인터넷 비즈니스에서 효과적인 소비자 관계관리 (Customer Relationship Management)를 위한 데이터 마이닝 기법의 응용에 대한 연구*

김 충 영
서울시립대 경영학부 부교수
E-mail : cnkim27@uos.ac.kr

장 남 식
서울시립대 경영학부 조교수
E-mail : nchang@uos.ac.kr

김 상 욱
메타비경영연구원 선임연구원
E-mail : sukim@metab.co.kr

본 연구에서는 고객 세분화를 위하여 고객프로필과 사이트 접속자료를 통합, 분석하는 분석적 CRM을 시도하였다. 실제 고객 데이터를 분석하여 고객의 특성과 기호, 방문행태 등을 이해할 수 있다면 이를 기반으로 고객 세분화(segmentation)가 가능할 것이다. 예를 들어 고객의 거주지, 재산정도, 교육수준, 연령 등 인적정보를 토대로 동일 사이트에 접속하는 고객의 공통점을 찾게 된다면 이들 고객에 접근할 수 있는 적절한 마케팅 미디어가 무엇인지, 어느 페이지에 홍보물을 게재하는 것이 효과적일 것인가 등을 결정하는 데 도움을 줄 수 있을 것이다. 한편 웹 기반 마이닝의 핵심은 웹으로 부터의 자료를 어떻게 하면 효율적으로 수집할 것인가, 또한 이렇게 수집된 자료를 다양한 (multiple) DB와 어떻게 통합하고 분석하여 필요한 정보를 추출할 것인가 일 것이다. 본 연구에서는 실제 인터넷 사업자의 사용자 그룹의 비율에 따라 구성된 패널을 활용하여 효율적인 자료수집 방안을 모색하였다. 패널 구성원에 대한 웹 데이터를 수집함으로써 신뢰성과 대표성을 확보하면서 분석대상 자료의 양을 적절한 수준으로 유지할 수 있었다. 또한 고객자료 분석에서는 OLAP과 데이터 마이닝 기법(의사결정나무)을 동시에 사용하여 그 분석 결과를 비교함으로써 각 기법의 결과를 상호 확인하고 보완할 수 있었다. 이 결과는 데이터 마이닝 기법에 의해서 발견된 패턴을 분석하고 확인하는 작업에서 OLAP이 유용하게 사용될 수 있다는 과거 연구의 주장을 확인하였다.

* 이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음.(KRF-2000-041-C00329)

1. 서 론

(1) 고객관계관리

Global 환경에서 날로 심화되고 있는 기업간 경쟁 속에서 고객관계에 대한 적절한 이해와 관리가 더욱 중요하게 대두되고 있다. 고객관계관리(Customer Relationship Management: CRM)란 고객의 인구통계학적, 실적, 라이프스타일 등에 대한 데이터와 정보를 분석하여 고객의 요구를 파악하고 이를 마케팅, 서비스, 영업 등에 전략적으로 활용하는 프로세스이다. 즉, 개별고객의 특성에 맞는 마케팅을 통해 기존 고객의 유지(Retention)는 물론 새로운 시장기회를 개발하는 확장(Extension)의 개념을 기반으로 하는 것이다(Gartner Group, 2000). CRM 적용 작업의 예를 들면 소비자의 행태 분석이나 만족도 평가에 활용할 수 있으며 고객계층의 세분화와 특성분석, 고객의 수익성 평가 및 고객개발 등도 가능하다. 일반적으로, CRM은 1) 고객 데이터의 수집, 저장 및 분석을 하는 고객 분석단계, 2) 분석된 정보를 마케팅 계획에 연결하는 전략화 단계, 3) 다양한 고객접점을 통하여 마케팅을 수행하는 단계 등 세 단계로 나누어 볼 수 있는 데 이 중에서도 고객을 제대로 이해하는 기반을 마련하는 고객 분석단계가 CRM의 필수적인 작업이라 할 수 있다 (Berry & Linoff, 2000).

최근 정보기술의 발달과 인터넷 사용의 확산으로 대 고객접점이 다양해지고 고객과의 양방향 대화(Communications)가 가능해짐에 따라 기존 CRM 작업을 확대하여 웹 기반 환경에서 수집되는 고객 관련 정보를 추가적으로 분석하려는 시도가 이루어지고 있다. 특히 웹사이트의 보편화는 고객에 관한 정보를 종전보다 빠르고 편리하게 수집하고 이용할 수 있게 하였다. 이는 고객 개개인의 성향과 취향을 밝혀내고 획일적인 서비스가 아닌 고객과의 일대일 관계를 발전시켜 나가는 이른바 개인화 전략을 가능하게 하였다(Mobasher et al., 2000; Spiliopoulou, 2000).

(2) CRM과 데이터 마이닝

CRM은 마케팅, 콜센터, 영업분석 등 다양한 분야의 업무와 밀접한 관계가 있다 (Harmon et al., 2001). CRM이 갖추어야 할 주요 기능으로는 1) 고객 세분화와 분류 등과 같은 고객 분석기능, 2) 캠페인 계획과 관리기능, 3) 고객 대응기능, 4) 전략컨설팅 및 시스템 통합기능 등을 들 수 있다 (김재문, 2001). 현재 Gartner Group 등 여러 컨설팅 사업자들이 CRM에서 사용되는 기능을 다양한 형태로 분류하고 있으나, 그 내용을 종합하면 1) 운영적 CRM, 2) 분석적 CRM 3) 협업적 CRM으로 분류할 수 있다(오라클 솔루션 연구회, 2001). 운영적 CRM은 마케팅, 영업, 고객서비스 등의 비즈니스 프로세스의 자동화를 위하여 응용시스템을 통합, 연계시키는 데 초점을 맞추는 반면 협업적 CRM은 고객과의 정보교환이 원활하게 이루어지도록 콜센터, 웹 상에서의 서비스 제공 등을 관리하는 기능을 의미

한다. 분석적 CRM은 데이터웨어하우징(Data Warehousing), 데이터 마이닝(Data Mining), OLAP(On-Line Analytical Processing) 등의 기술을 이용하여 운영적 CRM에서 수집된 자료를 토대로 필요한 정보를 추출하고 고객 세분화를 통하여 고객에 대한 이해를 높이며 나아가 고객의 행태를 예측하는 기능을 의미하는 것으로 CRM에서 핵심기반이 되는 부분이라 할 수 있다.

이렇게 CRM의 중요성이 부각되면서 이에 대한 기술적 솔루션으로서 데이터 마이닝 작업도 함께 주목받고 있다. 데이터 마이닝은 대용량의 데이터에 대한 실시간 분석이 가능할 뿐 아니라 데이터에 내재되어 있는 패턴을 발견할 수 있다는 장점이 있다. 데이터 마이닝 이란 자동화되고 지능을 갖춘(Automated and Intelligent) 데이터베이스 분석기법으로 90년대 초반부터 지식발견(Knowledge Discovery)라는 이름으로 소개되었는 데 일반적으로 대량의 데이터로부터 새롭고 의미 있는 정보를 추출하여 마케팅 등 전략적 의사결정에 활용할 수 있게 하는 작업으로 정의 할 수 있다(Berry & Linoff, 1997; Fayyad et al., 1996).

