응용통계연구 제 14권 2호, 2001년, pp. 261–271

웹 로그(Web Log) 데이터 분석 방법에 관한 연구

김석기[1] 안정용[2] 한경수[3]

요 약

정보 공유와 비즈니스 수행 등의 매체로서 World Wide Web의 이용이 보편화됨에 따라 다양하고 방대한 데이터를 웹을 통하여 얻을 수 있게 되었으며, 이러한 데이터로부터 유용한 정보를 추출하기 위한 데이터 분석과 활용은 많은 분야에서 중요한 사안으로 인식되고 있다. 본 연구에서는 웹 로그(web log) 데이터로부터 정보를 추출하기 위한 과정 및 방안에 대해 살펴보고자 한다. 로그 데이터의 특징과 통계 데이터와의 차이점, 데이터 수집 및 사전 처리 과정, 추출할 수 있는 정보 및 분석 방법 등을 제시하고 로그 데이터 분석 예제를 제시한다.

주요용어: 웹 로그, 로그 분석, 연관 규칙.

1. 서론

최근 들어 World Wide Web(이하 웹)은 정보의 공유는 물론 비즈니스(business)를 수행하기 위한 매체로서 확장되고 있으며, 사용자 역시 그 수적인 면에서 매우 빠른 속도로 증가하고 있다. 한국인터넷정보센터(KRNIC) 통계보고서에 따르면 1999년 2월 현재 340여만 명이었던 우리나라 인터넷 사용자 수는 2000년 6월 현재 1,580여만명 이상으로 증가하여 전국 인구 대비 37%가 인터넷을 이용하고 있는 것으로 집계되어 보고되고 있으며, 전 세계 인터넷 사용자는 2000년에 3억명, 2005년에는 7억명을 훨씬 초과할 것으로 CIA(Computer Industry Almanac)사는 보고하고 있다. 이처럼 많은 사용자들이 웹(또는 인터넷)을 이용함으로써 다양하고 방대한 양의 데이터가 웹을 통하여 생산되고 있으며, 이러한 데이터로부터 유용한 정보를 추출하기 위한 데이터 분석과 활용이 여러 분야에서 중요한 사안으로 인식되고 있다.

웹 상에서 사용자들의 활동에 관한 기초적인 정보는 웹 서버에 의해 자동적으로 수집되어 '웹 로그(web log)'라는 형태로 저장, 관리되어 진다. 웹 로그 데이터(이하 로그 데이터)는 웹 서버에 접속한 사용자를 인식할 수 있는 웹 페이지 방문 정보들을 포함하고 있으며, 일반 통계 데이터와는 다른 특징을 가지고 있다(로그 데이터의 특징에 관해서는 2장에서 간단히 소개된다).

1) (573-719) 군산시 개정동 413, 군산간호대학 간호정보공학센터, 연구원
E-mail: Sisyphus@kcn.ac.kr
2) (590-711) 남원시 광치동 720, 서남대학교 컴퓨터정보통신학부, 조교수
E-mail: jyahn@tiger.seonam.ac.kr
3) (561-756) 전주시 덕진동 664-14, 전북대학교 수학 통계정보과학부, 교수
E-mail: kshan@stat.chonbuk.ac.kr

기본적으로 사용자들의 웹 페이지 방문은 시간에 따라 순차적으로 이루어지며, 로그 데이터는 주로 사용자들의 웹 페이지 방문 패턴을 발견하기 위한 목적으로 이용된다. 따라서 웹 페이지 방문 순서와 방문 시간은 로그 데이터의 분석을 위한 중요한 요소이다. 웹 페이지 방문 순서 정보를 이용하는 대표적인 분석은 순차 패턴(sequential pattern) 발견과 연관 규칙(association rules) 탐색 등을 들 수 있으며 Agrawal과 Srikant(1995), Mobasher 등(1996), Pei 등(2000)에서 연구가 이루어졌다.

