세션과 객체 정보를 이용한 개인화된 로그 추출기법

김민숙, 박명순
고려대학교 컴퓨터 과학 기술 대학원 컴퓨터 공학과

kms246@korea.ac.kr
myongsp@korea.ac.kr

A Personalized Extracting Method using Session and Object Information

Min-Sook Kim, Myong-Soon Park
Dept. of Computer Science and Technology, Korea University

요 약

웹 로그에는 개별 사용자를 식별할 수 있는 사용자 정보와 세션 정보가 포함되어 사용자 식별과 해당 URL은 알 수 있지만, 그 URL에 해당하는 페이지내에 어느 객체에 관심이 있어 클릭하는지 알 수 없고, 페이지내에서 외부 사이트로의 링크 부분을 클릭했을 시 로그 파일에 기록이 되지 않는다. 본 연구에서는 세션과 사용자 중심의 로그 기록 방식에 객체를 추가함으로써 복잡하고 다양해지는 객체 요소(동영상, 오디오, 플래시 등)가 포함된 웹사이트에서는 객체 중심의 로그 기록 방식이 고객의 행동 패턴을 분석하여 세분화된 개인화 서비스에 보다 효율적임을 관찰하였다.

1. 서론

사용자로부터 요청된 파일이 웹 서버를 통해 처리되는 과정을 기록하고 있는 파일을 웹 로그(Web Log) 라고 부른다. 이러한 웹 로그 데이터는 사용자의 선호도 및 행동양식을 파악하여 마케팅 전략을 수립하는데 기반이 되는 중요한 데이터이다[1].

웹 로그를 표현하는 형식은 웹 서버의 종류에 따라 IIS의 CLF(Common Log Format), W3C Extention 등과 같이 여러 가지 형식이 사용되고 있다. 이러한 웹 로그 데이터는 파일 이름, 방문 페이지 등과 같이 물리적인 형태로 저장될 뿐만 아니라, 웹 서버에서 이루어지는 모든 작업이 기록되는 것으로 그 양이 방대하다. 이 로그의 데이터 정제 과정에서 주로 하는 역할은 분석에 불필요한 웹 페이지 내의 그림 정보 등을 제거하고, 사용자 IP로 사용자를 식별하고 요청 URL을 통해 단순히 평균 접근수, 평균 요청수, 그리고 평균 사용자 등과 같은 요약 정보를 추출할 수 있지만, 개인화된 정보를 추출하기는 어렵다. 그 이유는 웹 로그 파일에 저장된 사용자 IP를 이용하여 정확한 개인을 구별하기 어렵기 때문이다. ISP에서 제공되는 IP 정보는 유동 IP인 관계로 하나의 IP가 동시에 여러 명에 의해 사용되거나 여러 개의 IP가 한명의 사용자에 의해 사용되는 경우도 있다.

현재 상용화된 마이크로소프트의 IIS 서버[7] 또는 Apach 서버와 같은 웹 서버나 위세아텍, 웹 트렌즈와[8] 같은 웹 로그 분석 툴들은 웹 로그 데이터를 생성하는 단계에서 웹 로그의 분석에 무의미한 파일들을 파일 확장자(예:gif, jpg)에 따라 일률적으로 제거하도록 되어 있고 사용자의 선호도 및 행동양식을 파악하기 위해 요청 URL이 유일한 요소로 사용하고 있다.

최근에 전자상거래나 콘텐츠 사이트가 늘어나면서 웹 로그 정보를 수집하여 사용자 패턴분석이나 사이트 구조 분석을 위한 작업은 마케팅 전략을 수립하는데 상당히 중요하다[3]. 따라서 URL에서 좀더 세부적인 객체 단위로 분석이 들어가야 한다. 하지만 지금까지의 웹 로그 분석 툴들은 사용자가 해당 페이지 내에 관심있는 부분의 클릭 이벤트에 의한 이미지 객체나, 플래시, 동영상 객체들이 반영되지 못하기 때문에 사용자의 구체적인 관심 사항에 대한 패턴을 알 수가 없다. 사용자가 웹 사이트에 접속해서 자주 방문하는 페이지의 패턴을 분석하고 구체적인 관심부분을 알아 인터넷 마케팅 전략을 수립하기 위해서는 사용자가 웹 페이지를 방문해서 떠날 때까지의 모든 패턴을 알고 있어야 한다.

