The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 831-840

A Study of Web Usage Mining for eCRM

Hyuncheol Kang[1], Byoung Cheol Jung[2]

Abstract

In this study, We introduce the process of web usage mining, which has lately attracted considerable attention with the fast diffusion of world wide web, and explain the web log data, which is the main subject of web usage mining. Also, we illustrate some real examples of analysis for web log data and look into practical application of web usage mining for eCRM.

Keywords : Web Usage Mining, eCRM, Web Log, Click Stream

1. 서론

최근 인터넷이 빠르게 확산되면서 수많은 웹사이트와 온라인(on-line) 기업들이 생겨나고 있다. 특히 인터넷은 기업이 고객과 접촉할 수 있는 새로운 수단으로써 기업의 홍보나 서비스를 제공하는 기능을 수행할 뿐만 아니라 사업을 위한 중요한 도구가 되고 있다. 따라서 인터넷을 통한 고객과의 커뮤니케이션 및 관계유지는 웹사이트를 운영하고 있는 많은 기업들의 주요 관심사항이 되고 있으며, 기존의 CRM(Customer Relationship Management)에 대응되는 eCRM이 새롭게 주목받고 있다.

웹사이트를 관리하고 있는 각 기업의 웹서버에는 웹사이트 방문자의 모든 행위들이 웹로그라는 기록으로 남게 되는데, 아직까지도 대부분의 기업들은 이를 제대로 활용하지 못하고 있는 실정이다. 그러나 최근 웹로그 데이터의 저장, 처리, 분석 등에 대한 기술적·학문적 발전이 이루어짐에 따라 이러한 방문자의 정보를 이용하여 보다 효율적인 서비스 제공 및 사업기회의 획득을 수행하기 위해 노력하려는 움직임이 여러 기업에서 시도되고 있다. 즉, 웹사이트 운영자는 방문자가 어느 곳으로부터 자신의 사이트에 접속하는지 또는 어떤 웹페이지들을 주로 보는지와 같은 사용자에 대한 정보를 더 많이 얻고자 하며, 이를 통해 방문자의 특성에 맞는 서비스를 제공함으로써 신뢰를 쌓아가고자 한다.

웹 마이닝(web mining)은 웹컨텐츠 및 로그 데이터를 가공하고 분석하는 포괄적인 의미를 가지고 있는데, 이는 웹컨텐츠(web contents) 마이닝과 웹유시지(web usage) 마이닝으로 크게 구분된

1) Senior Lecturer, Department of Mathematics, Hoseo University, Asan 336-795, Korea.
E-mail : hychkang@office.hoseo.ac.kr
2) Post Doctorial Researcher, Department of Economics, Korea University, Seoul, 136-701, Korea.
E-mail : bcjung@kustat.korea.ac.kr

다(Srivastava et. al., 2000; Cooley et. al., 1999). 웹컨텐츠 마이닝은 텍스트(text) 마이닝의 한 분야로 웹사용자가 원하는 정보를 빠르고 정확하게 찾을 수 있도록 도와주는 기법을 의미하며, 주로 야후와 같은 검색엔진에서 활용되고 있다. 웹유시지 마이닝은 데이터마이닝의 한 분야로 웹서버 로그로부터 웹 사용자의 의미 있는 접속패턴을 발견하고 이를 통해 웹사이트의 개선 및 고객에 대한 차별적 서비스 제공 등을 수행하고자 하는 것으로 eCRM의 주된 도구로 사용되고 있다.

본 연구에서는 웹유시지 마이닝의 주요 개념과 제 과정을 소개하고 웹유시지 마이닝의 주요 대상이 되는 웹로그 데이터의 형태와 분석과정을 자세히 설명하고자 한다. 또한 웹유시지 마이닝의 실제 활용방안 및 몇 가지 사례를 제시할 것이다.