한편, 기존의 고객 데이터 외에 인터넷 사용이 확산되면서 웹 로그 데이터, 웹사이트 컨텐츠 데이터와 같은 다양한 종류의 고객정보가 웹 상에서 수집됨에 따라 고객 관련 데이터의 양이 폭발적으로 증대되어 데이터 마이닝에 대한 요구가 상대적으로 증가하고 있다(O'Keefe & Mceachan, 1998). 그러나, 데이터의 폭발적인 증가와 자료 형태의 다양성으로 인하여 데이터 마이닝을 활용하는 자료분석에 많은 어려움을 겪고 있으며 적절한 자료수집 방법의 부재로 고객분석 기반의 구축 조차 효과적으로 수행되지 못하는 이면 또한 내재하고 있다.

(3) 자료수집

그 동안 대부분의 인터넷 사업자들은 고객자료를 수집하여 마케팅에 활용하거나 판매를 통해 수익을 낼 수 있다는 판단 하에 가능하면 많은 고객자료를 확보하려고 노력하였다. 그 결과는 수집되는 자료의 양이 폭발적으로 증가하게 하였으며 이로 인해 데이터 관리와 분석에도 많은 어려움이 야기되었다. 이렇게 Cache-Busting을 이용하는 자료수집 방법은 맹목적으로 이루어지기 때문에 구체적인 분석을 위해서는 수집된 자료를 다시 정리해야 하는 어려움이 따르게 되었다(Mogul & Leach, 1997). 실제로 인터넷 접속인구의 급증, 접속 가능한 페이지의 증가, 새로운 홈페이지의 등장, 요청되는 페이지의 증가 등을 고려한다면 유명사이트의 경우 하루에도 수백만의 접속과 이에 따른 기가바이트급의 자료저장이 필요하게 될 것이다. 이는 사이트에 접속하는 사용자에 관한 분석을 통해 가치있는 정보를 추출하는 데 요구되는 시간과 비용의 증가를 의미한다. 따라서 고객 (사용자) 집단에 대한 대표성이 있고 신뢰할만한 데이터를 효율적으로 수집하고 관리하는 방법이 절실히 요구된다.

(4) 연구의 목적

본 연구의 목적은 인터넷 상에서 효과적인 CRM 구현을 위해 필요한 고객분석 작업에 데이터 마이닝 기법을 응용하는 것이다. 구체적으로 접속자의 인구 통계적인 자료와 사이트 방문 자료를 기반으로 그들의 특성을 분석하여 계층을 분류하는 작업을 시도하며 특정 사이트에 방문하는 접속자들을 분류하는 모델을 제시하고자 한다. 또한 웹데이터에 대한 데이터 마이닝의 실제 적용도를 높이기 위하여 이에 대한 배경과 개념 등을 소개하고 실제 데이터를 수집하여 패턴을 발견하고 발견된 패턴의 확인 작업을 수행하고자 한다.

CRM을 효과적으로 수행하기 위해서는 고객분석 기반의 구축이 필수적이다. 이러한 목적으로 실제 데이터(고객의 인구통계학적 데이터와 웹접속 데이터)를 수집하고 통합하며 이를 분석해 보는 것이 본 연구의 핵심이 되는 부분이다. 특히 웹 환경에서 새로이 자료를 수집할 때, 데이터의 대표성과 신뢰성을 유지하면서 분석 가능한 수준의 데이터 양을 확보하는 방안을 모색하였다. 이를 위해 고객의 구성비를 반영하여 패널구성비를 결정하고 이들 패널의 특정사이트 접속현황을 측정하였다. 이 방법은 웹 환경에서 폭발적으로 증가하는 데이터를 효율적으로 수집, 관리하는 데 하나의 대안으로 고려될 수 있을 것이며 이러한 방법을 통해 분석결과의 신뢰성 제고에도 도움을 줄 수 있을 것으로 보인다.

2. 연구배경

(1) 국내 · 외 동향

데이터 마이닝 분야는 90년대 초반부터 Knowledge Discovery, Information Discovery, Automated Knowledge Acquisition 등의 이름으로 다양한 지능을 갖춘 데이터 분석 기법들이 소개되어 왔다. 이들 기법들은 통계적 분석이나 인공지능에 기반을 둔 것으로 OLAP과 데이터 웨어하우징 등의 등장과 함께 다량의 데이터에 대한 실시간 분석의 필요성이 크게 부각되면서 지난 몇 해 동안 각종 학회(e. g., 한국경영정보학회, Knowledge Discovery Workshop, International Conference of Knowledge Discovery and Data Mining)와 전문서적(e. g., Journal of Data Mining & Knowledge Discovery, Journal of Intelligent Information Systems) 등을 통하여 많은 연구들이 소개되고 있다(김충영 외, 2002). 그러나 대부분의 연구들이 데이터 마이닝 기법을 이용하여 분류적 예측(Classificatory Prediction)모형을 구축하고 모형의 예측 정확도를 비교하는 데 초점을 맞추고 있다. 이 외에도 상황적 변수 등이 예측 모형의 정확성에 영향을 주는 가를(Kim & McLeod, 1999) 조사하거나 예측력을 높이기 위해 여러 기법을 통합적으로 사용하는 등의 새로운 시도가 있기는 하지만 분류적 예측 외 다른 작업(예: Clustering, Affinity)에 대한 시도는 활발하게 이루어지고 있지 못하다. 한편 CRM 분야에서 데이터 마이닝의 응용에 대한 연구는 최근 들어

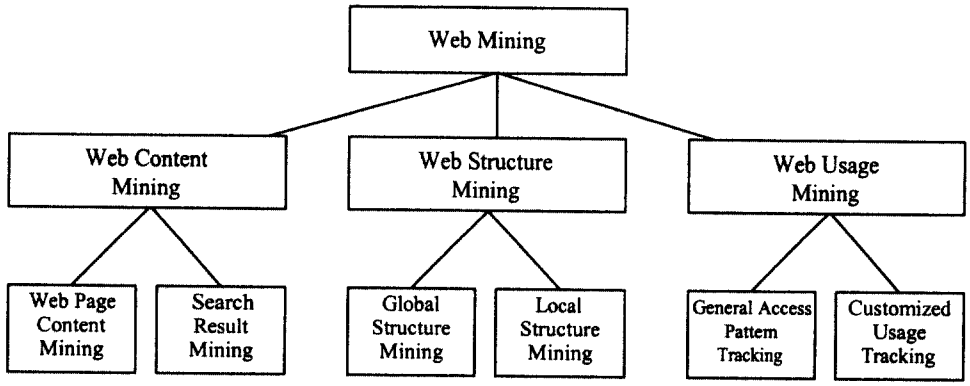
학회(경영정보학회 등)를 통해서 발표되고 있으나, 대부분이 연구 틀(Framework)이나 전략 등을 제시하는 수준에 머물고 있으며 (김병곤, 2001: 이선조 외 2001; Cooley et al., 2000; Mobasher et al., 2000; Srivastava et al., 2000). 실제 데이터를 활용하는 경험적인 연구는 미흡한 편이다.

