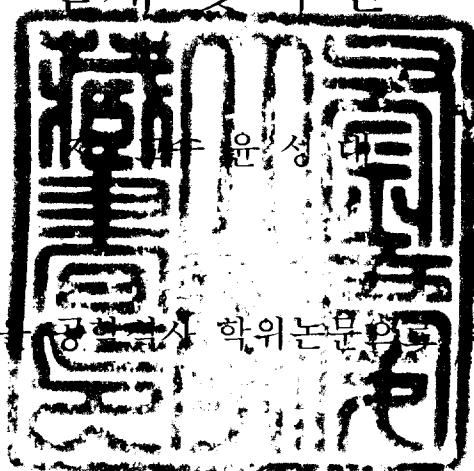


공학석사 학위논문

커뮤니티 점유시간을 이용한
상품추천 시스템의
설계 및 구현



이 논문은 공학석사 학위논문으로 제출함

2005년 8월

부경대학교 산업대학원

전산정보학과

천 소 영

천소영의 공학석사 학위논문을 인준함

2005년 6월 17일

주 심 공학박사 김 창 수



위 원 이학박사 박 홍 복



위 원 이학박사 윤 성 대



<차 례>

표 차례	ii
그림 차례	iii
Abstract	iv
I. 서 론	1
II. 관련연구	4
2.1 통계적인 방법	4
2.2 웹마이닝	9
2.3 웹관련기술(JSP)	12
III. 제안하는 추천 시스템	13
3.1 모니터링 에이전트	15
3.2 분석 에이전트	21
3.3 상품추천 에이전트	23
IV. 시스템 구현 및 성능평가	24
V. 결론 및 향후 과제	31
참고문헌	33

<표 차례>

<표 1> 회원 기본 정보 테이블	16
<표 2> 커뮤니티 가입 정보 테이블	17
<표 3> 커뮤니티 접속 로그 정보 테이블	18
<표 4> 회원 관심 카테고리 테이블	19
<표 5> 커뮤니티 카테고리 대분류 테이블	20
<표 6> 커뮤니티 카테고리 중분류 테이블	20
<표 7> 시스템 구축 환경	24
<표 8> 식 (1) 적용한 결과	25
<표 9> 식 (2) 적용한 결과	26
<표 10> 추천된 상품에 대한 선호도	29
<표 11> Precision 측정 결과	30

<그림 차례>

(그림 1) WebTrends를 사용한 로그 분석	5
(그림 2) JSP Processing	12
(그림 3) 시스템 구성도	14
(그림 4) 카테고리 계층도	21
(그림 5) 로그인시 상품 추천 화면	27

A Design and Implementation of Products Recommendation System Using the Residence Time in Community

SoYoung Cheon

*Dept. of Computer and Information
Pukyung National University*

Abstract

According to the development of the computer network and the expansion of the e-commerce, a product purchasing rate using internet is growing rapidly. In past decade, there are many researches about an efficient product recommendation system.

Now, the content-based recommendation method and the collaborative filtering method are frequently used to recommend a product on internet. Both methods have two drawbacks, the requirement of purchasing history and an omission of customer information in recommendation of a product for a customer.

Therefore, we propose the product recommendation system using the residence time in community to solve problems mentioned above, in this paper.

The experimental result shows that the proposed method is

more efficient than existing methods in the point of the rate satisfaction.

I. 서 론

인터넷 환경의 급속한 발달과 보급으로 인터넷 인구가 폭발적으로 증가함에 따라 전자상거래 시장이 급속하게 성장하고 있다. 또한 전자상거래의 발전은 인터넷의 대중화를 통하여 가속화되고 있으며 이로 인해 전자상거래 분야가 활발히 진행되면서 고객의 유치·관리 및 서비스에 대한 연구가 진행되고 있다. 이러한 전자상거래 시장의 증가에 따른 사용자들의 다양한 요구에 대응하기 위해 개개인의 취향과 최근 관심사에 맞추기 위한 개인화는 전자상거래의 중요한 핵심기술로 떠올랐다. 현재 인터넷의 많은 전자상거래 업체들의 주된 목표는 기업이 가진 많은 상품들을 변화하는 사용자의 관심도에 맞추어 사용자에게 추천함으로써 추천 상품의 히트율을 증가시켜 고객의 만족도와 기업의 이익을 극대화하는 것이다.

쇼핑몰 등에서 자신에게 맞는 상품들을 추천 받는 서비스는 이미 해외 시장에서는 어느 정도 일반적인 일이 되었으며, 국내에서도 이미 그 자리가 차지하고 있는 실정이다. 철저하게 고객의 필요와 욕구에 밀착된 인터넷 서비스를 제공한다는 개인화(Personalize)의 취지에 비추어 볼 때, 이와 같은 추천 시스템은 필수적으로 입지를 굳힐 것으로 전망된다. 이런 상황이라면 고객에게 어느 만큼의 개인화 서비스를 제공하느냐가 앞으로 쇼핑몰의 성패를 좌우하게 될 것임에 틀림없다.

개인화 서비스란, 무분별하고 과다하게 제공되는 정보들에 대해 여과 기술을 이용하여 정제함으로써 사용자 개인의 취향에 알맞은 정보를 제공하는 서비스를 말한다. 이러한 개인화 서비스는 추천의 정확도에 있어

서 한계를 보이고 있다는 문제점이 드러나고 있다.

또한 사용자들 간의 유사도를 계산하는데 있어서 항목간의 중요도, 즉 가중치를 고려하지 못하는 문제점을 지니고 있을 뿐만 아니라 공통된 선호도 항목이 적은 경우 발생하는 희소성(sparsity)문제와 사용자의 수가 증가함에 따라 수행속도가 저하되는 범위성(scalability)문제로 인하여 효율성이 떨어지는 문제점들을 지니게 된다. 이러한 문제점들에 대해 현재 까지 고객이 관심을 보일만한 것들을 추천하는 많은 연구가 이루어져 왔으며 그 중 대표적인 것은 내용기반(Content based) 추천과 협업(Collaboration)을 통한 추천 등이 있다.

내용기반 추천은 정보검색 분야에서 그 기원을 찾을 수 있는 것으로 사용자가 과거에 구매를 하였거나 관심을 보인 상품의 프로파일과 유사한 상품간의 비교를 통하여 추천이 이루어진다. 예를 들면 휴대전화를 많이 구입한 고객인 경우 새로 나온 휴대전화를 보여주고 책을 많이 구매한 고객에게는 책을 추천하는 방식이다. 이 방식은 과거에 고객이 구매한 내역을 바탕으로 추천이 이루어지므로 각 개인의 구매히스토리를 필요로 하고 고객에게 제한된 추천이 이루어 질 수 있는 문제점이 있다[1,2].

