사용자의 선호도를 반영하는 영화추천시스템의 개발†

이세호[a], 이강은[a], 황옥삼[a], 노상욱[a]

Developing Movie Recommendation System Reflecting Movie Viewers' Preferences

S. Lee[a], G. Lee[a], O. Hwang[a], S. Noh[a]

[a]School of Computer Science and Engineering, The Catholic University of Korea
Yeokgok 2-Dong, Wonmi-Gu, Buchon-si, 420-743, Korea
Tel: +82-2-2164-4579, E-mail: sunoh@catholic.ac.kr

[a]가톨릭대학교 컴퓨터정보공학부
부천시 원미구 역곡2동, 420-743
Tel: +82-2-2164-4579, E-mail: cis@catholic.ac.kr

Abstract

기존의 영화정보제공 시스템에서는 사용자에게 영화에 대한 정보를 전달할 때 단순히 새로운 영화에 대한 정보를 전달하는데 그치고 있다. 이러한 정보시스템은 사용자에 대한 기호나 성향을 고려하지 않기 때문에, 사용자에게 필요하고 적절한 정보를 제공하지 못하는 문제점이 있다. 따라서, 본 논문은 정보 제공의 효율성을 높이기 위하여 사용자의 영화 선호도가 반영된 영화추천시스템을 설계 및 구현한다. 다양한 사용자로부터 수집한 기본정보에 데이터 분류도구를 적용하여 사용자에 대한 일정한 기호 또는 성향을 추출한다. 결과적으로 추출된 정보를 대상 사용자들에게 SMS로 제공하여 각자의 기호나 성향을 고려한 정보를 얻을 수 있도록 한다.

Keywords:
사용자 선호도; 데이터 마이닝; 영화정보제공 시스템

I. 서론

최근 주5일 근무제로 인하여 여가에 대한 사람들의 관심이 증가하고 있다. 특히, 영화를 보면서 개인의 여가생활을 즐기는 사람들이 날로 증가하고 있는 추세이다. 영화진흥위원회 (http://www.kofic.or.kr) 에 따르면 작년도 서울 관객 수만 40,089,040명으로 전년도(2006년 대비)에 비해 16.6% 증가된 수치를 보여주고 있다. 이러한 증가 추세에도 불구하고, 영화 상영관들의 고객에 대한 서비스나 홍보 등은 아직도 일방적이고 획일적인 방법에서 벗어나지 못하고 있다. 현재 영화 홍보는 TV광고에 의존하고 있으며, 고객이 자신이 원하는 정보를 찾기 위해서는 홈페이지 등을 검색하는 등 고객이 자신의 시간과 노력을 투자하여야만 원하는 정보를 찾을 수 있다. 이런 한계점을 극복하기 위하여 자신의 노력이나 시간을 투자하지 않고 자신이 원하는 취향과 성향에 맞는 맞춤정보서비스를 추천해주는 기술이 요구된다.

본 논문에서는 특정한 영화시스템 안에서 고객의 최소한의 정보를 데이터베이스에 입력하며, 입력한 정보로부터 개인의 선호도를 추출하는 기법을 제안한다. 이러한 기법을 통하여 사용자가 원하는 서비스를 지능적으로 추천 받을 수 있도록 할 것 이다.

II. 관련연구

사용자의 대한 최소한의 정보로부터 사용자의 영화에 대한 선호도를 추출하기 위하여 기계학습

† 본 논문은 2007년도 컴퓨터정보공학부 학부특성화 사업의 지원으로 이루어졌음.

알고리즘을 적용한다. 본 논문에서는 기계학습 알고리즘으로 결정트리 방식, 베이지안 분류 방식을 활용한다.

조건에 따라 데이터를 분할-정복(divide and conquer) 하는 방식으로 자료를 분류하여 특정한 개념 클래스를 얻을 수 있도록 만든 트리를 결정 트리(decision tree)라하며, 속성집합으로부터 결정 트리를 생성하는 알고리즘을 결정 트리 알고리즘이라 한다. 결정 트리 알고리즘은 빠르고 규칙으로의 전환이 쉬우며 이해하기가 쉬운 것이 특징이다 [2][3]. 이 기법은 현재 상태에서 어떤 속성 하나를 선정하여 데이터를 분류하였을 때 예측할 수 있는 엔트로피(entropy)와 정보 이득(information gain)에 따라 속성의 우선순위를 결정하여 트리를 생성하며, 결과적으로 얻은 트리의 형태로 각각의 이용자에 적합한 선호도의 지식을 축적하고자 한다. 나무구조 형태의 표현은 결과적으로 획득한 지식을 이해하기 쉬운 장점을 가진다.

다른 모든 분류기와 비교하면 베이지안 분류기(Bayesian classifier)는 최소 오류 율을 갖는다. 그러나 실제로는 클래스 조건 독립성이라는 가정에 따른 부정확성과 가용 확률 데이터의 부족으로 인하여 항상 최대의 정확도를 나타내지는 않으며, 정확도를 비교하기 위하여 적용될 수 있다. 결정트리와 신경망 분류기에 비교하면 일반적인 응용 도메인에서는 대등한 결과를 제공하며, 베이지안 이론을 쓰지 않는 다른 분류기와 달리 이론적 근거를 제공하는 면에서 유용하다. 본 논문에서는 영화선호도의 추출을 위하여 단순 베이지안 분류를 적용하며, 설문조사 자료로부터 어떠한 규칙성이 존재하는가를 알아보고자 한다.

콘텐츠 제작을 위해서는 무선인터넷 플랫폼을 갖추어야 하는데 J2ME Wireless Toolkit [14]은 무선 소프트웨어 개발을 쉽게 시작할 수 있게 간결하고 직관적인 이용자 인터페이스를 갖은 최고의 환경이며, 단말기로 다운로드 되어 수행되는 응용프로그램(Application)을 PC의 윈도우환경에서 개발하는 도구로서 에뮬레이터 환경에서 좀 더 신뢰성 있는 결과를 추론하기 위해 사용하게 된다. 실제 이용자에게 편의를 제공하기 위하여 무선인터넷 플랫폼으로 사용자의 선호도를 전송하게 된다.

