韓國 컴퓨터情報學會 論文誌
第8卷 第4號, 2003. 12.

2003-8-4-2-6

웹 마이닝을 이용한 개인 광고기법에 관한 연구

김 은 수*, 송 강 수**, 이 원 돈***, 송 정 길****

A Study on Personalized Advertisement System Using Web Mining

Eun-Soo, Kim*, Kang-Soo, Song**, Won-Don, Lee***, Jung-Gil, Song****

요 약

최근 전자상거래의 발전과 인터넷 사용자의 급증으로 온라인 상에서 수많은 광고들이 서비스되고 있다. 하지만 이러한 광고서비스는 사용자들의 성향 분석을 기초로 하기보다는 해당 광고의 일방적 서비스에 그치고 있다. 따라서 많은 웹사이트들이 해당 광고의 효율적 서비스를 위해 개인화된 광고서비스를 원하고 있고 해당 서버의 로그 분석을 통한 서비스를 연구 및 시행하고 있다. 본 논문에서는 서버측 로그데이터의 분석이 아닌 로컬 시스템의 로그데이터를 이용하여 사용자의 선호도와 성향을 분석한다. 또한 해당 사이트별 분류 카테고리를 만들어 해당 분류의 가중치를 부여함으로써 개인화된 광고 시스템을 제안하려고 한다. 사용자의 선호도 분석은 웹 개인화 기법 중 협업 필터링의 대상이 되는 사용자 선호도 정보를 방문 사이트 분류에 사용하고 학습에이전트의 대상이 되는 인터넷 사용자의 행동을 해당 사이트의 방문횟수로 가정하여 사용자의 성향분석을 시도하였다. 사용자의 선호도를 벡터로 표현하고, 성향분석 결과를 단순 적용형태가 아닌 연속적 데이터로 간주하였으며 이전 데이터와 이후 데이터의 성향분석 변화를 제안하는 기법을 이용하여 새롭게 분석하고 피드백 시킴으로써 지속적인 갱신과 적용을 할 수 있도록 제안하였다. 이러한 결과를 통해 해당 분류의 광고들을 선정하고 선정된 광고에 사용자 성향분석과 동일한 과정을 적용시킴으로써 차별화된 광고 서비스를 제공할 수 있는 방법을 제시하였다.

* 한국 과학기술정보연구원 위촉연구원
** 충남대학교 컴퓨터과학과 석사과정
*** 충남대학교 컴퓨터과학과 교수
**** 한남대학교 정보통신·멀티미디어공학부 교수

Abstract

Great many advertisements are serviced in on-line by development of electronic commerce and internet user's rapid increase recently. However, this advertisement service is stopping in one-side service of relevant advertisement rather than doing users' inclination analysis to basis. Therefore, want advertisement service that many websites are personalized for efficient service of relevant advertisement and service through relevant server's log analysis research and enforce. Take advantage of log data of local system that this treatise is not analysis of server log data and analyze user's Preference degree and inclination. Also, try to propose advertisement system personalized by making relevant site tributary category and give weight of relevant tributary. User's preference user preference which analysis is one part of cooperation fielder ring of web personalized techniques use information in visit site tributary and suppose internet user's action in visit number of times of relevant site and try inclination analysis of mixing form. Express user's preference degree by vector, and inclination analysis result uninterrupted data that simplicity application form is not regarded and techniques that propose inclination analysis change of data since with move data use and analyze newly and proposed so that can do continuous renewal and application as feedback Sikkim. Presented method that can choose advertisements of relevant tributary through this result and provide personalized advertisement service by applying process such as user inclination analysis in advertisement chosen.

▸ Keyword : eCRM, 웹개인화, 데이터마이닝, 웹마이닝, 로그분석, 광고

Ⅰ. 서 론

인터넷의 급속한 발달과 함께 전자상거래 사이트와 포탈(potal) 사이트 간의 고객 유치 경쟁이 치열해 지고, 인터넷 사이트들은 고객을 끌어들이기 위해 경쟁적이고 다양한 광고와 뉴스를 제공하여 보다 많은 수의 회원을 확보하기 위해 노력하고 있다. 하지만, 대부분 인터넷 사이트에 제공되는 획일적인 광고와 무분별한 뉴스 및 정보 서비스가 오히려 회원들의 인터넷 사용을 불편하게 하고, 고객의 많은 관심을 끌지 못하여 네트워크의 효율적인 사용을 저해하는 정보공해에 지나지 않는 상황에 이르렀다[1][2].

이러한 이유로 사용자별로 차별화된 맞춤 서비스를 제공하여 사용자 만족을 극대화 시키는 개인화(personalization)의 중요성이 크게 부각되고 있다. 웹 사용자에게 개인화된 서비스를 제공하는 다양한 서비스 형태들이 제시되고, 이러한 서비스를 제공하기 위하여 고객에 대한 정확한 성향 분석을 통해 인터넷상에서 최적화된 성능을 보일 수 있는 개인화 기법들이 필요하게 되었다.

본 논문에서는 개인화된 광고 서비스를 위하여 고객들의 취향을 분석하고 분석결과를 광고 선택에 반영하는 개인화된 광고시스템을 제안하고자 한다. 이러한 분석을 위해서 웹 로그 데이터로부터 사용자의 정보를 수집해 분석해야 한다. 분석방법으로는 서버 측 로그파일 분석과 사용자 측 로그 파일 분석으로 양분시킬 수 있다. 서버 측 로그파일을 분석하여 이룰 수 있는 서비스는 단일 서버 내에서의 개인화가 가능하며 이때 개인화는 맞춤형 포탈 서비스 와 같은 것이 있다. 그러나 사용자 측에 있는 URL접속기록을 분석하면 분석하는 대상의 범위가 서버에서의 분석과 비교할 때 더 개인의 성향에 가까운 분석이 가능하다.

