고객관계관리를 위한 데이터마이닝 통합모형에 관한 연구

송임영° 오영덕 이태석 신기정 김경창
홍익대학교, 충주대학교, 한국과학기술정보연구원
{iysong°, kckim}@cs.hongik.ac.kr, rdoh@chungju.ac.kr, {tsyi, kjshin}@kisti.re.kr

An Integrated Data Mining Model for Customer Relationship Management

Imyoung Song° R.D. Oh T.S. YI K.J. Shin K.C. Kim
Hongik Univ, Chungju National Univ, KISTI

요 약

본 논문은 웹 서버에 의해 자동으로 수집되는 로그 파일로부터 고객 가치 판단 기준을 고객의 행동 기반에 두고 군집화 기법을 이용하여 고객을 세분화하고 세분화 결과에 의사결정나무를 적용함으로써 고객을 분류하는 통합 모형을 제안하였다. 또한, 분류된 고객들의 주 서비스 활용 패턴을 분석하기 위하여 연관규칙기법을 적용하여 고객의 과학기술정보 활용의 연관성을 분석함으로써, 과학정보포털 서비스를 제공하는 사이트 이용자의 분류군에 해당하는 정보와 인터페이스를 제공하는 새로운 방법에 대하여 연구하였다.

고객 관리 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.

1. 서 론

오늘날 디지털 정보기술의 발달로 정보관리와 활용에 대한 인식이 높아지면서 효과적인 정보관리와 정보 활용 방안에 대한 연구가 활발해지고 있다. 기업들은 신속하고 정확한 마케팅 전략과 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.

웹 사이트에서 고객에게 알맞은 정보를 제공하는 전략을 세우기 위해서는 고객 개개인의 행동 패턴에 대한 정보가 필요하다. 이와 같은 정보를 기반으로 고객 분류군의 특성에 맞는 동적인 웹 페이지 구성이나 링크정보를 제공할 수 있다.

본 논문은 웹 서버에 의해 자동으로 수집된 로그 파일에 데이터마이닝 기법을 적용하여 유용한 정보를 얻고자 한다. 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.

로그 파일 분석은 사용자가 자신의 사이트에 방문한 경우 로그 파일에 흔적을 남기게 되며 이러한 방문자의 정확한 데이터를 기반으로 고객 분석을 통하여 마케팅 피드백을 할 수 있는 고객 분석 방법이다.[1]

본 논문은 고객이 이용하는 정보 서비스 관리, 고객 세분화 및 고객 데이터를 데이터마이닝 기법에 의한 고객이 속한 분류군에 맞는 정보와 인터페이스 제공 등이 목적이다.

고객 데이터를 분석하는데 있어 단일 데이터마이닝 기법만을 이용한 데이터 분석이 아닌, 좀 더 깊이 있는 알고리즘 사이에 비교를 통하여 정확한 예측을 할 수 있는 기법이 요구된다. 이러한 고객 관계 관리를 위하여 데이터마이닝 기법이 많이 사용되고 있다. 이중에 단일 데이터마이닝 기법을 사용하는 방법보다는 군집화, 의사결정나무, 연관규칙기법 등의 타 기법들과 결합을 통하여 예측의 정확도를 향상 시키는 방법들이 소개되고 있다. 본 논문에서는 고객 가치 판단 기준을 고객의 행동 기반에 두어 데이터마이닝 기법 중에 군집화를 이용하여 고객을 세분화하고, 세분화 결과에 의사결정나무를 적용함으로써 고객을 분류하는 통합 모형을 제안하였다. 분류된 각 고객 군별 정보 활용 연관성을 분석하기 위하여 연관규칙을 활용하여 각 분류군에 해당하는 고객들의 활용 서비스, 컨텐트와 워드&컨텐트 연관성을 분석하였다.

따라서 본 논문은 군집화와 의사결정나무 그리고 연관규칙기법 등을 이용하여 결합 모형을 얻어 웹에서의 고객이 이용하는 정보 활용 패턴을 분석해 내는 기법을 제시하고, 고객의 분류군에 맞는 서비스를 제공하고자 한다.

본 연구는 5개의 장으로 구성되어 있으며, 각 장의 내용은 다음과 같다. 1장은 서론부분, 2장은 관련 연구를 기술한다. 3장은 실제 데이터를 이용한 제안 모형에 대하여 설명하고 4장은 데이터마이닝을 적용한 동적인 정보 서비스 결과에 대하여 설명하고 5장은 결론을 맺는다.