본 연구에서는 로그 데이터로부터 정보를 추출하기 위한 과정 및 방안에 대해 살펴보고자 한다. 로그 데이터는 일반적으로 분석에 불필요한 데이터를 많이 포함하고 있으며, 적절한 분석을 위해서는 기본적인 로그 데이터 이외에 추가적인 정보가 요구된다. 또한 로그 데이터의 분석을 위해서는 먼저 웹 서버에 대한 사용자들의 방문 정보를 개개인의 그룹으로 구성하는 세션 구분(session identification) 작업이 수행되어져야 한다. 기존의 연구들에서는 방문객의 웹 사이트 체류 시간을 기준으로 세션을 구분하고 있다. 이러한 세션 구분은 실제 방문객의 방문 특성에 근거한 구분이라 할 수 없다. 본 연구에서는 세션 구분을 위해 방문객이 웹 서버에 접속할 때 발생하는 로그를 사용한 세션 구분에 대하여 제안하고, 기존의 방법과 비교하고자 한다.

2장에서 로그 데이터의 특징과 일반적인 통계 데이터와의 차이점에 관해 살펴보고, 3장에서는 수집 방법 및 사전처리 과정을 다룬다. 4장에서는 로그 데이터에서 추출할 수 있는 정보 및 분석 방법 등을 제시하고 분석 예제를 제시한다.

2. 로그 데이터의 특징 및 통계 데이터와의 차이점

로그 데이터는 웹 서버에 대한 모든 방문객들의 접근을 기록한 데이터로 HTTP 프로토콜의 일부로 명시된 Common Log Format 또는 Extended Log Format을 따라 저장된다. 저장된 로그 데이터는 그림 2.1과 같이 웹 서버에 접속한 방문객의 IP(internet protocol) 주소, 접근 시각, 접근 방법, 대상 URL, 전송 프로토콜, 에러 코드, 전송 바이트 수와 같이 방문객를 인식할 수 있는 정보와 웹 페이지에 대한 방문 정보들을 포함하고 있으며, 다음과 같은 특징을 가지고 있다.

Common Log Format

```
165.194.11.110 - - [11/Jul/2000:21:06:32 +0900] "GET /image/nnimage/head1.jpg HTTP/1.1" 200 7592
165.194.11.110 - - [11/Jul/2000:21:06:32 +0900] "GET /defaultms.asp HTTP/1.1" 200 18514
165.194.11.110 - - [11/Jul/2000:21:06:32 +0900] "GET /image/nnimage/statinfo.jpg HTTP/1.1" 200 7376
```

Extended Log Format

```
01:30:23 210.117.171.66 GET /lyahn/sampling/ 302
01:30:23 210.117.171.66 GET /lyahn/sampling/Default.htm 200
01:31:35 210.182.144.220 GET /Emp/Guide/1042.html 200
01:31:44 210.182.144.220 GET /Emp/images/midback.gif 200
```

그림 2.1: 로그 데이터 Format

- 데이터를 수집하기 위한 특별한 절차가 필요 없다.

  로그 데이터는 일반적으로 방문객이 데이터의 발생을 인식하지 못하는 사이에 생성되어 수집되며, 데이터를 수집하기 위하여 방문객들에게 어떠한 요구도 할 필요가 없다.

- 분석에 불필요한 정보를 많이 포함하고 있다.

  로그 데이터에는 분석에 이용되지 않는 데이터가 많이 포함되어 있으며, 이러한 데이터는 일반적으로 분석에 이용되는 데이터보다 훨씬 많다. 따라서 로그 데이터를 분석할 때 이러한 정보를 제거하기 위한 정제(cleaning) 작업이 요구된다.

- 분석을 위해서는 추가적인 정보가 필요하다.

  로그 데이터에서 각각의 방문자들은 IP 주소(예를 들어, 210.182.144.220)를 통하여 구분된다. 그러나 IP 주소만으로는 방문자에 대한 정보를 얻기가 어려우며 지역별, 기관별 방문 현황 등을 분석하기 위해서는 IP 주소 할당 정보인 WHOIS 정보가 필요하다. WHOIS 정보는 특정 IP 주소를 어떤 기관 또는 단체가 사용하고 있는가를 보여주는 목록으로 이해하면 된다. 두 번째로 요구되는 추가적인 정보는 로그 데이터에 포함되어 있는 불필요한 정보를 정제하기 위해서 웹 관리자에 대한 정보, 웹 검색 엔진 정보 등이 필요하며, 웹 서버 방문객에 대한 방문 정보를 구분하는 세션 구분(session identification)에 대한 구체적인 방법 또한 필요하다. 이것은 방문자들의 방문 경로를 체류 시간을 기준으로 구분하는 작업으로 한번의 방문으로 인한 경로인지 여러 번의 방문으로 인해 생성된 경로인지를 구분하게 된다(3.2절 참조). 쉽게 설명하면, 어떤 방문객의 세션이 하루에 3회였다면 3번 방문했다는 의미이다.

