The Korean Communications
in Statistics Vol. 11 No. 1, 2004
pp. 93-106

A Development Study of Tool for Web Log Analysis

Seungbae Choi[1], Changwan Kang[2], Kyukon Kim[3], Jongkwan Son[4]

Abstract

Recently, many data of various types is gained with development of computer in many fields. Especially, web log data generating in web site furnish beneficial information on an organization. The enterprise's destiny is swayed by according as how these information gaining from the web site utilize. In this paper, for the purpose of obtaining useful information, we present a tool is called WebBizi for web log analysis. This will be helpful to enterprise working the web site.

Keywords : WebBizi, WebTrends, WiseLog, Client Relationship Management.

1. 서론

현대는 인터넷이라는 매체를 통하여 많은 지식을 공유하고 있다. 이러한 매체를 통하여 얻어지는 데이터는 상상할 수 없을 정도로 실로 방대하며 거대하다. 이러한 수많은 데이터에서 목적한 바를 이루기 위하여 적절히 정제하여 필요한 정보를 얻게 된다. 사이트를 운영하고 있는 기업에 대한 성공 여부는 이러한 정보를 어떻게 잘 활용하는가에 달려 있다. 즉, 어느 고객이 언제 어떤 목적으로 사이트를 방문했는가 등에 대한 내용을 알아내고, 이러한 고객의 성향을 어떻게 잘 파악하느냐에 따라 기업의 사활이 걸려 있다고 해도 과언이 아니다. 따라서 기업은 고객 데이터에 관심을 가지게 되었고, 고객의 행동패턴을 연구하는데 심혈을 기울이고 있다. 이런 상황에서 웹사이트를 운영하고 있는 기업에서는 웹 로그 분석을 이용하여 고객의 관심도와 그들의 행동을 파악하는데 최선을 다하고 있다. 고객의 성향을 파악하고 연구하는 분야로서 고객관계관리(CRM: Customer Relationship Management)가 있다. 웹사이트에서 CRM을 전개할 때, 가장 먼저 고려해야 할 사항은 사용자가 어디에 관심이 있고, 웹사이트를 어떤 방식으로 운용하고 있는지에 대한 것이다. 이러한 것과 관련된 분석이 웹 로그 분석으로서 사용자가 사이트에 언제 접속했으며, 어떤 페이지를 참조했으며, 어떤 검색 엔진으로 사용했는지에 대한 사용자의 행동 패턴을 분석하고 연구하는 것이다(전옥선, 2002).

1) Full-time Lecturer, Department of Information Statistics, Dongeui University, 24, Kaya-Dong, Busanjin-Gu, Busan 614-714, Korea, E-mail: csb4851@hyomin.dongeui.ac.kr
2) Associate Professor, Department of Information Statistics, Dongeui University, 24, Kaya-Dong, Busanjin-Gu, Busan 614-714, Korea.
3) Professor, Department of Information Statistics, Dongeui University, 24, Kaya-Dong, Busanjin-Gu, Busan 614-714, Korea.
4) Graduate student, Department of Information Statistics, Dongeui University, 24, Kaya-Dong, Busanjin-Gu, Busan 614-714, Korea.

실제 얻어진 웹 로그 데이터를 적용하여 웹 로그 분석을 하기 위하여, 강창완 등(2001)은 전국 아르바이트 정보, 구인/구직 검색, 경매 서비스 등을 운영하고 있는 인터넷 업체인 (주)알바누리(www.albanuri.co.kr)의 웹사이트에서 얻어진 웹 로그 데이터를 이용하여 구인/구직에 대한 내용에 초점을 두고 웹 로그 분석을 수행하였다. 또한 김석기 등(2001)은 웹 로그 데이터로부터 정보를 추출하기 위한 과정 및 방안에 대해서 고찰하고, 로그 데이터 분석 예제를 통하여 데이터 수집 및 사전 처리 과정과 추출할 수 있는 정보 및 분석 방법 등을 제시하고 있다.

최승배 등(2002)은 웹 로그 데이터를 분석하기 위한 툴로서 비교적 시장성이 큰 웹 로그 분석기인 웹트렌즈(WebTrends)와 와이즈로그(WiseLog)를 고려하여 두 분석기의 장단점과 각 분석기들이 가지고 있는 특징들을 비교하고, 동의대학교 정보통계학과의 홈페이지 내에서 학술정보 사이트를 개설하여 얻어진 웹 로그 데이터를 가지고 두 분석기에 의해서 웹 로그 분석을 실시하고 비교 분석하였다. 현재 웹사이트에서 생성되고 있는 웹 로그 데이터를 분석하기 위한 몇몇의 분석 툴이 시판되고 있거나 웹 상에서 분석할 수 있도록 제공되고 있다. 그러나 이러한 분석기들은 대부분이 영문판이고 가격대도 고가라는 이유 등으로 최근 들어 한글판 웹 로그 분석의 개발이 활발히 진행 중에 있다. 웹 로그 및 데이터 분석 업체인 넷스루(www.nethru.co.kr)는 웹 로그와 운영계 데이터를 연동하여 분석하는 시스템의 구축이 활발히 진행되고 있다고 밝혔다. 또한 웹 서버 소프트웨어 전문 개발 업체인 이너버스(www.innerbus.com)에서 어떠한 서버 운영체계(OS)에서도 자유자재로 사용할 수 있고, 순수 객체지향 프로그래밍언어(JAVA)와 확장형 생성언어(XML)로만 제작된 웹 로그 분석기를 개발했다고 밝혔다. 이러한 추세에 발 맞추어 본 연구에서도 보다 사용하기 편리한 한글판과 가격적인 측면에서 저렴한 분석기 개발의 필요성을 절실히 느끼게 되어 기존의 웹 로그 분석기와 버금가는 한글판 웹 로그 분석기('WebBizi')를 개발하였다. 본 연구에서는 개발된 'WebBizi'에 대해서 1) 기능적인 측면, 2) 방법론적인 측면, 그리고 3) 통계적인 측면에서 소개와 함께 'WebBizi'에 대한 사용법을 설명한다.

