The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 39-51

Internet Survey System Construction and Utilization of Web log Data[1)]

- APM[2)] SURVEYOR 1.5 -

Kyung Joon Cha[3)], Jae Woo Jung[4)]

Abstract

In this paper, we propose a poll system on the internet. This poll system makes use of APM which are useful in web and PC using for a server. These tools are all free to obtain, so we can construct this system with the minimum computing environment and the minimum cost. We mention merits and demerits about internet survey and propose how to overcome using this system, and utilize web log data to get an additional information of panel. Finally, we suggest extensibilities of internet survey and the proposed system

Keywords ; Internet Survey, Web usage mining, Apache, PHP, MySQL

1. INTRODUCTION

컴퓨터 보급의 증가와 인터넷 인구의 확산은 쇼핑몰의 개념부터 마케팅의 개념까지 바꾸어 놓았다. 이러한 전반적인 추세에 의해 여론조사의 방법도 기존의 면접조사, 우편조사, 전화조사 등의 방법에서 인터넷 서베이(survey) 방법이 주목받게 되었다. 기관 및 단체에서도 설문 조사부분에서 웹(web)을 이용하는 경우가 많아지고 통계를 공부하지 않은 사람도 자신의 사이트에 소규모적인 서베이 프로그램을 전시해 놓고 있다. 하지만 이는 극히 단순한 빈도 분석에 불과하고 상업적인 사이트를 운영하는 단체조차도 몇몇을 제외하고는 실시간으로는 빈도분석에서 크게 벗어나지 않는 방법만을 제시하고 있다. 이는 실시간으로 분석하기에는 프로그램화 하기 어려운 부분과 서버(server)운영에 따른 고정 비용의 증대 그리고 상업성 가치에 따른 결과로 보여진다. 이에 본 논문에서는 최소한의 비용과 관리자에게 편리한 환경으로 좀 더 나은 프로그램을 만들어 웹 상에서 실행해보고 설문 과정 중 로그 파일(log file)을 분석, 결과를 보여주고자 한다. 웹 언어로 만들어

1) This work was supported by Hanyang University, Korea, made in the program year of 2001
2) Apache+PHP+MySQL
3) Professor, Department of Mathematics, Hanyang University, Seongdong-ku, Hangdang-dong 17, Seoul, E-mail : kjcha@hanyang.ac.kr
4) Graduate Student, Department of Mathematics, Hanyang University, Seongdong-ku, Hangdang-dong 17, Seoul,
E-mail : legna-j@hanmail.net

진 통계 패키지는 향후 통계 저변 확대에도 큰 몫을 하리라 본다.

2. 'APM SURVEYOR 1.5' TOOL 에 대한 소개

본 장에서는 본 시스템에 이용된 APM 과 JAVA 및 Web log 파일 분석에 대해 간략히 소개하고자 한다.

2.1 Apache

Apache 웹서버(web server)는 1995년 당시에 가장 인기 있는 웹서버였던 NCSA httpd 1.3버전을 기반으로 하고 있으며 현재 전 세계에서 가장 많은 서버에 웹서버로 탑재되어 운용되고 있다. 이는 웹서버가 반드시 갖추어야 할 안정성과 활용성면에서 다른 웹서버에 비해 훨씬 우수한 성능을 보여주고 있기 때문이다. 그리고 한 가지 인기를 얻는 요인은 Apache 웹서버의 강점인 패치 파일(patch file)을 통해 지속적으로 성능이 향상된다는 점과 소스(source)까지 무료로 완전 공개 프로그램이라는 점을 들 수 있다 [이승혁 (2000) 참조]. 또한 Apache는 Linux뿐 아니라 Unix, Sun Solaris, IBM, HP등 대부분의 유닉스 기반 운영 체계 및 Windows-NT를 지원하며 본 시스템에서 시험 운행해본 Windows 98 운영체계에서도 사용할 수 있다. 따라서 인터넷 서베이 프로그램의 호환성 문제의 해결에 관해서도 어느 정도 충족시켜 줄 수 있는 서버라 볼 수 있다.

2.2 PHP

PHP는 한마디로 사용자와의 상호 작용을 통한 다이내믹한 웹 페이지(web page)를 한층 더 쉽게 구현 할 수 있도록 도와주는 스크립트 언어이다. 홈페이지를 제작할 때 흔히들 많이 쓰는 자바 스크립트(Java Script)가 사용자의 browser상에서만 실행되는 반면에 PHP는 사용자가 HTML(Hyper Text Markup Language) 폼을 통해 입력한 값을 웹서버 상에서 처리한 후 그 결과를 HTML과 같은 형태로 가공하여 다시 사용자의 browser에 전달하는 서버쪽 언어로서 Windows-NT의 ASP처럼 HTML코드와 함께 프로그래밍이 가능한 스크립트 언어이다. 그러므로 보통 server-side HTML-embedded scripting language라는 말로 표현한다. 이런 PHP의 장점으로는 배우기 쉽고 개발속도 및 실행속도가 빠르며 호환성 측면에서 대부분의 DB(Data Base)와 운영체계를 지원한다 [이승혁 (2000) 및 정진호(2000) 참조]. 본 시스템에서는 각종 통계량을 구하는데 필요한 연산 과정을 수행하였다.

2.3 MySQL

MySQL은 트랜잭션(transactions)이나 트리거(triggers) 등의 기능을 지원하고 있지는 않지만 그만큼 다른 DB보다 속도가 빠르고 표준 SQL(Structured Query Language)문을 충실히 지원하며 무엇보다 PHP와 함께 연동하여 사용하기에 가장 좋은 DB로 평가받고 있다. MySQL은 멀티 쓰레드(multi thread)를 지원하며 PHP는 물론 C나 C++, Java, Perl, Python, TCL등과 함께 사용할 수 있도록 각각에 대한 API(Application Programming Interface) 함수를 지원한다. 또한 Linux, Unix, Solaris, SGI, AIX, FreeBSD 및 Windows-NT 그리고 본 시스템에서 사용된 Windows 98

환경 등 거의 모든 운영체제를 지원한다. 또한 사용자가 사용하기 쉽도록 윈도우즈상에서 DB 수정이 용이하도록 설계되었다 [이승혁 (2000) 및 문정혁(2000) 참조].