- 일반적으로 대용량이다.

  로그 데이터는 웹 서버에 대한 모든 사용자들의 접근을 기록하기 때문에 -웹 사이트별로 차이는 있지만- 일반적으로 대용량인 특징이 있다. 따라서 대용량의 데이터를 처리하기 위한 기법이 필요하다.

로그 데이터의 위와 같은 특징들은 기존에 통계학 분야에서 이용되어 왔던 통계 데이터와의 차이를 분명하게 보여 준다. 일반적으로 과거의 통계학 분야에서 다루어지던 데이터는 행렬 형태로 표현되는 비교적 소규모의 데이터가 주를 이루어 왔으며, 검정(testing)에 필요한 속성(예를 들면 정규성, 등분산성 등)인 동질성(homogeneity)을 가지고 있는 특징이 있다.

그러나 로그 데이터와 같이 현실 세계에서 나타나는 데이터의 대부분은 실험 데이터와는 달리 이와 같은 속성을 반드시 갖지는 않으며, 분석에 바로 이용될 수 있는 형태도 아니다. 따라서 이러한 데이터를 분석하기 위해서는 사전 처리가 불가피하며 Friedman(1997), Hand(1997) 등에서 지적하듯이 데이터로부터 어떤 패턴(patterns)을 발견하기 위한 적절한 분석 방법의 개발이 필요하다.

3. 로그 데이터 수집 및 사전 처리

3.1. 로그 데이터 수집

로그 데이터를 수집하는 방법은 Srivastava 등(2000)이 설명한 바와 같이 데이터의 수집 위치에 따라 server 수준, client 수준, proxy 수준으로 구분할 수 있으며, Microsoft IIS(Internet Information Server)에서 server 수준의 로그 데이터를 수집하는 방법은 3가지로 구분할 수 있다. 첫 번째는 웹 서버에서 제공하는 방법을 이용하여 로그를 수집하는 방법이며, 두 번째는 사용자가 웹 서버에 접속되는 시점에서 세션이 생성된다는 점을 이용하여 로그를 수집하는 방법이다. 마지막 방법은 분석자가 관심을 가지는 특정 페이지에 로그를 기록할 수 있는 스크립트 코드를 삽입하는 것으로 이 방법은 정제된 로그 데이터를 수집할 수 있어 분석이 용이하다는 장점이 있는 반면 모든 웹 페이지에 프로그램을 해 주어야 하는 단점이 있다.

본 연구에서 제시하는 예제에서는 위의 방법들 중 첫 번째와 두 번째 방법을 병행하여 로그 데이터를 수집하고, 수집된 데이터는 분석 절차를 용이하게 하기 위해 데이터베이스에 저장하여 이용한다.

3.2. 사전 처리

로그 데이터를 분석하기 위해서는 여러 처리 과정이 필요하다. 첫째, 수집된 데이터는 정제 과정을 거쳐야 된다. 로그 데이터는 일반적인 방문객들의 정보만을 저장하는 것이 아니며 웹 관리자나 개발자의 작업 내용 및 검색 엔진인 로봇(robot)에 의한 조회 기록까지 저장된다. 또 웹 페이지에 대한 정보 이외에도 웹 페이지에 삽입된 이미지나 오디오와 같은 미디어에 대한 참조 정보 역시 포함 되어있다. 따라서 정제 과정이 필수적으로 요구되며, 정제 과정을 통해 생성된 데이터는 분석의 대상이 되는 문서 자료들로만 구성 되어있다. 또한 웹 서버 관리자와 개발자의 로그 정보는 물론 웹 검색 엔진에 의해 탐색된 정보 역시 제거된 자료이다(로그 기록자가 web robot인지의 판단은 기록된 로그 정보 중 방문객의 웹 브라우저 정보를 나타내는 `HTTP_USER_AGENT` 속성을 이용하여 결정한다).

