한국지능정보시스템학회논문지 제13권 제3호 2007년 9월(pp.83~99)

고객관계관리를 위한 통합 데이터마이닝 모형 연구

송임영
홍익대학교 컴터공학과
(iysong@cs.hongik.ac.kr)

이태석
한국과학기술정보연구원
(tsyi@kisti.re.kr)

신기정
한국과학기술정보연구원
(kjshin@kisti.re.kr)

김경창
홍익대학교 컴퓨공학과
(kckim@cs.hongik.ac.kr)

오늘날 디지털 정보기술의 발달로 정보관리와 활용에 대한 인식이 높아지면서 효과적인 정보관리와 정보 활용방안에 대한 연구가 활발해지고 있다.

본 논문은 웹 서버에 의해 자동으로 수집되는 로그 파일로부터 고객 가치 판단 기준을 고객의 행동 기반에 두고 고객을 분류하여 과학정보포털 서비스를 이용하는 이용자의 각 분류군에 해당하는 정보와 인터페이스를 제공할 수 있는 통합 모형을 제안하였다.

고객관리 측면에서 본 논문은 정보 서비스를 제공하는 웹 사이트의 기존고객을 분류하여 패턴을 분석함으로써 고객 위주의 사이트 운영정책과 동적 인터페이스를 제공하기 위한 웹사이트 활용 방안을 제시하였다. 또한, 고객의 지속적인 관리와 각 고객 분류군별에 맞는 서비스를 제공하고 고객의 관리에도 기여할 수 있을 것이다.

논문접수일 : 2007년 02월 게재확정일 : 2007년 07월 교신저자 : 김경창

1. 서론

오늘날 디지털 정보기술의 발달로 정보관리와 활용에 대한 인식이 높아지면서 효과적인 정보관리와 정보 활용방안에 대한 연구가 활발해지고 있다. 신속하고 정확하게 여러 가지 상황에 대한 적절한 의사결정을 위한 의미 있는 고급 정보 혹은 지식들이 필요할 수밖에 없다. 이러한 디지털 정보 욕구를 만족시키기 위해 다양한 연구 및 활동을 유도하는 곳이 과학정보포털서비스를 제공하는 사이트이다.

인터넷상에서 일련의 서비스를 제공하는 정보포털 사이트는 많은 사용자와 함께 많은 정보와 다양한 볼거리를 제공하면서 구조가 점차 복잡해지게 되었다. 한 사이트 내에서 제공되는 정보가 많아지면 그만큼 이용할 수 있는 정보도 많아지게 되지만, 단편적인 정보를 효율적으로 찾기는 어려워진다.

웹 사이트에서 고객에게 알맞은 정보를 제공하는 전략을 세우기 위해서는 고객 개개인의 행동 패턴에 대한 정보가 필요하다. 이와 같은 정보를 기반으로 고객의 욕구를 충족시킬 수 있는 개인화된 서비스 제공을 위해 고객의 특성에 맞는 동적인 웹 페이지 구성이나 링크정보를 제공할 수 있다. 이런 필요성에서 인터넷에서의 고객관계관리(Customer Relationship Management : CRM)에 대한 연구가 활발히 이루어지고 있으며(박주석, 2000), 전자상거래 관련 연구에서 고객관계관리 연구가 많이 시도되고 현재 많은 사이트에서 상용화되고 있는 것처럼 디지털 정보를 제공하는 과학정

보포털사이트에서 사용자들의 요구사항과 특징들을 파악하여 고객들에게 필요한 콘텐츠와 정보를 제공한다면 더욱 적극적인 관심을 보일 것이다.

본 논문은 웹 서버에 의해 자동으로 수집된 로그 파일에 데이터마이닝 기법을 적용하여 유용한 정보를 얻고자 한다. 사용자가 자신의 사이트에 방문한 경우 로그 파일에 흔적을 남기게 되며 이러한 방문자의 정확한 데이터를 기반으로 한 로그 파일 분석을 통한 고객 분석은 마케팅 피드백을 할 수 있는 고객 분석 방법이다(Zhu, Greiner and Haubl, 2003).

웹 서버의 로그 파일에는 많은 트랜잭션이 일어나고 데이터들의 누적이 끊임없이 진행된다. 웹 페이지를 구성할 때 이렇게 누적된 데이터들의 변화를 관찰해 얻은 정보를 의사결정의 중요한 참고자료로 삼는다.

이러한 고객관계관리를 위하여 고객 데이터를 분석하는데 있어 단일 데이터마이닝 기법을 사용하는 방법보다는 좀 더 깊이 있는 데이터 분석을 위하여 군집화, 의사결정나무, 신경망 등 타 기법들과 결합을 통하여 고객 분석의 정확도(accuracy)를 향상 시키는 방법들이 소개되고 있다.

특히, 연관규칙 기법을 통해 찾아낸 규칙들은 고객들의 서비스 활용 패턴이나 웹에서의 사용자 접근 패턴 분석에 사용될 수 있다. 하지만, 정확한 고객 분류 없이 전체 고객을 대상으로 고객의 행동 패턴을 분석함으로써 고객 속성별 특징에 맞는 정확한 활용 패턴을 분석할 수 없는 문제가 있다. 이러한 문제 해결을 위하여 고객 속성별 고객을 세분화하고 분류함으로써 보다 정확한 고객 분석이 가능해질 수 있다.

본 논문에서는 고객가치 판단 기준을 고객의 행동 기반에 두어 데이터마이닝 기법 중에 군집화를 이용하여 고객을 세분화하고, 세분화 결과에 의사결정나무를 적용함으로써 고객을 분류하고 분류된 각 고객 군별 정보 활용 연관성을 분석하기 위하여 연관 규칙 기법을(R. Agrawal, T. Imielinkski and A. Swami, 1993) 활용하여 각 분류군에 해당하는 고객들의 활용 서비스, 콘텐트와 키워드 및 콘텐트 연관성이 분석 가능한 통합 데이터마이닝 모형을 제안하였다.

이 통합 데이터마이닝 모형을 통하여 고객 분석의 정확도 향상이라는 관점에서 기존의 단일 데이터마이닝 모형이 가졌던 문제점인 보다 정확한 고객 분류군별 행동 패턴과 서비스 활용 패턴을 분석하여 고객이 속한 분류군에 맞는 정보 서비스와 인터페이스를 제공하는 방법을 제안하고자 한다.

본 논문은 5개의 장으로 구성되어 있으며, 각 장의 내용은 다음과 같다. 제 1장은 서론부분, 제 2장은 지금까지 수행되어온 관련 연구와 데이터마이닝을 위한 웹 로그 파일 분석에 관하여 기술한다. 제 3장은 제안 통합 데이터마이닝 모형을 설계하고 제 4장은 과학정보포털 서비스를 제공하는 사이트를 이용하는 고객들의 한 달 웹 로그 파일을 통합 데이터마이닝 모형에 적용하여 동적인 정보 서비스 결과에 대하여 설명하고 제 5장은 결론을 맺는다.

2. 관련 연구

2.1 인터넷 비즈니스 기반의 CRM

CRM과의 통합을 통해 인터넷 비즈니스는 단지 저렴한 매스 비즈니스 수준에서 정교하게 타겟팅된 수익성 높은 비즈니스의 모습으로 변화될 수 있다. 이들간의 연계와 상호 보완적인 기능의 통합은 더 이상 새로운 시장이 생겨나기 어려울 정

도로 포화상태에 이른 시장에서 기업의 생존과 지속적인 성장을 위한 통합된 고객중심 마케팅 전략 실행으로의 혁명을 실제로 가능하게 할 것으로 보인다(방성희, 1998).

