2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

웹 로그 분석을 위한 전처리기의 설계

김건량*, 이도헌*
*전남대학교 전산학과
e-mail:glkim@dbcore.chonnam.ac.kr

Design of a Preprocessor for Web Log Analysis

Geon-Lyang Kim*, Doheon Lee*
*Dept of Computer Science, Chonnam National University

요 약

최근들어 인터넷 쇼핑몰의 활성화로 인한 고객의 행동 패턴 분석의 필요성이 증가하고 있다. 본 논문에서는 고객의 행동 패턴 분석 방법 중의 하나로 데이터마이닝 기법을 이용한 웹 로그 분석을 소개한다. 웹 로그에는 고객의 접근 시간, 접근한 웹 페이지, 접근 시 사용한 브라우저 등 많은 정보가 포함되어 있는데, 마이닝 기법을 적용하기 위해서는 우리에게 필요한 정보만을 추출하고 적용하기 편리한 형태로 변환해야 한다. 본 논문에서는 마이닝 기법을 적용하기 위해 필요한 정보를 추출하고 적절한 형태로 변환하는 작업을 수행하는 전처리기의 설계를 제안한다. 본 논문에서 제안하는 전처리기로 구축된 트랜잭션을 통하여 원하는 항목과 범위에 대해서 연관 규칙을 얻을 수 있다.

1. 서론

웹 로그 분석은 웹 서버에 접근한 사용자의 정보가 기록되는 로그 파일을 분석함으로써 단순히 웹 사이트에 방문한 사용자의 수를 아는 것 이상의 정보, 즉 구체적인 방문객의 정보를 알고자 할 때나 기간별 분석, 사용자 분석, 페이지 분석 등 다양한 분석을 하고자 할 때 유익하다. 웹 서버에 대한 모든 접근들은 웹 서버에 의해 로그 파일로 기록이 되는데, 보통 사용자의 접근 시간, 사용자 아이디, 사용자의 IP주소, 요청한 웹 페이지, 접근 시 사용한 OS와 브라우저 등 많은 정보를 포함하고 있다. 본 논문은 NCSA 계열의 CLF(Common Logfile Format)형식의 웹 서버 로그 파일을 대상으로 한다.

웹 로그를 분석하기 위해서 연관 규칙 탐사, 순차 패턴, 클러스터링, 분류 등 다양한 데이터 마이닝 기법을 적용할 수 있다. 그런데 이러한 데이터 마이닝 작업 이전에 데이터를 적절한 형태로 가공하는 전처리 작업이 필요하다. 전처리기를 위한 요소 작업으로는 먼저 웹 로그 파일에서 그림 파일 등의 필요 없는 데이터는 삭제하고 마이닝을 적용할 대상이 될 데이터만 추출하는 데이터 추출, 각 사용자의 데이터들을 식별해내는 사용자 식별, 각 사용자들의 정보를 세션으로 식별해내는 세션 식별, 한 세션에서 클라이언트의 브라우저와 프록시 서버의 캐시로 인해 로그 파일의 중간중간에 기록되지 않은 웹 페이지들을 채워주는 패스 컴플리션, 마이닝 기법을 적용하기에 알맞은 트랜잭션 단위를 식별해 내는 트랜잭션 식별 등이 있다.[1]

전처리 과정 중 사용자를 식별하는 방법에서 많은 연구가 필요하다. 사용자를 식별하는 방법에는 쿠키, 사용자 아이디, IP주소를 이용하는 방법 등이 있다. 그러나, 쿠키는 지워지거나 쿠키 사용을 허용하지 않는 클라이언트들이 있다는 문제가 있고, IP 주소를 이용하는 방법은 여러 사용자가 한 PC를 공유하거나 한 사용자가 여러 PC를 사용할 수 있기 때문에 이용할 수 없다. 사용자 아이디를 이용하는 방법은 사용자가 다른 사용자의 아이디를 사용할 수 있는 한계점이 있다. 본 논문에서는 웹 사이트에 방문한 고객들을 식별할 때 고객 아이디를 사용한다.

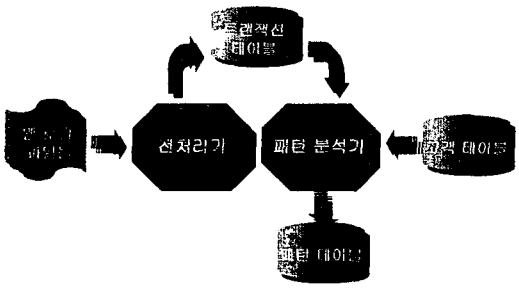
트랜잭션 식별 단계에서는 관리자가 원하는 연관 규칙이 있음에도 불구하고 일괄적으로 트랜잭션을 구성함으로 관리자의 요구사항을 수용할 수 없는 단

점이 있다. 본 논문에서는 관리자의 요구사항에 따라 각기 다른 트랜잭션을 구성함으로써 관리자가 원하는 연관 규칙을 생성할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 시스템의 전체적인 구조를, 3장에서는 전처리기의 구조를, 4장에서는 트랜잭션 식별 프로세서의 구체적인 처리 과정을 살펴보고, 마지막 5장에서 결론과 향후 연구 방향을 중심으로 마무리하고자 한다.