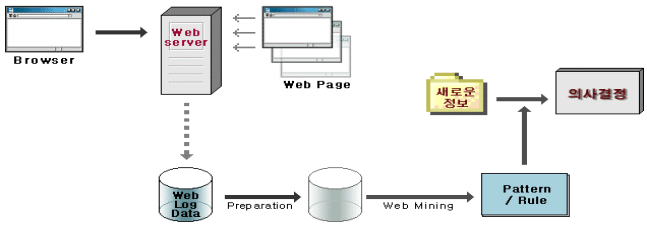
본 연구에서는 사용자가 웹 페이지를 방문해서 떠날 때 까지의 사용자 정보 및 세션 정보 그리고 클릭 이벤트에 의한 객체 중심의 데이터를 DB에 실시간으로 저장하여 동적으로 분석할 수 있도록 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 일반적인 웹 로그 분석시스템의 방법 및 문제점에 대해 설명하고 3절에서는 제안한 부분의 실험환경 및 결과에 대해 설명하고 4절에서는 결론을 내리고 향후 연구 과제를 제시하였다.

2. 기존 시스템의 웹 로그 수집 및 분석 방법

웹 로그 수집 방법은 크게 두가지로 구분된다.

첫째, 웹서버상에 기록되는 웹 로그 파일로부터 로그 정보를 수집하는 방법이다.[3] 이 방법은 기존의 웹 로그 분석 기법에서 주로 사용되던 방법으로 사용자가 웹 사이트를 방문하면 웹 서버는 요청된 웹페이지를 전달하면서 그와 관련된 모든 로그 데이터(include 파일,링크된 이미지파일 등)를 저장하거나 필터를 통해 단지 확장된 웹 로그 데이터만을 저장한다. 웹 로그 데이터는 파일에 저장되었다가 데이터 정제 과정을 통해 정제된 후에 OLAP이나 데이터 마이닝을 통해 분석된다. 이 웹 로그 수집 방법은 추가적인 작업없이 쉽게 로그 정보를 얻을 수 있는 장점이 있다. 하지만 기존의 로그 파일에서 정보를 추출하기 위해 로그 파일을 직접 접근하여 불필요한 내용을 제거하는 전처리 과정이 필요한 문제점이 있다[2]. 다음 (그림1)은 기존의 웹 로그 수집 및 분석 구성도이다.[4]



(그림1) 기존의 웹 로그 수집 및 분석 구성도

둘째, 에이전트를 이용하여 웹 로그 정보를 수집하는 방법으로 중소규모의 웹 서버를 운영하는 사이트에서 주로 사용하는 방법이다. 이 방법은 로그 정보를 수집하는 에이전트를 웹 서버상에 두고 웹 로그는 원격지 서버에서 수집한다. 이 방법은 웹 서버상에 로그 파일을 남기지 않으므로 디스크 저장 공간이 절약되는 장점이 있다. 하지만 원격지의 서버로 전송되는 과정에서 보안상의 문제가 있다.[5]

기존 웹 로그 수집 및 분석을 하기 위한 방법으로는 웹 로그를 변환하기 위한 전처리 과정이 필요하고 실시간으로 웹 로그를 분석하기 어려운 문제점이 있으며[2], 고객이 자신의 ID를 밝히는 경우 그 내용을 로그파일의 임의의 부분에 기재해야 하지만 기존의 웹 로그에는 고객의 ID를 밝히는 부분이 기재되지 않도록 되어있다[6]. 따라서 사용자를 식별하기 위한 방법으로 IP정보만을 가지고 파악해야 하는데, 이 정보만으로는 개인별 행동 패턴을 분석하기 어렵다.

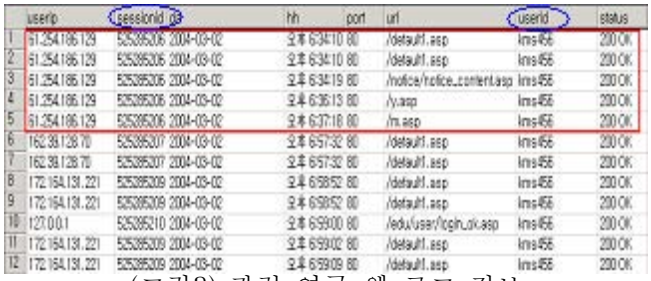
이러한 부분을 해결하기 위해 기존의 연구에서는 웹 로그 정보에 개별 사용자를 식별할 수 있는 사용자 정보와 세션 정보가 포함되어 개인화된 로그 정보를 저장하도록 하였다[2]. 그러나 이 경우 사용자 식별과 해당 URL은 알수 있지만 다음과 같은 문제가 있다.