2.4 Java Script

Java Script는 PHP처럼 서버를 거쳐서 실행이 되어야 그 결과를 알 수 있는 서버쪽 언어(server-side script language)가 아니라 서버를 거치지 않고도 사용자의 browser 내에서 바로 호출 및 실행이 가능한 클라이언트쪽 언어(client-side script language)이므로 응답이 매우 빠르며 서버에는 전혀 부담을 주지 않게 된다. 이러한 자바 스크립트를 이용하면 사용자가 텍스트 입력박스에 입력한 값이나 리스트 박스 등에서 선택한 값을 알아낼 수 있으므로 이들 값을 CGI(Common Gateway Interface) 프로그램에 전송하기 전에 허용되지 않은 문자나 값을 입력한 경우, 또는 반드시 입력을 해야 하는 항목에 입력하지 않고 넘어간 경우를 체크하여 사용자에게 재 입력을 요구 할 수 있다. 다시 말하면 에러처리 부분에 상당히 유의한 언어로 평가 할 수 있다. 따라서 CGI 프로그램 언어로서의 PHP가 갖는 단점을 보완하는 측면에서 Java Script는 상당한 가치와 의미를 갖는다고 볼 수 있다 [이승혁 (2000) 참조]. 이에 본 프로그램에서도 각종 에러처리부분은 Java Script로 처리하였다.

2.5 Web log 파일 분석

Web log 파일 분석은 web usage mining의 일종으로 이 또한 web mining의 일 부분으로 분류된다. Web usage mining이란 사용자 등록정보 및 server에 남겨진 log 파일을 데이터로 하여 각종 통계기법 및 data mining 기법을 사용하여 필요한 정보를 얻는 방법이다 [Cooley, R. Mobasher, B. and Srivastava, J.(1997), Cooley, R. Mobasher, B. and Srivastava, J.(2000)]. 이는 인터넷 인구의 확산에 따라 대부분의 거래가 offline에서 online 상으로 이동함에 따라 등장하게 된 방법으로 현재 사용자 성향을 알아보는 방법으로 각광받고 있는 방법이다. 이중 web log 파일은 사이트를 방문한 사람의 정보를 그들이 인식하지 못하는 상황 하에서 성향을 파악 할 수 있다. 그러나 log 파일의 양이 방대하여 분석 전에 자동적으로 data cleaning작업이 필요하며 각종통계기법 및 offline상의 data mining기법을 online 상으로 이동해야하는 어려움이 따른다. 이러한 web log 파일의 특성을 이용하여 인터넷 서베이에서는 패널(panel)들에게 부담을 주지 않고 그들의 성향을 파악할 수 있다. 본 시스템에서는 web log를 분석하여 설문응답자의 응답시간에 대한 기초통계량을 제시하고 응답자들의 접속위치와 설문결과에 대한 관심도를 측정하였다.

3. 'APM SURVEYOR 1.5'에 대한 소개

본 시스템은 위에서 제시한 web tool을 이용하여 만들어졌다. 다시 말하면 컴퓨터를 관리자에게 친숙한 Windows 98 환경과 웹상에서 계속 업그레이드되어 최신형을 무료로 구할 수 있는 APM을 사용하였다. 물론 환경 자체를 Windows 98이 아닌 Windows-NT나 Linux, Unix에서도 사용할 수 있고 더 나은 성능을 보여줄 수 있으나, user-friendly 차원에서 Windows 98 환경에서 구현하였다. 또한 상용 패키지를 이용하여 결과의 재분석 및 고급분석을 위해 EXCEL을 활용하여 각종 자료를 1분 간격으로 자동 refresh하여 저장하도록 하였다. [표2]는 본 시스템에 대한 이용환

경을 나타낸 것이다. 본 시스템을 구성하는 화면은 다음 [표3]과 같으며 이를 위해 50개의 html문서와 각종 연산을 위한 23개의 PHP 파일이 이용되었고 5개의 DB를 구축하였다. 데이터 베이스 시스템(Data Base System)또한 윈도우상에서 관리자의 계정으로 쉽게 볼 수 있으며 새로운 DB 생성 및 자료의 수정, 보안이 용이하다.

[표1] APM SURVEYOR 1.5 이용 환경

구분	Server	Basic Language	Error 처리	Data Base	자료 정리
사용 환경	Windows 98	PHP	Java	MySQL	EXCEL 2000

[표2] 화면 구성

구분	Page / DB directory	세부 내용
관리자 화면	사용자 등록정보	사용자가 회원등록시 등록정보를 확인할 수 있는 화면 [그림1]
	설문 응답시간에 대한 정보	응답 시간에 대한 기초 통계량 및 극단 값을 제외한 평균 제공 [그림2]
	접속 위치	접속위치별 빈도수 및 막대그래프 제공 [그림2]
	설문 결과에 대한 관심도	설문 결과(14개)화면에 대한 패널들의 접속횟수를 트리구조로 제시 [그림14]
	Raw Data	Excel을 이용하여 raw data 표시 [그림1]
사용자 화면	회원 등록	회원등록 화면
	로그인	설문 취지 설명 및 로그인을 위한 화면
	설문 응답	설문 내용 제시
	각종 결과	각종 통계량 제시 [그림3]~그림[13]
	Raw Data	온라인 상에서 raw data 표시
Database	member	회원정보에 대한 DB구축
	poll	설문 응답결과에 대한 DB구축
	prelogdb	웹 로그파일을 1차 정제후 DB구축
	timedb	prelog DB에서 응답 시간에 관한 것만을 추출하여 기초 통계량과 함께 DB 구축
	zipcode	회원등록시 필요한 우편번호에 대한 DB 구축

4. 'APM SURVEYOR 1.5'의 특징

본 장에서는 인터넷 서베이의 장단점을 요약하고 'APM SURVEYOR 1.5'를 이용하여 설문조사를 실시한 결과를 바탕으로 인터넷 서베이의 장단점에 대해 어떻게 보완 및 발전시켰는지에 대해 논하고자 한다.

4.1 인터넷 서베이의 장단점

지금까지의 김영원, 변종석(2000), 김연형, 오민권(2000), 이계오(2000), 이종수, 제병환(2000), 이해용, 김기환(2000) 등에 의해 발표된 논문에 기술된 인터넷 서베이에 대한 장, 단점을 요약하면

다음 [표3]과 같이 정리해 볼 수 있다.

[표 3] 인터넷 서베이의 장, 단점

장점	단점
비용의 저렴성	고정 비용의 증대
표본 추출틀 구성의 용이성	특정 표본추출 틀의 결여
소요 시간의 단축	대표성 결여
웹으로서의 장점 (그래픽이나 음성 동화상의 이용가능)	웹으로서의 문제점 (browser 나 호환성의 문제)
설문 응답의 용이성	분석 및 설문지의 단순화
재 질문이 가능	소프트웨어의 추가 사용

[표3]에서 보듯이 인터넷 서베이의 장점과 단점을 요약하면 장점으로 지적한 부분도 상황에 따라 단점으로 간주될 수 있는 부분이 많다. 그러므로 본 프로그램은 장점부분을 최대한 부각하고 단점 부분을 보완하는 방향으로 시스템을 구동하였다.

