IE Interfaces
Vol. 16, No. 1, pp. 54-62, March 2003.

기계학습 기반의 웹 마이닝을 이용한 고객 세분화에 관한 연구

이건창[1†] · 정남호[2]

[1]성균관대학교 경영학부/ [2](주)크레듀 기획 · 컨설팅팀

Identification of Customer Segmentation Strategies by Using Machine Learning-Oriented Web-mining Technique

Kun-Chang Lee[1] · Nam-ho Chung[2]

[1]School of Business Administration, Sung Kyun Kwan University, Seoul, 110-745

[2]Planning & Consulting Department, Credu Co., Seoul, 100-814

With the ubiquitous use of the Internet in daily business activities, most of modern firms are keenly interested in customer's behaviors on the Internet. That is because a wide variety of information about customer's intention about the target web site can be revealed from IP address, reference address, cookie files, duration time, all of which are expressing customer's behaviors on the Internet. In this sense, this paper aims to accomplish an objective of analyzing a set of exemplar web log files extracted from a specific P2P site, and identifying information about customer segmentation strategies. Major web mining technique we adopted includes a machine learning like C5.0.

Keywords: web mining, web log, customer segmentation, machine learning, C5.0 Algorithm

1. 서 론

최근의 인터넷에 관한 연구는 인터넷상에서의 소비자 행위 연구에 대한 비중이 점점 커지고 있다. 그 이유는 인터넷이 제공하는 환경 자체가 전통적인 환경과 매우 다르기 때문에 고객의 행위를 효과적으로 파악할 필요가 더욱 커지고 있기 때문이다. 특히, 인터넷 사이트들이 개인화되고 맞춤화된 정보를 제공하기 위해서는 웹 사이트 고객의 행위를 보다 구체적이고 전략적으로 연구할 필요가 있다. 또한 전자상거래의 단점을 보완, 높은 광고비를 지불하지 않고도 충성도 높은 네티즌을 회원으로 확보하기 위하여 커뮤니티를 운영하는 많은 기업들이 개인화에 대해 많은 관심을 보이고 있다. 하지만 성공적인 개인화를 이루기 위해서는 현 고객들에 대한 특성을 분석하고 이를 웹 마케팅에 반영하여야 한다. 이러한 배경 하에 최근에 급부상하고 있는 웹 로그 파일 분석(web log file analysis)은 매우 시기 적절한 분석 방법이라 말할 수 있다. 왜냐하면 일반적으로 웹 사이트를 방문한 사람들의 방문 데이터에는 누가 언제 무엇을 요청했고 무엇을 가져갔는지, 웹 서버에 얼마나 많은 사람이 왔는지, 어디에서 왔으며 무엇을 제일 좋아하는지, 무엇을 제일 싫어하는지, 가장 오래 보는 또는 가장 많이 보는 페이지는 무엇인지 등등의 방문자 기록이 정리된 자료가 제공되기 때문이다. 특히 웹 로그 파일 분석은 최근에 새로운 화두로 떠오른 데이터 마이닝(data mining) 개념과 결합되어 웹 마이닝(web mining)으로 불리기도 하는데, 그 방법론 자체가 기존의 통계적인 방법과 많이 다르기 때문에 새로운 지식의 발견이 가능하다. 더욱이 컴퓨터 사이언스 분야와 마케팅 분야, 통계 분야 등에서 많은 연구가 진행되어 오고 있으며 효과적으로 웹 마이닝을 할 수 있는 구체적인 알고리듬이 제시되고 있다. 이에 본 연구의 목적은, 첫째, P2P 사이트를 대상으로 웹 고객의 유형을 데이터 마이닝 기법을 통하여 세분화하고, 둘째, 세

본 논문은 2001년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2001-005-C20010)
†연락저자 : 이건창 교수, 110-745 서울시 종로구 명륜동 3가 53번지 성균관대학교 경영학부, Fax : 02-745-4566, e-mail : leekc@skku.ac.kr
2002년 7월 접수, 1차 수정 후 2002년 11월 게재 확정.

분화된 집단별로 특징을 분석하여 고객 세분화 전략을 구축하는 방법을 소개하고자 한다.

본 연구 결과는 최근에 각광받고 있는 CRM 전략(Customer Relationship Management Strategy)의 한 유형으로, 웹 로그를 이용하여 어떠한 식으로 고객 세분화가 구체적으로 가능한지 제시할 수 있는 특징을 가지고 있다.

본 연구의 구성을 살펴보면, 2장에서는 웹 로그에 대한 기본 개념 및 웹 마이닝의 관점에서 고객 유형 세분화에 대한 기존 연구를 고찰한다. 3장에서는 본 연구의 연구 대상 및 전처리 과정을 제시하고, 4장에서는 구체적인 웹 마이닝을 통하여 고객 유형을 세분화하고 이 결과를 바탕으로 전략을 제시한다. 마지막으로 5장에서는 결론 및 향후 연구 방향에 대해 언급한다.

2. 기존 연구

2.1 웹 로그에 대한 이해

웹 마이닝 과정을 이해하기 위해서는 웹 로그에 대한 선행 이해가 필수적이다. 웹 로그란 사용자가 웹 사이트를 이용할 때 남는 '로그' 라는 형태의 기록을 의미하며, 이를 축적하여 분석하는 것을 웹 로그 분석 혹은 웹 마이닝이라고 한다. 웹 마이닝은 분석 목적에 따라 단지 로그 정보를 간략히 분석하는 수준으로 한정시키는 경우도 있고, 다양한 분석기법을 동원하여 로그 안에 숨어 있는 다양한 정보를 파악하는 확장된 개념으로 파악하기도 한다.