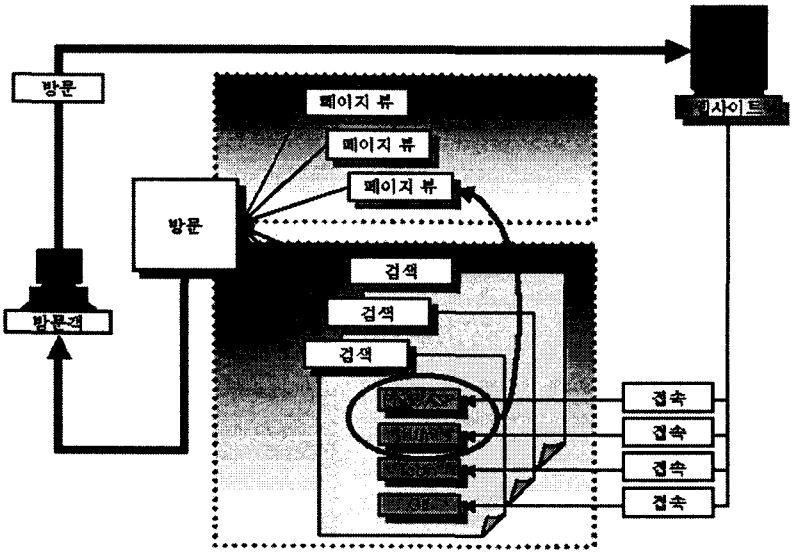
본 연구의 구성은 다음과 같다. 2절에서는 웹 로그 분석에 대해서 간단히 알아보고, 3절에서는 'WebBizi'의 특징을 소개하고 기존의 웹 로그 분석기 log-analyzer(영문판)와 몇몇의 측면에서 비교해 본다. 'WebBizi'의 기능 및 사용법의 설명을 위해서 실제 얻어진 웹 로그 데이터를 이용해서 'WebBizi'에 의한 웹 로그 분석을 4절에서 수행한다. 그리고 마지막 절에서 결론을 맺는다.

2. 웹 로그 분석

2.1 웹 로그 데이터

사용자가 웹사이트에 접속하기 위하여 웹브라우저에서 웹사이트의 주소(URL)를 입력하는 데, 이를 웹사이트를 보기 위한 요청(request)이라고 한다. 이러한 요청에 의해서 해당 사이트를 호스팅하는 웹 서버로 인터넷을 통하여 보내진다. 특정 사이트에 대한 모든 요청들은 해당 웹 서버에 '로그 데이터' 파일로 저장되어 진다. [그림 1]은 웹 로그 데이터가 어떻게 만들어지는 가를 보여주고 있다. [그림 1]에서 볼 수 있듯이 인터넷 상의 사용자들이 '.php', '.asp', '.html', '.icon', '.gif' 등의 확장자를 가진 페이지들을 클릭 함으로써, 웹 서버에 클릭 흐름도 및 사용자들의 상태(OS버전, 브라우저 버전)들이 저장된다. 웹 상에서 이러한 과정을 통하여 얻어진 정보를 웹 로그 데이터라고 한다. 웹 로그 데이터는 사용자가 웹 페이지를 액세스할 때마다 기록되는 것으로, 사용자의 IP와 액세스한 파일 및 시간 등의 정보가 남아 있다. 일반적으로 특정 웹 페이지를 보기 위한 사용자의 요구로, 웹 서버는 해당 웹 페이지와 관련된 여러 파일 등에 접근하게 된다. 따라서 사용자가 요청하는 특정 웹 페이지뿐만 아니라 해당 웹 페이지와 관련된 이미지 파일, 이미지 데이터, 모든 연관 파일 등에 대한 정보가 로그 파일에 저장되는 것이다. 로그 파일 분석을 통한 웹 트래킹 측정은

웹사이트 방문자들의 다양한 사이트 방문형태를 알려주기 때문에 사이트 관리를 효율적으로 개선해주며, 전략적으로 사업을 수행할 수 있도록 한다. 이런 웹 로그 트래킹을 측정하는 단위는 히트(Hits), 페이지뷰(Page View), 체류 시간(Duration Time), 세션(Session), 방문자(Visitor) 등이 있다. 로그 파일은 웹 서버가 지정하는 곳에 위치하며 웹 서버관리자는 웹 서버를 설치할 때 로그 파일의 위치와 기록방법 등을 지정할 수 있다.



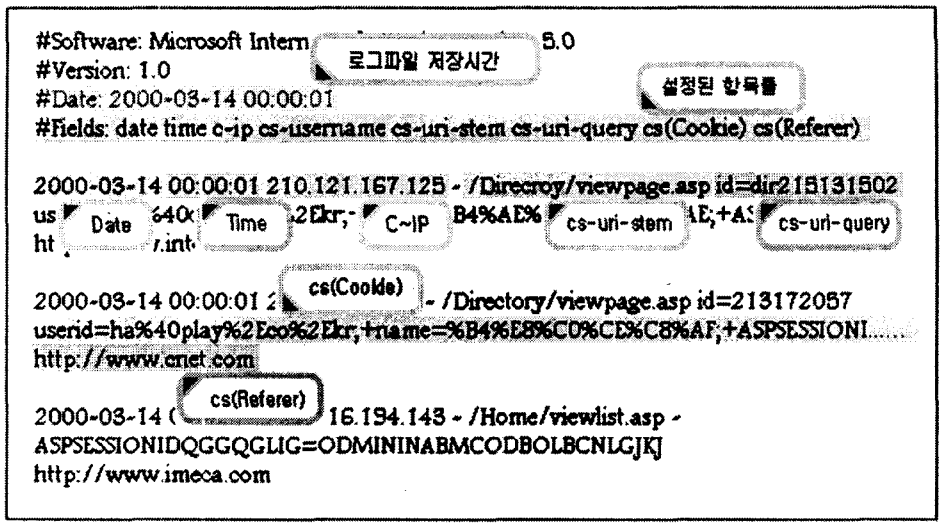
[그림 1] 웹 로그 데이터의 생성과정(출처 http://webpro.co.kr/log2.htm)

[그림 2]는 [그림 1]의 과정을 통하여 얻어진 웹 로그 데이터의 형식 중에서 IIS형식의 로그 데이터 예를 보여 주고 있다. 로그 파일의 형식에는 다음과 같이 크게 3가지로 구분할 수 있다.

1) 공통 로그 파일형식(CLF: Common Logfile Format) : CLF는 보통 Transfer 로그 파일 또는 Access 로그 파일로 불리며, 파일이름은 access_log 또는 access_log.1과 같이 저장된다. CLF는 웹 서버의 원조라고 할 수 있는 NCSA 계열의 웹 서버에서 기본으로 생성되는 로그 파일형식으로서 대부분의 웹 서버에서 지원하고 있다.

2) 인터넷정보서버(IIS: Internet Information Server)형식 : IIS는 Windows NT에서 가장 많이 사용되는 웹 서버 소프트웨어로서 자체적으로 분석도구를 제공하고 있다. 로그 파일형식은 NCSA 계열의 로그 파일과는 다르며, 파일의 기록기간을 단위별로(일별, 월별) 설정할 수 있고, 이는 IIS 관리자에서 실행할 수 있다.

