웹로그 마이닝을 이용한 개인화 광고 서비스 기법

김석훈† · 김은수††

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 개인화된 광고 서비스를 가능하게 하는 비교적 간단한 적응형 시스템을 설계하고, 그 성능을 실험하였다. 개인의 성향을 시스템에 가장 효율적으로 반영하기 위하여 개인 컴퓨터의 히스토리 파일을 원시 데이터로 하여 정제후 사용하여 이 파일을 이용하므로 해당 서버를 방문한 자에 한해서만 성향을 파악할수 있는 단점을 극복하여 고객이 다른 서버의 방문 기록도 활용하므로 좀 더 현실성 있는 성향 파악이 가능하게 하였다.

키워드 : 웹로그 마이닝, 개인화, 웹로그. 광고

Personalized Advertisement Service Method Using Web Log Mining

Seok-Hun Kim† · Eun-Soo Kim††

ABSTRACT

Numerous internet pop advertisement are being provided according to the rapid development of e-commercial and a rise in users. However, it has not been based on analysis of users' incli nation but just one-sided providing. With that reason, many web-site provider want to advertis e more efficient and distinguished Internet-advertisement as analyzing Server's Log accessed. I n this thesis, we have studied and tested relatively simply adoption system to provide personali zed advertisement service. In order to influence personal disposition to system as the most effe ctive way, it first of all uses History files as source data and after refining it, it can search no t only visitors' inclination but also the others' visit-list on the other server. As a result of it, i t can make advertisement more reality and activity.

Keywords : WebLog Mining, Personalized, WebLog, Advertisement

1. 서 론[1)]

한국인터넷정보센터[2]에서 발표한 결과에 의하면, 인터넷을 이용한 전자상거래를 비롯한 생활의 여러 분야에서 인터넷 이용률이 계속적으로 증가하고 있다[1,2]. 또한 인터넷을 통한 정보의 폭발적 증가로 인해 하루에 생성되는 정보와 데이터 량을 감시하는 것조차 불가능해졌고, 통계기법이나 간단한 질의만으로도 충분했던 과거와는 달리 유용한 정보를 찾아내는 것은 상당한 시간적, 기술적 노력이 필요해졌다. 이로 인해 시스템 안에 수록된 방대한 과거 데이터로부터 의미있는 결과를 끌어내기 위한 좀더 전략적인 방법을 모색하게 되었다[3].

최근 인터넷 광고 시스템이 지닌 가장 큰 특징

† 준 회 원: 한남대학교 대학원 컴퓨터공학과 박사과정 (교신저자)
†† 비 회 원: 한남대학교 대학원 컴퓨터공학과 공학박사

논문접수: 2004년 7월 15일, 심사완료: 2004년 11월 10일

중의 하나는 바로 고객의 관심과 행동 양식을 파악하여 사용자 만족을 극대화시키는 개인화(personalization)[5] 서비스 이다. 이러한 개인화 서비스를 위한 데이터 분석 요구와 웹 구조 및 로그 등의 분석을 위한 웹 마이닝[8]에 관한 연구가 활발히 진행되고 있으나, 웹에서 입력되는 대용량 데이터에 대한 분석은 아직도 미약한 상태이다[2,5]. 그 이유는 웹에서 획득된 데이터의 신뢰도가 낮아 좋은 분석 결과를 기대하기 힘들며 통계와 같은 기존의 분석 방법을 적용하기에는 많은 어려움이 있기 때문이다.

본 논문에서는 개인의 성향을 시스템에 가장 효율적으로 반영하기 위하여 개인 컴퓨터의 히스토리 파일을 이용하여 데이터 수집 및 분석하여 데이터를 정제, 발견, 변환하고자 한다. 또한 이렇게 전처리된 데이터로 고객 세분화, 고객 취향의 카테고리 분류를 위하여 야후(Yahoo)의 검색 시스템을 이용한 방법으로 성향 카테고리를 획득하여, 사용자별로 차별화된 광고 서비스가 가능한 적응형(adaptive) 시스템을 설계하고, 실험하여 평가 하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 관하여 서술하고, 제 3장에서는 개인 광고 시스템을 위한 제안한 방법 및 알고리즘에 대해 기술한다. 제 4장에서는 3장에서 언급한 알고리즘과 성향 분석 기법을 이용한 시스템을 설계하였다. 5장에서는 실험을 통한 분석 결과를 기술하고, 마지막으로 제 6장에서 결론 및 향후 연구방향에 대하여 서술한다.

2. 관련연구

2.1. 웹 마이닝을 위한 로그 분석

로그화일 분석은 사용자가 어떤 사이트를 방문한 경우 서버의 로그화일에 흔적을 남기게 되며 이러한 방문자의 정확한 데이터를 기반으로 고객 분석을 통하여 마케팅 피드백을 할 수 있는 고객 분석 방법이다[15]. 이 로그 화일의 분석 결과를 이용하여 웹 사이트 내에서 가장 빈번히 접근되는 페이지나 사용자의 이동 패턴 등을 파악할 수 있으므로 웹 사이트의 정보를 효과적으로 전달하기 위해 로그 화일을 이용하고 있다. 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[10].