첫째로, 개인화에 필요한 페이지 내에 어느 객체에 관심이 있어 클릭 하였는지는 알 수 없다. 예를 들어 공지사항에 5개의 목록이 있을 경우 그중 어떤 것을 클릭하였는지 로그 기록이 없으며 같은 위치에 image가 일정 시간으로 변하면서 링크가 이루어질 때 어떤 image(객체)를 클릭 하였는지 로그가 기록되지 않는다. 이는 개인화에 중요한 요소이므로 문제가 있다.

둘째로, 페이지내에서 외부 사이트로의 링크부분이다. 예를 들어 특정 image나 객체를 클릭하면 외부 사이트로 접속 후 browser를 닫으면 다시 메인 페이지로 되돌아 오는데 이 경우 해당 URL은 기존 서버에서 제공되는 로그 방식과 관련 연구 모두에서 전혀 로그파일에 기록되지 않는 중대한 문제가 있

다. 이는 객체를 추가하여 로그를 기록함으로서 해결 할 수 있다.

셋째로, 페이지 내에서 미디어 파일을 재생하였을 경우 기록된 로그가 include된 파일인지 사용자의 요청에 의해 재생된 파일인지 알 수 없다[6]. 이 또한 객체중심의 로그 기록 방식으로 해결할 수 있다.



	userip	sessionid	hh	port	url	userid	status
1	61.254.186.129	525285206 2004-03-02	오후 6:34:10	80	/default1.asp	kms456	200 OK
2	61.254.186.129	525285206 2004-03-02	오후 6:34:10	80	/default1.asp	kms456	200 OK
3	61.254.186.129	525285206 2004-03-02	오후 6:34:19	80	/notice/notice_content.asp	kms456	200 OK
4	61.254.186.129	525285206 2004-03-02	오후 6:36:13	80	/y.asp	kms456	200 OK
5	61.254.186.129	525285206 2004-03-02	오후 6:37:18	80	/m.asp	kms456	200 OK
6	162.39.128.70	525285207 2004-03-02	오후 6:57:32	80	/default1.asp	kms456	200 OK
7	162.39.128.70	525285207 2004-03-02	오후 6:57:32	80	/default1.asp	kms456	200 OK
8	172.164.131.221	525285209 2004-03-02	오후 6:58:52	80	/default1.asp	kms456	200 OK
9	172.164.131.221	525285209 2004-03-02	오후 6:58:52	80	/default1.asp	kms456	200 OK
10	127.0.0.1	525285210 2004-03-02	오후 6:59:00	80	/edu/user/login_ok.asp	kms456	200 OK
11	172.164.131.221	525285209 2004-03-02	오후 6:59:02	80	/default1.asp	kms456	200 OK
12	172.164.131.221	525285209 2004-03-02	오후 6:59:09	80	/default1.asp	kms456	200 OK

(그림3) 관련 연구 웹 로그 정보

따라서 본 연구에서는 기존 연구의 이러한 문제점을 해결하고자 관련 연구의 session과 URL 중심[2]에 클릭 이벤트에 의한 객체를 추가하여 로그 파일에 기록하고자 한다. 이를 실험을 통하여 관련 연구에서 누락된 해당 URL 로그가 전처리 과정을 거치지 않고 로그 파일에 기록되는 것을 검증하고 추가로 기록된 객체 로그가 개인화 서비스에 성능이 향상되는지를 분석하고자 한다.

3. 실험 환경 및 결과

본 연구의 실험은 사용자가 서버에 접속하면 생성되는 고유의 session id를 기준으로 사이트를 떠날 때까지를 하나의 session 단위로 묶어 사용자가 클릭한 패턴을 기존 연구처럼 여러개의 레코드가 로그 파일에 기록되는 것이 아니라 한개의 레코드로 로그 파일에 기록한다.