3) WWW컨소시움(W3C: World Wide Web Consosium)형식 : W3C는 IIS의 확장된 로그 파일 형식으로서 사용자가 로그 항목의 위치와 내용을 지정할 수 있다. 사용자가 직접 로그 항목을 결정할 때 주의할 점은 로그 항목의 부족으로 인해 로그 분석이 어려워질 수 있다는 것이다. 따라서 W3C 로그 형식을 사용할 때는 시스템이 지정해 주는 초기값을 사용하는 것이 좋다.



[그림 2] IIS 로그 파일 형식(출처: http://www.machi.pe.kr/document/log_anal/log_5.htm)

로그 파일 형식의 예로서 CLF의 기록 예는 다음과 같다.

203.241.197.160 - - [12/May/2003:13:16:39 +0800] "GET /index.html HTTP/1.0" 200 2050.

위의 예는 다음과 같은 내용을 알려준다. 203.241.197.160이라는 IP주소를 가진 사용자가 2003년 5월 12일 16시 12분 39초(그리니치표준시로부터 8시간 떨어진 곳-KOREA)에 GET의 방법으로 /index.html에 대하여 서비스를 요청하였고, 이것은 HTTP 버전 1.0 프로토콜에 의해서 성공적으로(200) 연결이 되어 졌고, 이동한 총 데이터의 양은 2,050Byte임을 나타내고 있다.

2.2 웹 로그 데이터의 구성요소

웹 서버에 따라서 한 개가 아닌 여러 개의 로그 파일을 만들 수도 있다. 로그 파일은 다음과 같이 크게 4 가지로 분류할 수 있다(전성훈, 최영희, 2001).

2.2.1 Access 로그 파일

'Access 로그 파일'은 'Transfer 로그 파일'이라고도 하며 일반적인 사이트 방문 기록 등을 모두 기록하기 때문에 웹사이트 방문시간 및 방문경로 등을 파악할 수 있다. [그림 2]에서 Date, Time, C~IP, cs-uri-stem, cs(Cookie)는 2000년 3월 14일 0시0분 1초에 210.121.167.125에서 /Direcroy/viewpage.asp 페이지를 클릭했다는 것을 의미한다. Access Log는 웹 사용자들이 웹 서버에 접속했을 때 누가, 언제 웹사이트를 방문했는지를 알려주는 요소로서 [그림 2]에서 cs(Referer)를 제외한 부분이라고 할 수 있다.

2.2.2 Error 로그 파일

'Error 로그 파일'은 웹 서버에서 발생하는 모든 에러와 접속실패를 시간과 내용 두 가지로 기록한다. 즉, 홈페이지 관리자의 실수로 이미지가 깨지거나 링크를 잘못 연결한 경우 등에 의해서 발생하는 모든 에러와

접속실패 등을 필드에 기록한다. 이러한 'Error 로그 파일'을 토대로 웹사이트 컨텐츠 관리를 효율적으로 수행할 수 있는 관리 정보를 파악할 수 있다. 'Error 로그 파일'은 [그림 2]에는 나타나 있지 않지만 요청한 홈페이지가 없거나 링크가 잘못되는 등의 에러가 발생했을 경우에는 [그림 2]의 내용에 'Error 로그 파일'의 부분이 추가된다.

2.2.3 Referrer 로그 파일

'Referrer 로그 파일'은 사용 중인 웹 서버를 소개해준 사이트와 소개받는 페이지를 화살표로 기록하며 [그림 2]의 cs(Referer)가 포함되어 있는 마지막 부분에 해당되는 것으로써 cs(Referer)는 www.imeca.com에서 사이트에서 접속했음을 의미한다. 즉, 'Referer 로그 파일'은 웹 사용자들이 어느 경로를 통해 사이트에 접속했는가에 대한 정보를 제공한다. 이것은 어떤 웹사이트에 링크를 하고 있는 홈페이지의 주소를 파악하거나 검색 엔진으로부터 어떠한 키워드를 통해 어떤 방문자가 방문을 하였는지 등에 대한 정보를 알려준다. 따라서 'Referer 로그 파일'의 부분을 잘 파악함으로써 고객관리와 같은 마케팅에 적용할 수 있는 장점이 있다.

2.2.4 Agent 로그 파일

'Agent 로그 파일'은 사용자가 사용한 웹 브라우저와 운영체제(OS)에 대한 정보를 기록하는 것으로서 '브라우저 로그 파일'이라고도 한다. 'Agent 로그 파일'은 웹 브라우저의 이름, 버전, OS, 화면 해상도 등의 정보를 제공해 준다. 이것은 마케터와 개발자, 디자이너들에게 고객의 사용 환경을 이해시키고 보다 나은 인터페이스를 구현할 수 있도록 해준다. 'Agent 로그 파일'은 [그림 2]의 부분에서 나타나고 있지 않지만 웹 서버에서 'Agent 로그 파일'을 선택했을 경우 [그림 2]의 내용에 'Agent 로그 파일'의 부분이 추가된다.

3. WebBizi

WebBizi는 본 연구에서 개발한 웹 로그 분석기의 고유명칭이다. 'WebBizi'에 대한 기능 및 사용법은 4절에서 세부적으로 소개한다.

3.1 WebBizi의 구현환경

WebBizi는 Visual Basic으로 개발되었으며, WebBizi에 대한 구현환경은 [표 1]에 주어져 있다.

[표 1] 개발 도구

운영체제	웹 서버	개발언어
Window NT Server	IIS	Visual Basic

3.2 기능적인 측면

일반적으로 웹 로그 데이터에 대한 분석은 어떤 방문객들에 대한 특정 페이지에 대한 행동패턴을 알아내기 위한 것이다. 몇 년 전까지만 해도 기존의 웹 로그 분석기에 의한 분석은 주로 빈도분석을 위주로 실행되어 왔다. 최근에 들어서 일부 웹 로그 분석기들은 연관성분석과 군집분석 등 많은 기능들이 추가되고 있는 실정에 있다. 본 연구에서 개발한 WebBizi는 연관성분석과 분류분석의 기능을 강화하였는데, 특히 분류

분석은 기존의 분석기에 볼 수 없는 기능이다. 이 분석기능은 4.7절에서 소개되고 있는 바와 같이 방문자, 페이지, 히트, 에러수를 몇 개의 집단으로 분류시켜 주는 기능을 수행한다. 예를 들면, '히트수-사용자지정 분류분석'에서 사용자가 히트한 빈도수를 이용하여 몇 개의 그룹으로 분류하는 기능을 말하는데, 몇 개의 그룹으로 분류하는 것을 옵션으로 지정함으로써 원하는 집단 수로 분류할 수 있다. WebBizi는 웹 로그 데이터를 이용하여 페이지분석, 파일분석, 방문자분석, 에이전트분석, 에러분석, 연관성분석, 분류분석 등을 수행할 수 있도록 설계되어 있다. WebBizi에서 수행 가능한 기능들을 몇 가지만 간단히 소개하면 다음과 같다.