(1) 액세스 로그(Access Log)

웹 사이트 방문자는 웹 브라우저를 통해 해당 사이트를 방문하게 되는데, 이때 브라우저가 서버에 파일을 요청한 기록을 시간과 IP 등의 정보와 함께 남긴 것을 액세스 로그라고 한다. 서버로부터 브라우저에 파일이 전송된 기록이므로 액세스 로그를 트랜스퍼 로그(Transfer Log)라고도 한다. 현재 액세스 로그를 기록하는 표준은 NCSA의 "Common Log Format"을 따르는데 실제 기록 예는 <그림 1>과 같다[10].

```
208.247.40.78 - - [10/Jan/2003:19:12:48 -0300] "GET /index.html HTTP/1.0" 200 1622
```

<그림 1> CLF 로그화일의 실제 기록 예

(2) 에러 로그(Error Log)

로그항목에서 에러 로그는 중요한항목 중 하나이다. 에러에 대한 정보를 기록하지 않을 경우 무엇이 잘못되었는지, 어디에 잘못이 있는지를 알수 없게 된다. 에러 로그는 웹 서버의 오작동에 대한 모든 정보를 포함하고 있다. 에러 로그가 발생하는 경우는 화일이나 이미지들을 잘못 링크하여 존재하지 않는 화일인 경우나 CGI 프로그램이 정상적으로 작동하지 않은 경우 서버의 퍼미션(Permission) 설정을 제대로 부여하지 못하여 정상적으로 서버에 기록되지 못한 경우 등이다[10].

(3) 레퍼럴 로그(Referrer Log)

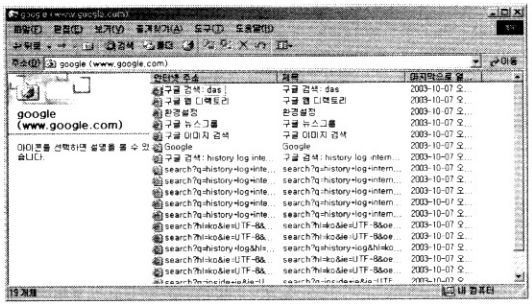
트렌스퍼 로그를 보면 사용자가 해당 사이트에서 어떤 페이지를 보았는지 알수 있는 반면에 레퍼럴 로그에는 그 페이지를 보기 위해서 어떤 페이지를 거쳐왔는지에 대한 기록이 남아 있다. 이 로그를 살펴보면 사용자들의 웹 사이트를 찾아오기 위해 어떤 검색엔진을 이용하는지에 대한 정보를 알 수 있고, 사이트의 구조상 어떤 페이지들이 Navigation을 도와주는 역할을 하는 페이지들인지 알 수 있다. 만약 어떤 사용자가 야후(Yahoo) 검색엔진을 이용해 사이트를 찾아 왔다면 레퍼럴 로그에는 <그림 2>와 같은 기록이 남을 것이다.

```
http://search.yahoo.com/bin/search?p=data+mining+websites → /index.html
```

<그림 2> 레퍼럴 로그 실제 기록 예

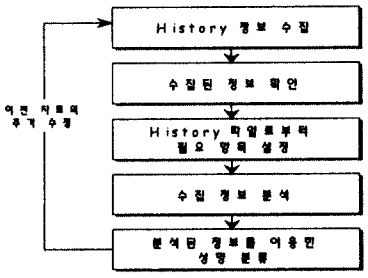
2.2. 개인의 컴퓨터에 있는 로그파일 분석

개인 컴퓨터에 남아있는 URL 접속기록은 Internet 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 3> Hstory 정보

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. <그림 4>는 수집된 History 로그를 분류하고 정의된 설정을 기반으로 자동으로 분석하여 사용자의 성향을 분류하는 로그 분석과정을 도식화 한 것이다.



<그림 4> Hstory 정보 분석과정

2.3. 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 1> 서버측의 로그파일 분석과 사용자측 로그파일의 분석 비교

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

3. 제안한 방법 및 알고리즘

3.1. 클라이언트 데이터 수집 및 정제

사용자의 성향 분석을 위한 데이터는 개인 컴퓨터에 있는 히스토리(history) 데이터를 사용한다. 이때 히스토리 데이터를 수집한 후 정제과정에서 필요한 데이터만 전처리 과정을 통하여 선별한다.

(1) 데이터 수집

데이터 수집은 사용자 컴퓨터에 저장된 웹 사이트 접속 기록 데이터가 기록된 히스토리 화일을 이용한다. 이때 얻을 수 있는 정보는 방문 URL, 접근 횟수, 최종 접속 날짜 등으로 구성되어 있다. Win32 계열의 운영체제에서 인터넷 익스플로러를 사용하는 경우 물리적인 히스토리 화일의 위치는 <그림 5>와 같다.