2. 웹 로그 분석 시스템의 구조

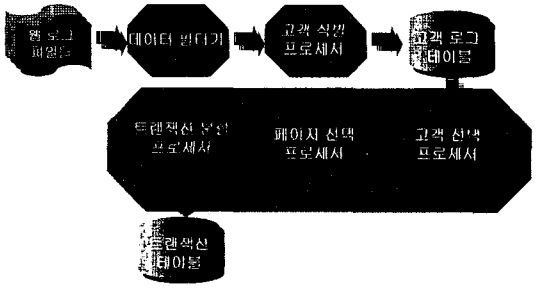
시스템은 웹 로그 파일을 입력값으로 하여 전처리를 한 후에 관리자의 요구가 반영된 각 파라미터에 따라 작업을 행하고 해당하는 데이터를 트랜잭션 테이블(Transaction table)에 저장한다. 패턴 분석기(Pattern Analyzer)는 트랜잭션 테이블의 데이터와 고객의 신상 정보가 저장된 고객 테이블(Customer Table)의 데이터와 함께 마이닝 기법을 적용하여 패턴을 분석한 후에 패턴 테이블(Pattern Table)에 저장한다. 패턴 테이블에 저장된 지식은 메일이나 우편, 사이트에 방문한 고객에게 웹 페이지를 추천해 주는 등 고객 지원을 위해 쓰이게 된다. 이러한 전체적인 작업은 관리자가 일정한 기간을 정하여 오프라인에서 주기적으로 일괄 처리한다.



[그림 1] 웹 로그 분석 시스템의 구조

3. 전처리기의 구조

전처리기는 웹 로그 파일을 입력값으로 한다. 웹 서버가 생성하는 로그 파일에는 접근 로그, 에러 로그, 참조 로그, 에이전트 로그 네 가지 종류가 있다. 먼저 접근 로그(access log)는 클라이언트의 접근 시간과 접근한 IP주소, 웹 전달의 성공 여부 등을 기록하고, 에러 로그(error log)는 웹 서버에서 발생하는 모든 에러와 접속 실패 정보를, 참조 로그(referrer log)는 사용중인 웹서버를 소개해준 사이트와 소개받은 페이지를, 에이전트 로그(agent log)는 사용자가 사용한 OS, 브라우저 등의 에이전트 정보를 기록한다. 구체적인 전처리기의 구조는 [그림 2]와 같다.



[그림 2] 전처리기의 구조

접근 로그 파일의 형식은 다음과 같이 IP주소, IdentityCheck 여부, 등록된 사용자 이름, 날짜와 시간, 요청한 메소드, 요청한 URL, 전송 프로토콜, 웹 전달 성공 여부, 전송 데이터량으로 구성된다.

210.25.22.130 - ACE [07/Jun/1999:17:02:33 +0900] "GET /A.html HTTP/1.1" 200 159

데이터 필터기는 접근 로그에서 등록된 고객 아이디, IP주소, 접근한 날짜와 시간, 요청한 페이지를 추출하는 역할을 수행한다.

고객 식별 프로세서는 고객이 접근한 웹 페이지들을 고객별로 분류한다. .htaccess와 .htpasswd를 이용한 사용자 인증방법을 사용했을 경우 접근 로그 파일에는 웹 페이지에 접근할 때마다 고객 아이디가 기록이 되기 때문에 [표 1]과 같이 쉽게 고객을 식별할 수 있다.

[표 1] 고객 로그 테이블

CID	IP	Date	Time	Page
:	:	:	:	:
ACE	210.25.22.130	07/Jun/1999	17:02:33	A.html
ACE	210.25.22.130	07/Jun/1999	17:05:01	B.html
ACE	210.25.22.130	07/Jun/1999	17:07:26	F.html
ACE	210.25.22.130	07/Jun/1999	17:11:12	I.html
ACE	210.25.22.130	07/Jun/1999	17:17:03	G.html
:	:	:	:	:
mon	210.25.22.130	07/Jun/1999	17:02:23	A.html
mon	210.25.22.130	07/Jun/1999	17:02:30	C.html
mon	210.25.22.130	07/Jun/1999	17:04:21	L.html
mon	210.25.22.130	07/Jun/1999	17:05:20	P.html
mon	210.25.22.130	07/Jun/1999	17:10:24	E.html
mon	210.25.22.130	07/Jun/1999	17:15:45	T.html
mon	210.25.22.130	07/Jun/1999	17:18:42	V.html
mon	210.25.22.130	07/Jun/1999	17:19:11	Y.html
:	:	:	:	:

트랜잭션 식별 프로세서는 고객 식별이 끝난 후에 트랜잭션 테이블(Transaction Table)을 구축하는 작업을 수행한다. 트랜잭션은 접근 페이지들간의 연관 규칙을 얻기 위해 필요한 접근 페이지들의 모음으로 트랜잭션의 범위에 따라 또는 트랜잭션을 구성하는 항목들에 따라 얻어지는 연관 규칙의 의미가 달라진다. 본 논문이 제안한 전처리기는 관리자가 원하는 트랜잭션을 구성하기 위해 인터페이스를 통해서 원하는 작업을 선택하여 파라미터를 넘겨주면, 엔진은 선택한 파라미터에 해당하는 작업만을 처리한다.

