The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 167-174

On the Design of Statistical Software in the Network Environment

Beom Soo Han[1], Jeong Yong Ahn[2] and Kyung Soo Han[3]

Abstract

Computer network provides a powerful infrastructure for information sharing and the development of the statistical software with new concepts. In this paper, we discuss the design concepts of the statistical software in the network environment.

Keywords : Statistical Software, Network Environment, Customer Relationship Management, Personalized Service, Statistical Database.

1. 서론

우리는 '현대 사회는 네트워크 사회이다'라는 말을 흔히 접하고 있다. 이 말은 네트워크(network) 기술이 현대 사회 전반에 얼마나 많은 영향을 미치고 있는지를 잘 반영해 주고 있다. 실제로 컴퓨터 네트워크(computer network), 정보 기술(information technology)과 같은 컴퓨터 환경의 급속한 발달은 교육과 연구, 그리고 비즈니스(business) 분야를 포함한 우리 생활의 전반에 많은 변화를 가져오고 있다. 교육에 있어 멀티미디어 컨텐츠(multimedia contents)를 포함하는 전자 교재(electronic text)와 각종 소프트웨어의 활용은 이미 일반화되어 가고 있는 현상이며, 연구에 필요한 많은 정보들이 웹(Web)을 통하여 공유되고 있다. 또 전자 결재, 전자 상거래 등을 통하여 비즈니스(business) 분야에서도 매우 큰 비중을 차지하고 있다.

이러한 환경은 통계학 분야에도 많은 변화를 가져오고 있다. 우수해진 컴퓨팅 성능(powerful computing power)에 힘입어 통계학자들은 Bootstrap, MCMC(Markov Chain Monte Carlo), 비모수적 분석 기법, 데이터 마이닝(Data Mining) 등과 같은 분야의 연구에 과거보다 쉽게 접근하고 있으며, 웹과 같은 네트워크를 통하여 실생활에서 발생하는 대용량의 데이터를 쉽게 접할 수 있게 되었다.

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

Cameron(1997)과 West 등(1998)에서 언급하듯이 컴퓨터 네트워크의 발전은 통계학 분야에 많은 기회와 새로운 문제를 제공하고 있다. 웹 상에서 발생하는 데이터로부터 정보를 추출하기 위한 웹 데이터 마이닝(Web data mining)은 그 좋은 예이며, 통계학의 기본적 영역인 데이터 수집과 관리 분야에 대한 체계적인 연구, 통계정보 서비스 방안, 데이터를 분석하기 위한 소프트웨어의 개발 및 활용 방안 등과 같은 여러 가지 연구 과제를 부여해 주고 있다.
본 연구에서는 일반화된 네트워크 환경을 고려한 통계 소프트웨어(statistical software)의 설계 방안에 관해 논의해 보고자 한다. 2장에서 기존 통계 소프트웨어의 이용 환경 및 문제점, 네트워크 환경의 장점 등에 대해 살펴보고, 3장에서 통계 소프트웨어를 설계할 때 고려해야 할 개념들을 논의한다.