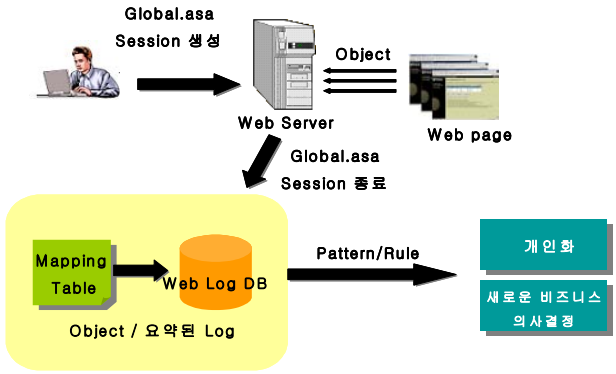
또한 기존 방식과 비교하여 session 단위의 흐름[2]이 효율적인지를 분석하고 여기에 클릭 이벤트에 의한 객체 로그를 추가함으로써 기존방식[2]에서 누락된 해당 URL 로그가 기록되는지를 구현하고 이것이 개인화 서비스에 성능이 향상되는지를 분석하고자 한다.

3.1 실험환경

본 실험의 시뮬레이션을 위하여 다음의 장비를 구성하였다.

- H/W : Pentium-4 2.0, 512Memory
  10/100MB Ethernet LAN
- Server : Windows 2000, IIS 5.0
- 대상site : 218.153.73.243
- DB : SQL Server 2000

실험은 client가 서버에 접속하면 기존 IIS Server에서 자동 제공하는 Text방식의 W3C Ext 로그 기록 파일과 관련연구의 기록방식인 session 단위와 사용자 중심의 로그 파일[2]에 본 연구에서 추가한 객체단위의 로그파일을 동시에 기록하는 방식을 취하였다. 비교를 위하여 기록될 때 발생하는 트래픽은 고려하지 않았다. 단지 기존 연구에서 기록되는 로그 항목에 누락된 URL이 본 연구에서 객체를 추가한 로그 항목에는 기록이 되는지를 테스트하였고 또한 앞에서 제기한 문제점이 해결되는지를 비교 분석하였다.



(그림4) 객체 중심의 로그 수집 및 분석 구성

구현 방법은 (그림4)와 같이 시스템 설계 후 해당 페이지내에 분석하고자 하는 객체를 클릭할 때 로그가 기록되는 분석 스크립트를 삽입하는 방식을 선택하였다.

	userip	sessionid	da	hh	port	url
1	219.133.25.228	525285199	2004-03-02...	오후 5:08:19	80	/Default.asp
2	172.164.131.221	525285204	2004-03-02...	오후 6:23:08	80	/Default.asp
3	61.254.186.129	525285206	2004-03-02...	오후 6:34:10;오후 6:34:19;오후 6:36:13;...	80	/default1.asp/notice/notice_content.asp/y.asp/m.asp
4	127.0.0.1	525285207	2004-03-02...	오후 6:57:31	80	/default1.asp
5	127.0.0.1	525285211	2004-03-02...	오후 7:03:20	80	/default1.asp

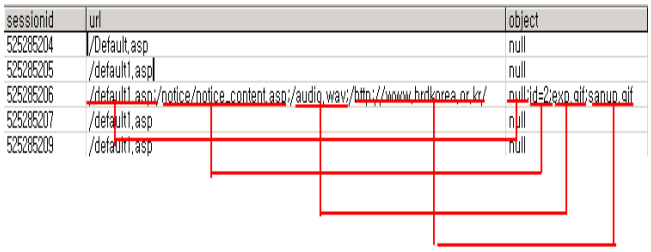
userid	object	bytes	status
kms456	null		200 OK
kms456	null		200 OK
kms456	null;id=2;exp.gif;sanup.gif		200 OK
kms456	null		200 OK
kms456	null		200 OK

(그림5) 제안한 웹 로그 정보

(그림5)는 본 연구에서 객체를 추가한 방식인데 기존 연구[2]에서는 (그림3)과 같이 한 사용자가 동일 session에서 5개의 웹페이지를 보면 5개의 로그 레코드가 생성되므로 행위패턴을 알려면 5개의 로그 레코드를 모두 읽어야 가능하지만 본 연구에서는 1개의 레코드로 가능하고 여기에 (그림6)과 같이 해당 url과 객체를 1:1로 매핑하여 기록함으로서 한 session에서 어떤 객체를 클릭하여 어떤 페이지(url)를 보았는지 알 수 있으므로 기존 연구[2]와 비교하여 보다 효율적이다.

