olken, F. (1993). Random Sampling from Databases, PhD Dissertation, University of California at Berkeley.

[11] Schwarz, C.J. (1997). StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling, *Journal of Statistics Education*, Vol. 5, No. 2, http://www.amstat.org/publications/jse

[12] Shoshani, A. (1997). OLAP and statistical databases: similarities and differences, *Proceeding of the ACM PODS*, 185-196.

[13] Short, T.h. and Pigeon, J.G. (1998). Protocols and Pilot Studies: Taking Data Collection Projects Seriously, *Journal of Statistics Education*, Vol. 6, No. 1, http://www.amstat.org/publications/jse

[14] 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.

[2000년 2월 접수, 2000년 6월 채택]

Data Collection and Management on the World Wide Web : Evaluating system for Lecture

Jeong Yong Ahn [1] Seung Hyun Choi [2] Kyung Soo Han [3]

ABSTRACT

Data collection, management, and analysis to furnish information are very important in these modern days. In this paper, we discuss the methods of data collection and management on the World Wide Web and introduce an evaluating system for lecture.

Keywords: Data collection and management; Web application; Evaluating system for Lecture.

1) Assistant Professor, Division of Computer Science and Information Communications, Seonam University.
E-mail: jyahn@tiger.seonam.ac.kr
2) Department of Computer Science and Statistics, Chonbuk National University.
E-mail: shchoi@stat.chonbuk.ac.kr
3) Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University.
E-mail: kshan@stat.chonbuk.ac.kr