2. 통계 소프트웨어와 네트워크 환경

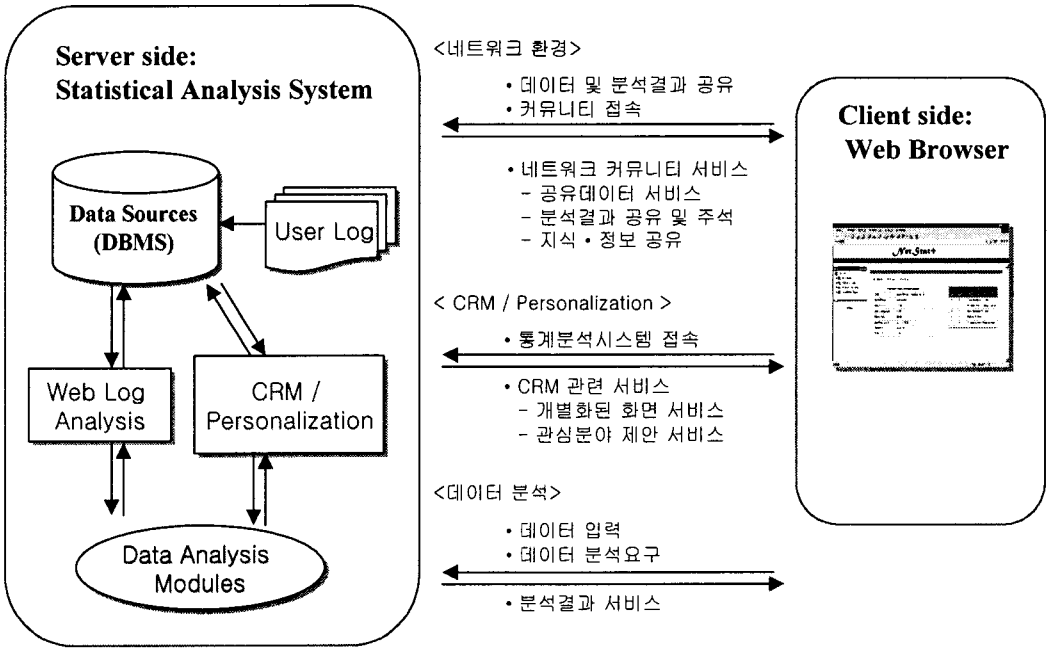
현재 많이 이용되고 있는 통계 소프트웨어로는 SAS, SPSS, S-Plus 등을 들 수 있다. 이러한 통계 소프트웨어들은 일반적으로 개인용 컴퓨터에 개별적으로 설치되어 운영되고 있으며 데이터를 다양한 형태로 입력받아 분석하는 것을 목적으로 한다. 예를 들어, 사용자로부터 데이터를 직접 입력받을 수도 있고, 미리 파일 형태로 저장되어 있는 데이터를 활용할 수도 있으며, 최근에는 데이터베이스에 저장된 데이터에 대한 분석도 가능하다. 그러나 이러한 소프트웨어들의 몇몇 문제점들이 새롭게 부각되고 있다.
첫째, 개인용 컴퓨터에 독립적으로 설치되어 사용되고 있는 통계 소프트웨어들의 경우에는 급변하는 환경을 반영하기에는 소프트웨어의 배포, 유지보수 및 관리 등에 어려움이 있다. 예를 들어, SAS의 경우 우리는 주기적으로 'setinit'과 같은 환경 파일을 갱신해 주어야 하며, 다른 소프트웨어들도 좀 더 최근의 버전(version)을 이용하고자 할 때 각 사용자마다 소프트웨어 자체를 다시 설치해야 한다.
둘째, 데이터베이스에 저장된 데이터를 이용하기 위해서는 데이터베이스의 테이블에 대한 구조를 명확히 알아야 한다. 최근에 많은 기업들은 데이터 웨어하우스(Data Warehouse)를 구축하고 있다. 이것은 아주 단순하게 표현하면, 데이터를 데이터베이스에 저장하여 관리, 활용한다는 것이다. 이와 같은 데이터 환경을 감안할 때, 현재 이용되고 있는 통계 소프트웨어 환경은 사용자가 데이터베이스에 저장된 테이블의 구조, 데이터베이스 접근 방법 등의 자세한 정보를 알아야 하는 제약이 있으며, 데이터베이스 관리 시스템(DBMS)이 지원하는 다양한 분석관련 기법들의 이용도 쉽지 않다고 할 수 있다.
셋째, 사용자 개개인의 입장에서 그 동안 자신이 이용한 것에 대한 어떠한 히스토리(history)도 없으며, 다른 이용자들과 정보의 공유가 불가능하다.

이러한 문제점들은 사용자들에게 불편을 주는 것은 물론 데이터 분석이 어렵고 복잡하다는 인식을 심어주고 있다. 따라서 이러한 문제들을 해결할 수 있는 ASP(Application Service Provider), 개인화된 서비스(Personalized Service), 데이터 및 분석 결과의 공유 등과 같은 새로운 개념을 갖는 통계 소프트웨어의 설계와 개발이 요구되고 있다.
현대 사회의 정보 기술의 발달은 이러한 요구를 해결할 수 있는 적절한 환경을 제공해 주고 있다. 특히 컴퓨터 네트워크의 발달에 힘입어 위에서 제시한 새로운 개념들을 통계 소프트웨어에 적용하는 것이 가능해지고 있으며, Hahn(1985)의 연구는 통계 소프트웨어 설계 및 개발에 좋은 모델(model)이 되고 있다
일반적으로 네트워크를 활용한 소프트웨어는 다음과 같은 장점을 갖는다.

