로그파일 분석 능력과 한계 습득 필요

라이브러리 웹사이트를 만들고 유지하는데 시간과 돈과 모든 노력을 투자하고 난 다음에는 그곳을 이용하고 있는 사람들이 누구인가 하는 것이 가장 알고 싶을 것이다. 그에 더해 방문자들이 가장 자주 이용하는 기능과 자원이 무엇인지도 궁금할 것이다. 어떤 노력을 해야 하는지 살펴봤다. <편집자>

이들은 사용량에 대한 문제이고 도서관원들은 이미 사용량 데이터를 수집해본 경험을 가지고 있다. 예를 들어 도서관원들은 이용횟수를 계측하기 위한 방법으로 조회 데스크에 문의한 질문수를 세어본다. 조회 데스크와 마찬가지로 라이브러리 웹사이트는 서비스 포인트를 나타낸다.

라이브러리 웹사이트를 만들고 유지하는데 시간과 돈과 모든 노력을 투자하고 난 다음에는 그곳을 이용하고 있는 사람들이 누구인가 하는 것이 가장 알고 싶을 것이다. 그에 더해 방문자들이 가장 자주 이용하는 기능과 자원이 무엇인지도 궁금할 것이다. 사람들이 어떻게 웹사이트를 찾았을까? 그들이 검색엔진을 사용했을까?

그러나 웹사이트 서비스 포인트는 전자식이어서 사용량을 계측하는 새로운 방법이 요구된다. 웹서버 테크놀러지와 데이터 서버 레코드의 기초를 이해하는 것은 사용량 계측 기법을 발전시키는데 있어 좋은 출발점이다. 그 다음에 웹사이트 통계를 이해하는데 도움이 되는 기존의 소프트웨어를 조사해보고 각 시스템에 적합한 소프트웨어를 찾아볼 수 있다.

웹서버 로그 파일들

인터넷상에서의 모든 거래는 브라우저 클라이언트로부터의 요구와 컴퓨터 서버로부터의 대응 활동으로 이루어져 있다. 개별 클라이언트/서버 거래는 서버 로그 파일이라고 불리는 서버 상에 기록된다. 그것의 가장 기본적인 형식을 공통 로그 파일이라 부른다. (주의: 다음 보기에서 로그 기재사항들은 대체로 한 줄의 텍스트로 나타난다.)

공통 로그 파일(Common Log File)

공통 로그 파일 포맷은 월드 와이드 웹 협회에서 정한 기준이다. 공통 로그 파일의 기재사항 구문은 다음과 같이 나타난다.

remotehost rfc931 authuser [date] "request" status bytes

공통 로그 파일의 각 구성요소를 나누어 보면 그 자체의 의미를 가지고 있다.

- remotehost: 웹서버에 접속한 컴퓨터의 이름
- rfc931: 원격 사용자의 이름. 이 필드는 비어있는 경우가 종종 있다.
- authuser: 원격 사용자의 로그인. 이 필드는 비어있는 경우가 종종 있다.
- [date]: 접속을 요청한 날짜와 시간.
- Status: 요청에 의해 생성되어 클라이언트에게 반환된 오류 코드.
- Bytes: 클라이언트에게 응답한 문서의 바이트 크기.

전형적인 로그 기재사항은 이와 같이 나타날 것이다.

gateway.iso.com - - [10/MAY/1999:00:10:30 -000] "GET/class.html HTTP/1.1" 200 10000

위 보기에서 원격 호스트는 gateway.iso.com이다. 그 다음 두 필드, rfc931필드와 authuser필드는 비어있다(위에서는 대시로 나타나 있다). 요청이 있었던 것은 1999년 5월 10일 밤 12시 10분이었다. 요청된 파일은 class.html이었다. 반환된 오류 코드는 200(상태 좋음)이었고, 요청된 파일의 크기는 10,000 바이트였다.

공통 로그 파일이 표준이 될 수 있기는 하지만, 로그 파일들의 편차가 존재한다. 리퍼러 로그 파일이나 에이전트 로그 파일에 추가 정보가 저장되어 있을 수 있다.