2. 관련연구

2.1 데이터마이닝의 정의

데이터마이닝 기법이란 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 여기서 정보는 묵시적이고 잘 알려져 있지 않지만 잠재

적으로 활용가치가 있는 정보를 말한다. 다시 말해 데이터마이닝이란 기업이 보유하고 있는 일일 거래자료, 고객자료, 상품자료와 기타 외부자료를 포함하여 사용 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 실제 경영의 의사결정 등을 위한 정보로 활용하고자 하는 것이다.[2]

2.2 데이터마이닝 주요기법

데이터마이닝의 여러 기술들은 웹에서 사용자 접근 패턴 분석 기법과 연관성을 가지며, 체계적으로 규칙을 생성해 낸다. 또한 사용자 접근 패턴 분석 결과로 얻어진 정보가 유용하게 쓰일 수 있는 현실적인 응용 분야 중 하나이다.

데이터마이닝은 알고자 하는 정보에 따라 작업 유형이 결정된다. 작업유형은 크게 연관규칙(association rule), 연속규칙(sequence), 분류규칙(classification), 데이터 군집화(clustering) 등 4가지 유형으로 나누어진다. 그리고 이 네 가지 작업유형을 지원하는 데이터마이닝 기법은 전통적인 통계기법(예: 회귀분석, 판별분석), 의사결정나무, 신경망, 동시발생매트릭스(Co-Occurrence Matrix), k-평균 군집화(K-Means Clustering)기법 등이 있다.[2]

위에서 열거한 여러 가지 기법들 가운데 본 연구에서 적용한 군집화, 의사결정나무, 연관규칙에 대해 살펴보겠다.

군집화(Clustering)란 주어진 데이터 집합을 서로 유사성을 가지는 몇 개의 군집으로 분할해 나가는 과정으로, 하나의 군집에 속하는 데이터 점들 간에는 서로 다른 분류 내의 점들과는 구분되는 유사성을 갖게 된다.[3]

군집화는 크게 분할(partitioning)접근과 계층적(hierarchical)접근으로 나눌 수 있다. 분할 접근은 범주함수를 최적화 시키는 k개의 분할영역을 결정해 나가는 방법으로, 본 논문에서는 분류의 무게중심점을 대표 값으로 분할해 나가는 k-means 방법을 사용하였다.

k-means 알고리즘은 군집 수 K를 미리 정하고 중심점으로부터 거리를 계산하여 군집을 구하는 방법으로 이 기법은 N개의 속성으로 구성되는 각각의 레코드를 벡터로 표시하여 N차원의 데이터 공간에 나타낼 때, 유사한 특성을 갖는 레코드들은 서로 근접하여 위치한다는 가정에 근거하고 있다. 또한, 사전에 정해진 어떤 수의 클러스터를 통해서 주어진 데이터 집합을 분류하는 간단하고 쉬운 방법이다.

의사결정나무는 데이터마이닝의 분류 작업에 주로 사용되는 기법으로, 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 분류별 특성을 속성의 조합으로 나타내는 분류모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는데 사용된다.

의사결정나무의 대표적인 알고리즘들에는 CHAID, CART, C4.5 등이 있으며, 본 논문에서 사용한 알고리즘은 CHAID이다.

CHAID(Chi-squared Automatic interaction Detection)는 통계적 유의성 검정으로 각 표본에서 특정 경우에 관측된 도수와 기대도수와의 사이의 표준화된 차의 제곱합으로 정의된 것으로 검정 통계량을 사용하여 분할 변수를 선택하는 알고리즘으로 의사결정트리에 사용된 모든 변수는 범주형 변수로 모두 범주형 자료일 때 독립성을 검정하는 알고리즘이다. 그리고, 한 개의 부모 노드가 셋 이상의 자식 노드를 가질 수 있어 해석이 편리하다.

연관규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙이다. 임의의 단위에 발생한 사건들의 묶음을 트랜잭션이라 하고, 대용량의 트랜잭션들이 데이터베이스에 누적된 환경에서 사건 부류 혹은 트랜잭션간의 상호 관계를 발견하는 작업을 연관규칙 기법으로 정의할 수 있다.