· 소프트웨어의 유지보수 및 배포의 용이.
· 데이터 관리의 편리성과 시공간에 무제약
· 정보지식 공유 편리

3. 통계 소프트웨어의 설계

새로운 개념의 통계 소프트웨어는 데이터의 분석뿐만 아니라 지식(knowledge) 또는 정보(information)의 공유 개념이 필수적이라고 생각한다. 이러한 개념은 컴퓨터 네트워크 환경을 이용하지 않고는 실현하기가 어렵다. 따라서 한경수와 최숙희(2000)에서 지적하는 바와 같이 네트워크 환경에 기반한 통계 소프트웨어 설계의 새로운 패러다임이 요구된다.



<그림 1> 통계 소프트웨어 구성도

<그림 1>은 네트워크 환경을 고려한 통계 소프트웨어 설계에 관한 대략적인 구성도이다. CRM(Customer Relationship Management)/개인화된 서비스(personalized service), 그리고 데이터 베이스(database)의 활용 등 전체 시스템 구조에서 서버 측의 역할이 많이 강조되고 있는 것을 그림에서 볼 수 있다. 각각의 개념들은 통계 소프트웨어 설계 및 구현에 있어서 매우 중요하게 고려해야할 개념이다. 본 연구에서 제안하고자 하는 이러한 설계 개념들을 각 항목별로 자세히 살펴보면 다음과 같다.

3.1 ASP(Application Service Provider) 개념의 도입

네트워크 기술의 발전은 소프트웨어를 임대하는 개념의 ASP 서비스를 가능하게 하였다. ASP는 사용자가 필요한 소프트웨어들을 개인의 컴퓨터에 설치할 필요 없이 고성능의 서버에 네트워크를 통해 접속하여 사용할 수 있는 서비스를 말한다. 이러한 ASP는 소프트웨어를 서버에서 총체적으로 제공하고 관리함으로서 사용자가 소프트웨어를 쉽게 이용하는 것을 가능하게 한다. 결과적으로 사용자는 자신의 시스템에 이상이 생겼다거나 또는 최근 버전(version)을 이용하기 위해서 소프트웨어를 다시 설치할 필요가 없기 때문에 매우 편리하게 소프트웨어를 이용할 수 있고 자신의 시스템 환경의 제약으로부터 해방될 수 있다. 예를 들면, 최근에 개발된 소프트웨어를 이용하기 위해서 자신의 컴퓨터를 값비싼 컴퓨터로 바꿀 필요가 전혀 없다.
ASP 환경하에서 사용자는 웹 브라우저 등을 통해 서비스 서버에 접속하여 필요한 소프트웨어를 사용할 수 있게 된다. 따라서 소프트웨어의 유지 보수가 필요 없이 네트워크와 웹 브라우저만 있다면 언제 어디에서나 동일한 환경의 소프트웨어를 사용할 수 있으며 다음과 같은 장점이 있다.

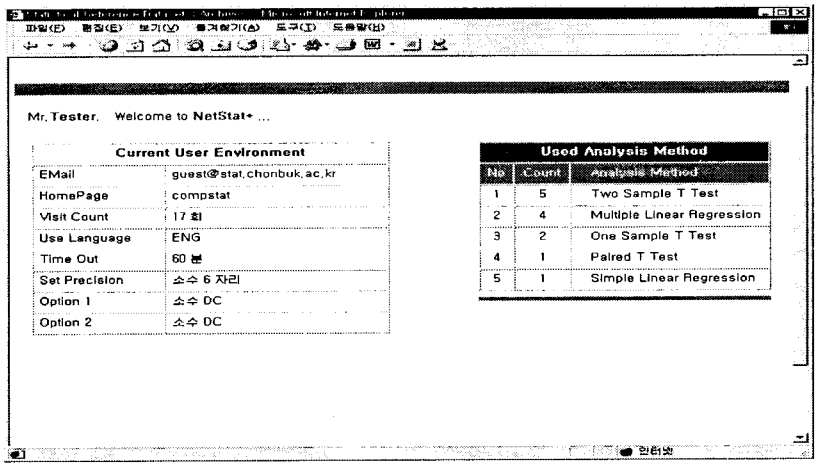
(ⅰ) 사용자의 편리
- 장소에 무관하게 서버에 접속하여 동일한 환경으로 소프트웨어를 사용
- 데이터를 서버의 일정한 장소에 보관할 수 있으므로 데이터 관리가 편리