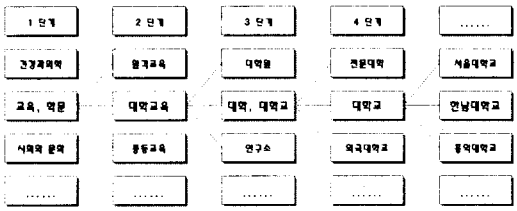
C:\Documents and Settings\Administrator\Local Settings\History

<그림 5> URL 로그화일 위치

(2) 사용자 성향 카테고리 발견

사용자 성향 카테고리를 발견하기 위하여 히스토리 화일의 모든 URL의 카테고리를 발견하면 된다. 각 URL의 카테고리를 발견하기 위해서 야후(Yahoo)의 검색시스템을 이용한다. 왜냐하면

야후의 디렉터리 서비스는 현재 많은 검색엔진에서 이용하고 있으며 이것은 각각의 홈페이지의 성격을 분석하여 분류별로 나누어진 것을 이용함으로써 표준적인 카테고리 데이터를 얻을수 있기 때문이다. 별도 자체 다른 검색 시스템을 설계하고 설치하여 유지하기 위해서는 너무 복잡하고 방대하므로 중소규모 시스템에 이용하기 위하여 이 방법을 선택하였다. 야후의 분류방식은 14개의 대 분류로 구성되어 있으며, 각각의 대 분류는 그 안의 소분류로 나뉘어져 있다. 예를 들어 한남대학교의 URL인 http://www.hannam.ac.kr은 야후에서 그 분류가 <그림 6>과 같다. 한남대학교의 야후에 의한 대분류부터 "교육, 학문 > 대학교육 > 대학 > 대학교"에 위치하고 있고, 리스트로 표현한 것이다.



<그림 6> 한남대학교 분류 예

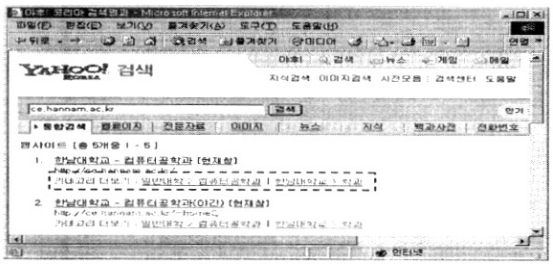
이 방법을 구현하기 위하여 Wget 프로그램을 이용한다. 사용한 정보는 URL 로그화일이며 이것을 가지고 야후에 질의 하여 정보를 수집한다.

Wget을 사용하여 야후에 검색어로 http://ce.hannam.ac.kr을 입력하여 사용하는 방법은 <그림 7>과 같다.

```
Wget http://kr.search.yahoo.com/search?p=http://ce.hannam.ac.kr
```

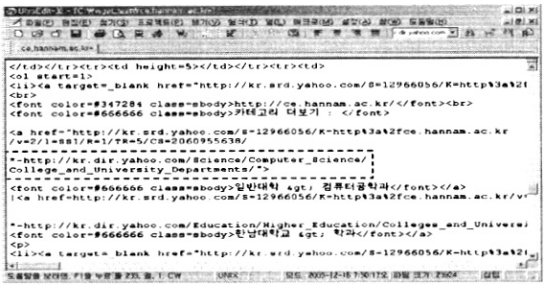
<그림 7> Wget 프로그램 사용 예

위의 명령을 수행결과는 HTML 형식으로 수신하여 저장한다. 예를 들어 "한남대학교 컴퓨터 공학과" 분류 정보를 발견하기 위한 야후 검색결과 화면은 <그림 8>과 같다.



<그림 8> 검색 결과 페이지

<그림 9>는 <그림 8>의 소스 리스트이다. 분류 정보는 <그림 9>의 리스트 중 에서 "*-http://kr.dir.yahoo.com/Science/Computer_Science/College_and_University_Departments/" 이다.



<그림 9> 분류 정보 화일 데이터

3.2. 사용자 성향 카테고리 결정

개인별 카테고리 집합은 서로 교집합이 없다고 가정하고, 전체집합을 Σ라 하고, 카테고리 집합을 Ai라 <그림 10>과 같이 가정한다.

전체집합 : U
부분집합 : A

$$U = \sum_{i=1}^{N} A_i$$

<그림 10> 전체와 부분집합

즉, 전체 집합을 U라고 하고 이것의 부분집합을 A라고 정의하며, n(U)=n(A)+n(B)+n(C)+n(D)+n(E)+n(F)+n(G)를 만족한다고 가정한다. 각 카테고리별 Weight Ai 는 n(Ai)/n(U)로 표시하고, 분류비율 범위