둘째, 사전 처리를 수행하기 위해서는 웹 서버에서 기본적으로 수집되는 로그 이외에 WHOIS 정보와 웹 구조에 대한 정보가 필요하다. WHOIS 정보는 각국의 인터넷 정보 센터에서 제공되는 IP 주소 할당 정보로 국내에 배정된 IP 주소에 대한 등록 정보는 한국인터넷정보센터(http://ipwhois.nic.or.kr)에서 조회할 수 있다.

셋째, 사전 처리 절차에서는 웹 로그에 대한 분석을 용이하게 수행하기 위해 정제된 로그를 구조화하는 작업이 수행된다. 이 단계에서 이루어지는 주요 작업 중 하나가 세션 구분으로 순차 패턴 발견, 연관 규칙 탐사, 군집 분석 등을 위한 자료로 사용된다.

세션을 구분하는 일반적인 방법은 방문자 로그의 최대 유휴 시간(maximum idle time)을 이용하는 것이다(김광용 2000; Lee 등 1998; Mobasher 등 1996; Fu 등 1999). 이 방법은 초기 유휴 시간을 설정하고 학습 자료(training data)의 분석을 통한 최적 시간을 산출하여 방문자 로그를 세션(session)별로 구분한다. 그러나 이러한 방법은 방문자 로그를 세션별로 구분하는데 있어서 실제 방문 형태에 기반한 것이 아니라 추론에 의해 구분한다는 문제를 가

지고 있다. 즉, 실제로 방문자가 웹 방문 중 특정 페이지를 오랜 시간 참조할 경우 새로운 세션이 아님에도 불구하고 새로운 세션으로 구분하게 되며, 역으로 방문 시간 간격이 작을 경우 새로운 세션을 구분하지 못하는 단점을 가지고 있다(그림 3.1 참조).

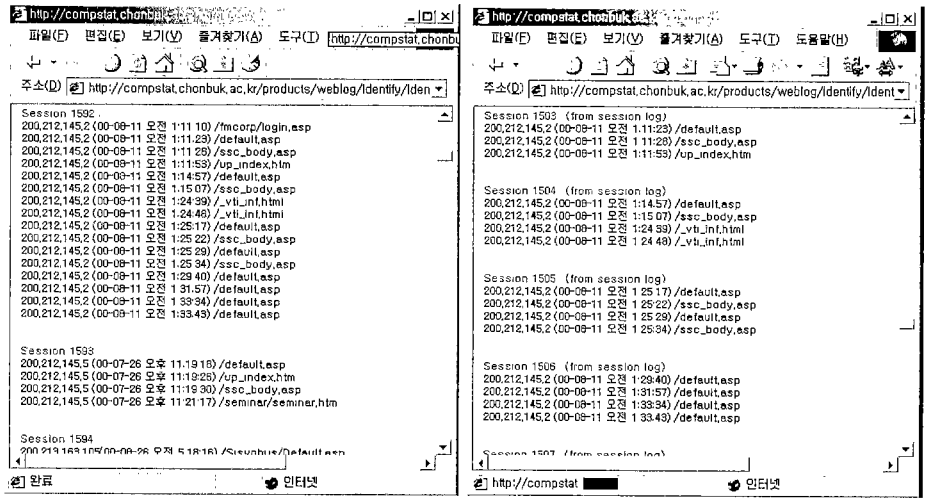

그림 3.1: 세션 구분 : 최대 유휴 시간 이용(왼쪽), 세션 로그 정보 이용(오른쪽)

본 연구에서는 이러한 문제를 해결하기 위해 방문객이 웹 서버에 접속할 때 발생되는 세션 정보를 이용하여 방문자 세션을 구분하고, 세션 정보가 존재하지 않는 로그에 대해서는 최대 유휴 시간을 적용하는 방법을 사용하였다.