4.2 'APM SURVEOR 1.5'의 장점

본 시스템의 장점은 다음을 들 수 있다.

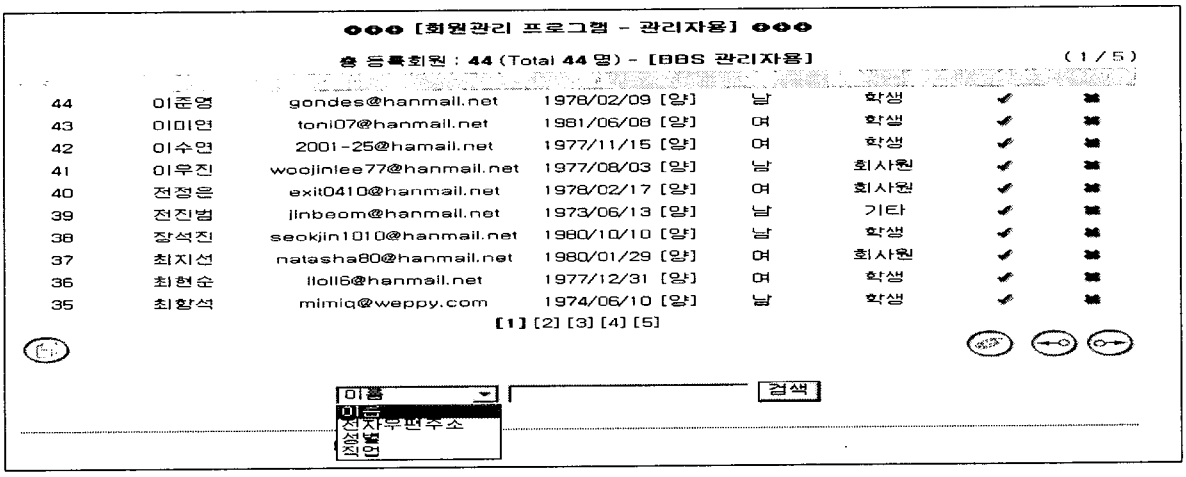
첫 번째로 비용문제의 해결이다. 인터넷 서베이라고 하면 비용이 저렴하다는 것을 첫 번째 이유 중에 하나로 꼽는다. 다른 매체에 비해 인건비같은 비용의 문제가 적다는 것은 사실 일수 있다. 하지만 단점으로 제시되었던 고정 비용의 문제와 표본 수나 질의 향상을 위한 비용은 전형적인 설문조사와 비교할 때 그리 저렴하다고 판단할 수 있는 문제는 아니다. 실제로 많은 네티즌들(netizen)의 아무런 대가 없는 성실한 답변을 기대하기가 힘든 상황이고 경품의 질에 따라 네티즌들의 성실성에서도 차이가 나는 상황이다. 또한 시스템을 운영하기 위하여 유지되는 서버의 가격, 그에 따른 업그레이드(upgrade) 비용, 패널 관리비용이 크다는 것이다. 물론 자본이 풍부하여 최고의 사양과 응답자에게 풍부한 혜택으로 좋은 결과를 위해 투자하는 것은 좋지만 설문조사를 지속적으로 하는 단체가 아니면 이 부분도 큰 문제가 된다. 이에 본 시스템은 서버용 컴퓨터를 이용하지 않고 가격이 저렴하고 누구든지 편하게 사용 할 수 있는 개인용 PC로 구현하였으며 그와 관련된 소프트웨어도 쉽게 무료로 구할 수 있고 업그레이드 속도가 빠른 PHP와 MySQL을 사용하였다. 패널 문제도 자체 연구실 및 조교, 수학과 학회('HUMAN') 그리고 인터넷 동호회('왕모') 회원으로 패널 구축을 하였다. 큰 규모를 필요로 하지 않는 서베이에 대해서는 비용면에서 저렴하며 성실성의 문제도 어느 정도는 해결되는 방법을 사용하였다. 또한 실제 상용화하고 있는 사이트에서는 표본 수의 증가에 따라 비용이 증가된다. 참고로 인터넷 서베이를 대행해주는 유료 사이트의 비용은 아래 [표4]와 같이 상황에 따라 비용의 많은 차이가 있다.

[표 4] 인터넷 유료 설문조사 사이트 가격 (2001년 4월 기준)

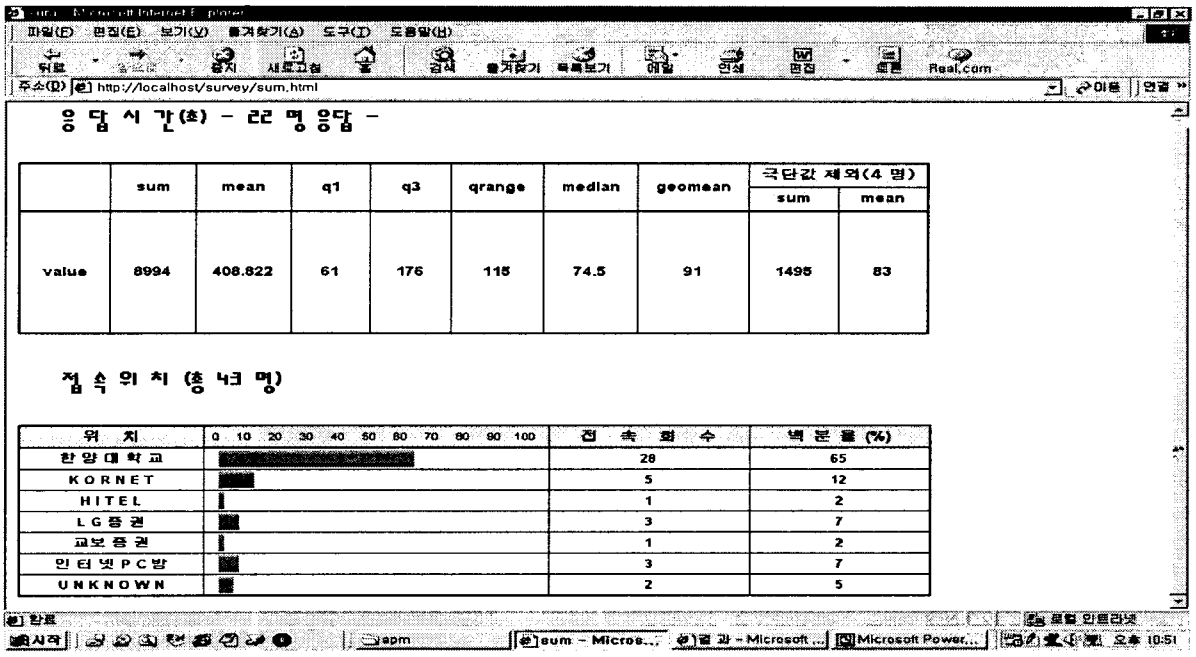
회사	비용 책정방법	실시간 분석	비고
A	문항수*8만원, 표본수*천원	단순 회귀분석 까지가능	설문목적/내용에 따라 추가비용 소요. 고급분석은 off-line에서 분석후 전송
B	20문항 400표본기준 500~800만원	빈도 분석 및 비율	
C	표본당 3천원~2만원	빈도 분석 및 비율	
D	20문항시 표본당 7천원	빈도분석 및 교차분석	