산업계의 동향을 보면 최근까지 CRM 도입을 위하여 데이터 마이닝이 시도되고 있는 분야는 은행, 보험사 등의 금융권과 일부 유통업, 이동통신 등으로 매우 제한된 수요층을 보이고 있다(김충영 외, 2002). 이러한 편중현상은 위 업종들이 다른 업종에 비하여 기업과 고객간의 거래가 지속적으로 유지되며 따라서 분석을 할 만한 신뢰성 있는 데이터가 상대적으로 제대로 관리되고 있기 때문이라 사료된다.

한편 인터넷 사용의 확산과 정보통신 기술의 발달은 고객 데이터의 추적, 수집, 관리 기술의 진보를 가져왔다. 특히, 웹 환경에서는 개별고객의 방문내역 및 거래 내역을 손쉽게 획득할 수 있다는 이점 때문에 데이터 마이닝의 응용이 적극적으로 시도되고 있다. 최근에는 "Clickstream Data Analysis"라는 진보된 개념도 등장하고 있으며, 국내외 일부 시장조사 기관에서 이러한 개념을 이용하여 인터넷 기업을 대상으로 시장조사 및 고객성향 분석 등에 데이터 마이닝을 응용하고 있다. 그러나, 현재 추진되고 있는 웹 환경에서의 데이터 마이닝 연구는 시장조사 기업, 데이터 베이스 업체, SI업체, 통계 소프트웨어 업체 등이 각각의 입장에서 접근하고 있기 때문에 단편적인 자료분석 자체에 머무는 경향이 있어 구체적인 CRM으로 발전되고 있지 못하다고 판단된다. 이와 더불어 웹을 통해 수집되는 데이터 형태의 다양성과 방대한 규모 등으로 인하여 수집된 데이터를 분석 가능한 형태로 전환 하는 데 필요한 비용과 시간 및 적절한 방법론의 부재 등이 실질적인 연구 성과를 가져오는 데 장애요인이 되고 있다.

(2) 데이터 마이닝과 웹마이닝

웹상에서의 사용 가능한 정보자원이 증가함에 따라 데이터 마이닝 기술을 World Wide Web(WWW)으로 확장한 개념인 웹 마이닝에 대한 연구가 최근 많은 주목을 받고 있다. 웹 마이닝의 세부적인 연구 사항들에 대해서 체계적인 틀을 세운, Etzioni (1996)의 정의에 의하면 "웹 마이닝이란 데이터 마이닝 기술을 이용하여 웹 문서나 웹 서비스로부터 자동적으로 정보를 발견하거나 추출해내는 과정"을 뜻한다. 일반적으로 웹마이닝은 사용 목적에 따라, 구조마이닝(Structure Mining), 활용마이닝(Usage Mining), 컨텐츠마이닝(Content Mining) 등의 세가지 영역으로 나눌 수 있다 (Strivastava, et al., 2000). 그러나 실제 웹 마이닝 작업을 할 때 세가지 영역이 독립적으로 이루어지는 경우도 있지만, 대부분 혼합되어 작업이 이루어지는 것이 일반적이다.



〈그림 1〉 웹마이닝의 분류체계(Strivastava, et al., 2000)

웹 구조마이닝은 웹 페이지의 링크 구조로부터 정보를 추출하고, 웹 페이지간의 유사성과 관계에 대한 정보를 추출하는 과정이다. 따라서 웹 구조마이닝은 특정 웹 페이지가 다른 어떤 웹 페이지로부터 링크가 되는가? 또는 어떤 웹 페이지가 다른 웹 페이지로 링크 되고 있는가? 등의 정보를 추출해 내는 과정이다. 즉, 웹 문서내의 하이퍼링크 구조에 초점을 두고 웹에 링크 된 하위구조의 모델을 발견하는 과정으로 유사한 웹 페이지를 묶거나 서로 다른 사이트간의 관계와 유사성에 관한 정보를 얻는데 유용하다. 웹 구조마이닝은 외부 웹 페이지간의 하이퍼링크를 분석 대상으로 하는 외부적 구조(Global Structure) 마이닝과 웹 사이트 내의 하이퍼 링크를 분석대상으로 하는 내부적 구조(Local Structure) 마이닝으로 구성된다.

웹 컨텐츠마이닝은 웹 페이지를 구성하고 있는 텍스트, 이미지, 기타 다른 형태의 컨텐츠로부터 정보를 추출해내는 과정이다. 어떤 웹사이트가 특정 주제를 가장 잘 다루고 있는가?, 사용된 언어는 무엇인가?, 특정 주제와 관련된 웹 페이지는 어떤 것들인가? 등에 대한 유용한 정보를 추출해내는 과정이다. 검색엔진, 추천엔진 등이 컨텐츠마이닝을 사용하여 웹에서 필요한 정보를 정확하게 찾을 수 있도록 도와준다.

웹이 계속 발전하면서 HTML 문서 대신에 구조화된 메타 데이터를 사용할 수 있는 XML 문서의 사용이 증가하고 있지만, 아직은 보편화되지 않았으며, HTML 파일과 같이 웹에 존재하는 정보들은 이질적이며, 비구조적이기 때문에 이들로부터 자동적으로 데이터를 추출하고, 조직화하여 관리하는 것은 매우 어렵다. 웹 페이지의 구조정보, 분류, 필터링과 같은 고급기능의 구현이 어려운 이유가 여기에 있다.

최근의 연구들은 보다 지능적인 도구 개발에 초점을 두고 있는데, 지능형 에이전트 접근법과 데이터베이스 접근법으로 구분된다. 지능형 에이전트 접근법은 웹 기반 정보를 자율적으로 발견하고 구조화 시킬 수 있는 인공지능 시스템을 개발하는 것이다.

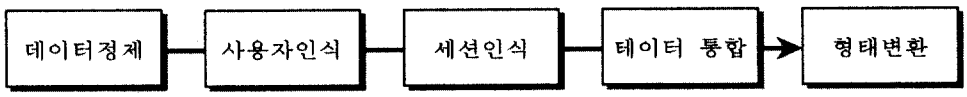
반면 데이터베이스 접근법은 이질적이고 비구조적인 웹 정보를 구조화된 데이터베이스로 통합하고 여기에 OLAP과 데이터 마이닝 기법을 적용하여 분석하는 것이다.