(1) '페이지분석' : a) 페이지뷰수 추이(시간별 페이지뷰수의 추이), b) 많이 방문한 페이지(페이지별로 많이 방문한 페이지), c) 오래 머문 페이지(사용자들이 페이지에서 머문 시간이 많은 페이지), d) 많이 시작한 페이지(사용자들이 홈페이지로 들어올 때 많이 들어오는 페이지), e) 많이 이탈한 페이지(사용자들이 홈페이지에서 이탈할 때 많이 이탈하는 페이지).

(2) '파일분석' : a) 히트수 추이(시간대별로 히트수 추이를 보여줌), b) 많이 히트된 파일(파일별로 많이 히트된 추이).

(3) '방문자분석' : a) 방문수 추이(시간대별로 방문수 추이를 보여줌), b) 페이지뷰수 높은 방문자(방문자 중 페이지 뷰수가 높은 방문자들을 보여줌), c) 히트수 높은 방문자(방문자 중에서 히트수가 높은 사람들 순으로 보여줌), d) 방문수 높은 방문자(방문자중 방문수가 높은 사람들을 보여줌), e) 에러경험 높은 방문자(방문자중 에러경험이 높은 사람들을 보여줌), f) 오래 머문 방문자(방문자중 홈페이지에 오래 머문 사용자들을 보여줌).

(4) '분류분석' : 데이터를 적절한 정의에 의해서 몇 개의 집단으로 분류하는 기능, 즉 특징에 맞게 방문자, 페이지, 히트, 에러수를 몇 개의 집단으로 분류시켜 주는 기능.

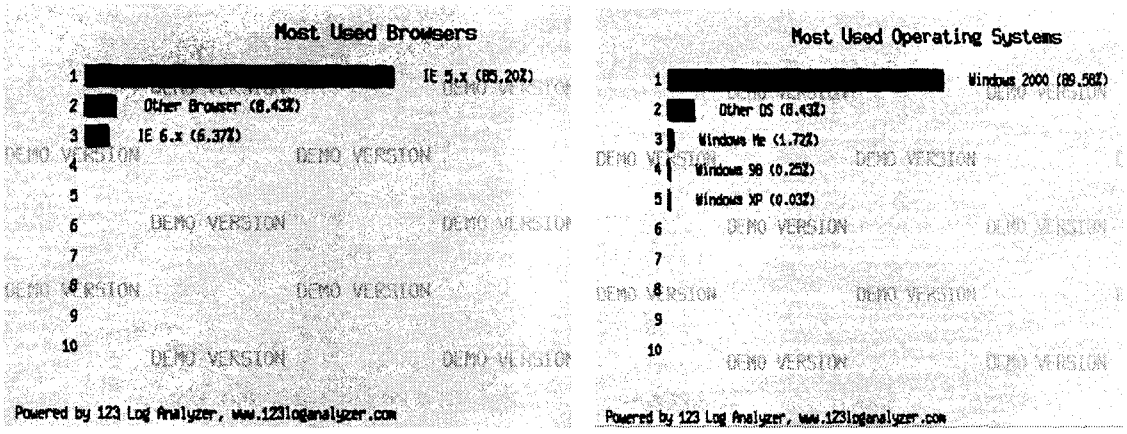
3.3 방법론적인 측면

본 절에서는 방법론적인 측면에서 WebBizi에 대한 특징을 소개하고자 한다. WebBizi는 기존의 분석기와는 달리 다음과 같은 내용으로 웹 로그 데이터의 효율적인 분석 방법에 초점을 맞추고 있다.

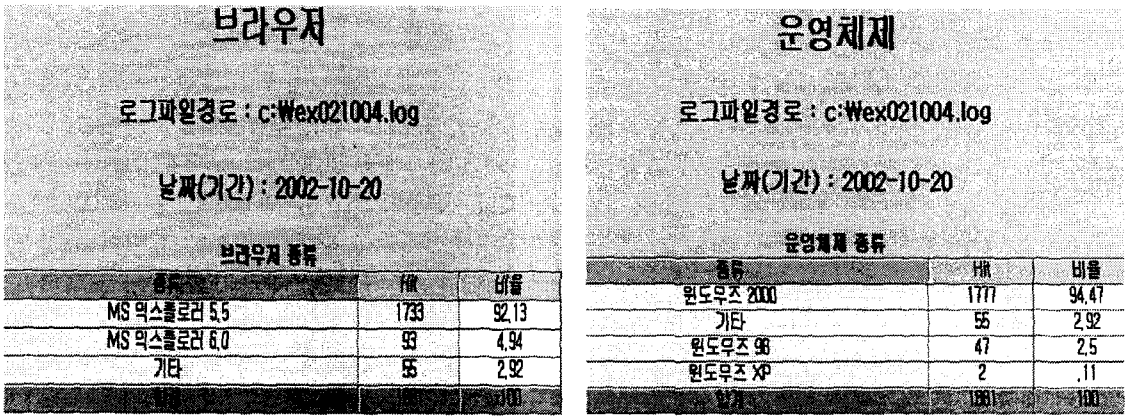
(1) 연관성 분류 분석법을 추가시켜 웹 로그 데이터에서 온라인 상의 사용자들의 행동패턴을 파악하고자 하였다. 또한 기존의 웹 로그 분석기는 웹 로그 데이터의 제약성으로 인하여 다양한 통계적 분석법을 적용하기에 아주 힘이 들었다. WebBizi는 이러한 문제점들을 보완하는 방법으로 파생변수를 추가시켜 보다 다양한 분석법들을 도입하였다.