협업을 통한 추천 방식은 고객이 구매한 상품들을 비교하여 다른 고객의 구매 패턴과의 유사도를 측정하여 유사한 구매를 보인 다른 고객이 많이 구매한 상품들을 추천하는 방식이다. 예를 들어 고객 A와 B가 있는데 고객 A는 DVD와 책을 구매했고 고객 B는 DVD와 CD만 구매했다고 하고 두 고객 간의 유사도가 시스템에서 정한 임계치 이상이 된다면 고객 A에게는 CD를 추천하고 고객 B에게는 책을 추천하는 방식이다. 이 방식은 시스템 구축 초기에 일정량 이상의 고객 데이터가 축적되어야 추천을 시작할 수 있다는 단점과 고객이나 상품이 추가될 경우 다른 모든 고객과의 유사도를 다시 계산해야 하기 때문에 고객이나 상품이 많아

실 경우 성능에 문제가 생길 수 있다는 단점이 있다[3,4,5,6].

본 논문에서 제안하는 방법은 포털 사이트에 가입한 고객이 과거에 구매 내역이 없더라도, 커뮤니티 페이지에서의 점유시간을 모니터링하고 분석한 뒤, 분석된 내용에 대해 쇼핑몰 카테고리와의 연관성 규칙을 적용하여 클러스터링 한 후, 고객이 가장 활발히 활동하는 커뮤니티 카테고리와 관련이 있는 상품을 고객에게 추천하기 위한 것이다. 이 방법은 고객이 관심을 갖고 있는 커뮤니티가 시간이 지남에 따라 변하더라도 그 변화에 맞춰 유동적인 추천이 가능하고 고객자신의 커뮤니티 접근시간을 데이터마이닝하므로 다른 고객과의 유사도를 고려함으로써 생기는 성능 저하 문제를 줄일 수 있는 장점이 있다.

대부분의 고객이 가입초기에 채택하게 되어 있는 관심분야가 무의미하게 선택되어 진다는 점과 선택을 하더라도 관심카테고리의 우선순위를 측정할 수 없는 부분에 대해 고객이 가입하여 접근하게 되는 커뮤니티페이지의 접근 빈도와 시간을 측정하여 고객이 관심도를 반영하도록 하여 다음 로그인 할 때는 관련된 상품을 추천하도록 한다.

또한 각 커뮤니티와 전자상거래 상품이 카테고리별로 클러스터링 되는 특징을 이용하고, 가장 최근에 가입한 커뮤니티와 각 커뮤니티에 머무는 시간을 마이닝하여 사용자가 다음 로그인시 가장 관심 있는 카테고리의 상품을 추천하는 방법이다.

본 논문 이후의 구성은 다음과 같다. 2장에서는 관련 연구 분야인 통계적인 방법의 분석유형을 살펴보고, 웹 데이터의 패턴 발견을 위한 기법과 웹 관련 기술에 대해 살펴본 후, 3장에서는 제안하는 기법인 커뮤니티에서의 점유시간을 이용한 상품추천 시스템에 대해 기술하고, 4장에서는 시스템 구현에 대해 알아본다. 마지막으로 5장에서는 본 시스템을 적용함으로써 얻게 되는 기대효과를 기술하고 향후 과제를 제시한다.

II. 관련연구

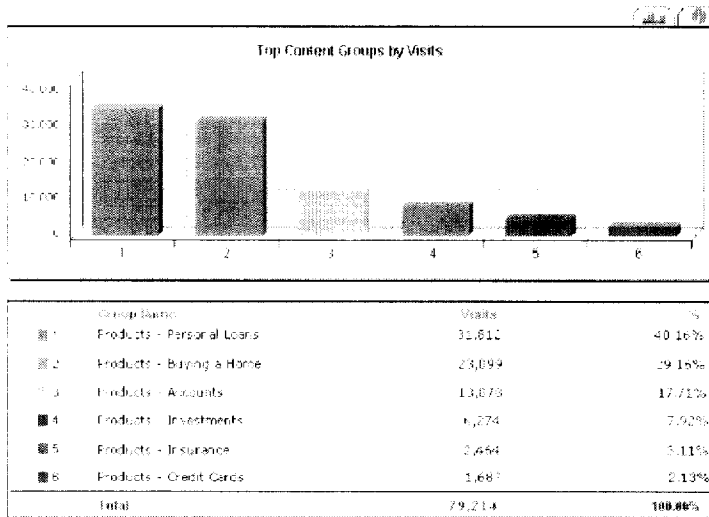
본 장에서는 관련연구로 통계적인 방법과 웹마이닝 그리고 웹관련 기술(JSP)에 관해 살펴본다.

2.1 통계적인 방법

웹 액세스 패턴을 추출하는 가장 단순한 형태는 통계적인 방법들을 이용하는 것이다. 예를 들어, 일정 기간 동안 가장 많이 액세스된 페이지들의 순위를 알아내는 것과 같은 형태이다. 그림 1은 웹로그 분석 툴인 WebTrends라는 툴을 사용하여 고객들이 가장 많이 액세스한 페이지를 분석하여 그래프 형태로 나타낸 것이다.

웹 서버는 사용자들의 접근을 로그파일에 기록한다. 이것은 단순히 접속이라는 데이터로, 모든 방문객과 모든 페이지에 대해 기록한다.

로그파일을 분석해서 알 수 있는 내용은 대부분 로그파일 분석 툴의 기능에 달려있으며, 간혹 분석 툴에 따라서는 보고서의 내용을 정의할 수 있는 경우가 있다. 웹서버 관리자는 결과보고서의 내용을 현재 관리하고 있는 웹서버의 현실에 맞도록 해석해야 한다. 로그파일을 통하여 다음과 같은 분석을 할 수 있다.



(그림 1) WebTrends를 사용한 로그 분석

가. 기간별 분석

방문객의 웹서버에 대한 접속, 방문 및 에러발생회수 등을 기간별 비율과 함께 살펴볼 수 있다. 일별 평균과 최대 접속, 방문 및 에러발생회수 등도 웹서버 관리에 많은 도움을 준다. 이것은 월별, 요일별, 일별 및 시간대별로 분류하면 접속이 가장 집중되는 월, 요일, 일 및 시간대를 한눈에 파악할 수 있다. 예를 들어 주말에 트래픽이 집중되는 웹서버의 경우, 주말 특별대책반 운영도 검토할 수 있다. 특히 뉴스 또는 광고 등을 통해 웹서버의 홍보가 이루어진 날과 평일을 비교하여 볼 수도 있으며, 이것은 뉴스나 광고의 효과측정에 유일한 데이터가 된다.

나. 방문객 분석

웹서버는 사용자, 즉 방문객을 위해 존재한다. 사용자는 등록여부에 따라 일반사용자와 등록사용자로 구분하거나, 또는 접속방법에 따라 모뎀사용자와 전용선사용자로 구분하여 분석할 수 있다.