방문자의 세션 구분을 위해 최대 유휴 시간을 사용한 경우 그림 3.1의 왼쪽과 같이 하나의 세션으로 구분하는 반면에, 세션 로그 정보를 같이 사용할 경우 그림 3.1의 오른쪽과 같이 4개의 세션으로 구분함을 알 수 있다. 이러한 결과는 실제 방문객이 웹사이트를 탐색하는 과정에서 다른 사이트로 이동했다가 되돌아 온 경우 또는 웹 브라우저를 종료했다가 다시 방문한 경우 등에 해당한다. 따라서 세션 로그 정보에 근거한 세션 구분이 최대 유휴 시간을 이용한 세션 구분에 비하여 방문객의 특성을 반영하는 구분 방법이라 할 수 있다.

4. 로그 데이터 분석 방법과 분석 예제

4.1. 로그 데이터 분석 방법

로그 데이터는 2장에서 서술한 바와 같이 일반적으로 대용량인 특징이 있다. 따라서 로그 데이터를 분석할 때 이러한 특징을 고려해야 한다. 대용량 데이터를 분석할 때 이용할 수 있는 방법은 여러 가지가 있을 수 있겠으나 몇 가지만 제시해 보면,

- 통계적 관점 : 표본을 추출하여 이용하는 방법
- 컴퓨팅 관점 : 병렬 또는 분산 컴퓨팅(parallel/distributed computing) 방법

- 데이터베이스 활용 관점 : 요약 정보를 갖는 테이블(summary table)을 이용하는 방법 등을 들 수 있다.

로그 데이터는 일반적으로 방문객들의 웹 페이지 방문 패턴을 발견하기 위해서 분석되어지며, 분석 시 추출하고자 하는 정보가 무엇인가에 따라 적용할 수 있는 분석 기법들은 여러 가지가 있다. 로그 데이터에서 추출할 수 있는 정보와 이를 위해 이용되는 분석 기법을 정리해 보면 다음과 같다.

- 트래픽(traffic) 정보

  웹사이트를 방문할 때 트래픽의 잦은 발생은 방문자 수를 감소시키는 직접적인 원인이 된다. 따라서 트래픽 발생의 원인을 탐색하고 해결함으로서 방문 고객과 사이트를 효율적으로 관리 할 수 있다. 트래픽 정보는 트래픽의 시점과 종류 등의 빈도를 계산하여 얻어진다.

- 연관 규칙(association rules) 탐색

  연관 규칙의 탐색은 웹 페이지간의 상관 정보를 얻기 위한 방법으로 웹사이트를 효율적 구성하는데 도움을 준다. 연관 규칙 탐색의 알고리즘에 관한 세부적인 내용은 Agrawal과 Srikant(1994), Mobasher 등(1996)을 참조하기 바란다.

- 순차 패턴(sequence patterns) 발견

  순차 패턴 발견은 웹 페이지를 방문하는 순서 정보를 이용하여 몇 개의 대표적인 패턴을 찾기 위한 방법으로 웹사이트의 효율적 구성에 도움을 주며, 그 결과는 다른 분석(예를 들어, 분류(classification))에도 이용될 수 있다. 순차 패턴에 관한 알고리즘 및 세부적인 내용은 Agrawal과 Srikant(1995), Pei 등(2000)을 참조하기 바란다.

- 군집 분석(cluster analysis)

  웹 페이지 또는 방문객들을 몇 개의 군으로 나누어 특성을 파악하기 위한 방법으로 방문객 관리를 위해 가장 많이 활용되는 기법이다. 군집 분석 알고리즘은 Hartigan과 Wong(1979)의 k-평균 알고리즘과 신경망(Neural Network) 이론의 하나인 Kohonen(1984)의 코호넨 네트웍(Kohonen Network)이 주로 이용되고 있으며 Guha 등(1998), Zhang 등(1996)에서 대용량 데이터를 위한 알고리즘이 개발되었다.

이러한 분석을 통하여 방문자들에 대한 접속 유휴 시간 측정, 트래픽(traffic)에 대한 정보, 웹사이트 재구축 전략 수립 등에 대한 정보를 획득할 수 있고, 특히 웹 마케팅(marketing)을 수행하는 사이트에서는 교차 판매(cross selling) 계획 수립, 캠페인 효과에 대한 평가, 개인화(personalization)된 서비스(예를 들어, 개인별 Web 페이지 서비스와 상거래에 있어서 개인별 제안 등)를 제공할 수 있는 기본 자료로 활용할 수 있기 때문에 로그 데이터의 적절한 분석과 활용이 필수 요소로 부각되고 있다.