2. 웹로그 데이터

인터넷 사용자가 특정 웹사이트를 방문하여 웹페이지를 클릭하거나 특별한 요청에 대해 웹서버가 응답할 때마다 그 사이트를 관리하고 있는 서버(server)에는 로그(log)라고 불리는 레코드(record)들이 저장된다. 이와 같이 쌓여지는 일련의 레코드들의 집합을 웹로그 또는 클릭스트림(click stream) 데이터라고 한다. 따라서 웹로그를 통해 웹사이트 관리자는 누가 언제 무엇을 요청했는지를 알 수 있고, 얼마나 많은 사용자가 왔는지 그리고 어디에서 왔는지, 가장 오래 보는 페이지와 가장 많이 보는 페이지가 무엇인지 등을 알 수 있다. 이와 같은 로그 데이터가 저장된 로그파일은 웹서버가 지정하는 곳에 위치하며, 보통 웹서버 관리자가 웹서버를 설치할 때 로그파일의 위치와 기록방법 등을 지정하게 된다. 현재 널리 사용되고 있는 웹서버 소프트웨어로는 NCSA(www.ncsa.uiuc.edu), W3C(CERN, www.w3.org), MS IIS(www.microsoft.com), Netscape (www.netscape.com), Apache(www.apache.org), WebSite(website.oreilly.com) 등이 있다(WebLog 2000; Sane Solutions, 2000).

이들 웹서버마다 자체적으로 독특한 로그파일의 저장형식을 제안하고 있지만, 대부분 CLF (Common Log Format)라고 불리는 표준 로그파일 형식 및 ECLF(Extended Common Log Format Standard)를 기본적으로 지원하고 있으며, 웹유시지 마이닝은 주로 이 CLF 및 ECLF에 저장되는 정보를 이용하여 이루어진다. CLF에는 다음과 같은 7개의 필드가 저장되며, ECLF에는 여기에 Referrer와 User-Agent 두 개의 필드가 추가된다.

(1) Host : 사용자의 인터넷 주소이며, 도메인(domain) 이름 또는 IP(Internet Protocol) 주소로 기록된다. 웹서버는 초기에 이 호스트를 '10.119.195.208'과 같은 수치 IP 주소로 받아들이며, 대부분의 웹서버는 이 주소를 'www.webmania.co.kr'과 같은 텍스트 도메인 이름으로 확인할 수 있는 기능을 가지고 있다. 그러나 이와 같은 역방향 주소 조회는 웹서버가 매 접속 때마다 도메인 이름을 역추적해야 하므로 웹서버에 상당한 부하를 주게 된다.

(2) Ident(RFC931) : 이 필드는 Identd(Identification daemon)라는 프로토콜을 지원하는 클라이언트 애플리케이션이 제공하는 중재 ID이다. 현재 이 인증 스키마를 사용하는 웹브라우저는 거의 없기 때문에 대부분의 웹서버에는 보통 '-'로 기록된다.

(3) AuthUser : 웹서버에 등록된 사용자 이름이다. 만약 현재의 사용자가 등록된 사용자가 아닌 경우 '-'로 기록된다.

<표 2.1> 로그 레코드의 예

```
210.119.195.208   --- dragon [25/Jun/2000:04:02:12 +0900]
     (1)        (2)   (3)              (4)
"GET /download/index.html HTTP/1.1" 200 16621
            (5)                     (6)  (7)
"http://www.webmania.co.kr/download/ -> index.html"
                        (8)
"Mozilla/4.0 (compatible; MSIE 5.5; Windows 98)"
                        (9)
```

(4) Time : 접속일자와 시간을 기록한 필드로 [dd/mon/yyyy:hh:mm:ss x\#\#:\#\#]와 같은 형식으로 저장된다. 마지막의 'x\#\#:\#\#'에서 x는 '+' 또는 '-' 부호이며, '\#\#:\#\#'는 그리니치 표준시로부터의 시차를 나타낸다.

(5) Request : 이 필드는 'GET 및 POST 등의 명령어', '실제 요청대상의 파일 이름', '전송 프로토콜 및 버전' 등 세 개의 세부 필드를 기록한다. 가장 일반적인 HTTP(Hyper-Text Transfer Protocol, 인터넷에서 하이퍼텍스트 문서를 교환하기 위해 사용하는 통신규약) 명령어로는 GET(웹서버에게 객체를 요청)과 POST(웹브라우저에서 얻은 정보를 웹서버로 보냄)가 있다.

(6) Status : 접속상태와 데이터의 이동 현황을 기록하는 것으로 100, 200, 300, 400, 500과 같은 5개의 카테고리로 구분된다. 예를 들어, 200은 '성공', 300은 '무시', 400은 '에러'를 나타낸다.