위탁자 로그 파일(Referrer Log File)

많은 서버들은 위탁자 사이트에 대한 정보를 기록한다. 즉 방문자가 현재의 웹사이트에 있는 어떤 페이지를 요청하기 직전에 방문했던 곳의 URL에 대한 정보를 기록한다.

08/02/99, 12:02:35, http://ink.yahoo.com/bin/query?p=" sample+log+file" &b=21&hc=0&hs=0, 999.999.999.99, jaz.med.yale.edu

위의 보기에서 조회 페이지는 검색엔진, ink.yahoo.com이었고, 요청된 페이지를 찾는데 사용된 검색은 "sample log file"이었다(많은 웹디자이너들과 마케팅 담당자들은 사용자들을 자신들의 사이트로 이끌어 주는 검색어에 관심을 가지고 있다). 요청을 한 컴퓨터의 인터넷 프로토콜(IP) 주소, 999.999.999.99도 이곳에 기록되어 있다는 점에 주목하기 바란다.

에이전트 로그 파일(Agent Log File)

세번째 유형의 기록은 에이전트 로그이다. 에이전트 로그는 브라우저와 방문자가 사용한 운영체계를 기록한다. 또한 여기에는 귀하의 웹사이트를 찾아내는데 사용된 스파이더의 이름이나 로봇의 이름이 기록될 것이다. Northern Light 검색엔진으로부터의 적중자료의 예가 에이전트 로그에 기록되면 다음과 같을 것이다.

07/09/99, 13:59:24, , 999.999.99.99, scooby.northernlight.com,
crawler@northernlight.com, Gulliver/1.2

날짜, 시간, 인터넷 프로토콜(IP) 주소에 대한 표준정보에 더하여 crawler@northernlight.com은 이 적중자료가 crawler 사이트에서 왔다는 것을 말해주고 있다. 어떤 웹 브라우저에서 출원된 적중자료에는 모질라 4.0과 같은 브라우저명과 버전이 나타나 있다. 이것은 방문자의 브라우저가 넷스케이프 4.0이었음을 의미하는 것이다(모질라는 넷스케이프의 코드명이었고, 지금도 개방 소스 넷스케이프 코드를 따르는 브라우저에 사용되고 있다). 그러나 브라우저 정보는 항상 신뢰할 수 있는 것은 아니다.

공통 로그 파일, 위탁자 로그 파일, 에이전트 로그 파일들은 때로 하나의 로그 파일로 결합되어 있다. 귀하의 웹 서버가 사용하고 있는 포맷이 무엇이든지 간에, 귀하가 맨 먼저 해야할 것은 어떤 유형의 로그 파일을 생성시킬 것인가를 결정하는 일이다. 서버 책임자는 사용된 포맷이 무엇인지 귀하에게 말해 줄 수 있어야 한다. 그에 더해, 어떤 데이터를 기록할 것인가를 결정하는 옵션이 로그 파일에 있을 것이다. 그리고 필요에 따라 이들 옵션을 이용하여 수집된 데이터를 늘리거나 줄일 수 있다.

웹서버의 로그 파일들을 통해 알 수 있는 것들 몇 가지를 요약해 보자.

- 귀하의 사이트의 어떤 페이지가 요청되었나.
- 요청한 컴퓨터의 IP 주소.
- 요청한 날짜와 시간.
- 파일 전송이 성공했는지 여부.
- 귀하의 사이트로 오기 전에 요청자가 방문했던 마지막 페이지.
- 귀하로 사이트로 이끈 검색어.

로그 파일의 한계

로그 파일들은 서버 관리자들이 서버요청을 측정하는데 도움이 되도록 고안되었다. 그들은 이것을 하는데 아주 능숙하다. 그러나 로그 파일들은 사람들이 사이트를 이용하는 방법을 기술하도록 고안되어 있지는 않다.

예를 들어 세 명의 다른 사람들이 하나의 파일을 세 번 요청한 것인지 아니면 한 사람이 그 파일을 세 번 요청했는지를 항상 구별할 수 있는 것은 아니다. 이것은 사람들이 인터넷 서비스 제공업체(ISP)에 전화접속을 하여 인터넷을 이용하는 것이 일반적이기 때문이다. 동적 번지지정 방식에서 인터넷 서비스 제공업체는 인터넷에 연결하고 있는 사용자에게 IP 주소를 할당한다. 그러나 처음 사용자가 접속을 끊으면 그 IP를 새로운 사용자에게 다시 할당할 것이다.

