통계 및 데이터마이닝 기법을 이용한 웹 사이트 분석

류창수* • 서용무**

Analysis of E-biz Site Using Statistics and Data Mining Techniques

요 약

인터넷 기술의 발달과 인터넷 비즈니스의 발전으로 인해 오늘날 사람들은 더욱 많은 시간을 인터넷 상에서 보내고 있다. 사용자가 기업의 웹 사이트를 방문한 기록은 웹 로그 파일이라는 형태로 기업의 서버에 남게 되는데 이러한 로그 파일을 이용해 고객의 행동을 더욱 잘 이해하는 것이 매우 중요한 경쟁력의 요소로 자리 잡게 되었다. 이제까지는 웹 로그를 분석하기 위해 웹 로그 분석 도구를 이용해 왔는데, 경영 의사 결정에 도움이 되는 지식을 발견하기 보다는 단순한 기술적인 통계량을 구하는데 그쳤다. 본 연구에서는 통계와 데이터마이닝 기법을 웹 데이터에 적용하여 경영 의사 결정에 도움이 되는 의미 있는 정보를 추출한다. 이를 위해 실제 인터넷 기업의 데이터를 기반으로 하여 대량 데이터를 데이터마이닝을 위해 전처리 하는 과정과 준비된 데이터를 분석하는 과정을 소개한다. 웹 사이트의 분석은 경영 지식을 찾아내기 위한 과정으로 개별 사이트가 처한 상황에 따라 분석과정이 상이해 질 수 있기 때문에 실제 기업의 데이터를 가지고 분석해 나가는 과정을 보이는 것은 의미 있는 연구라 생각된다.

Key words : 웹, 웹 마이닝, 전자상거래, 연관규칙

I. 서론

인터넷 기술의 발달과 인터넷 비즈니스의 발전으로 인해 오늘날 사람들은 더욱 많은 시간을 인터넷 상에서 보내고 있다. 이때 사용자가 기업의 웹 사이트를 방문한 기록은 웹 로그 파일이라는 형태로 기업의 서버에 남게 되는데 이러한 로그 파일을 이용

* 고려대학교 경영학과 석사과정 (manpha@dreamwiz.com)
** 고려대학교 경영학과 교수 (ymsuh100@dreamwiz.com)

해 고객의 행동을 이해하려는 노력이 계속 되어 왔다. 초창기의 로그 분석 도구들은 기술적 통계량을 분석하여 주로 웹 사이트의 트래픽을 분석하였으나 그 기능이 제한되어 있어 인터넷으로부터 유용한 지식을 이끌어 내지 못했다. 이러한 한계를 극복하기 위해 데이터마이닝 기법을 웹 로그 파일에 적용하여 잠재되어 있는 패턴을 찾아내고 이를 기업의 의사결정에 이용하는 연구가 필요하다. 데이터마이닝 기법을 월드 와이드 웹 (World Wide Web)에 적용하는 것을 Web Mining 이라고 하는데[Kosala & Blockeel 2000], 여기에서 얻은 결과를 이용해서 웹 사이트의 디자인을 개선하거나, 타겟 광고의 효과를 측정하거나 관련 상품 추천을 통해 수익을 증대하는 등에 사용할 수 있다.

현재 인터넷은 사용자에 대한 많은 정보가 매일 매일 쌓이는 장소가 되고 있으며 이러한 데이터를 활용하는 것이 기업의 의사결정에 매우 중요한 요소가 되었다. 그러나 웹 로그로부터 의미 있는 정보나 패턴을 추출하는 자동화 도구는 없는 상태이며 웹 로그 파일을 분석하기 위해 전처리 (preprocessing)하는 과정 또한 몇 가지 경험론적인 방법만이 사용되고 있는 실정이다. 그 만큼 현재 기업 현장에서는 아직까지 실제적으로 웹 마이닝 기법이 사용되지 않고 있다.

본 연구에서는 실제 기업에 대한 웹 마이닝 기법 적용을 적용하기 위한 계획 수립과 분석 과정을 제시하고 분석한 결과를 바탕으로 실제 기업의 의사 결정에 도움이 되는 지식을 얻어 내는 것을 연구의 목적으로 한다. 이를 위해 로그 데이터를 전처리 하고 데이터 베이스에 로드시키는 과정을 거친 후 통계분석, 에러분석, 연관성 규칙 분석 등의 작업을 수행하였다. 논문의 구성은 다음과 같다. 먼저 제 2 장에서는 웹 마이닝에 대한 기존 연구들을 문헌 조사를 통해 알아보고 웹 마이닝에 사용된 데이터마이닝 알고리즘을 살펴본다. 3 장에서는 실제 기업의 데이터에 대하여 분석을 실시하고 그 결과를 해석한다. 4 장에서는 결론 및 추후 연구 과제를 제시한다.

II. 관련 문헌 연구

2.1 Web Mining 의 정의

Web Mining 이란 월드 와이드 웹(World Wide Web)으로부터 유용한 정보를 발견하고 분석하는 과정으로 포괄적으로 정의 될 수 있다. 정보를 발견하고 추출하기 위해서 데이터마이닝 기법을 이용하기 때문에 Web Mining 을 데이터마이닝과 월드 와이드 웹의 교집합이라 할 수 있다. Web Mining 은 아직 분명하게 정의된 용어는 아니고 여러 가지 연구 분야에서 다양한 의미로 쓰이고 있다. 또한 기존 인공지능 분야의 많은 연구 영역들이 Web Mining 의 범주로 포함되고 있으며 그 영역이 아직 확장되고 있는 중이다. 본 논문에서는 Kosala 등이 제시한 일반적인 정의를 따르기로 하겠다. Kosala 등은 Web Mining 을 웹 문서와 서비스로부터 자동으로 정보를 발견하고 추출하기 위해 데이터마이닝 기법을 이용하는 것으로 웹 데이터로부터 미리 알려지지 않은 유용한 정보나 지식을 발견하는 과정으로 정의하였다 [Kosala & Blockeel 2000].

2.2 Web Mining 의 분류

Web Mining 은 기술적 관점에서는 데이터

마이닝 기법을 웹에 적용시키는 것이다. 이 때의 적용 대상이 무엇이냐에 따라 Web Mining 을 Web Content Mining, Web Structure Mining, Web Usage Mining 으로 세분할 수 있다

2.2.1 Web Content Mining

Web Content Mining 웹 상에 존재하는 content, data, documents 등으로부터 유용한 정보를 추출하는 일련의 작업을 일컫는다. Web Content Mining 을 Information Retrieval 측면에서 본다면 어떤 정보를 필요로 하는 사용자가 자신이 원하는 정보를 쉽게 찾을 수 있게 하거나, 필요 없는 정보를 쉽게 걸러낼 수 있게 하는 것을 도와 줄 목적으로 Web Content Mining 이 사용되는 것으로 볼 수 있고, 데이터 베이스 측면에서 본다면 일반적인 키워드 검색 외에 좀 더 정교한 질의어를 사용할 수 있도록 웹 상의 data 를 모델링 하는데 Web Content Mining 이 사용된다고 할 수 있다.

2.2.2 Web Structure Mining

웹사이트들은 서로 복잡하게 링크로 얽혀 있는데 이러한 링크 구조 내에 잠재하는 모델을 찾으려고 하는 작업이 Web Structure Mining 이다. 즉 Web Structure Mining 은 하이퍼링크의 토폴로지(topology)에 기반한 모델로서 서로 다른 웹사이트 간의 유사성이나 관계를 파악하는데 사용된다. 웹 링크의 구조를 파악하면 어떤 사이트의 특징을 발견할 수도 있는데 예를 들어 어떤 주제에 대해 많은 웹 페이지들이 어떤 한 사이트를 링크하고 있는 경우 이 사이트를 authority 사이트라고 하며, 어떤 웹 페이지가 많은 authorities 사이트를 링크하고 있는 경우 이를 Hubs 라고 한다.