활용마이닝은 사용자가 이동한 경로, 방문한 페이지, 각각의 페이지에 머문 시간, 웹 페이지에 접근을 유도한 페이지, 검색을 마치고 나가게 된 페이지에 대한 정보추출 과정이다. 즉, 사용자 브라우징 패턴과 접속 패턴에 대한 데이터 마이닝 과정이라고 말할 수 있다. 최근에는 웹 기반 애플리케이션들과 인터넷마케팅 분야에서 웹 개인화(Web Personalization)에 대한 연구 및 기술개발이 활발히 이루어지고 있다. 웹 개인화는 개별적인 사용자 수준의 관심사와 선호, 웹 행동분석, 사용자 정보를 기반으로 구현되는데, 주로 컨텐츠의 개인화, 실시간 추천, 개인화 된 광고와 이벤트를 제공하는 캠페인 관리의 목적으로 활용한다. 일반적으로 웹 활용마이닝의 경우 대부분의 연관규칙, 순차패턴 분석, 분류, 군집화 등과 같은 데이터 마이닝 기법들이 그대로 이용되며, 웹 페이지간의 연관관계, 시간간격에 따른 순차패턴, 접근패턴에 따른 사용자의 분류 등과 같은 유용한 정보를 자동으로 발견한다.

(3) 웹데이터와 데이터 사전처리 과정(Pre-processing)

데이터 사전처리 과정은 다양한 사용 가능한 데이터 소스에 포함된 활용 정보와 컨텐츠 정보 그리고 구조 정보를 패턴발견을 위한 데이터 추출로 전환시키는 과정이다. 로그 데이터가 저장되면 이를 분석 가능한 데이터로 변환하는 작업이 필요하다. 로그에 저장되는 것은 원시적인 형태의 데이터다. 따라서 단순히 로그파일만을 가지고 분석할 수 있는 범위는 한정적이다. 또한 데이터 분석에 있어서 가장 중요한 요소중의 하나는 데이터의 질(quality)로서, 초기 로그 데이터는 이러한 요건을 충족시키지 못하고 있는 것이 대부분이다. 따라서 분석에 적합한 형태로 데이터를 변환시키고, 정제하는 과정은 필수적이라고 할 수 있다. 〈그림 2〉는 로그 데이터의 전처리 단계를 위한 단계이다.

데이터의 정제(Cleaning Data) : 서버 로그 데이터에서 분석에 필요하지 않은 항목을 정제하는 기술로 데이터 마이닝 뿐만 아니라 웹로그 분석에도 아주 중요한 기술이다. 일반적으로 사용자가 웹서버에 접속하게 되면 HTML 태그 때문에 불필요한 이미지 파일 등을 웹 서버에 요청하게 된다. 웹 활용마이닝의 경우 사용자의 행동을 분석하는 것이 목적이기 때문에 분석에 필요한 부분을 제외한 gif, jpeg, jpg, map 등을 포함한 로그 파일을 정제하게 된다. 이러한 과정을 통해 로그 데이터의 용량이 일반적으로 1/10에서 1/40정도로 축소되는 효과를 가져온다.



〈그림 2〉 데이터 전처리과정

사용자 인식(User Identification) : 웹마이닝 분석을 위해서는 사용자 식별이 반드시 이루어져야 한다. 그러나 사용자 컴퓨터의 캐시, 기업의 방화벽, 프락시 서버 등으로 인해 사용자 식별 작업에 많은 어려움이 있다. 로그분석을 할 경우 개별 사용자의 궤적에 관한 정보를 필요로 하며, 일반적으로 IP 주소와 브라우저의 종류를 이용하여 개별 사용자를 식별하게 된다.

세션 인식(Session Identification) : 세션 인식은 사용자의 시간초과 유무를 점검하여 정제하는 과정이다. 일반적으로 로그분석을 할 경우, 웹사이트 내에서 체류 시간이나 방문 경로를 파악할 경우 한번의 방문으로 발생한 데이터를 분석하게 된다. 또한 유동 IP나 프락시 서버를 사용할 경우 발생하는 문제인 여러 사람이 동일한 IP를 사용하는 것을 구분할 필요가 있다. 따라서 웹사이트를 방문하여 일정시간이 경과한 후에 발생되는 움직임에 대해서는 개별 사용자로 구분하게 된다.

데이터 통합 (data integration) : 성명, 주소 등 개인의 인적 자료를 정제된 웹로그 데이터와 통합하는 과정을 의미한다. 인식된 로그데이터를 해당고객의 등록정보와 연결시켜 고객 개인별 접속 성향분석이 가능하게 된다.

형태변환 (transformation) : 데이터 마이닝 분석에 필요한 데이터 형태로 전환하는 과정이다. 사전에 정의된 필요 정보와 기법에 적합하도록 데이터의 항목과 기간을 선택해야 한다.

(4) 분석방법(데이터 마이닝&OLAP)

N개의 독립된 속성으로 구성된 데이터를 N차원(Dimension) 공간에서 분석한다는 개념을 기반으로 하는 On-Line Analytical Processing(OLAP) 도구는 데이터를 다차원으로 분석하며, 결과를 주로 이해하기 쉬운 차트의 형태로 제공한다. 또한 시-구-동, 연도-분기-월 등과 같이 속성 내에 계층관계가 형성되어 있을 경우 Drill-Down/Up 기능을 이용하여 심도있는 분석을 가능하게 한다(조재희 1996). 예를 들어 OLAP도구를 이용하여 "특정지역의 제품 판매량을 분기별, 연령대별로 보여주시오."라는 조회를 할 경우 이것은 3차원(판매량, 기간, 연령) 질의에 해당하며, 만약 1사분기의 판매량이 평균치 이상으로 월등할 경우 1사분기를 Drill-Down하면 1~3월간의 심층 분석결과를 제공한다. 이렇듯 OLAP은 흥미있는 정보를 다양한 형태로 제공하는 장점을 지니고 있다. 그러나 위의 예에서와 같이 '기간'과 '연령' 등과 같이 판매에 영향을 미치는 기준이 되는 속성들을 사전에 파악할 수 없는 경우에는 원하는 결과를 얻기 위해서 수많은 시도와 실패를 필요로 한다는 단점도 있다. 이것은 OLAP이 원천적으로 미리 가설(정보)을 세우고 데이터를 통해 그 가설을 확인하는 방식이기 때문이다. 즉, 사용자가 자신의 경험에 비추어 가설을 세우고, 여기에 맞추어 질의를 만들어 실행하며, 결과를 검토하여 가설이 맞는지를 확인하는 것이다. 만약 가설이 틀리다는 결과가 나온 경우에는 확인되는 결과가 나올 때까지, 혹

은 사용자가 가설이 틀렸다고 생각할 때까지 질의를 반복한다. 이러한 방식은 데이터 상에 속성(차원)들이 너무 많아 적절한 가설을 세울 수 없는 상황에서는 적용하기 곤란하며, 가설이 확인되거나 부정되는 것 외에는 새로운 정보가 거의 생성되지 않는다. 이에 반해 사용자가 미리 가설을 세우지 않아도 알고리즘이 가능한 가설을 스스로 생성하고 이를 검증하는 데이터 마이닝은 가설을 발견하는 방식으로, 사용자의 가이드를 거의 받지 않고 정보를 찾는다 [장남식, 1999]. 이러한 이유로 지금까지 데이터 마이닝은 OLAP과는 별개의 개념으로 인식되고 사용되어 왔으며 일반적으로 데이터 마이닝을 통해 발견된 규칙을 OLAP을 활용하여 분석하고 확인하는 데 사용되어 왔다. 그러나 데이터를 구성하는 속성의 수가 그리 많지 않은 경우나 데이터 내에 분류를 위한 결정적인 핵심 (Key) 속성을 포함하는 경우 OLAP은 단순히 데이터 마이닝을 보완하는 이상의 결과를 제공할 수도 있을 것이다.