일반적 의미의 로그 분석은 로그 데이터를 이용하여 트래픽(traffic)을 파악하고, 이 트래픽이 지닌 의미를 분석해 나가는 것이라고 할 수 있다. 로그 데이터를 이용하여 웹 사이트의 페이지뷰(pageview), 사용자별 페이지뷰, 접속 장소 및 방식, 시간별 페이지뷰, 방문자 수 등에 대한 현황 및 추세를 분석하는 것이다. 웹 사이트의 클릭 흐름을 분석하는 것 역시 이 범주에 들어간다. 사용자가 웹 사이트를 방문하는 경로와 서핑(surfing)하는 경로에 대한 분석을 통하여 웹 사이트가 지닌 문제점을 찾고, 사용자가 웹 사이트에서 무엇을 원하는지 보다 구체적으로 파악하는 것이다. 그러나 이러한 분석은 웹 로그에 대한 1차원적인 정보 외에는 추출이 불가능하기 때문에 웹 로그가 가지고 있는 최소한의 정보밖에는 추출할 수가 없다. 이에 반해, 확장된 의미의 로그 분석은 단지 로그 데이터뿐 아니라 웹 사이트에서 보유하고 있는 고객 등록 정보, 구매 정보, 외부환경 정보 등을 복합적으로 사용하는 분석을 말한다. 이러한 분석을 통하여 사용자 특성별로 웹 사이트의 이용, 구매에 대한 보다 폭넓은 분석이 가능하다. 본 연구에서 언급하는 웹 마이닝 개념은 후자에 속하며, 특히 웹 사이트를 이용하는 고객들의 인구통계적인 특성을 바탕으로 고객 세분화하는 것을 주요 목표로 하고 있다. 이러한 웹 로그 파일에 대한 이해를 돕기 위하여 일반적인 웹 로그 파일의 예를 〈그림 1〉에 소개하였다(※ 본 논문에서 제시하고 있는 로그 파일은 마이크로 소프트사의 IIS 서버의 경우이다. Apache 서버인 경우에는 로그 파일의 형태가 약간 다르다).

설정된 항목을 바탕으로 로그 파일의 내용을 설명하면 다음과 같다.

1) 날짜(date) : 활동이 발생한 날짜를 의미한다.
2) 시간(time) : 활동이 발생한 시간을 의미한다.
3) 사용자 IP 주소(ip) : 웹 서버에 접근한 IP 사용자의 주소이다. IP와 사용자를 일대일로 매치시킬 수는 없지만(PC 통신과 같은 proxy 서버를 이용하여 접근한 경우 IP가 동일하기 때문임), IP만 잘 분석해도 많은 결과를 얻을 수 있다.
4) 사용자 이름(username) : 사용자의 이름을 나타내는데, 일반적으로 로그 파일에 남지 않는다.
5) 접근 방식(method) : 클라이언트가 시도한 작업으로, 예를 들

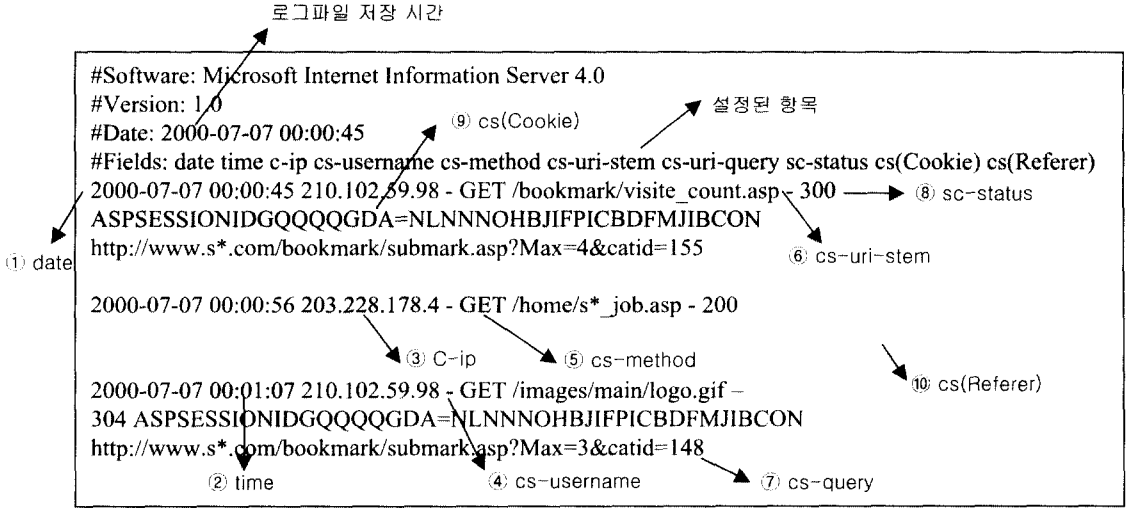


그림 1. 웹 로그 파일의 예.

면 GET 방식과 POST 방식 등이 있다. GET 방식은 일반적인 하이퍼 링크를 선택한 경우가 해당되며, POST 방식은 ID나 패스워드를 입력하고 '확인' 단추를 눌렀을 때 발생하는 방식이다.

6) URI 스템(uri-stem) : 사용자가 요청한 페이지에 대한 모든 내용이 담겨 있다. 예를 들어 사용자가 default.asp를 열었을 때 default.asp를 이루고 있는 default.asp, 이미지 파일(gif, jpg 등), Include 파일(css, js 등), 자바 애플릿(cg), CGI 파일 등의 이름이 모두 기록된다. 그러나 이미지라든지 Include 파일, 자바 애플릿과 같은 파일은 해당 페이지 안에 있다는 이유로 모두 기록되므로 실제 분석시에는 삭제하는 것이 일반적이다.
7) URI 쿼리(uri-queri) : 사용자가 시도한 질의가 남는다. 즉, 해당 파일에서 필요한 파라미터가 남는 부분이다. 예를 들어 위의 예제에서는 submark.asp를 실행시키기 위해 Max=3&catid=8 이라는 추가적인 부분이 기록되어 있음을 알 수 있다. 대부분의 경우 간과하기 쉽지만 URI Stem을 제대로 파악하기 위해 꼭 필요한 항목이다.
8) HTTP 상태(status) : http 관점에서 본 작업 상태로, 성공 또는 에러를 나타낸다.
9) 쿠키(cookie) : 보내거나 받은 쿠키 정보를 의미한다. 쿠키는 웹 사이트에서 사용자의 하드디스크에 집어넣는 특별한 텍스트 파일로, 웹 사이트와 사용자의 컴퓨터를 매개해주는 정보를 담고 있는 소량의 파일을 말한다.
10) 참조 페이지(uri-referer) : 사용자가 현재 페이지로 오기 전의 바로 전 페이지를 의미한다. 즉, 검색 엔진에서 사이트를 찾아 현재의 사이트로 이동했다면 검색 엔진에 나타난 페이지가 참조 페이지가 되고, 현재 페이지는 스템(stem)이 된다. 이 정보는 특히 배너 광고 효과 등을 파악하는 데 요긴하게 사용된다.

2.2 웹 마이닝에 대한 기존 연구

기존의 웹 마이닝에 대한 연구는 〈그림 2〉와 같이 웹 마이닝을 웹 콘텐츠 마이닝, 웹 구조 분석 연구 그리고 웹 사용자 마이닝의 세 가지 부류로 나누어 연구를 하고 있다(Han & Kamber, 2000).