따라서 본 논문에서는 사용자 측 URL 접속기록을 통해 사용자별로 성향분석을 하고 분석된 결과를 통해 사용자별로 차별화된 광고서비스가 가능한 방법을 제시하려고 한다. 또한 데이터 수집 및 분석 과정에서 기존에 사용되던 데이터 마이닝 기법과 웹 개인화 기법의 일부분들을 가정과 적용 알고리즘에 응용함으로써 분석의 신뢰도를 높인다. 그리고 사용자의 성향은 항상 변화의 가능성을 내포하고 있기 때문에 시간의 흐름에 따라 변하는 사용자의 성향분석결과를 제시된 방법에 의해 재분석함으로써 능동적인 개인화 서비스가 가능하도록 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 관하여 서술하고, 제 3장에서는 로그분석을 통한 시스템 제안 및 사용자 성향 분석에 대해 기술한다. 제 4장에서는 3장에서 언급한 알고리즘과 성향 분석 기법을 이용하여 시스템 설계 및 실험을 통한 분석 결과를 기술하고, 마지막으로 제 5장에서 결론 및 향후 연구방향에 대하여 서술한다.

Ⅱ. 관련연구

1. eCRM(Electronic Customer Relationship Management)

인터넷 기반의 온라인 CRM을 eCRM이라고 하며, 고객에 대한 이해와 접근방식은 오프라인 CRM과 동일하나, 고객정보획득 및 커뮤니케이션 방법에는 차이가 있다.[16].

즉, eCRM이란 e-Business 환경에서 적용되는 CRM을 의미한다. 따라서 eCRM은 기존의 오프라인 CRM과 달리 인터넷을 통해 고객 데이터를 수집하고, 고객과 커뮤니케이션할 수 있다는 특징이 있다[17][19].

또한, eCRM은 e-Business의 특징인 실시간 반응(real-time reaction), 실시간 가격 책정(real-time pricing) 등을 할 수 있다는 장점이 있다. 또한, eCRM은 고객과 회사간의 공간적, 시간적, 물리적 장벽을 극복함으로써 글로벌 관점에서 고객관리를 할 수 있다. eCRM의 개념을 보다 명확하게 규정하기 위하여 도식화한 eCRM의 개념적 위상은 <그림 1>과 같다[18][19].

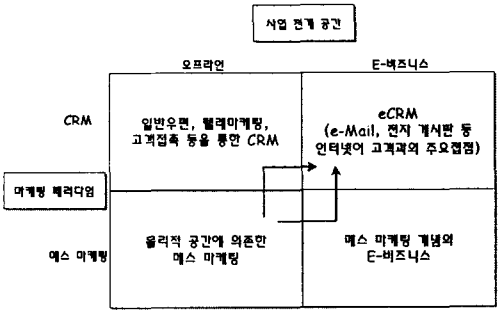

그림 1. eCRM의 개념적 차원
Fig. 1 Concept enemy dimension of eCRM

2. 데이터 마이닝(Data Mining)

2.1 데이터 마이닝의 개념

데이터 마이닝이란 대량의 데이터로부터 통계적 기법, 인공지능 기법, 패턴인식 등을 이용하여 데이터간의 숨겨져 있는 상호 관련성, 분류 및 군집화, 추정 및 예측 등 유용한 정보를 추출하여 의사 결정에 적용하는 과정이라 할 수 있다[10]. 데이터 마이닝은 기업 경영, 과학 및 의료 등 다양한 분야에서 단순한 데이터 수준이 아닌 보다 고급 수준의 정보 즉 지식의 획득, 활용 및 효율적인 분배의 측면에서 최근 그 중요성이 급격히 부각되고 있다. 데이터 마이닝을 위한 기법으로는 통계학적인 기법, 연관성 측정, 클러스터링, 의사결정나무, 신경망 모형과 같은 기법들이 있다[9][11].

2.2 데이터 마이닝 기법

(1) 통계적 기법(Statistical techniques)

주어진 문제를 과거의 유사한 사례의 통계를 바탕으로 상황에 맞게 응용하여 해결해 가는 기법이다. 과거의 통계를 이용하여 새로운 상황을 설명하거나, 과거의 통계데이터로 새 해답을 평가하거나, 또는 새로운 상황을 이해하기 위해서나, 새로운 문제에 대한 적당한 해답을 만들기 위해 추정하는 것을 의미한다. 이러한 방법은 과거의 전문가 시스템에서 사용하던 지식 (정형화된 Rule)의 추론을 통해서 해를 얻는 방법보다는 단순하면서도 문제 영역이 잘 정형화되지 않는 분야에서는 좋은 접근법이라 할 수 있다[9].

(2) 연관성 측정(Associations)

연관성 측정은 어떤 특정 문제에 대해 아직은 일어나지 않은 답(예를 들어, 예/아니오)을 얻고자 하는 예측(Prediction)의 문제나 고객들을 특정목적에 따라 분류(Segmentation)하는 문제가 아니라, 상품 혹은 서비스의 거래 기록(Historical) 데이터로부터 상품간의 연관성 정도를 측정하여 연관성이 많은 상품들을 그룹화하는 클러스터링의 일종으로서, 동시에 구매될 가능성이 큰 상품들을 찾아냄으로써 시장바구니 분석(Marcket Basket Analysis)에서 다루는 문제들에 적용될 수 있다[10].

(3) 클러스터링(Clustering)

어떤 목적변수(target)를 예측하기보다는 고객수입, 고객연령과 같이 속성이 비슷한 고객들을 묶어서 몇 개의 의미 있는 군집으로 나누는 것으로, 대용량의 데이터가 너무 복잡할 때 이를 구성하고 있는 몇 개의 군집을 나누어 살펴봄으로써 전체에 대한 윤곽을 잡는 기법이다[9][10].

(4) 의사결정나무(Decision Trees)

의사결정나무는 분류 및 예측에 있어서 자주 쓰이는 기법으로 DM의 응답여부 등에 영향을 미치는 변수들과 변수들의 상호작용을 누구나 쉽게 이해할 수 있도록 굳이 통계학적인 용어를 쓰지 않더라도 설명이 가능한 기법으로 데이터 마이닝을 언급할 때마다 빠지지 않고 소개되는 분석기법이다[11].

(5) 신경망 모델(Netutal networks)

신경망 모델은 신경생리학 분야에서 두뇌의 활동을 이해하고자 하는 목적 하에 신경의 작업을 설명하려는 시도에서 출발하여 생물학적 프로세스를 컴퓨터를 이용하여 모형화하려는 노력에서 비롯된 것으로, 80년대 이후 생물학적 활동의 모형발전과 더불어 컴퓨터 성능의 진보, 신경망 이론에 대한 통계학적인 접목으로 인해 빠르게 진보되면서 최근에는 데이터 마이닝에 있어서 유용한 기법이 되고 있다. 〈표 1〉은 현재 사용되고 있는 데이터 마이닝 기법들간의 상대적인 장점에 대한 분석결과이다[9].