(5) 웹데이터와 분류적 패턴분석

분류는 각 데이터를 미리 정해진 몇 개의 그룹 중 하나에 할당하는 것을 말한다 (Fayyad et al, 1994). 이는 사용자 그룹이 미리 규정되어 있다는 점에서 군집화와 구분된다. 웹 마이닝에서는 특정한 클래스나 카테고리에 포함 된 사용자 프로파일을 개발하는데 사용된다. 사이트의 운영자는 보유하고 있는 사용자들 중 다른 경쟁 사이트로 이탈할 사용자들은 누구이며 그 특성은 어떤 것인지, 또는 특정 제품을 구매하는 사용자들은 누구이며 어떤 특성을 갖는지 등의 내용을 알고 싶어한다. 예를 들면, 웹 서버의 로그파일을 분류하면 /news/sports/golf에서 기사를 검색한 사용자의 30%가 40-50대이고 강남에 거주한다는 흥미로운 규칙을 발견할 수 있다. 이런 경우 사용할 수 있는 웹마이닝 기법이 분류이며 의사결정나무(Decision Tree)와 같은 귀납적 알고리즘을 이용하여 분석을 수행할 수 있다.

패턴분석 과정은 웹 마이닝의 마지막 단계로, 패턴발견 단계에서 찾아낸 규칙이나 패턴 중에서 유용하지 않거나 의미없는 규칙과 패턴을 여과하는 단계라고 할 수 있다. 또한 유용하고 의미 있는 규칙과 패턴을 찾아냈다면 그것을 이해하고, 해석하는 단계이기도 하다. 가장 일반적인 형태의 패턴분석 방법으로는 OLAP과 시각화(Visualization)가 있다. OLAP은 다차원(Multidimensional) 데이터 베이스 안에 저장되어 있는 웹 활용 데이터를 선택적으로 추출해 내고, 여러 관점으로 데이터를 볼 수 있도록 해준다. 시각화 기법에서는 패턴을 그래프로 표현하고 이를 다른 색깔들로 값을 할당하여 구분함으로써 데이터 내에서 전반적인 패턴이나 성향을 나타낸다. 컨텐츠와 구조에 관한 정보는 특별히 많이 활용되는 웹 페이지, 컨텐츠의 종류, 특정한 하이퍼링크 구조에 일치하는 웹 페이지의 패턴을 추출하는데 사용될 수 있다.

3. 연구방법

(1) 데이터 수집과 사전처리 과정

국내 인터넷 전문 조사업체인 A사는 인터넷 센서스를 통해 인터넷 프로파일의 변화를 지속적으로 모니터링하고 이용자 중심 측정방법으로는 조사하기 어려운 인터넷 이용행태 및 인터넷 시장구조를 분석해 왔다. 이러한 분석 결과를 토대로 국내 인터넷 이용자의 구성비를 반영해 패널 구성비율을 결정하고 Random Digit Dial(RDD)방식을 적용, 전화를 통한 리쿠르트를 통해 전국규모의 패널을 선정했다. RDD는 난수 생성을 통해 무작위로 대상자를 추출, 샘플링에서 발생할 수 있는 표본오차를 제거하는 선진적인 통계조사 방식으로 이렇게 패널의 조건에 일치하는 대상자가 선정되면 패널 ID와 패스워드를 PcMeter와 함께 발행하게 되는데 PcMeter는 로그 트래킹 소프트웨어로서 각 패널 구성원들의 PC에 상주하면서 패널 구성원이 인터넷에 접속하면 가동하여 그 구성원의 인터넷 서핑경로를 A사의 데이터베이스 서버에 실시간으로 전송하게 된다.

본 연구에서 사용된 데이터는 A사에서 제공한 734명의 패널 구성원들에 관한 인구통계학적 데이터와 이들 패널들이 2001년 7월 1일부터 7월 31일 한 달간 한 번 이상 접속한 신문사 사이트(디지틀조선, 한국일보, 동아닷컴, 조인스닷컴)에 대한 로그 데이터를 사용하였다. 특히 데이터 사전처리 과정을 통해 본 연구에서 필요로 하는 패널들의 ID와 접속사이트에 대한 데이터를 우선적으로 웹 로그 데이터에서 추출하였고, 이들을 패널 ID를 매칭키로 하여 패널들의 인구통계학적 데이터와 병합(Merge) 하였다. <표 1>은 패널들이 한 달간 신문사 사이트의 채널에 접속한 횟수와 비율을 상위 채널별로 보여주는 통계치이다.

(2) 분석 기법

본 연구에서는 신문사 채널 유형 중 가장 접속 빈도가 높은 '뉴스'와 '방송연예' 채널

〈표 1〉 신문사 채널 방문율

채 널 유 형	분 포
뉴 스	8,589 (23.9%)
방송연예	4,753 (13.2%)
스 포 츠	3,374 (9.4%)
커뮤니티	2,340 (6.5%)
기타 (부동산, 법률, 라이프 등등)	16,960 (47.0%)
합 계	36,016 (100.0%)

〈표 2〉 자료의 구성필드

속 성	속성 값
채널유형	뉴스, 방송연예
성 별	남자, 여자
나 이	10대초반, 10대후반, 20대초반, 20대후반, 30대초반, 30대후반, 40대초반, 40대후반, 50대초반
교 육	고재, 고졸, 대재, 대졸, 대졸이상
직 업	무직/기타, 블루, 주부, 학생, 화이트
도 시	군/읍, 대도시, 중소도시
주 소	강/충/제, 경기/인천, 경상도, 서울, 전라/광주

에 접속하는 방문자들의 인구통계학적 특성을 OLAP과 데이터 마이닝의 의사결정나무를 통해 분석하여 분류규칙을 도출하고 상호비교 하였다. 이 기법들은 고객의 프로파일이나 접속행위에 근거한 조건문(If-Then-Else) 형식의 규칙을 이용하여 고객세분화 혹은 고객개인화를 시도하는 가장 일반적인 방법으로서 생성되는 규칙이 이해하기 쉽고 친숙하다는 장점이 있다. 또한 생성된 규칙을 토대로 OLAP의 결과가 어느 정도 데이터 마이닝의 결과를 보완할 수 있는가를 조사하였다. OLAP과 데이터 마이닝 툴로는 LEXKEN사의 PowerPlay6.5와 Rulequest사의 See5.0을 각각 사용하였다.