둘째, 표본 추출의 용이성을 부각시키기 위해 MySQL을 이용하여 DB를 구축하였다. 패널을 데이터 베이스화 시키면 [그림1]과 같이 특정집단 즉 학생이나 직장인 혹은 특정 나이만을 추출하여 설문 조사를 할 수 있다. [그림1]은 전체 회원중 일부분을 제시한 그림이고 [그림1] 하단에 있는 검색기능을 이용하여 특정집단만을 선별 할 수 있다. 이는 PHP 프로그램에서 한 줄의 코딩(coding)만으로도 더 세분화된 집단을 추출 할 수 있다. 또한 동일한 질문을 동일 패널에게 질문할 수 있어 네티즌들의 성향변화도 측정할 수 있다. 이는 패널과의 지속적인 관계 유지를 할 수 있어 향후 web mining이나 CRM(Customer Relationships Management)등으로 활용될 수 있다.

[그림 1] 회원관리 화면



셋째로 대표성 결여의 문제에 대한 해소 안을 들 수 있다. 대표성 결여의 문제는 인터넷 서베이가 가지는 가장 큰 문제로 보고 있다. 자칫하면 인터넷 서베이만의 문제라고 오해할 수 있는 문제지만 이 부분은 모든 설문조사가 가지는 공통된 문제라고 볼 수 있다. 하지만 이 대표성 문제에서도 인터넷이라는 매체가 가지고 있는 문제점이 패널들의 비 성실성, browser가 가지는 문제점 즉 browser 자체에서 『새로 고침』을 눌렀을 때 응답수가 증가하는 부분을 들 수 있다. 이에 본 시스템은 패널 구축후 한 사람이 두 번 설문 조사를 할 수 없게 주민등록 번호로 password화 시키고 두 번 입력 시는 분석 자체에 입력이 되지 않고 에러처리 할 수 있게 하였다. 또한 성의 없는 답변을 체크하는 부분은 세 가지 방법으로 확인할 수 있다. 본 시스템은 설문 문항이 짧아 이중 두 번째 경우로 실행하였다. 우선은 첫 번째로 설문 문항이 길 경우 똑같은 질문을 두 번 넣고 보기 순서를 바꾸는 방법이다. 이는 패널의 무성의한 대답을 찾고자하는 방법으로 인성검사에서도 활용되는 부분이다. 그래서 결과를 산출하는 과정 중에 두 질문의 답이 틀리면 분석 결과에서 제외시키는 방법이다. 두 번째로 이와 비슷한 방법으로 기존에 DB로 구축되어있는 패널 부분에서 패널의 특성(성별, 직업등)으로 저장되어 있는 값과 설문조사시 패널에 관련된 질문을 비교하여 틀린 부분이 있으면 무성의로 간주하고 계산결과에서 제외시키는 방법을 사용했다. 세 번째 방법으로는 [그림2]와 같이 web log를 활용하여 설문 응답 시간에 대한 통계량을 추출하여 너무 짧은 시간에 응답 한 경우를 제외시키는 방법이다. 이는 설문 문항수가 많을때 유용하게 쓸 수 있는 방법으로최종후, 강현희(2000)에서 제시한 방법을 응용하였다. [그림2]는 일주일간의 web log 파일을 분석한 결과이다.



응답시간(초) - 22 명 응답 -

	sum	mean	q1	q3	qrange	median	geomean	극단값 제외(4 명) sum	극단값 제외(4 명) mean
value	8994	408.822	61	176	115	74.5	91	1495	83

접속위치 (총 43 명)

위치	0 10 20 30 40 50 60 70 80 90 100	접속회수	백분율(%)
한양대학교		28	65
KORNET		5	12
HITEL		1	2
LG증권		3	7
교보증권		1	2
인터넷PC방		3	7
UNKNOWN		2	5

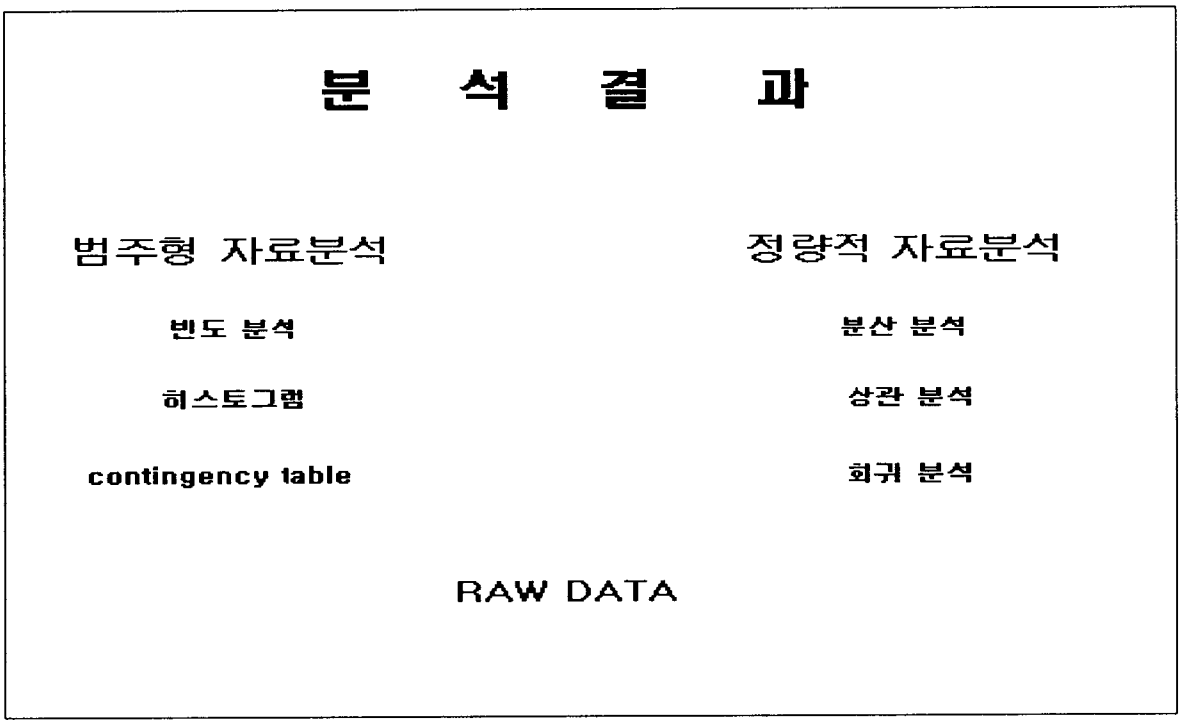
[그림 2] Web log data를 이용한 설문 응답 시간 및 접속 위치 정보