정적 번지지정 방식으로 인터넷에 직접 연결하지 않는다면, IP를 통해 개인을 식별할 수는 없다. 어떤 로그 분석 소프트웨어는 짧은 시간동안 같은 IP 주소로부터의 요청들을 조사함으로써 어떤 사이트의 개별 방문자를 식별해 보려 한다. 그러나 아직까지 그 다음날 그 사람이 다른 IP로 다시 연결할 때 구별하지는 못한다.

고속기억작용이 어떻게 로그 파일들에 영향을 미치는가에 대한 관심이 있다. 고속기억작용이 일어나는 것은 웹 페이지에 방문하여 브라우저가 그 페이지를 기억장치에 저장할 때이다. 다음 번에 같은 웹주소(URL)를 요청하면, 브라우저가 기억장치에서 그 URL를 찾아낼 것이다. 그것이 저장되어 있다면, 서버에 요청하지 않고 기억장치에서 그 페이지를 찾아낼 것이다.

그러므로 그 사이트를 이용하고 있지만 그 서버의 로그 파일들에는 기록되지 않을 것이다. 인터넷 서비스 제공업체들도 고속기억장치를 이용하고 있다. 그것 때문에 문제가 더 심각하다.

주의해야 할 것 한 가지는 로그 파일은 서버 상의 특정 파일들에 대한 요청들을 측정하는 것이지 정확한 사용량을 측정하는 것이 아니라는 것이다. 요청 횟수는 개별 방문자들의 수로 바꾸어질 수 없으며, 그 숫자도 고속기억작용 때문에 사용량을 전부 반영하지는 못한다. 사용량 계측은 로그 파일에 나타나 있는 것으로부터 외삽법에 의한 추정을 요구하고 있기 때문에 어느 정도의 오차가 있음을 의미한다. 웹사이트의 사용량에 대한 보다 정확한 정보를 얻고자 한다면, 앙케이트나 쿠기와 같은 다른 조사 수단을 써야한다.

프라이버시에 대한 배려

로그 파일의 한계에 관한 좋은 소식은 이들이 사용자의 프라이버시를 적어도 어느 정도는 보호하고 있다는 것이다. 로그 파일에는 사용자의 이름, 집, 전자우편 주소가 절대로 기록되지 않는다. 그런 정보는 웹사이트에서 사용자에게 등록을 요청하고 나서, 그 다음에 방문하기 위해 매번 로그인을 요청할 경우에만 기록될 수 있다.

이렇게 한 사이트는 그 후에 사용자 프로파일 데이터베이스에 로그 파일들을 연결하여 개별 사용량에 대한 보고서를 만들 수 있다. 로그 파일들만을 분석하는 것은 여러 그룹의 사용자들에 대한 데이터를 제공할 수 있을 뿐이다.

동적 번지지정은 특정 IP 주소로 연결하고 있지 않기 때문에 개별 사용자들을 드러내지 않는 다는 것을 기억하자. 그러나 인터넷에 직접 연결한 사람은 누구든지 변하지 않는 특정 IP 주소를 가지게 된다. 비록 그들의 이름이 기록되지는 않지만, 어떤 개인의 IP 주소에 접속하면 인터 상에서 그들이 어떤 활동을 하는지 나타난다.

현재로서는 로그 파일에 들어있는 정보를 어떻게 처리할 것인지를 규정한 법은 없다. 그러나 로그 파일들에 개인의 IP 주소에 대한 정보가 들어 갈 수 있기 때문에, 대출 기록들이 기밀인 것과 마찬가지로 이들도 기밀 자료로 간주되어야 한다. 라이브러리에서 로그 파일들을 이용하여 발표하는 어떤 데이터에도 개인의 IP 주소는 드러나지 않아야 한다. 데이터는 항상 특정 국가의 사용자들이나 조직내 사용자 대 조직외 사용자와 같은 대규모 그룹들의 사용량 수준에서 발표되어야 한다.