(3) 데이터의 구성 및 가공

사전처리 과정을 통해 총 12개의 변수(속성)로 구성된 데이터 세트를 생성했다. 그러나고객 개개인을 구분하기에 불필요하거나 변수 값의 구성상 의미가 없다고 판단되는 것(각각의 빈 값 정도, 분포 등을 토대로 점검하였음)들을 제외한 후, 최종 7개의 변수를 선정하였으며 목적변수로는 '채널유형'이 사용되었다 <표 2>. 또한 연속형 변수 값을 취했던 '나이'는 분석 목적에 합당하게 9개의 구간으로 분리하였으며, 나머지 변수들의 값은 최초 수집 그대로의 형태로 사용하였다.

데이터 마이닝의 의사결정나무를 이용한 분류분석을 위하여 <표 3>에서 보듯이 전체 13,342개의 방문횟수(레코드)를 70% (9,339)와 30% (4,003)로 임의로 분리하여 모형 추정용(Training)과 모형 시험용(Testing) 데이터 세트를 구성하였다. 모형 추정용 데이터는 예측모형을 구축하기 위하여, 모형 시험용 데이터는 모형의 예측력을 시험하기 위하여 사용하였다. 또한 임의 표본 추출과정에서 발생하는 데이터의 치우침(Bias)을 예방하고, 결과의 일반화를 도모하기 위해 위의 방법을 5회 반복하여 5개의 데이터 세트를 만들었다.

〈표 3〉 분석 데이터 세트의 구성

	뉴 스	방송연예	합 계
모형 구축용(70%)	6,012	3,327	9,339
모형 시험용(30%)	2,577	1,426	4,003
합 계	8,589	4,753	13,342

4. 결과분석

(1) OLAP 분석

OLAP분석을 위해 'Measure' 값으로 각각의 채널유형에 대해 '사람수' <그림 3>와 '선택률'을 생성하였다. 이 때 '사람수'란 채널방문 횟수를 의미한다. 분석은 가능한 모든 속성 조합을 고려하여 <표 4>에서와 같이 6가지의 1차원 분석과 15가지의 2차원 분석으로 분리하였으며 그 결과 <표 5>와 같은 규칙들이 생성되었다.

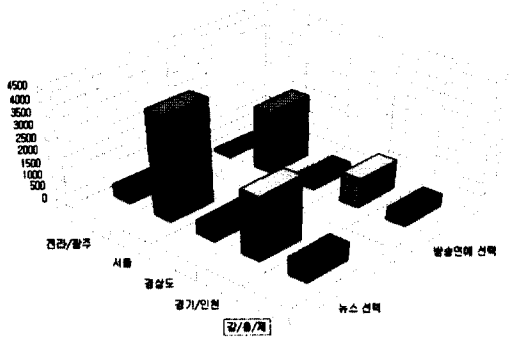
규칙의 성립조건은 규칙에 해당하는 사례

〈표 4〉 속성조합

	성 별	나 이	교 육	직 업	도 시	주 소
성 별	1차원	2차원	2차원	2차원	2차원	2차원
나 이		1차원	2차원	2차원	2차원	2차원
교 육			1차원	2차원	2차원	2차원
직 업				1차원	2차원	2차원
도 시					1차원	2차원
주 소						1차원

〈표 5〉 OLAP산출 규칙

	차 원	규칙 수	규 칙 예 시
뉴 스	1차원	10	주소='전라/광주'
	2차원	19	나이='20대초반' & 교육='대졸이상'
방송연예	1차원	3	나이='10대초반'
	2차원	20	직업='블루' & 도시='중소도시'



	사람수	뉴스 선택	방송연예 선택
강/충/제	1216	722	494
경기/인천	3774	2574	1200
경상도	984	583	401
서울	6747	4196	2551
전라/광주	621	514	107
주소	13342	8589	4753

〈그림 3〉 1차원 분석 결과

중 하나의 부류 값('뉴스' 또는 '방송연예')이 전체 사례의 75% 이상이며, 포함하는 사례의 수가 최소 40개 이상인 경우로 설정하였다. 아래의 <그림 3>은 '뉴스'채널을 방문한 고객에 대한 규칙으로 '주소'를 통한 1차원 분석 결과이다. '전라/광주' 지역에서 뉴스채널을 접속한 방문자의 수는 전체 621명 중 514명으로 82.8%에 달하며, 사례의 수는 514개이다. 이는 사전에 정의한 성립 조건을 만족하기에 규칙으로 만들어 진다.

(2) 데이터 마이닝 분석

모형구축에 사용한 Rulequest사의 See5.0은 Quinlan이 1993년 발표한 C4.5 알고리즘을 보완한 프로그램이다 (Quinlan, 1993). 모형의 과잉맞춤(overfit)으로 인한 예측력 저하를 방지하기 위해 마디의 최저 순수도 및 최소 관측 개수를 이용한 가지치기 (Pruning) 방식을 택하였다. 이 값들은 모형 구축 전에 미리 정하는데, 마디의 최저 순수도는 마디를 구성하는 사례들에서 한 종류의 부류 값에 속한 사례의 비율이 사전에 정의한 최저 순수도보다 커지면 나무의 확장을 중지하는 방식이며, 최소 관측개수는 끝마디에 포함된 사례의 개수가 정의된 값 이하가 되면 확장을 중지하는 방식이다. 따라서 최저 순수도의 값을 낮게 정의할수록, 또한 최소 관측개수의 값을 크게 할수록 나무의 구조는 단순화된다. 본 연구에서는 위의 두 가지 중 먼저 만족되는 기준에 도달하면 나무의 확장을 멈추었다. 또한 OLAP의 결과와의 비교를 위해 최저 순수도를 75%, 최소 관측개수를 40으로 정하였다. 입력변수 선정에 있어서는 의사결정나무는 분류나 예측에 영향력이 없는 변수들을 모형 구축 시 자체적으로 배제시키므로 수집한 모든 변수들을 입력변수로 고려하였다.