III. 지능형 영화추천 시스템

본 논문에서 제안하는 지능형 영화추천 시스템의 구조는 그림 1과 같다. 우선 이용자의 개인 프로파일정보는 설문지를 통하여 수집된다. 수집된 정보는 데이터베이스 인터페이스를 통하여 관련 테이블에 입력한다. 저장된 정보로부터 기계학습 알고리즘을 통하여 사용자의 선호도 조건을 추출한다. 결과적으로 추출한 영화 선호도는 개인 정보의 특정한 조건에 맞는 이용자들에게 모바일SMS를 통하여 제공할 것이다.

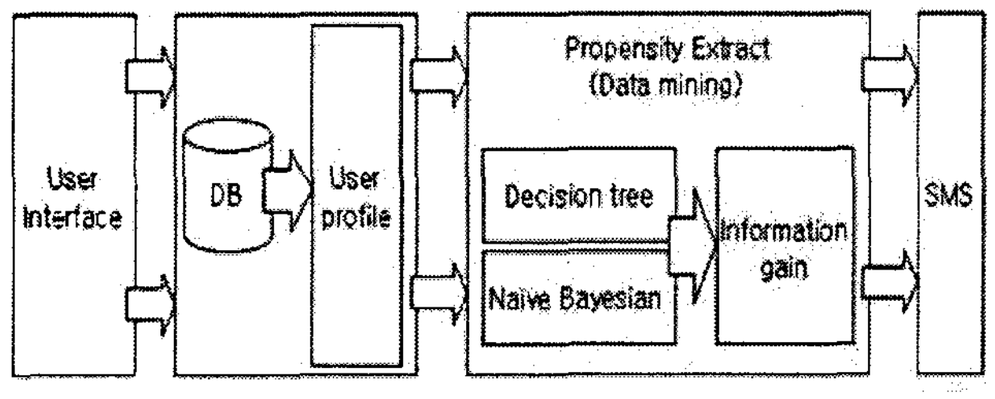


그림 1- 지능형 영화추천 시스템 구조도

1. 사용자 인터페이스

이용자가 프로그램과 직접적으로 대화하는 사용자 인터페이스(User Interface)와 데이터베이스에 대한 정보를 저장할 메타정보를 설계한다. 그림 2는 이용자 인터페이스의 메뉴를 도식화한 것이다.

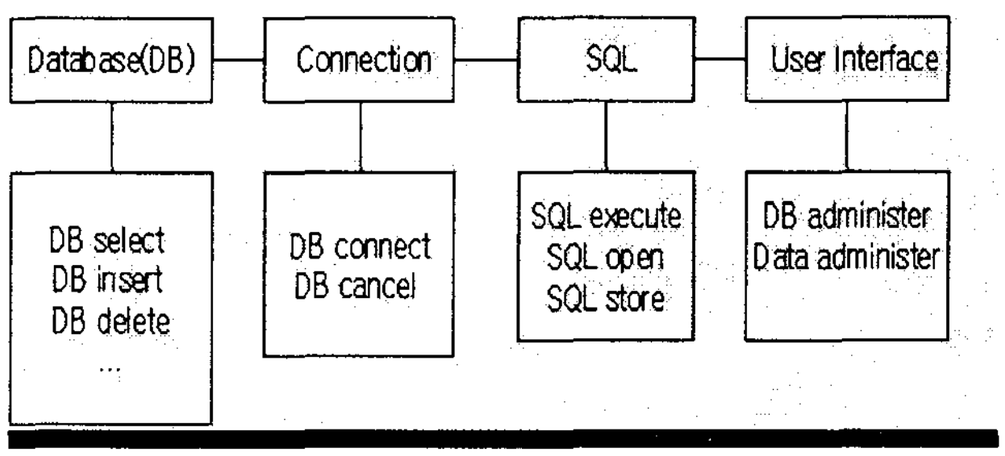


그림 2- 사용자 인터페이스에 대한 상세도

데이터베이스에 접속한 후, 데이터베이스에 존재하는 테이블 이름과 속성을 보여주는 기능이 필요하다. 테이블 이름을 관리자들이 이해하기 쉽게 보여주며, 테이블의 모든 튜플(tuple)을 나열한 사용자 인터페이스 화면을 그림 3에 나타낸다. 그림 3의 좌측에 표시된 트리 부분은 속성들의 계층구조를 정의한 온톨로지(ontology) 생성 부분을 표시한다. 테이블을 선택하면 테이블의 속성을 보여주며, 접속된 데이터베이스로부터 주어진 질의(query)에 대한 결과 값을 그림 3의 우측 하단 부분에 보여준다.

그림 3은 이용자인터페이스의 기능을 나타낸 것이다. 왼쪽의 트리 부분은 현재 선택된 데이터베이스에 저장된 테이블과 속성을 보여주도록

한 것이고, 오른쪽 부분은 수집된 정보를 데이터베이스에 손쉽게 입력하여, 각 테이블 대한 데이터와 속성을 보여주는 인터페이스를 나타낸다. 즉, 여러 데이터베이스 노드 중 하나의 노드를 마우스로 선택하면 선택한 데이터베이스 테이블에 대한 수집된 데이터정보가 오른쪽 테이블에 표시되도록 사용자 인터페이스를 구성하였다.