연관규칙을 탐사하는 문제는 기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 갖는 항목집합들의 모든 집합들인 빈발항목집합(large itemset)을 찾아내어 연관규칙을 생성하는 단계로 이루어진다.[4]

- 지지도는 특정 항목 집합의 통계적 중요성을 나타내는 수치 값으로, 예를 들면, '전체 트랜잭션에 대해 커피와 프림을 함께 구매한 트랜잭션 수의 비율'로 측정된다.
- 신뢰도는 연관규칙의 강도를 나타내는 척도이다. 예를 들면, '커피와 프림을 구매한 고객들 중에 설탕을 함께 구매한 트랜잭션의 비율'로 측정된다.

본 논문에서는 기존의 데이터마이닝 기법들을 결합하는 방법을 사용하여 보다 정확하게 고객 특성별 고객 분류를 하여 차별화된 정보와 인터페이스를 제공함으로써 고객의 만족도 증가와 고객의 지속적인 관리에도 기여할 수 있을 것이다.

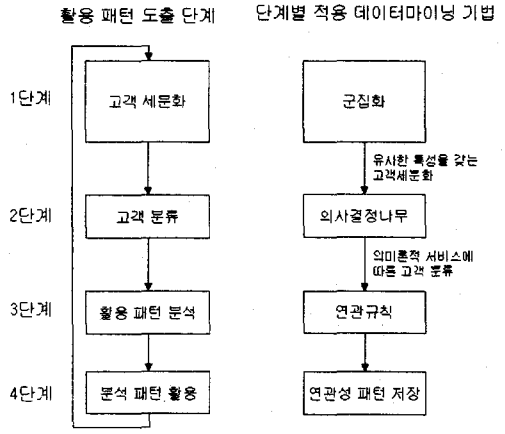
3. 고객관계관리를 위한 데이터마이닝 통합모형 연구

이 장에서는 고객 행동 기반 세분화 및 고객 분류를 위한 통합모형을 제시한다.

데이터마이닝을 위해 과학정보포털 서비스 제공 사이트 DB의 웹 로그 파일을 사용했다. 과학정보포털 서비스 제공 사이트는 수많은 고객들이 방문하여 이들은 디지털 컨텐트 분야에 대한 소비에 큰 관심을 가진 사람들이다. 따라서 각각 사용자 특성에 맞는 사용자 위주의 적극적인 관리 시스템의 운영이 필요하다. 앞으로 제시할 데이터마이닝 모형의 목적은 웹 활동 기록을 토대로 고객 군에 맞는 차별화된 정보와 인터페이스를 제공함으로서 적극적이고 능동적인 사이트 운영이 될 수 있도록 하는 것이다.

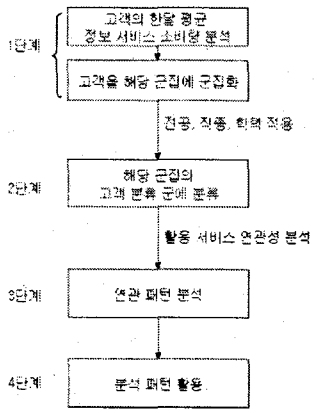
(그림 1)은 본 논문에서 제안한 통합 모형을 이용하여 서비스 이용 패턴을 도출하는 단계를 설명하는 데이터마이닝 모형 활용 절차이다. 전체 4단계로 구성되어 있으며 활용 패턴 도출 단계는 각 단계별 수행 작업 과정에 대하여 설명하고 단계별 적용 데이터마이닝 기법은 각 단계별 작업에 대한 데이터마이닝 적용 기법에 대하여 설명하였다.

(그림 2)는 사이트를 이용하는 고객에게 데이터마이닝 모형을 적용하는 절차를 설명하였다.



(그림 1) 데이터마이닝 모형 활용 절차

고객의 한달 log in count와 use count를 이용하여 한달 평균 정보서비스 소비량을 분석하여 해당 군집에 군집화를 수행한다.



(그림 2) 사이트 고객 데이터마이닝 모형 적용 절차

3.1 전처리(preprocessing)

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.

3.2 행동기반 고객 세분화를 위한 군집화

본 논문에서는 고객 세분화를 수행하기 위하여 군집화를 이용하였다.

과학정보포털 서비스를 제공하는 사이트는 이윤창출을 목적으로 하는 사이트가 아니므로 얼마나 자주 접속하고 얼마나 자주 제공하는 정보서비스를 사용하는지를 고객 가치 판단 기준으로 정의하였다.