본 연구는 웹 사용자 마이닝에 해당하는데, 이러한 웹 마이닝의 연구 분야는 어떠한 알고리듬을 써서 분석을 하느냐에 따라 다시 군집 및 분류 규칙(Clustering & Classification Rules), 연관 규칙(association rules) 분석, 순차적인 패턴(sequential patterns) 등 세 부분으로 나누어 생각해 볼 수 있다. 이 중에서 특히, 웹 로그가 일종의 문자열이라는 측면에서 기존의 기계학습(Machine Learning) 방법의 일종인 C5.0과 연관 규칙을 파악하는 Apriori 기법이 매우 활발하게 사용되고 있다(Agrawal & Srikant, 1994; Park *et al*, 1995; Savasere *et al*., 1995).

군집 및 분류 규칙의 대표적인 예인 C5.0은 Quinlan(1986)이 ID3에 이어 개발한 귀납적 학습 방법의 하나이다. C5.0 이전에 귀납적 학습 방법의 예로는 기계학습(machine learning)이라는 이름 하에 발전해온 CLS, ACLS, ID3 등과 통계학에 기반을 둔 CART(Classification And Regression Tree), CHAID(Chi-square Automatic Interaction Detection) 등이 있다(Berry & Linoff, 2000). C5.0 에서는 학습 자료들을 분류하기 위해 사전에 정의된 등급(class)과 속성(property) 간의 관계를 파악하여 단계적으로 의사결정 트리(decsion tree)를 형성한다. 특히, C5.0은 기존의 ID3가 가지고 있었던 문제점인 의사결정 트리 생성에서 사용되는 gain criterion이 가지는 편향성 문제와 숫자형 변수를 다룰 수 없는 문제점을 개선한 것이다(이상호 & 지원철, 1998). 한편, C5.0은 생성된 의사결정 트리가 지나치게 많은 단계와 리프(leap) 노드를 가질 경우에 학습된 의사결정 트리의 일반화 능력을 제고하기 위하여 리프 노드를 제거하는 방법인 프루닝(pruning)을 시행한다. 일반적으로 프루닝을 통하여 예측력은 향상될 수 있으나 오류율도 증가하기 때문에 C5.0 에서는 오류 기반 프루닝을 통하여 오류율의 증가를 통제한다. 또한, C5.0 에서는 의사결정 트리가 해석하기 난해하다는 점을 해결하기 위하여 자동적으로 if-then 규칙을 생성해 주는 특징이 있다.

C5.0과 더불어 웹 마이닝에서 많이 사용되는 기법은 연관 규칙이다. 이 분야는 대규모의 슈퍼마켓 등에서 판매한 물품들의 상호 연관성을 찾아내어 고객이 미리 정한 지지도와 신뢰도를 바탕으로 연관 규칙을 찾아내는 것이다. 예를 들어 기저귀를 사는 사람이 맥주를 동시에 구매하는지를 알고자 한다고

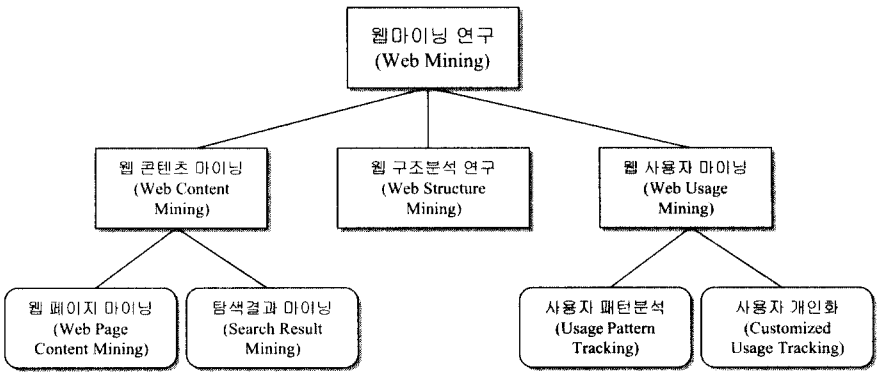


그림 2. 웹 마이닝에 대한 응용분야 연구.

가정하자. 이를 연관 규칙을 이용하여 일정한 규칙을 추출해 보니 '기저귀-맥주'(지지도=15%, 신뢰도=75%)로 분석되었다. 여기서 지지도 15%의 의미는 전체 고객 중에서 15%가 기저귀와 맥주를 동시에 구매한다는 것이고, 신뢰도 75%의 의미는 기저귀를 구매한 고객 중에서 75%만이 맥주를 구매한다는 것이다. 지지도는 전체 구매 고객 중에서 연관 규칙에 관련된 항목들을 구매하는 정도를 나타내고, 신뢰도는 특정 물품과 물품 사이의 연관성을 나타내므로, 먼저 최소한 주어진 지지도 이상의 물품들을 찾아내고 그 다음으로 물품들 사이의 신뢰도를 측정하게 된다. 슈퍼마켓 등에서 판매된 물품들의 판매 트랜잭션으로 구성된 데이터베이스를 '바스켓 데이터베이스(basket-database)'라 부르고, 대부분의 연관 규칙 탐사 알고리듬이 이 분야에서 연구되었다(Srikant & Agrawal, 1995). 연관 규칙을 탐사하는 방법은 크게 두 가지 부분으로 나누어진다. 첫째, 최소 지지도 이상으로 발생하는 항목을 찾아내는 방법. 둘째, 이들 사이의 신뢰도를 측정하여 기준점(Threshold) 이상의 값을 보일 때만 연관 규칙을 만들어내는 것이다. 이를 효율적으로 해결하고자 하는 연구가 Apriori 알고리듬(Agrawal & Srikant, 1994; Agrawal *et al.*, 1993a; 1993b)이다.

끝으로 순차적인 패턴이란 일정 시간동안 시간에 따라 순차적으로 발생하는 거래를 말한다(Srivastava *et al.*, 1999). 순차적인 패턴을 인식하는 알고리듬에 대한 연구는 비교적 다양한 방면으로 이루어져 왔다. 대표적인 연구자로서 grawal & Srikant(1994)는 순차적인 고객의 거래 자료에서 일정한 패턴을 인식하는 알고리듬을 제시하였으며, 이후 Mannila & Meek(2000)는 웹에서의 사용자들의 거래 자료를 순차적인 패턴(Sequential patterns)으로 인식하는 방법론을 개발하였다. 또한 Mannila *et al.*(1997) 등은 순차적으로 발생하는 일련의 사건을 부분적인 집합으로 함께 묶어 에피소드(episodes)라 새롭게 정의하고 이를 활용하여 일정한 패턴을 분석하고자 하였다. 이들은 에피소드를 기본으로 연관 규칙을 도출하는 것이 기존의 다른 방법보다 그 효율성이 우수함을 증명하였다.