5개의 모형 시험용(Test) 데이터 세트에 대한 예측모형의 정분류율과 모형이 제시한 규칙은 각각 <표 6>, <표 7>에 종합되었다. '뉴스' 방문자를 '뉴스' 방문자로 정분류한 비율은 평균 80.2%, '방송연예' 방문자를 '방송연예' 방문자로 정분류한 비율은 평균

〈표 6〉 정분류율 비교

	정 분 류 율			
	뉴스 → 뉴스		방송연예 → 방송연예	
세트1	80.0%	평균 80.2%	67.7%	평균 67.2%
세트2	81.9%		67.0%	
세트3	84.7%		65.3%	
세트4	77.9%		66.0%	
세트5	76.4%		70.2%	

〈표 7〉 의사결정나무 모형이 제공하는 규칙

	차 원	규칙 수	규 칙 예 시
뉴 스	1차원	8	주소='전라/광주'
	2차원	6	직업='화이트' & 도시='중소도시'
	3차원	5	나이='20대초반' & 교육='대졸이상' & 성별='남자'
	4차원	2	나이='20대후반' & 주소='경기/인천' & 성별='남자' & 도시='대도시'
방 송 연 예	1차원	2	나이='10대초반'
	2차원	12	직업='블루' & 도시='중소도시'
	3차원	11	나이='30대초' & 교육='대졸이상' & 직업='주부'
	4차원	4	나이='20대후반' & 주소='경기/인천' & 성별='여자' & 도시='대도시'

67.2%로 '뉴스' 방문자를 정분류한 비율이 '방송연예' 방문자를 정분류한 비율에 비해 평균 13%정도 높게 나타났다. 발견된 규칙의 수에 있어서는 '방송연예'가 총 28개로 '뉴스'에 비해 7개 더 많았으며, 1차원 규칙에 비해 2,3,4차원 규칙들의 비율이 상대적으로 높았다. 이 같은 결과는 '나이', '주소' 등과 같이 데이터 자체에 '뉴스' 방문자 분류에 결정적인 역할을 하는 변수들이 비교적 분명하게 존재했다는 점과 '뉴스' 표본의 수가 '방송연예' 표본의 수에 비해 거의 2배 정도 많았다는 점에 기인한다고 판단된다.

(3) 결과 비교(OLAP vs. 데이터 마이닝)

다차원 분석 방식인 OLAP은 이론적으로는 3차원 분석 이상도 가능하다. 그러나 이

경우 테이블이나 챠트 형식으로 제공되는 정보들은 대부분 상당히 복잡하고 때에 따라서는 조잡하기까지 하다. 따라서 3차원 이상의 분석을 통해 필요한 정보를 추출하기는 현실적으로 어렵다고 판단된다. 2차원 분석 또한 데이터를 구성하는 속성의 수가 증가할수록 필요 정보를 발견하는 작업은 힘들어질 수 밖에 없다. 예를 들어 속성의 수가 10개라면 2개의 가능한 조합의 수는 10C2, 즉 45개나 되며 속성의 수가 더욱 증가할수록 조합의 수는 기하급수적으로 늘어난다. 이러한 이유로 OLAP은 가설(정보)발견 위주의 조회에 기반한 데이터 마이닝과는 달리 가설확인 중심의 조회방식이라고 인식되어 왔다. 그러나 본 연구의 대상인 신문사 채널 방문자의 특성분석과 같이 속성의 수가 많지 않은 경우 OLAP은 가능한 속성의 조합을 모두 검토함으로써 부분적으로나마 가설발견 역할도 수행할 수 있다.

의사결정나무는 개체의 속성과 부류로 구성된 데이터로부터 순환적 분할(Recursive Partitioning)방식을 이용하여 나무모형을 구축하는 기법으로, 나무의 가장 상단에 위치하는 뿌리마디(Root Node), 속성의 분리기준을 포함하는 내부마디(Internal Nodes), 마디와 마디를 이어주는 가지(Link), 그리고 최종 분류를 의미하는 잎(Leaves)으로 구성되며, 분류나 예측에 주로 사용한다(Quinlan, 1986). 그러나 뿌리마디 또는 상위마디 구축을 위해 속성을 결정할 때 사례들을 분류하는데 결정적인 역할을 하는 속성 값의 존재로 인해 불필요한 나머지 속성 값들이 함께 선정되곤 한다. 이럴 경우 최종적으로 모형의 구축을 통해 제시되는 규칙들의 차원은 불필요하게 늘어나게 된다. 그럼에도 불구하고 의사결정나무는 기법의 특성상 3개 이상의 속성(차원)으로 구성되는 규칙들도 무리없이 찾아내기 때문에 OLAP이 지닌 근원적인 문제를 극복할 수 있다.

OLAP 및 데이터 마이닝을 통한 분석결과를 비교, 요약하면 다음과 같다. 첫째, "IF 나이 = 10대초반 THEN 방송연예"와 같은 1차원 규칙은 OLAP의 경우 모두 발견된 반면 의사결정나무에서는 일부 발견되지 않았다. 이것은 위에서 언급한 나무모형 구축과정의 특성상의 문제에 기인한 것으로 발견되지 않은 일부 규칙은 다른 속성과 조합된 즉, 2차원 또는 3차원의 규칙으로 대부분 발견되었다.

둘째, 데이터 마이닝을 통해 발견된 "IF 직업 = 화이트 & 도시 = 중소도시 THEN 뉴스"와 같은 2차원 규칙은 OLAP에서는 모두 발견되었으며, 그 외의 다수의 2차원 규칙도 발견되었다. 이것은 OLAP이 조합이 가능한 2개 속성으로 구성되는 모든 조합을 검색하기 때문에 가능하다.

세째, "IF 나이 = 20대초반 & 교육 = 대졸이상 & 성별 = 남자 THEN 뉴스"와 같이 데이터 마이닝을 통해 발견된 규칙은 "IF 나이 = 20대초반 & 교육=대졸이상 THEN 뉴스"와 같이 OLAP에서 발견된 규칙에 '성별=남자' 등과 같은 하나 이상의 조건을 추가함으로써 OLAP 규칙을 구체화시킬 수 있고 결과적으로 발견된 규칙의 신뢰성을 증가시킨다.

네째, OLAP을 이용해 '교육'과 '직업'을 조합하여 분석하면 '교육 = 대졸이상 & 직업 = 주부'라는 조합에서는 특정 패턴을 찾을 수가 없다. 따라서 이러한 조합으로 생성된 규칙은 근원적으로 배제된다. 그러나 여기에 '나이 = 30대초'라는 속성을 추가하면 "IF 교육 = 대졸이상 & 직업 = 주부 & 나이 = 30대초 THEN 방송연예"라는 규칙을 발견할 수 있다. 이것이 바로 가설발견에 있어 데이터 마이닝이 OLAP에 비해 확실한 우위를 점할 수 있는 특징이다. 따라서 데이터 마이닝은 분석대상에 대하여 알려진 가설이 없는 경우 또는 완전히 새로운 패턴을 찾으려는 경우에 적합하며 OLAP에 비해 훨씬 자유롭게 사용할 수 있다.

선행연구(Cooley et al., 2000)에서 주장되었듯이 OLAP은 발견된 패턴을 분석하고 확인하는 데 효과적인 것으로 나타났다. 다만 데이터를 구성하는 속성의 수가 제한적인 경우에는 OLAP은 가설확인 뿐만 아니라 데이터 마이닝을 보완하는 이상의 역할을 수행할 수 있다고 판단된다. 특히, 데이터 마이닝에서 간과되기 쉬운 추가 가설의 발견이 가능하기 때문에 데이터 마이닝 작업 이전에 분석방향을 정하는 데 유용하게 사용될 수 있다고 판단된다.