표 4. 데이터 마이닝 기법간의 상대적인 장점
Table. 1 Relative Strenth among data-minings

통계적 기법	인공신경망	사례기반 추론	의사결정 나무	클러스터링
• 풍부한 통계 전문가 • 다양하고 풍부한 통계 소프트웨어 • 결과에 대한 수학적 설명(증명) 가능	• 넓은 문제 영역 • Value Prediction • Classification • Clustering • 다양한 데이터 형태를 처리 • Categorical Variables • Continuous Variables	• 다양한 데이터 형태 지원 • 결과에 대한 설명 용이 • 결과 도출 이유를 구체적인 예를 통해 설명	• 우수한 설명력 • 결과에 대한 rule을 생성 • 상대적으로 우수한 실행 속도 • 해당 답색에 Divide and Aonquer 전략을 사용함으로 수행 속도 빠름 • 많은 항목(Variable)을 갖는 문제영역에 적합	• 데이터의 범주에 관한 지식이 없어도 학습가능 • 데이터가 갖는 고유한 특성을 직관적으로 찾아내는 것이 가능

3. 웹 개인화(Web Personalization)

3.1 웹 개인화의 개념

웹 개인화는 웹에서 제공되는 기본 화면을 사용자의 취향에 맞게 편집하여 볼 수 있는 기능을 비롯해 사용자의 스타일에 맞는 정보를 선별하여 볼 수 있게 해주는 기법을 의미한다. 또한 전자 상거래 업체에서 사용되는 사용자의 개인적 취향에 따라 자신의 페이지를 구성하고 사용자의 구매

기록, 취향에 맞는 제품을 추천 받을 수 있는 기능들까지 포함한다[4][13].

웹 사이트에서의 개인화는 일련의 가치교환 과정이다. 사용자가 자신의 선호, 관심, 구매경험과 같은 정보를 웹 사이트에 제공하면 웹사이트는 사용자가 제공한 자료를 기초로 사용자에게 가장 알맞은 정보를 제공한다. 개인화를 통해서 웹 사이트 운영자는 사용자에 관한 자료를 얻고 사용자의 지속적인 이용이나 구매를 얻어낼 수 있게 되며 사용자는 자신에게 가장 알맞은 정보를 편리한 방법으로 얻을 수 있게 된다. <그림 2>는 웹 개인화의 기본구조에 대하여 설명하고 있다.

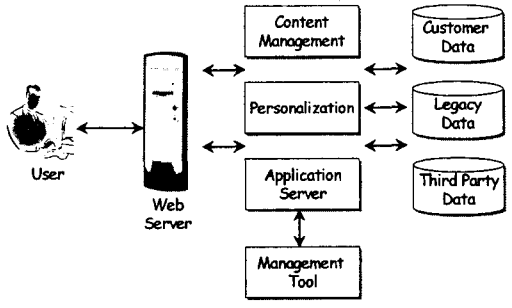

그림 2. 웹 개인화의 기본구조
Fig. 2 Structure of Web Personalization

3.2 웹 개인화의 기법

아직까지는 개인화 방법에 대한 분류 기준이 아직 확립되지 않았지만, 일반적으로 개인화 방법에는 규칙기반 필터링(Rule-based filtering), 협업 필터링(Collaborative filtering), 학습 에이전트(Learning agent) 가 있다. 각각의 방법들은 실행 방법과 비용에서 차이를 가지고 있으며, 한 가지 방법만 사용하는 것이 아니라 위에서 말한 것들 중 두세 가지 방법을 혼합해서 사용한다. 따라서 각 방법들의 비용과 이로 인해 나타나는 직접·간접적인 효과를 정확하게 판단하고 적절한 방법을 선택하는 것이 매우 중요하다[13].

(1) 규칙기반 필터링(Rule-based filtering)

규칙기반 필터링은 인터넷 사이트 사용자에게 연속적인 질문을 던지고, 이에 대한 사용자의 반응과 기존에 존재하는 사용자의 구매나 인구통계정보 등을 활용해서 적절한 아이템(상품, 웹 페이지, 광고 등)을 추천하는 기법으로 사용자에 의해서 입력되고, 사용자에 의해서 생성된 데이터를 활용해서 세밀한 분석과 추론과정을 통해 규칙을 생성하게 된다[4].

(2) 협업 필터링(Collaborative filtering)

협업 필터링은 사용자들의 선호도 정보를 바탕으로 유사한 성향을 가지는 다른 사용자에 의해 높은 선호를 보인 구매 아이템 등을 사용자에게 추천하는 방식이다. 협업 필터링은 사용자의 방문 사이트, 구매정보 등 개인적인 로그 정보들에 의한 개인적 선호도와 연령대, 성별, 거주지 등 일반적 선호도를 기반해서 패턴을 만들어낸다. 그리고 사용자가 선호하지 않은 아이템들에 대해서 같은 분류에 속해 있는 다른 사용자들이 보인 선호도의 가중 평균값의 순서대로 정리하여 높은 선호도의 아이템들을 추천한다[13].

(3) 학습 에이전트(Learning agent)

학습 에이전트는 인터넷 사용자의 행동에 초점을 맞추어 사용자의 성향과 관심을 찾아내고 이에 따라 사용자에게 적절한 정보를 제공하는 기법이다. 학습 에이전트는 주로 사용자의 구매나 장바구니 정보를 통해서 알 수 있는 구매횟수, 사이트 접속 횟수나 시간 등을 이용해서 사용자의 성향과 관심, 인터넷 패턴을 분석하기 때문에 일정량을 넘는 고객의 사용 행태가 관찰되고 학습되어야만 적절한 정보를 제공할 수 있는 단점이 있다. 반대로 사용자의 인터넷 행동에 초점을 맞추기 때문에 사용자에게 많은 정보를 입력받지 않아도 되는 장점이 있다. 학습 에이전트는 규칙기반 필터링과 마찬가지로 사이트가 능동적으로 사용자에게 행동을 요구하게 되고, 사이트가 에이전트를 동작시켜 에이전트가 전달(push)하는 정보를 사용자가 보게 된다[4].

