Some thoughts for the design of new statistical services in network society

Kyungsoo, Han (Chonbuk National University, Korea)
and
Sookhee, Choi (Woosuk University, Korea)

Abstract

The spectacular growth of computer technology is obvious to all. Perhaps the ubiquitous personal computer is in some ways the least significant aspect of this revolution. One cannot appreciate the important value of information technology before one has become accustomed to the use of it. It is time to design the new statistical services in network society. In this talk, it will be addressed some aspects to improve old statistical package.

1. Introduction

The real world has primarily provided us with the more complex data rather than normal, independent identically distributed data like as we have treated in the usual statistical text. For statistical data analysis with available computing in the past, one has seemed to inevitably assume normality and linearity in statistical model with some approximation sense.

Computing power has been doubling every year, and this trend is safely predicted to continue for the next several years. The modern statisticians with powerful computing ability have a general feeling that these assumptions are basically unrealistic and more complex statistical models are much more likely to be required to accurately describe reality. They have studied many computer-based methods such as bootstrap, nonparametric regression, Markov Chain Monte Carlo, etc., but which could not have existed without recent powerful computers. What is undeniable is that the advent of more advanced computing environment will make it possible to study the statistical methods that are unthinkable at this time. The rapid development of computer has made statistics one of the most exciting professions.

We have often, at least in our country, seen that many statistical consultants are frequently using the classical methods rather than newly devised ones, which have been programmed by using S-PLUS or MATLAB. Elder and Pregibon (1996) remarked that many of those never made it into commercial statistical package systems such as, SAS, SPSS and therefore never made it into the mainstream of methods used by non-statisticians. It seems to us that one of the most important things that we statisticians should solve is the leadership and direction in using these methods in practice.

In this talk, we discuss some issues about the design of new statistical services for the cyber community with special statistical purpose in the computer network. Section 2 deals with some desirable properties of statistical services. In section 3 we introduce a prototype service, called NetStat, as an implementation of the design. And we conclude that it is time to renovate the way of designing and implementing the statistical software.

2. Statistical services in network society

What do we think about the features of computing environment in this era in designing the main statistical software such as, SAS, S-PLUS? In our opinion, it was probably assumed that a person should exclusively use the statistical software installed in one computer with one's knowledge about computer and statistics. In this philosophy of software design, it was greatly emphasized the speed of computation and deeply considered in designing the software. But even now we often see that the bulk of software have been designed without considering the importance of computer network, Internet or World Wide Web, being the culprit of the Information Age.

Friedman (1997) said that every time a technology has increased the effectiveness by a factor of ten, one should completely rethink how to apply it. He explained it with one example of the historical progression form walking to driving, and even to flying. In each stage, the speed has increased roughly by a factor of ten. However, such purely quantitative increase has completely reoriented our thinking on the use of transportation in our society.

Does a computer with the Internet connection increase in effectiveness by a factor of ten compared to a personal computer without networking? Obviously the answer is yes and many people will probably say that the effectiveness certainly is greater than hundreds times and may not be measured if the wireless network is also considered. Emphasizing the importance of network, this society may be now called the network society because of being connected so many people and so many computers through computer networking. In the network society, it may be thought one of the most important features to collaborate with other people, through the network, in accomplishing a task or in solving a problem.

In the mainframe age, a computer vendor should have made everything including hardware, operating system and some application software. From the IBM compatible PC age, we attained the valuable experience of standardizing and assembling hardware and software components. It has been proved the component-based architecture to be the fast, easy and reliable method of developing applications, and its importance will be emphasized more and more in network age.

We have been concerned about so much, but more or less useless, information related computer operation and maintenance, which are indifferent to our main interest. If our computer is certainly connected to server computer installed many application programs, which providers constantly maintain and upgrade, we may take advantage of the application service. It may be called a *statistical service* to be able to use statistical software through network

One of the earliest attempts at disseminating statistical methods over the Web is Xlisp-Stat. The UCLA Electronic Textbook (http://ebook.stat.ucla.edu/textbook) is an outstanding example of Xlisp-stat at work. A number of vendors have begun developing Web-based interface to their existing computing platforms, including XploRe (http://www.galaxy.gmu.edu/~xplore) and S-PLUS. Statlets (http://www.statlets.com) and WebStat (http://www.stat.sc.edu/webstat/) are the statistical computing packages written in the form of a Java applet for use on the World Wide Web.

The Omega project (http://www.omegahat.org) began in 1998, through discussion among designers responsible for three current statistical languages (S, R, Lisp-Stat), with the idea of working together on new directions with special emphasis on web-based software, Java, the Java virtual machine, and distributed computing. Omegahat software is being developed in the project and emphasizes a component-based approach. This means that not all modules need to be implemented in Java. CORBA (Common Object Repository Broker Architecture) is a system to run programs in a distributed heterogeneous environment. It provides a method for calling functions whose language and host processing machine are unknown and irrelevant to the calling routine.