2.2.3 Web Usage Mining

Web Usage Mining 은 사용자들이 웹과 상호 작용하는 동안 축적된 정보를 바탕으로 사용자의 행동을 예측하는 기법이라고 할 수 있다. 사용자들의 행동기록이 웹 서버에 로그파일의 형태로 기록되고 이것으로부터 session 정보를 구한 후 의미 있는 패턴을 찾아내는 작업이 일반적인 Web Usage Mining 의 절차이다. Web Usage Mining 은 크게 두 가지 방향으로 연구 되었다. 그 중 첫번째는 일반적인 액세스 패턴을 찾는 작업으로, 기존의 웹 로그 분석 도구들이 가진 한계를 극복하려는 시도에서 시작되었고 두 번째는 개별 사용자의 사용 패턴을 분석하여 차별화 된 서비스를 제공하려는 노력으로서 Mobasher 등은 최근에 개인의 사용 패턴을 학습한 후 각 개인의 선호에 따라 웹 사이트 자체가 적응하는 시스템을 개발하기 시작하였다[Cooley et al 1999b].

2.3 Web Usage Mining 에 대한 기존 연구

본 논문은 실제 인터넷 기업의 데이터를 기반으로 Web Usage Mining 을 실행하여 경영 의사 결정에 실제적인 도움을 줄 수 있는 의미 있는 정보를 추출하는데 있다. 먼저 웹 로그파일에 대해 살펴보고 기존의 Web Usage Mining 에 관한 연구들을 통해 Web Usage Mining 에 대한 전체적인 개요를 살펴보겠다.

2.3.1 로그파일(Log Files)

사용자가 웹사이트를 방문하게 되면 서버에 대한 요청기록이 여러 형태로 남게 된다. 이러한 것들을 Server Log Files 라고 하며

여기에는 access log, error log, agent log, referrer log 등이 있다.

Access Log: 웹 사이트 방문자는 웹 브라우저를 통해 해당 사이트를 방문하게 되는데 이때 브라우저가 서버에 파일을 요청한 기록을 시간과 IP 등의 정보와 함께 남긴 것을 access log 라고 한다. 서버로부터 브라우저에 파일이 전송된 기록이므로 access log 를 transfer log 라고도 한다. 현재 access log 를 기록하는 표준은 NCSA 의 "Common Log Format" 따르는데 구체적인 내용은 표 2.1 에 나타난 바와 같다.

<Access Log 의 예, Common Log Format>

163.152.84.134 - - [01/May/1998:00:10:00 +091800] "GET H.html HTTP/1.0" 200 1000

표 2.1 : Access Log 의 내용

필드 이름	각필드의 내용	예
Host Field	Domain name or IP address	163.152.84.134
Identification Field	Identical field (hyphen)	-
AuthUser Field	AuthUser (ID or Password)	-1
Time Stamp	Date, time and GMT	[01/May/1998: 00:10:00 +091800]
HTTP Request	HTTP Requests	H.html HTTP/1.0
Method	Method of transaction (Get or Post with file name)	GET
Status Code Field	Status or error code (200 for success)	200
Transfer Volume	Size in bytes	1000

- Host Field

서버에 페이지를 요청한 클라이언트의 도메인 네임 혹은 IP 주소를 나타낸다. DNSLookup 기능이 켜져 있는 경우에는 DNS 를 이용해 IP 주소를 domain name 으로 전환하게 되고 그렇지 않은 경우에는 IP 주소가 기록된다. 방문자가 ISP 업체를 통해 접속한 경우에는 고정 IP 를 갖는 것이 아니라 ISP 업체에서 동적으로 IP 를 할당하기 때문에 domain name 을 알기 어려운 문제가 있다.

- Identification Field

FTP, ARCHIE, telnet 등에서 사용되던 Identification Protocol 로서 현재는 사용되지 않기 때문에 거의 모든 경우 hyphen 으로 되어있다.

- AuthUser Field

웹 서버에 의해 보호된 디렉토리에 password 로 접근 했을 때 남는 기록이다.

- Time Stamp

날짜는 'DD/Mon/YYYY' 로 표현하고 시간은 'HH:MM:SS' 로 표현한다. GMT 는 +, - 를 써서 표준 GMT 와의 시간차를 나타낸다.

- HTTP Request

"GET" : 브라우저가 Get 이라는 method 를 이용해 파일을 요청했으며 이때 요청된 파일의 URL 은 H.html 이다. HTTP/1.0 은 HTTP Protocol 의 버전이 1.0 임을 뜻한다.

- Status Code Field

요청된 파일이 제대로 전송되었는지를 나타낸다.

- Transfer Volume

요청된 파일의 전송 크기로서 Method 가 Get 이고 status 가 200 인 때만 값을 가진다.Common Log Format 의 마지막 field 이다.

이외에도 Referrer Log, Agent Log, Cookies 등이 Transfer Log 에 포함되어 기록될 수 있다.

Error Log: 브라우저가 서버에 요청한 파일을 다운로드 받지 못하게 될 때 error 가 발생하게 되고 이것이 error log 에 기록된다. 이 로그 파일을 분석하면 웹 사이트에 어떤 문제가 있는지 진단할 수 있고, 어떤 상황에서 사용자들이 접속을 도중에 끊어 버리는지 파악하는데 도움을 준다.

Referrer Log: Transfer Log 를 보면 사용자가 해당 사이트에서 어떤 페이지를 보았는지 알 수 있는 반면에 Referrer Log 에는 그 페이지를 보기 위해서 어떤 페이지를 거쳐 왔는지에 대한 기록이 남아 있다. 이 로그를 살펴보면 사용자들의 웹 사이트를 찾아오기 위해 어떤 검색엔진을 이용하는지에 대한 정보를 알 수 있고, 사이트의 구조상 어떤 페이지들이 Navigation 을 도와주는 역할을 하는 페이지들인지 알 수 있다. 만약 어떤 사용자가 Yahoo!검색엔진을 이용해 사이트를 찾아 왔다면 Referrer Log 에는 다음과 같은 기록이 남을 것이다.

http://search.yahoo.com/bin/search?p=data+mining+websites → /index.html

Agent Log: 방문자가 사용하는 브라우저의 이름과 버전에 대한 정보가 기록된다. 또한 검색 로봇이 사이트를 방문한 경우에는 그 로봇에 대한 정보가 기록된다. Agent Log 의 예는 아래와 같다.

Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+98) MetaCrawler/1.2b libwww/4.0D

2.3.2 User Session Files 에 대한 연구

2.3.2 .1 User 와 Session 을 파악할 필요성

앞장에서 언급한 웹 서버에 기록된 로그 파일들을 Raw Log Files 라고 하는데 Web Usage Mining 을 위해서는 이 Raw Log Files 을 적당한 형태로 가공할 필요가 있다. 이러한 작업이 Web Usage Mining 의 Preprocessing 에 해당하며 그 결과로 나오는 것이 User Session File 이 된다. 정확한 결과를 얻기 위해서는 Preprocessing 단계에서의 계획과 방법론이 정확해야 한다.

2.3.2.2 User Session File

Web Usage Mining 의 입력물은 user session file 이 된다. Web Usage Mining 이 개별 사용자의 웹 사이트 방문기록으로부터 의미 있는 정보를 찾아내는 과정이므로 Web Usage Mining 의 입력물인 user session file 에는 누가 웹 사이트를 방문했고, 어떤 페이지를 어떤 순서대로 보았으며, 얼마동안이나 그 페이지를 보았는지에 대한 기록이 담겨 있어야 한다. 즉 user session file 에는 한 사용자가 일회 방문동안 거친 모든 웹 페이지에 대한 방문 정보가 담겨 있어야 하는 것이다. 그러나 이것은 이상론이며 일반적으로 이러한 요구사항이 모두 충족되지는 않는다.