넷째 호환성이 용이한 web tool을 사용하였다. 위에서 설명하였듯이 대부분의 프로그램이 운영체계 환경에 따라 실행이 안 되는 경우가 많고 응답자 입장에서도 browser 버전이나 환경에 따라 보여지는게 틀려질 수도 있다. 이에 대해 본 시스템은 호환성이 좋은 Apache, PHP, MySQL을 이용하였고 이는 윈도우에서 사용된 명령어 그대로를 Unix나 Linux에서 사용할 수 있어 호환성 문제를 좀 더 줄일 수 있도록 만들었다.

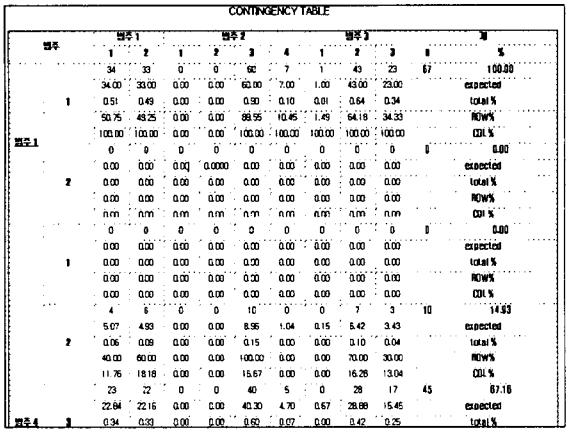
다섯째, 다양한 분석기법을 보여주었다. 최근 들어 빈도분석에서 좀 더 많은 분석을 실시간으로 보여주는 사이트들이 생겨나고 있지만 아직도 대부분의 사이트는 실시간으로는 빈도 분석만을 위주로 하고 있다. 이에 본 시스템에서는 범주형 자료와 정량적 자료에 대한 분석을 구별하고, 빈도분석 뿐만 아니라 범주형 자료분석에는 [그림 4]와 [그림 5]와 같이 빈도 분석과 히스토그램, contingency table, 각 문항별 Anova Table을 제시하였고, 정량적 자료분석에는 [그림 6] ~ [그림 12]에 나타난 것과 같이 기초 통계량, 각 문항별 Anova Table, Correlation Matrix, Simple Linear Regression을 분석에 활용하였다. 그리고 각종 검정 통계량 및 자유도를 결과에 첨부시켜 유의성을 확인할 수 있게 하였다. 그래프 부분도 histogram과 더불어 scatter plot, residual plot, linear plot을 볼 수 있게 하였다. 각각의 내용은 [표4]에 간략히 요약해 놓았다. 본 시스템에서는 가장 기본적인 환경 즉 Windows 98 환경을 서버로 하고 다른 라이브러리(library)를 이용하지 않고 기본 함수와 프로그램 자체 연산을 통해 위와 같은 결과를 제시하였다. 참고로 본 시스템을 이용한 설문 주제는 'PC방 이용 실태조사'이고 6개의 객관식 문항, 2개의 주관식, 총 8개의 질문과 패널의 정보를 얻기 위하여 4개의 질문을 포함 시켰다. 여기에 사용된 질문은 아하넷(www.ahanet.co.kr)을 참고하였다. [그림 3]은 분석 결과의 첫 화면이다.

[표 5] APM-SURVEYOR 1.5 분석내용

분석 결과 window		분석 내용
범주형 자료	Contingency Table	응답자 부류별로 각 문항에 응답한 빈도, total%, col%, row%
	Anova Table	클릭(click)한 문항에 관한 내용과 응답자 부류에 따른 응답빈도수, histogram, 자유도, 검정통계량 값(χ^2)
정량적 자료	Basic Statistics	SAS의 proc univariate 으로 추출할 수 있는 모든값
	Anova Table	문항 내용과 응답자 부류별 빈도수, S(M)SE, S(M)STR, SSTO, 자유도, 검정통계량 값(F-value)
	Correlation Matrix	correlation matrix와 검정통계량 값(T-value)
	Linear Regression	회귀식과 S(M)SE, S(M)SR, SSTO, 자유도, R-square 및 검정 통계량, 변수 변환($\sqrt{X}, X^2, \log X$)
	Scatter Plot	scatter plot
	Residual Plot	residual plot(표준화된 잔차)
	Linear Plot	linear plot 과 scatter plot