(2) 처리속도측면에서 WebBizi는 기존의 웹 로그 분석기보다 빠르다는 장점을 가지고 있다. 기존의 웹 로그 분석기들 중에서 시스템 구축에 있어서 웹 데이터의 저장 방식을 데이터베이스로 저장하는 방법을 사용하는 경우가 있다. 이러한 경우에는 대용량의 웹 데이터를 처리할 수 있다는 측면에서 장점이 될 수도 있지만 웹 로그 데이터의 로딩시간이 많이 걸린다는 점과 데이터의 용량에 따른 부하가 발생할 수 있다는 단점을 있다. 따라서 WebBizi는 웹 로그 데이터를 파일로 저장함으로써 데이터베이스를 사용하여 저장하는 방법의 단점을 피하였다. 또한 WebBizi는 사용하고 있는 데이터 파일 형식의 측면에서 처리속도의 최적화를 위하여 IIS형식을 사용하였다. 이렇게 함으로써 데이터 분석 시 웹 로그 파일 포맷 형식에 대한 비교 연산 부분을 제거할 수 있기 때문에 IIS 웹 로그 파일 분석을 빠르게 수행할 수 있다. 따라서 본 연구에서 제안된 WebBizi는 아파치 형식의 웹 로그 데이터는 다룰 수 없다는 단점이 있기는 하지만 처리 속도적인 측면에서는 [표 2]에 주어져 있는 바와 같이 기존의 웹 로그 분석기에 비해 괄목할 정도로 빠르다는 것을 알 수 있다.

[그림 3]과 [그림 4]는 각각 기존의 분석기 log-analyzer(www.123loganalyzer.com)와 본 연구에서 개발한 분석 툴인 WebBizi를 사용하여 사용자들이 '많이 사용한 브라우저'와 '많이 사용한 운영체제'의 'Agent 로그 파일'을 분석하는 측면에서 분석된 결과를 보여주고 있다. 두 분석 툴에 의한 수행상의 비교결과에서 큰 차이를 보이고 있지 않음을 알 수 있다. WebBizi와 비교하기 위한 기존의 분석기로 log-analyzer를 선택한 이유는 무료로 제공되고 있어 손쉽게 사용할 수 있기 때문이다. 웹 로그 분석기는 어떤 알고리즘 또는 어떤 방법론으로 작성되었는가에 따라서 클릭 수는 다소 차이가 있을 수 있다.



[그림 3] log-analyzer 분석기에 의한 분석결과
(좌 : 많이 사용한 브라우저, 우 : 많이 사용한 운영체제)



[그림 4] WebBizi 분석기에 의한 분석결과
(좌 : 많이 사용한 브라우저, 우 : 많이 사용한 운영체제)

그리고 데이터의 분석 처리 시간(computing time)과 분석의 기능적인 측면에서 기존의 분석기와 비교한 결과가 [표 2]에 주어져 있다. 'WebBizi'는 기존의 툴과 처리속도를 비교했을 때, 주어진 로그 데이터에 대해서 전체 로그 분석 수행 소요시간이 'WebBizi'는 30초 소요된 반면, 기존의 분석 툴의 처리속도는 3분 정도가 소요되었다. 기존의 분석 툴보다 처리속도인 측면에서 빠르다는 점이 'WebBizi'의 가장 큰 장점으로 꼽을 수 있다. 그리고 기능적인 측면에서 기존의 분석 툴에서 가지고 있지 않은 연관성분석과 분류분석기능

을 추가하여 보다 유용한 분석이 가능하도록 고안되었다는 측면이 본 논문에서 제안한 분석 툴의 또 다른 장점이 될 수 있다. 참고로 WebBizi와 log-analyzer 모두 IIS 형식 로그 파일로 분석을 하였다. 'WebBizi'와 기존의 분석기(log-analyzer)를 비교하기 위하여 동의대학교 정보통계학과 학술정보사이트에 접속한 사용자들에 의해서 얻어진 웹 로그 데이터를 사용하였다.

[표 2] WebBizi와 기존의 분석기의 비교

구분	WebBizi	log-analyzer
속도	30초	3분
분석기능	연관성분석, 분류분석기능이 있음	연관성분석, 분류분석기능이 없음
로그 파일 분석형식	IIS 형식	IIS 형식

마지막으로 본 논문에서 제안한 툴의 타당성 검토를 위하여 분석결과를 수작업을 통하여 모든 기능에 대해서 확인한 결과 차이를 보이지 않았음을 밝혀 둔다.

3.4 통계적인 측면

WebBizi는 얻어진 정보(웹 로그 데이터)를 이용하여 웹 로그 분석을 통하여 사이트를 운영하고 있는 기업체에 보다 나은 운영과 의사결정을 위하여 효율적으로 사용할 수 있다. 예를 들면, 다양한 통계적 기법이 적용된 분석결과에 의하여 고객들을 세분화하여 경영자로 하여금 고객들을 어떻게 관리하는 것이 기업의 이윤을 극대화할 수 있는가를 조언함으로써 경영자의 의사결정에 효율성을 기할 수 있다. 또한 WebBizi는 최근 부각되고 있는 이메일을 이용하여 효율적으로 고객관리를 할 수 있는 인터넷 마케팅인 eCRM(electronic Customer Relationship Management)에 적극적으로 적용 가능하다는 점을 들 수 있다. 즉, WebBizi를 이용하여 구매가능도, 구매성향, 선호도조사 등 고객들의 성향정보를 수집하여 고객들의 패턴분석 등 다양한 통계적 분석을 통하여 마케팅 전략 수립한다. 이를 기초로 하여 이메일을 통해서 어떤 제품 또는 어떤 기업에 대한 광고, 홍보 및 다양한 프로모션활동을 수행하여 고객의 구매동기를 유발시키거나 기업에 대한 인식을 새롭게 할 수 있는 계기를 만들 수 있을 것이다. 이러한 측면에서 볼 때, WebBizi 뿐만 아니라 웹 로그 분석기의 개발에 대한 연구는 분석기의 기능 자체가 통계와 밀접한 관련이 있기 때문에 보다 나은 기능들을 연구해 나감으로써 사회가 요구하는 바에 부응할 수 있으리라 기대되며, 통계적으로 매우 의미가 있다고 생각한다. 따라서 웹 로그 분석기의 개발에 대한 연구는 전산의 한 분야인 웹 서버와의 연계를 통한 통계분석의 응용성을 제시하고 있고, 요즘 부각되고 있는 웹 로그 데이터와 통계적 기법의 결합에 의한 통계의 한 응용분야로서 가능성을 보여 주고 있다고 할 수 있다.