2.3.2.3 User Session File 획득의 어려움

Pitkow 는 그의 연구에서 웹 서버 로그에서 user 와 session 을 파악하는데 걸림돌이 되는 것에 관해 지적했다. 그에 따르면 웹 서버 로그로부터 개별 사용자의 정확한 방문기록을 찾아내기 힘든 이유는 local caching 과 proxy server 때문이다[Pirolli et al 1996]. 대부분의 웹 브라우저는 웹 탐색의

속도 개선을 위해 방문했던 페이지들을 저장하게 되는데 사용자가 back 버튼을 사용하게 되면 바로 전 페이지를 보여주기 위해 또 다시 웹 서버에 그 파일을 요청하는 대신 저장되어있던 페이지를 보여주게 된다. 이를 local caching 이라고 하며 local caching 이 일어나게 되면 실제적으로 사용자는 해당 페이지를 방문 하였지만 웹 서버 로그에는 방문기록이 남지 않게 되어 사용자의 정확한 방문 경로를 추적하기 어려운 문제가 생긴다. 또한 웹 서버 로그에는 같은 proxy server 에서 나온 페이지에 대한 요청은 모두 같은 식별자(즉 요청한 클라이언트의 ip 주소)를 가지게 되기 때문에 정확한 사용자의 방문 기록을 얻기 힘들게 된다. 이러한 제약에도 불구하고 웹 서버 로그로부터 되도록 정확한 User Session File 을 얻으려는 노력이 여러 연구자들에 의해 이루어졌다. 그 대표적인 방법은 사용자의 컴퓨터에 저장된 Cookie 파일을 이용하는 방법과 Java Agent 를 사용자의 컴퓨터에 설치하여 모든 방문 페이지를 강제로 서버에 요청하게 하여 local caching 이 일어나지 않게 하는 방법이다. 그러나 이러한 방법은 사용자의 협력을 요구하며 사용자의 컴퓨터에 저장된 Cookie 파일을 이용하는 방법은 사용자가 임의로 Cookie 파일을 지울 수 있는 위험성과 더불어 사생활 침해의 소지가 있기 때문에 현실화 되기 어려운 문제가 있다[Cooley et al 1999a]. 결국 별도의 소프트웨어 설치나 사용자의 특별한 협력이 없어도 정확한 User Session File 을 얻어내는 방법이 필요한데 이제까지의 연구에서는 뚜렷한 해법을 찾아내지 못했으며 단지 Cooley 등이 일련의 연구에서 최대한 정확한 User Session File 을 얻어내는 경험론적인 방법(heuristic solution)을 제시하였을 뿐이다.

2.3.2.4 User Identification

웹 서버 로그에는 개별 IP 가 기록되어 있으므로 일단은 IP 가 다른 경우 사용자가 다른 것으로 가정해 볼 수 있다. 그러나 proxy server 나 firewall 이 이용되는 경우, IP 가 같더라도 실제 사용자는 다른 경우가 많을 수 있기 때문에 이를 해결할 방법이 필요하다. 먼저 IP 가 같다고 하더라도 상이한 운영 체계나 웹 브라우저를 이용해 사이트에 접근한 경우에는 다른 사용자로 볼 수 있다. 또한 동일한 IP 에 동일한 운영체계와 브라우저를 이용해 접근한 사용자라고 하더라도 웹 사이트의 토폴로지(topology)에 의해서는 직접 연결되지 않는 경로를 연달아 방문한 기록은 서로 다른 사용자의 방문으로 취급할 수 있다.

2.3.2 .5 Session Identification

User Identification 의 목적은 웹 사이트에 대한 방문기록을 개별 사용자의 방문 기록으로 나누는 것이다. 그러나 오늘날의 대부분의 상용 사이트에는 하루에도 같은 방문자가 여러 번 방문을 하게 된다. 이러한 이유로 개별 사용자의 방문 기록을 의미 있는 그룹으로 나눌 필요가 있다. 이렇게 나뉜 의미 있는 방문 기록들의 그룹을 session 이라고 한다. Session 을 구하는 방법은 일정한 시간 범위내의 연속된 방문기록으로 session 을 구분하는 것이다. Cooley et al.[1997]은 사용자의 페이지간 방문기록의 시간차이가 30 분을 넘을 경우 session 을 분리하는 방법을 사용하였고, Catledge & Pitkow.[1995]는 25.5 분을 사용하였다.

2.3.3 사용자의 웹 항해를 도와주는 에이전트의 구현

초창기 웹 마이닝 연구는 사용자의 웹에서의 항해를 도와주는 에이전트를 구현하는 것에서 시작하였다. Pazzani et al.[1996]은 사용자가 3 점 척도에 따라 방문하는 페이지를 평가하고 에이전트가 이를 학습하여 어떤 링크가 사용자에게 흥미 있을지를 예측하는 시스템(Syskill& Webert)을 구현 하였다. Syskill& Webert 의 작동 원리는 다음과 같다. 사용자가 3 점 척도 (hot list, warm list, cold list)에 따라 방문한 페이지를 평가한 후 여러 사용자들의 평가를 HTML 소스와 같이 저장하여 평가에 대한 요약을 만든다. 다음에는 각 사용자의 과거의 모든 평가를 요약하여 사용자 프로파일이 학습된다. 이제 방문 되지 않은 웹 페이지의 HTML 소스를 분석하여 이것을 사용자 프로파일과 매치 시켜서 사용자의 흥미도에 대한 평가 등급을 정한 후 표시한다. 즉 HTML 소스에 나타난 단어를 분석하여 hot list 에 자주 나타나는 단어와 유사한 단어가 많이 나타날 경우 흥미도가 높게 평가되는 것이다. 아래 그림 2.1 은 사용자가 방문하지 않은 페이지에 대한 링크에 대해 사용자의 흥미 정도를 미리 평가하여 보여주는 화면이다.

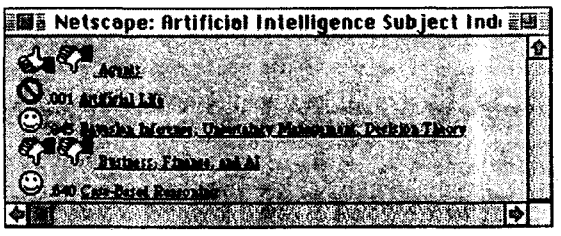


그림 2.1: 사용자에게 흥미 있는 페이지를 예측한 화면

2.3.4 웹 사이트 액세스 패턴에 따른 사용자 클러스터링

Fu et al.[1999]은 웹 사용자를 액세스 패턴에 따라서 클러스터링 하였다. 웹 로그로부터 session 을 얻어내고 이것을 계층구조를 이용해 일반화 하여 사용자 클러스터를 얻는데 사용하였다. 예를 들어 어떤 사용자가 baby furniture, baby toys, diapers 에 관한 웹 페이지를 방문하였다면 이 사용자를 expecting parents 라는 클러스터에 분류할 수 있으며 또한 이러한 미리 분류된 클러스터를 이용하여 어떤 방문자가 baby furniture, baby toys 를 방문하고 있다면 diapers 로 가는 동적 링크를 제공할 수 있다고 제시하였다.

2.3.5 동적링크의 제공과 웹사이트의 개인 맞춤화

Yan et al.[1996]은 웹 로그를 분석하여 사용자 클러스터링을 하고 이를 이용해 동적 링크를 제공하는 시스템을 구현하였다.