이러한 알고리듬을 적용한 구체적인 연구 사례를 살펴보면, Igor(2000) 등은 웹에서 사용자들의 거래 자료 형태가 관계형 데이터베이스 형태로 일정하게 고정되어 있지 않은 경우 이를 처리할 수 있는 확률 기반의 군집 프레임워크를 제안하였다. 이들은 이를 통하여 고객들의 웹 사이트 방문 형태를 다양하게 분류하였을 뿐만 아니라 실제 혈액 분류 사례에 적용하여 그 우수성을 검증하였다. 또한 Fu(1999) 등은 웹 사용자의 접속 패턴을 군집화 하고 이를 통하여 일정한 규칙을 제공하는 방법론을 개발하였으며, McCallum(2000) 등은 사용자들의 웹 데이터를 다차원의 데이터 집합으로 관리하고 있는 경우 이를 효율적으로 군집화 할 수 있는 방법론을 제시하였다.

본 연구에서는 웹 마이닝을 위해 연관 분석과 순차적 패턴 보다는 기계학습 방법인 C5.0을 주로 많이 사용하고 있다. 그 이유는 특정 웹 페이지 간의 연관 관계에 관심이 있는 것이 아니라 특정 페이지 또는 특정 서비스를 이용하는 고객의 특성을 파악하여 이들 고객을 특성별로 세분화하는데 관심이 있기 때문이다. 또한, C5.0은 해당 규칙을 If-Then 규칙의 형태로 제공하기 때문에 일반적인 연관 규칙보다는 해석이 용이하다는 장점이 있다. 그러나 C5.0이 정확한 패턴 매칭을 요구하고 있기 때문에 실제 사례가 없는 노드를 작성하거나 도출된 규칙이 매우 크거나 지나치게 세분화 될 가능성이 있다. 따라서 도출된 규칙에 대해서는 충분한 검토를 한 후 고객 세분화의 판단 기준으로 삼아야 한다.

2.3 웹 마이닝 솔루션 현황

현재 국내외에는 많은 웹 마이닝 솔루션이 등장하고 있다. 그러나 대부분의 솔루션이 개발 초기 단계이고, 아직까지 많이 사용되지 않아 구체적인 기능은 널리 알려져 있지 않다. 현재 국내에 알려진 웹 마이닝 상용 솔루션은 크게 두 가지 측면으로 나누어 생각해 볼 수 있는데, 하나는 웹 로그 자료를 기술통계 분석을 하여 방문 빈도, 자주 노출되는 사이트, 고객들의 평균 이용시간 등을 일목요연하게 정리하여 보여주는 것이다. 다른 하나는 본 연구와 같이 솔루션 안에 데이터 마이닝 알고리듬을 포함하고 있어 전 처리한 자료에 대해 분석을 실시하고 이로부터 유의한 결론을 도출하는 것이다. 현재 개발된 국내의 솔루션들은 대개 전자에 해당하는 것이 많은데, Web Log(www.weblog.com), ClickAnalyzer:(www.weblog.com), ClickCast (www.netrix21.co.kr) 등이 여기에 해당한다. 후자에 대한 것들은 대학 연구소와 합동으로 운영하는 기업에서 시제품 형태로 나온 것들이 많이 있는데, WebTrends Log Analyzer (webtrends.greenmart.co.kr), WiseItech(www.wise.co.kr)의 iTransform과 iPersonalizer가 여기에 해당하며, 숭실대학교가 참여하고 있는 ECminer(www.ecminer.com)의 Web Analyzer도 후자에 해당한다.

3. 연구 대상 선정 및 전처리

본 연구의 목적은 웹 고객의 로그 파일을 중심으로 고객 유형을 몇 개의 집단으로 세분화하고 이들 특성을 분석하여 고객의 유형에 따른 고객 세분화 전략을 제안하는 것이다. 이를 위하여 첫째, 분석 대상을 선정하였으며, 둘째, 이 분석 대상에서 고객들의 정보가 포함된 웹 로그를 추출하였다. 셋째, 전처리 작업을 통하여 웹 로그를 분석이 가능한 데이터의 형태로 만들었다.

3.1 분석 대상의 선정

본 연구에서는 고객 세분화 전략을 파악하기 위해 P2P 서비스를 제공하는 E 사이트를 분석 대상으로 선정하였다. E 사이트는 개인, 지식, 비즈니스 포탈형 비즈니스 모델을 제공하는 P2P 기반의 지식 포탈 사이트이다. 이 사이트에서는 현재 각 분야별 전문가 2천여명을 확보해 놓고 이들을 통해 국내뿐 아

니라 국제적 P2P 기반의 지식 커뮤니티(knowledge community)를 구축하고 있다. 특히, 웹 사이트를 효과적으로 분석하기 위해서는 E 사이트의 특성을 분석하는 것이 중요하다. E 사이트의 가장 큰 특징은 P2P 소프트웨어를 제공하고 있는 것이다. 이는 전 세계의 수많은 컴퓨터를 직접 연결함으로써 거대한 지식/정보 저장 공간을 형성하고, 고객의 익명성을 보장함으로써 지식/정보의 무한 공유 환경을 제공하는 서비스의 일종이다. 또한 고객 세분화의 일종인 마이웹 기능을 제공하여 회원에 가입하면 자신만의 정보를 관리할 수 있는 서비스가 제공된다. 그리고 커뮤니티에서는 고객들이 원하는 커뮤니티를 만들 수 있으며 이를 통하여 정보 및 지식을 공유하고 있다. 마지막으로 '지식정보마당'에서는 각 커뮤니티 자료, 웹 폴더 공개자료, 및 인터넷 사이트의 모든 자료를 찾아주고, 빠르고 쉽게 지식 정보에 접근할 수 있도록 '지식탐색 트리'를 제공한다.

본 연구에서 사용된 웹 로그는 1주일 동안의 자료로서 총 레코드 수는 5,200건이며, 이때 고객 수는 240명이다. 그리고 E 사이트의 웹 로그 필드는 크게 고객 정보(user infomation), 서비스(service), 커뮤니티(community), 콘텐츠(contents), 방문시간(duration), 사이트(url) 등으로 구성되어 있다.