(7) Bytes : 사용자가 실제로 웹서버에서 가져간 데이터의 양을 기록한 것으로 단위는 바이트이다. 이는 앞의 Status 코드가 200 카테고리인 경우에만 기록되며, 그 외의 경우(즉, 실제 전송된 데이터의 양이 0인 경우)에는 '-'로 기록된다.

(8) Referrer : 이 필드는 사용자의 요청이나 링크의 원래 소스(source)를 나타내기 위해 전송하는 텍스트 문자열을 기록한다. 즉, 웹서버를 소개해 준 사이트와 소개받은 페이지를 화살표로 기록한다. 따라서 이 필드를 이용하면 사용자가 어디에서 그 웹사이트로 연결되었는지를 알 수 있으며, 이는 온라인 광고 또는 홍보 등을 고려할 때 중요한 평가자료로 사용될 수 있다.

(9) User-Agent : 사용자(클라이언트)의 요청을 만든 소프트웨어 및 운영체제(OS)의 이름과 버전 등이 기록되는 필드로써 이를 브라우저 로그파일이라고도 한다.

3. 웹로그 데이터의 사전처리

웹로그 데이터는 일정 기간동안 사용자들이 접속 또는 요청한 내용을 기록한 일종의 거래(transaction) 데이터이므로 이를 실제로 분석하기 위해서는 다음과 같은 여러 단계의 사전처리(pre-processing) 과정이 필요하다.

3.1 필터링 및 필드 선택

웹서버는 수많은 사용자들이 접속하는 상황을 개별 레코드로 기록하기 때문에 매우 많은 양의 로그 데이터가 생기게 된다. 실제로 옥션(www.auction.co.kr), 다음(www.daum.net), 야후(www.

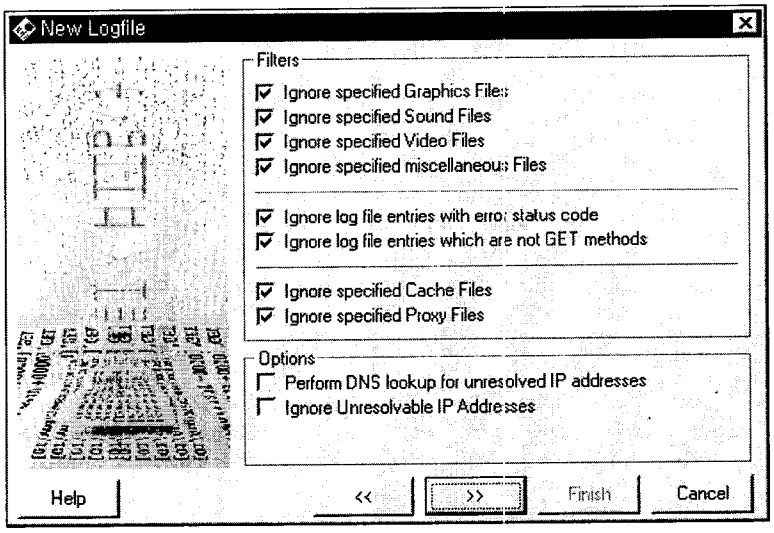
yahoo.co.kr) 등 국내의 대형 웹사이트들의 경우 하루에 수십 기가 바이트(Giga byte) 이상의 로그 데이터가 쌓이고 있는 실정이다. 따라서 분석에 불필요한 레코드들을 제거하거나 필요한 필드만 선택하는 것은 원활한 분석을 위해 매우 중요하다. 이러한 필터링 과정에서는 Request 필드의 파일 이름의 확장자가 gif나 jpg인 이미지 파일, cgi인 스크립트 파일, avi나 mov인 오디오 및 비디오 파일 등인 레코드들을 제거하는 작업이 포함되기도 하는데 이러한 파일들은 대부분 특정 웹페이지의 일부를 구성하고 있는 구성요소에 불과하므로 사용자의 행동패턴과는 무관한 로그 레코드들일 가능성이 많기 때문이다. <그림 3.1>은 웹마이닝 소프트웨어 중 하나인 EasyMiner (www.mineit.com)의 필터링 옵션 대화상자를 보여주고 있다.