효율적인 CRM을 위해서 꼭 필요한 것 즉, 고객에 대한 특성 정보와 구매 의사결정 기준 그리고 구매 행위에 대한 정보를 획득하고 관리하는 과정에는 상당한 비용과 시간이 소요되며, 그 보다는 최초에 입력되는 고객 정보는 신뢰성이 부족하여, 지속적으로 갱신되지 않으면 CRM을 통해 얻을 수 있는 이점은 거의 없다. 만일 데이터가 최신의 정보를 정확하게 반영하지 못한다면 이를 기반으로 한 의사결정은 오히려 역효과를 가져올 수도 있기 때문이다.

이러한 데이터의 신뢰성을 제고하고 지속적으로 갱신하기 위한 가장 효과적인 방법이 인터넷의 상호 작용적이고(interactive) 쌍방향적인 커뮤니케이션을 통한 고객 정보 및 반응정보의 관리이다. 이렇게 인터넷 마케팅과 연계되어 구현되는 CRM은 기존의 CRM과 근본적으로 다르다. 먼저, 고객 정보를 획득하는 과정과 그 질이 달라지는 측면이다.

인터넷 마케팅은 인터넷에 개설된 홈페이지나 쇼핑몰에 통신망을 통해 고객이 접속하여 인증되면서 시작되는데, 일단 접속이 되면 고객은 의식하지 못하지만 고객의 모든 행동을 모니터링 할 수 있다.

이렇게 수집된 데이터가 각종 통계, 데이터마이닝 기법으로 분석되어 고객 개개인을 타겟으로 한 수준 높은 마케팅 조치가 설계될 수 있기 때문에 기존의 CRM과 그 격이 달라진다.

데이터의 획득과는 반대편에서 인터넷을 통해 얻어낸 고객 개개인의 정보를 활용하는 측면을 살펴보면 앞서 설명한 바와 같이 고객이 인터넷에 회원 확인을 통해 접속하게 되면 데이터베이스에 저장된 개인 마케팅 정보와 그 고객이 전에 구매한 실적, 관심 정보 등이 고객이 요청하는 제품 또는 서비스에 따라 적절하게 배치되어 구매 의사결정을 지원함으로써 궁극적으로는 개인 고객을 목표로 하는 가상 상점을 실시간으로 만들어 낼 수 있다.

2.2 웹 데이터 마이닝

어느 웹 사이트에 심심해서 방문을 했던지 그렇지 않으면 구체적인 목적을 갖고 방문을 했던지 간에 기업의 웹 사이트에 한번이라도 들렸던 사람들이 사이트에서 행한 모든 일은 기록으로 남게 된다. 웹 사이트 설정이 어떻게 되었는지에 따라서 다르긴 하지만 방문자가 어떤 경로를 통해 들어왔는지 알 수 있다(Lingras, 2002).

예를 들면, 어떤 검색엔진을 통해 들어왔으며, 어떤 키워드로 검색했는지에 관한 정보들에 대한 로그파일이 남게 된다. 또한 쿠키(사용자 인증에 관한 사용자 브라우저와 서버간의 주고받은 기록)는 홈페이지 방문자의 이동경로 혹은 그 사용자가 이전에 한번 들렀던 적이 있는 사용자인지의 여부를 알려주게 된다. 그러나 그 보다도 더욱 중요한 것은 방문한 고객의 정보가 기록된다는 것이다(WEBLOG사, http://www.weblog.com/kr/faq.html).

웹 사이트에서 수집한 고객등록 정보들과 그 내용들에 대한 마이닝을 통해 기업은 인구통계학적 소비자 선호도를 발견하여 특정 광고나 배너를 포지셔닝 할 수 있도록 하는 기초자료를 추출할 수 있다. 새로운 데이터나 정보가 웹 사이트를 통해 수집되면 이 정보들은 지속적으로 데이터웨어하우스로 통합되어 향후의 의사결정에 도움을 주는 분석결과를 제공하고, 데이터베이스 마케팅과

전략기획을 위한 자료로 활용되는 것이다. 또한 웹 사이트 데이터마이닝을(M. Spiliopoulou, 1999; R. Kosala and H. Blockeel, 2000; S. Madria, S.Bhowmick, W. Ng and E.Lim, 1999)통해서 온라인상에서 제공하는 서비스와 제품간의 연관관계를 밝혀내어, 적절한 제품이 적절한 서비스와 함께 판매가 되고 있는지의 여부를 밝혀내게 될 수도 있다.

데이터마이닝을 통해서 기업은 웹 사이트상의 패턴을 의미 있는 정보로 종합해내고, 인터넷상의 고객들과 예상치들을 이해하고 연관시킬 수 있게 된다.

데이터와 웹이 제공하는 방대한 사업지식의 흐름에 근거한 웹 마이닝은 온라인 고객과의 관계를 생성하고 유지시키며 생산성 있는 온라인 상점의 최전선을 구축하는데 있어 결정적 열쇠가 되는 것이다.

2.3 웹 로그 분석

웹 로그 파일 분석은 사용자가 자신의 사이트에 방문한 경우 로그파일에 흔적을 남기게 되며 이러한 방문자의 정확한 데이터를 기반으로 고객 분석을 통하여 마케팅 피드백(feedback)을 할 수 있는 고객 분석 방법이다(B. Mobasher, N. Jain, E. Han and J. Srivastava, 1996). 이러한 로그파일 분석을 통하여 주요 고객층, 고객 구매패턴, 주 구매시간, 구매 탐색경로 등의 데이터를 추출할 수 있다. 이러한 데이터를 기반으로 인터페이스 설계나 상품 레이아웃 등의 설계, 고객 서비스의 강화 등의 다양한 대 고객 마케팅을 펼칠 수 있다.

로그 파일 분석 결과를 비즈니스에 전략적으로 활용하는 방안은 크게 세 가지로 나눌 수 있다. 첫 번째는 새로운 신규 사업개발을 들 수 있다. 신규 사업 개발은 고객의 다양한 요구를 예측하여 새로운 사이트 개발 및 새로운 시장 기회를 창출하는 것이다. 두 번째는 마케팅 및 광고 전략으로서의 활용이다. 기존의 매스마케팅이나 매스광고처럼 무차별적으로 광고 집행을 하는 것이 아닌 방문자의 방문 경로 및 페이지뷰(Page View), 영향이 많은 타깃 페이지(Target Page)에 집중적으로 타깃에 맞는 광고를 집행하여 효과를 보는 것을 말한다. 세 번째로 최적의 환경에서 사용자들이 사이트를 탐색하고 방문하도록 서버 및 회선 등의 기술적 자원 및 수행능력 계획을 수립할 수 있다.

로그분석으로부터 얻을 수 있는 분석데이터 정보는 방문자 트래픽(traffic)정보, 방문경로 정보, 방문자의 시스템 환경정보, 방문자에 관한 방문정보, 사이트 열람정보, 그리고 페이지 방문정보 등으로 구분할 수 있다.

본 논문에서는 웹 서버에 저장된 웹 로그 파일을 이용하여 사이트 열람정보와 페이지 방문 정보를 이용하여 웹 페이지 사용 패턴을 분석하고, 고객이 웹 사이트를 이용하면서 발생되는 로그 데이터와 고객이 직접 작성한 등록정보 등에 의해 얻어지는 데이터를 사용하여 사용자의 사용 패턴을 분석하여 차별화된 정보 서비스와 인터페이스를 제공하고자 한다.

2.4 개인화 서비스 현황 분석

많은 사이트에서 다양한 개인화 서비스를 제공한다. 개인화의 기본은 사이트의 근간이 되는 콘텐트와 커뮤니티, 커머스에서 고객이 어떤 부분을 소유하고 관리하고 싶은지의 니즈를 파악하고 그것을 지원할 수 있는 서비스를 구현하는 것이다. 개인화를 도입한다는 것은 사이트의 기본 아키텍처를 건드리는 큰 변화이다. 사이트가 진화할수록

개인화는 더욱 중요해진다. 사이트의 성장과 함께 유저의 니즈는 더욱 복잡해지고 세분화되는데 이런 다양한 니즈를 충족시켜주기 위해 개인화된 서비스와 콘텐트를 제공한다(이경윤, 2004).