(ⅱ) 유지 보수 및 관리의 편리
- 소프트웨어의 설치, 원활한 업그레이드 등의 유지보수 관리가 불필요
- 별도의 웹 서버 등의 관리가 불필요
- 체계적인 데이터 백업 지원
- 상시적인 시스템 모니터링 지원
- 전문가에 의한 소프트웨어 및 시스템 관리
- 확장성 요구의 신속한 대응
- 최신, 최고의 정보 기술을 제공

(ⅲ) 비용 절감
- 초기 소프트웨어 도입 비용 감소
- 사용한 만큼의 사용료 지급으로 비용 감소

3.2 CRM/개인화된 서비스 개념의 도입

CRM/개인화된 서비스는 고객 관리 및 서비스에 대한 개념으로 비즈니스 분야에서 많이 이용되고 있다. 이들 개념의 주 목적은 고객 개개인의 특성을 파악하고, 그 특성을 이용하여 고객에게 가장 적합한 서비스를 제공한다는 것이다. 예를 들면, 각 개인의 선호도 및 성향에 맞추어진 웹 페이지를 제공하거나 별도의 지정된 맞춤형 서비스를 제공하는 것이다.
통계 소프트웨어를 이용하는 사용자 역시 서비스를 이용하는 고객으로 간주할 수 있다. 그러므로 통계 소프트웨어를 설계할 때 사용자 개개인과 원활한 상호 관계를 유지하고 편리성을 제공하기 위해 CRM/개인화된 서비스를 지원할 수 있는 방안을 고려하는 것이 좋다. <그림 2>는 사용자가 통계 분석 서비스를 제공하는 시스템에 접속했을 때, 사용자의 기존 사용 이력과 주로 이용하는 분석 기법들을 보여주는 개인화된 시스템의 예이다.



<그림 2> 개인화된 서비스의 예

이러한 CRM/개인화된 서비스 개념의 도입을 통하여 사용자들에게 통계 소프트웨어 또는 데이터 분석이 어렵다는 인식을 해소시킬 수 있으며 사용자들에게 자신과 관련된 정보를 제공할 수 있다. 이것은 사용자 입장에서 매우 바람직한 서비스라 할 수 있다. 자신이 이용한 내용에 대한 히스토리(history)의 제공은 추후 소프트웨어 이용에 많은 도움을 줄 수 있기 때문이다. CRM/개인화된 서비스 개념을 활용하는 세부적인 방법으로는 다음과 같은 사항을 고려할 수 있다.

(ⅰ) 사용자 인터페이스의 개인화 서비스
- 사용자가 주로 이용하는 분석기법 위주의 화면 구성
- 사용자 군집 분류 및 개별화된 사이트 구성
- 사용자의 성향 및 수준을 고려한 통계 분석결과 출력

(ⅱ) 추천 및 제안 기능(메시징 서비스: e-mail)
- 개인별 관심분야 데이터와 관련 분석 기법 소개
- 향상된 고급 분석 기법의 소개 및 제안

(ⅲ) 협업(Collaboration) 기능
- 공동 목표를 가진 팀(조직)의 운영 지원
- 정보 지식 및 자료의 공유

3.3 데이터베이스의 활용

현재 이용되고 있는 대부분의 통계 소프트웨어는 사용자 개개인이 자신의 컴퓨터에서 이용할 수 있도록 구성되어 있기 때문에 데이터 저장 및 이용 방법에 파일 형태가 적합하였다. 그러나 ASP, CRM/개인화된 서비스 개념의 도입과 데이터 및 분석 결과의 공유 등을 고려할 때 데이터베이스의 활용 방안은 필수적인 설계 개념이다. ASP 형태의 소프트웨어는 서버에서 사용자들의 데이터를 관리해 주어야 하며, 사용자들간에 데이터를 공유하여 활용할 수 있도록 데이터를 분야별, 규모별, 형태별로 체계적으로 관리해야 한다. 파일을 이용하여 데이터를 관리한다면 이러한 일들은 어렵고 복잡하다.