3.2 사용자 식별

사용자 식별이란 로그 레코드들의 사용자를 유일하게 구별해 주는 과정을 말한다. 가장 확실한 사용자 식별방법 중 하나는 사용자가 웹사이트를 방문할 때 로그인(log-in) 과정을 거치도록 하는 것이다. 이렇게 하면 로그 데이터의 AuthUser 필드에 사용자의 유일한 ID가 기록되기 때문에 특별한 사용자 식별과정을 거치지 않고서도 사용자를 식별할 수가 있다(이런 이유로 실제 많은 웹사이트들이 로그인 시스템으로 변경하는 추세이다). 그러나 이 경우 사용자가 웹사이트 방문에 거부감을 가지게 할 수 있다는 위험이 따르고, 한 사용자가 복수 ID를 사용하거나 여러 사용자가 단일 ID를 사용할 수 있다는 문제점을 가지고 있다.

만약 로그인을 통한 사용자 인증과정이 없는 웹서버의 경우에는 로그파일에 저장되어 있는 Host와 User-Agent 필드를 이용하여 사용자를 구별하는 방법을 사용할 수 있다. 그러나 이 방법은 같은 프록시(proxy) 서버에 의해 웹사이트에 접속하는 사용자는 모두 동일한 IP 주소를 갖기 때문에 정확한 사용자 식별을 보장받지는 못하게 된다. 또한 인터넷 전용선 제공회사를 이용하는 많은 사용자들은 유동(floating) IP를 사용하기 때문에 이에 따른 문제도 발생하게 된다.

<그림 3.1> EasyMiner의 필터링 옵션 대화상자



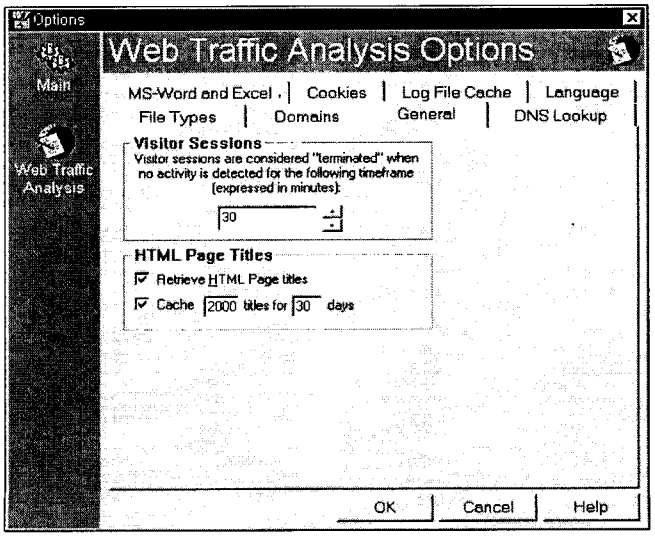
다른 대안으로는 쿠키(cookie, 웹사이트와 사용자의 컴퓨터 사이에서 통신을 매개해주는 기능을 하며, 사용자의 컴퓨터에 텍스트 문자열로 저장된다)를 이용할 수 있으나, 쿠키를 이용하는 방식도 사용자에 의해 쿠키가 지워지거나 조작될 수 있다는 문제점을 가지고 있다. 또한 쿠키는 사용자의 컴퓨터에 저장되며 저장된 쿠키를 웹사이트가 읽어들이는 방식을 취하게 되므로, 쿠키가 고의로 사용자의 정보를 빼낼 수 있는 통로 역할을 할 수도 있기 때문에 사용자의 정보보안에 대한 문제점이 지적되고 있다. 그러나 쿠키는 현재까지 웹서버가 사이트를 다시 방문하는 사용자를 식별하고 프로파일러가 한 웹사이트에서 다른 웹사이트로 사용자를 추적하는 중요 수단이 되고 있다(Kimball & Merz, 2000).

<표 3.1> 세션 ID 생성 예

User ID	Session ID	Time	Request
apple	1	20NOV2000:13:53:31	/main.jsp
apple	2	20NOV2000:15:31:11	/main.jsp
apple	2	20NOV2000:15:31:30	/mypage/magic/magic_skin1.jsp
apple	2	20NOV2000:15:31:30	/mypage/magic/magic_skin2.jsp
dragon	1	20NOV2000:13:53:13	/magazine/sr/sr_02_01.html
dragon	1	20NOV2000:13:56:06	/mypage/magic/magic_skin1.jsp
dragon	1	20NOV2000:14:03:09	/mypage/magic/magic_skin1.jsp
dragon	1	20NOV2000:14:04:15	/skincare/basicinfo/sys19.html
dragon	1	20NOV2000:14:13:20	/skincare/secret/now_01.html
dragon	1	20NOV2000:14:16:21	/skincare/secret/now_02.html
dragon	2	20NOV2000:17:44:27	/makeup/hottip/hottip_01.html
dragon	2	20NOV2000:17:47:53	/makeup/letsmakeup/sys2.html
...	...	...	...