일반사용자 중 등록되어있는 사용자가 등록사용자이므로, 일반사용자는 등록사용자를 포함하고 있다. 따라서, 일반사용자는 있는 그대로 분석하지만, 등록사용자는 별도의 필터를 거쳐 일반사용자와는 다른 분석을 해야 한다. 만약 유료서비스 또는 회원제 웹서버를 운영하고 있다면, 등록사용자의 사용시간, 사용 데이터량 등이 차후 종시제 또는 종량제의 기준이 된다. 또한 고객센터 차원에서 등록사용자가 즐겨 찾는 페이지 또는 카테고리에 대한 DB 구축도 시도해 볼 수 있다. 일반사용자의 경우 그 지역에 따른 국가별 분포도와 국내 지역별 분포도는 사이버공간의 웹서버를 운영하는 관리자에게 중요한 데이터가 된다. 특히 사용자의 도메인이름에서 추출해 낼 수 있는 업종별 분석도 나름대로의 의미를 부여할 수 있다.

모뎀 사용자와 전용선 사용자는 웹서버가 RAS(Remote Access Service)와 같은 자체모뎀을 사용하지 않는 한 특별히 구별할 방법은 없다. 그러나, 모뎀접속 서비스를 제공하는 국내의 ISP(Internet Service Provider)가 가지고 있는 IP를 참조하면 국내 ISP를 통한 모뎀사용자를 간단히 알아낼 수 있다. 물론 전용선 사용자도 국내 ISP에 가입하였다면, 소속 ISP를 알아낼 수 있을 것이다. 이는 사용자에 대한 홍보매체를 선택하는데 중요한 역할을 한다.

다. 페이지 분석

웹서버의 페이지들에 대한 분석은 웹서버 관리자에게 필수적인 사항이다. 따라서 로그분석 틀은 이러한 페이지분석에 필요한 데이터를 제공해야 한다. 페이지분석은 페이지별 또는 디렉토리별로 하거나 메뉴별로 시도해 볼 수 있지만 메뉴별 분석의 경우 메뉴와 그 하위메뉴와의 관계는 매우 복잡해서 특정 메뉴가 두 개 혹은 그 이상의 메뉴로부터 연결되어 있을 경우도 있지만, 반대로 하나의 하위메뉴가 두 개의 상위메뉴와 연결되는 경우도 있을 수도 있다. 하지만 이러한 경우라도 홈페이지의 주 메뉴에 대한 분석은 가능하다.

페이지분석은 사용자가 가장 많이 찾는 페이지, 가장 오래 읽는 페이지, 가장 최근에 방문한 페이지, 가장 먼저 또는 끝으로 즐겨 찾는 페이지 등을 보여줄 수도 있으며, 이를 디렉토리로 구분하여 같은 내용의 분석도 가능하다. 이는 웹서버의 업그레이드를 계획할 때 작업의 우선순위를 결정하는 데 유용하게 활용할 수 있다. 페이지는 텍스트, 이미지, 오디오, 비디오 등과 같은 MIME의 타입으로 의해 분류할 수 있다.

라. 디렉토리 분석

모든 페이지정보는 디렉토리로 구분되어 있다. 디렉토리는 웹서버의 메뉴와 직접적인 관계를 가지고 있으므로, 웹서버관리자는 파일단위보다는 디렉토리단위의 관리에 주의를 집중해야 한다.

마. 메뉴분석

모든 웹서버는 메뉴를 사용하며, 주 메뉴는 하위메뉴를 가지고 있어서 방문객이 웹서버가 만들어 놓은 메뉴를 따라 웹서버를 검색하게 된다. 이러한 메뉴는 디렉토리 또는 페이지와는 별도로 관리된다. 디렉토리 또는 페이지가 물리적 분류라면, 메뉴는 그러한 물리적 분류 위에 형성된 논리적 분류인 것이다[10].

2.2 웹 마이닝

Web Data로부터 사용 패턴을 발견하기 위한 Data Mining 기법으로는 협업 필터링(Collaborative Filtering), 사례기반추론(Case-Based Reasoning), 군집분석(Clustering), 인공신경망(Neural Network), 유전적 알고리즘(Genetic Algorithm), 퍼지이론(Fuzzy Theory)등과 같은 알고리즘이 적용된다.

협업필터링을 이용한 추천 방식은 고객이 구매한 상품들을 비교하여 다른 고객의 구매 패턴과의 유사도를 측정하여 유사한 구매를 보인 다른 고객이 많이 구매한 상품들을 추천하는 방식이다. 예를 들어 고객 A가 DVD와 책을 구매했고, 고객 B는 DVD와 CD를 구매한 경우 두 고객 간의 유사도가 시스템에서 정한 임계치 이상이 된다면, 고객 A에게는 CD를 추천하고 고객 B에게는 책을 추천하는 방식이다. 이 방식은 시스템 구축 초기에 일정량 이상의 고객 데이터가 축적되어야 추천을 시작할 수 있다는 단점과 고객이나 상품이 추가될 경우 다른 모든 고객과의 유사도를 다시 계산해야 하기 때문에 고객이나 상품이 많아질 경우 성능에 문제가 생길 수 있다는 단점이 있다.

군집화(Clustering)란 이질적인 원소로 구성되어 있는 모집단을 여러 개의 동질적인 부 그룹(Sub Group) 혹은 군집(Cluster)으로 나누어 세분화하는 것을 의미한다. 군집화(Clustering)와 분류(Classification)의 다른 점은 군집화는 미리 정의된 부 집단이 없다는 것이다. 그런데 분류는 사전에 어느 집단에 속하는지 밝혀져 있는 자료를 근거로 개발된 모형을 바탕으로 새로운 원소나 레코드를 미리 정의된 집단에 할당함으로써 모집단을 나누는 방법이다. 군집화에서는 사전에 정의된 부 집단이나 예제 자료가 있을 수 없다. 단지 주어진 상황에 따른 자기들끼리의 유사성에

의해 나누어진다. 군집화를 통하여 나누어진 각각의 부 집단에 어떤 의미를 부여하는 것은 분석자 자신에게 달려 있다. 예를 들어 증상에 따라 나누어진 군집들은 각기 서로 다른 질병을 나타낼 수 있다. 나무 잎이나 줄기의 속성에 따라 나누어진 군집들은 나무의 종류가 서로 다르다는 것을 의미한다.

군집화는 가끔 다른 데이터마이닝이나 모델링의 선행 작업으로 이루어진다. 예를 들어 군집화는 시장 세분화 과정의 첫 번째 단계의 작업이 될 수 있다. 어떤 판촉활동 혹은 홍보 전략이 모든 고객에게 효과가 있을까를 생각하는 대신에 고객을 구매습관에 따라 여러 개의 부 집단으로 나눈 다음 각 부 집단에 대해 어떤 판촉활동이 최선일까를 생각하는 것이 더 효과적이다.