$$\text{range(Ai)=ratio}[(\sum_{i=0}^{i} R_{(Ai-1)}),\ (\sum_{i=1} R_{(Ai)})]$$

로 표시한다. 이것을 표현한 알고리즘은 (그림 11)과 같다.

```
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 11> 비중과 비중범위를 결정하기 위한 알고리즘

<표 2>는 알고리즘 <그림 11>을 적용하여 이러한 각각의 부분집합을 계수하여 각각의 비중을 구하고, 그 비중에 따라서 비중 범위를 계산하는 과정을 보여주고 있다. 개인별 성향 테이블에서 초기의 분류값 α에 대한 비중(weight)을 구하고, 범위를 구하고, 그 다음 분류값 β의 비중과 범위를 구하는 과정이다. 이러한 방법으로 모든 분류값에 대한 비중과 범위를 구한다.

<표 2> 개인별 성향 테이블

카테고리	광고금액(E)	비 중(Φ)	범 위
A	α	$\frac{\alpha}{\sum E}$	$[0, \frac{\alpha}{\sum E})$
B	β	$\frac{\beta}{\sum E}$	$[\frac{\alpha}{\sum E}, \frac{\alpha+\beta}{\sum E})$
...	...	...	...
X (last)	χ	$\frac{\chi}{\sum E}$	$[1-\frac{\chi}{\sum E}, 1]$

예를들면 불특정 사용자의 개인화 서비스를 위한 사용자 성향 분석을 위해 <그림 12>와 같은 분류정보를 모았다고 가정하자.

이때 <그림 11>의 알고리즘을 이용하여 <표 2>와 같이 각각의 분류별 개수정보가 수집되고 수집된 정보를 기초로 하여 분류정보를 분석하게 된다.

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

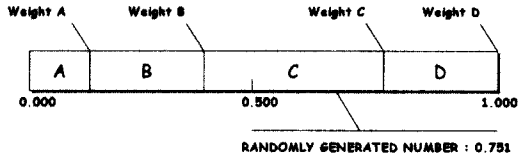
<그림 12> 비중과 비중범위를 결정하기 위한 분류정보 예

<그림 12>의 분류정보를 분석하면 <표 3>과 같이 분류별 비율 및 비율범위를 산출해낼 수 있다.

<표 3> 개인별 카테고리 테이블

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]
분류 D (last)	55	55/200 = 0.275	[0.725, 1.000]

분류된 정보를 비율범위별로 구분해 보면 (그림 13)과 같이 나타낼 수 있는데 각각의 분류 중 분류 C가 가장 넓은 범위를 차지하고 있음을 알 수 있다. 이때 난수를 발생시켜 발생된 난수가 포함되는 범위의 분류를 선택하게 된다.



<그림 13> 난수를 이용한 한 개의 분류 결정

<그림 13>에서 확인할 수 있듯이 광고 카테고리를 선정하기 위해서 난수(Random No)를 발생시켜 하나를 선정한다. 위의 예에서는 난수가 0.751이 발생되었고, 그러므로 카테고리 C가 선정된 것을 나타낸다.

3.3. 사용자 성향 수정

사용자의 성향을 주기적으로 업데이트 하여 새로운 성향을 파악하기 위하여 <그림 14>와 같은 과정을 수행한다.

δ: 이전 값
ε: 이후 값
업데이트 되어야 할 분류 비율

$$\text{category ratio } (\Phi) = \frac{\delta+\varepsilon}{\sum\delta+\sum\varepsilon}$$

<그림 14> 연속적인 사용자 성향 데이터에 대한 분석 기법

<표 4>는 위의 식을 적용하기 이전과 이후의 값의 비율을 보여준다.

<표 4> 이전값과 이후값 문제 및 해결

	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

3.4. 광고 결정 기법

광고 결정은 사용자의 성향 분석을 통해 얻은 결과에 따라 각각의 분류별 광고를 정의하고 각각의 광고에 대해 <표 4>와 같은 계산법을 적용하여 분류별 광고의 비중과 비율범위를 결정짓는