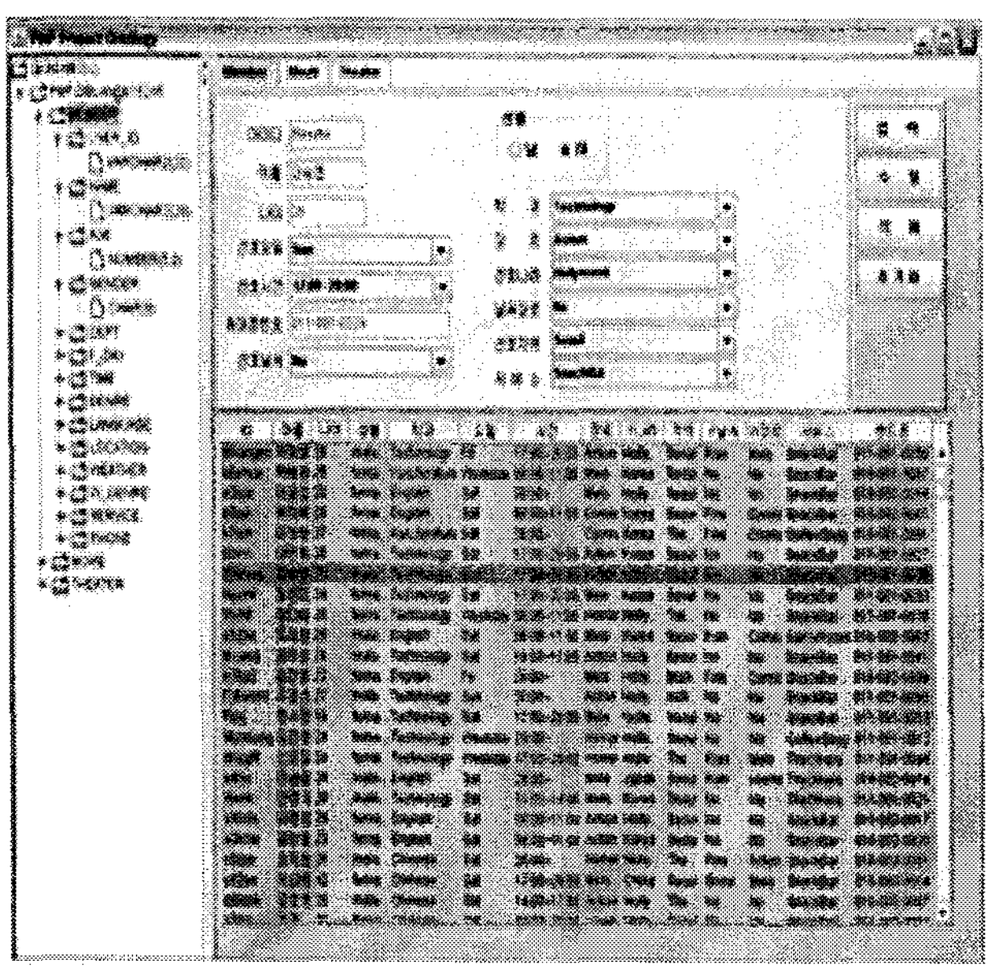


그림 3- 사용자 인터페이스

2. 선호도를 반영한 추론방식

본 논문에서는 각 개인에 대한 프로파일 정보는 설문지로 정보를 수집하였다. 설문지로부터 획득된 이용자에 대한 정보를 직접 입력하였으며, 이러한 정보는 성별, 연령, 학과, 지역, 선호요일, 선호시간, 선호장르, 선호국가, 날씨에 따른 영화장르, 서비스 정보 등으로 구성된다. 기계학습 알고리즘은 공개 소프트웨어인 WEKA(Waikato Environment for Knowledge Analysis) [2]를 사용하였으며, 사용자 인터페이스를 통하여 입력 받은 데이터는 WEKA의 파일양식인 .arff로 변환하였다. 설문지를 통하여 획득한 정보에서 사용자에 대한 영화 선호도를 추출하기 위하여 WEKA에서 제공하는 기계학습 알고리즘으로 C4.5 결정 트리 알고리즘과 단순 베이지안 분류기를 사용하였다. 또한, 클러스터(cluster) 기능을 이용하여 수집정보의 군집화 부분을 가시적으로 표현한다.

단순 베이지안 분류기의 결과로 도출된 확률 값만으로 직접 어떤 영화를 추천하는 것은 어렵다. 따라서 본 논문에서는 추론의 결과로 얻은 각 속성의 확률로부터 고객의 선호도에 맞는 추천 속성을 계산한다. 영화장르는 액션, 멜로, 호러, 코믹의 4개의 속성으로 낸다.

다음 표는 베이지안 분류에 따른 각 속성들의 확률 값을 나타낸 것이다.

표-1 에 각 장르의 확률

Genre			
Action	**Melo**	**Horror**	**Comic**
0.282	0.387	0.134	0.197

표-2 성별에 따른 장르의 확률

Gender				
장르 / 성별	**Action**	**Melo**	**Horror**	**Comic**
Male	0.494	0.180	0.333	0.456
Female	0.506	0.820	0.667	0.544

표-3 학과에 따른 장르의 확률

Department				
장르 / 학과	**Action**	**Melo**	**Horror**	**Comic**
Liberal Arts	0.202	0.143	0.213	0.338
⋮	⋮	⋮	⋮	⋮
Nursing	0.101	0.160	0.106	0.046

표-4 나이에 따른 장르의 평균 및 표준편차

Age				
장르 / 수치	**Action**	**Melo**	**Horror**	**Comic**
Mean	23.582	22.908	23.378	23.491
Standard	3.245	2.879	2.520	2.486

표-5 지역에 따른 장르의 확률

Location				
장르 / 지역	**Action**	**Melo**	**Horror**	**Comic**
Seoul	0.530	0.531	0.561	0.525
Incheon	0.096	0.133	0.146	0.068
Bucheon	0.205	0.204	0.073	0.254
TheOthers	0.169	0.133	0.220	0.153

표-6 실험 데이터

Dept	**Location**	**Age**	**Gender**	**Genre**
Nursing	Seoul	24	Female	?

위의 표-6의 실험데이터를 적용한 결과는 다음과 같다.