The XML (extensible Markup Language) has been recognized as a standard of information delivery between applications in network environment. The MathML (Mathematical Markup Language) as an application of XML is intended to facilitate the use and re-use of mathematical and scientific content on the Web, and for other applications such as computer algebra systems, print typesetting.

In current HTML documents, all mathematical expressions have been represented as images and lose the flavor of semantic. If mathematical expressions in statistics are represented as MathML, some statistical knowledge may be programmed and saved in database. It will be necessary for statistics education on the Web to support XML. A StatML (Statistical Markup Language) may come into being to represent the standard result of statistical analysis. From the above viewpoint, we may have to keep an eye on Microsoft's .Net Framework, which emphasizes the common language runtime engine and XML.

3. A Prototype – NetStat (http://compstat.chonbuk.ac.kr/netstat)

First of all, it should be remarked that main paradigm of software development in network environment is 3 or multi-tier architecture, which consists of the presentation layer, business layer and data layer. In the presentation layer, everything of user interface is related and classified as thin or fat client according to existence of application in client. The computational logic of an application or the business logic of a corporation is not frequently changed and requests the fast transaction in many cases and thus usually operated in server machine. In some cases, it demands the business layer to be divided into multi-tier layer. The usual database management system is frequently used in data layer and manages especially all data pertinent to the application.