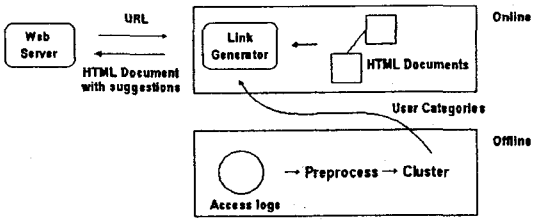


그림 2.2: 동적링크 생성시스템의 구조

그림 2.2 에서 보듯이 이 시스템은 오프라인 모듈과 온라인 모듈로 나뉘어져 있는데 오프라인 모듈에서는 사용자의 액세스 로그로부터 주기적으로 작업을 하여 클러스터를 분류하는 작업을 하고 온라인 모듈에서는 사용자의 partial session 과 매치되는 클러스터를 찾아서 동적 링크를 생성하는 일을 하고 하게 된다. 예를 들어 A 라는 클러

스터가 남성복, 가전제품, 스포츠 용품에 관한 페이지를 포함하고 있는데 사용자가 남성복 페이지를 방문한 후 가전제품 페이지를 방문하였다면 이들 두 페이지에 대한 방문기록이 업데이트 되어 partial session 에 반영되고 이 session 은 클러스터 A 와 매칭이 되게 되고 클러스터내의 페이지이면서 아직 방문 되지 않은 스포츠 용품 페이지에 대한 링크를 제공하게 디는 것이다. Cooley et al.[1999b]은 웹 개인화(web personalization)에 대해 다룬다. 웹 개인화란 사용자의 웹 환경에서의 경험을 사용자의 취향에 맞게 만드는 모든 활동을 말하는 것으로 사용자의 항해를 즐겁게 하거나 사용자에게 필요한 정보를 신속하게 전달하는 등의 활동을 포함한다. 기존의 collaborative filtering 은 사용자의 프로파일을 획득하기 위해서 사용자의 자발적인 협조가 필요한 방법이었지만 이 논문에서는 사용자의 usage data 로부터 실시간으로 개인화 하는 방법에 대해서 다룬다. 개인화의 과정을 전체적으로 살펴보면 Preprocessing stage, Usage mining stage, Recommendation stage 로 나뉜다. Preprocessing 단계에서는 사이트 파일과 서버 로그를 이용하여 데이터 정제작업, 사용자와 트랜잭션 파악, 지지도 필터링 등을 통하여 최종적으로 트랜잭션 파일을 얻게 된다. 서버로그의 preprocessing 에 관한 자세한 방법과 문제점에 대해서는 이 논문의 2.3.2 를 참고하기 바란다. Usage mining 단계에서는 사용자의 행동에 대한 특성을 얻어내기 위해 사용자 트랜잭션 파일에서 연관규칙 탐사를 하게 되면 빈발항목집단을 얻게 되는데 이것으로부터 트랜잭션 클러스터링과 Usage clustering 을 하게 된다. 전자는 각 클러스터가 바슷한 액세스 패턴을 가진 사용자들을 포함하도록 하는 것이고 후자의 경우는 트랜잭션의 클러스터를 구하기보다 URL 들의 클러스터를 구하는 방법이다. Recommendation 단계는 온라인 프로세스로서 동적으로 이루어 지게 된다. 이 단계에서는 사용자의 현재 session 에 대하여 사용자가 다음에 방문하리라 여겨지는 페이지들의 링크를 동적으로 제공하게 된다.

2.3.6 OLAP 기법의 Web Mining 에의 응용

Zaiane et al.[1998]과 Bucher & Mulvenna.[1998]는 웹 로그 데이터를 이용해 데이터 큐브(data cube)를 만들어서 OLAP(Online Analytical Processing)기법을 적용하였다. 웹 로그 데이터에 대한 데이터 큐브를 만들면 단순한 질의어로는 알아낼 수 없는 심층적인 분석이 가능한 장점이 있다. 또한 고객 데이터와 웹 로그 데이터를 결합한 형태로 materialized view 를 만들고 난 후 이를 이용해 마케팅 전략에 사용할 수 있는 중요한 발견을 할 수 있다.

2.3.7 방문패턴 시각화와 디자인 개선

웹 사이트에 대한 사용자들의 방문 패턴을 시각화 한 모델은 디자이너나 개발자에게 웹 사이트의 구조와 문제점에 대한 직관을 제공한다. Cadez et al.[2000]은 웹 사이트 상에서의 행동이 유사한 방문자끼리 클러스터를 구성하고 각 클러스터내에서 개별 사용자의 행동을 시각화 하는 시스템을 (WebCANVAS) 구현하였다. 여러 사용자 클러스터와 개별 사용자의 행동 양식을 일목요연하게 시각화 함으로써 사이트 관리자가 사용자들이 사이트를 이용하는 특징을 직관적으로 알 수 있게 하였다. [Spiliopoulou 2000]에서는 웹 사이트 디자이너가 방문 패

턴 탐사를 통해 웹 사이트가 원래 의도되어 진 데로 사용되고 있는지를 평가할 수 있게 하는 방법을 제시하였다. 디자이너가 두 페이지간 또는 세 페이지간의 패턴을 탐사한 후 이것들이 사이트 구조상 당연한 것인지 아닌지를 파악할 수 있게 한 것이다.

2.4 분석에 사용되는 연관규칙 알고리즘

본 논문에서는 web usage mining 을 위해 웹 로그 파일에 대하여 데이터마이닝 기법 중 연관 규칙탐사 기법을 적용 하였다. 이번 절에서는 연관 규칙 탐사에 대해 간략히 설명하겠다.

2.4.1 연관 규칙 탐사(장바구니 분석)

대형 할인점, 백화점 또는 인터넷 쇼핑몰 등의 경우 많은 고객들이 산 상품 거래 정보를 모아 두었다가 이를 분석하면 마케팅 활동에 이용할 수 있는 유용한 정보를 얻을 수 있다. 예를 들어, "상당수의 고객들이 목탄을 사면 라이터도 같이 사더라." 라는 구매 물건들 사이에 연관성 규칙을 알아낼 수 있다. 이런 연관성 규칙은 매장에서 상품을 진열하거나 끼워 팔기에 이용할 수가 있다. 이처럼 주로 고객들의 거래 자료를 분석하여 상품들 사이의 연관성 규칙을 알아내는 방법을 연관 규칙 탐사 또는 장바구니 분석이라고 한다.

2.4.2 연관 규칙의 정의

연관 규칙은 보통 A ⇒ B 로 나타낸다. 이는 어떤 고객이 A 라는 물건을 사면 B 라는 물건도 산다는 것을 의미한다. 여기서, A 는 반드시 하나의 물건을 가리키지는 않으며 복수의 물건도 가능하지만, B 는 대부분의 경우 하나의 물건이다. 이러한 연관성 규칙을 마케팅 활동에 활용하기 위해서는, 그 규칙에 관련된 상품들이 전체 고객 중 상당 %를 차지하는 고객들의 거래 내역에서 발견되어야 하며(지지도), 동시에 A 상품을 산 고객 중에서 상당 %의 고객이 B 상품을 샀어야 한다(신뢰도). 따라서 연관성 규칙을 찾기 위한 알고리즘에서는 사용자가 미리 정한 지지도와 신뢰도의 수치를 입력할 것을 요구한다. 이렇게 찾아낸 연관 규칙에 대해 다시 lift 값을 계산하여 그 값이 1 보다 큰 연관 규칙만을 활용하게 된다. 다음은 지지도, 신뢰도 그리고 lift 에 대한 정의이다.

① 지지도(support): 연관성 규칙 A ⇒ B 의 support 가 x%라는 것은 전체 고객 중 x%의 고객이 물건 A 와 B 를 함께 구매했다는 것을 의미한다.

② 신뢰도(confidence): 연관성 규칙 A ⇒ B 의 confidence 가 y%라는 것은 A 를 산 고객 중에서 y%의 고객이 B 를 샀다는 것을 의미한다.

③ lift: 연관성 규칙 A ⇒ B 의 lift 는 P(B|A) / P(B)로 정의된다. 즉, A 를 산 고객 중에서 B 를 산 고객의 %를, 전체고객 중에서 B 를 산 고객의 %로 나눈 값이다.