P(Action|Nursing,Seoul,24,female)=

0.101 x 0.530 x 0.551 x 0.506 x 0.282 =0.004

P(Melo|Nursing,Seoul,24,female)=

0.160 x 0.531 x 0.648 x 0.820 x 0.387=0.017

P(Horror|Nursing,Seoul,24,female)=

0.106 x 0.561 x 0.597 x 0.667 x 0.134=0.003

P(Comic|Nursing,Seoul,24,female)=

0.046 x 0.525 x 0.581 x 0.544 x 0.197=0.002

장르에 대한 확률 값을 적용한 결과 멜로의 값이 다른 값들 보다 높으므로 실험데이터의 경우 "멜로" 영화를 결정하여 추천하게 된다.

IV. 실험 결과

설문조사는 가톨릭대학교에 재학중인 학생들을 대상으로 한 달 (2007년 4월1일~5월1일)동안 수집된 데이터이며, 이용자의 정보로 성별, 연령, 학과, 지역, 선호요일, 선호시간, 선호장르, 선호국가, 날씨에 따른 장르, 서비스 정보 등을 수집하였다.

실험의 목적은 영화관 이용객의 패턴과 선호도를 분석하여, 이용객의 선호현황 패턴에 맞추어 이용객에게 원하는 정보를 지능적으로 추천하고자 하는 것이다.

실험을 위하여 설문조사 결과를 영화추천 시스템의 데이터베이스에 저장하였다. 데이터베이스에 저장한 테이블은 member, movie, theater로 구성되며, 그림 3에 나타난 사용자 인터페이스를 활용하여 사용자에 대한 정보는 member 테이블에 저장하였으며, 영화에 대한 정보는 movie 테이블에, 영화관에 대한 정보는 theater 테이블에 각각 저장하였다.

member테이블에 수집된 이용자정보를 귀납학습 알고리즘에 입력하여 사용자에 대한 선호도를 조사하였다. 그림 4에 나타난 바와 같이 영화 사용자의 선호도 특징은 남자는 액션, 여자는 멜로를 선호하였고, 선호요일은 토요일이 55.71%, 선호시간은 16:00-20:00 시간대가 42.14%, 장르는 멜로 38.93%, 선호국가는 미국 52.86%, 날씨영향장르는 영향을 받지 않는다 67.86%, 서비스는 매점 72.86%, 선호지역은 서울 55.00%를 나타냈었다.

이와 같은 선호도의 추출결과로부터 다음과 같은 특징을 정리할 수 있었다. 선호요일과 선호시간, 선호국가, 선호서비스부분에서는 압도적으로 토요일, 16:00-20:00시간대, 미국, 매점을 선호하였고, 날씨에 영향을 많이 받을 것이라고 추측하였지만 날씨의 영향은 많이 받지 않았고, 남자는 액션을 여자는 멜로를 선호하는 결과가 나왔다. 그림 4는 설문조사 결과 나타난 원천자료의 분포를 나타내며, 결정트리 학습 알고리즘을 이용하여 선호도를 추출한 결과는 그림 5와 같다.

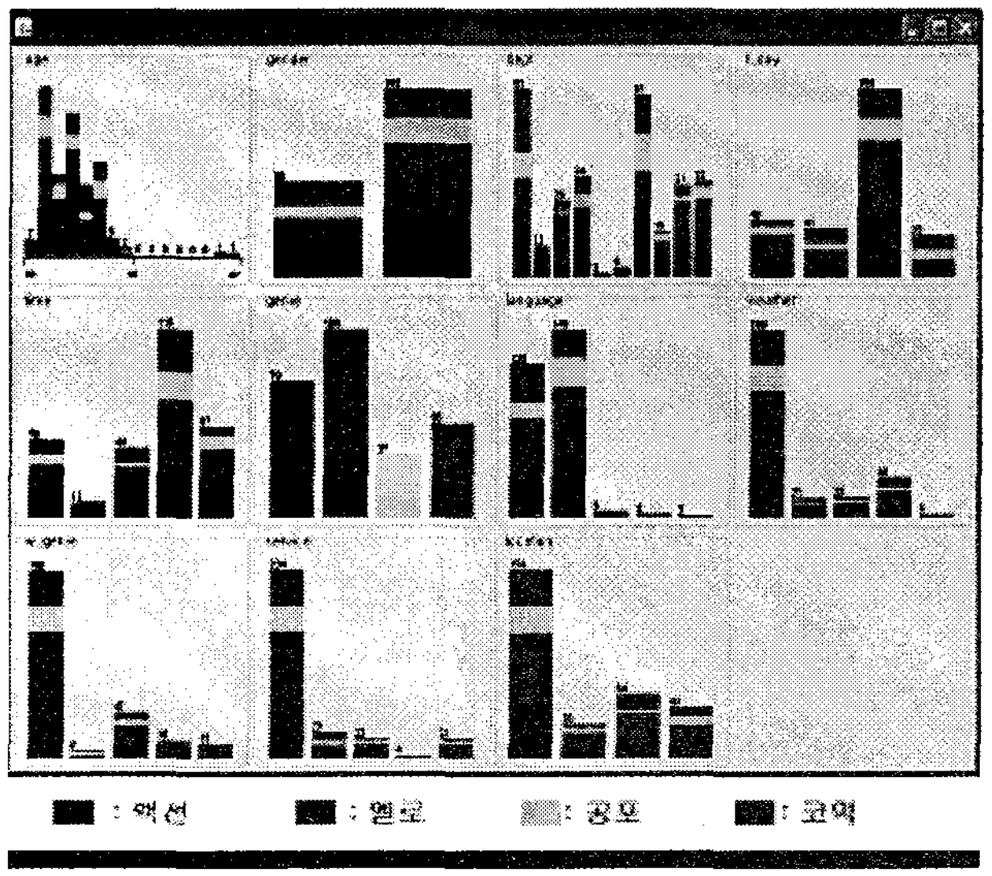


그림 4- 설문조사 결과 수집된 자료의 분포

그림 5는 영화 사용자의 성별, 나이, 학과, 선호위치, 선호장르 5가지 속성에 대한 분포를 나타내며, 그림5는 장르에 따른 분류의 선호도를 나타내는 결정트리이다.