4.2. 로그 데이터 분석 예제

이 절에서는 로그 데이터를 분석한 간단한 예제를 제시한다. 분석에 이용된 로그 데이터는 전북대학교 통계학과에서 운영하고 있는 COMPSTAT 서버의 데이터이며, 대용량의 데이터로 취급하기엔 약간 무리가 있다(데이터베이스의 데이터 파일의 크기 : 대략 630 megabyte). 그러나 일반적인 대용량 데이터의 특징을 감안하여 데이터베이스에 요약 정보를 생성하여 분석에 이용하고 있다.

예제는 로그 데이터로부터 추출할 수 있는 기초 통계량과 연관 규칙을 포함하고 있다. 연관 규칙의 탐색은 기존의 방문자 로그의 최대 유휴 시간(maximum idle time)을 이용하는 세션 구분 방법과 본 연구에서 제안한 방문객의 웹 서버 접속 시에 발생하는 로그를 사용한 세션 구분 방법을 비교 제시한다.

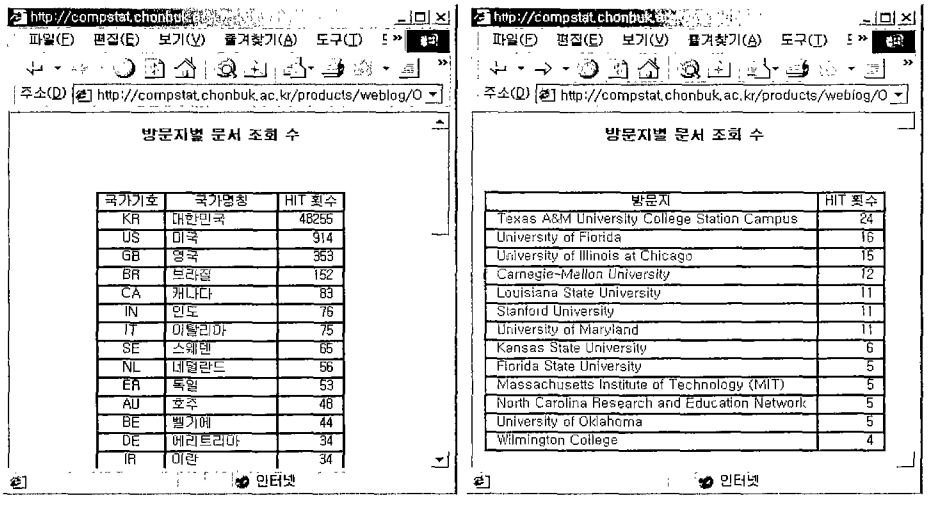

그림 4.1: 웹사이트를 방문한 국가(왼쪽), 기관(오른쪽)별 현황

그림 4.1은 WHOIS 정보를 이용하여 웹사이트를 방문한 방문 IP 주소를 방문지(국가, 기관)별로 구분한 것이며, 그림 4.2는 웹 구조에 따른 문서 조회 수에 대한 결과와 조회 빈도가 높은 문서들에 대한 분석 자료이다.

웹 페이지 방문 순서, 세션 구분 등의 정보를 이용하는 연관 규칙(association rule) 탐색에서는 그림 4.3에서 볼 수 있듯이 전체 세션 중 특정 항목에 대한 참조 비율인 support(A), support(A,B)와 특정 항목(A) 참조 후 다른 항목(B)을 참조한 비율인 confidence(B|A)에 대한 분석 결과를 제시한다. 그림 4.3의 왼쪽은 최대 유휴 시간을 30분으로 세션을 구분한 후 연관 규칙을 계산한 결과이고, 그림 4.3의 오른쪽은 세션 구분 과정에서 세션 로그 정보를 활용하여 세션을 구분한 후 연관 규칙을 계산한 결과이다. 두 경우가 서로 상이한 결과를 나타내고 있다. 이 결과는 본 연구에서 제안한 세션 구분 방법이 최대 유휴 시간으로 세션을 구분하는 방법에 비해 세션을 잘 구분하고 있음을 보여주고 있다.