4. 웹 로그 분석(Web Log Analysis)

4.1 웹 로그파일 분석

로그파일 분석은 사용자가 어떤 사이트를 방문한 경우 서버의 로그파일에 흔적을 남기게 되며 이러한 방문자의 정확한 데이터를 기반으로 고객 분석을 통하여 마케팅 피드백을 할 수 있는 고객 분석 방법이다[15]. 이 로그 파일의 분석 결과를 이용하여 웹 사이트 내에서 가장 빈번히 접근되는 페이지나 사용자의 이동 패턴 등을 파악할 수 있으므로 웹 사이트의 정보를 효과적으로 전달하기 위해 로그 파일을 이용하고 있다[29].

이렇게 얻어낸 정보를 바탕으로 인터넷 비즈니스에 전략적으로 활용하고 고객의 다양한 요구를 예측하여 새로운 사이트 개발 및 새로운 시장 기회를 창출하거나 마케팅 및 광고 전략으로서 활용한다. 그리고 최적의 환경에서 사용자들이 사이트를 탐색하고, 방문하도록 서버 및 회선 등의 기술적 자원 및 수행 능력 계획을 수립할 수 있다.

4.2 웹 로그파일 형식

(1) CLF(Common Logfile Format)[15]

CLF[A. Luotonen]는 웹서버의 원조라 할 수 있는 NCSA 계열의 웹서버에서 기본으로 생성되는 로그파일형식으로, 표준로그파일은 보통 Transfer 또는 Access Logfile이라 불리며 파일이름은 access_log와 같이 정한다. 〈그림 3〉은 CLF 로그파일의 실제 기록 예이다.

```
203.247.40.78 - - [11/Jan/2003:19:12:48 +0900] "GET /index.html HTTP/1.0" 200 1622
```

그림 3. CLF 로그파일의 실제 기록 예
Fig. 3 Actuality recording recording example of CLF log file

(2) IIS(Internet Information Server)[15]

MS IIS 로그파일형식은 Windows NT에서 가장 많이 사용되는 웹서버 S/W로서 자체적으로 분석도구를 제공한다. 로그파일 형식은 NCSA 계열의 로그파일과는 다르며, 파일의 기록기간단위 즉 일별, 월별 등의 환경설정도 가능하다. 〈그림 4〉는 IIS 로그파일의 실제 기록 예이다.

```
203.247.40.76, -, 03-03-13, 11:59:48, W3SVC3, EB2M, 203.247.40.78, 254, 332, 391, 897, 200, 0, GET, /default.htm, -,
```

그림 4. IIS 로그파일의 실제 기록 예
Fig. 4 Actuality recording recording example of IIS log file

4.3 개인의 컴퓨터에 있는 로그파일 분석

개인 컴퓨터에 남아있는 URL 접속기록은 Inetrnet Explorer의 경우 history 디렉토리에 저장된다. 저장된 접속정보는 사용자가 접속한 URL주소와 접속페이지 제목 그리고 최종 접속 날짜와 시간으로 분류된다. 또한 접속한 해당 URL 페이지의 등록 정보로부터 접속횟수 정보등과 같은 사용자의 인터넷 항해 정보를 발견할 수 있어서 이것을 이용하여 사용자의 인터넷 성향과 사용자 관점의 정보를 수집할 수 있다.

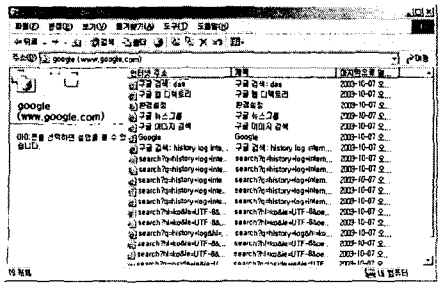

그림 5. History 정보
Fig. 5 History Information

수집된 사용자의 History 정보들은 서버에서 확인할 수 있는 웹 로그보다 그 형태가 단순 하지만, 사용자의 다양한 서버에 대한 접속여부와 접속횟수 등을 알아볼 수 있는 개인 정보로서, 개인의 사용성향 및 관심도를 측정하는데는 매우 유리하다. 다음은 수집된 History 로그를 분류하고 정의된 설정을 기반으로 자동으로 분석하여 사용자의 성향을 분류하는 로그 분석과정을 도식화 한 것이다.

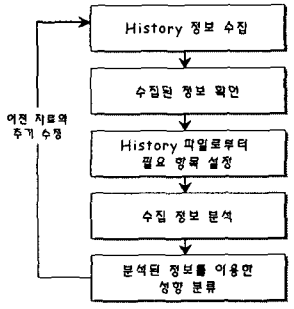

그림 6. History 정보 분석과정
Fig. 6 History Information Analysis Process

4.4 서버의 로그파일과 개인의 컴퓨터의 로그파일의 비교

서버에 남아 있는 로그 파일은 그것이 단일 서버 내에서 사용자별 행동 양식을 알아 낼 수 있다. 그러나 이러한 사용자가 다른 서버에서 다른 서비스를 받을 때는 다른 행동 양식(패턴)을 보일수도 있다. 그러므로 이러한 로그 파일을 이용하면 단일의 서버에서 개인화에 유리하다. 그러나 개인의 URL접속기록을 이용하여 그것의 분류를 통한다면 개인의 정보 요구를 분석하는데 이용하는 데이터의 범위가 크므로 서버의 로그파일을 분석하는 것보다 비교적 좋은 분석의 결과를 도출할 수 있다.

표 2. 서버측의 로그파일 분석과 사용자측 로그파일의 분석의 비교
Table. 2 Comparism between Server LOG File and Client's Local LOG File

	서버측 로그파일 분석	사용자측 로그파일 분석
용도	사이트의 개인화 (단일의 사이트에서 개인화에 적합)	일반적 서비스의 개인화 (사용자의 관심 정도에 따른 복합적 서비스 가능)

III. 시스템 제안 및 사용자 성향 분석

1. 개인화 광고 시스템의 데이터 수집

사용자의 성향 분석을 위한 데이터 수집 방법으로 웹 로그 데이터를 사용하게 된다. 이때 웹 로그 데이터를 수집하는 기법은 주로 웹 서버를 경유한 로그 정보를 중심으로 이루어져 왔다. 그러나 본 논문에서 제안하는 기법은 개인의 컴퓨터에서 웹 사이트를 방문한 로그 데이터를 수집함으로써 개인화에 보다 더 근접한 방법을 제안하고 실험하고자 한다.