다. 이중 가중치가 높은 광고에 대해 우선선택권이 부여되며 이는 해당광고 선택되어 서비스되는 것을 의미한다. 광고선정은 전단계에서 카테고리가 결정되면 해당 카테고리에 <표 5>의 테이블에서 난수(Random No)를 이용하여 범위중 하나를 선택한다.

<표 5> 카테고리별 광고 테이블

광고주	광고금액 (E)	회당 광고비	비 중 (Φ)	범 위
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.

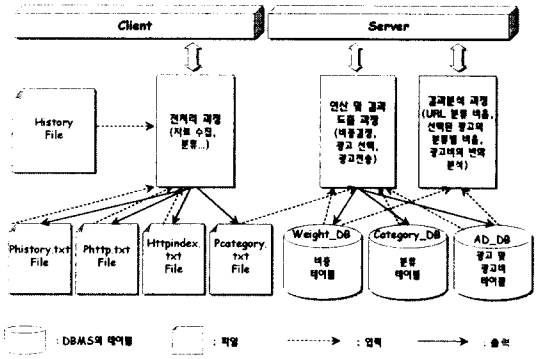
<표 6> 개인화 광고 테이블의 예

광고주	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]

4. 개인화 광고 시스템 설계

4.1. 시스템 구성

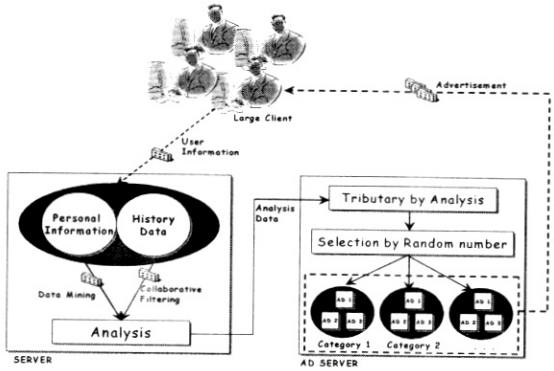
본 논문에서 제안한 개인화 광고 시스템은 그 처리 과정을 기능에 따라 크게 정보 수집 및 분류 과정, 전처리 과정과 결과 분석 과정으로 구분하여 각각 하나의 시스템 안에서 독립적으로 수행되도록 설계하였다. 본 시스템은 클라이언트와 서버부분의 작업으로 구성한다. 클라이언트 부분은 개인별 성향 데이터의 수집, 정제, 카테고리 분류를 수행하고, 서버 부분에서는 크게 분석 기능과 광고 기능을 수행한다. 분석기능에서는 여러 클라이언트로부터 사용자 성향 데이터를 수집, 정리, 분류, 분석을 하고, 광고기능에서는 광고 데이터 수집, 광고 선택, 광고 전송 서비스, 결과 표시(visualization) 등을 수행한다. 개인화된 광고시스템의 구성도는 <그림 15>와 같다.



<그림 15> 시스템 구성도

4.2. 시스템 설계

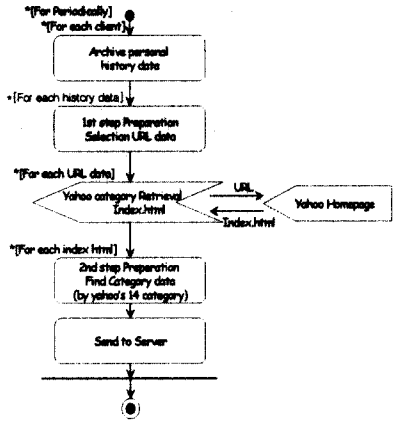
본 시스템은 클라이언트의 사용자 성향 정보와 히스토리 데이터를 전처리 과정을 거쳐서 서버에서 URL 히스토리 화일을 분류정보에 적합한 형태로 변환하고, 사용자 선호도 데이터를 생성하기 위하여 제안한 분류 정보 알고리즘을 광고 서버에서 적용하였다. 그리고, 사용자별 성향 분석과 비중결정, 광고 전송기능을 웹 브라우저를 통하여 볼 수 있도록 광고 분석 시스템을 설계하였다. <그림 16>은 개인화 광고 시스템의 흐름도이다.



<그림 16> 전체 시스템 흐름도

(1) 클라이언트 처리과정 설계

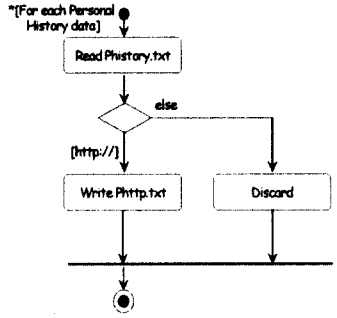
사용자 히스토리 데이터를 수집하여 전처리 과정을 거쳐 분류하여 서버에 전송하는 클라이언트 처리과정을 UML의 Activity 다이어그램을 이용하여 모델링한 결과는 <그림 17>과 같다.



<그림 17> 사용자 정보의 수집, 정제 및 분류 과정(Client)

(2) 1단계 전처리 과정 설계

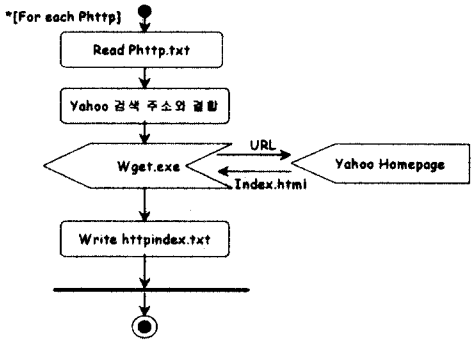
개인 히스토리 데이터를 수집과정을 거친후, 사용자 히스토리 데이터가 저장된 Phistory.txt 화일을 읽어들인다. 여기에서 http://로 시작하는 데이터만을 정제하여 Phttp.txt 화일로 저장하는 과정을 UML의 Activity 다이어그램을 이용하여 모델링한 결과는 <그림 18>과 같다.



<그림 18> 전처리 과정

(3) 야후 카테고리 수집

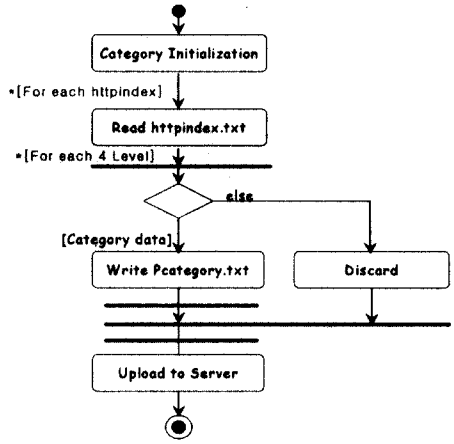
야후 사이트의 많은 양의 카테고리 깊이를 측정해 오기 위해 Wget을 사용하였지만 카테고리의 분류가 많고 많은 양의 하위 카테고리까지 추출하기에 걸리는 시간 등을 고려해서 본 논문 에서는 1차 대분류인 14개 카테고리정보만 추출하였다. Wget 프로그램을 실행시킬때 아규먼트로써 야후와 URL을 입력하고, 리턴시에는 야후의 카테고리가 포함된 홈페이지의 소스를 html 형식으로 가져와 httpindex.txt 화일로 저장하는 과정을 UML의 Activity 다이어그램을 이용하여 모델링한 결과는 <그림 19>와 같다.