2.4.1 아마존의 개인화 서비스

많은 개인화 사례의 모범이 되고 있는 아마존은 현재에도 계속적으로 진화하고 있다. 아마존은 "Your store"라는 개인화 메뉴를 글로벌 메뉴에 위치시키고 로그인시에 고객의 이름을 메뉴에 삽입하여 친근감을 주고 있다. 또한 로그인 전에는 "Top Seller"를 로그인 후에는 Recommendation For you(당신을 위한 추천 상품)을 전면에서 제공한다. 각각의 추천항목에 대해서는 상단에 왜 그 항목을 추천하였는지에 대해 설명하고는 부분을 생성하여 고객이 원하는 정보를 제공받고 있음을 애기해 주고 있다. 아마존은 고객과의 협업 필터링(Collaborative filtering)에 의해 개인화 서비스를 제공하고 있다.

2.4.2 일본 Goopas의 개인화 서비스

Goopas 서비스는 일본의 지하철에서 2003년 2월부터 제공되고 있는 모바일 서비스이다. 이 서비스는 일본 지하철역 자동 개찰기를 통과하는 시점에 맞춰 지하철 이용자에게 유용한 정보를 이동전화를 통해 메일로 제공하는 서비스이다. 일정구간을 왕복 이용하는 학생, 직장인 등 정액권 구입고객을 대상으로 먼저 이용자는 이용을 위해 정약권 구입과 이용자 정보를 사이트에 제공한 후 정액권 고유번호를 구파스 사이트에 입력한 뒤 이용자가 선호하는 콘텐트 정보를 선택하고 이용자가 자동 개찰기를 통과하는 시점에 맞춰 자신의 위치, 목적지, 선호정보 및 특성에 따라 자신에게 맞는 쇼핑정보, 할인쿠폰, 이벤트 정보 등의 콘텐트, 광고를 하루 4회 제공한다.

2.4.3 Daum의 개인화 서비스

Daum은 전체 타브이 로그인 박스에서 개인화를 실시하였다. 사용자의 성별에 따라 로그인 박스 좌측의 백그라운 칼라를 파랑과 분홍으로 달리 가져갔으며 남성이 로그인 했을 때와 여성이 로그인 했을 때 제공되는 서비스에 있어서도 로그인 박스 하단의 텍스트 프로모션 영역과 추천! Daum 세상은 동일하지만, 그 아래에 제공되는 메노는 서로 다른 콘텐트를 제공하였다. Daum은 고객이 기본적으로 제공하는 profile인 성별을 기본 rule로 정하여 개인화서비스를 제공하는 규칙기반 필터링(Rule-based filtering)방법을 사용하였다. 검색에 있어서는 "리모컨" 기능을 두어 관심 키워드를 간편하게 입력하여 반복적으로 확인할 수 있도록 하였다.

2.4.4 쇼핑몰의 개인화 서비스

GS 이숍은 로그인 사용자들에게 나만의 쇼핑 내역이라는 메뉴를 주어 주문관련정보, 쿠폰 등 혜택정보, 자신의 문의내역, 나만의 지식방, 맞춤쇼핑, 단골매장, 최근 이벤트 등의 내용을 제공하고 있다. 특히 나의 지식방, 단골매장 등은 자신이 직접 등록한 지식이나 단골매장을 보여주는 곳으로써 고객이 사이트에 대한 충성도를 높일 수 있는 공간이다. 또한 생일이나 기념일 등에는 고객 한 사람 만을 위한 쿠폰이 주어짐으로 고객에게 특별한 혜택임을 강조한다.

2.4.5 이동통신사의 개인화 서비스

KTF의 "팝업 서비스"는 휴대폰 바탕화면에 원

하는 서비스를 설정 후 핸드폰 폴더만 열면 내가 원하는 정보와 서비스가 바로 나오는 서비스로 고객의 성향에 맞춰 차별화 된 콘텐츠를 제공한다.

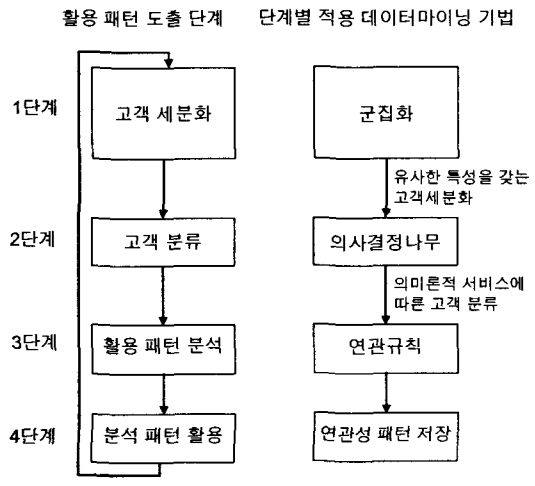
3. 고객관계관리(CRM)를 위한 통합 데이터마이닝 모형 설계

이 장에서는 고객 행동 기반 세분화 및 고객 분류를 수행하여 고객관계관리를 위한 통합 데이터마이닝 모형을 제시한다.

통합 데이터마이닝 모형 개발은 고객이 원하는 서비스 패턴, 규칙을 찾아내고 모형화하여 각 고객별 차별화된 서비스 및 콘텐트를 제공하는 것이 궁극적인 목적이다.

[그림 1]은 고객관계관리를 위한 통합 데이터마이닝 모형 설계 과정에서 각 단계별 적용되는 데이터마이닝 기법과 각 단계별 데이터마이닝 기법 적용에 따른 활용 패턴을 설명하고 있다.

통합 데이터마이닝 모형은 4단계로 구성되어



[그림 1] 통합 데이터마이닝 모형 설계 절차

있으며, 모형에 적용할 데이터는 과학정보포털서비스를 제공하는 사이트를 이용하는 고객들의 웹 로그 데이터로 해당 사이트의 특성을 고려하여 통합 데이터마이닝 모형을 설계하였다.

3.1 행동기반 고객 세분화를 위한 군집화

통합 데이터마이닝 모형 도출 첫 번째 단계는 군집화 기법을 이용하여 고객 세분화를 수행하는 단계로 유사한 특성을 갖는 고객들로 세분화한다. 군집화란 주어진 데이터 집합이 가지고 있는 이질적인 특성을 유사성을 바탕으로 동질적인 군집으로 분할하는 기법이므로 고객의 특성 및 성향을 반영한 고객 세분화는 고객의 행동 패턴 정보의 전략 수립에 기반이 될 수 있다.

고객 세분화를 위한 고객의 행동기반 정보는 기업이 보유하고 있는 가장 정확한 고객 정보이며, 모든 고객이 보유하고 있는 정보이므로 전체 고객에게 적용이 가능하여 시스템화 할 수 있다. 따라서 행동기반 고객 세분화를 수행하기 위하여 고객의 행동을 담고 있는 변수를 생성한다.

과학정보포털 서비스를 제공하는 사이트는 이윤창출을 목적으로 하는 사이트가 아니므로 얼마나 자주 접속하고 얼마나 제공하는 서비스를 이용하였는지를 고객 세분화 가치기준으로 삼을 수 있다. 고객 행동을 설명하기에 충분한 형태로 도출된 변수로 만들어진 세분화 결과라야만 의미있는 세분화가 될 수 있는 것은 너무도 자명하다.

따라서 고객의 한달 로그인 횟수와 서비스 활용 횟수를 이용하여 한달 평균 정보서비스 소비량을 분석하여 해당 군집에 군집화를 수행한다.