본 논문에서 제안하는 데이터 수집 방법은 사용자측 로그에 남아있는 URL을 검색한다. 그리고 검색된 정보를 이용해 URL의 분류를 찾은 후 분류정보를 개인화에 이용한다. 어떤 사람이 특정한 분류에 많은 관심을 보였다면 해당 분류는 통계학적으로 보았을 때 높은 비율을 차지 할 것이다. 이러한 귀납적 방법에 근거하여 사용자의 성향을 판단하고 판단결과를 여러 서비스에 적용시킬 수 있다.

2. 분석을 위한 분류결정 알고리즘 적용

사용자의 로그 정보들이 데이터 풀에 있을 때 로그정보를 이용하여 분류별 정보를 얻어내야 한다. 이때 수집된 정보에서 의미있는 정보를 추출하기 위해 데이터 마이닝 기법(통계적 기법)에 기초하여 다음과 같은 분류결정 알고리즘을 제시하였다. 이러한 일련의 과정은 위에서 언급한 학습 에이전트에 기반 하여 사용자의 인터넷 활동 중 주로 관심있어 하는 분야 중 한 개를 선택하는 기법이다.

```
Weight = 0.0
Total # of counts = A
Do until the last category
Count of a category = a[i]
Ratio = a[i]/A
Weight[i] = Weight[i] + ratio
Loop
```

그림 7. 비중과 비중범위를 결정하기 위한 알고리즘
Fig. 7 Algorithm to decide weight and weight extent

불특정 사용자의 개인화 서비스를 위한 사용자 성향 분석을 위해서 분류정보를 모았다고 가정하자. 이때 〈그림 7〉과 같은 알고리즘을 이용하여 〈그림 8〉과 같이 각각의 분류별 개수정보가 수집되고 수집된 정보를 기초로 하여 분류정보를 분석하게 된다.

```
분류 A: 25, 분류 B: 45, 분류 C: 75, 분류 D: 55
```

그림 8. 비중과 비중범위를 결정하기 위한 분류정보 예
Fig. 8 Tributary information example to decide weight and weight extent

〈그림 8〉의 분류정보를 분석하여 보면 〈표 3〉과 같이 분류별 비율 및 비율범위를 산출해낼 수 있다.

표 3. 통계적 마이닝 예
Table. 3 Example of Stochastic Mining

분류명	분류개수(E)	분류비율(Φ)	분류 비율범위
분류 A	25	25/200 = 0.125	(0.000, 0.125)
분류 B	45	45/200 = 0.225	(0.125, 0.350)
분류 C	75	75/200 = 0.375	(0.350, 0.725)
분류D(last)	55	55/200 = 0.275	(0.725, 1.000)

분류된 정보를 비율범위별로 구분해 보면 〈그림 9〉와 같이 나타낼 수 있는데 각각의 분류 중 분류 C 가 가장 넓은 범위를 차지하고 있음을 알 수 있다. 이때 난수를 발생시켜 발생된 난수가 포함되는 범위의 분류를 선택하게 된다. 단, 난수를 예측할 수 없고 발생한 난수는 항상 최대영역의 분류를 선택하게 되지는 않는다. 하지만 각각의 분류영역에 비례하여 볼 때 가장 넓은 범위를 차지하는 분류가 선택될 확률은 매우 높다. 따라서 분류의 선택 빈도수도 함께 증가하게 된다.

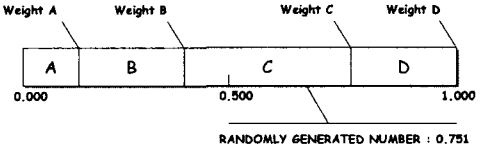

그림 9. 난수를 이용한 한 개의 분류 결정
Fig. 9 Selection of One Cateroy Using Random Number

〈표 3〉과 〈그림 9〉에서 선택된 분류가 사용자의 성향을 분석하는데 매우 중요한 근거로 작용한다. 서론에서 언급했던 웹 개인화의 협업필터링과 학습에이전트의 대상인 사용자들의 선호도와 사용자의 행동을 방문사이트의 분류와 방문횟수로 가정한다.

그리고 분류결정 알고리즘의 적용결과를 이용하여 해당 사이트의 선호도와 방문횟수에 따라 선호하는 광고 분류를 결정지을 수 있다.

3. 사용자의 성향 분석 기법

사용자의 성향 데이터가 계속해서 입력될 때 이전 값과 이후 값의 문제가 생길 가능성이 있다. 만약에 이전의 값을 모두 무시하고 이후의 값만을 적용시킬 경우 성향을 결정하는데 있어 이전 성향들이 무시된다. 이전 값과 이후 값이 서로 같은 비율로 영향을 미치게 하기 위해서는 다음의 과정이 필요하다.

이전 값: δ
이후 값: ε
업데이트 되어야 할 분류 비율
category ratio $(\Phi) = \frac{\delta+\varepsilon}{\sum\delta+\sum\varepsilon}$

그림 10. 연속적인 사용자 성향 데이터에 대한 분석 기법
Fig. 10 Updating Client's Before and Afert Data

다음 〈표 4〉는 위의 식을 적용하기 이전과 이후의 값의 비율을 보여준다. 이러한 결과를 가지고 이전 값이나 이후 값 중 좀더 영향력을 가졌으면 하는 값에 알맞은 상수를 곱하여 영향을 줄 수 있다.

이때 적용하는 상수는 사회통계학적 기법이나 심리학적 요인들을 고려하여 정의하는 것이 좋다.

표 4. 이전값과 이후값 문제 및 해결
Table. 4 Example of Two Data Set(Before and After)

	분류 A	분류 B	분류 C	분류 D
이전 값	20	80	10	70
이후 값	80	30	100	80
이전의 비중	20/180 = 0.11	80/180 = 0.44	10/180 = 0.05	70/180 = 0.38
이후의 비중	80/290 = 0.27	30/290 = 0.10	100/290 = 0.34	80/290 = 0.28
제안하는 기법	(20+80)/ (180+290) = 0.21	(80+30)/ (180+290) = 0.23	(10+100)/ (180+290) = 0.23	(70+80)/ 180+290) = 0.31

4. 개인화 광고 테이블을 이용한 광고 결정

사용자의 성향 분석을 통해 얻은 결과에 따라 각각의 분류별 광고를 정의하고 각각의 광고에 대해 다음과 같은 가중치 계산법을 적용하여 분류별 광고의 비중과 비율범위를 결정짓게 된다.