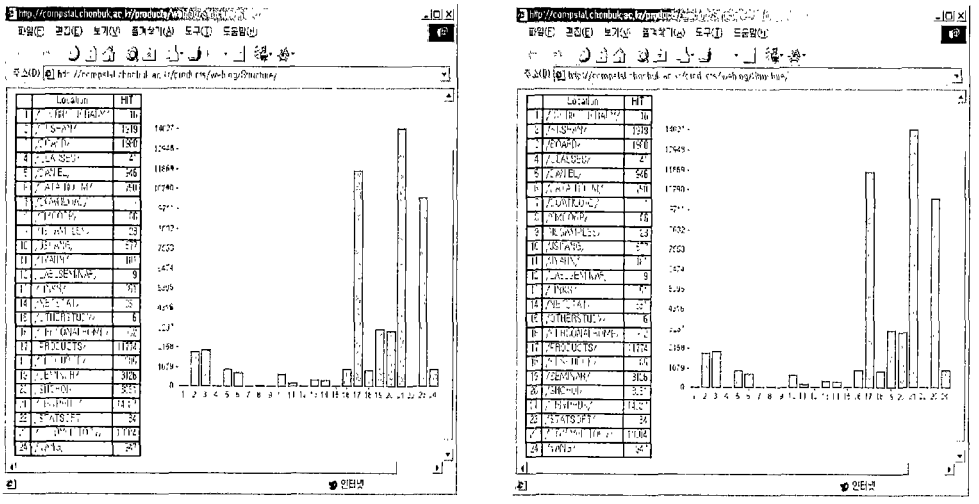

그림 4.2: 웹 구조에 따른 문서 조회 수(왼쪽)와 Hit Ranking(오른쪽)

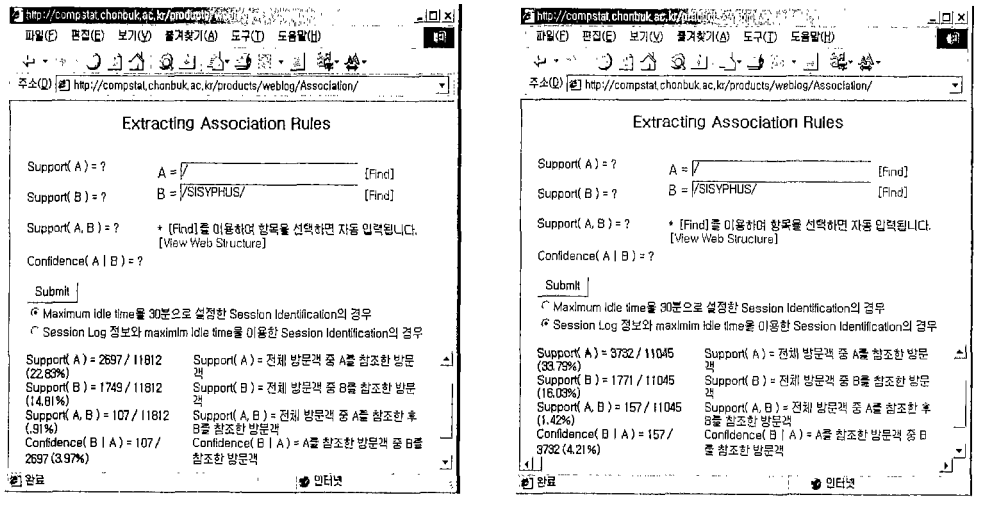

그림 4.3: 연관 규칙 : 최대 유휴 시간 구분(왼쪽), 세션 로그 정보로 구분(오른쪽)

5. 결론

본 연구에서는 로그 데이터의 특징 및 일반적인 통계 데이터와의 차이점, 정보를 추출하기 위한 과정 및 방법에 대해 살펴보고 연간 규칙 탐색, 순차 패턴 발견 등에 활용하는 방문자의 세션을 구분하기 위한 방법을 제안하였다.

본 연구에서 제안한 방법을 이용하여 세션을 구분하면 특정 페이지를 오랜 시간 참조할 경우 새로운 세션으로 구분하게 되는 기존의 방법보다 더 정확한 결과를 얻을 수 있다.