행동기반고객세분화를 위해 우선 고객의 행동을 담고 있는 변수를 생성하였다. 고객 행동을 설명하기에 충분한 형태로 도출된 변수로 만들어진 세분화 결과라야만 의미있는 세분화가 될 수 있는 것은 너무도 자명하다. 따라서, 군집 형성 변수는 log in count, use count를 이용하였다. 두 가지 변수 선정의 근거는 제공하는 정보서비스를 많이 이용하는 고객이 우수 고객이라고 볼 수 있기 때문이다.

고객 세분화 방법은 통계 프로그램 SPSS12.0으로 K-means 알고리즘을 이용하여 4종류의 군집으로 군집화 시켰다.

4개의 군으로 고객 군을 분류한 근거는 각 고객의 평균값을 기준으로 4개의 군으로 분석하고자 하였다.
평균 보다 월등히 높은군, 평균보다 월등히 낮은군, 평균에 근접한 고객군, 평균과 비교하여 log in 횟수보다 정보 서비스 활용 횟수가 높은 군으로 분류하였다. 정보서비스 활용 횟수보다 log in 횟수가 낮은 군에 대한 분류를 하지 않은 이유는 분석 데이터 자체가 log in 사용자에 대한 데이터만 분석을 하였고 사이트 특성상 원하는 정보를 찾고 소비하기 위한 방문이 목적일 것이다. 따라서, log in을 하고 사용하는 고객들은 최소한의 정보서비스는 소비할 것이라는 가정을 하고 4개의 군으로 군집을 형성하였다.

군집화된 4개의 고객군은 다음과 같은 특성을 가진다.

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

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

<표 1>은 4개의 군으로 군집화한 결과이다.

<표 1> 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 2>는 최종 군집 중심 결과이다.

<표 2> 최종 군집 중심

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

첫 번째 군은 평균보다 월등히 log in count와 use count가 높은 군에 해당하며, 평균과 비교하여 고객들의 log in count는 4.3배 정도 높지만 use count는 평균과 비교하여 8배 이상 높음을 확인할 수 있다. 즉, 가장 사용 빈도가 높은 군이면서 한번 접속하면 많은 정보 서비

스를 소비하는 군임을 알 수 있다.

3군은 평균보다 큰 값을 가지는 고객군으로 평균과 비교하여 log in count는 3배, use count는 4.2배 정도 높음을 알 수 있다. 2군은 평균에 가장 근접한 고객군으로 평균과 비교하여 log in count와 use count가 1.67배 정도 높음을 알 수 있다. 4번째 군은 평균과 비교하여 두 변수에 대한 값이 월등히 적은 고객군에 해당한다.

즉, 사이트 방문 빈도가 높을수록 log in count에 비해 use count가 높은 것을 알 수 있다.

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 3번째 군을 선택하여 의사결정 트리를 이용하여 고객을 분류하고 예측하였다.

또한 3번째 군집은 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.

3.3 고객 분류를 위한 의사결정나무의 형성

유사한 특징을 가진 세분화된 고객군에 대한 활용 서비스를 파악하기 위하여 우선 각 고객군을 의사결정트리 기법을 이용하여 분류하였다.

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

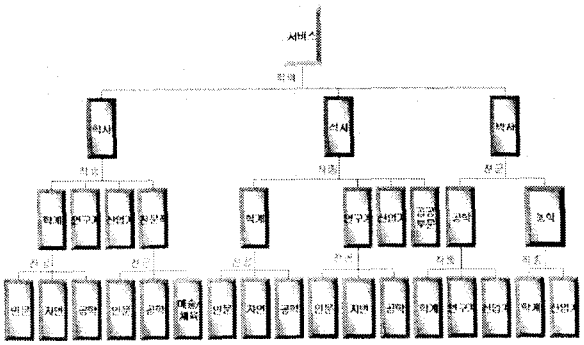
의사결정나무 적용에 목표변수는 활용 서비스를 파악하기 위함이므로 과학기술정보 포털 사이트에서 제공하는 서비스의 서비스 코드를 사용하였고, 웹 로그 데이터 분석 시 고객의 속성 중 분류 군별 차별화된 정보 서비스 제공에 크게 영향을 줄 수 있다고 판단되는 전공, 학력, 직종 변수를 고객군을 분류하는 예측 변수로 사용하였다.