두 가지 변수 선정의 근거는 사이트에 많이 접속하고 제공하는 정보서비스를 많이 이용하는 고객을 우수 고객의 가치 판단 기준으로 삼을 수 있

기 때문이다.

고객 세분화 알고리즘은 k-평균 군집화를 수행한다.

이 방법은 군집의 수를 미리 정하고, 각 개체가 어느 군집에 속하는 지를 분석하는 방법으로서 대량의 데이터 군집분석에 유용하게 이용되는 방법이다.

개인 또는 개체 중에서 유사한 것들을 몇몇의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적인 분석방법이다.

본 논문에서는 k-평균군집화 기법을 2차원 공간에 적용하여 중심 좌표에 근접한 로그인 횟수, 정보 서비스 활용 횟수를 가진 고객들을 군집화하여 분석한다. 각 고객의 평균값을 기준으로 다음과 같은 특성을 가진 4개의 군으로 군집화하기 위하여 k의 수를 4로 지정하여 군집화를 수행하였다.

군집화의 결과 유사도가 비슷한 고객들이 같은 군집에 속하게 되며, 군집화된 4개의 고객군은 다음과 같은 특성을 가진다.

- log in count and use count가 평균 보다 월등히 높은 군
- log in count and use count가 평균과 비슷한 값을 가지는 군
- log in count에 비해 use count의 값이 높은 군
- log in count and use count가 평균 보다 월등히 낮은 군

log in 횟수에 비해 정보서비스 활용 횟수가 낮은 군에 대한 분류를 하지 않은 이유는 분석 데이터 자체가 log in 사용자에 대한 데이터만 분석하고 사이트 특성상 원하는 정보를 찾고 활용하기 위한 방문이 목적일 것이다. 따라서, log in을 하고 사용하는 고객들은 최소한의 정보 서비스는 활용할 것이라는 가정을 하고 4개의 군으로 군집을 형성한다.

군집화 제외 대상은 데이터의 일반적인 경향에서 벗어나는 고객들로 예외나 잡음으로 고려하여 VIP 고객, 잠재 이용자로 분류한다.

3.2 고객 분류를 위한 의사결정나무의 형성

통합 데이터마이닝 모형 설계 두 번째 단계는 고객 분류 단계로 의미론적 서비스에 따른 고객 분류 단계이다.

유사한 특징을 가진 고객군으로 세분화된 고객군에 대한 서비스 이용 패턴을 파악하기 위하여 각 고객군을 의사결정나무를 적용하여 분류한다.

행동 기반 고객 세분화 결과를 의사결정나무에 적용하여 목적 변수에 맞는 고객군으로 분류하여 고객들을 특정한 규칙을 이용하여 고객 특징을 예측함으로써 분류된 고객에 대한 분류군별 정보 분석이 가능하기 때문이다.

의사결정나무 적용의 목표변수는 활용 서비스를 파악하기 위함이므로 과학기술정보 포털 사이트에서 제공하는 서비스의 서비스 코드를 사용하고, 웹 로그 데이터 분석 시 고객의 속성 중 분류군별 차별화된 정보 서비스 제공에 크게 영향을 줄 수 있다고 판단되는 전공, 학력, 직종 변수를 고객군을 분류하는 예측 변수로 사용한다.

본 논문에서 사용하는 의사결정나무 알고리즘은 CHAID로 다지분리를 수행하는 알고리즘으로 결과트리를 가지치기 할 필요가 없는 장점이 있다. 형성된 의사결정나무가 거대해지면 해석의 어려움이 존재하게 되므로, 트리의 depth는 전공, 학력, 직종 등 3가지 변수를 모두 적용하여 3으로 한다.

목적변수로 사용하는 서비스 코드는 의미론적으로 그룹화하여 처리한다. 사이트에서 제공하는

200여개의 서비스코드를 목적변수로 사용하면 고객을 분류할 때 각 서비스 코드의 factor가 너무 많으므로 서비스 코드를 의미론적으로 그룹화하여 고객을 분류한다.

서비스 코드를 목적 변수로 하여 고객을 분류하므로 고객이 많이 소비하는 서비스에 따라서 고객의 해당 분류군이 달라지며 달라진 분류군에서 분석된 서비스 활용 패턴에 따라 정보와 인터페이스를 제공받게 될 것이다.

3.3 연관 규칙 기법을 이용한 정보 활용 연관성 분석

통합 데이터마이닝 모형 설계 세 번째 단계는 고객이 활용한 서비스와 콘텐트들 간의 연관성 분석과 활용 패턴을 분석하기 위해 연관규칙 기법을 적용하여 주 이용 서비스 연관성을 분석한다.

연관규칙 기법을 통해 찾아낸 규칙들은 고객들의 서비스 활용 패턴이나 웹에서의 사용자 접근 패턴 분석에 사용될 수 있다. 하지만, 정확한 고객 분류 없이 전체 고객을 대상으로 고객의 행동 패턴을 분석함으로써 고객 속성별 특징에 맞는 정확한 활용 패턴을 분석할 수 없는 문제가 있어 이러한 문제 해결을 위하여 고객 속성별 고객을 세분화하고 분류함으로써 보다 정확한 고객 분석이 가능해질 수 있다.

과학기술정보 포털 사이트에서 제공하는 200여개의 서비스 중에서 사이트 이용 고객들이 주로 이용하는 서비스는 60여개 정도인 것을 알 수 있다. 따라서, 각 분류군에 속하는 고객 중 고객 3분의 1이상이 사용한 서비스에 대해서만 분석을 하도록 한다. 연관관계도출에서 빈도수가 적은 서비스가 전체 서비스의 40%를 차지하므로 의미 있는 연관관계를 가지는 지지도의 수준을 40%로 설정한다.

의미있는 서비스 이용 패턴 선택을 위한 지지도와 신뢰도 기준은 다음과 같다.

- 각 분류 군에 해당하는 고객들이 사용한 서비스 60여 종류를 분석하여 같은 분류군 고객 중 고객 3분의 1이상이 사용한 서비스에 대해서만 분석. 연관관계도출에서 빈도수가 적은 서비스가 전체 서비스의 40%를 차지하므로 의미가 있는 연관관계를 가지는 지지도의 수준을 40%로 설정
- 서비스 코드가 두 개 이상이며 네 개 이하인 연관규칙을 선택하고, 임계치는 지지도 40%, 신뢰도 50%로 지정하여 연관 규칙을 도출

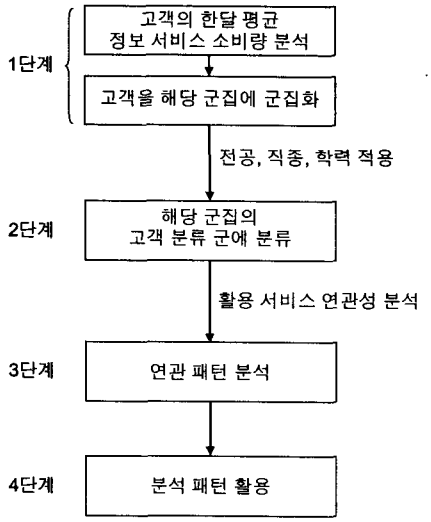
연관규칙추출에서 항목수가 증가함에 따라 추출되는 연관규칙의 수가 기하급수적으로 증가하므로 도출된 연관규칙에서 타겟이 되는 서비스를 기준으로 의미 있는 연관규칙을 추출하도록 한다.

각 군에서 분석할 서비스 연관성은 다음과 같다.

- 해당 군에서 가장 많이 활용되는 서비스 연관성 분석
- 해당 군에서 가장 적게 활용되는 서비스 연관성 분석
- 해당 군에서만 활용되는 서비스

타깃 서비스에 대한 연관규칙 추출 기준은 다음과 같다.