구체적인 예로 어떤 백화점에서 고객을 효율적으로 관리하기 위해 고객들을 몇 개의 부류로 구분하려 한다고 가정해 보자. 이 백화점에서는 지금까지 고객의 최근 구매경력, 구매빈도, 구매액 등을 기준으로 고객을 VIP고객, 우수고객, 일반고객으로 분류하여 관리한다. 그러나 이벤트나 프로모션 대상을 선정하는데 있어 현재의 고객 분류는 너무나 추상적이고 고객의 개인별 특성을 전혀 고려하지 못하고 있어 활용도가 그다지 높지 않다. 고객의 구매이력 뿐만 아니라 인류통계학적 데이터, 라이프스타일 데이터, 그리고 지불유형 등의 다양한 데이터에 군집화 기법을 이용하여 몇 개의 군집을 나눌 수 있다. 그 군집 중 다음과 같은 군집이 있다고 하자.

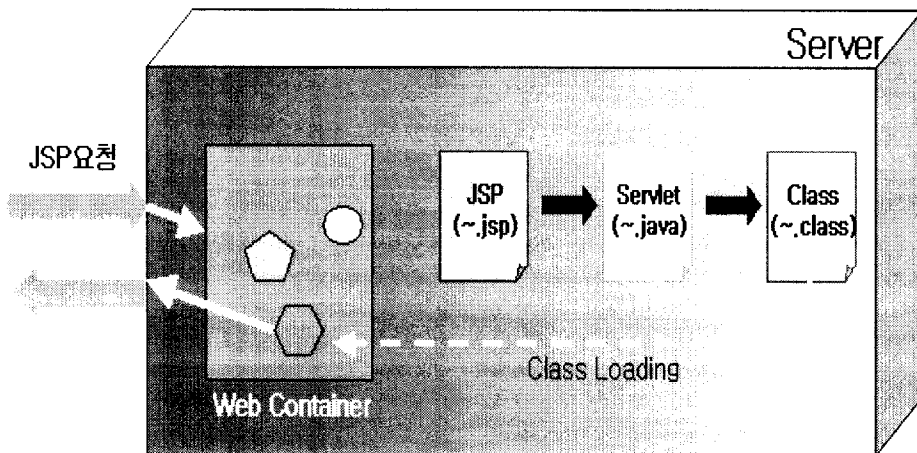
- 소득이 200만원 이상이고, 자녀가 없으며, 연령이 30대
- 교육수준이 높으며, 자녀는 모두 출가했고, 연평균 구매액이 200 ~ 300만원 정도

이와 같은 군집에서 고객들의 특성을 파악할 수 있다면, 적어도 이 군

집에 속한 고객들에게 어린이 의류나 장난감 카탈로그를 발송하는 것은
의할 수 있다[5].

2.3 웹 관련 기술 : JSP(Java Server Page)

JSP는 동적 Web Page를 쉽고 빠르게 개발하는 방법을 제공한다. Java Family의 한 부분으로 Web 기반의 Application을 빨리 개발할 수 있도록 해주며, 특정 Web Server나 Platform에 독립적인 서비스를 제공한다. JSP는 표준 HTML Page에 Java 언어를 사용하여 프로그램 할 수 있도록 하는 것이다. JSP 기술의 핵심은 Presentation과 Business Logic을 분리하는데 그 목표를 둔다. JSP는 Servlet에 비해 Web Page생성이 간단하고, JSP는 최종적으로 class File로 만들어지게 되므로 Platform에 독립적이며 이식성이 좋다. JSP의 수행 과정은 그림 2와 같다.



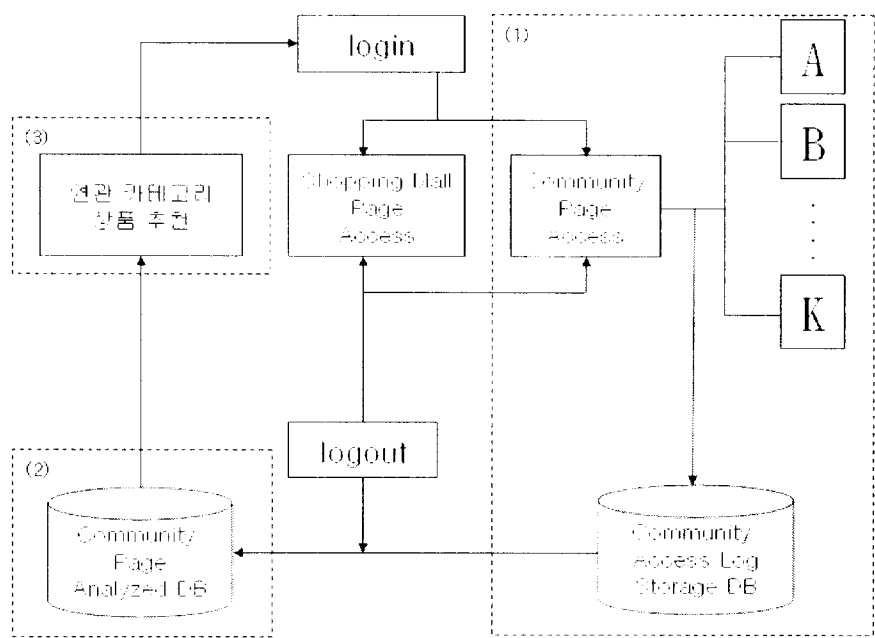
(그림 2) JSP Processing

Ⅲ. 제안하는 추천시스템

본 논문에서 제안하는 기법의 기본적인 개념은 포털 사이트에 가입한 사용자가 로그인 한 후, 자신이 가입되어 있는 커뮤니티 페이지의 점유 시간을 기록하고, 이를 커뮤니티의 카테고리 분야별로 나누어진 클러스터의 총 점유시간과 비교하여 각 카테고리에 대한 선호도를 데이터베이스에 저장한다. 그리고 고객이 해당 사이트에 다시 로그인 하면 고객의 선호도가 가장 높은 카테고리의 관련 상품을 추천하는 것이다.

시스템은 그림 3과 같이 세 가지 에이전트로 구성된다. 첫 번째로 커뮤니티 페이지에 접속하여 각 커뮤니티 페이지에서의 점유 시간을 기록하는 단계인 모니터링 에이전트가 있고, 그리고 이를 분석하여 선호도가 가장 높은 카테고리를 얻어내기 위한 분석 에이전트가 있다. 마지막으로 고객이 로그아웃 한 후, 다음 번 로그인 시에 선호도가 높은 카테고리 관련된 상품을 추천하는 상품추천 에이전트로 구성된다.

그림 3에서 (1)영역은 고객자신이 등록한 커뮤니티 페이지를 액세스하여 점유하는 시간을 기록하는 모니터링 하고, (2)영역은 수집된 데이터를 분석하며, (3)은 고객이 다음 번 로그인 할 때 관심도가 높은 분야의 상품을 추천하는 영역이다.