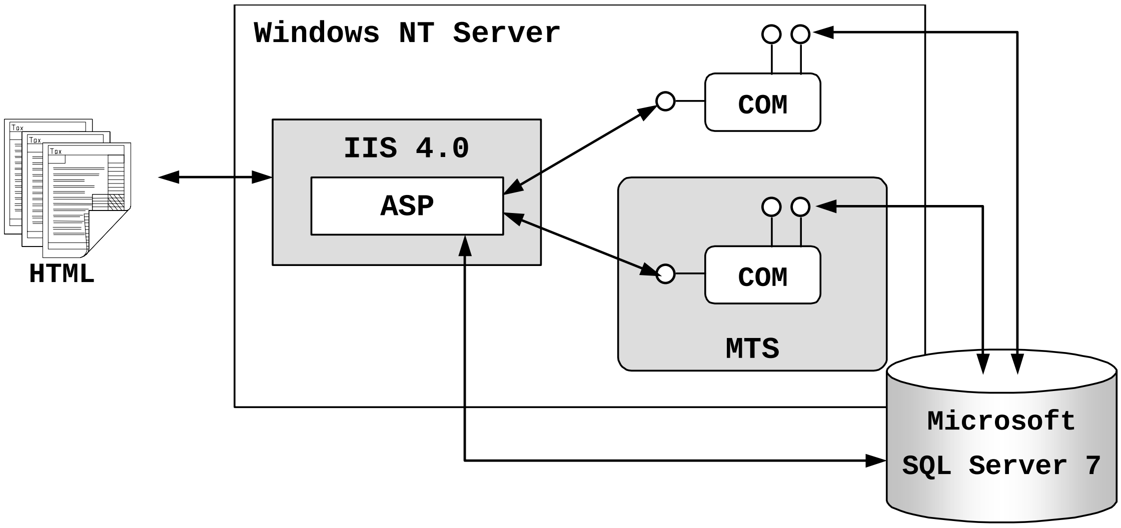


Figure 1: 4-tier architecture

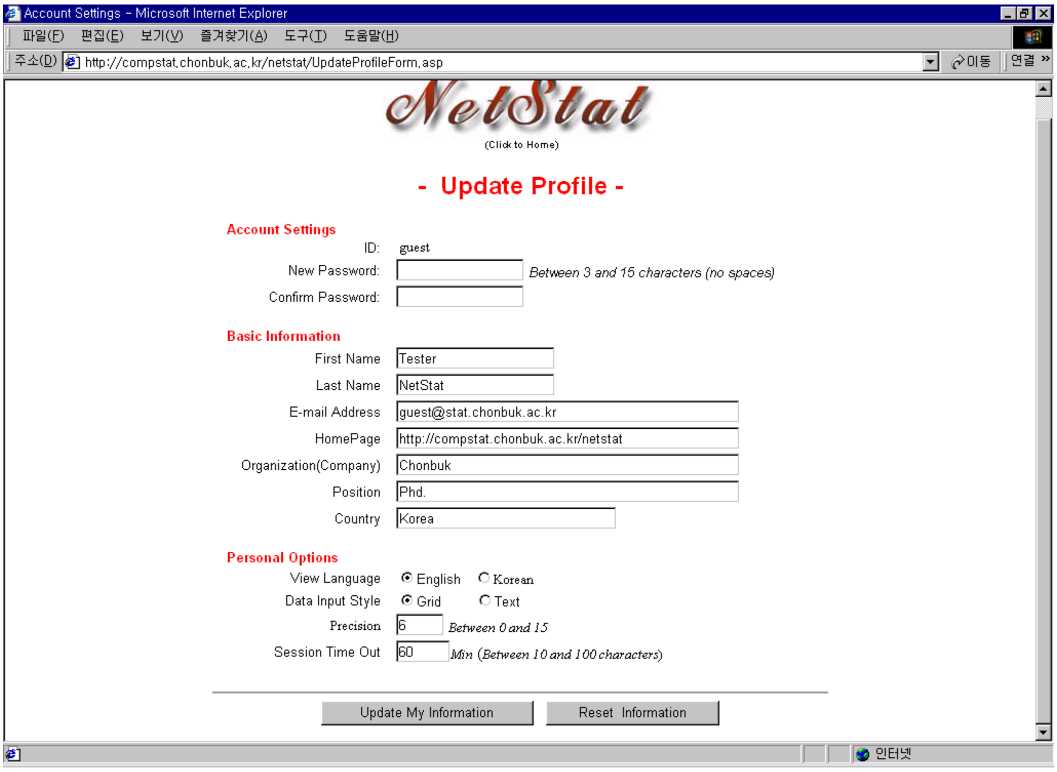
We are developing a prototype of statistical service with the 4-tier architecture shown in Figure 1. The presentation layer is a HTML web page. The request of user's web page is delivered to web server, Microsoft's Internet Information Server. Active Server Pages (ASP) engine delivers it to the application server, Microsoft Transaction

Server (MTS), to acquire the transformed information and generate a dynamic web page. MTS controls software modules, so called COM components, which carry out some statistical computation and select data from database. The tools used in this development are summarized in Table 1.

Layer	Function	Tool
Presentation	Web Page Authoring	Frontage, Visual InterDev, Active Server Pages(VB Script)
Business	COM programming	Microsoft Visual C++
	Library	IMSL
Data	DBMS	Microsoft SQL Server

Table 1: Developing tools

Figure 2 shows how users set up the preferences according to one's computer environment such as, Microsoft Office 2000, language and numerical precision, etc. Both Korean and English are currently supported, but other language can be easily added because the web page can be dynamically generated from the terminology dictionary stored in database.


Account Settings - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
주소(D) http://compstat.chonbuk.ac.kr/netstat/UpdateProfileForm.asp
이동
연결
NetStat
(Click to Home)
Update My Information
Reset Information
인터넷

- Update Profile -

Account Settings

ID: guest

New Password: Between 3 and 15 characters (no spaces)

Confirm Password:

Basic Information

First Name Tester

Last Name NetStat

E-mail Address guest@stat.chonbuk.ac.kr

HomePage http://compstat.chonbuk.ac.kr/netstat

Organization(Company) Chonbuk

Position Phd.