- 지지도와 신뢰도가 가장 높은 것을 선택, 같으면 리프트 값이 높은 것 선택
- 여러 종류의 연관규칙에서 하위 관계(relation)의 서비스를 포함하는 패턴을 선택
- 같은 서비스 카테고리에 포함되는 서비스 패턴과 다른 서비스 카테고리에 속하는 서비스에 대한 패턴을 각각 추출



[그림 2] 모형 설계과정에 적용되는 주요 변수와 분석 기준

- 분류체계를 이용하여 서비스를 상위 서비스에서 하위 서비스로 분류하여 상위 서비스가 아닌 같은 레벨의 하위 서비스에 대해서만 분석

위와 같은 기준으로 온라인상에서 고객들이 활용하는 서비스, 콘텐트 및 키워드의 연관관계를 분석하여 웹 공간에서 다양한 웹 서비스들에 대한 관심 있는 접근 패턴을 찾아내어 고객 군별 차별화된 정보서비스와 인터페이스를 제공하고자 한다.

[그림 2]는 이미 설명한 통합 데이터마이닝 모형 설계 과정에서 적용되는 주요 변수와 분석 기준을 각 단계별로 설명하고 있다.

본 논문에서는 의사결정나무만 적용하여 고객 분류를 수행하는 모형이 아닌 군집화를 적용하여 고객 세분화를 수행하고 세분화 결과에 의사결정나무를 적용하여 고객을 분류하여 서비스 이용 패턴을 파악하는 통합 데이터마이닝 모형을 제안함으로써 예측의 정확도를 향상시켜, 웹상에서는 사이트 간에 이동이 쉽고 사용자들은 유동성이 강함으로 사용자들의 요구사항과 특징들을 잘 이해해서 고객 군별 활용 서비스에 따른 서비스 방안을 강구하고자 하였다. 이는 고객을 분류하여 각 분류군의 정보 이용자별 성향에 맞는 정보 서비스와 인터페이스 제공으로 고객 만족감을 극대화할 수 있다.

3.4 통합 데이터마이닝 모형 설계 타당성

본 논문에서는 고객관계관리를 위하여 고객 데이터를 분석하는데 있어 단일 데이터마이닝 기법을 사용하는 방법보다는 좀 더 깊이 있는 데이터 분석을 위하여 군집화, 의사결정나무, 연관규칙기법 등 타 기법들과 결합을 통하여 고객 분석의 정확도(Accuracy)를 향상 시키는 통합 데이터아미이닝 모형을 설계하였다.

<표 1>은 고객관계관리에서 가장 많이 적용하는 기법인 의사결정나무와 신경망만을 적용한 단일 데이터마이닝 모형의 문제점을 비교 설명하여

<표 1> 모형 적용 가능 데이터마이닝 기법 비교

적용 기법	군집화와 의사결정나무	의사결정나무	신경망
적용 결과	고객세분화 및 고객 분류	고객 분류	고객 분류
분석	- 고객행동기반을 기준으로 유사한 특징을 가진 고객들로 군집화하여 고객 분류 - 행동 패턴이 비슷한 고객들의 서비스 활용 패턴 분석 가능	- 고객행동기반을 고려하지 않고 고객 분류 - 사이트에 많이 접근하고 서비스를 많이 활용하는 고객의 활용 서비스만 분석 가능	- 고객행동기반을 고려하지 않고 고객 분류 - 출력 결과에 대한 설명 능력의 부족으로 결과에 대한 신뢰성, 수용성, 응용성을 주지 못함 - 서비스 연관성 분석에 대한 고객 분류군의 고객 특성을 설명할 수 없음

본 논문에서 제안하는 통합 데이터마이닝 모형 설계의 타당성을 제시하고자 한다.

의사결정나무와 신경망모형은 분류와 예측을 하는데 효과적으로 많이 쓰이는 데이터마이닝 기법이다.

어떤 적용에서는 왜 이런 결정을 하게 되었는지 설명하는 것도 중요하며 의사결정나무분석은 이러한 경우에 유용하다. 예를 들어, 특정 정보를 가장 많이 소비하는 고객은 어떤 학력에 어떤 직종에 종사하며, 무엇을 전공했는지를 설명할 수 없는 신경망분석보다 이유를 설명해 줄 수 있는 의사결정나무 분석이 더 유용하다(S. Mitra, S. K. Pal and P. Mitra, 2002).

또한 신경망분석은 결과에 대한 설명 능력의 부족으로 결과에 대한 신뢰성, 수용성, 응용성은 주지 못하여 서비스 연관성 분석에 대한 고객 분류군의 고객 특성을 설명할 수 없는 단점이 있다.

본 논문에서는 고객들의 정보 활용 연관성을 분석하기 위하여 고객의 행동 기반으로 고객 세분화를 수행하고 세분화 결과에 의사결정나무를 적용함으로써 고객 분석의 정확도 향상이라는 관점에서 기존의 단일 데이터마이닝 모형이 가졌던 문제점인 보다 정확한 고객 분류군별 행동 패턴과 서비스 활용 패턴을 분석하여 고객이 속한 분류군에 맞는 정보 서비스와 인터페이스를 제공하는 방법을 제안하고자 한다.

4. 통합 데이터마이닝 모형 적용 및 검증

본 논문에서 제안하는 통합 데이터마이닝 모형을 적용하기 위하여 과학정보포털 사이트의 웹 로그 파일을 이용하였다. 과학정보포털 사이트는 수많은 고객들이 방문하며 이들은 디지털 콘텐트 분야에 대한 소비에 큰 관심을 가진 사람들이다. 따라서 각각 사용자 특성에 맞는 사용자 위주의 적극적인 관리 시스템의 운영이 필요하다. 본 논문에서 제시한 통합 데이터마이닝 모형의 목적은 웹 활동 기록을 토대로 고객 군에 맞는 차별화된 정보와 인터페이스를 제공함으로서 적극적이고 능동적인 사이트 운영이 될 수 있도록 하는 것이다.

4.1 적용 데이터 및 분석도구

과학정보포털 서비스 제공 사이트의 한 달 웹 로그 파일을 연구 목적에 맞도록 필요한 항목을 추출하고, 변형하여 EXCEL로 변환하여 구축하였다.

로그데이터에 대해 세션아이디를 사용하여 사용자별로 세션을 분류하고 세션 내에 유효한 고객 ID를 사용하여 이용자 별로 분류하였다.

그리고, 분류군별 차별화된 정보 서비스를 제공하는데 주요 변수로 판단한 전공, 직종, 학력이 기타군에 해당하는 고객은 제외시켰다.

데이터마이닝 기법을 이용한 고객의 서비스 활용 패턴을 분석하기 위해 유사한 특징을 가진 고객들로 묶어서 몇 개의 의미 있는 군집으로 나누기 위해 SPSS 12.0을 사용하여 군집분석을 실시하고, 고객의 주 사용 서비스를 분석하기 위하여 Answer tree 2.0을 사용하여 의사결정나무 분석을 실시하였다. 분류된 고객들의 정보 서비스 활용 패턴을 파악하기 위하여 연관규칙기법을 적용하였으며, SAS사의 Enterprise Miner 8.1을 이용하였다.

4.2 통합 데이터마이닝 모형 적용

통합 데이터마이닝 모형 적용의 첫 단계인 고객 세분화를 수행하기 위하여 과학정보포털 사이트를 이용하는 고객의 한달 로그인 횟수와 서비스

활용 횟수를 이용하여 한달 평균 정보서비스 소비량을 분석하여 해당 군집에 군집화를 수행하였다.

고객 세분화 방법은 통계 프로그램 SPSS 12.0을 사용하여 k-평균 군집화를 수행하였다.

<표 2>은 각 고객 당 데이터 분석 결과이며 <표 3>은 최종 군집 중심 결과이다.