(그림 3) 시스템 구성도

3.1 모니터링 에이전트

그림 3의 (1)단계는 커뮤니티 페이지에서의 점유시간을 기록하는 단계로서, 사용자가 로그인 한 후 자신이 가입한 커뮤니티 페이지에 액세스하게 되면, 커뮤니티 코드와 회원 아이디, 점유시간 등이 기록되게 된다.

커뮤니티 페이지에 액세스하게 되면, 접근한 커뮤니티 코드와 시간 정보를 세션에 저장해 둔 다음, 고객이 해당 커뮤니티를 벗어나게 되면 점유한 시간을 계산하여 데이터베이스에 저장하게 된다.

예를 들어 고객이 A라는 커뮤니티 페이지에 접근하면 커뮤니티 코드 값과 접근 시간을 세션에 저장하게 된다. 그 다음 고객이 액세스한 커뮤니티에서 필요한 정보를 읽거나 쓰기를 한 후, 고객이 다른 페이지로 이동하거나 B라는 커뮤니티로 이동할 경우, 이전에 머물러 있었던 A 커뮤니티에서의 점유시간 정보를 데이터베이스에 저장하고, B 커뮤니티에 대한 코드 값과 액세스 시간 정보를 세션에 저장하게 된다. 고객이 커뮤니티 페이지를 계속 접근 하는 동안 이 과정을 반복한다.

3.1.1 회원 가입 테이블

회원 가입 테이블은 회원이 포털 사이트에 가입 시 등록되는 회원 자신의 기본 정보가 저장되는 테이블이다.

표 1은 회원가입 테이블의 구조로서 사용자의 id, 비밀번호, 성명, 전자우편 등으로 구성되어 있다.

<표 1> 회원 기본 정보 테이블

필드명	데이터형	크기	설명	비고
mem_id	문자열	varchar(15)	회원아이디	primary key
mem_name	문자열	varchar(10)	회원이름	
mem_passwd	문자열	varchar(15)	비밀번호	
mem_idno1	문자열	char(6)	주민번호1	
mem_idno2	문자열	char(7)	주민번호2	
mem_email	문자열	varchar(30)	email	
mem_tel	문자열	varchar(13)	전화번호	
mem_addr	문자열	varchar(100)	주소	
mem_regdate	날짜	date	가입일자	
mem_moddate	날짜	date	수정일자	

3.1.2 커뮤니티 가입 정보

표 2는 회원이 포털 사이트에 가입한 후, 커뮤니티 페이지에 가입을 하면 해당하는 커뮤니티 가입일자와 그에 대한 정보사항들이 저장되는 테이블로써 각 회원이 어느 커뮤니티에 언제 가입했는지 알 수 있는 테이블이다.

<표 2> 커뮤니티 가입 정보 테이블

필드명	데이터명	크기	설명	비고
seq	숫자	integer	번호	primary key
mem_id	문자열	varchar(15)	회원아이디	
comm_code	문자열	varchar(10)	커뮤니티코드	
comm_regdate	날짜	date	가입일자	
comm_deldate	날짜	date	탈퇴일자	

3.1.3 커뮤니티 접속 로그 정보

표 3은 회원이 포털사이트 내에 있는 커뮤니티에 가입한 후 커뮤니티 페이지로 접속하면 접속일자와 접속시간이 기록되는 테이블이다.

<표 3> 커뮤니티 접속 로그 정보 테이블

필드명	데이터명	크기	설명	비고
seq	숫자	integer	레코드번호	primary key
mem_id	문자열	varchar(15)	회원아이디	
comm_code	문자열	varchar(10)	커뮤니티코드	
logging_date	날짜	date	접속일자	
residence_time	날짜	date	접속시간	

3.1.4 회원 관심 카테고리

표 4는 회원의 커뮤니티 로그 정보를 분석하여 카테고리별로 관심도를 구한 값이 저장되는 테이블이다.

<표 4> 회원 관심 카테고리 테이블

필드명	데이터명	크 기	설 명	비 고
seq	숫자	integer	번호	primary key
mem_id	문자열	varchar(15)	회원아이디	
cat1_code	문자열	char(5)	카테고리(대)	
cat2_code	문자열	char(5)	카테고리(중)	
per_inter	문자열	integer	관심도	

3.1.5 카테고리의 대분류, 중분류 코드

표 5와 표 6은 커뮤니티 카테고리를 클러스터링하는 기준이 되는 대분류와 중분류 코드가 저장되는 테이블이다.

<표 5> 커뮤니티 카테고리 대분류테이블

필드명	데이터명	크기	설명	비고
cat1_code	대분류코드	char(5)	대분류	primary key
cat1_name	대분류명	varchar(20)	대분류명	

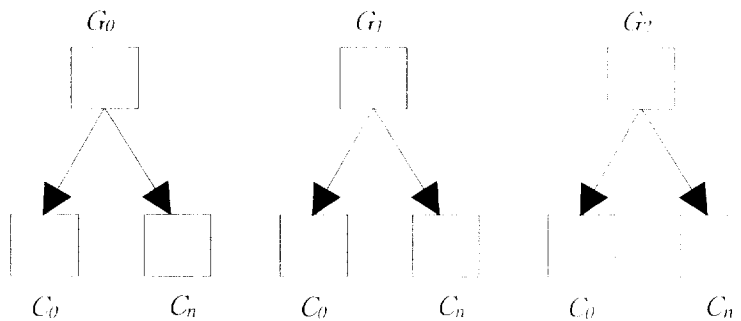
<표 6> 커뮤니티 카테고리 중분류테이블

필드명	데이터명	크기	설명	비고
cat2_code	중분류코드	char(5)	중분류	primary key
cat2_name	중분류명	varchar(20)	중분류명	

3.2 분석에이전트

그림 3의 (2)단계에서는 모니터링 에이전트를 통하여 데이터베이스에 저장된 커뮤니티 가입일자, 접근일자, 체류시간 정보를 사용하여 고객이 가장 활발히 활동하는 커뮤니티페이지를 분석한다.

커뮤니티를 이동할 때마다 그 액세스한 시간과 체류시간을 기록하고, 로그아웃 시에 커뮤니티 카테고리별로 그룹화하여 관심도를 분석한 뒤 고객의 커뮤니티 DB에 관심도를 반영한다.



(그림 4) 카테고리 계층도

그림 4에서 보는 바와 같이 그룹 G_0 카테고리에 속한 커뮤니티 $C_0, C_1, \dots, C_n$ 의 커뮤니티 점유시간은 모두 그룹 G 카테고리에 대한 관심도에 반영된다.

$$C_{u,j} = \frac{R_{u,comm[j]}}{\sum_{i=1}^n R_{u,comm[i]}} \times W + \frac{F_{u,comm[j]}}{\sum_{i=1}^n F_{u,comm[i]}} \times (1 - W) \quad (1)$$

$$G_{u,i} = \sum_{j=1}^n C_{u,i,j} \quad (2)$$

식 (1)은 분석 시에 사용할 패턴추출 알고리즘이다[7]. 식 (1)에서 $C_{u,j}$ 는 사용자 u 의 커뮤니티 j 에 대한 선호도를 나타내며, $R_{u,comm[j]}$ 는 사용자 u 가 커뮤니티 j 에서 체류한 시간을 나타낸다. $F_{u,comm[j]}$ 는 사용자 u 가 커뮤니티 j 에 접속한 회수이다. W 은 커뮤니티 점유시간에 적용할 가중치이고, $(1 - W)$ 은 커뮤니티 접속회수에 적용할 가중치이다.