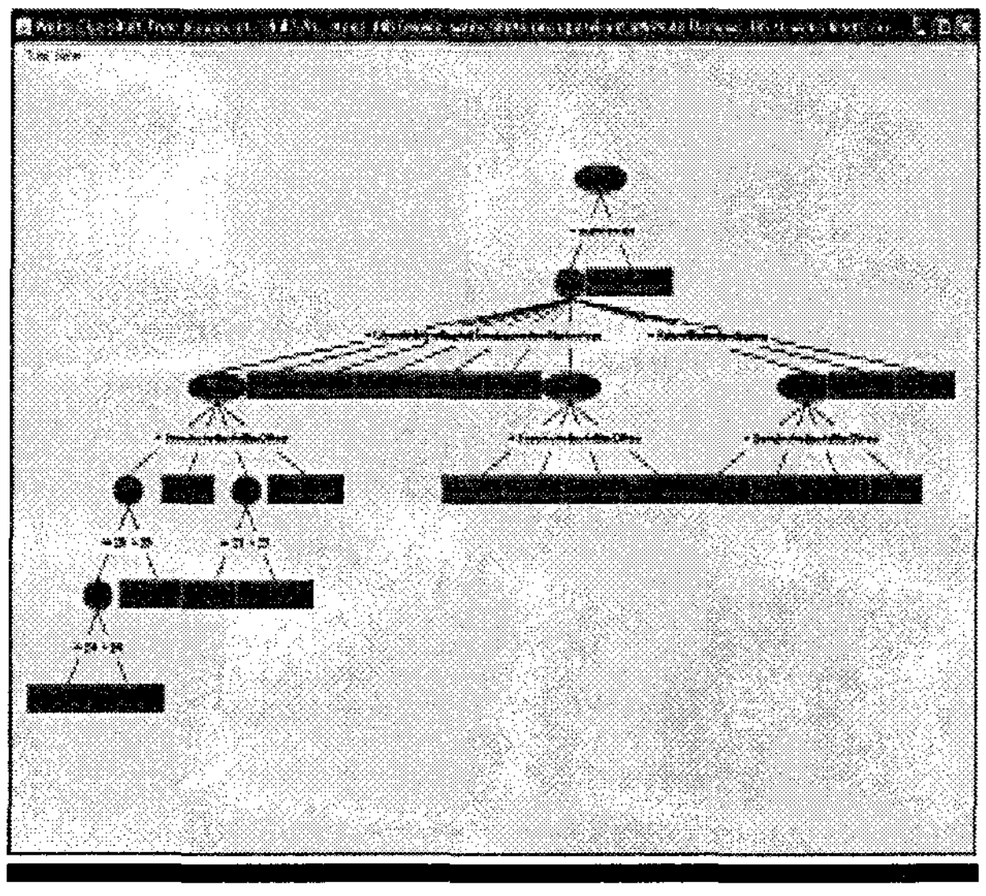
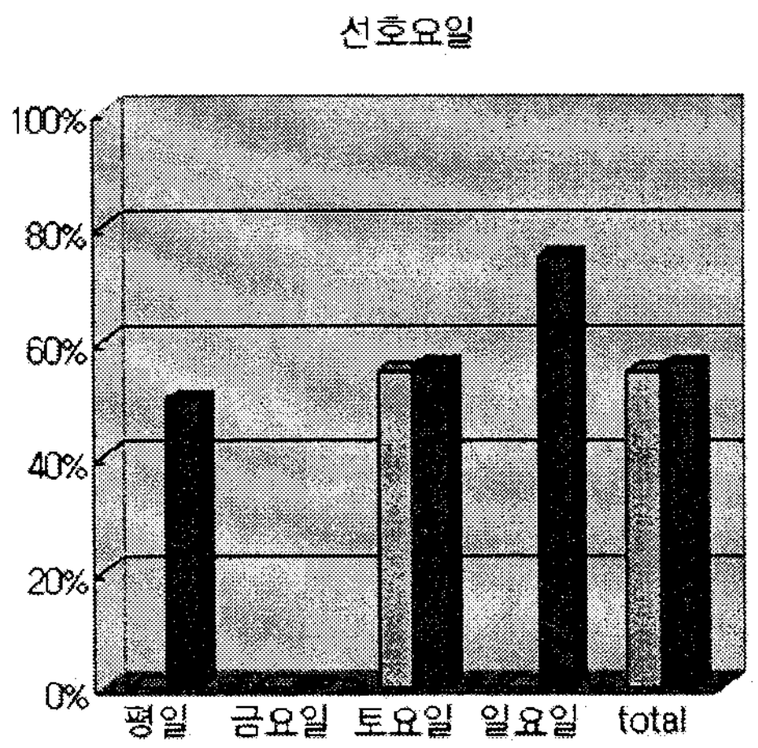