Country Korea

Personal Options

View Language ● English ○ Korean

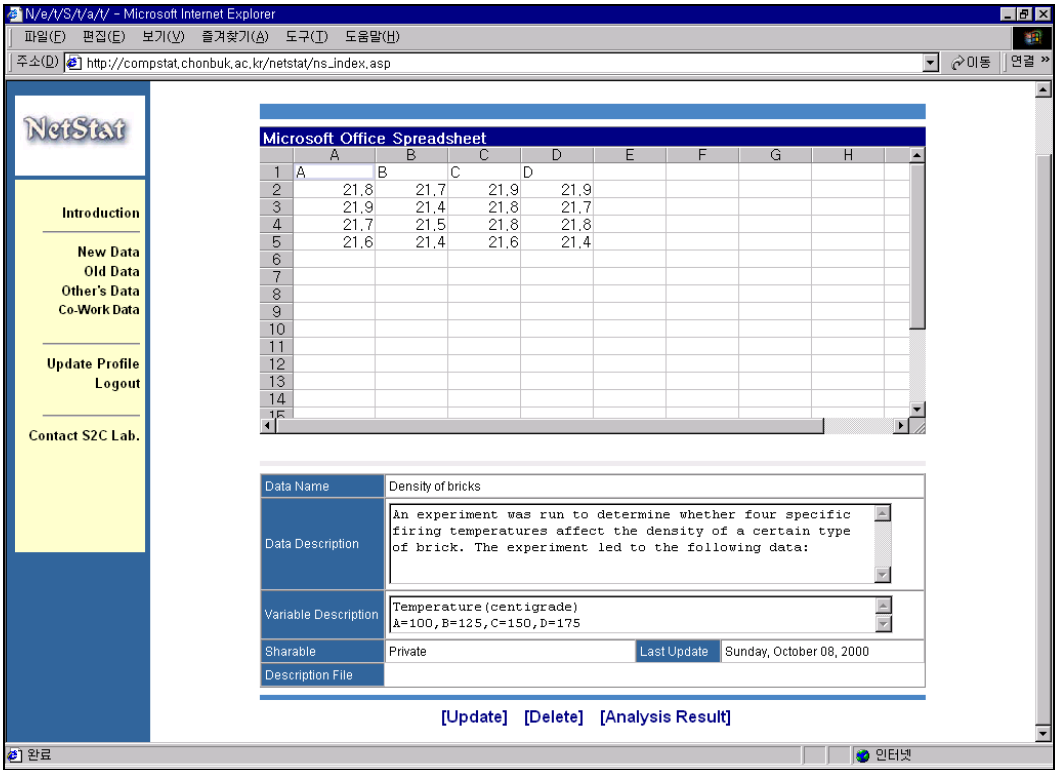
Data Input Style ● Grid ○ Text

Precision 6 Between 0 and 15

Session Time Out 60 Min (Between 10 and 100 characters)

Figure 2: Personal environment setting

A user can input his data according to the statistical techniques such as, t-test, simple regression and one-way ANOVA. Some metadata including data description, variable description and other file can be included to promote the ease of understanding data. Figure 3 shows an example of data input.


N/e/t/S/t/a/t/ - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
주소(D) http://compstat.chonbuk.ac.kr/netstat/ns_index.asp
이동
연결
NetStat
Introduction
New Data
Old Data
Other's Data
Co-Work Data
Update Profile
Logout
Contact S2C Lab.
Microsoft Office Spreadsheet

	A	B	C	D	E	F	G	H
1	A	B	C	D				
2	21.8	21.7	21.9	21.9				
3	21.9	21.4	21.8	21.7				
4	21.7	21.5	21.8	21.8				
5	21.6	21.4	21.6	21.4				
6								
7								
8								
9								
10								
11								
12								
13								
14								

Data Name	Density of bricks		
Data Description	An experiment was run to determine whether four specific firing temperatures affect the density of a certain type of brick. The experiment led to the following data:		

Variable Description	Temperature(centigrade)
	A=100,B=125,C=150,D=175		

Sharable	Private	Last Update	Sunday, October 08, 2000
Description File			

Figure 3: Data input

As a statistical graph tool, the graph component of Microsoft Excel 2000 or Java applet are allowed to selected by users. Considering the complex shape of statistical graph and the variant support of client machines, much more graphic Java applets should be developed in the future.

Note that the result of statistical data analysis can be also used as input in other application program and saved in database to be used in another analysis. To achieve this purpose, the analysis results are necessary to be represented as XML documents with statistical meaning.

One of the recent trends that we notice is that it is rapidly increasing the use of statistical methods in the analysis of data and education in the various fields such as government, industry, and academic institutions. However, it is quite often that the statistical methods are misused due to the lack of understanding of statistics and adopting an appropriate statistical model. Therefore, we believe that diverse statistical services rather than statistical portal site should be developed in the near future to remedy the misuse of statistical methods by opening the result of analysis and exchanging opinions among users including experts in statistics.

4. Conclusion

The design of statistical service, the next generation of statistical software, should be sincerely considered to reflect the effectiveness of network environment. This point has been mostly overlooked in existing statistical software because the network was not well introduced as today. It may not be possible to make useful one by simple extension. We have not seen a statistical service suitable in network society. It is time to start the development of this statistical service. The relevant computer topics considered as the research tools include numerical linear algebra, numerical optimization, data structures, algorithm design, machine architecture, programming methodology, database management, parallel architectures and programming, etc.

An electronic statistical text appropriate to network society may be developed using some propositions here, saving the steps of each learning process, and being analyzed by teachers.

We have used the relational database, but we had some difficulties to represent data structure and metadata. Chambers (2000) remarked that a single observation on some variables could correspond to multiple observations on some other variables. In this case, the relational data model does not work well. It may be more suitable to consider an object-oriented database in the future study.

References

1. Chambers, J. M. (2000). Users, Programmers, and Statistical Software, Journal of Computational and Graphical Statistics, 2000,9(3).
2. Elder, J. F. and Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases, In Advances in Knowledge Discovery and Data Mining, Ed. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, The MIT Press, 83-116.
3. Friedman, J. H. (1997). Data Mining and Statistics : What's the Connection?, 29th Symposium on the Interface : Computing Science and Statistics, http://www.stat.rice.edu/interface97.html.