식 (2)는 식 (1)에서 구해진 각각의 커뮤니티 선호도를 그룹화하여 고객이 가장 선호하는 커뮤니티 카테고리를 분석하는 식이다. 식 (2)에서 $G_{u,i}$ 는 사용자 u 의 카테고리 그룹 i 에 대한 선호도이고, $C_{u,i,j}$ 는 카테고리 그룹 i 에 속한 커뮤니티 j 에 대한 사용자 u 의 선호도이다.

본 논문에서 제안하는 방법은 고객이 가입하여 활동하는 커뮤니티 중에 가장 선호하는 커뮤니티를 분석해 내는 것이므로, 고객의 체류시간에 가장 높은 가중치를 주고, 다음으로 커뮤니티에 액세스한 빈도, 커뮤니티에 가입한 날짜순으로 가중치를 부여한다.

각각의 커뮤니티에 대한 관심도를 그룹화 하여 카테고리별 관심도를 구할 경우, 유사한 커뮤니티에 여러 개 가입하여 활동하더라도 카테고리별로 그룹화되어 분석 시에 오류를 줄일 수 있다.

3.3 상품추천 에이전트

분석 에이전트단계에서 분석된 가장 관심도가 높은 커뮤니티 카테고리를 바탕으로 고객이 다음 로그인 시에 상품 추천을 받을 수 있도록 한다.

커뮤니티에 가입하여 활동한 데이터가 없는 고객의 경우 회원가입 시에 선택한 관심분야를 “회원 관심 카테고리 Table”에 등록 하고 이를 토대로 관련된 상품을 추천하도록 한다. 그 외에 본 논문에서 목표로 하고 있는 고객이 커뮤니티에 가입하고 액세스한 로그 데이터가 있는 경우에는, 커뮤니티 활동 DB 분석단계에서 계산된 카테고리를 기반으로 로그인시에 상품 추천을 받게 된다.

Ⅳ. 시스템 구현 및 성능평가

본 시스템에서는 고객이 회원가입 시에 관심분야로 채택한 카테고리를 기반으로 상품을 추천하는 경우와 커뮤니티 관심도 계산에 의해 추천받은 상품을 비교해 본다.

본 시스템의 사용언어는 Java j2sdk1.4와 JSP 2.0이고, 웹서버는 resin 2.1.14를 사용하였으며, 데이터베이스는 mySql 3.23.53을 사용하였다. 그 외 기타 구축환경은 표 7과 같다.

<표 7> 시스템 구축환경

구 분	구축환경
기종	펜티엄 4 PC
O/S	windows 2000 Server
RDBMS	mySql 3.23.53
개발 Language	Java j2sdk1.4, JSP 2.0
Web Server	resin 2.1.14
JSP Container	inside resin

본 논문의 실험을 위한 데이터로는, 포털사이트의 카테고리 대분류 코드와 중분류 코드를 사용하였다. 그리고, 고객의 커뮤니티 액세스 로그 기록은 20명의 고객이 30일 동안 접속하여 활동한 커뮤니티 로그 기록을 사용하였다.

표 8은 “park9977” 아이디를 사용하는 고객의 커뮤니티별 액세스 회수

와 점유시간에 대해 식 (1)을 사용하여 선호도를 구한 예이다. 이렇게 구해진 값에 식 (2)를 적용하여 관심도가 높은 카테고리를 구하게 된다.

<표 8> 식 (1)을 적용한 결과

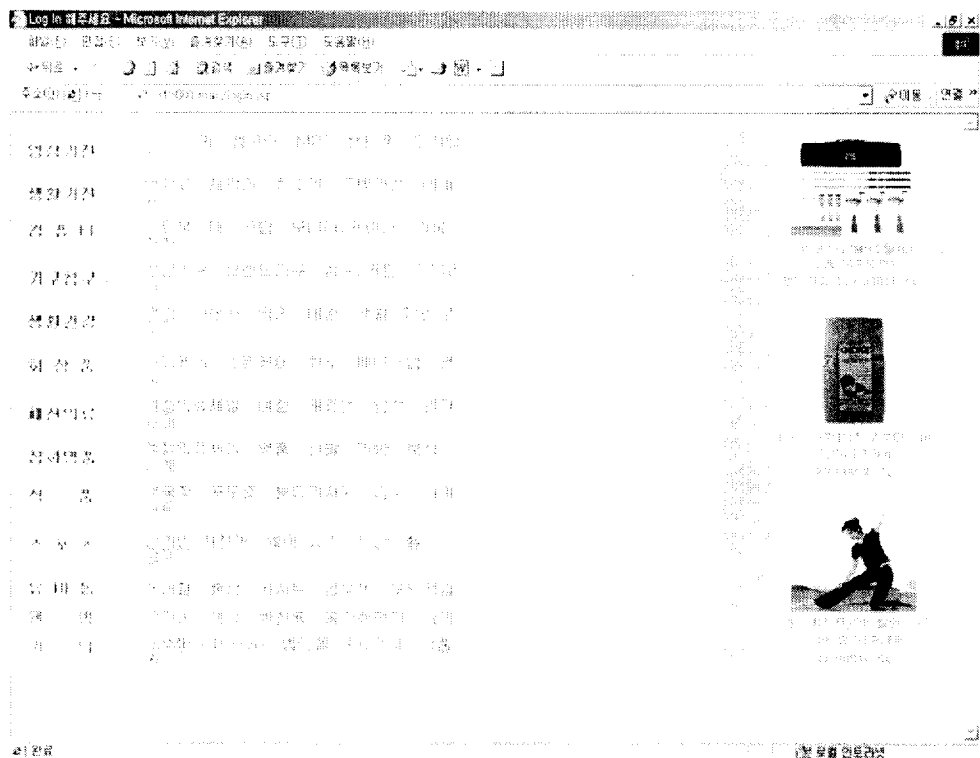
회원아이디	커뮤니티코드	커뮤니티접속회수	점유시간	선호도
park9977	B1	27	12425	29.88
park9977	A1	20	18912	32.09
park9977	D1	8	5219	11.93
park9977	C1	4	2884	6.39
park9977	D2	4	4209	8.49
park9977	A3	1	487	1.2
park9977	B2	1	159	0.4
park9977	F1	0	0	0

표 9는 표 1에서 구해진 값에 대해서 식 (2)를 적용하여 고객이 가입한 커뮤니티에 대해 같은 카테고리 내의 커뮤니티에 대해 접속회수와 점유시간을 계산하여 해당 카테고리 중에 가장 선호도가 높은 순서대로 나열한 것이다.