로그 데이터는 웹 사용자 개개인에 최적화 된 서비스와 웹에 게시될 정보들의 효율적인 배치 및 마케팅에 직접 활용할 수 있는 정보를 제공한다. 따라서 최근에 부각되고 있는 eCRM(electronic Customer Relationship Management)을 위한 분석에 다른 데이터들과 통합되어 활용되고 있으며 인터넷을 기반으로 한 마케팅 및 정보 서비스에 있어서 중요한 위치를 담당하게 될 것으로 예상된다.

참고문헌

[1] 김광용 (2000). Web Information Center와 Internet Survey, *Internet Survey Workshop 논문집*, 111-122.

[2] Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules, *Proceedings of the International Conference on Very Large Data Bases*, 487-499.

[3] Agrawal, R. and Srikant, R. (1995). Mining Sequential Patterns, *Proceedings of the International Conference on Data Engineering*, http://www.almaden.ibm.com/cs/people/srikant

[4] Friedman, J.H. (1997). Data Mining and Statistics : What's the Connection?, *Proceedings of the International Conference on the Interface : Computing Science and Statistics*, http://www.stat.rice.edu/interface97.html

[5] Fu, Y., Sandhu, K. and Shih, M.Y. (1999). Clustering of Web Users Based on Access Patterns, *Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profilling*, http://www.acm.org/sigkdd/proceedings/webkdd99

[6] Guha, S., Rastogi, R. and Shim, K.S. (1998). CURE: An Efficient Clustering Algorithm for Large Databases, *Proceedings of ACM SIGMOD International Conference on Management of Data*, 73-84, http://cs.kaist.ac.kr/ shim/

[7] Hand, D.J. (1997). Intelligent Data Analysis : Issues and Opportunities, *Intelligent Data Analysis*, Vol. **2**, No. 2, 1-14.

[8] Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS 136: A K-means clustering algorithm, *Applied Statistics*, Vol. **28**, 100-108.

[9] Kohonen, T. (1984). *Self-Organization and Associative Memory*, Springer-Verlag.

[10] Lee, D.H., Seo, D.Y., Kim, N.H. and Lee. J.Y. (1998). Discovery and Application of User Access Patterns in The World Wide Web, *Proceedings of the 4th World Congress on Expert Systems*, 321-327.

[11] Mobasher, B., Jain, N., Han, E.H. and Srivastava, J. (1996). Web Mining: Pattern Discovery from World Wide Web Transactions, Technical Report 96-050, Department of Computer Science, University of Minnesota, Minneapolis.

[12] Pei, J., Han, J., Mortazavi-asl, B. and Zhu, H. (2000). Mining Access Patterns Efficiently from Web Logs, *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, http://db.cs.sfu.ca/sections/publication/kdd

[13] Srivastava, J., Cooley, R. W., Deshpande, M. and Tan, P.N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations*, Vol. **1**, Issue 2, http://www.acm.org/sigs/sigkdd/explorations/

[14] Zhang, T., Ramarkrishnan, R. and Livny, M. (1996), BIRCH: An Efficient Data Clustering Method for Very Large Databases, *Proceedings of ACM SIGMOD International Conference on Data Management*, http://citeseer.nj.nec.com/zhang97birch.html

[2001년 1월 접수, 2001년 6월 채택]

Web Log Data Analysis

Seok Ki Kim[1] Jeong Yong Ahn[2] Kyung Soo Han[3]

ABSTRACT

Recently, World Wide Web is generally used for the knowledge sharing and business processing. As a byproduct of human activities in the space such as web, the data with very large scale have been collected, and analysis of the data is rasing the main issues. In this paper, we discuss the methods and processing for analysis of web log data, and present an example of the web log data analysis.

Keywords: Web Log; Log Analysis; Association Rule.

1) Researcher, Nursing Information Engineering Center, Kunsan College of Nursing. E-mail: Sisyphus@kcn.ac.kr
2) Assistant Professor, Division of Computer Science and Information Communications, Seonam University. E-mail: jyahn@tiger.seonam.ac.kr
3) Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University. E-mail: kshan@stat.chonbuk.ac.kr