5. 결 론

본 연구에서는 CRM을 위한 기초작업으로서 고객프로필과 사이트 접속자료를 통합하여 분석하였다. 데이터는 무료 e-mail 서비스와 다양한 Community를 통해서 상당한 양의 고객 데이터를 확보하고 있는 인터넷 사업자의 실제 데이터를 사용하였다. 인터넷 사업자의 실제 데이터를 사용한 이유는 다음 두가지 때문이다. 첫째, 최근 많은 인터넷 사업자들이 그들의 고객에 대해서 좀 더 이해하려는 노력을 기울이고 있다. 실제로 고객의 특성과 기호, 방문행태 등을 이해할 수 있다면 이를 기반으로 고객 세분화(Segmentation)가 가능할 것이다. 예를 들어 고객의 거주지, 재산정도, 교육수준, 연령 등 인적정보를 토대로 동일 사이트에 접속하는 고객의 공통점을 찾게 된다면 이들 고객에 접근할 수 있는 적절한 마케팅 미디어가 무엇인지, 어느 페이지에 홍보물을 게재하는 것이 효과적일 것인가 등을 결정하는 데 도움을 줄 수 있을 것이다. 따라서, 인터넷을 이용하는 사업자에게 웹 기반 마이닝을 통해 고객에 대한 이해를 제공할 수 있는 방안을 제공한다는 것은 큰 의미가 있다고 하겠다. 둘째, 웹 기반 마이닝의 핵심은 웹으로 부터의 자료를 어떻게 하면 효율적으로 수집할 것인가, 또한 이렇게 수집된 자료를 다양한 DB와 어떻게 통합하고 분석하여 필요한 정보를 추출할 것인가 일 것이다 (Mitchell, 1999). 본 연구에서는 실제 인터넷 사업자의 사용자 그룹의 비율에 따라 구성된 패널을 활용하여 효율적인 자료수집 방안을 모색하였다. 패널 구성원에 대한 웹 데이터를 수집함으로써 신뢰성과 대표성을 확보하면서 분석대상 자료의 양을 적절한 수준으로 유지할 수 있었다.

또한 고객자료 분석에서는 OLAP과 데이

터 마이닝 기법(의사결정나무)을 동시에 사용하여 그 분석 결과를 비교함으로써 각 기법의 결과를 상호 확인하고 보완할 수 있었다. 이 결과는 데이터 마이닝 기법에 의해서 발견된 패턴을 분석하고 확인하는 작업에서 OLAP이 유용하게 사용될 수 있다는 과거 연구의 주장을(Cooley et al., 2000) 다시 확인하는 것이다. 특히, OLAP은 핵심변수를 효과적으로 선별할 수 있기 때문에 본격적인 데이터 마이닝 기법의 응용에 앞서서(혹은 후에라도), 데이터 마이닝에서 간과되기 쉬운 추가 패턴의 발견에 사용될 수 있는 가능성을 보여주었다.

참 고 문 헌

김재문, E-비즈니스 모델에 맞는 eCRM 구축 실행가이드, 거름, 2001.

김충영, 장남식, 김준우, "이동통신 해지고객 예측모형의 비교분석에 관한 연구", 경영정보학연구, 제12권 제1호, 2002.

김병곤, 최성, 박순창, "e-Business에서 eCRM 전략", Proceedings, 경영정보학회 2001 추계학술대회, 서울, 2001. 11.

오라클솔루션 연구회, 정보화의 새로운 패러다임: e-비즈니스 시스템, 교우사, 2001.

이선조, 지태창, 김현도, 박동규, 박종덕, 양재영, "자동추천 알고리즘을 이용한 원투원 마케팅 시스템", Proceedings, 경영정보학회 2001 추계학술대회, 서울, 2001. 11.

장남식, 홍성완, 장재호, *데이터마이닝*, 대청미디어, 1999.

조재희, 박성진, *데이터웨어하우징과 OLAP*, 대청미디어, 1996.

ITG 기술정보서비스팀, Technical Report, LG-EDS 시스템, pp.2-25, July 1996.

Berry, M., and Linoff, G., *Data Mining Techniques: For marketing, Sales, and Customer Support*. New York, NY: John Wiley and Sons, 1997.

Cooley, R., B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," Technical Paper, Department of Computer Science and Engineering University of Minnesota, 2000.

Etzioni, O., "The World Wide Web: Quagmire or gold mine," Communications of the ACM, Vol. 39, No. 11, 1996.

Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smith, "From Data Mining to Knowledge Discovery." In Advance in Knowledge Discovery and Data Mining,(eds) Fayyas U. M., G. Piatetsky-shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/MIT Press, CA., pp.1-34, 1996.

Gartner Group, "Website Personalization Strategies/Technologies," Conference Presentation, SPG1WebSite400Cabrams, 2000.

Harmon, P., Rosen, M., and M. Guttman, *Developing E-Business Systems and Architectures: A Manager's Guide*, Morgan Kaufmann Publishers, San Francisco, 2001.

Kim, Choong N. and McLeod Jr., "Expert, Linear Models, and Nonlinear Models of Expert in Bankruptcy Prediction, "Journal of Management Information Systems, Vol. 16. No. 1, 1999.

Mitchell, T. M., "Machine Learning and Data Mining," Communications of the ACM, Vol. 42, No.11, 1999.

Mobasher, B., Colley, R., and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Communications of the ACM, Vol. 43, No. 8, 2000.

O'keefe, R. M. & Mceachern T., "Web-based Customer Decision Support Systems," Communications of the ACM, Vol. 41 No. 3, 1998.

Quinlan, J. R., "Induction of Decision Trees," *Machine Learning*, 1986.

Quinlan, J. R., C4.5: *Performance for Machine Learning*, Morgan Kaufmann Publisher, San Mateo, California , 1993.

Raguram Sasisekharan and V. Seshadri, "Data Mining and Forecasting in Large-Scale Telecommunication Networks," AT&T Bell Lab, 1996.

Spiliopoulou, M., "Web Usage Mining for Web Site Evaluation," Communications of the ACM, Vol. 43, No8, 2000.

Srivastava, J., R. Cooley, M. Deshpande, P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD, Jan. 2000.

A Study on the Application of Data-Mining Techniques into Effective CRM (Customer Relationship Management) for Internet Businesses

Choong-Young Kim · Nam-Sik Chang · Sang-Uk Kim

Abstract

In this study, an analytical CRM for customer segmentation is exercised by integrating and analyzing the customer profile data and the access data to a particular web site. We believe that effective customer segmentation will be possible with a basis of the understanding of customer characteristics as well as behavior on the web. One of the critical tasks in the web data-mining is concerned with both 'how to collect the data from the web in an efficient manner?' and 'how to integrate the data(mostly in a variety of types) effectively for the analysis?' This study proposes a panel approach as an efficient data collection method in the web. For the customer data analysis, OLAP and a tree-structured algorithm are applied in this study. The results of the analysis with both techniques are compared, confirming the previous work which the two techniques are inter-complementary.