2.4.3 연관 규칙 탐색 알고리즘

트랜잭션 데이터베이스에서 연관 규칙을 찾는다는 것은 사용자가 명세한 최소 지지도(minimum support)와 최소 신뢰도(minimum confidence) 보다 큰 지지도와 신뢰도를 갖는 항목집단(itemset)을 찾는 것이라고 말할 수 있고, 최소 지지도보다 큰 지지도를 갖는 항목 집합들을 빈발하다라고 말한다. 이런 경우에 항목들의 집합 X 를 빈발 항목집합(frequent itemset)이라고 한다. 우리가 최소

지지도에 관심이 있는 이유는 트랜잭션 데이터베이스에 대해 관심 있을 정도로 자주 일어나는 항목만을 고려해야 하기 때문이다. 일반적으로 연관 규칙 탐사는 빈발항목을 찾고, 빈발항목 집합을 이용해 규칙을 생성하는 두 단계로 나누어진다. 대표적인 알고리즘인 Apriori[Agrawal & Srikant 1994] 알고리즘을 예로 들어 설명하면, 데이터 베이스를 읽어가면서 각 패스(pass)마다 찾아낸 빈발 항목집합이 다음 패스의 후보 항목 집합을 생성하게 된다. 후보집합 생성 시 apriori-gen 이라는 함수를 사용하는데, apriori-gen 함수는 빈발 항목을 조인하는 것과 모든 빈발 항목집합의 부분집합 또한 빈발 항목집합 이라는 특성을 이용, 전지(pruning)하는 2 개의 단계로 이루어져 있다. apriori 알고리즘의 실행 시 k 개의 항목을 가진 후보 항목 집합을 생성하기 위해 (k-1)개의 항목들로 구성된 빈발 항목집합을 이용한다. 그림 2.3 은 최소 지지도가 2 일 때 전체 트랜잭션 데이터 베이스 D 에서 1 개짜리 빈발항목 집합 L_1 과 후보 항목 집합 C_1 을 구하는 것을 예를 들어 보인 것이다.

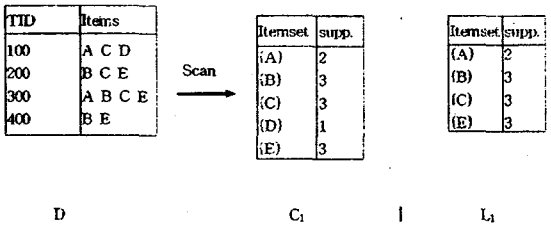


그림 2.3: 연관규칙 탐사 1(*minsup =2*)

전체 데이터 베이스를 스캔하여 한 개짜리 후보항목 집합(C_1)을 구한 후 이것의 지지도를 계산하면 그림 2.3 과 같이 1 개짜리 빈발항목 집합(L_1)을 얻는다. 이제 apriori-gen()함수를 이용해 L_1 에서 C_2 를 구해보자. L_1 을 서로 조인하면 {AB}{AC}{AE}{BC}{BE}{CE}를 얻게 된다. 여기에 '모든 빈발항목집합의 부분집합도 빈발하다'라는 특성을 적용해 전지하면 C_2 는 {AB}{AC}{AE}{BC}{BE}{CE}가 된다.

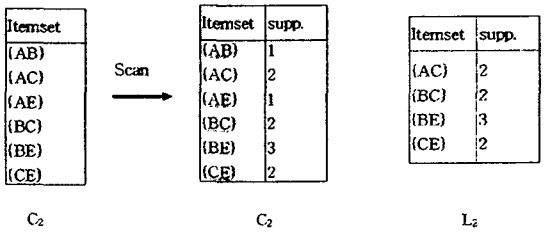


그림 2.4: 연관규칙 탐사 2(*minsup =2*)

두 번째 패스에서 2 개 짜리 후보항목 집합의 지지도를 계산하면 그림 2.4 와 같이 L_2 를 얻게 된다. 여기에 후보생성 알고리즘인 apriori-gen()을 적용하면, L_2 를 서로 조인하여 {ABC}{ABE}{ACE}{BCE}를 얻게 되고, {ABC}의 경우 부분집합인 {AC}가 L_2 에 속하지 않음을 알 수 있다. 이런 식으로 '모든 빈발항목집합의 부분집합도 빈발하다'라는 특성을 적용하면 {BCE}를 C_3 로 얻게 된다.

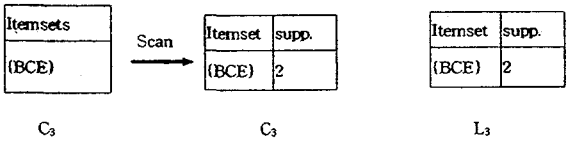


그림 2.5: 연관규칙 탐사 3(minsup =2)

세 번째 패스에서 3 개 짜리 후보항목 집합의 지지도를 계산하면 그림 2.5 와 같이 L_3 를 얻게 된다. 그런데 L_3 에서 더 이상의 후보항목 집합을 얻어낼 수 없게 되어 알고리즘은 종료된다.

2.4.4 연관 규칙이 활용되는 분야

이처럼 연관 규칙은 주로 함께 구매되는 상품간의 연관성을 찾아내는데 사용된다. 연관 규칙 탐사를 장바구니 분석(market basket analysis)이라고 하는 이유가 여기에 있다. 연관 규칙은 판내 상품간의 연관성, 통신 회사의 다양한 서비스간의 연관성, 웹 사이트에서 특정 방문 페이지간의 연관성 등을 이해하기 쉽게 파악할 수 있다. 이런 특성을 이용해 관련 상품을 가까운 곳에서 찾을 수 있도록 제품 진열장의 배치를 결정하거나, 특정 서비스 이용고객에게 새로운 서비스를 제공하거나, 웹사이트의 구조를 재편하는 등 비즈니스 목적에 맞게 다양하게 활용할 수 있는 장점이 있다.

III. 실제 인터넷 기업을 대상으로 한 Web Usage Mining 분석

3.1 대상 사이트 개요

분석 대상 사이트는 회원을 대상으로 다이어트 정보를 제공하는 사이트이다. 다양하고 건전한 정보를 제공하는 것이 사이트의 주된 목적이기 때문에 방대한 웹 페이지들로 구성 되어 있으며 표 3.1 과 같이 컨텐츠의 디렉토리가 매우 많고 복잡한 전형적인 컨텐츠 사이트의 모습을 보여 주고 있다.

표 3.1: 분석 대상 사이트의 카테고리 구성

커뮤니티 디렉토리	개인 일기장
	연령별 일기장
	친구 일기장
	소모임
	추억의 일기장
	이벤트
다이어트 정보	
식이요법	식단작성
	칼로리 정보
운동요법	운동방법
	부위별 운동법
미용정보	피부관리
	피부치료
	몸 각 부분 관리
기타	다이어트 사전
	다이어트 클리닉

참고로 대상 사이트의 시스템 환경은 아래 표 3.2 와 같다.

표 3.2: 분석대상 사이트의 시스템 환경

웹 서버	사용 언어	데이터 베이스	운영 체계
IIS 4.0	ASP, HTML	SQL SERVER	Windows NT4.0

3.2 분석 대상 데이터 설명 및 로그 포맷 설명

분석 대상 사이트의 로그 파일 형식은 확장된 W3C 형식을 쓰고 있다. IIS 서버에서는 사이트 관리자가 로그에 남길 필드를 선택할 수 있는데, 대부분의 사이트는 로그 파일의 크기에 대한 부담 때문에 모든 필드를 선택하고 있지는 않다. 대상 사이트의 경우 단지 6 개의 필드만 선택해서 로그에 남기고 있지만 매일 50MB 정도 크기의 로그 파일이 생성 되고 있다. 표 3.3 은 각 필드에 대한 설명이다.

표 3.3: 로그 파일 각 필드에 대한 설명

필드	설명
Date	서버에 해당 파일을 요청한 날짜
Time	서버에 해당 파일을 요청한 시간
c-ip	파일을 요청한 클라이언트의 ip
cs-method	클라이언트가 파일을 요청한 형식
cs-uri-stem	요청된 파일의 url
sc-status	파일 요청에 대한 상태 코드

3.3 분석과정

그림 3.1 은 web usage mining 의 전체적인 분석 과정을 나타내고 있다. 사용자가 웹 사이트를 방문한 기록이 웹 서버에 남게 되고 이것을 관계형 데이터 베이스에 로드 하여 통계적 분석과 데이터마이닝 기법을 적용하여 web usage mining 을 하게 된다.