귀하의 라이브러리에서 로그 파일들을 어떻게 분석하기로 결정을 하든지 간에 라이브러리 웹사이트에는 어떤 데이터가 수집되어 있고, 그 데이터를 볼 수 있는 사람들은 누구이며, 그 데이터가 어떻게 이용될 수 있는지에 대한 일람표를 전부 수록해야 한다.

로그 분석 소프트웨어의 선택

로그 파일들은 수천 줄의 텍스트로 이루어져 있다. 단순히 읽기만 해서는 유용한 정보를 발췌해낼 수 없다. 로그 파일들을 분석하려면 스프레드시트나 각 사이트에 맞게 특별히 고안된 데이터베이스로 그 목록들을 다운로드 해야한다. 이런 유형의 응용 프로그램을 만드는 것은 시간이 많이 걸리고 아주 어렵다. 로그 파일에 포함되어 있는 데이터를 처리하고 분석할 수 있도록 고안된 소프트웨어를 사용하는 것이 그 파일들을 분석하기 위한 훨씬 더 일반적인 접근 방법이다.

분석 소프트웨어를 고려하기 전에 각 라이브러리의 웹사이트 사용량에 대해 정말로 알고자 하는 것이 무엇인지 확인해야 한다. 로그 분석에 이용할 수 있는 소프트웨어에는 무료 소프트웨어와 상업적 소프트웨어가 있다. 이들 각자는 장점과 단점을 가지고 있다.

일반적으로 상업적 소프트웨어는 기능이 더 많고, 그래픽 기능이 강화되어 있으며, 어떤 수준의 고객 지원을 제공하고 있다. 만약 필요한 것이 그리 복잡한 것이 아니라면 모든 기능을 다 탑재한 분석 패키지보다는 기능이 보다 단순하고 값이 더 저렴한 대안품이 적합할 것이다.

로그 분석 소프트웨어는 로그 파일에서 정보를 수집, 발췌, 표시하는데 도움을 줄 수는 있다. 그러나 그 소프트웨어가 아무리 정교하다 하더라도 그 로그 파일 내에서 기존에 이용할 수 있었던 것에 내용을 더하거나 개선할 수는 없다. 로그 파일의 목록들은 로그 분석 소프트웨어가 무엇을 할 수 있는가 하는 것에 있어서 근본적인 제한 요소이다.

무료 로그 분석 소프트웨어

명칭: 아날로그 3.31
제작자: 케임브리지 대학 통계 연구소, 스테판 터너.
URL: http://www.statslab.cam.ac.uk/~Esret1 /analog/
가격: 무료
고객 지원: 없음.
로그 파일 포맷: 여러 포맷을 지원하기 위해 구성가능.
플랫폼: 윈도, 매킨토시, 유닉스, 기타.

아날로그는 스테판 터너에 의해 개발된 아주 인기 있는 무료 로그 분석 소프트웨어이다. 아날로그는 사용자의 규격에 맞게 구성할 수 있는 표준 보고서를 제공한다. 그리고 웹 서버에 대한 요청인 일반적 개요(General Summary)를 제공한다.

중요한 기능은 요청 보고서(Request Report)이다. 이 보고서는 웹사이트 상에서 요청을 가장 많이 받은 웹페이지들을 표시하

고, 요청을 가장 많이 받은 것부터 가장 적게 받은 것을 표시한다. 요청 보고서는 요청수, 파일을 요청한 마지막 날짜, 파일명을 기재한다.

일반적 개요, 요청 보고서 이외에도 아날로그는 월별, 일별, 시간별 개요를 표시해준다. 이렇게 함으로써 가장 바쁜 달과 주중 가장 바쁜 날, 하루 중 가장 바쁜 시간을 아는데 도움이 될 수 있다. 또한 아날로그는 서버의 웹페이지들을 가장 많이 요청한 컴퓨터의 도메인명을 표시할 수 있다. 예를 들어 아날로그는 어떤 사이트를 방문한 사람들의 수를 알아내기 위해 아무런 시도도 하지 않고 미국의 학술 사이트들에서 35%의 요청이 왔다는 것을 나타낼 수 있다.

아날로그는 널리 이용되고 있다. 아날로그는 다양한 플랫폼에서 운용되고 있고 많은 로그 파일을 인지할 수 있다. 그러나 고급 그래픽 기능을 제공하지 않는다. 아래의 두 제품들도 무료로 인터넷에서 다운 받을 수 있다.