3.2 웹 로그의 전처리

웹 로그 파일은 그 수가 많으나 실질적으로 분석 대상이 되는 정보를 가지고 있는 레코드 수는 소수에 불가하다. 또한, 분석 목적에 따라 로그 파일 중에서 원하는 항목을 선택하여 분석하여야 한다. 따라서 데이터베이스에 저장되어 있는 웹 로그 파일에서 정보를 가져다주지 못하는 로그는 제거되고 의미 있는 로그만을 전처리하여 접속자별로 로그를 배열한다. 이때 웹 로그에서 발생한 자료 중에서 이미지 파일(gif, jpg 등), include 파일(css, js 등), 자바 애플릿(cg), CGI 파일 등은 아무 의미가 없기 때문에 삭제해야 한다. 배열된 로그는 분석에 용이한 데이터로 전환하는데, 이때 집합(aggregate), 여과(filtering), 통합(merge) 기법을 이용하게 된다. 분석하고자 하는 데이터를 결정하였으면 이를 활용하여 데이터베이스에서 패턴과 규칙을 도출하기 위하여 다양한 분석 방법을 통하여 분석을 실시하게 된다. 이때 연관 규칙 분석, 순차적 패턴 분석, 군집 및 분류 분석이 이용된다. 마지막으로 분석된 결과에서 의사결정에 도움이 되는 지식을 획득하게 되는 것이다.

본 연구에서는 이렇게 전처리가 완료된 로그 파일로부터 고객의 특성을 추출하기 위하여 기계학습의 일종인 C5.0 방법을 사용하였다. 본 연구에서 CHAID나 CART보다는 C5.0을 이용한 이유는 기계학습의 경우 자료의 분포에 대한 가정을 하지 않고 엔트로피(entropy)를 이용하여 자료를 분류하는 기준이 해석에 더욱 용이할 것으로 판단하였기 때문이다.

4. 웹 마이닝에 의한 고객 세분화

본 연구에서는 웹 로그에서 고객의 유형을 세분화하기 위하여 주어진 웹 사이트를 서비스별, 커뮤니티별로 세분화하고 이에 따른 고객의 인구통계적인 특성을 파악하여 고객 세분화 전략을 구축하였다.

4.1 제공 서비스별 고객 세분화

현재 E 사이트에서 제공하는 전체 서비스 수는 총 80개이고, 이 중에서 주로 사이트 등록 및 삭제와 관련된 '파일 서비스'가 27개, 지식 및 정보를 제공하거나 최신 용어를 검색하는 '지식 서비스'가 18개, 커뮤니티를 형성하거나 이 속에서 게시판을 이용하는 '커뮤니티 서비스'가 35개로 구성되어 있다. 하지만 분석 결과 세분화 대상 고객이 실제 사용한 서비스는 총 24개 뿐으로 파일 관련 서비스 4개, 지식 관련 서비스 4개, 커뮤니티 관련 서비스 16개로 분석되었다. 따라서 실제 제공하는 서비스의 25% 정도만 사용되고 있으며, 파일 관련 서비스나 지식 관련 서비스보다는 커뮤니티 관련 서비스가 많이 활용되고 있는 것으로 분석되었다. 따라서 본 연구에서는 커뮤니티 서비스를 한번 이용한 고객을 0으로 가정하고, 두 번 이상 이용한 고객을 1로 가정하여 C5.0 분석을 실시하였다. 분석 결과를 〈그림

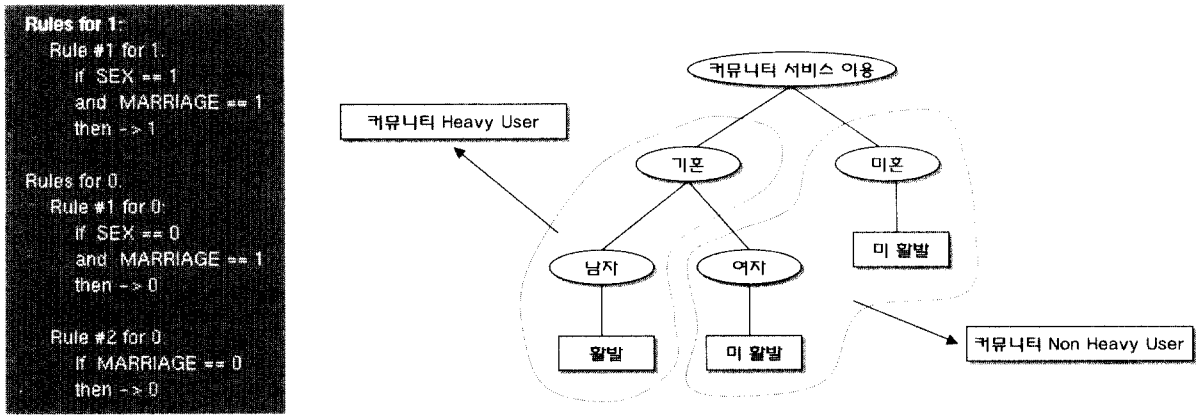


그림 3. C5.0 분석 결과 및 분류 그림.

3〉 에 제시하였다.

분석 결과 성별이 여자(0)이고 결혼한 사람(1)은 커뮤니티 서비스를 많이 사용하지 않았으며, 결혼을 하지 않은 사람도 커뮤니티 서비스를 많이 사용하지 않았다. 또한 성별이 남자이면서 결혼한 사람들이 커뮤니티 서비스를 많이 사용함을 알 수 있다. 따라서 고객들이 빈번하게 사용하는 커뮤니티 서비스만을 분석한 결과, 기혼이면서 남자인 경우는 커뮤니티 주요(Heavy) 고객으로 볼 수 있고, 미혼인 경우는 모두 비주요 고객이며, 여자인 경우는 결혼 여부와 상관없이 비주요 고객으로 분석되었다. 이것은 현재 운영중인 커뮤니티의 특성이 기혼인 남자 위주로 구성되어 있을 가능성이 높기 때문에 E 사이트에서 보다 다양한 커뮤니티가 운영될 수 있도록 하면 고객이 전반적으로 많아지고 이용 빈도도 더 높아질 것으로 분석된다.

또한 커뮤니티를 많이 이용하는 고객들에 대한 인구통계적 특성을 파악하기 위하여 직업을 종속변수로 두고 C5.0 분석을 실시하였다. 분석 결과가 〈그림 4〉 에 제시되어 있다.