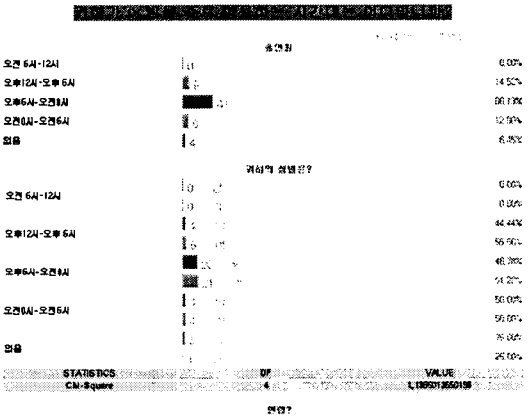


[그림 3] 분석 결과 첫 화면

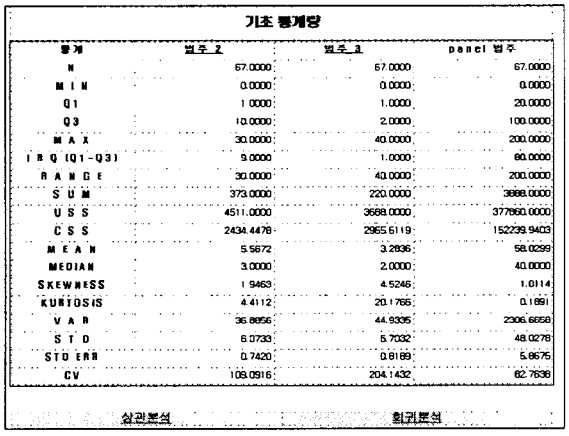

CONTINGENCY TABLE

범주		범주 1		범주 2			범주 3			계	
		1	2	1	2	3	4	1	2	3	%
		34	33	0	0	62	7	1	43	23	67	100.00
		34.00	33.00	0.00	0.00	60.00	7.00	1.00	43.00	23.00		expected
	1	0.51	0.49	0.00	0.00	0.93	0.10	0.01	0.64	0.34		total %
		50.75	49.25	0.00	0.00	86.55	10.45	1.49	64.18	34.33		ROW%
		100.00	100.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00		COL%
범주 1		0	0	0	0	0	0	0	0	0	0	0.00
		0.00	0.00	0.00	0.0000	0.00	0.00	0.00	0.00	0.00		expected
	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		total %
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		ROW%
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		COL%
		0	0	0	0	0	0	0	0	0	0	0.00
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		expected
	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		total %
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		ROW%
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		COL%
		4	6	0	0	10	0	0	7	3	10	14.93
		5.07	4.93	0.00	0.00	8.95	1.04	0.15	5.42	3.43		expected
	2	0.06	0.09	0.00	0.00	0.15	0.00	0.00	0.10	0.04		total %
		40.00	60.00	0.00	0.00	100.00	0.00	0.00	70.00	30.00		ROW%
		11.76	18.18	0.00	0.00	15.87	0.00	0.00	16.28	13.04		COL%
		23	22	0	0	40	5	0	28	17	45	67.16
		22.84	22.16	0.00	0.00	40.30	4.70	0.67	26.88	15.45		expected
범주 4	3	0.34	0.33	0.00	0.00	0.60	0.07	0.00	0.42	0.25		total %

[그림 4] 범주형 자료에 대한 contingency table



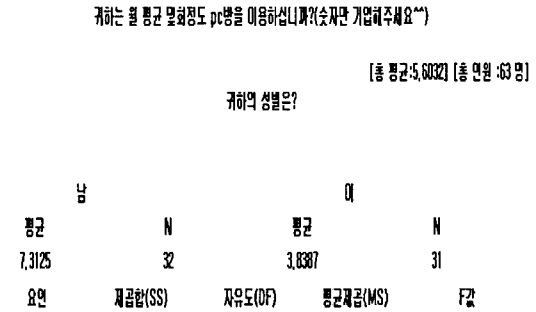
[그림 5] 범주형 자료에 대한 문항별 Anova Table


기초 통계량

통계	범주 2	범주 3	panel 범주
N	67.0000	67.0000	67.0000
MIN	0.0000	0.0000	0.0000
Q1	1.0000	1.0000	20.0000
Q3	10.0000	2.0000	100.0000
MAX	30.0000	40.0000	200.0000
IRQ (Q1-Q3)	9.0000	1.0000	80.0000
RANGE	30.0000	40.0000	200.0000
SUM	373.0000	220.0000	3888.0000
USS	4511.0000	3688.0000	377660.0000
CSS	2434.4478	2965.6119	152239.9403
MEAN	5.5672	3.2836	58.0299
MEDIAN	3.0000	2.0000	40.0000
SKEWNESS	1.9463	4.5246	1.0114
KURTOSIS	4.4112	20.1766	0.1891
VAR	36.8856	44.9335	2396.6658
STD	6.0733	6.7032	48.0278
STD ERR	0.7420	0.8189	5.8675
CV	109.0916	204.1432	82.7638

상관분석 회귀분석

[그림 6] 정량적 자료에 대한 기초 통계량


귀하는 월 평균 몰입정도 pc방을 이용하십니까?(숫자만 기입해주세요^^)

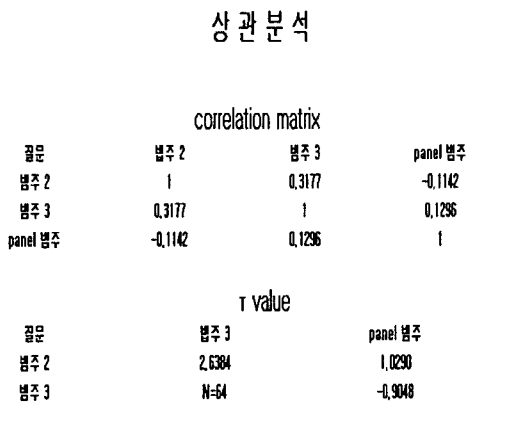
[총 평균:5.6032] [총 인원 :63 명]

귀하의 성별은?

남		여	
평균	N	평균	N
7.3125	32	3.8387	31

요인	제곱합(SS)	자유도(DF)	평균제곱(MS)	F값
처리	190.0119	1	190.0119	5.2564
잔차	2205.0685	61	36.1487	
계	2395.0794	62		

[그림 7]정량적 자료에 대한 문항별 Anova Table


상관분석

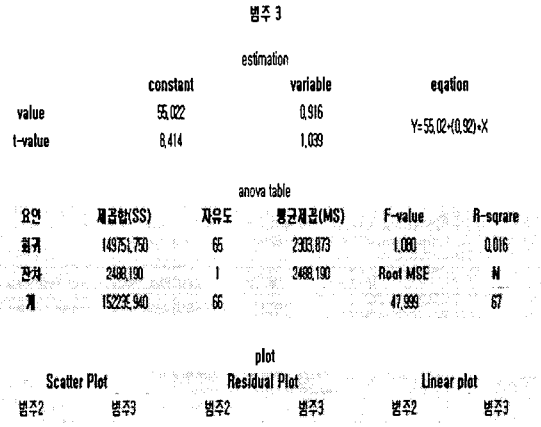
correlation matrix

질문	범주 2	범주 3	panel 범주
범주 2	1	0.3177	-0.1142
범주 3	0.3177	1	0.1296
panel 범주	-0.1142	0.1296	1