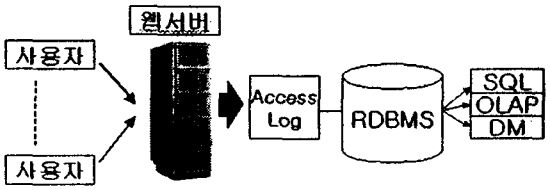


그림 3.1: 분석과정

우선 원본 로그파일을 획득한 후 이를 정제(cleaning)하는 작업이 필요하다. 텍스트 편집기를 이용하여 로깅 과정에서 필드가 깨어졌거나 에러가 있는 항목을 제거 한다. 데이터의 정제 작업이 끝나면 로그 파일을 DBMS 의 로더 유틸리티를 이용하여 데이터 베이스의 테이블 형태로 만들 수 있는데, DBMS 의 로더 유틸리티를 사용하기 위해서는 먼저 그림 3.2 에서와 같이 해당 테이블을 정의하여 두어야 한다.

```
create table nt
(
view_date varchar2(15),
view_time varchar2(15),
c_ip varchar2(20),
cs_method char(6),
cs_uri_stem varchar2(4000),
sc_status char(8));
```

그림 3.2: 테이블 만들기의 예

테이블이 정의되었으면 로더 유틸리티를 이용해 그림 3.3 에서와 같이 로그 파일을 해당 테이블로 로드시킨다.

```
load data
infile *
into table nt
fields terminated by ' '
(view_date,view_time,c_ip,cs_method
,cs_uri_stem,sc_status)
begindata
```

그림 3.3: 로딩을 위한 콘트롤 파일의 예

로그 파일이 관계형 테이블로 만들어 진 다음에는 분석에 필요 없는 레코드들을 삭제하여야 한다. 클라이언트(웹 브라우저)가 웹 서버에 파일을 요청한 기록들이 로그 파일에 남아 있게 되는데 문서 페이지를 요청하게 되면 태그에 포함되어 있는 *.gif, *.jpg 등의 그림 파일이 자동으로 요청되기 때문에 이런 그림 파일에 대한 요청 기록은 삭제하여야 한다. 삭제는 그림 3.4 에서와 같이 SQL 질의어를 사용해 가능하다.

```
DELETE *
FROM nt_log
WHERE url Like '*.jpg' Or url Like
'*.gif' Or url Like '*.css';
```

그림 3.4: data cleansing 을 위한 sql 질의어

data cleansing 이 끝나면 user session 을 구하여야 한다. 2 장에서는 user session 에 대한 기존의 연구를 기술하였는데 본 논문에서는 기존 연구에서 일반적으로 쓰인 방법대로 접근 ip 와 시간을 이용해 user session 을 구했다. user session 을 구하기 위해서는 프로그래밍을 할 필요가 있는데 본 논문에서는 다음과 같이 asp 로 프로그램을 작성하여 사용

자 IP 별로 분류한 후 방문 페이지간의 접속 시간이 30 분 이상이 날 때 session 을 나누는 방법을 사용하였다.

<user session 을 구하기 위한 프로그램의 일부>

```
...
//변수 선언
Dim
AllRec,rows,i,cols,view_date,view_time,ip,se_0,se_1,strSql,time_0,time_1,diffdate,id,seq,session_id


session_id = "session-0"
id = 1
Response.write "AA04<br>"
i=1
//레코드 마다 각 필드를 변수로 받아온다
Do until Rs.Eof
    se_1 = se_0
    time_1 = time_0
    view_date =Rs(0)
    view_time =Rs(1)
    ip =Rs(2)
    seq =Rs(3)
    se_0 = ip
    time_0 = view_date & " " & view_time

    If time_1 = "" Then
      time_1 = view_date & " " & view_time
    End If

    diffdate = DateDiff("n",time_1,time_0)

//페이지뷰 간의 시간차가  30 분을 초과하
//면 session 을 1 증가 시킨다
    If se_0 <> se_1 Or Cint(diffdate) > 30 Then
      id = id + 1
      session_id = "session-" & id
      time_1 = time_0
    End If
//update 를 이용하여 레코드마다 session_id
를 찍어준다.
    strSql = "Update nt_log Set session_id='" &
session_id & "' ,datediff = '" & diffdate & "'
Where seq = " & seq
    DbCon.Execute strSql,3,adCmdText
    Response.write ip & ":::" & view_date & " "
& view_time & ":::" & session_id & " ::: " & i &
"<br>"

    i = i+1
    Rs.MoveNext
loop
  Rs.Close
  DbCon.Close
  Set Rs = Nothing
  Set DbCon = Nothing
%>
```

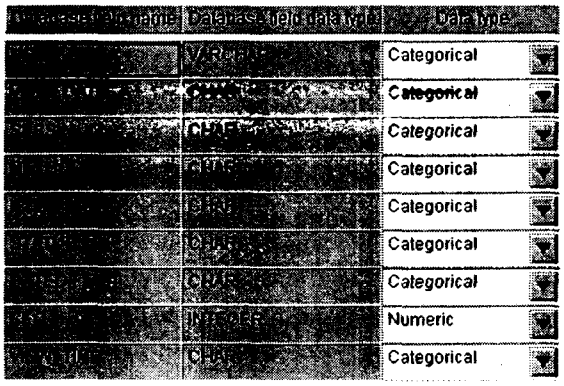
3.4 분석내용

3.4.1 웹 페이지간 방문의 연관성 분석

Web Usage Mining 을 위해 우선 전처리를 거친 웹 로그 데이터를 관계형 데이터 베이스인 IBM DB2 에 로드 하였다. 방문자가 웹 사이트를 방문할 때 어떤 페이지를 자주 보는지를 알아내기 위해 데이터마이닝 기법 중의 하나인 연관규칙 탐사를 실시하였다. 연관규칙 탐사에는 IBM Intelligent Miner 가 사용되었다. 먼저 데이터 베이스에 로드된 로그 데이터를 Intelligent Miner 에서 불러들여 데이터 마이닝 작업에 필요한 데이터 오

브젝트로 만든다. 표 3.4 는 Intelligent Miner 가 로그 데이터를 읽어들인 모습이다.

표 3.4: 데이터 오브젝트



데이터 오브젝트를 만들었으면 연관규칙 분석을 실시한다. 이때 그림 3.5 에서와 같이 아이템 필드는 URL 페이지를 선택하고 트랜잭션 필드는 sessionid 를 선택한다.

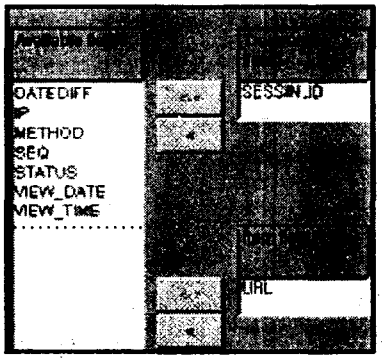


그림 3.5: 필드선택

연관 규칙 탐사를 하기전에 파라미터를 지정해주어야 하는데 support 와 confidence 를 변형 시켜가면서 여러 번의 실행을 해 보아야 한다. Rule length 는 2 를 선택하여 두 개 페이지 간의 연관성을 분석하였다. Intelligent Miner 는 연관 규칙의 결과를 결과 화면 창에서 그래프의 형태로 보여준다. 그림 3.6 에는 규칙의 길이를 2 로 하고 최소 신뢰도(minimum confidence)를 25%로 하였을 때의 연관 규칙 결과를 예시하였다.