분석 결과 흥미로운 것은 커뮤니티 사용 고객 중 결혼한 41세 이하의 고객 중에는 컴퓨터 관련 직종이 많고, 41세 이상은 세일즈/마케팅이 많다는 것이다. 여자의 경우에는 27세 이하에서 직업을 가지고 있으며, 특히 22세 이하는 학생이 많았다. 그리고 전체적으로 연령이 많을수록 컴퓨터 관련 직종이 많아, E 사이트의 주요 고객 중에는 컴퓨터 관련 업종에 근무하는 사람이 많음을 알 수 있다. 따라서 직업별로 구분된 커뮤니티 서비스를 빈번히 이용하는 고객에 대하여 그룹 1은 전문직/컴퓨터 공학(기혼 41세 이하, 미혼 25세 이상 여성), 그룹 2는 대학생/대학원생(남자 27세 미만, 여자 22세 미만), 그룹 3은 세일즈/마케팅 & 서비스/고객지원(남자 41세 이상, 여자 25세 이하)의 세 그룹으로 세분화가 가능하다. 이들 그룹의 고객은 상호 유사한 직업을 가지고 있고 커뮤니티 서비스를 활발히 이용하기 때문에 이들에게 공통적인 마케팅 서비스 제공이 필요하다.

4.2 커뮤니티별 고객 세분화

E 사이트에서는 117개의 커뮤니티를 13개의 카테고리로 나누어 활동하고 있으나 카테고리명과 활동하고 있는 커뮤니티의 특성이 상이한 것이 있어, 이를 10개 그룹으로 재분류하여 분석 대상을 확정하였다. 〈표 1〉 에는 확정되어진 커뮤니티에 대한 카테고리 및 비율이 제시되어 있다.

분석 대상인 238명은 한 개의 커뮤니티에도 가입하지 않은 고객부터 최대 7개의 커뮤니티에 가입한 고객으로 구성되어 있다. 따라서 고객 세분화 대상을 선정하기 위하여 커뮤니티에 한 번 이상 가입하고 있는 77명을 분석 대상으로 하였으며, 이들이 가입하고 있는 커뮤니티의 수는 총 53개이다. 세부적으로 살펴보면 P2P 비즈니스 모델 연구회(18명), SCJP 자격증(11명), 엽기동(9명), 그램쟁이/웹천재(8명), 웃음이 있는 곳(12명)이 활발히 참여되고 있는 커뮤니티로 분석되었다. 본 연구에서는 커뮤니티별 고객을 세분화하기 위한 분석 방법은 다음과 같다.

첫째, 성별, 나이, 직업 등의 인구 통계학적인 특성을 가입한 커뮤니티의 카테고리와 연결하여 C5.0을 이용하여 특징을 추출하고자 한다. 이때 인구 통계학적인 특징을 입력변수로 하고 가입한 커뮤니티를 출력변수로 하면 어떠한 인구 통계학적인 특성을 가진 고객이 특정 커뮤니티에 가입하는지 파악이 가능하다.

둘째, 역으로 커뮤니티를 입력변수로 하고 인구 통계학적인

표 1. 커뮤니티에 대한 카테고리 및 비율

커뮤니티 명	빈도	비율
컴퓨터와 인터넷	39	37.9%
스포츠	22	21.4%
비즈니스	21	20.4%
e-Biz	7	6.8%
사회문화	5	4.9%
지식경영	4	3.9%
기타	3	2.9%
자연과학	1	1.0%
엔터테인먼트	1	1.0%

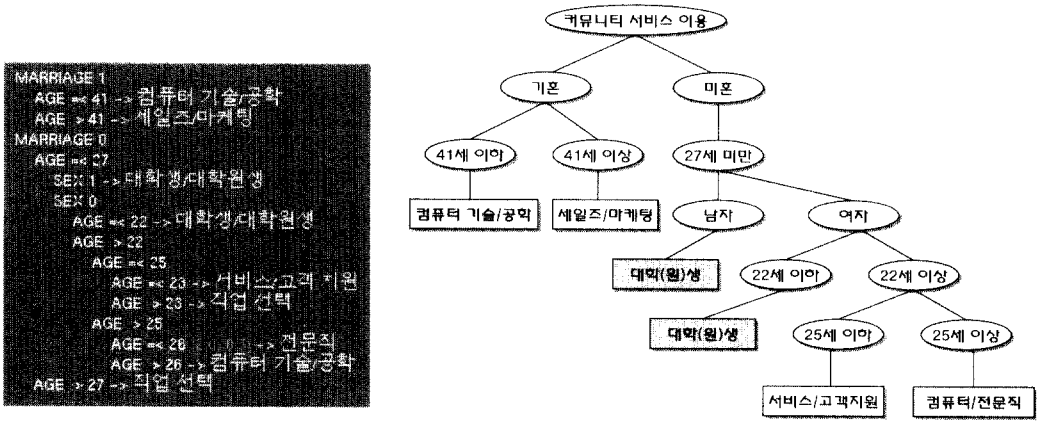


그림 4. 인구통계적 특성 분석 결과 및 의사결정 트리.

특성을 출력변수로 분석을 실시하여 해당 커뮤니티를 구성하고 있는 고객의 인구 통계학적인 특성을 파악한다.

따라서 두 방법을 바탕으로 커뮤니티별 고객 세분화를 실시하였다. 먼저, C5.0을 이용하여 특징을 추출하기 위해서는 커뮤니티를 카테고리화 해야 하는데, 〈표 2〉와 같이 카테고리화하였다.

〈그림 5〉에는 C5.0을 사용하여 인구 통계학적인 특징을 입력변수로 하고 가입한 커뮤니티를 출력변수로 하여 분석한 결과와(a), 커뮤니티를 입력변수로 하고 인구 통계학적인 특성을 출력변수로 분석을 실시한 결과(b)가 제시되어 있다.

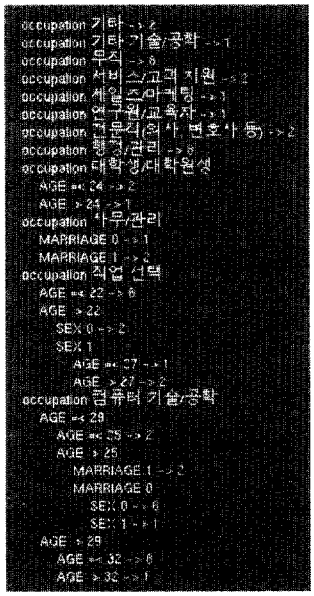
〈그림 5〉 (a)의 분석 결과를 살펴보면 'e-Biz 커뮤니티'에 가입하고 있는 고객의 경우, 직업이 기술/공학, 세일즈/마케팅, 연구원/교육자 등 다양하게 구성되어 있다. 또한 전문직이며 24세 이하의 대학생, 27세 이상의 기혼 남자, 직업이 컴퓨터와 관련된 25세 미만, 기혼인 경우 25세 이상 등이 컴퓨터와 인터넷 관련 커뮤니티에 많이 가입하고 있는 것으로 분석되었다. 반면에 22세 미만이거나 25세 이상의 미혼 여성의 경우에는 스포츠 관련 커뮤니티에 많이 가입하고 있다. 특히 직업이 컴퓨터/공학 계통인 경우이고 32세 이상이면 거의 e-Biz 관련 커뮤니티에 가입하고 있는 것으로 분석되었다.