웹로그 데이터 제공 : www.beauty-i.co.kr

<그림 3.2> WebTrends의 General 옵션 대화상자



3.3 세션 식별

세션(session)이란 사용자가 한 웹사이트를 방문하여 일련의 연속적인 행동을 수행한 후 접속을 중단할 때까지의 과정을 의미한다. 사실 로그 데이터 자체는 여러 사용자의 접속상황이 단지 시간 순서에 의해서 기록된 것이기 때문에 사용자가 언제 새로운 접속을 시도하여 언제 그 접속을 종료하였는지에 대한 정보가 존재하지 않는다. 웹로그 데이터를 분석하는 경우 대부분 세션이 분석단위가 되므로 <표 3.1>에서와 같이 동일한 사용자 내에서 세션 ID라고 불리는 일련의 일련번호(즉, 방문번호)를 추가해 주어야 한다.

세션을 식별하기 위해 몇 가지 방법이 제안되어 있지만, 현재 일반적으로 사용되는 방법은 Time 필드의 시간간격을 이용하는 것이다. 즉, <표 3.1>과 같이 먼저 ID와 Time 필드를 키(key)로 하여 로그 데이터를 정렬(sort)하고, 동일 ID 내에서 일정 시간 이상의 시간간격이 발생하면 새로운 세션 ID를 부여하는 것이다. 이 때 대부분의 상용 소프트웨어에서는 시간간격의 디폴트 설정값으로 30분을 사용하고 있는데, <그림 3.2>는 웹마이닝 소프트웨어 중 하나인 WebTrends(www.webtrends.com)의 옵션 대화상자 중 세션 구분을 위한 시간간격을 설정하는 화면을 보여주고 있다.