<표 9> 식(2)를 적용한 결과

회원아이디	카테고리코드	선 호 도
park9977	A	33.29
park9977	B	30.28
park9977	D	20.42
park9977	C	6.39
park9977	F	0

그림 5는 고객이 로그인 한 후 선호도가 높은 카테고리 와 관련이 있는 상품을 추천 받은 화면이다.



(그림 5) 로그인 시 상품 추천 화면

본 시스템의 평가는 쇼핑몰을 이용한 경험이 있는 사용자 20명과 그들의 커뮤니티 점유시간 데이터, 그리고 각 카테고리별로 분류된 20여 카테고리의 상품 데이터를 이용하였다. 실험에 대한 평가는 무작위로 추천된 상품과 초기 프로파일 생성 시에 제시하는 구매율이 높은 상품에 대한 선호도, 그리고 제안하는 시스템을 통해 추천된 상품에 대한 선호도를 비교 평가 하였다.

평가 방법으로는 정보검색 분야에서 보편적으로 사용되는 평가척도인 Precision을 응용하여 사용하였다. Precision은 전체 추천된 상품 개수 중에서 사용자가 선호하는 상품 개수의 비율로, 여기에서는 추천된 상품에 대해서 고객이 선호하는 경우의 클릭 수를 반영하였다.

표 10은 추천된 상품에 대해 선호도를 측정한 결과이다. A는 무작위로 추천된 상품에 대한 선호도이고, B는 많이 팔린 상품에 대한 선호도이며, C는 제안하는 추천 기법으로 추천된 상품에 대한 상품의 선호도를 나타낸다.

사용자 중에는 많이 팔린 상품에 대한 선호도를 높게 평가한 경우도 있지만, 대부분의 사용자가 선호하는 커뮤니티와 관련된 상품에 대해 선호도를 높게 평가하였다.

표 11은 표 10의 결과에 대한 Precision 측정 결과이다. 고객에게 추천된 전체 상품 수에 대해 A, B, C의 방법에 대한 선호도 클릭수를 반영하여 나타내었다.

$$Precision = \frac{|A \cap B|}{|A|} \quad (3)$$

식 (3)에서 A는 시스템이 추천한 상품리스트이고, B는 사용자가 선호하는 상품 리스트이다.

<표 10> 추천된 상품에 대한 선호도

회원아이디	선호카테고리	상품추천회수	추천 A	추천 B	추천 C
park9977	A	39	7	12	20
pa98k2	E	11	1	3	7
mitocondira	F	16	2	6	8
miso376	B	3	0	1	2
bookinet	A	20	4	10	6
stormgf	D	32	5	17	10
adwinnie	F	7	1	2	4
pslc	J	27	3	14	10
ds24	O	40	5	18	17
pakch	I	46	7	16	23
hrhanny	C	23	2	10	11
plushost	F	16	1	9	6
mudoin21	K	38	5	12	21
promise1013	F	15	3	4	8
arrasha	J	5	1	0	4
kikiki0110	H	44	7	12	25
rjsah97	N	29	4	12	13
y72031	G	5	0	1	4
jin7447	G	14	3	7	4
q0200	L	36	6	18	12

추천 A : 무작위 상품 추천

추천 B : 구매율이 높은 상품 추천

추천 C : 제안하는 시스템의 상품 추천

<표 11> Precision 측정 결과

구 분	추천 A	추천 B	추천 C
Precision	0.144	0.395	0.461

실험결과, 제안하는 시스템에서 추천된 상품의 선호도가 높게 나타났다. 또한 커뮤니티 점유시간에 의해 발생한 데이터를 이용한 본 실험을 통해, 새로운 구매데이터가 쌓일 때까지의 오랜 시간과 비용발생 없이도 성능을 평가할 수 있었다.

기존의 포털 사이트에 가입된 고객이 쇼핑몰 페이지로 로그인 할 경우 최다판매량을 기록한 상품이나 최근의 인기 상품을 추천 받게 된다. 하지만 이것은 고객 개인의 선호도가 반영된 것이 아니기 때문에 고객 관심을 갖고 있는 것에 대한 상품 추천이라고 보기 어렵다.

개인의 선호도에 맞춘 상품 추천을 위해서는 사용자의 모든 정보를 적당한 표현 방법으로 저장한 후 이 정보로부터 선호도 정보를 추출하고, 여기에 고객의 웹로그 정보를 분석하여 적용하는 것이 바람직하다.

본 논문에서 구현한 방식만을 포털사이트에 적용한다는 것은 사용자의 부분적인 웹 활동에만 초점을 맞춰 상품을 추천하는 것과 같다. 하지만 이를 기존의 포털 사이트에서의 운용되고 있는 개인화 분석 시스템에 부가적으로 사용하게 된다면 고객 개인의 기호에 맞춘 맞춤형 인터넷을 제공하는 역할을 할 수 있다.

V. 결론 및 향후 연구과제

인터넷 환경의 발달로 인터넷 인구가 기하급수적으로 증가함에 따라 전자상거래 시장이 빠른 속도로 성장하고 있다. 이로 인해 전자상거래 분야에 있어, 고객의 유치·관리 및 서비스에 대한 연구가 진행되고 있다. 이러한 전자상거래 시장의 증가에 따른 사용자들의 다양한 요구에 대응하기 위해 개인화 기술은 전자상거래의 중요한 부분으로 떠올랐다. 현재 전자상거래 업체들의 최대 관심사는 기업이 가진 많은 상품들을 변화하는 사용자 개개인에 맞추어 추천함으로써 추천된 상품의 히트율을 증가시켜 고객의 만족도와 기업의 이익을 극대화하는 것이다.

사용자 개인의 취향에 알맞은 정보를 제공하는 개인화 서비스는 추천의 정확도에 있어서 한계를 보이고 있다는 문제점이 드러나고 있다.

또한 사용자들 간의 유사도를 계산하는데 있어서 가중치를 고려하지 못하는 문제점과 희소성(sparsity)문제, 사용자의 수 증가에 따른 수행속도가 저하되는 문제들을 안고 있다.

본 논문에서는 상품추천 성능의 향상을 위해 포털사이트의 고객이 로그인 한 후, 각 커뮤니티페이지에 접근한 시간을 측정하여 가장 관심을 보이는 커뮤니티를 도출해 내어 커뮤니티카테고리와 관련된 상품을 추천하는 방법을 제안하였다.

기존의 추천 시스템에서는 협력필터링을 통한 추천과 내용 기반의 추천방법으로 고객의 성향에 맞는 상품을 추천하도록 하고 있으나 이러한 방법을 적용하기 위해서는 고객 개개인의 구매 내역이 일정 수준이상 축적되어야 가능하다는 단점이 있다.