사용자가 웹 사이트에 처음 접속했을 시 메인 화

면인 default1.asp를 호출하고 이 URL이 로그 테이블에 기록이 된다. 이때 처음 웹 사이트에 접속할 시에는 어떠한 클릭 이벤트가 일어나지 않았기 때문에 웹 로그 내의 객체에는 null 값이 디폴트로 기록이 된다. 이어서 notice/notice_content.asp를 호출했고 이것이 로그에 기록이 된다. 이때 객체값을 보면 id=2인 부분을 호출 했음을 알 수 있다. 이는 공지사항의 2번째 목록을 선택했음을 보여준다.



sessionid	url	object
525285204	/Default.asp	null
525285205	/default1.asp	null
525285206	/default1.asp;/notice/notice_content.asp;/audio.wav;/http://www.hrdkorea.or.kr/	null;id=2;exp.gif;sanup.gif
525285207	/default1.asp	null
525285209	/default1.asp	null

(그림6) URL과 객체의 1:1로 매핑 결과

객체 단위로 웹 로그에 기록이 되면 첫째로, 개인화에 필요한 페이지내에 어느 객체에 관심이 있어 클릭하는지 알 수 있다. 예를 들어 뉴스관련 공지사항이 있을 경우 사용자가 어떤 것을 클릭하였는지 로그 기록이 있으므로 어떤 사항에 관심이 있는지 보다 더 구체적으로 알 수 있다. 그리고 같은 위치에 image가 일정 시간으로 변하면서 링크가 이루어질때 어떤 image(객체)를 클릭하였는지 로그가 기록되므로 이는 개인화에 중요한 요소가 된다.

둘째로, (그림7)의 타원안에 있는 한국산업인력공단 이미지 객체를 클릭하여 외부 사이트(서버)에 접속 후 browser를 닫으면 다시 메인 페이지로 되돌아 오는데, 이 경우 웹 서버에서 제공되는 로그 파일과 기존연구[2] 모두 로그가 기록되지 않는다. 이는 해당 객체를 클릭하면 매핑 테이블을 이용하여 (그림6)처럼 객체와 url을 1:1로 매핑하여 로그 테이블에 누락된 url을 기록함으로서 로그 분석에 효율적이다.

셋째로, (그림7)의 타원안에 있는 무료체험 강의실 객체를 클릭하면 다른 페이지(url)로 이동을 하지 않고 해당 미디어파일(오디오,동영상)이 재생되는데 이 경우 오디오 및 동영상 파일이 기록 되긴 하지만 그 페이지에서 include 되어 기록된 파일인지 사용자의 요청으로 클릭 이벤트에 의해 기록된 파일인지 분별할 수가 없다. 본 연구에서는 (그림7)의 강의실 객체(exp.gif)를 클릭하면 미디어파일(audio.wav)이 (그림6)처럼 로그 테이블에 기록 되도록 하였다. 따라서 사용자가 어떤 미디어에 관심이 있는지 알 수 있기에 개인화 서비스에 효율적임을 알 수 있다.

(그림7) 해당 웹사이트의 객체 링크 부분

4. 결론

본 연구에서는 로그 기록방식에 객체를 추가함으로써 고객들의 행동 패턴을 분석하여 세분화된 개인 서비스를 위해 효율적인 로그 기록방식을 찾으려고 하였다.

여기에 기존의 session 단위에다 객체를 추가함으로써 날로 복잡하고 다양해지는 객체 요소(동영상, 오디오, 플래시 등)가 포함된 웹사이트내 에서는 객체 중심의 로그 기록 방식이 개인화 서비스에 보다 효율적임을 관찰하였다. 그러나 대상이 되는 객체마다 분석 스크립트를 삽입해야 하는 단점이 있다. 이에 앞으로 분석 대상이 되는 객체명을 삽입하면 자동으로 분석 스크립트가 링크되는 연구가 필요하다.

5. 참고문헌

[1] R. Cooley and J. Srivastava,, "Data Preparation for Ming World Wide Web Browsing Patterns," Int'l Journal of Knowledge and Information Systems, Vol. 1, No. 1, 1999

[2] 김재홍, 세션정보를 이용한 XML 기반의 개인화된 로그 추출기법, 창원대학교 석사학위 논문, 2002

[3] 전성훈, 최현희, eCRM실무지침, 삼각형 프레스, pp.142-158, 2001

[4] 인윤선, 정안모, 김명. 비즈니스 인텔리전스를 위한 지능적 웹 로거

[5] WiseLog for Web, Nethru Technologies, http://www.nethru.co.kr/wiselog/index.html, 2002

[6] 2N9Soft Logger. http://logger.co.kr, 2003

[7] http://www.microsoft.com

[8] http://www.webtrends.com