이중 가중치가 높은 광고에 대해 우선선택권이 부여되며 이는 해당광고 선택되어 서비스 되는 것을 의미한다.

표 5. 개인화 광고 테이블 생성을 위한 가중치 계산
Table. 5 Calulation of Weights for Advertisement Table

AD Name	Ad Amount (E)	Ad unit Cost	Ad Ratio (Φ)	Ad Weignt Range
Ad A	α	α'	$\frac{\alpha}{\sum E}$	$[0, \frac{\alpha}{\sum E})$
Ad B	β	β'	$\frac{\beta}{\sum E}$	$[\frac{\alpha}{\sum E}, \frac{\alpha+\beta}{\sum E})$
...	...		...	...
Ad X (last)	χ	χ'	$\frac{\chi}{\sum E}$	$[1-\frac{\chi}{\sum E}, 1]$

알맞은 광고가 선택이 되면 광고가 다 나간 뒤 해당 광고의 총 금액을 한 회의 광고비만큼 빼고 전체 비용에서 차지하는 비중을 다시 계산하여 비중의 범위를 재설정하고 다음의 광고의 요청이 있을 때 신규로 적용된 비중의 범위로 새로운 광고를 선택한다.

표 6. 개인화 광고 테이블의 예
Table. 6 Example of Personalized Advertisement Table

광고 ID	총 광고비	회당 광고비	비 중	범 위
Adver A	1000	1	1000/10000 = 0.1	(0.0, 0.1)
Adver B	2000	1	2000/10000 = 0.2	(0.1, 0.3)
Adver C	3000	1	3000/10000 = 0.3	(0.3, 0.6)
Adver D	4000	1	4000/10000 = 0.4	(0.6, 1.0)

Ⅳ. 시스템 설계 및 실험

1. 개인화된 광고 시스템 시나리오

광고 시스템에는 크게 두 가지 측면이 있다. 하나는 사용자가 접속한 URL정보를 기반으로 데이터 풀을 만드는 것이며 생성된 풀에 통계학적인 기법을 적용하여 한 개의 광고를 선택하는 것이다.

본 논문에서 제안한 시스템을 실험하기 위하여 사용자용 프로그램(Category Retriever)을 사용하였다. 사용된 프로그램을 통해 사용자의 접속 URL 정보를 추출하고, 추출된

정보를 분류결정 알고리즘을 적용하여 한 개의 분류 정보를 고르고 그 분류 중에서 한 개의 광고를 고르는 서버측 프로그램을 네트워크 환경에서 실험하였다. 〈그림 11〉은 개인화 광고 시스템의 사건 추적도 이다.

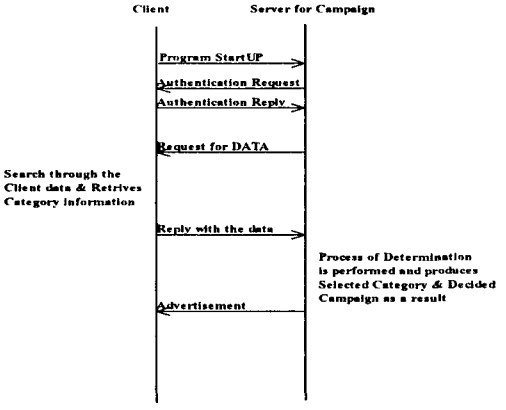

그림 11. 개인화된 광고 시스템 사건 추적도
Fig. 11 System Event Trace Diagram

사용자의 웹 로그 데이터와 사용자의 개인 정보를 가져와 사용자의 사이트 방문 패턴을 분석하고, 사이트의 접속 횟수에 대한 데이터를 이용한 선호도를 분류한 뒤 각 난수를 발생하여 각 선호 비율에 비례한 광고를 제공하도록 하였다. 다음은 개인화된 광고시스템의 구성도이다.

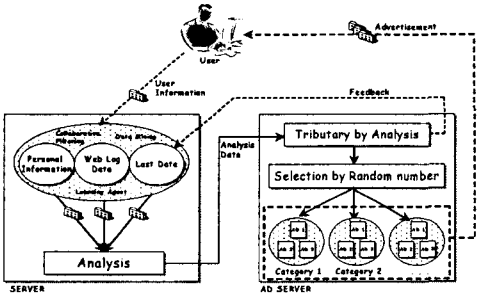

그림 12. 시스템 구성도
Fig. 12 System Structure Overview

2. 수집된 정보 수집 및 분류

어떤 광고에 대한 선호도를 분석하기 위해 사용자의 개인 정보 또는 웹 로그 데이터를 이용하였다. 사용자의 개인 정보는 광고 시스템에 접속 후 개인 정보를 입력함으로써 획득될 수 있고, 웹 로그 데이터는 사용자의 시스템에 남겨져 있는 history data를 사용하게 된다.

네트워크 상의 각각의 컴퓨터는 모두 각기 다른 고객이라고 생각하고 성향 결정과 광고의 실험을 수행하였으며 3장에서 제시한 사용자별 성향 분석 기법을 적용하였다. 다음 〈표 7〉은 사용자 컴퓨터에서 검색해낸 URL을 가지고 그것의 분류를 추출한 후 각각의 사용자가 분류에 대하여 비율에 대한 실험 결과이다.

표 7. URL 분류 비율
Table. 7 URL Category Ratio

Category \ UserID	PC008	pc009	pc013	pc017	pc019	pc021
A(1)	8.78	7.25	16.52	5.36	7.07	7.65
B(2)	9.46	7.73	0.00	30.93	12.45	15.51
C(3)	0.00	0.97	0.00	0.41	0.46	0.87
D(4)	43.24	10.14	25.22	13.20	23.02	27.61
E(5)	2.03	10.63	9.57	10.72	8.45	9.83
F(6)	0.00	0.48	0.00	0.00	0.21	0.34
G(7)	8.11	28.50	21.74	8.25	9.52	4.97
H(8)	0.68	4.83	9.57	6.60	5.74	3.29
I(9)	3.38	10.14	7.83	2.68	9.23	6.97
J(10)	1.35	0.00	0.00	1.86	2.31	0.39
K(11)	12.16	0.00	5.22	0.82	4.20	2.13
L(12)	8.11	3.86	0.87	5.77	10.66	13.63
M(13)	1.35	4.83	0.00	12.78	5.43	4.86
N(14)	1.35	10.63	3.48	0.62	1.25	1.95

〈표 7〉에서 나타난 정보는 전체의 정보가 아니고, 전체 정보 중 특징적인 정보 몇 개를 추출하여 나타낸 것으로서 한 명의 고객(한대의 컴퓨터)에서는 20회 이상의 각기 다른 광고를 볼 수 있었고, 광고는 14개의 항목에 무작위로 위치하고 있으며 단 각각의 항목의 광고비의 총합이 445단위이상 597단위이하로 하고 분류의 광고비의 평균이 527단위 이고 모든 광고는 한번의 광고 당 5단위씩 떨어지게 설정하였다.