<그림 19> 야후 카테고리 수집

(3) 2단계 전처리 과정 설계

데이터베이스에 저장된 카테고리 DB에는 1단계 14개의 항목 아래로 4단계까지 설계되어 있고, 영어로 된 분류정보만을 가져오도록 설계하였다. 검색된 주소로 httpindex.txt로 저장된 리턴 정보 화일에서 분류 정보만 선별하여 카테고리 데이터인 Pcategory.txt로 저장한 후 서버로 업로드 하는 과정은 <그림 20>과 같다.



<그림 20> 2단계 전처리 과정

5. 개인화 광고 서비스 프로토타입 구현

본 프로토타입 시스템은 클라이언트의 사용자 성향 정보와 히스토리 데이터를 전처리 과정을 거쳐서 서버에서 전처리 파일을 생성하도록 VC++ Enterprise 6.0을 이용하여 구현하였고, 사용자별 성향 분석과 비중결정, 광고 전송기능을 웹 브라우저를 통하여 볼 수 있도록 ASP와 MS SQL Server 2000을 이용하여 광고 분석 시스템을 구현하였다.

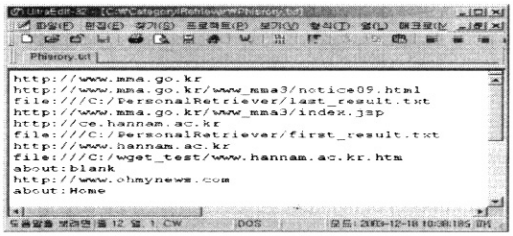
5.1. 클라이언트에서의 정보 추출기

본 시스템에서 클라이언트 사용자 측에서 이용하는 URL 로그 분류 정보 추출기의 동작 모습을 구현한 결과이다.

<그림 21> 분류 정보 추출기 실행 화면

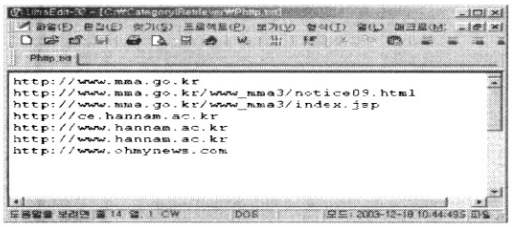
5.2. 전처리 화일 생성 구현

분류 정보 추출기에서 먼저 전처리 과정을 실행시키면 History 화일을 입력으로 들어가고, 출력데이터로 Phistory.txt 화일이 <그림 22>와 같이 생성된다.



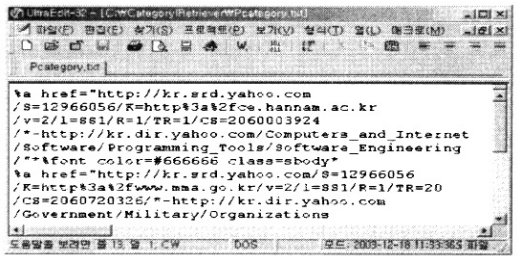
<그림 22> 전처리된 Phistory.txt 화일 생성

Phistory.txt 화일에서 'http://'로 시작되는 웹서핑 정보들만 선택하여 정제되어 전처리된 Phttp.txt 화일을 <그림 23>과 같이 생성한다.



<그림 23> 정제된 Phttp.txt 화일 생성

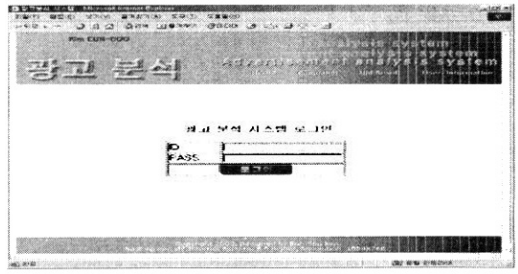
생성된 Phttp.txt 화일의 레코드들을 입력으로 받아 Wget.exe를 실행시켜 야후 검색 페이지를 가져와 Httpindex.txt 화일로 저장한다. Httpindex.txt 화일의 소스 내용중 카테고리 정보만 추출하여 ASP 소스를 결합하여 저장된 Pcategory.txt 화일은 <그림 24>와 같다.



<그림 24> 결합된 Pcategory.txt 화일 생성

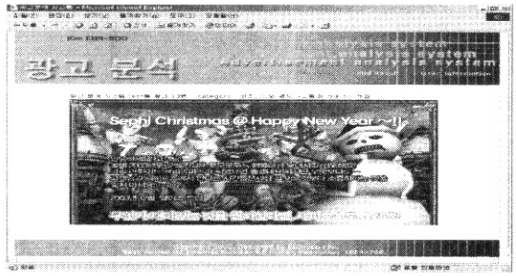
5.3. 광고 분석 시스템 구현

전처리 과정에서 최종적으로 생성된 Pcategory.txt 화일을 서버로 업로드 하면 분류를 카운트한뒤 비중 결정 알고리즘에 의해 비중을 결정하여 Weight_DB 테이블에 저장하고, 저장된 데이터는 Category_DB 테이블의 카테고리 정보와 매핑 한뒤 랜덤 함수를 발생시켜 카테고리를 선택한뒤 미리 입력된 AD_DB의 광고 데이터를 가져와 다시 한번 난수를 발생시켜 광고를 결정하여 광고를 전송한 후 각 테이블은 다시 업데이트되고 다음의 비중결정을 위한 자료로 활용된다.



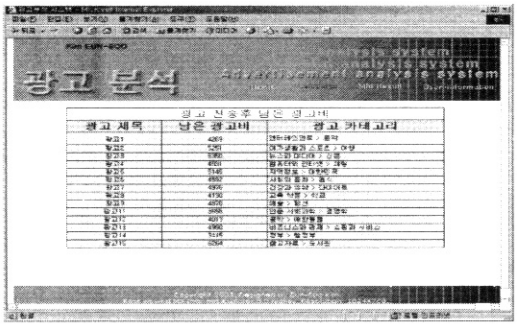
<그림 25> 웹에서의 분석을 위한 광고 분석 시스템

Campaign 메뉴를 클릭하면 <그림 26>과 같이 알고리즘에 의해 결정된 광고를 출력하게 된다.



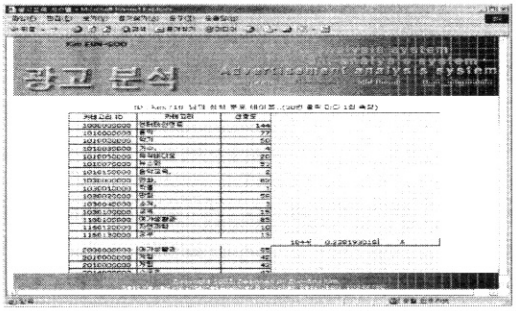
<그림 26> 사용자별 광고전송 화면

Mid Result 메뉴를 클릭하면 광고 전송후 남은 광고비에 대한 테이블을 볼 수 있다. <그림 27>은 Mid Result 메뉴를 클릭한 결과 화면이다.



<그림 27> 광고 전송 후 남은 광고비 테이블