4. WebBizi에 의한 로그 데이터 분석

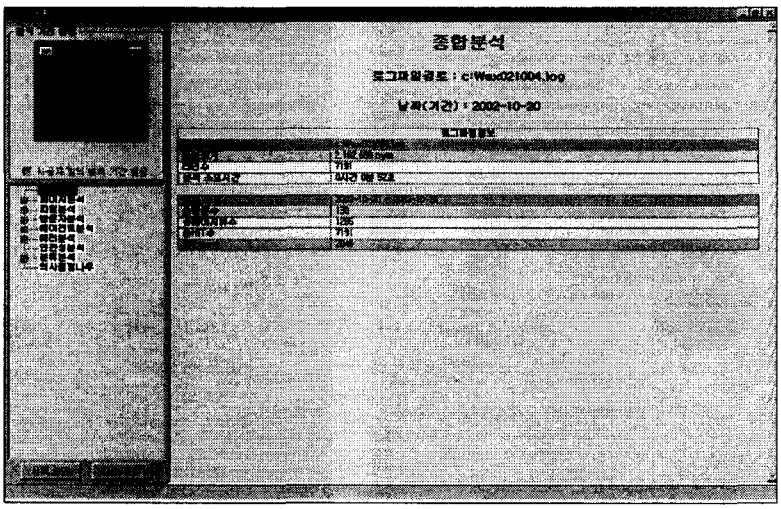
본 절에서는 WebBizi의 사용법 및 기능을 설명하기 위하여 실제 사이트를 개설하여 얻어진 웹 로그 데이터를 이용하여 웹 로그 분석을 실행해 본다. 지면 상 WebBizi가 수행할 수 있는 모든 기능을 소개할 수는 없기 때문에 단지 몇몇의 기능만을 소개한다.

4.1 분석 데이터

개발한 분석기의 사용법을 설명하기 위하여 사용된 데이터는 동의대학교 정보통계학과에 재학 중인 400명의 학생 중에서 개설된 학술정보 사이트에 가입한 161명을 대상으로 2002년 3월 30일부터 6월 27일까지 운용하여 얻어진 웹 로그 파일이다(최승배, 임승범, 2002). 분석 데이터의 크기는 2M이고, 사용된 운영체제는 윈도우 2000이다.

4.2 종합분석

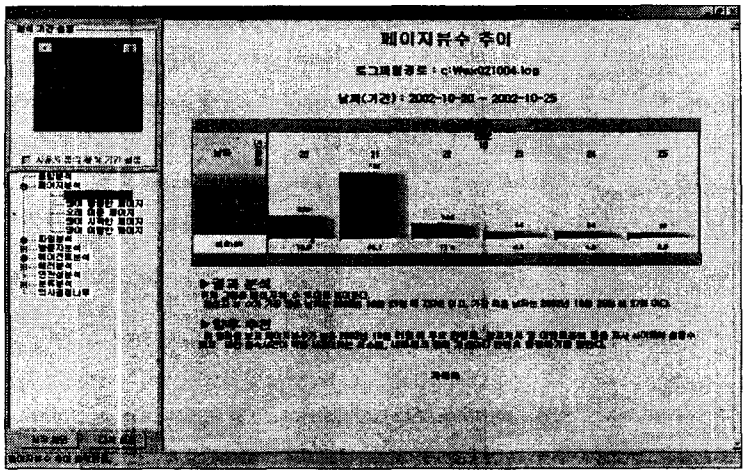
종합분석은 '파일의 경로', '파일 크기', '파일의 line수', '분석 시간'과 같은 파일의 세부사항을 보여주고, '분석기간', '총방문 수', '총페이지뷰 수', '총히트 수', '총에러 수'와 같은 기본적인 분석결과를 보여준다.



[그림 5] 결과 출력화면

4.3 페이지분석

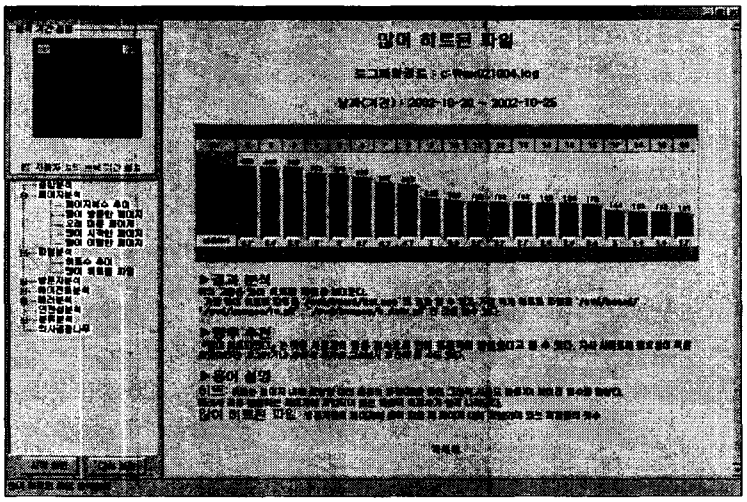
페이지 분석에는 '페이지뷰수 추이', '많이 방문한 페이지', '오래 머문 페이지', '많이 시작한 페이지', '많이 이탈한 페이지'의 결과를 볼 수 있는데, 여기에서는 [그림 6]에서 보여주고 있는 것과 같이 페이지뷰수 추이에 대해서만 소개하도록 한다. 페이지뷰수 추이는 시간대(하루일 경우는 24시간, 이틀 이상일 경우에는 일자로 나온다)별로 보여 준다. 이러한 결과를 통하여 어떤 시간대에 어떤 페이지에 접속이 많았는지에 대한 분포를 알 수 있다. 또한 [그림 6]의 화면 하단에 있는 아래의 '자세히' 버튼을 클릭함으로써 분석일자에 대한 시간대의 페이지 정보를 알 수 있다.



[그림 6] 페이지뷰수 추이

4.4 파일분석

파일분석은 '히트수 추이', '많이 히트된 파일'의 결과가 나타난다. '히트 수 추이'는 시간대(하루 일 경우는 24시간, 이틀 이상일 경우에는 일자로 나타난다)별로 보여 준다. 이를 통하여 어떤 시간대에 어떤 페이지에 히트가 많은지에 대한 분포를 알 수 있다. '많이 히트된 파일'은 페이지에 있는 히트가 많이 방문했다는 것을 알 수 있다. 만약 히트가 많이 발생하여 트래픽 및 머문 시간이 길어진다면 필요성이 적은 히트를 줄이는 것이 바람직할 것이다. 아래의 '자세히' 버튼을 클릭하면 표의 번호를 통해 많이 히트된 파일의 정보를 알 수 있다. [그림 7]은 '많이 히트된 파일'에 대한 수행결과를 보여 주고 있다.