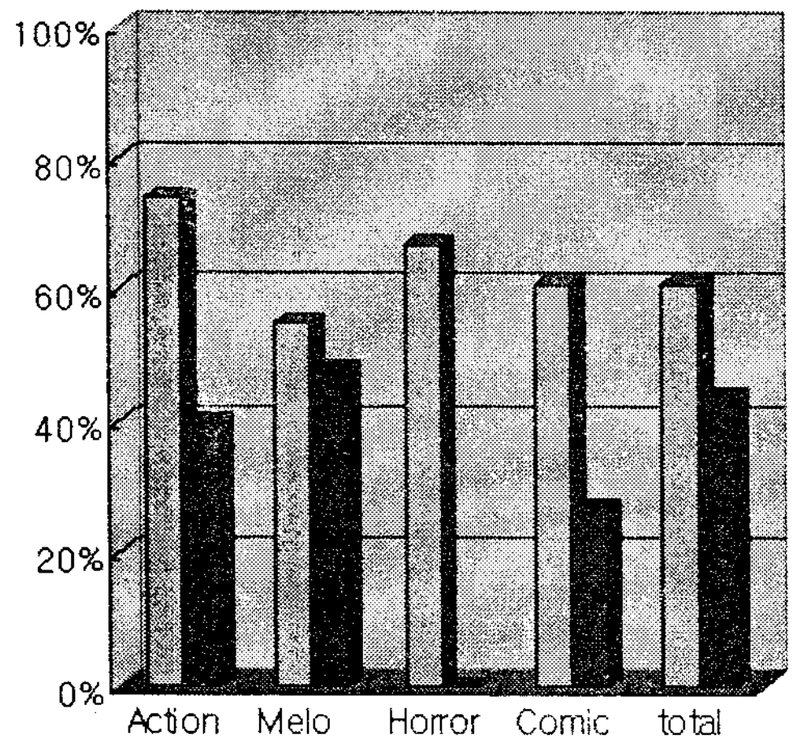
그림 5- 영화 사용자의 선호도를 나타내는 결정트리

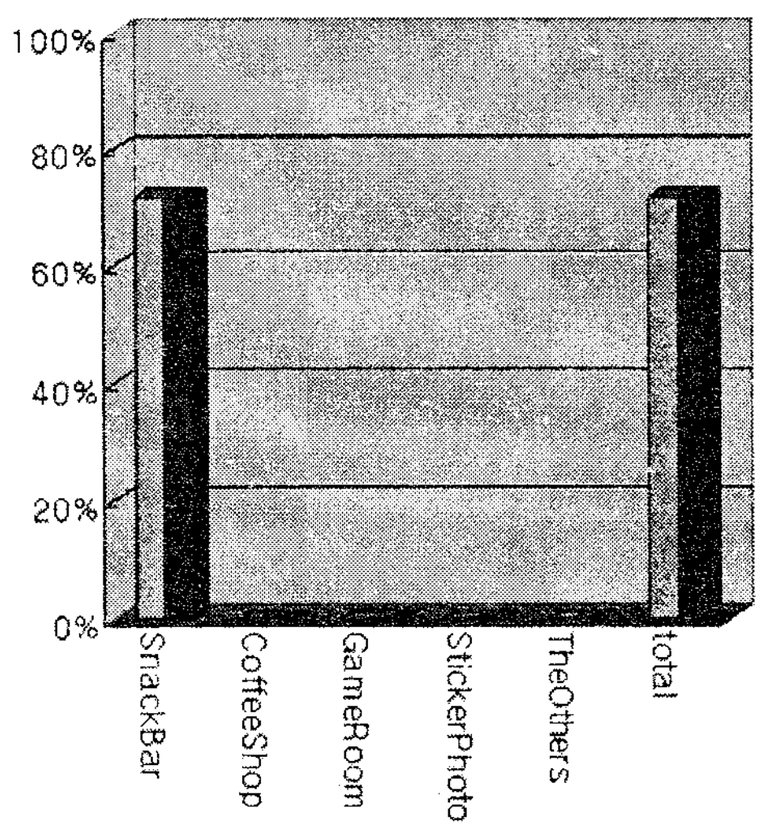
그림 5에서 장르에 의한 추천된 값에 대한 결정트리의 예를 들면, 생성된 하나의 규칙은 다음과 같이 나타낼 수 있다.

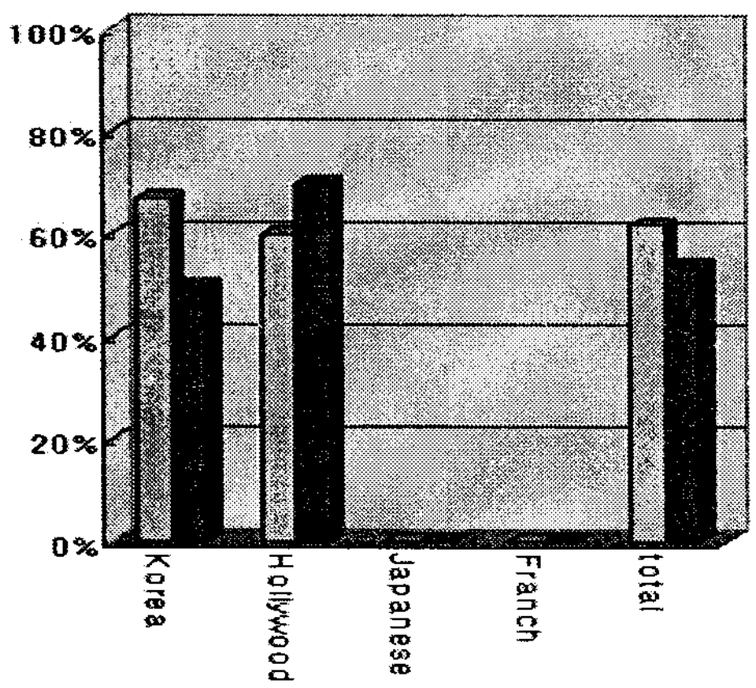
If gender = male
and dept = LiberalArts
and location = Seoul
and age = <= 24
then recommendation = Action

이와 같은 방법을 통해서 각각의 데이터는 해당장르로 분류 된다.