본 논문에서 사용한 의사결정나무 알고리즘 CHAID는 다지분리를 수행하는 알고리즘으로 결과트리를 가지치기할 필요가 없는 장점이 있다. 형성된 의사결정나무가 거대해지면 해석의 어려움이 존재하게 되므로, 트리의 depth는 3가지 변수를 모두 적용하여 3으로 하였다.

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

목적변수로 사용한 서비스 코드는 의미론적으로 그룹화하여 처리하였다. 전체 200여개의 서비스코드를 목적변수로 사용하면 고객을 분류할 때 각 서비스 코드의 factor가 너무 많으므로 서비스 코드를 의미론적으로 그룹화하여 고객을 분류하였다.

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



(그림 3) 3군의 의사결정트리 결과

의미론적으로 그룹화하여 사용한 서비스를 이용하여 분류 고객들이 자주 사용하는 서비스에 대한 연관관계는 연관규칙 기법을 이용하여 조사하였다.

4. 동적인 정보 서비스

4.1 연관 규칙 기법을 이용한 정보 활용 연관성 분석

이 장에서는 온라인상에서 고객들이 활용하는 서비스와 서비스간의 연관관계를 분석하여, 웹 공간에서 다양한 웹 서비스들에 대한 관심 있는 접근 패턴을 찾아내어 고객 군별 차별화된 정보서비스와 인터페이스를 제공하고자 한다.

3군에 대한 의사결정트리 기법에 따른 분류군은 21개이다. 그 중 다음의 분류군 선택 기준으로 분류군을 선택하여 해당 군의 정보 서비스 이용 연관성을 연관 규칙 기법을 적용하여 분석하였다.

- 활용측면에 따른 분석 결과를 전체 고객에 일반화시킬 수 있는 고객군은 군집 선택 기준과 같이 적정 고객군은 5~10%로 판단하였다.
- 제일 마지막 level에 속하는 고객 군이 3가지 예측 변수를 모두 적용한 결과로 가장 잘 분류된 고객 군이라 판단하고 활용 측면 분석 분류군은 트리의 최대 depth인 3 level에 속하는 분류군을 선택하였다.

의사결정트리 적용 결과인 21개의 분류군에서 위의 조건에 해당하는 3개의 분류군에 대하여 연관규칙 기법을 적용하여 정보 서비스 이용 패턴을 분석하였다.

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

과학기술정보 포털 사이트에서 제공하는 200여개의 서비스 중에서 해당 분류군들의 고객들이 이용한 서비스는 60여개 정도인 것을 알 수 있다. 그중 분류군에 속하는 고객 중 고객 3분의 1 이상이 사용한 서비스에 대해

서만 분석을 하였다. 연관관계도출에서 빈도수가 적은 서비스가 전체 서비스의 40%를 차지하므로 의미 있는 연관관계를 가지는 지지도의 수준을 40%로 설정하였다.

따라서, 서비스 코드가 두개 이상이며 4개 이하인 연관규칙을 채택하였으며, 임계치는 지지도 40%, 신뢰도 50%로 지정하여 SAS Enterprise Miner8.1을 이용하여 연관 규칙을 도출하였다.

<표 3>은 가장 기본적인 서비스에 해당하는 통합검색 서비스에 대해서 3개의 분류군의 연관규칙 기법 적용 결과에서 의미있는 연관 규칙을 추출한 결과이다.

<표 3> 통합검색 연관규칙결과

지지도	신뢰도	연관규칙
75	88	통합검색 == > 국내연구보고서 & 국내학술지 & 해외학술지
77	87.5	통합검색 ==> 국내학술지 & 해외학술지 & 국내회의자료
88	80	통합검색 == > 국내연구보고서 & 국내학술지 & 국내학위논문

세 개의 분류군에서 통합검색의 서비스 연관 패턴 중에서 지지도와 신뢰도가 가장 높은 패턴만 정리한 결과이다. 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 국내연구보고서와 국내학술지 그리고, 국내학위논문 서비스를 함께 소비할 가능성이 가장 높은 것으로 분석할 수 있는데, 전체 트랜잭션에서 통합검색과 국내연구보고서, 국내학술지, 그리고 해외학술지가 함께 활용될 확률은 75%이며, 통합검색이 활용될 때 국내연구보고서, 국내학술지, 해외학술지가 활용될 확률은 88%이다.

따라서, 통합검색 결과를 제공할 때 국내연구보고서와 국내학술지 그리고, 국내학위논문에 대한 서비스를 함께 제공함으로써 이용 고객들의 정보 검색이 편리성과 활용하고자하는 서비스에 대한 depth를 줄여줄 수 있다.