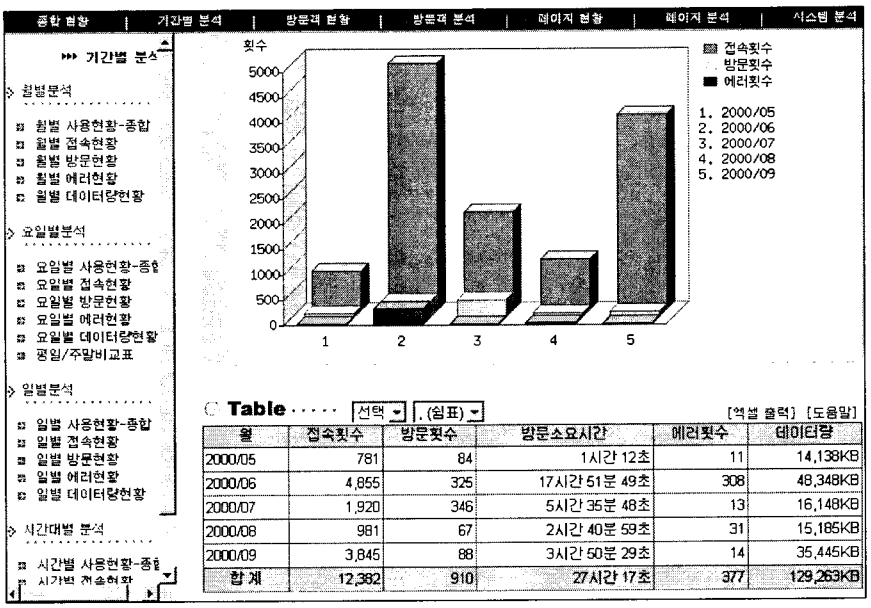
3.4 분석용 데이터 생성

세션 ID를 생성한 후에는 분석에 불필요한 레코드를 제거하는 등 다시 한번 로그 데이터를 정리하고 이후 분석과정에 적합하도록 데이터를 가공해 주어야 한다. 이때 필요한 작업 중 하나는 모든 웹페이지들을 부수적 페이지와 컨텐츠 페이지로 구분하는 것이다. 여기서 부수적 페이지란 사용자가 그 페이지를 방문했는지의 여부가 방문패턴을 분석함에 있어 특별한 정보가 되지 않는 페이지를 의미한다. 예를 들어, 어떤 웹사이트 홈페이지나 그 내부의 링크 페이지 등은 다른 페이지를 이동하기 위한 경로에 해당하므로 방문패턴과는 큰 관련이 없을 가능성이 많다. 이에 반해 컨텐츠 페이지는 특별한 내용을 담고 있어서 방문패턴과 밀접한 관련이 있는 페이지를 의미한다. 이와 같이 웹페이지들을 분류한 다음 분석자는 분석목적에 따라 컨텐츠 페이지만을 분석할 것인지 아니면 모든 페이지들을 분석에 포함시킬 것인지를 결정하고 이에 따라 분석용 데이터를 만들어 주어야 한다. 또한 어떤 분석방법을 사용할 것인가에 따라 데이터를 재배열 하는 등 추가적인 가공을 해주어야 한다.

4. 웹로그 데이터의 분석 및 활용

웹서버 관리자의 일차적인 관심은 웹사이트 사용에 대한 여러 가지 기초통계에 있을 것이다. 즉, 웹사이트에 얼마나 많은 방문객이 왔는지, 어떤 웹페이지가 가장 많이 또는 가장 적게 읽혔는지, 주로 어디에서 방문하고 있는지, 가장 오래 보는 페이지는 무엇인지 등이 주로 관심이 된다. 현재 사용되고 있는 상용 웹마이닝 소프트웨어는 주로 이러한 기초통계를 출력하는 기능 중심으로 되어 있는데, Analog(www.analog.cx), EasyMiner(www.mineit.com), NetTracker(www.sane.com), WebLog(www.weblog.com; www.weblog.co.kr), WebSuxess(www.exody.net), WebTrends(www.webtrends.com; www.webtrends.co.kr) 등이 널리 알려진 소프트웨들로서, <그림 4.1>은 이 중 WebLog의 분석보고서 중 요일별 접속현황 통계를 보여주고 있다.

<그림 4.1> WebLog의 분석보고서 예



<표 4.1> 연관성 규칙발견의 결과 예

# of pages	Support	Confidence	Sequence	Rule
3	5.12	91.80	67.31	A & B → C
4	4.20	93.88	73.25	A & B & D → C
4	2.19	88.89	53.73	A & B & E → C
…	…	…	…	…

A : skincare/skinxfile/skinproblem_01_01
B : skincare/secret/essentials_01
C : skincare/basicinfo/sys19
D : skincare/aboutskin/skintype_01
E : makeup/makeuptechnique/makeup_flow_06

웹로그 데이터를 분석함에 있어 주요 관심 사항 중 하나는 웹페이지들 간의 연관성(즉, 동일한 세션(방문) 내에서 어떤 페이지들이 함께 요청되는 경향이 있는가?)을 파악하는 것이다. 연관성 규칙발견(association rule discovery)은 웹페이지들 간의 연관성을 분석하는 기법으로, 이 때 연관성을 재기 위해 주로 사용되는 측도(mesurement)로는 지지도(support), 신뢰도(confidence), 시차 연관성(sequence) 등이 있다. 즉, n을 전체 사용자 세션의 수, $n(A)$를 A라는 페이지를 방문한 세션의 수, $n(A,B)$를 A와 B를 모두 방문한 세션의 수, $n(A \rightarrow B)$를 A를 먼저 방문한 후에 B를 방문한 세션의 수라고 할 때, 이들은 다음과 같이 계산된다.

$$Support(A \rightarrow B) = n(A, B)/n$$
$$Confidence(A \rightarrow B) = n(A, B)/n(A)$$
$$Sequence(A \rightarrow B) = n(A \rightarrow B)/n(A)$$

예를 들어, <표 4.1>에서 웹페이지 A와 B를 방문한 사용자 중 91.8%가 페이지 C도 방문하는 경향이 있다는 것을 알 수 있다. 이와 같은 연관성의 측도는 표 3.1과 같은 형태의 웹로그 데이터를 분석에 바로 사용할 수 있다는 장점을 가지고 있으나, 웹페이지가 많은 경우 모든 조합을 계산하기 위해 매우 많은 계산시간을 필요로 하므로 이를 효율적으로 계산하기 위한 많은 알고리즘이 제안되어 있다(Agrawal et. al., 1993, 1996; Han et. al., 1997).