3. 정보 분석

사용자별로 〈표 7〉의 정보를 기반으로 광고를 선택된 분류별 비율의 실험 결과는 〈표 8〉과 같다. 〈표 8〉의 광고가 선택된 분류별 비율의 실험결과와 〈표 7〉의 입력에 대한 결과와 비슷한 비율로 나타난 것을 확인할 수 있었다. 본 논문에서 제안한 가정에 의하면 모두 같은 비율로 나와야 정상이지만 광고주의 광고비가 모두 소진하면 그 항목의 광고는 광고가 서비스가 되지 않는다는 결과를 확인하였으며, 광고의 비용이 무한하다면 이러한 차이는 나타나지 않았다.

표 8. 광고가 선택된 분류별 비율
Table. 8 Selected Advertisemnt Ratio

Category \ UserID	PC008	pc009	pc013	pc017	pc019	pc021
A(1)	9.09	5.94	13.79	5.32	5.34	8.26
B(2)	7.44	8.91	0.00	14.89	7.63	10.74
C(3)	0.00	3.96	0.00	2.13	1.53	3.31
D(4)	14.05	2.97	9.48	7.45	7.63	7.44
E(5)	2.48	14.86	9.48	12.77	6.11	7.44
F(6)	0.00	0.00	0.00	0.00	3.82	3.31
G(7)	3.31	16.83	13.79	4.26	5.34	6.61
H(8)	3.31	5.94	18.97	9.57	5.34	5.79
I(9)	8.26	4.95	9.48	4.26	6.87	7.44
J(10)	14.05	0.00	0.00	7.45	13.74	5.79
K(11)	19.01	0.00	9.48	1.06	10.69	5.79
L(12)	11.57	1.98	0.86	6.38	14.50	10.74
M(13)	0.83	7.92	0.00	24.47	8.40	8.26
N(14)	6.61	25.74	14.66	0.00	3.05	9.09

4. 분석 결과

본 논문에서 제안한 광고 시스템에서는 광고가 시간대별(Hit Count)에 따라 각 분류별 총 광고 금액의 변화를 광고주의 광고금액의 hit(시간)수에 대한 변화량의 형태에 따라 크게 인기 있는 항목, 비인기 항목 그리고 중간 정도의 인기를 가지는 항목으로 분류하였다.

각 항목의 총 투자한 금액의 hit에 따른 변화추이에서 보듯이 초기에 총액이 450~600 사이의 값을 가지고 있으면서 시작을 하였고, 어떠한 항목이 먼저 0으로 값이 내려가면 다른 항목의 값을 자동으로 선택하게 하는 시스템이다. 특이하게 F 항목은 실험결과에서 나타나듯이 처음부터 광고가 선택이 안되다가 다른 것들이 모두 0에 수렴(convergence)한 뒤에 광고가 서비스되는 것을 보였다.

항목 B와 D는 인기 있는 항목으로 분류할 수 있는데, 그 이유는 히트수가 370회가 될 때까지 모든 광고비가 소진되었음을 나타내었고, 전체의 14%를 차지하는 것을 나타내었다. 370회에서 610회 사이에 떨어진 모든 중간 정도의 인기를 가지고 있는 항목은 총 9개이며, 전체의 64%를 차지한다. 또 가장 인기 없는 항목 또한 3개이고, 전체의 21%를 차지하는 것을 <그림 13>과 같이 실험결과를 통하여 확인하였다.

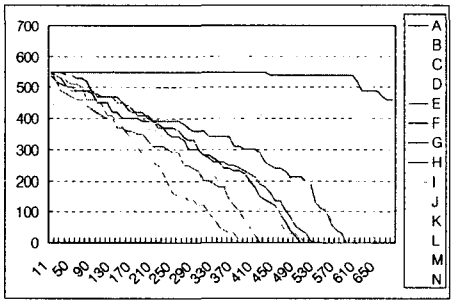

그림 13. 시간의 흐름에 따른 광고비 변화 실험결과
Fig. 13 The Change of Total Investement for each Category accrding to Time Flow

V. 결론 및 향후 연구방향

본 논문에서는 개인화된 광고 서비스를 위한 데이터 수집기법으로 사용자 측 로컬시스템의 웹 로그파일을 이용하여 사용자의 선호도와 성향을 분석하였고, 수집된 정보의 분류 및 군집화를 위한 방법으로 분류결정 알고리즘을 적용하여 사용자별로 차별화된 광고 서비스가 가능한 방법을 제안하고 실험하였다.

또한, 새로운 방문자에 대한 사용자들의 선호도 정보나 행동양식에 대한 예측을 기초 데이터로 하여, 고객들에게 선정된 광고에 사용자 성향분석과 동일한 과정을 적용시킴으로써 고객이 어떤 상품을 구매하고 싶어하는지를 먼저 예측하여 적절한 광고를 보낼 수 있게 되는 것이다. 이는 인터넷이 기업의 대고객 접점으로서 중요성이 증가되는 현시점에서 쇼핑몰이나 웹사이트 광고기업의 eCRM을 위한 윈투윈 마케팅의 중요한 기초 데이터로 활용될 수 있다.

향후 연구과제로는 분류 데이터 풀의 이전 값과 이후 값을 적용하는데 있어서 인구 통계적, 심리적 영향을 참고하여 알맞은 상수를 찾을 수 있게 되면 사용자의 성향을 분석하는데 있어서 좀더 발전된 형태의 연구가 필요하고, 온톨로지와 사용자 프로파일을 이용한 개인화 광고 시스템과 동적인 웹 정보집합에서 유전자 알고리즘(Genetic algorism)을 이용한 최적화된 정보 서비스의 연구가 필요하다.