고객이 포털사이트에 가입할 당시의 채택한 관심분야를 바탕으로 상품 추천을 할 경우, 고객이 회원정보를 변경하지 않은 한 변화하는 고객의 관심분야를 반영할 수 없는 반면 본 논문에서 제안하는 알고리즘을 적용하면 고객이 최근에 가장 활발하게 관심을 보이는 커뮤니티 카테고리를 분석해내므로 그에 관련된 상품을 추천할 수 있다.

앞으로 선호도 계산 알고리즘을 더욱 보완하여 더욱 더 향상된 커뮤니티 선호 알고리즘을 적용하고, 관련 카테고리에서 최적의 상품을 선택해서 추천할 수 있도록 할 계획이다.

[참고문헌]

- [1] Marko Balabanovic and Yoav Shoham, Fab, "Content based, Collaborative Recommendation", Communications of the ACM, Vol.40, No.3, pp. 66-72, March, 1997.
- [2] Basu, C., Hirsh, H., and Cohen, W., "Recommendation as Classification, Using Social and Content Based Information in Recommendation", in Proceedings of the American Association for Artificial Intelligence, pp. 714-720, 1998.
- [3] Sawar, B. M., karypis, G., Konstan, J. A., and Riedl, J., "Item based Collaborative Filtering Recommender Algorithms", Accepted for Publication at the WWW10 Conference, 2001.
- [4] Breese, J. S., Heckerman, D., and Kadie, C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", in Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pp. 43-52, 1998.
- [5] Lin, W., Ruiz, C., and Alvarez, S. A., "Collaborative Recommendation via Adaptive Association Rule Mining", International Workshop on Web Mining for E Commerce (WEBKDD2000), Held in Conjunction with the Sixth International Conference on Knowledge Discovery and Data Mining (KDD2000), 2000.
- [6] Billsus, D., and Pazzani, M. J., "Learning Collaborative Information Filters", in Proceedings of the Fifteenth International Conference on Machine Learning, pp. 46-53, 1998.

- [7] Cohen, W. W. Web_Collaborative Filtering: "Recommending Music by Crawling The Web", The International Journal of Computer & Telecommunications Networking, Vol.33 No.1-6, pp. 685-698, 2000.
- [8] Sarwar, B. M., Konstan, J. A., Rorchers, A., Herlocker, J., Miller, B., and Riedl, J., "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System", in Proceedings of the ACM Conference on Computer Supported Cooperative Work(CSCW), 1998.
- [9] Ahmad, M., and Wasfi, A., "Collecting User Access Patterns for Building User Profiles and Collaborative Filtering", in Proceedings of the 1999 International Conference on Intelligent User Interfaces, pp. 57-64, 1999.
- [10] Melville, P., Mooney, R., Nagarajan, R., "Content-Boosted Collaborative Filtering for Improved Recommendations", in Proceedings of the Eighteenth National Conference on Artificial Intelligence, pp. 187-192, 2002.
- [11] Good, N., Schafer, B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., Riedle J., "Combining Collaborative Filtering with Personal Agents for Better Recommendation", in Proceedings of the AAAI Conference, pp. 439-446, 1999.
- [12] Q. Li and B. M. Kim, "An Approach for Combining Content-based and Collaborative Filters", in Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, pp. 17-24, 2003.
- [13] Q. Li and B. M. Kim, "Clustering Approach for Hybrid

- Recommender System”, in Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence(WI 2003), pp. 33-38, 2003.
- [14] Fu, X., Budzik, J., and Hammond, K. J., “Mining Navigation History for Recommendation”, in Proceedings of the 2000 International Conference on Intelligent User Interfaces, pp. 106-112, 2000.
- [15] Soboroff, I., Nocholas, C., “Combining Content and Collaborative in Text Filtering”, in Proceedings of the IJCAI Workshop on Machine Learning in Information Filtering, pp. 86-89, 1999.
- [16] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J., “Getting to know you : Learning New Use Preferences in Recommender System for Groups of Users”, in Proceedings of the Seventh International Conference in Intelligent User Interfaces, pp. 127-134, 2002.
- [17] 고병진, 유영훈, 조근식, “개인화 추천시스템을 위한 효율적 연관 규칙 방법”, 2002년 한국정보처리학회 추계학술대회 2002. 10.
- [18] 강귀영, 조동섭, “사용자 프로파일을 이용한 웹페이지 추천”, 2001년 한국정보과학회 춘계학술발표회 2001. 4.
- [19] 김준호, 김준태, “카테고리 정보를 이용한 추천 성능의 향상”, 2003 한국정보과학회 춘계학술대회 2003. 4.
- [20] 이은영, 조동섭, “개인화 상품 추천을 위한 협력 필터링 에이전트”, 2001년 한국정보과학회 춘계학술대회 2001. 4.
- [21] 신형섭, 이충세, “클러스터의 점유시간을 이용한 웹 페이지 추천 기법”, 2002년 한국정보과학회 추계학술대회 2002. 10.

감사의 글

남들보다 늦게 대학원을 입학하여 아득하게만 느껴지던 졸업이 이제 얼마 남지 않았다고 생각하니 그 동안 저에게 많은 도움을 주신 분들이 머리 속에 떠오릅니다.

부족한 점이 너무나 많은 저를 따뜻하게 대해주시고 마지막까지 격려와 지도를 아낌없이 베풀어 주신 윤성대 지도교수님께 진심으로 감사드립니다. 논문 심사를 위해 바쁘신 와중에도 많은 가르침을 주신 김창수 교수님과 자상하게 심사해 주신 박홍복 교수님께 감사의 말을 전하고 싶습니다.

대학원에 입학한 저에게 열심히 하라고 하시며 격려해주신 경상대학의 김용덕 교수님, 연구실에 갈 때마다 친오빠처럼 맞아주시는 박현호 선배님, 부족한 부분을 많이 채워주신 황순환 선배님, 세심하게 항상 챙겨주시는 윤종찬 선배님, 세심하게 조언을 해주신 고석범 선배님, 친절하게 저를 챙겨주시는 박상일 선배님에게 정말 감사드립니다. 그리고 저와 2년 반을 함께하며 서로에게 많이 힘이 되어준 이정미 선생님과 항상 즐거운 모습으로 대해준 유은아 선생님, 밝고 활기차게 생활하는 귀녀, 너무 착해서 걱정되는 정애를 비롯해 2년 반 동안 함께 했던 연구실의 선배님들과 후배님들께 정말 감사드립니다.

끝으로, 제가 힘들 때 많이 위로해주고 힘이 되어준 그 이와 항상 저의 편이 되어주시고 저의 선택을 믿어주시며 묵묵히 뒷바라지 해주신 어머니께 사랑과 존경의 마음을 담아 이 논문을 드립니다.

2005년 6월 30일

대학원 생활을 정리하면서 -