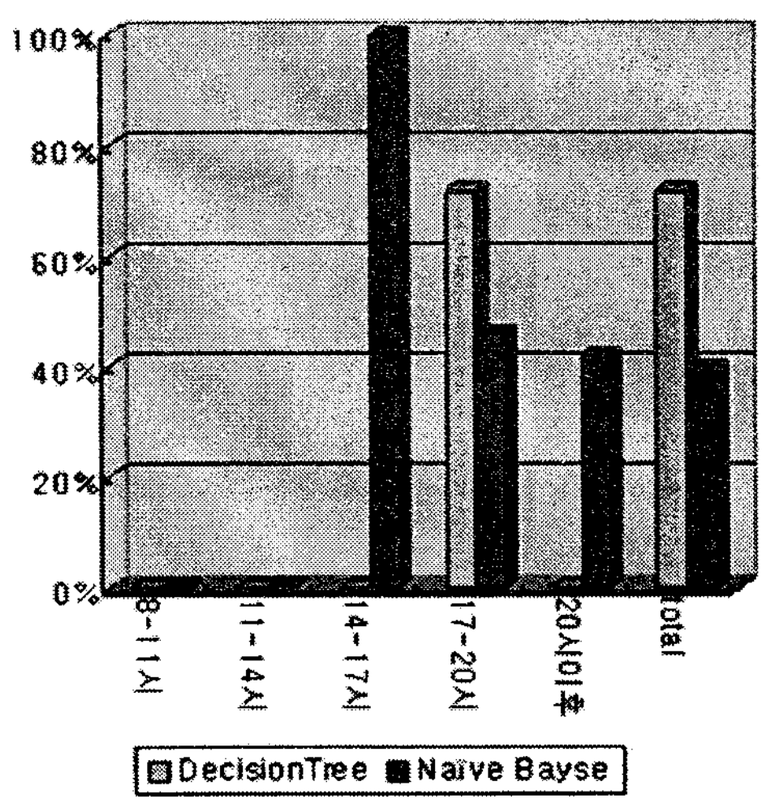


그림 6- 결정트리와 베이지안 분류기의 영화 선호도 추출에 대한 정확도

위의 질의조건으로 수행한 결과 장르별 결정트리 적중률은 액션: (58/78) 74%, 멜로: (107/194) 55%, 공포: (2/3) 67%, 코믹: (8/13) 61% 이었으며, 장르별 베이지안 분류기를 이용한 정확도는 액션: (22/54) 40%, 멜로: (88/182) 48%, 공포: (측정불가) 0%, 코믹: (12/44) 27%의 정확도의 결과가 나왔다.

선호국가별 결정트리의 적중률은 한국: (79/159) 67%, 미국: (107/178) 60%, 일본: (측정불가) 0%, 프랑스: (측정불가) 0% 결과가 나왔으며, 선호국가별 베이지안 분류기를 이용한 정확도는 한국: (79/159) 50%, 미국: (72/120) 70%, 일본: (0/1) 0%, 프랑스: (측정불가) 0%의 결과가 나왔다.

선호요일 별 결정트리의 적중률은 평일, 금요일, 일요일은 측정불가였으며, 토요일은 (156/280) 55% 결과가 나왔고, 베이지안 분류기의 정확도는 평일: (1/2) 50%, 금요일: (측정불가) 0%, 토요일: (155/274) 56%, 일요일: (3/4) 75% 결과가 나왔다.

부가서비스 별 결정트리의 적중률은 SnackBar: (204/280) 72%, 베이지안 분류기의 정확도는 SnackBar: (204/280) 72%의 결과가 나왔다.

시간 별 결정트리의 적중률은 17-20시: (204/280) 72%, 베이지안 분류기의 정확도는 14-17시: (1/1) 100%, 17-20시: (103/224) 46%, 20시 이후: (8/19) 42%의 결과가 나왔다.

영화 선호도에 대한 추천 결과를 종합하면, 전체장르별 적중률은 각각 61%, 44%, 전체선호국가별 적중률은 각각 62%, 54%, 전체선호요일 별 적중률은 각각 55%, 56%, 전체부가서비스 별 적중률은 각각 72%, 72%, 전체시간 별 적중률은 각각 72%, 40%의 결정트리와 베이지안 분류기의 정확도 결과가 나왔다.

그림 7은 군집 분류기법을 적용하여 얻은 선호장르와 학부에 따른 장르선택 패턴의 자료유형과 군집상태를 나타낸다. 객체들의 군집은 한 군집 내의 객체들은 서로 높은 유사성을 지니지만 다른 군집에 있는 객체들과는 매우 상이하게 형성된다. 그림 7의 x축은 선호장르를 나타내며, y축은 학부를 나타낸다.