참고문헌

[1] 전자신문, "http://www.etnews.co.kr/"
[2] 코리아 인터넷 마케팅센터, "http://www.webpro.co.kr/"
[3] 오픈타이드, "Analytical eCRM 소개", 2001.
[4] 웹 개인화, "http://www.personalization.co.kr"
[5] Berry, M. J. and G. Linoff, "Master Data Mining: The Art and Science of Customer Relatinship Management", John Wiley & Sons, 2000.
[6] Ogata H., Kaneko M., Hakamazuka A., Orito A. and Sato K., "eCRM Service", NEC TECHNICAL JOURNAL, pp. 69~72, 2001.
[7] B.Mobasher, N. Jain, E-H Han, and J. Srivastava, "Web Mining Pattern Discovery from World Wide Web Transactions", Tech Report. 1996.
[8] Magdalini Eirinaki, Michalis Vazirgiannis, "Web mining for web personalization" ACM

Transactions on Internet Technology(TOIT), Volume 3 Issue 1, February 2003.

[9] Michael Goebel, Le Gruenwald, A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations Newsletter, Volume 1 Issue 1, June 1999.

[10] B. Mark, eds., "Data Mining : Getting the Nuggets", Special Issue, IEEE Export, Oct.1996.

[11] U. Fayyad and R. Uthurusamy, eds., "Data Mining", Special Issue, Communications of the ACM, 39(11), Nov. 1996.

[12] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Ting, "Web usage mining : Discovery and applications of usage patterns from web data", SIGKDD Explorations, vol.1, no. 1-2, pp12~23, 2000.

[13] Maurice D. Mulvenna, Sarabjot S. Anand, Alex G. Buchner, "Personalization on the Net using Web mining",Communications of the ACM, Volume 43 Issue 8, August 2000.

[14] Gustavo Rossi, Daniel Schwabe, Robson Guimaraes, Designing personalized web applications Proceedings of the tenth international conference on World Wide Web, April 2001.

[15] A. Luotonen, The common log file format, "http://www.w3.org/"

[16] Elsenpeter, R. C. and Velte, T. J., e-Business A Beginner's Guide, McGraw-Hill, Berkeley 2001.

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

[18] 김재문, E-비즈니스 모델에 맞는 eCRM 구축 실행 가이드, 거름, 2001

[19] Brown, S. A., Customer Relationship Management, John Wiley & Sons, 2000.

[20] Wendy Gersten, Rudiger Wirth, Dirk Arndt, "Predictive modeling in automotive direct marketing: tools, experiences and open issues" Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, August 2000

[21] Vinod Anupam, Richard Hull, Bharat Kumar, "Personalizing E-commerce applications with on-line heuristic decision making", Proce edings of the tenth international conference on World Wide Web, April 2001

[22] Andreas Rauber, Alexander Muller-Kogler, "Integrating automatic genre analysis into digital libraries", Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, January 2001.

[23] Usama Fayyad , Gregory Piatetsky-Shapiro , Padhraic Smyth, The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM, v.39 n.11, p.27-34, Nov. 1996.

[24] 김형도, "개인화된 웹 광고를 지원하기 위한 요구 통합조정 체계의 설계", 한국정보처리학회 논문집 제 6권 제 6호, 1999.

[25] 김재형, 노효원, 김남호, 정정화, "인터넷 비즈니스 기반의 고객관계관리(CRM)을 위한 웹 로그 분석에 관한 연구", 한국정보처리학회 춘계 학술발표 논문집 제 7권 제 1호, 2000.

[26] 장형진, 최성, 한정란, 이기민, "데이터마이닝을 이용한 eCRM", 한국정보처리학회 학회지, 2001.

[27] 박남섭, 김정범, 이윤정, 김태윤, "eCRM을 위한 지능형 배너 광고 관리 및 분석 시스템의 설계 및 구현", 한국정보과학회 춘계 학술발표 논문집, pp.319~321, 2002.

[28] 박성준, 김주연, 김영국, "분산 이기종 인터넷 쇼핑몰 환경에서의 벡터 모델 기반 개인화 서비스 시스템", 한국정보과학회 논문집, 제 8권 제 2호, pp.206~218, 2002.

[29] 정현섭, 양재영, 최중민, "개인화된 웹 네비게이션을 위한 온톨로지 기반 추천 에이전트", 한국정보과학회 논문집, 제 30권 제 1호, pp.40~50, 2003.

[30] 이치훈, 고세진, 김용환, 이필규, "웹 마이닝과 협력적 정보 여과를 이용한 개인화 서비스의 성능 개선 방안", 한국정보과학회 춘계 학술발표 논문집, pp.63-65, 2000.

저 자 소 개

김 은 수

1994년 서울산업대학교 시각디자인과(이학사)
1997년 서울산업대학교 대학원 시각디자인과(이학석사)
2000년 ~ 현재 한남대학교 대학원 컴퓨터공학과 (박사과정)
2001년 ~ 현재 한국과학기술정보연구원 위촉연구원
〈관심분야〉 웹디자인, 에니메이션, 웹마이닝, 개인화

이 원 돈

1979년 서울대학교 화학과 졸업 (이학사)
1982년 일리노이대(Urbana) 화학과 졸업(이학석사)
1986년 일리노이대(Urbana) 전산학과 졸업(이학박사)
1987년 텍사스대(Arlington) 전산공학과 조교수
1987년 ~ 현재 충남대학교 컴퓨터 과학과 교수
〈관심분야〉 design and application of neural networks, the development of optimization techniques, and machine leaning

송 강 수

2002년 충남대학교 컴퓨터과학과 (이학사)
2002년 ~ 현 재 충남대학교 대학원 컴퓨과학과 (석사과정)
〈관심분야〉 bio informatics, machine leaning

송 정 길

1966년 한남대학교 수학과(이학사)
1982년 홍익대학교 대학원 전자계산학과(이학석사)
1988년 중앙대학교 대학원 전자계산학과(이학박사)
1990년 ~ 1991년 University of illinois 객원교수
1979년 ~ 현재 한남대학교 정보통신 · 멀티미디어 공학부 교수
〈관심분야〉 XML, 웹서비스, 웹마이닝, 개인화, 분산처리시스템