명칭: wwwstat
URL: http://www.ics.uci.edu/ pub/Websoft/wwwstat/
제작자: 로이 필딩
가격: 무료
고객 지원: 없음.
로그 파일 포맷: 공통 로그 파일 포맷
플랫폼: 유닉스

명칭: http-분석기(Analyzer) 2.01
URL: http://www.netstore.de/Supply/http-analyze/index.html
제작자: 렌트 아 구루
가격: 교육용이나 개인용은 무료.
고객 지원: 없음.
로그 파일 포맷: 공통 로그 파일, 일부 확장 로그 파일 포맷.
플랫폼: 유닉스

상업적 로그 분석 소프트웨어

명칭: 웹 트렌즈 로그 분석기
URL: www.Webtrends.com
제작자: 웹트렌즈社
가격: 399달러, 이 이상의 값비싼 제품도 이용 가능함.
고객 지원: 전화, 전자우편, 웹서버, 온라인 FAQ와 문서에 의한 무료 기술 지원.
시험기간: 14일간 무료.
로그 파일 포맷: 30개의 로그 파일 포맷을 인식.
플랫폼: 윈도 95/98/NT

웹트렌즈는 로그 분석과정을 단순화 하고자 한 강력한 소프트웨어 팩키지이다. 로그 프로파일과 보고서들은 메뉴에 따라 조작하는 구조의 시스템에서 만들어지고 편집된다. 이 과정을 쉽게 할 수 있도록 마법사와 온라인 도움말을 이용할 수 있다. 웹트렌즈를 이용하면 몇 개의 서버를 망라하는 다중의 로그 파일들을 다룰 수 있다.

개별화 된 보고서를 만드는 일은 보고서 마법사를 이용하면 쉽게 할 수 있다. 보고서 생성 모듈일 때 일반 통계, 호출된 자원, 방문자들과 인구 통계, 활동 통계, 기술 통계, 리퍼러/키워드, 브라우저/플랫폼 섹션에서 표와 그래픽 생성을 결정할 수 있다.

표와 그래프를 포함시키는 것은 마법사 과정에서 박스에 체크만 하면 되기 때문에 쉽다. 그래프는 파이 챠트, 막대 그래프나 선 그래프로 좀 더 개별화시킬 수 있다. 보고서는 HTML, 마이크로소프트 워드, 마이크로소프트 엑셀 문서로 생성될 수 있다.

웹트렌즈 보고서들 중 어떤 것들은 무료 소프트웨어에서 제공된 것과 유사하다. 예를 들어 웹트렌즈는 웹사이트에서 가장 많은 요청을 받는 페이지들 중에서 보고서를 만들 것이다. 그래프가 포함되어 있고 그 파일의 주소들을 제목에 의해 구별할 수 있다는데 주목하자.

웹트렌즈에는 무료 소프트웨어에서 이용할 수 있는 것보다 더 많은 보고서들이 있다. 웹트렌즈는 호출된 자원 영역 단독으로 입구 페이지, 출구 페이지, 사이트 통로, 다운로드된 파일들, 폼들에 대한 표와 그래프를 생성한다. 다른 보고서 섹션들도 확장된 기능들로 가득하다. 리퍼러/키워드 섹션에는 귀하의 사이트에 적중자료를 보낸 최상위 엔

진들과 각 사이트를 찾아낸 검색어들이 표시된다.

웹트렌즈 보고서에는 여과기능이 있어 특정 데이터를 제외시키거나 포함시킬 수 있다. 예를 들어 보고서에서 그 IP 주소를 배제함으로써 라이브러리 근무자들에 의해 생성된 요청들을 선택해서 제외시킬 수 있다. 다른 필터들은 자료의 한 페이지만을 나타낸다거나, 한 주의 특정한 날 혹은 하루 중 특정한 시간에 대한 자료를 나타내거나, 아니면 특정한 리퍼러 페이지를 나타낼 수 있다. 이 기능은 제시된 자료의 양을 조절하는데 편리하고, 보고서를 특정한 주제에 맞춰 보다 세밀하게 정리하는데 도움이 된다.