<표 2> 각 고객당 데이터 분석

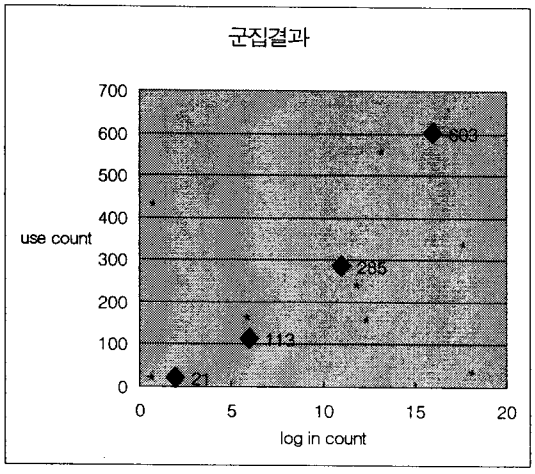
Sample data	Max use	Avg	Std
Log in_count	91	3.6783	5.5467
Use_count	926	67.278	111.07

<표 3> 최종 군집 중심

분류군 / 변수	1	2	3	4
login_count	16	6	11	2
use_count	603	113	285	21

[그림 3]은 최종 군집 중심 결과를 차트로 표현하여 설명하였다.

군집 중심좌표에서 벗어나 있는 고객들은 outlier 대상이다.



[그림 3] 최종 군집 중심 결과

1군은 평균보다 월등히 로그인 횟수와 서비스 활용 횟수가 높은 군에 해당하며, 평균과 비교하여 고객들의 로그인 횟수는 4.3배 정도 높지만 서비스 활용 횟수는 평균과 비교하여 8배 이상 높음을 확인할 수 있다. 즉, 가장 사용 빈도가 높은 군이면서 한번 접속하면 많은 정보 서비스를 소비하는 군임을 알 수 있다.

3군은 평균보다 큰 값을 가지는 고객군으로 평균과 비교하여 로그인 횟수는 3배, 서비스 활용 횟수는 4.2배 정도 높음을 알 수 있다. 2군은 평균에 가장 근접한 고객군으로 평균과 비교하여 로그인 횟수와 서비스 활용 횟수가 1.67배 정도 높음을 알 수 있다. 4군은 평균과 비교하여 두 변수에 대한 값이 월등히 적은 고객군에 해당한다.

즉, 사이트 방문 빈도가 높을수록 로그인 횟수에 비해 서비스 활용 횟수가 높은 것을 알 수 있다.

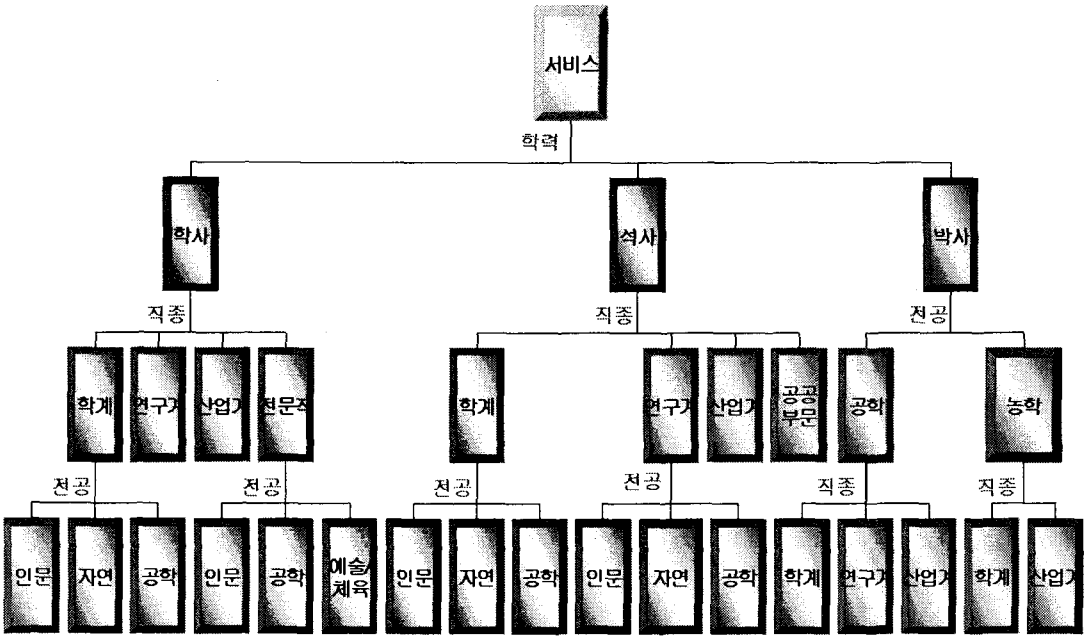
4개의 군집 결과에서 하나의 군집을 선택하여 의사결정나무에 적용하였다.

보통 데이터마이닝 프로젝트를 위해 사용하는 샘플의 크기는 모집단의 5~10% 비율의 데이터를 추출하여 사용한다. 따라서 4개의 군으로 분류된 군집 중에서 방문 횟수와 서비스 활용 횟수가 평균 보다 큰 군에 해당하며 분포 고객이 6.1%를 가지는 3군을 선택하여 의사결정나무를 이용하여 고객을 분류하고 예측하였다.

또한 3군은 두 변수에 대한 평균값과 월등히 큰 차이를 보이는 고객군도 아니며 접속한 횟수와 비례하여 정보 서비스를 이용하는 고객군으로 일반적으로 사이트를 이용하는 고객군의 특징을 가질 것으로 판단할 수 있다.

데이터의 일반적인 경향에서 벗어나는 고객들은 예외나 잡음으로 고려하여 VIP 고객, 잠재 이용자로 분류하였다.

통합 데이터마이닝 모형 적용 두 번째 단계로



[그림 4] 3군의 의사결정나무 결과

유사한 특징을 가진 세분화된 고객군에 대한 활용 서비스를 파악하기 위하여 하나의 고객군을 의사결정나무 기법을 이용하여 분류하였다.

[그림 4]는 Answer tree 2.0을 이용하여 고객 세분화 군집 중 3군에 의사결정나무를 적용한 결과를 조직도로 표현한 것이다.

목적변수로 사용한 서비스 코드는 의미론적으로 그룹화하여 처리하였다.

통합 데이터마이닝 모형 적용 세 번째 단계는 의미론적으로 그룹화하여 사용한 서비스를 이용하여 분류 고객들이 자주 사용하는 서비스에 대한 연관관계를 연관규칙 기법을 이용하여 조사하였다.

의사결정나무 적용 결과인 21개의 분류군에서 분류군 선택 기준에 적합한 3개의 분류군에 대하여 연관규칙 기법을 적용하여 정보 서비스 이용 패턴을 분석하였다.

연관규칙 기법 적용 군은 학력이 학사이며 직종이 학계, 전공이 공학인 분류군과 학력이 석사이고 직종은 연구계, 전공은 공학인 분류군, 학력이 박사이며 전공은 공학, 직종은 학계인 분류군이 해당된다.

사이트의 특성상 원하는 정보를 찾기 위하여 사이트를 방문하는 목적을 가진 고객이 많으므로 세 개의 분류군 모두 정보 검색 서비스에 집중하는 것을 확인할 수 있다. 각 분류군마다 고객의 속성이 다르므로 다른 분류군과 비교하여 활용되는 서비스와 활용되지 않는 서비스를 분석할 수 있다.

의미있는 서비스 이용 패턴 선택을 위한 지지도와 신뢰도 기준을 적용하여 서비스 코드가 두개 이상이며 4개 이하인 연관규칙을 채택하였으며, 임계치는 지지도 40%, 신뢰도 50%로 지정하여 SAS Enterprise Miner8.1을 이용하여 연관 규칙

을 생성하였다.

<표 4>는 세 개 분류군의 연관 규칙 결과이다.