User Information 메뉴를 클릭하면 사용자 성향의 분포를 볼 수 있다. <그림 28>은 20번 클릭 후 1회씩 측정한 결과를 보여준 화면이다.



<그림 28> 사용자별 성향 분포 테이블

6. 실험 및 결과분석

6.1. 실험환경

본 시스템의 연구 시나리오를 검증하기 위하여 한남대학교 컴퓨터공학과 실습실의 컴퓨터 20대를 선정하여 고객이라 가정하고, 모아진 데이터를 기반으로 하여 최소 1000회 광고를 나가게 하였다. 사용자의 성향을 분석한 데이터를 수집하는 서버는 Windows 2000 Server를 사용하였으며, DBMS는 MS-SQL 2000을 사용하였다.

사용자 환경은 Pentium Ⅲ급의 PC로 모두 30대에서 고객 프로그램을 이용하여 고객 컴퓨터에 저장되어 있는 성향을 분석하였고, 그 결과를 서버로 업로드 하여 사용자에게 광고를 선택하여 보내는 상황을 가정하여 근거리 통신망에서 실험을 하였다.

6.2. 정보 수집 및 분류과정의 결과 분석

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 히스토리 data를 사용하게 된다.

네트워크 상의 각각의 컴퓨터는 모두 각기 다른 고객이라고 생각하고 성향 결정과 광고의 실험을 수행하였으며 3장에서 제시한 사용자별 성향 분석 기법을 적용하였다.

<표 7> URL 분류 비율

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>의 사용자의 컴퓨터의 URL 분석을 통한 분류에 대한 비율이고, 전체 정보 중 특징적인 정보 몇 개를 추출하여 나타낸 것으로 각각의 분류에 대한 호감도이다. 평균적으로 D 항목이 다른 것에 비하여 높은 비율을 가지고 있는 이유는 D 항목이 컴퓨터와 인터넷에 관한 항목고, 수집한 데이터가 컴퓨터공학과 학생들이 주로 이용하는 컴퓨터 실습실에서 수집하였기 때문이다.

또한, 데이터가 같은 그룹의 컴퓨터에서 수집되었기 때문에 모두 동일한 수치를 가지고 있지는 않지만 비슷한 성향을 보이고 있다는 것을 분석하였다.

6.3. 수집된 정보에 의한 광고 선택 비율 과정의 결과 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.

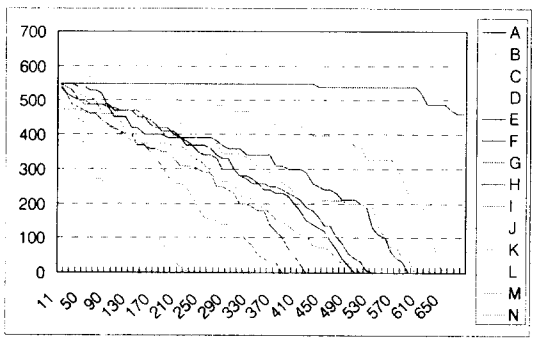
<표 8> 광고가 선택된 분류별 비율

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	14.85	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

6.4. 총 광고비의 시간에 따른 인기 항목의 결과 분석

본 논문에서 제안한 광고 시스템에서는 광고가 시간대별(Hit Count)에 따라 각 분류별 총 광고 금액의 변화를 광고주의 광고 금액의 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 하였다. 특이하게 F 항목은 실험결과에서 나타나듯이 처음부터 광고가 선택이 안 되다가 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 29>와 같이 실험결과를 통하여 확인하였다.



<그림 29> 시간의 흐름에 따른 광고비 변화 실험결과

7. 결론 및 향후 연구방향

본 논문에서는 개인화된 광고 서비스를 가능하게 하는 비교적 간단한 적응형 시스템을 설계하였고, 그 성능을 실험하였다. 개인의 성향을 시스템에 가장 효율적으로 반영하기 위하여 개인 컴퓨터의 히스토리 파일을 원시 데이터로 하여 정제 후 사용하여 해당 서버를 방문한 자에 한해서만 성향을 파악할 수 있는 단점을 극복하여 고객이 다른 서버의 방문 기록도 활용하므로 좀 더 현실성 있는 성향 파악이 가능하게 하였다.