웹 사용자의 패턴을 파악하고 분류하기 위해 주로 사용되는 다른 기법으로는 군집분석을 들 수 있다. 웹로그 데이터에 대한 군집분석에서는 보통 세션이 분석대상이 되므로, 먼저 <표 4.2>에서와 같이 세션 프로파일 행렬로 분석용 데이터를 변형해 주어야 한다. 즉, <표 4.2>에서 1은 해당 세션에서 그 페이지가 요청되었음을 의미하고 0은 요청되지 않았음을 의미한다. 이와 같은 세션 프로파일 행렬에 대해 유클리드 거리를 이용한 통상적인 군집분석 알고리즘(예를 들면, k-평균 군집분석)을 적용하는 것이 일반적이지만, 웹페이지들 간의 구조를 고려한 가중거리를 사용하기도 한다(Joshi & Krishinapuram, 2000). <표 4.3>은 1,094개의 세션에 대해 군집분석을 수행한 결과로서 각 군집의 중심을 제시한 것으로, 예를 들어 군집 1에 속하는 세션은 294개이며 이 중 43%인 126개의 세션에서 IBBO라는 웹페이지가 요청되었다는 것을 알 수 있다.

<표 4.2> 세션 프로파일의 예

User Id	S_ID	IBOO	IAAE	IAAF	IAAG	IAAB	…
apple	1	1	1	1	1	0	…
apple	2	0	0	0	0	1	…
dragon	1	1	0	1	1	0	…
dragon	2	1	1	0	0	1	…
670821	1	0	1	0	1	0	…
…	…	…	…	…	…	…	…

<표 4.3> 군집분석의 결과 예 : 군집 프로파일

	전체	군집 1	군집 2	군집 3	군집 4	군집 5
IBOO	0.45	0.43	0.41	0.35	0.57	0.47
IAAE	0.19	0.15	0.21	0.17	0.33	0.15
IAAF	0.17	0.13	0.20	0.12	0.31	0.16
IAAG	0.16	0.12	0.19	0.12	0.31	0.15
IAAB	0.13	0.13	0.11	0.08	0.25	0.12
…	…	…	…	…	…	…
군집 크기	0194	294	346	148	139	177

웹로그 데이터에 대한 분석결과는 일차적으로 웹페이지의 수정 및 재배치 등 웹사이트의 개선에 사용될 수 있다. 즉, 연관성 규칙발견이나 군집분석의 결과로부터 동일 세션(방문) 내에서 어떤 페이지들이 함께 요청되는 경향이 있는지 또는 어떤 경로를 통해 특정 페이지가 요청되었는지 등을 파악함으로써 사용자 측면에서 웹사이트를 개선해 나갈 수 있다. 한편 웹로그 데이터 분석을 이용한 개인화(pesonalization) 또는 추천시스템(recommendation system)이 현재 많은 관심을 가지고 연구되고 있는데(Schafer et. al., 2001; Herlocker et. al., 2000; Sarwar et. al., 2000; Mobasher et. al., 2000), 이는 방문자의 사용패턴에 근거하여 특정 웹페이지를 사용자마다 다르게 구성해 주거나 특정 페이지를 읽도록 추천하고자 하는 것이다. 또한 쇼핑몰을 운영하고 있는 웹사이트에서는 특별한 사용패턴을 가지는 방문자들이 주로 어떤 속성(예를 들어, 연령 및 성별 등)을 가지고 있으며 어떤 상품을 구매하는 경향이 있는지 등을 파악하여, 실시간 또는 전자메일을 이용한 상품추천에도 응용할 수 있다. Amazon, CDNOW, Drugstore, eBay, Movie- Finder, Reel 등 많은 국외 온라인 기업들이 추천시스템을 운영하고 있고 국내에서도 Yes24, 디지틀조선일보, 삼성캐피탈 등 일부 온라인 기업에서 추천시스템을 도입하고 있다.