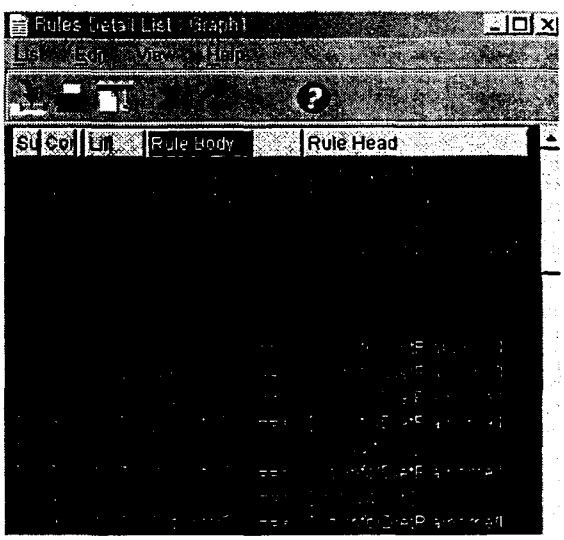


그림 3.6: 연관 규칙 결과 화면

웹 마이닝 분석은 내부 개발자의 도움이 필수적인데 수 많은 연관 규칙 중에서 의미가 있는 것을 골라내기 위해서는 웹 사이트 구조에 대한 지식이 필요하기 때문이다. 찾아낸 연관 규칙에 대해 리프트가 1 보다 큰 것 중에서 해당 사이트의 경영 의사결정에 도움을 줄 수 있는 의미 있는 결과는 표 3.5 와 같다.

표 3.5 연관규칙 결과

	연관 규칙	지지도	신뢰도
1	캘린더 → 먹은 음식 선택	4.3	69.4
2	식이요법메인 → 식이요법 Q&A	3.0	80.0
3	일기장 → 20 대 개인일기장	3.2	57.8
4	다이어트뉴스 → 살빼기정보 → 성공담실패담	2.8 3.5	48.2 58.0
5	살빼기정보 → 똑똑한 살빼기 → 자가진단 → 피부미용	3.2 5.3 3.5	73.3 69.2 56.6

표 3.5 의 결과를 분석하면 1 의 경우 다이어트 캘린더 페이지를 방문한 고객은 여러 가지 기능을 선택할 수 있는데 그 중 자신

이 그날 섭취한 음식을 기록하는 작업을 주로 하는 것으로 파악된다. 이는 다이어트 캘린더가 원래 목적대로 쓰여지고 있음을 나타내고 있다. 고객의 충성도를 계속적으로 유지하기 위해서는 캘린더 페이지에서 다양한 서비스를 제공할 필요가 있으며 특히 섭취한 음식의 칼로리를 계산해 주는 기능을 자세히 보강할 필요가 있다. 2 번 규칙의 경우 식이요법 페이지를 접근한 고객은 식이요법에 대한 질문과 답 페이지를 주로 보는 것으로 나타났다. 3 번 규칙은 개인 일기장을 방문한 고객은 주로 20 대일기장을 주로 보는데 이는 당 사이트의 주 고객이 20 대임을 감안할 때 당연한 결과로 생각된다. 그러나 개인 일기장의 메인 페이지의 경우 20 대 취향으로 꾸며져 있지는 않은데 이런 결과를 이용해 아예 메인 페이지를 20 대 취향으로 바꾸는 것도 고려해 봐야 한다. 4 번 결과의 경우 매일 매일 새로운 다이어트 관련 소식을 올리는 '다이어트 뉴스'를 본 고객은 주로 살빼기 정보나, 성공담 실패담 페이지로 이동하는 것을 알 수 있다. 5 번 결과를 보면 살빼기 정보 페이지를 방문한 고객의 상당수가 똑똑한 살빼기, 자가진단, 피부미용 등의 다양한 컨텐츠 카테고리로 이동하는 것을 보아 이 페이지가 일종의 hub site 역할을 하는 것을 알 수 있다. 이런 경우 살빼기 정보 페이지에 고객의 관심을 끌만한 이벤트나, 쇼핑몰 또는 고급정보로 바로가는 링크를 삽입하여 활용도를 높일 수 있다.

3.4.2 통계분석

로그 파일을 관계형 데이터 베이스의 테이블로 작성한 후에는 질의어와 통계 프로그램을 이용해 다양한 기술적 통계량을 얻을 수 있는 장점이 있다. 이러한 방법을 이용하면 굳이 상용 로그 분석 툴을 구입하지 않더라도 로그 분석 툴이 제공하는 기능들을 구현할 수 있다. 이번 절에서는 로그 데이터로부터 얻을 수 있는 여러 통계치 들을 분석하였다.

3.4.2.1 컨텐츠 도용 탐지

웹 서버에 저장된 액세스 로그 파일에는 ip 주소가 기록되어 있다. 서로 다른 방문자는 서로 다른 ip 주소를 남기게 되는데 이를 통계적으로 분석해 보면 웹 사이트에 대한 특정 사용자의 사용 행태를 알 수 있다. 물론 이러한 방법은 기존의 웹 로그 분석 도구들이 실행한 방법으로 사이트 전체의 사용행태에 대한 정보를 제공해 주지는 않지만 특이값이나 전체적인 통계를 알 수는 있다. 예를 들어 표 3.6 에는 특정일 방문자에 대한 통계가 나타나 있는데(IP 는 개인정보 보호를 위해 영문자로 표현하였다) 어떤 방문자는 다른 방문자에 비해 상당히 많은 웹 페이지를 요청하고 있다.

표 3.6: top10 접근 ip

IP	%	# requests
A	1.4679	962
B	0.692749	454
C	0.585938	384
D	0.550842	361
E	0.534058	350
F	0.502014	329
G	0.485229	318
H	0.480652	315
I	0.450134	295
J	0.405884	266

특히 첫번째 ip 의 경우 하루에 약 1000 여건에 달하는 문서에 액세스를 했는데 이

것은 매우 특이한 경우로 주의 깊게 관찰할 필요가 있다. 특히 내부 개발자의 ip 를 제거한 상태에서 특정 사용자가 하루에 1000 여 페이지에 달하는 문서를 요청했다는 것은 의심 할 만한 일이다. 테이블에 저장된 로그 파일에 대해 그림 3.7 에서와 같이 질의어를 이용해 분석해 보면 더 자세한 사항을 알 수 있다.

```
SELECT Nt_log.ip, Nt_log.url,
Nt_log.view_time
FROM Nt_log
WHERE (((Nt_log.ip)="210.123.45.83"));
```

그림 3.7: 특정 ip 가 방문한 페이지를 조회하는 질의어

분석 결과 해당 ip 는 1 시간 50 여분의 시간동안에 무려 962 페이지에 대해 전송 요청을 한 것으로 나타났다(표 3.7). 이는 1 분에 약 9 페이지에 달하는 전송량으로 정상적인 사용자라기보다는 에이전트 프로그램을 이용해 웹 사이트의 특정 부분을 대량으로 다운로드 해간 것으로 보인다.

표 3.7: 특정 ip 분석 대량의 자료가 다운로드 되어짐

Modified IP	Accessed page	View_time
xxx.xxx.xxx.xxx	a	8:22:14
xxx.xxx.xxx.xxx	b	8:22:17
xxx.xxx.xxx.xxx	c	8:22:20
xxx.xxx.xxx.xxx	d	8:22:22
xxx.xxx.xxx.xxx	e	8:22:24
xxx.xxx.xxx.xxx	f	8:22:25
xxx.xxx.xxx.xxx	g	8:22:29
xxx.xxx.xxx.xxx	h	8:22:31
xxx.xxx.xxx.xxx	i	8:22:32
xxx.xxx.xxx.xxx	j	8:22:34
xxx.xxx.xxx.xxx	k	8:22:35
xxx.xxx.xxx.xxx	l	8:22:36
xxx.xxx.xxx.xxx	m	8:22:38
xxx.xxx.xxx.xxx	n	8:22:45

이처럼 웹 서버 로그를 관계형 데이터 베이스의 테이블로 저장해 둔 경우에는 간단한 통계치로부터 의미 있는 분석을 할 수 있는 경우도 있다. 특히 최근 들어 인터넷 비즈니스의 발달로 인해 매일 수 많은 웹 사이트가 생겨나 신생 후발 업체가 기존 업체의 컨텐츠를 무단 도용하는 일이 많아져서 사회적 문제가 되고 있는 상황에서는 웹 사이트 관리자가 직접 이러한 분석을 할 수 있는 능력이 필요하다고 할 수 있다.