<표 4> 연관 규칙 결과

분류 군	연관 규칙
1군 → 학력 : 학사, 직종 : 학계, 전공 : 공학	3,669
2군 → 학력 : 석사, 직종 : 연구계, 전공 : 공학	7,327
3군 → 학력 : 박사, 전공 : 박사, 직종 : 학계	22,163

연관 규칙 결과에서 의미있는 패턴 분석을 위하여 각 군에서 분석한 서비스 연관성과 타겟 서비스에 대한 연관규칙추출 기준을 적용하였다.

<표 5>는 사이트의 가장 기본적인 서비스에 해당하는 통합검색 서비스에 대하여 각 분류군의 연관규칙 기법 적용 결과에서 지지도와 신뢰도가 가장 높은 패턴만 정리한 내용이다.

<표 5> 통합검색 연관 규칙 결과

분류 군	지지도	신뢰도	연관 규칙
1군	75	88	통합검색 → 국내연구보고서 & 국내 학술지 & 해외 학술지
2군	77	87.5	통합검색 → 국내학술지 & 해외 학술지 & 국내 회의자료
3군	88	80	통합검색 → 국내 연구보고서 & 국내 학술지 & 국내학위논문

<표 5>의 제일 마지막 패턴이 학력은 박사, 직종은 학계이며 전공은 공학에 해당하는 분류군의 통합검색과 함께 활용되는 서비스 연관 패턴이다. 연관 규칙 결과에 따르면 해당 군의 고객들은 통합 검색 서비스를 활용하면서 국내 연구보고서와 국내학술지 그리고, 국내 학위논문 서비스를 함께 소비할 가능성이 가장 높은 것으로 분석할 수 있는데, 전체 트랜잭션에서 1군은 통합검색과 국내 연구보고서, 국내학술지, 그리고 해외 학술지가 함께 활용될 확률은 75%이며, 3군은 통합검색이 활용될 때 국내연구보고서, 국내학술지, 해외학술지가 활용될 확률은 88%이다.

따라서, 3군에 해당하는 고객들에게 통합검색 결과를 제공할 때 국내 연구 보고서와 국내 학술지 그리고, 국내 학위논문에 대한 서비스를 함께 제공함으로써 이용 고객들의 정보 검색이 편리성과 활용하고자하는 서비스에 대한 depth를 줄여 줄 수 있다.

모든 서비스에 대하여 위의 방법으로 연관규칙을 추출하여 선택 서비스를 소비할 때 연관 서비스에 대한 정보를 함께 제공해 줄 수 있다.

키워드와 콘텐트 연관성과 콘텐트들 간의 연관성도 같은 방법으로 분석하였다.

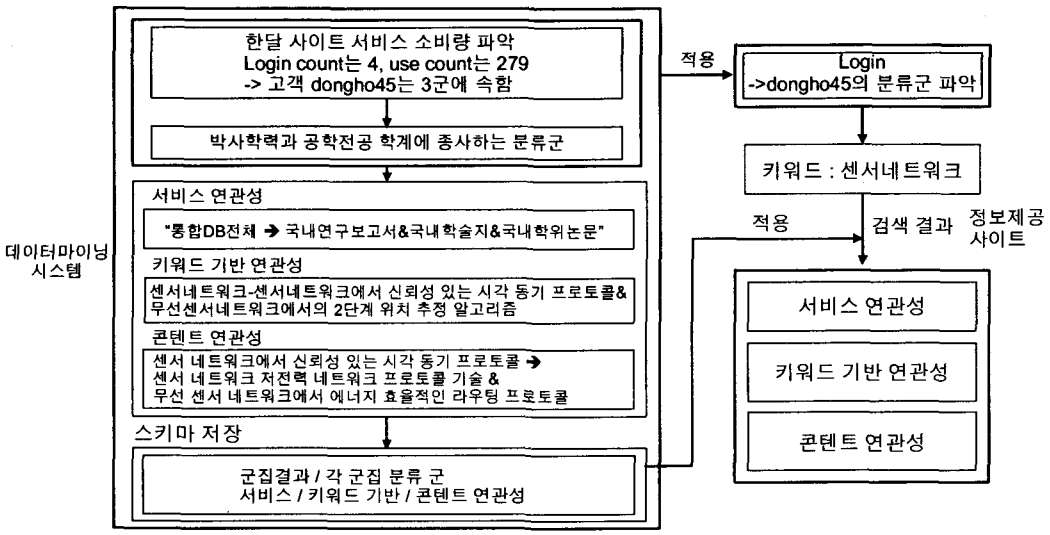
4.3 모형 검증

본 논문에서 제안하는 통합 데이터마이닝 모형의 검증은 방법론의 정확성을 증명하는 것으로 하였다. 실제 과학기술정보제공 사이트를 이용한 고객들의 한달 동안의 웹 로그 데이터를 적용하여 모형 적용의 각 단계에서 군집화에는 SPSS 12.0, 의사결정나무 기법은 Answer tree 2.0 그리고, 연관규칙은 Enterprise Miner 8.1을 활용하여 수행하였다.

수행결과 제안된 모형에 적절한 결과를 도출하였으므로 제대로 된 모형을 개발하였음을 검증한 것으로 한다.

4.4 통합 데이터마이닝 모형 적용에 따른 동적 서비스 제공

논문에서 제안하는 통합 데이터마이닝 모형을 과학기술정보 제공 사이트에 적용함으로서 어떻게 차별화된 정보 서비스와 인터페이스를 제공할



[그림 5] 통합 데이터마이닝 모형 적용 시스템

수 있는지를 설명한다.

고객관계관리를 위한 통합 데이터마이닝 모형 적용 정보 제공 정책은 다음과 같다.

- 서비스 연관성 : 고객들이 활용한 서비스 연관성을 기반으로 연관 서비스 제공
- 키워드와 콘텐트 연관성 : 키워드와 소비한 콘텐트를 분석하여 고객이 입력하는 키워드와 연관된 콘텐트 제공
- 콘텐트 연관성 : 고객이 활용하는 콘텐트들 사이의 연관성을 기반으로 연관 콘텐트 제공

연관규칙 기법을 적용하여 서비스, 키워드와 콘텐트 연관성 그리고 콘텐트 연관성을 분석하여 고객 분류군의 특성에 맞는 동적인 웹 페이지 구성이나 링크정보를 제공할 수 있다.

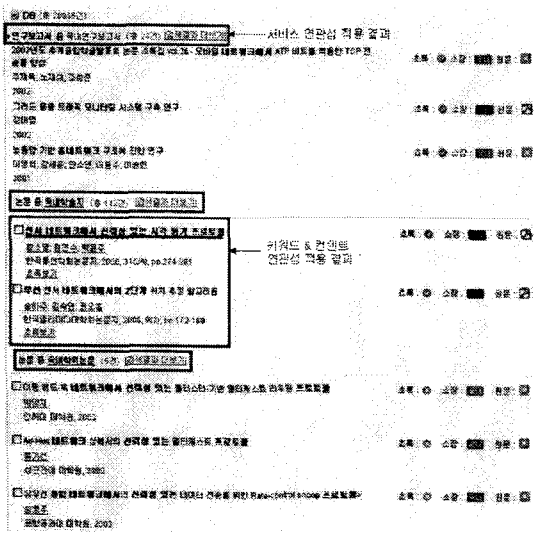
[그림 5]는 통합 데이터마이닝 시스템과 통합 데이터마이닝 모형 적용 과학기술정보 사이트의 정보 제공 정책 적용 절차를 ID가 dongho45인 고객의 정보서비스 활용으로 설명한 것이다.