〈그림 5〉 (b)의 분석 결과를 살펴보면, 'e-Biz' 커뮤니티 계열(1)인 경우에는 고객의 특성은 28세 미만이며, 여성의 경우 연구원/교육자이고, 남자인 경우 26세 미만의 학생이 많은 것으로 분석되었다. 또한, 28~30세인 경우는 기술/공학이 많으며, 30세 이상인 경우는 세일즈/마케팅이 많은 것으로 분석되었다. 가입한 커뮤니티가 컴퓨터 & 인터넷(2)인 경우에 고객의 특성을 살펴보면, 여자인 경우는 24세 미만의 서비스/고객지원에 종사하고 있으며, 남자인 경우는 학생이고, 24세 이상인 경우는 남녀 공히 컴퓨터 관련 직종에 많은 것으로 분석되었다. 가입한 커뮤니티가 지식경영(3)인 경우는 다른 커뮤니티보다는 직업이 사무/관리인 경우가 많은 것으로 분석되었다. 이는 현재 이루어지고 있는 공공기업의 지식경영의 영향에 의하여 지식경영 관련 정보 및 지식을 원하는 사무 및 관리직인 고객의 수가 높기 때문이라고 사료된다.

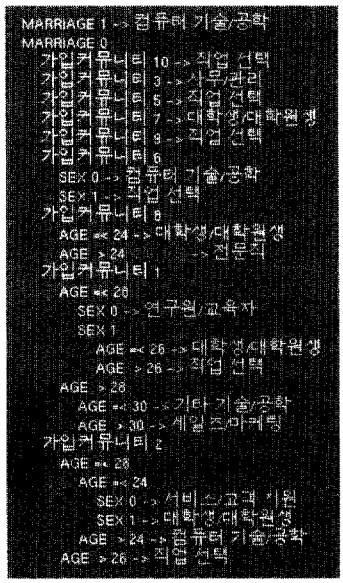
가입한 커뮤니티가 자연과학(4)인 경우는 한 개 사례가 있었으므로 지지도(support) 및 신뢰도(confidence)가 낮은 관계로 규칙에서는 제외하였다. 가입한 커뮤니티가 엔터테인먼트(5)인 경우는 직업이 특별히 분류되지 않았다. 가입한 커뮤니티가 스포츠(6)인 경우 여자 고객은 직업이 컴퓨터 기술/공학인 경우가 많았으며, 남자는 특정한 직업을 분류하기 어려웠다. 가입한 커뮤니티가 사회문화(7)인 경우는 전체적으로 대학생, 대학원생들이 많은 것으로 분석되었다. 가입한 커뮤니티가 비즈

표 2. 커뮤니티별 카테고리 분류 결과

커뮤니티 명	번호
e-Biz	1
컴퓨터와 인터넷	2
지식경영	3
자연과학	4
엔터테인먼트	5
스포츠	6
사회문화	7
비즈니스	8
기타	9



(a) 입력변수-인구 통계학적인 특성
출력변수-커뮤니티의 카테고리



(b) 입력변수-커뮤니티의 카테고리
출력변수-인구 통계학적인 특성

그림 5. 인구통계 및 커뮤니티 카테고리별 분석 결과.

니스인(8) 경우는 나이 24세 이하는 학생이 많았으며, 24세 초과인 경우는 전문직이 많았다.

이상의 분석 결과에 의해 세 개의 그룹으로 나눌 수 있다. 그룹 1은 커뮤니티가 e-Biz(1), 컴퓨터 & 인터넷(2), 스포츠(6), 비즈니스인(8)에 가입한 고객으로서 다양한 인구통계학적인 특성을 가진 고객으로 구성되어 있다. 따라서 이 그룹은 많은 사람들이 관심을 갖고 있는 부분이므로 연령별, 성별, 기혼 여부별로 보다 정교한 마케팅 전략을 구사해야 한다.

그룹 2는 나머지 커뮤니티에 가입한 고객들로, 구성되어 있는 고객들의 인구통계학적인 특성이 단순하다. 따라서 이들에 대해서는 추가적인 자료를 확보하여 전략을 제시하거나 그룹 2의 커뮤니티 활동에 변화를 제시하여 이들의 움직임을 분석할 필요가 있다.

그룹 3은 커뮤니티에 가입하지 않는 고객들로, 이탈 가능성이 매우 높다고 분석할 수 있다. 왜냐하면 E 사이트의 경우 커뮤니티를 중심으로 개인의 정보 및 지식을 주고받는 P2P 사이트이므로 커뮤니티에 가입하고 있지 않다는 것은 단순히 회원으로서 등록만 했다는 것이지 주요 고객으로 활동하고 있지 못하다는 뜻이다. 따라서 이러한 고객들은 이탈 가능성이 높은 것으로 사료된다.

이에 전체 고객 중에서 E 사이트를 계속적으로 사용하는 고객과 이탈하고자 하는 고객들을 분류할 필요가 있다. 전체 분석 대상인 238명의 고객을 대상으로 커뮤니티에 가입한 고객(1)과 커뮤니티에 가입하지 않는 고객(0)을 두 그룹으로 나누어 C5.0을 활용하여 규칙을 도출하고 이들의 특성 의사결정 트리로 비교 분석한 결과가 〈그림 6〉에 제시되어 있다.

〈그림 6〉을 살펴보면, 직종이 서비스/고객지원이며 남자인 고객은 이탈 가능성이 높은 것으로 분석되었으며, 컴퓨터 기술/공학 관련 직종에 종사하며 미혼자의 경우 이탈 가능성이 높은 것으로 분석되었다. 그밖에도 기타 기술/공학, 무역직/기능공, 무직, 연구원/교육자, 자영업, 주부, 초등학생 등은 회원으로 가입되어 있으나 그 사용 빈도가 적거나 관심이 적은 이탈 가능성이 높은 그룹으로 분류되었다.

5. 결론 및 향후 연구 방향

최근 인터넷에서 고객 세분화에 대한 연구가 드높아짐에 따라 어떠한 방법으로 고객을 세분화할 것인가에 대한 논의가 신중하게 이루어지고 있다. 본 연구에서는 최근에 각광받고 있는 신 기법인 웹 마이닝 기법을 이용하여 이러한 고객 세분화를 시도하였다. 이를 위하여 웹 로그에 대하여 정의를 하였으며 현재 웹 로그 분석이 이루어지고 있는 업계의 동향을 조사하였다. 또한 웹 마이닝에 대한 기존문헌 연구를 통하여 웹 마이닝이 이루어지고 있는 연구의 흐름을 분석하였다.