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

키워드와 컨텐트 연관성과 컨텐트들 간의 연관성도 같은 방법으로 분석할 수 있다.

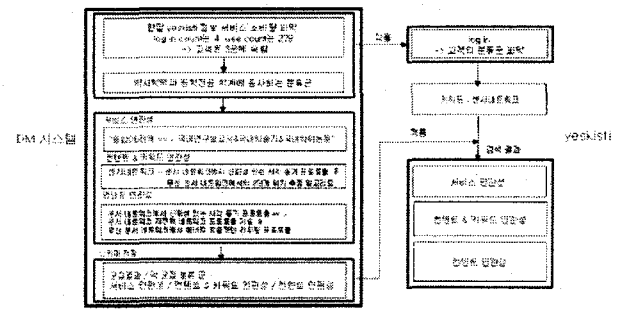
4.2 데이터마이닝 적용 정보 제공

데이터마이닝 적용 정보 제공 정책은 다음과 같다.

- 서비스 연관성 : 고객들이 활용한 서비스 연관성을 기반으로 연관 서비스 제공
- 키워드와 컨텐트 연관성 : 키워드와 소비한 컨텐트를 분석하여 고객이 입력하는 키워드와 연관된 컨텐트 제공
- 컨텐트 연관성 : 고객이 활용하는 컨텐트들 사이의 연관성을 기반으로 연관 컨텐트 제공

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

(그림 4)은 데이터마이닝 시스템과 데이터마이닝 적용 과학기술정보 사이트의 정보 제공 정책 적용 절차를 설명한 것이다.

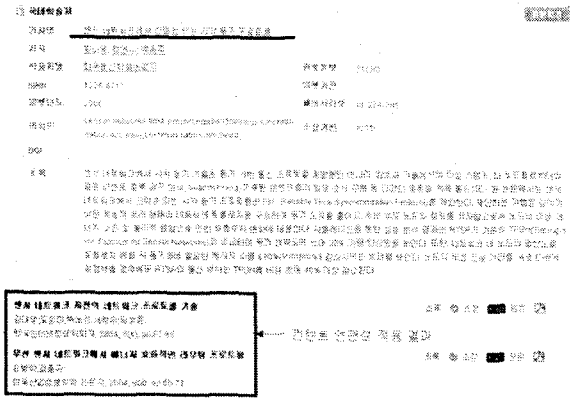


(그림 4) 데이터마이닝 적용 시스템

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

(그림 5)와 (그림 6)는 DM 적용 후의 정보 제공 정책을 적용한 결과이다.

(그림 5) 데이터마이닝적용 통합검색 결과



(그림 6) 데이터마이닝 적용 컨텐트 연관성 제공 결과

서비스 연관관계를 적용하여 통합검색과 함께 많이 활용되는 국내연구보고서, 국내학술지와 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 함께 많이 소비되는 연관성을 가진 컨텐트를 함께 제공할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 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 연관 규칙 기법을 적용하여 파악하고 차별화된 정보와 인터페이스를 제공할 수 있는 통합 모형을 제안하였다.

통합 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 컨텐트를 제공함으로써 추가적인 정보 제공과 컨텐트 선택의 depth를 줄일 수 있다. 또한, 소비 컨텐트와 함께 많이 소비된 컨텐트를 제공함으로써 추가적인 컨텐트 정보 제공과 함께 연관 컨텐트 소비를 유도하여 고객이 관련 컨텐트를 찾기 위한 depth를 줄일 수 있다.

정보이용자에게 적합한 웹 문서를 예측하여 추천해주는 시스템은 고객의 입장에서는 원하는 정보를 쉽게 찾을 수 있도록 하고, 웹 사이트 운영자의 입장에서는 불필요한 요청을 줄임으로서 서버의 부하를 줄여줄 수 있다.

본 논문에서는 고객의 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] Bamshad Mobasher, 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

[2] 장형진, 회성, 한정란, 이기인, "데이터마이닝을 이용한 eCRM", 정보처리학회보, 제 8권 제 6호 (2001.11)

[3] Michael J. A Berry, and Gorden Linoff, Data Mining Techniques : For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., 1997.

[4] Rakesh Agrawal and John C.Shafer, "Parallel Mining of Association Rules," IEEE Transactions on Knowledge and Data Engineering, Vol.8, No. 6, pp. 962-969, December 1996.