고객의 한달 평균 정보 소비량을 파악하여 고객을 군집화하고 군집별로 고객 분류를 수행한다. 분류군의 서비스, 콘텐트, 키워드와 콘텐트 연관성을 분석하여 군집결과, 각 군집 분류 군, 서비스 연관성, 키워드와 콘텐트 그리고 콘텐트 연관성 스키마를 저장하여 고객이 사이트에서 정보 서비스를 활용하기 위해 로그인하면 고객 분류군의 특성에 맞는 동적인 웹 페이지 구성이나 링크정보를 제공할 수 있다.

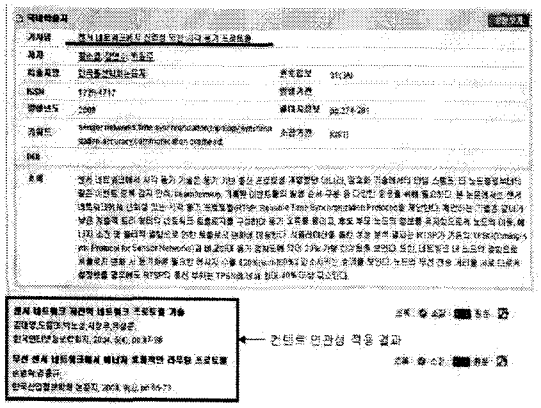
[그림 6]와 [그림 7]는 통합 데이터마이닝 모형 적용 후의 정보 제공 정책을 적용한 결과이다.

서비스 연관관계를 적용하여 통합검색과 함께 많이 활용되는 국내 연구보고서, 국내 학술지와 국내학위논문을 우선으로 제공한다.

또한 키워드 검색 결과로 해당 키워드로 많이 활용된 정보에 대하여 연관성 적용 결과를 보여주고 관심 정보를 선택하였을 때 선택 정보에 대한 정보와 함께 많이 소비되는 연관성을 가진 콘텐트



[그림 6] 통합 데이터마이닝 모형 적용후의 통합검색 결과



[그림 7] 통합 데이터마이닝 모형 적용후의 콘텐트 연관성 제공 결과

를 함께 제공할 수 있다.

이와 같이 실시간으로 축적된 데이터를 통합 데이터마이닝 기법으로 패턴을 분석하여 고객 행동에 대한 예측성을 높임으로서 고객에 대한 정확한 이해를 바탕으로 고객들에게 차별화된 정보 서비스와 인터페이스를 제공할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 전자상거래 관련 연구에서 고객관계관리 연구가 많이 시도되고 현재 많은 사이트에서 상용화 되고 있는 것처럼 디지털 정보를 제공하는 과학정보포털사이트에서 사용자들의 요구사항과 특징들을 파악하여 고객들에게 필요한 콘텐츠와 정보를 제공할 수 있는 통합 데이터마이닝 모형을 제안하였다.

과학기술정보를 제공하는 사이트의 특징상 고객의 행동기반을 고객 가치 판단의 기준으로 삼고 유사한 특성을 가지는 고객으로 군집화 기법을 적용하여 고객 세분화를 수행한 뒤, 의사결정나무를 적용하여 유사 특성을 가진 고객들의 주 이용서비스를 파악하여 고객 분류를 수행하였다. 또한, 분류된 고객 군별 정보 활용 서비스를 연관규칙 기법을 적용하여 파악하고 차별화된 정보와 인터페이스를 제공할 수 있는 통합 데이터마이닝 모형을 제안하였다.

통합 데이터마이닝 모형을 적용하여 분석된 연관성 패턴을 적용하여 서비스 연관성에 의한 연관 서비스를 제공함으로써 원하는 서비스를 소비하기 위한 depth를 줄일 수 있으며, 사용 키워드와 함께 많이 소비된 콘텐트를 제공함으로써 추가적인 정보 제공과 콘텐트 선택의 depth를 줄일 수 있다. 또한, 소비 콘텐트와 함께 많이 소비된 콘텐트를 제공함으로써 추가적인 콘텐트 정보 제공과 함께 연관 콘텐트 소비를 유도하여 고객이 관련 콘텐트를 찾기 위한 depth를 줄일 수 있다.

정보이용자에게 적합한 웹 문서를 예측하여 추천해주는 시스템은 고객의 입장에서는 원하는 정

보를 쉽게 찾을 수 있도록 하고, 웹 사이트 운영자의 입장에서는 불필요한 요청을 줄임으로서 서버의 부하를 줄여줄 수 있다.

본 논문에서는 고객 개개인의 정보와 행동 패턴을 분석하고 이를 활용하는 개인화 서비스를 제공하는 것이 아니라 고객의 분류군의 연관 패턴을 추천해주는 방식이다. 따라서, 고객 개개인의 행동 패턴을 분석하여 다음 행동이나 문서 추천 등에 대한 정보를 제공하는 웹 사이트 개인화에 대한 연구와 개발이 필요하다.

참고문헌

[1] 박주석, "정보기술과 마케팅의 변화 : CRM", *대한산업공학회*, Vol.7, No.2(2000), 28~33.

[2] 방성희, "CRM 이란 무엇인가", *월간 경영과 컴퓨터*, 1998.

[3] 이경윤, "웹사이트의 개인화 서비스 방안에 관한 연구", 이화여자대학교 디자인대학원, 2004.

[4] Mobasher, B., N. Jain, E. Han, and J. Srivastava, "Web Mining : Pattern Discovery from World Wide Web Transactions", *Technical Report TR96-050*, Department of Computer Science, University of Minnesota, 1996.

[5] Lingras. P., "Rough set clustering for Web mining", *FUZZ-IEEE Proceedings of the 2002 IEEE International Conference*, Vol. 2(2002), 1039~1044.

[6] Mitra, S., K. Pal, and P. Mitra, "Data mining in soft computing framework : a survey", IEEE Transactions on Neural Networks, Vol.13, No.1(2002), 3~14.

[7] Spiliopoulou, M., "Data mining for the Web", *Proc. 3rd European Conf. on Principles of Data Mining and Knowledge Discovery* (1999), 588~589.

[8] Agrawal, R., T. Imielinkski, and A. Swami, "Mining Associations between Sets of Items in massive Database", *Proc. Of the ACM-SIGMOD* 2003 Int'l conference on management of Data, Washington D. C., May 1993.

[9] Kosala, R. and H. Blockeel, "Web mining research : A survery", *ACM SIGKDD Explorations*, Vol.2, No.1(2000), 1~15.

[10] Madria, S., S. Bhowmick, W. Ng, and E. Lim, "Research issues in Web data mining", *Proc, 1st Int, Conf. on Data Warehousing and Knowledge Discovery,* (1999) 303~312.

[11] WEBLOG사, http://www.weblog.com/kr/faq.html.

[12] Zhu T., R. Greiner, and G. Haubl, "Learning a model of a web user's interest", *In the Int. Conf. on User Modeling,* Johnstown, USA, June, 2003.

[13] Zhu T., R. Greiner, and G. Haubl, "An effective complete-web recommender system", *In the Int. World Wide web Conf.*, Budapest, Hungrary, May, 2003.

Abstract

An Integrated Data Mining Model for Customer Relationship Management

Im-Young Song* · Tae-Seok Yi** · Ki-Jeong Shin** · Kyung-Chang Kim*

Nowadays, the advancement of digital information technology resulting in the increased interest of the management and the use of information has given stimulus to the research on the use and management of information.

In this paper, we propose an integrated data mining model that can provide the necessary information and interface to users of scientific information portal service according to their respective classification groups. The integrated model classifies users from log files automatically collected by the web server based on users' behavioral patterns.

By classifying the existing users of the web site, which provides information service, and analyzing their patterns, we proposed a web site utilization methodology that provides dynamic interface and user oriented site operating policy. In addition, we believe that our research can provide continuous web site user support, as well as provide information service according to user classification groups.

Key Words : Data mining, CRM, e-CRM, Decision tree, Clustering, Association rules, Personalize

* Dept. of Computer Engineering, Hongik University
** Korea Institute of Science and Technology