향후 연구과제로는 e-Learning, eCRM등 전체 시스템과 연계한 시스템 확장이 필요하고, 사용자 프로화일, 연관그룹 정보, 협동 시스템, 연관 상품 등을 고려하여 광고 이익모델을 다양화한 통합된 시스템 개발의 연구가 필요하다. 또한 카테고리 저장, 분류 및 검색을 위하여 XML을 이용한 좀 더 발전된 형태의 연구가 필요하고, 온톨로지를 이용한 광고 항목의 확장 검색 및 선택과 동적인 웹 정보 집합에서 유전자 알고리즘(Genetic algorism)을 이용한 최적화된 정보 시스템 구축에 관한 연구가 필요하다.l

참 고 문 헌

[1] 한국 전산원, "2003 한국 인터넷 백서", 2004.
[2] 한국 인터넷 정보센터, "http://www.nic.or.kr"
[3] 코리아 인터넷 마케팅센터, "http://www.webpro.co.kr/"
[4] 오픈타이드, "Analytical eCRM 소개 웹 마이닝을 중심으로", 2001.
[5] 웹 개인화, "http://www.personalization.co.kr"
[6] Chakrabarti, Soumen, "Mining the Web", Morgan Kaufmann Pub, 2002.
[7] Loton, Tony, "Web Content Mining With Java", John Wiley & Sons Inc, 2002.
[8] Thuraisingham, Bhavani, "Web Data Mining and Business Intelligence Analysis", CRC, 2003.
[9] Ries, Al/Trout, Jack, "Marketing Warfare", McGraw-Hill, 1997.
[10] 김형택, 민옥길, "효과적인 인터넷 마케팅을 위한 웹 로그 분석", 비비컴, 2001.
[11] R. Kosala, H.Blockeel, "Web Mining Research, A Survey", ACM SIGKDD, July, 2000.
[12] 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.
[13] Wang Jicheng, et al., "Web Mining: Knowl edge Discovery on the Web", IEEE, 1999, pp. 317.
[14] 유시호, 김경중, "웹 사용 마이닝을 위한 SA SOM+DT를 이용한 웹 데이터의 분류", 한국 정보과학회 추계 학술발표대회 논문집, 제 29권 제 2호, pp.346~348, 2002.
[15] 박성준, 김주연, 김영국, "분산 이기종 인터넷 쇼핑몰 환경에서의 벡터 모델 기반 개인화 서비스 시스템", 한국정보과학회 논문집, 제 8권 제 2호, pp.206~218, 2002.
[16] 정현섭, 양재영, 최중민, "개인화된 웹 네비게이션을 위한 온톨로지 기반 추천 에이전트", 한국정보과학회 논문집, 제 30권 제 1호, pp.40~50, 2003.
[17] 고윤희, 김현철, "TID List를 이용한 빈발항목의 효율적인 탐색 알고리즘", 한국 정보과학회 춘계 학술발표대회 논문집, 제 29권 제 1호, pp.136~138, 2002.
[18] 고수정, 최준혁, "연관 단어 마이닝을 사용한 웹문서의 특징 추출", 한국 정보과학회 추계 학술발표대회 논문집, 제 30권 제 4호, pp.351~361, 2003.

김 석 훈

2001 배재대학교 정보통신공학과 (공학사)
2003 한남대학교 대학원 컴퓨터공학과(공학석사)
2003~현재 한남대학교 대학원 컴퓨터공학과 박사과정
관심분야 : 모바일 컨텐츠, 모바일컴퓨팅, 웹마이닝, XML, e-Learning
E-Mail: shk@hannam.ac.kr

김 은 수

1994 서울산업대학교 시각디자인과 (이학사)
1997 서울산업대학교 대학원 시각디자인과(이학석사)
2004 한남대학교 대학원 컴퓨터공학과(공학박사)
2001~2003 한국과학기술정보연구원 위촉연구원
2004~현재 한남대학교 교수학습 지원센터 강의전담 교수
관심분야 : 모바일 디자인, 웹디자인, 컴퓨터교육
E-Mail: kimes@hannam.ac.kr