T value

질문	범주 3	panel 범주
범주 2	2.6384	1.0290
범주 3	N=64	-0.9048

[그림 8] Correlation Matrix


범주 3

estimation

	constant	variable	equation
value	55.022	0.916	Y=55.02+(0.92)*X
t-value	8.414	1.039	

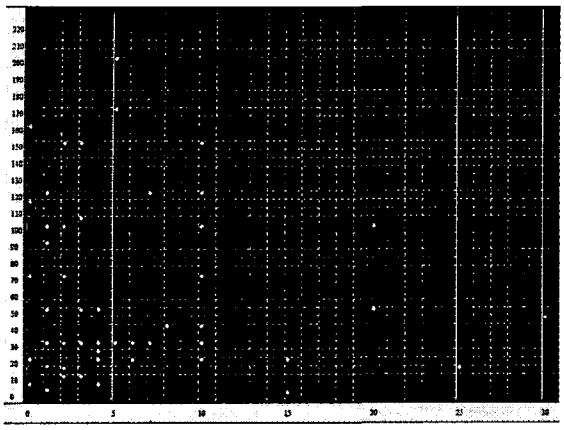
anova table

요인	제곱합(SS)	자유도	평균제곱(MS)	F-value	R-sqrare
회귀	14975.750	65	2303.873	1.060	0.016
잔차	2408.190	1	2408.190	Root MSE	N
계	15223.940	66		47.999	67

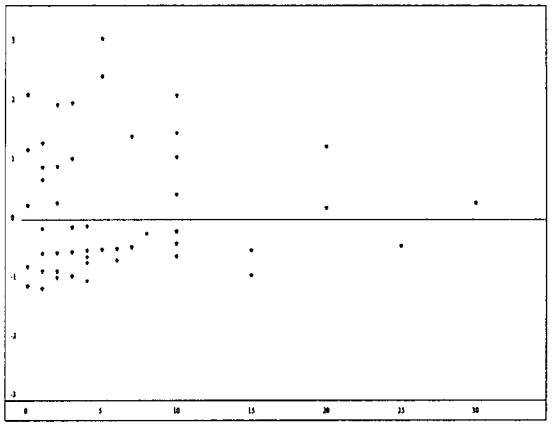
plot

Scatter Plot 범주2 범주3 | Residual Plot 범주2 범주3 | Linear plot 범주2 범주3

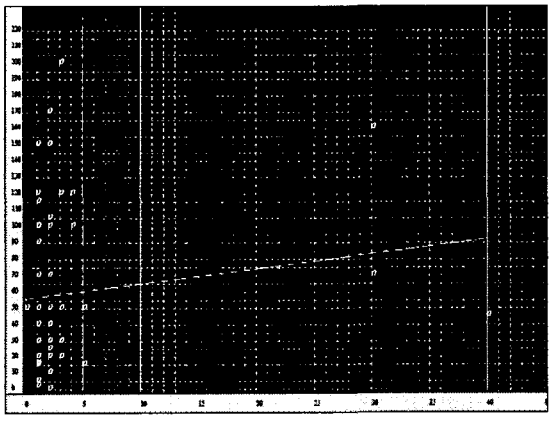
[그림 9] 범주 3에 대한 회귀분석 통계량



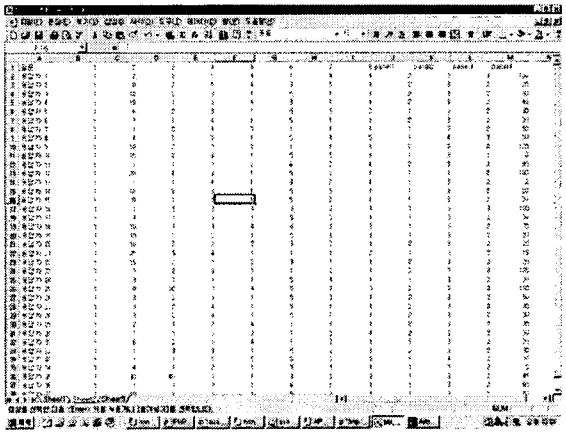
[그림 10] 범주 3에대한 scatter plot



[그림 11] Residual Plot



[그림 12] 범주 2에대한 linear plot

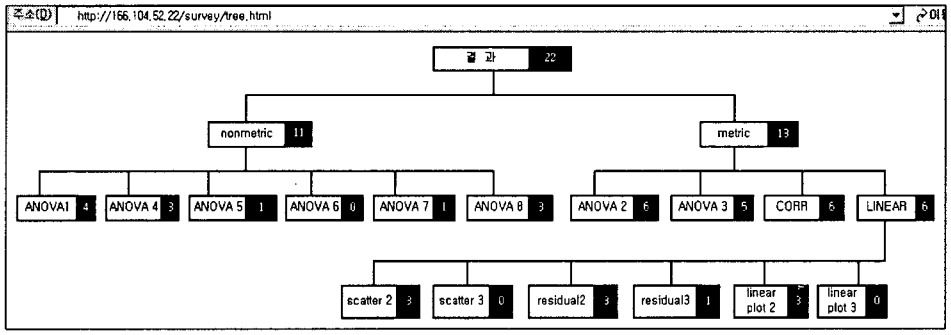


[그림 13] Raw data를 EXCEL로 불러들인 window

여섯째, 결과의 재분석 및 결과 데이터의 활용을 쉽게 하였다. 즉 대부분의 사용자가 쉽게 사용하는 EXCEL을 이용하여 각종 결과를 저장하였다. 이는 EXCEL 2000의 커리(query) 실행으로 웹상의 데이터를 엑셀 데이터로 변환하여 1분 간격으로 자동 refresh하며 확인할 수 있다. 또한 raw data를 재활용하여 기타 다른 소프트웨어(SAS, SPSS등)로 재분석 할 수 있다. 위 [그림13] 은 raw data를 1분 간격으로 EXCEL로 불러들이고 있는 화면이다.

일곱 번째로 사용자에 대한 보다 많은 정보를 얻기 위하여 web log 파일을 분석하여 위의 [그림2]에서는 설문 응답시간에 대한 평균 시간 및 접속자 위치를 분석하였고 다음 [그림14]와 같이 설문 검사 결과에 대한 관심도를 결과 page 를 방문한 인원을 카운트(count) 하여 tree구조로 보여주었다. [그림14]에서 보면 트리 모양은 결과 페이지를 트리모양으로 구조화한 그림으로 level이 낮은 page는 상위 레벨을 거쳐서 들어갈 수 있게 되어있다. 이는 Jiawei Han and Kamber, M.(2001)에서 소개된 FP tree를 수정한 그림으로 숫자부분은 해당 페이지를 방문한사람의 수를

나타낸다. 이는 로그파일 분석의 일부분만을 제시한 그림으로 각종 통계 기법 및 mining 기법을 이용하여 더 많은 정보를 얻을 수 있고 이는 향후 서베이와 CRM의 확장을 모색할 수 있을 것으로 보여진다.