또 CRM/개인화된 서비스를 지원하기 위해서는 먼저 사용자들에 대한 데이터를 확보, 저장하고 있어야 하며, 필요한 정보를 적절한 시점에 추출하여 활용할 수 있어야 한다. 사용자 개개인이 주로 이용하는 분석 방법이나 내용에 대한 히스토리(history) 데이터도 필요하다. 따라서 네트워크 환경에서 통계 소프트웨어를 효율적으로 이용하기 위해서는 데이터베이스의 사용이 필수적이며, 활용 방법으로는 다음과 같은 사항들을 고려할 수 있다.

- 데이터 및 분석 결과 저장 및 공유
- 데이터 분석 결과에 대한 여러 전문가들의 해석을 보관하고 공유
- 사용자들의 분석 방법, 분석 절차 등에 대한 히스토리 저장
- CRM/개인화된 서비스 지원을 위한 웹 로그 정보 및 요약 자료의 저장
  - 인공 지능 기법을 활용하는 다양한 지능형 분석 시스템의 구현에 필요한 대량의 지식 데이터 보관 및 관리

3.4 사용자 인터페이스 구성

사용자 인터페이스(user interface)란 사용자와 컴퓨터 사이의 상호 정보 교환을 지원하는 기능으로 인터페이스의 구성에 따라 사용자의 편의성은 크게 영향을 받게 된다. Chambers(2000)는 사용자 인터페이스를 구성할 때에 사용자들에 대한 이해를 바탕으로 설계할 것을 제안하고 있으며, Dix(1999)는 사용자 인터페이스의 중요한 디자인 논점으로 다음의 4가지를 제안하고 있다.

- 사용할 주요 대상에 대한 파악
- 사용자 개개인에 대한 구분과 이해
- 사용자가 접속하여 시작하는 위치에 대한 처리
- 사용자가 접속을 종료하고 떠나는 위치에 대한 처리

일반적으로 사용자 인터페이스는 사용자가 많이 접하고 사용해오던 방식으로 구현해야 친밀도를 높일 수 있고 직관적 사용이 가능하게 된다. 따라서 기존의 소프트웨어들과 유사하게 사용자 인터페이스를 구성하는 것이 바람직하다. 예를 들어 사용자에게 친숙한 SAS, SPSS 그리고 스프레드시트(MS Excel, Lotus 등) 프로그램의 데이터 입력 방식이나 메뉴의 형태를 선택적으로 사용할 수 있도록 구성하는 것이 좋다.
또 다른 중요한 사항으로 Web과 같은 분산 시스템에서는 각각의 처리 기능들이 어디에 위치하고 있는지에 따라 커다란 성능의 차이를 가져오므로 이러한 사항도 사용자 인터페이스를 구성할 때 함께 고려되어야 한다. 그리고 사용자들이 Web의 복잡성을 느끼지 않도록 프로그램 상의 현재 위치나 초기의 위치로 쉽게 돌아갈 수 있도록 고려하여 구성해야 한다.

3.5 지식 공유 개념의 도입

위에서도 언급한 바와 같이 현재 이용되고 있는 대부분의 통계 소프트웨어는 사용자 개개인이 자신의 컴퓨터에서 이용할 수 있도록 구성되어 있기 때문에 다른 사용자들과의 정보를 공유하는 것은 불가능하다. 예를 들어, 어떤 사용자가 입력하여 분석에 이용한 데이터를 다른 사용자는 이용할 수가 없으며,

분석 결과도 볼 수가 없다. 다시 말하면 지식의 공유 개념이 전혀 없다. 사용자 입장에서는 자신이 분석하고자 하는 데이터와 비슷한 형태의 데이터와 분석 결과를 제공받을 수 있으면 데이터 분석 및 해석이 훨씬 쉽고 간편할 것이다.
따라서 기존의 독립적으로 사용하던 통계 소프트웨어에서 탈피하여 각 사용자들의 데이터 및 분석 결과를 데이터베이스에서 관리하고 공유할 수 있도록 지원하는 개념이 필요하다. 이러한 개념의 도입은 사용자들의 통계 분석에 대한 오류를 줄이고 통계 분석 기법을 자연스럽게 적용하고 활용하는데 도움을 줄 수 있다.
지식 공유를 위한 방법으로는 사용자의 관심분야 별로 관련 데이터, 분석 기법, 분석 사례, 그리고 통계 전문가의 조언 등의 지식을 공유할 수 있는 네트워크 커뮤니티를 형성할 수 있도록 지원하는 것이다.