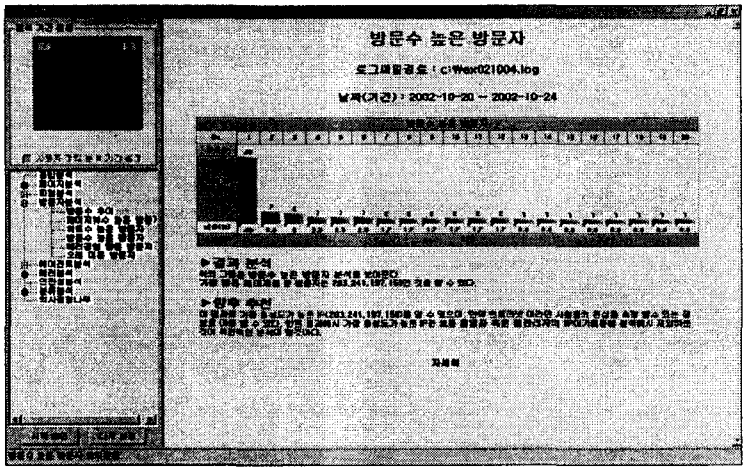


[그림 7] 많이 히트된 파일

4.5 방문자분석

방문자분석에는 '방문수 추이', '페이지뷰수 높은 방문자', '히트수 높은 방문자', '방문수 높은 방문자', '에러경험 높은 방문자', '오래 머문 방문자'의 결과를 보여주는 기능을 갖고 있다. '방문수 추이'는 시간대(하루

일 경우는 24시간, 이틀 이상일 경우에는 일자로 나타난다)별로 보여준다. 이 기능을 통하여 어떤 날짜의 어떤 시간대에 방문자가 많았는지에 대한 분포를 알 수 있다. '페이지뷰수 높은 방문자'는 어떠한 방문자가 많이 접속하였는지를 보여준다. 이를 통하여 자주 방문하는 고객(충성도 높은 고객)을 알 수 있다. '히트수 높은 방문자'는 어떠한 방문자가 많이 접속하였는지를 보여준다. 이 기능은 자주 방문하는 고객을 보여 준다. '방문수 높은 방문자' 또한 어떠한 방문자가 많이 접속하였는지를 보여준다. '에러경험 많은 방문자'는 어떤 방문자가 사이트에 접속하여 에러가 많았는지를 보여준다. '에러경험 많은 방문자' 화면의 하단에 있는 '자세히' 버튼을 클릭하면 표의 번호를 통해 에러 방문자의 IP 정보를 알 수 있다. '오래 머문 방문자'는 각 방문자가 얼마나 머물었는지를 보여준다. 머문 시간이 긴 사용자는 음악과 같은 용량이 큰 파일 등이 있는 페이지를 오랫동안 보았을 것으로 판단할 수 있다. 그렇지 않은데도 머문 시간이 길다면 서버의 과부화로 인하여 발생하는 트래픽일 수 있기 때문에 주의 깊게 검토해 보아야 할 것이다. 여기서도 화면의 하단에 있는 '자세히' 버튼을 클릭함으로써 오래 머문 사용자의 IP정보를 알 수 있다. [그림 8]은 '방문수 높은 방문자'에 대한 결과만을 보여 주고 있다.



[그림 8] 방문수 높은 방문자

4.6 연관성분석

방문자가 자사 사이트를 웹 서핑할 때 두 페이지간의 관계를 보여주는 분석이다. 연관성에 대한 척도로서 다음과 같은 것들이 있다(강현철 외, 2002).

(1) 지지도 : 두 페이지가 연결되었을 때의 확률로서 관련성이 있다고 판단되는 품목들을 포함하고 있는 거래나 사건의 확률을 의미하며, 다음의 공식에 의해서 얻어진다.

$$\text{지지도} = \frac{\text{품목 } A\text{와 } B\text{를 동시에 포함되는 거래수}}{\text{전체 거래수}}$$

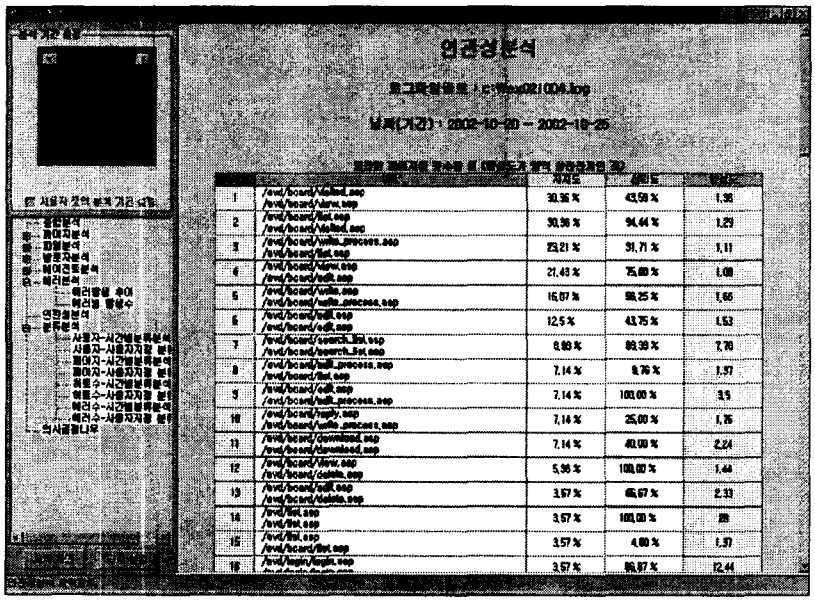
(2) 신뢰도 : 두 페이지가 떨어져 있을 경우의 확률로서 품목 A를 구매하였을 경우 품목 B를 구매할 가능성으로 조건부확률 P(B|A)로 나타낼 수 있고 대칭적이지는 않다. 이 신뢰도는 다음과 같은 공식에 의해서 얻어진다.

$$\text{신뢰도} = \frac{\text{품목 } A\text{와 } B\text{를 동시에 포함하는 거래수}}{\text{품목 } A\text{를 포함하는 거래수}}$$

(3) 향상도 : 1보다 크다면 두 페이지는 밀접한 관계를 가지고 있고, 1보다 작으면 두 페이지는 떨어지는 것이 자사 사이트에 좀 더 좋은 정보를 준다고 할 수 있다. 향상도는 실제의 신뢰도를 독립가정 하에서의 신뢰도로 나눈 값으로 다음에 의해서 얻어진다.

$$\text{향상도} = \frac{\text{품목 } A\text{와 } B\text{를 동시에 포함되는 거래수}}{\text{품목 } A\text{를 포함하는 거래수} \times \text{품목 } B\text{를 포함하는 거래수}}$$