[그림 14] 설문 결과 page 에 방문한 사람의 수

4.3. 'APM SURVEYOR 1.5'의 보완점

본 시스템에 대하여 다음과 같은 문제점을 생각해 볼 수 있다.

첫 번째로 최소한의 비용과 최소한의 환경을 목적으로 구축하려는 목적으로 인해 각종 프로그램을 무료로 얻을 수 있는 프리웨어 제품을 사용하였다. 이에 따라 시중에 나와 있는 상용프로그램보다 지원함수가 적어 프로그램의 길이가 길어지고 분석의 다양성 측면에서는 현재 주로 사용하고 있는 것보다는 많은 방법을 제시하였지만 참고 사이트[15] 같은 곳보다는 내용면에서 다소 부족함을 보였다. 물론 이 부분은 DB software및 컴퓨터 사양을 높게 하면 어느 정도 해소되는 부분이라 생각이 든다.

두 번째로는 서버용 컴퓨터를 서버로서 능력이 좋은 컴퓨터와 서버용 운용체제인 Unix나 Linux, Windows-NT등을 사용하지 않고 일반 PC에 Windows 98 환경을 이용함에 따라 수행 속도가 상대적으로 늦고, 동시 접속이 많은 경우에는 컴퓨터가 다운되는 경우도 발생하였다. 본 저자는 좀 더 좋은 컴퓨팅 환경에서 실험 해 볼 수도 있지만 본 시스템의 목적에 따라 최소한의 환경에서 시스템을 구동하였고 큰 문제는 발생하지 않았다. 상업적 목적으로의 활용은 본 프로그램과 좋은 컴퓨팅 환경으로 구축할 수 있을 것으로 보아진다.

세 번째로는 그래픽한 면이 기존의 상업적인 프로그램보다 떨어지는 흠을 가진다. 웹에서 이루어지는 프로그램은 비쥬얼한 면이 상당히 중요한 부분이다. 이는 본 시스템의 버전(version)을 향상함에 따라 좀더 나은 화면과 Visualization 기법을 동원하여 사용자에게 좀 더 친숙한 화면을 제공할 것이다. 본 시스템으로 모든 것을 할 수는 없다. 하지만 사용자의 목적에 따라 조금씩 수정,보완하면 좋은 설문조사 프로그램으로의 확장을 모색할 수 있으리라 본다.

5. 결론

인터넷인구의 급증을 염두 해 보면 표본 틀 문제가 해결되고 서베이의 한 분야로서 자리를 잡기보다는 대부분의 서베이가 인터넷에서 이루어질 것으로 보았다. 인터넷 서베이의 발전은 두 가지 측면에서 볼 수 있다[김광용(2000) 참조]. 하나는 현 시점과 같이 불특정 다수, 즉 네티즌을 대상으로한 서베이를 들 수 있고, 또 하나는 특정 다수, 즉 network화된 각종 기관이나 단체에서 소규모 인원을 대상으로 한 서베이의 발전이다. 우선 불특정 다수에 대한 서베이는 시스템 환경의 발전에 따라 대규모의 데이터 베이스와 상호 교환적인 서베이, 즉 화상을 통한 분석과 설문지에 대해서도 비쥬얼한 면이 부각되고 고급 분석의 실시간 화 및 1:1식의 대화면접이 활성화 될 것이다. 특정인원에 대한 서베이는 본 시스템에서 실행한 것과 같이 소규모인원에 대한 분석 또한 따로 발전되어 기업체나 단체에서 언제든 쉽게 소속인원들의 의사를 집계하기가 쉬워지고 분석 또한 용이해지리라 본다.

또한 인터넷 서베이는 서베이 과정중 동일인에 대하여 같은 질문을 할 수 있고, web log data 분석을 통하여 패널들의 성향 변화 과정, 바꿔 말하면 일반 기업에서 원하는 고객들의 성향 변화 과정을 분석할 수 있어 web mining 이나 CRM으로서의 통합과정도 확립된다고 볼 수 있다. 따라서 이러한 결과를 기반으로 하여 앞으로 web mining 이나 CRM으로의 확장을 모색하고자 한다.

참고문헌

[1] 김광용(2000). Web Information Center와 Internet Survey,
한국 조사 연구학회 「Internet Survey」 workshop 논문집, 111-122.

[2] 김영원, 변종석(2000). 인터넷 조사에서 표본 추출 동향 및 문제점,
한국 조사 연구학회 「Internet Survey」 Workshop 논문집, 19-35.

[3] 김연형, 오민권(2000). Internet Poll System,
「한국 통계학회 논문집」, 제 7권 3호 927-935.

[4] 문정혁(2000). 「MySQL & mSQL」, 한빛미디어

[5] 이계오(2000). 인터넷 여론조사의 현황과 전망,
한국 조사 연구학회 「Internet Survey」 Workshop 논문집, 1-17.

[6] 이승혁(2000). PHP 웹 프로그래밍 가이드, 마이트프레스.

[7] 이종수, 제병환(2000). 인터넷 환경과 리서치 방법론,
한국 조사 연구학회 「Internet Survey」 workshop 논문집, 93-99.

[8] 이해용, 김기환(2000). Internet Survey Methodology,
「한국 통계학회 논문집」, 제 7권 3호 945-953.

[9] 정진호(2000). 「PHP Web-DB Programming Guide」, 통일출판사.

[10] 최종후, 강현희(1998). 「설문조사 -처음에서 끝까지」, 자유아카데미.

[11] Jiawei Han, Kamber. M(2001). 「Data Mining: Concepts and Techniques」, MORGAN KAUFMANN.

[12] Srivastava, J., Clloey, R., Deshpande, M. and Pang-Ning Tan(2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations* Vol.1, Issue2.

[13] Cooley, R., Mobasher, B. and Srivastava, J(1997). Web Mining : Information and Pattern Discovery on the World Wide Web, In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).

[14] Cooley, R., Mobasher, B. and Srivastava, J(2000). Automatic Personalization Based on Web Usage Mining, *Communication of ACM*, Vol.43, Issue 8.

[15] 아하넷 설문 분석 시스템, http://www.ahanet.co.kr.

[2001년 7월 접수, 2001년 12월 채택]