웹트렌즈는 귀하의 사이트를 방문하는 사람들의 수를 구분하기 위하여 수학적 연산법을 이용한다. 로그 파일에서 개별 사용자들을 명확히 구분해 내는데는 어려움이 있고, 이 정보는 확실하지 않을 수도 있다. 웹트렌즈 자체는 사이트의 개별 방문자를 명확히 구분해 내는 유일한 방법은 인증(즉, 로그온과 암호)을 이용하는 것이라고 밝히고 있다. 인증을 요구하지 않는 사이트의 경우 웹트렌즈는 데이터베이스의 사용자 프로파일 정보와 사이트에서의 방문객 활동을 연결하는 기능을 제공하고 있다.

웹트렌즈는 사용하기 쉬운 기능들을 많이 제공하고 있다. 어떤 면에 있어서 이것은 저가 혹은 무료 유틸리티와 7,000~10,000달러에 달하는 아주 값비싼 소프트웨어 패키지 간의 가교역할을 한다. 웹트렌즈의 고급 기능들 중에는 귀하의 라이브러리에서 요구하고 있는 것들을 상회하는 것들이 있을 수도 있다.

다음의 두 가지 소프트웨어 패키지들은 웹트렌즈의 기능들과 같은 점을 많이 가지고 있다. 그 기능에는 사전 정의된 보고서와 개별화 할 수 있는 보고서, 자료 여과, 그래픽, 사용자에게 친근한 인터페이스와 같은 것들이 있다.

명칭: 넷인텔렉트 4.0
URL: http://www.Webmanage.com/
제작자: 웹메니지
가격: 295달러
고객 지원: 온라인 지도와 문서, 전화, 전자우편, 온라인에 의한 지원.
시험기간: 15일간 무료.
로그 파일: 45개의 로그 파일 인식.
플랫폼: 윈도 95/98/NT

명칭: 패스트스태츠
URL: http://www.mach5.com/fast/
가격: 99.95달러
시험기간: 25일간 무료
고객 지원: 전자우편을 통한 무료 기술지원. 전화지원 없음.
플랫폼: 윈도 95/98/NT

결론

라이브러리 웹사이트들은 중요성이 더해질 것이다. 이들은 극히 중요한 서비스 포인트와 돈과 스탭 시간의 투자를 나타내고 있다. 어떤 라이브러리에서 그 웹사이트 사용량을 측정하고자 한다면, 서버 로그 분석은 채택되어야할 도구 중 하나이다. 사용량에 대한 보다 더 상세한 정보를 얻고자 한다면, 라이브러리들은 앙케이트나 쿠키와 같이 데이터를 수집할 수 있는 다른 방법을 이용하여 조사해야 한다.

서버 로그들은 컴퓨터 서버 상의 통신량과 요구 부하를 측정하도록 고안되었고, 이런 용도로 기능을 잘 발휘하고 있다. 서버 로그 파일들을 사람들이 어떤 사이트를 어떻게 이용하는 가를 측정하기 위해 사용하고자 할 때, 이들은 그리 훌륭하게 기능하지 못한다.

그러나 이들은 귀하의 웹사이트 상에 있는 페이지들의 비교 사용량, 방문객을 귀하의 사이트로 위탁한 다른 사이트들, 다른 중요한 데이터들 가운데서 검색엔진이 사람들로 하여금 어떻게 귀하의 사이트를 찾도록 도왔는가에 대한 유용한 정보를 줄 수 있다.

비록 로그 분석이 완전하지 않지만 어차피 사용량 계측은 거의 존재하지 않는다. 예를 들어 도서관 문을 통과한 사람들을 셀 때, 그들이 책이나 잡지를 읽으러 그곳에 들렀는지 아니면 그저 화장실에 가려고 들렀는지 알 수 없다. 책을 대출할 때, 왜 그 책을 선택했는지 그것이 읽히기는 하는지 알 수 없다.

서버 로그 파일 분석은 결점이 있지만 사용량 계측에 필요한 방법이라는 것을 그와 같은 견지에서 보면 이해 할 수 있을 것이다. 중요한 것은 로그 파일 분석의 능력과 한계에 대해 배우는 것이다. 그리하면 그것이 만들어 내는 데이터의 이용법을 알 수 있게 될 것이다.