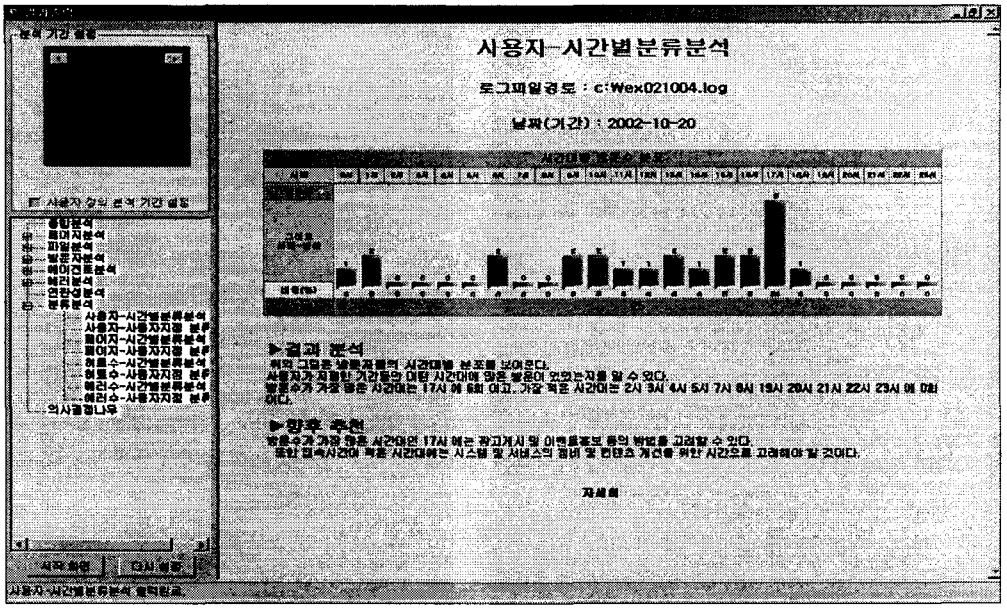
[그림 9]는 연관성분석에 대한 결과를 보여주고 있다.



[그림 9] 연관성분석

4.7 분류분석

분류분석은 방문자, 페이지, 히트, 에러수를 몇 개의 집단으로 분류시켜 주는 기능을 수행한다. 분류분석을 실행하면 분류분석 프로그램 자체에서 적절한 분류개수로 나누어준다. 또한 사용자가 직접 분류개수를 지정하여 데이터를 분류할 수 있도록 설계되어 있다. '분류분석'에는 '사용자-시간별 분류분석', '사용자-사용자지정 분류분석', '페이지-시간별 분류분석', '페이지-사용자지정 분류분석', '히트수-시간별 분류분석', '히트수-사용자지정 분류분석', '에러수-시간별 분류분석', '에러수-사용자지정' 분류분석을 수행할 수 있는 기능들이 있다. [그림 10]은 '사용자-시간별 분류분석'에 대한 출력결과를 보여 주고 있다.



[그림 10] 분류분석 출력결과(사용자-시간별 분류분석)

5. 결론

지금까지 동의대학교 정보통계학과의 학술정보사이트에서 얻어진 웹 로그 데이터를 적용하여 새롭게 개발한 웹 로그 분석기(WebBizi)를 사용하여 분석하여 보았고, 그것에 대한 특징 및 사용법에 대해서 간략히 설명하였다. 본 연구에서 개발한 웹 로그 분석기를 기존의 분석기와 비교하는 자체가 무리가 있다고 생각되지만 수행시간(computing time)과 기능적인 측면에서 기존의 분석기(log-analyzer)보다 뛰어나다는 장점을 가지고 있음을 볼 수 있었다. 본 연구는 앞으로 웹 서버에서 기본적으로 제공하는 웹 로그 데이터만이 아닌 인터넷 유저들의 개인 프로파일과의 연동을 통하여 보다 나은 질적인 분석을 위한 연구가 필요하리라고 생각된다. 그리고 본 연구에서 소개한 WebBizi에 기능적인 측면에서 보다 많은 보완이 필요하고 이에 대해서는 차후 연구과제로 남겨 두고자 한다. 또한 본 논문에서 소개한 분석 툴은 기능적인 측면에서 좀 더 보완만 한다면 사이트를 운영하고 있는 일반 기업들에게 유용한 정보를 제공할 수 있을 것으로 생각된다.

감사의 글

본 연구는 기술적인 면과 구성적인 면 등에서 많은 문제점이 있었다. 그럼에도 불구하고 자세하게 지적해 주신 결과, 보다 좋은 내용으로 본 연구를 거듭나게 해 주신 심사위원님들께 감사 드린다.

참고문헌

[1] 강창완, 김규곤, 최승배, 정민석, 박광준, 손승한, 임승범, 정경미 (2001). 구인/구직 웹사이트에 대한 웹 로그 분석, 「Journal of the Korean Data Analysis Society」, 2001, Vol.3, No.3, pp 331-342.

[2] 강현철, 한상태, 최종후, 김은석, 김미경 (2002). 「SAS Enterprise Miner 4.0을 이용한 데이터 마이닝 - 방법론 및 활용 -」, 자유아카데미.
[3] 김석기, 안정용, 한경수 (2001). 웹로그(Web Log) 데이터 분석 방법에 관한 연구, 「응용통계연구」, 14, 2, pp. 261-271.
[4] 전성훈, 최현회 (2001). 「eCRM실무지침」, 삼각형M&B.
[5] 전옥선 (2002). 효율적인 로그분석을 활용한 eCRM 마케팅전략, 「삼성 솔루션서비스그룹」, 보고서.
[6] 최승배, 임승범 (2002). 웹 마이닝 기법을 이용한 학술 정보 사이트 분석, 「Journal of the Korean Data Analysis Society」, Vol. 4, No. 4, pp 451-464.
[7] Michael Berry & Gordon Linoff (1997). Data Mining Techniques for Marketion, Sales, and Customer Support, Wiley Computer Publishing, John Wiley & Sons, Inc..
[8] Ralph Kimball (1996). The Data Warehouse Toolkit, Wiley Computer Publishing, John Wiley & Sons, Inc..
[9] Christopher Adamson & Michael Venerable (1998). Data Warehouse Design Solutions, Wiley Computer Publishing, John Wiley & Sons, Inc..
[10] www.1231loganalyzer.com.
[11] www.innerbus.com.
[12] www.nethru.co.kr.
[13] http://webpro.co.kr/log2.htm.

[2003년 6월 접수, 2004년 2월 채택]