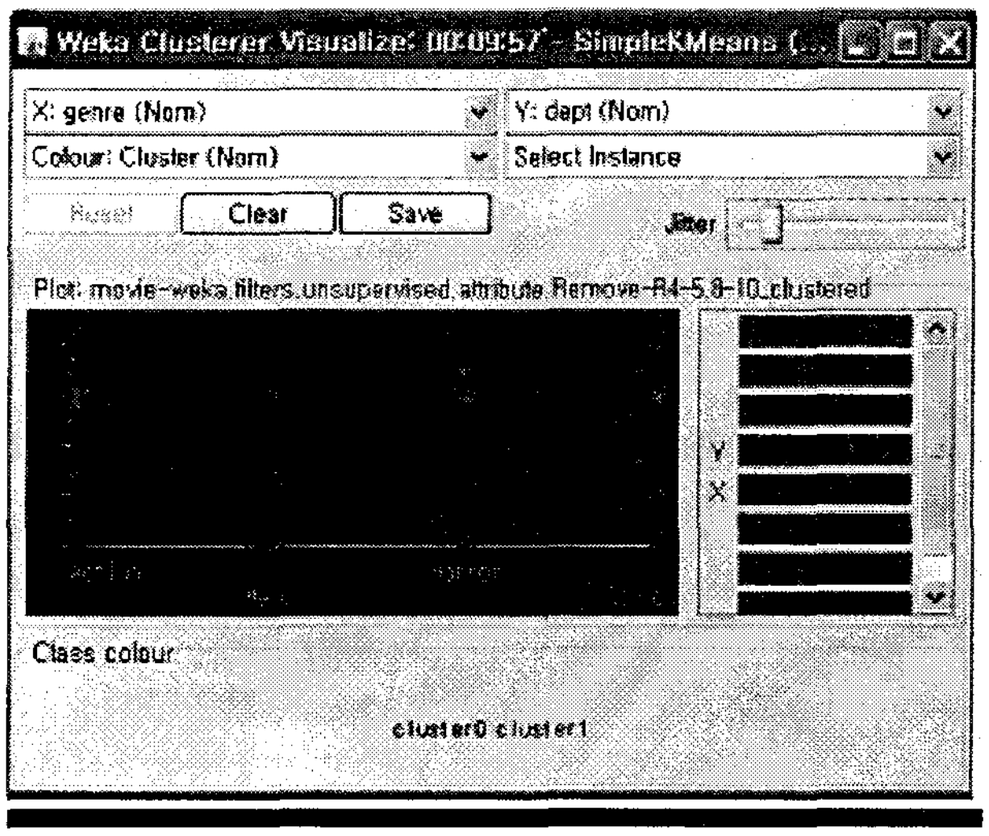


그림 7- 영화장르와 학과에 대한 군집 결과

특정한 사용자에 맞는 영화선호도는 Java 2 Micro Edition (J2ME)을 활용하여 사용자의 휴대전화나 PDA로 전송된다. 클러스터링과 적중률을 이용하여 장르별 선호도는 결정트리, 국가별선호도는 베이지안 분류기를 이용하여 추출한 결과를 모바일 SMS인 J2ME Wireless Toolkit을 이용하여 전송하였다. 획득한 수집정보를 가지고 새로운 회원에 대하여 추천하는 내용을 핸드폰 화면으로 개개인으로 보내 줄 수 있다. 그림 8은 각 회원에게 추출된 영화 선호도에 대한 정보를 새로운 회원에게 핸드폰을 통하여 추천하는 화면을 나타낸다.

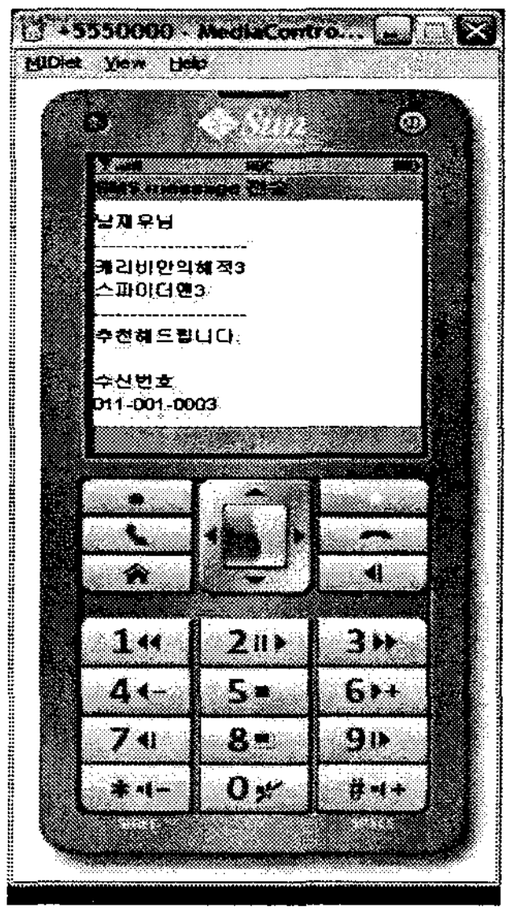


그림 8- 영화선호도를 핸드폰으로 전송하는 화면

V. 결론

본 논문은 사용자의 영화에 대한 선호도가 반영된 영화추천시스템을 설계 및 구현하였다. 영화추천시스템은 설문조사를 통하여 획득한 정보로부터 추출한 영화 선호도에 대한 정보를 새로운 사용자에게 제공할 수 있는 유연성을 가진다. 실험을 통하여 결정트리와 베이지안 분류기의 특징 추출에 대한 정확도를 비교할 수 있었으며, 생성한 규칙을 통하여 영화 사용자의 선호도에 미치는 속성을 확인할 수 있었다. 기계학습 알고리즘을 이용하여 컴파일 된 규칙의 유용성을 증명하였으며, 계속적으로 다양한 추천 시스템에 적용시켜 나갈 것이다.

VI. 참고문헌

[1] Morgan Kaufmann, (1993). "Quinlan, J. R., C4.5: Programs for machine learning"

[2] I. H. Witten and E. Frank, (2005). "Data Mining: Practical machine learning tools and techniques. 2nd Edition, San Francisco: Morgan Kaufmann Publishers"

[3] Han kamber. "Data Mining concepts and Techniques"

[4] Jiwei Han ,Micheline Kamber[공]저, 박우창[외]역 "데이터마이닝:개념 및 기법"

[5] Olivia Parr Rud,敎友社, (2003). "데이터 마이닝 Cookbook"

[6] 학위논문(석사) 이다운 (2004). "데이터마이닝 분류기법으로서의 베이지안 망에 관한 연구

= Bayesian Network classifiers in Data Mining"

[7] Pearl, J, (1994). "Bayesian Networks.. In Handbook of Brain Theory and Neural Networks, MIT Press."

[8] Schwarz, G. (1978). "Estimating the Dimension of a Model. The Annals of Statistics, 6" pp.461-464

[9] Wang, K. (2002). "UCI repository of machine learning data base"

[10]Boca Raton, Fla.: Chapman& Hall/CRC,(2004). "Bayesian Data Analysis"

[11]Chi, Albert Yu-Ming "The bayesian analysis of structural change in linear models"

[12]Raghu Ramakrishnan; Johannes Gehrke, McGraw-Hill Korea,(2003). "데이터베이스 시스템"

[13]임해철, 이석호 (1986). "확장형 베이지안정리를 이용한 추론기법," 정보과학회 봄 학술발표논문집.

[14]정영오, PCBOOK(2000)."모바일 자바PDA 핸드폰 프로그래밍 : J2ME 기반의 무선인터넷 개발"

[15]김성환,피어슨에듀케이션코리아."모바일 자바프로그래밍(J2ME 및 WAP프로그래밍)"