본 연구의 <표 3.1> 및 <표 4.1>~<표 4.3>은 국내 한 온라인 기업의 실제 웹로그 데이터를 분석한 결과 중 일부이며, 이 회사에서도 주기적인 웹로그 분석을 통해 웹페이지의 개선 및 웹사이트의 구조 변경 등을 수행하고 있다. 또한 연관성 규칙발견과 군집분석을 이용하여 웹페이지 및 상품을 추천하기 위한 시스템 구축을 준비하고 있다.

5. 결론

웹유시지 마이닝은 인터넷이 빠르게 확산됨에 따라 새롭게 주목받고 분야 중 하나이다. 그러나 웹유시지 마이닝은 주로 전산학 및 경영과학 분야에서 주로 연구되어 왔기 때문에 통계학 분야에서 널리 알려진 많은 분석방법들이 아직은 제대로 응용되지 못하고 있다. 즉, 대부분의 상용 소프트웨어들은 웹로그 데이터에 대한 기초통계분석 수준에 머물고 있으며(EasyMiner, 2000; Sane Solutions, 2000; WebLog, 2000), 연관성 규칙발견이나 군집분석 등이 일부 사용되고 있는 정도이다. 웹로그 데이터가 통계학 분야에서 기존에 다루어 왔던 데이터와는 다소 독특한 특성을 가지고 있으나, 주성분(principal component) 및 인자분석(factor analysis), 판별분석(discriminant analysis), 의사결정 나무분석(decision tree analysis) 등 많은 다변량 분석방법들은 웹로그 데이터에 대해서도 적절하게 사용될 수 있을 것이다. 따라서 다양한 통계분석방법들을 웹로그 데이터의 특성에 맞게 변형하고 보완하여 보다 정확하고 효율적인 분석을 수행할 수 있도록 통계학자들의 많은 관심과 연구가 요구되고 있다. 또한 현재 Intelligent Web Miner(www.ecminer.com) 등 몇 개의 추천시스템 솔루션이 출시되어 있으나, 이들 솔루션은 국내 기업의 환경 및 인터넷 사용자의 특성에 맞지 않는 부분이 많아서 이를 바로 적용하기에 어려운 점이 있으며 따라서 국내 현실에 적합한 알고리즘 및 솔루션에 관한 연구가 필요하다.

참 고 문 헌

[1] Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Associations between Sets of Items in Massive Database, *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data*, 207-216.

[2] Agrawal, R., Mannila, H., Strikant, R., Toivonen, H., and Verkamo, A. I. (1996). Fast Discovery of Association Rules, *Advances in Knowledge Discovery and Data Mining, AAAI Press*, 307-328.

[3] Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, 1, 5-32.

[4] EasyMiner (2000). *EasyMiner Version 1.3 Getting Started Guide*, www.mineit.com.

[5] Han, E. H., Karypis, G., Kuram, V., and Mobaser, B. (1997). Clustering Based on Association Rule Hypergraphs, *SIGMOD'97 Workshop on Research Issues on Data Mining and Knowledge Discovery*.

[6] Herlocker, J., Konstan, J., and Riedl, J. (2000). Explaining Collaborative Filtering Recommendations, *Proceedings of the ACM 2000 Conference on Computer Supported Cooperative Work*.

[7] Joshi, A., Krishinapuram, R. (2000). On Mining Web Access Logs, *SIGMOD Workshop on Data Management and Knowledge Discovery*, Dallas, TX.

[8] Kimball, R. and Merz, R. (2000). *The Data Webhouse Toolkit*, John Wiley & Sons, Inc.

[9] Mobasher, B., Cooley, R., and Srivastava, J. (2000). Automatic Personalization Based On Web Usage Mining, *Communication of ACM*, 43, 142-151.

[10] Sane Solutions (2000). *NetTracker 5.0 Professional User's Guide*, www.sane.com.

[11] Sarwar, B. M., Karpis, G., Konstan, J., and Riedl. J. (2000). Analysis of Recommender Algorithms for E-Commerce, *ACM E-Commerce 2000 Conference*.

[12] Schafer, J. B., Konstan, J., and Riedl, J. (2001). Electronic Commerce Recommender Applications, *Journal of Data Mining and Knowledge Discovery*, 5, 115-152.

[13] Srivastava, J., Cooley, R., Deshpande, M., and Ten, P. N. (2000). Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations*, 1, 12-13.

[14] WebLog (2000). *User's Guide*, www.weblog.co.kr.