P2P 실제 사이트의 웹 로그를 이용하여 분석한 결과 C5.0과 같은 웹 마이닝 기법은 주어진 웹 로그와 고객정보로부터 고객을 세분화할 수 있으며 그 결과 역시 기존의 웹 사이트 운영자들이 전혀 알고 있지 못하던 새롭고 유익한 정보로 판단되었다.

본 연구의 공헌점은 첫째, 기존에 많이 이루어지고 있던 고객 세분화 방법론에서 흔히 사용하던 설문지 대신에 실제 자료를 사용함으로써 새로운 방법론을 제시하였다는 점이다. 둘째, 웹 로그를 이용하여 고객 유형을 세분화하였다는 점이다. 현재 인터넷 사용인구가 증가하면서 이들이 웹상에서 이루어지는 행동 패턴에 관심이 높아지고 있다. 따라서 본 연구에서 실제 웹 로그를 바탕으로 고객의 유형을 세분화하였다는 것은 매우 적절한 것으로 분석된다.

본 연구의 한계점은 첫째, 특정 사이트에 대한 분석에 많이 치우쳐 고객 세분화를 위한 일반적인 방법론이 다소 부족하다.

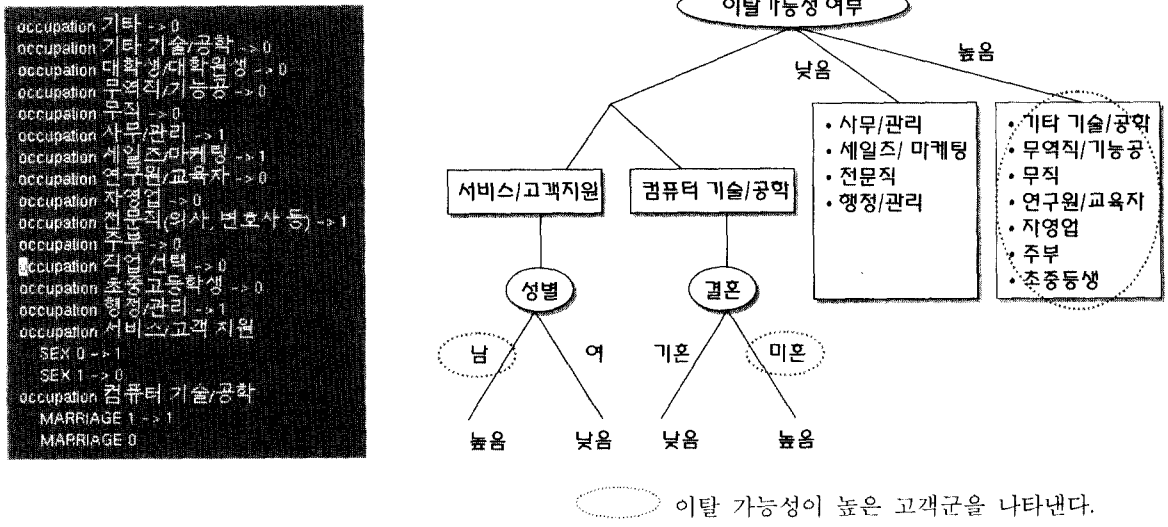


그림 6. E 사이트 사용 고객과 이탈 고객.

둘째, 본 연구에서는 웹 로그 파일의 수가 적어서 고객 유형을 세분화하고 이 결과를 일반화하기에는 무리가 있다. 셋째, 본 연구에서는 다양한 웹 마이닝 방법 중에서 C5.0을 주로 이용하였는데 연관 규칙이나 순차적 패턴 방법을 이용한다면 보다 더 다양한 규칙이 도출되었을 것으로 판단한다. 따라서 향후 연구 방향으로는 장기간의 웹 로그 자료를 수집하고 보다 고객 지향적인 일반적인 방법론의 개발이 요구된다고 하겠다.

참고문헌

이상호, 지원철 (1998), 귀납적 학습방법들의 분류성능 비교: 기업신용평가의 경우, *한국지능정보시스템학회논문지*, 4,(4),1-22.

Agrawal R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Dat abases, Proceedings of the 20th Inter national Conference on Very Large Databases,111-120.

Agrawal, R., Imielinski, T. and Swami, A. (1993a), Database Mining: A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering*, 5(6), 914-925.

Agrawal, R., Imielinski T., and Swami, A.(Washington, D.C., 1993b), Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD Conference on Management of Data, 207-216.

Berry, M.J.A., and Linoff, G. (1997), Data Mining Techniques: For Marketing, Sales, and Customer Support, Wiley Computer Publishing.

Fu, Y., Sandhu, K. and Shih, M.Y. (1999), Clustering of Web Users Based on Access Patterns, Proc. of the ACM SIGKDD international conference on Knowledge discovery and data mining, 95-104.

Han, J. and Kamber, M. (2000), Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

Igor V.C., Scott G. and Smyth, P. (2000), A general probabilistic framework for clustering individuals and objects, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 140-149.

Mannila, H. and Meek, C. (2000), Global partial orders from sequential data, Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining, 161-168.

Mannila, H., Toivonen, H. and Verkamo, I. (1997), Discovery of frequent episodes in event sequences, Department of Computer Science Series of Publications Report C-1997-15.

McCallum, A., Nigam, K. and Ungar, L.H. (2000), Efficient clustering of high-dimensional data sets with application to reference matching, Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178.

Quinlan, R. (1986), Induction of Decision Trees, Machine Learning, 1, 81-98.

Srikant R. and Agrawal, R. (1995), Mining Generalized Association Rules, Proceedings of the 21th International Conference on Very Large Databases, 407-419.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.N. (1999), Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining, 1-10

이 건 창

성균관대학교 경영학과 학사

한국과학기술원(KAIST) 경영과학과 경영정보시스템 석사

한국과학기술원(KAIST) 경영과학과 경영정보시스템 박사

현재: 성균관대학교 경영학부 교수

관심분야: 전자상거래, 퍼지인식도, 협상지원시스템, 지식경영, 인터넷 마케팅

정 남 호

경기대학교 경영정보학과 학사

성균관대학교 경영학과 경영정보시스템 석사

성균관대학교 경영학과 경영정보시스템 박사

현재: (주)크레듀 기획/컨설팅팀

관심분야: 인공지능기법을 이용한 의사결정, 지식경영, 퍼지 인식도, 지능형 에이전트, 인터넷 마케팅, 전자상거래, e-Learning