3.6 통계 분석 및 그래픽 라이브러리 개발

데이터 분석 및 그래픽 라이브러리는 통계 소프트웨어의 구현에 있어서 많은 시간과 노력이 요구되며 핵심이 되는 부분이라고 할 수 있다. 이러한 라이브러리의 특징은 관련된 이론이 바뀌지 않는 한 구현 알고리즘 상의 변화가 거의 없다는 것이다. 따라서, 데이터 분석 및 그래픽 라이브러리는 한 번 개발한 후에는 별도의 수정이 필요치 않게 되므로 재사용성이 매우 높은 부분이라 할 수 있다.
이러한 라이브러리 설계시 개발 언어나 도구 등의 변화에 따라 영향을 받지 않도록 다음과 같은 사항들을 고려하여 설계 및 구현해야 한다.

- 재사용을 위한 객체 지향(Object Oriented Programming) 개념과 모듈 조립 개념의 컴포넌트 개발 방법론
- 데이터 분석, 그래픽 라이브러리, 그리고 인터페이스 등의 분리
- 새로운 분석기법의 유연한 추가 방식

3.7 기타 고려 사항

이상에서 살펴본 것 외에 추가적인 고려 사항들을 살펴보면 다음과 같은 것들이 있다.

- 구조화된 문서를 정의하는 XML
- 객체지향 프로그래밍을 위한 디자인 패턴(Design Pattern) 기법
- 대용량 데이터에 대한 분석 및 처리 기법
- 데이터 분석 실습 등을 위한 예제 및 통계 이론 등 관련 문서 서비스

4. 결론

본 연구에서는 네트워크 환경에 기반하여 통계 소프트웨어를 설계할 때 고려해야할 몇몇 개념들을 제안하고 이러한 개념들을 간단히 정리해 보았다. 물론 이러한 개념 이외에도 많은 다른 요소들이 필요할 것으로 생각되며, 본 논문에서 제안된 개념은 극히 일부분일 수도 있다.
그러나 한 가지 분명한 것은 급속히 발전하는 정보 기술은 현대 사회의 모습을 바꿔가고 있으며 소프트웨어의 개발에도 과거와는 다른 개념의 도입을 요구하고 있다는 사실이다. 여러 다른 소프트웨어들과

비교하여 볼 때 통계 소프트웨어에는 지식의 공유 개념이 특히 필요하다고 생각한다. 왜냐 하면 통계 소프트웨어를 이용하기 위해서는 통계적 분석 기법들에 대한 전문적인 지식이 요구되기 때문이다.
본 연구에서 제안된 통계 소프트웨어 설계 개념의 실제적인 활용을 통하여 사용자들에게 편리성을 제공함은 물론 통계 소프트웨어 또는 데이터 분석이 어렵다는 인식을 덜어주는데 도움이 될 것으로 기대한다. 또한 개발된 통계 소프트웨어를 교육적인 목적에 활용함으로서 학습자에게 통계 데이터 분석에 대한 이해를 도울 수 있을 것이다.

References

[1] 한경수, 최숙희 (2000), Some thoughts for the design of new statistical services in network society, Statistical Research using Internet and Computer Conference, Rikkyo Univ., Tokyo, Japan.

[2] Cameron, M. (1997), Current influences of computing on statistics, Proceedings of The Ninth Korea and Japan Joint Conference of Statistics: Multivariate Analysis and Computing-KJCS'97, December 5-6, 1997

[3] Chambers, J. M. (2000), Users, Programmers, and Statistical Software(invited paper), The ASA Journal of Computational and Graphical Statistics, September 2000, Vol. 9, No. 3

[4] Dix, A. (1999), Design of User Interfaces for the Web (invited paper), User Interfaces to Data Intensive Systems-UIDIS'99, Edinburgh 5th - 6th September 1999

[5] Hahn, G.J. (1885), More Intelligent Statistical Software and Statistical Expert Systems: Future Direction, The American Statistician, Vol. 39, No. 1, 1-8.

[6] 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

[2001년 7월 접수, 2002년 1월 채택]