3.4.2.2 에러 분석

로그 파일의 필드 중 sc-status 는 파일 요청에 대한 상태 코드를 나타낸다. 웹 서버가 클라이언트가 요청한 대로 제대로 파일을 전송하였을 경우에는 200 의 상태 코드가 로그에 남게 된다. 상태 코드의 자세한 내용은 표 3.8 과 같다[W3C].

표 3.8: 상태코드의 내용

HTTP 상태 코드	
트랜잭션이 성공한 경우	
200	request가 성공적으로 완료되었음
201	request가 POST method이었으며 성공적으로 완료되었음
202	request가 서버에 전달되었으나 처리 결과를 알 수 없음. 배치 처리를 요한 경우
203	GET request가 실행되었으며 부분적인 정보를 리턴 하였음
204	request가 실행되었으나 클라이언트에게 보낼 데이터가 없음
트랜잭션의 redirection	
300	요구된 request가 여러 위치에 존재하는 자원을 필요로 하므로 response는 위에 대한 정보를 보낸다. 클라이언트는 가장 적당한 위치를 선택하여야 함
301	request에 의한 요구된 데이터는 영구적으로 새로운 URL로 옮기어 졌음
302	request가 요구한 데이터를 발견하였으나 실제 다른 URL에 존재함

304	If-Modified-Since 필드를 포함한 GET method를 받았으나 문서는 수정되지 않았음
오류 메시지	
400	request의 문법이 잘못되었음
401	request가 서버에게 Authorization: 필드를 사용하였으나 값을 지정하지 않았음. 서버는 WWW-Authenticate response header를 통해 가능한 인증 스킴을 보낸다.
402	request가 요구한 일은 비용이 요구되지만 request header의 Chargeto 필드에 아무값도 보내지 않았음. 현재는 구현되지 않았음
403	request는 금지된 자원을 요구하였음
404	서버는 요구된 URL을 찾을 수 없음
405	클라이언트는 자원을 액세스하기에 부적합한 method를 이용하였음.
406	요구된 자원을 발견하였으나 자원을 타입이 request header의 accept: 필드와 일치하지 않아서 전송할 수 없음
410	요구된 자원은 더 이상 활용가능하지 않음
500	서버에 내부적으로 오류가 발생하여 더 이상을 진행할 수 없음
501	요청된 request는 합법적이나 서버는 요구된 method를 지원하지 않음
502	클라이언트는 다른 서버(보조서버)로부터 자원 액세스를 요구하는 서버에 자원을 요구하였으나 보조 서버가 유효한 응답을 전달해오지 않았음
503	서버가 바쁘기 때문에 서비스를 할 수 없음
504	502의 오류와 유사하나 보조 서버의 응답이 너무 오래 지체되어 트랜잭션이 실패하였음

에러 분석 결과 약 25%가 302 에러를 내어서 에러 비율이 높았는데 이는 분석 대상 사이트의 잦은 구조개편으로 발생한 것으로 구조가 안정화 되면 줄어 들 것으로 보인다. 그림 3.8 에는 가장 자주 발생하는 상태 코드의 발생 비율을 표시하였다.

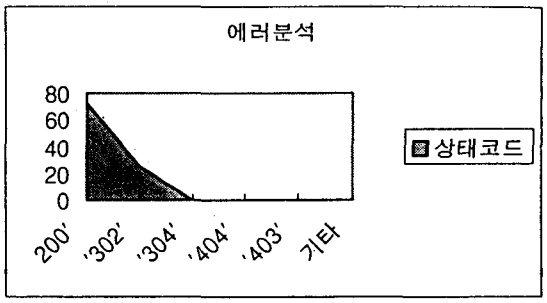


그림 3.8: 상태 코드의 발생 비율

IV. 결론

본 연구에서는 실제로 인터넷 비즈니스를 하고 있는 사이트를 대상으로 통계와 데이터마이닝을 이용하여 웹 사이트를 다양한 방법으로 분석하여 경영의사 결정에 도움이 될 수 있는 정보를 추출하였다. 이를 위해 웹 로그를 관계형 데이터 베이스에 로드한 후 IBM Intelligent Miner 를 이용하였다. 본 연구는 가공된 사이트가 아니라 복잡한 구조의 컨텐츠 사이트를 대상으로 웹 마이닝을 실행하였다는 점에 의의가 있으며, 웹 마이닝을 위한 데이터 전처리 과정과 분석 절차를 소개하였다. 연구 과정에서 컨텐츠 사이트 설계시 디렉토리 구조나 페이지 구성을 미리 웹 마이닝을 염두에 두고 설계하는 것이 예측 결과를 높일 수 있음을 알아내었다. 추후 웹 마이닝의 예측 결과를 높일 수 있는 사이트 설계방안에 대해 연구할 필요가 있다.

참고문헌

Agrawal, R. and R. Srikant, "Fast Algorithms for Mining Association Rules", *Proceedings of the VLDB Conference*, Santiago, Chile, September, 1994.

Bucher, A. G. and M. D. Mulvenna, "Discovering Internet Marketing Intelegence through Online Analytical Web Usage Mining", *SIGMOD Record*, 27(4), 1998, pp. 54-61.

Catledge, L. and J. Pitkow, "Characterizing browsing behaviors on the World Wide Web", *Computer Networks and ISDN Systems*, pages 385-392, 1996.

Cooley, R., B. Mobasher, and J. Srivastava, "Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns", *Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX-97)*, November, 1997.

Cooley, R., B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", *Journal of KAIS*, 1(1), 1999.

Cooley, R., B. Mobasher, and J. Srivastava, "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs", *Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, November, 1999.

Fu, Y., K. Sandhu and M.Y. Shih, "Clustering of Web Users Based on Access Patterns", *Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99)*, San Diego, CA, August, 1999.

Kosala, R. and H. Blockeel, "Web Mining Research: A Survey", *ACM SIGKDD*, July, 2000.

Mena, J., *Data Mining Your Web Site*, Butterworth-Heinemann/Digital Press, 1999.

Pazzani, M., J. Muramatsu, and D. Billsus, "Syskill & Webert: Identifying interesting web sites", *Proceedings of the AAAI-96.*

Pirolli, P., J. Pitkow, and R. Rao. "Silk from a Sow's Ear: Extracting Usable Structures from the Web", *Proceedings of the 1996 Conference on Human Factors in Computing Systems*, Vancouver, British Columbia, Canada, 1996.

Srikant R., and R. Agrawal, "Mining Sequential Patterns, Generalizations and Performance Improvements", *Proceedings of the Fifth Int'l Conference on Extending Database Technology (EDBT)*, Avignon, France, March, 1996.

Yan T., M. Jacobsen, H. Garcia-Monila, and U. Dayal. "From user access patterns to dynamic hypertext liniking". In *Fifth International World Wide Web Conference*, Paris, France, 1996.

Zaiane O. R., M. Xin, and J. Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", *Proceedings of the Advances in Digital Libraries Conf. (ADL'98)*, Santa Barbara, CA, April 1998, pp. 19-29.

Cadez I., D. Heckerman, C. Meek, P. Smyth and S. White, "Visualization of navigation patterns on a Web site using model-based clustering", *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, Pages 280 – 284.

Chi E. H., P. Pirolli and J. Pitkow, "The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a Web site", *Proceedings of the CHI 2000*

conference on Human factors in computing systems, 2000, Pages 161 – 168.

Spiliopoulou M, "Web usage mining for web site evaluation", *CACM 43, 8*, August, 2000, Pages 127 – 134.

W3C HTTP/1.1: Status Code Definitions , http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html