다음 장에서 본 논문이 제시하는 전처리기의 트랜잭션 식별 프로세서의 처리 과정에 대해서 자세히 살펴보도록 한다.

4. 트랜잭션 식별 프로세서

본 논문이 제안하는 전처리기의 트랜잭션 식별 프로세서는 관리자의 요구에 따라 각기 다른 트랜잭션 테이블을 구축할 수 있다. 관리자가 원하는 트랜잭션 테이블을 구축하기 위해 시스템이 제공하는 선택적인 작업은 다음과 같다.

첫 번째, 고객에게 페이지를 추천할 때 모든 고객 또는 특정 고객 집합의 행동 패턴만을 고려하여 페이지를 찾을 수도 있기 때문에 고객에 따른 분석을 생각해 볼 수 있다. 일괄 처리를 일주일 단위로 한다고 가정하면 고객 선택 프로세서는 모든 고객에 대해서는 일주일 동안 모든 고객 각각이 접근한 웹 페이지들이, 특정 고객 집합에 대해서는 일주일 동안 고객 집합에 속하는 각 고객들이 접근한 웹 페이지들이 트랜잭션의 항목들이 되도록 한다. 특정 고객 집합은 그 집합에 속하는 고객들을 관리자가 직접 선택하여 지정하거나, 특정 속성을 가지는 고객들을 선택할 수 있도록 한다. 예를 들어, 취미가 영화감상인 고객들의 고객 아이디가 A, B, C, D, E라면 [표 2]와 같이 이 고객 집합에 속하는 고객들의 접근한 웹 페이지들이 각 트랜잭션의 항목들이 된다.

[표 2] 트랜잭션 분할 테이블

CID	URL
A	D, F, G, I, O, P, R, U, X
B	C, D, E, I, K, O, P, R, U, X, Y
C	A, C, D, F, I, K, L, O, R, T, U
D	C, E, I, K, L, M, N, O, R, T, U, V, W
E	C, D, K, L, M, N, O, Q, T, U, W

두 번째, 페이지의 속성에 따른 트랜잭션을 구성할 수 있다. 고객은 보통 상품을 소개한 페이지를 원하는데, 상품을 소개한 페이지를 제외한 다른 페이지들은 상품을 소개한 페이지를 가기 위해 거치는 페이지이다. 그러므로 관리자의 분석 방법에 따라, 모든 페이지에 대해서 마이닝을 적용할 것인가? 아니면 상품을 소개한 페이지에 대해서만 마이닝을 적용할 것인가?를 구분하여 처리할 수 있다. 페이지 선택 프로세서는 디폴트로 모든 페이지에 대해서 처리하고, 상품을 소개한 페이지를 처리하고자 할 때는 페이지 프로파일 테이블(Page Profile Table)을 참고하여 상품을 소개한 페이지만을 추출한 후 트랜잭션을 구축하게 된다. 만약 [표 2]에서 상품을 소개한 페이지만을 추출하여 연관 규칙을 생성하고자 할 때 구축되는 트랜잭션은 다음 [표 3]과 같다.

[표 3] 트랜잭션 분할 테이블

CID	URL
A	D, F, G, I
B	C, D, E, I, K
C	A, C, D, F, I, K, L
D	C, E, I, K, L, M, N
E	C, D, K, L, M, N

세 번째, 관리자가 지정한 기간 단위로 트랜잭션을 구성할 수 있다. 보통은 전체 기간 동안 각 고객에 대해서 성립하는 사건간의 연관 규칙만을 고려한다. 그러나, 어떤 연관 규칙은 전체 기간 동안 각 고객에 대해서는 신뢰도가 높지 않지만, 특정 기간 동안 특별히 강한 신뢰도를 가질 수 있고, 그러한 정보는 마케팅 전략에 유용하게 사용할 수 있다. 예를 들어 컴퓨터 모니터를 구매한 고객의 70%가 삼일 안에 보안기를 구매한다는 규칙을 얻었다고 하자 이럴 경우 모니터를 구매한 고객에게 삼일 내에 보안기에 대한 가격과 제품 정보 등을 알린다면 좋은 마케팅 전략이 될 수 있을 것이다. 또한, 고객이 로그인하고 로그아웃하는 동안 연관된 규칙을 얻고 싶다면 세션 단위로 트랜잭션을 분할한다. 세션(Session)은 한 고객이 한 사이트에 로그인하고 로그아웃할 때까지 기간을 말한다. 그러나, 로그인 시점과 로그아웃 시점을 식별하기는 어렵기 때문에 일정 시간동안 고객이 웹 서버에 페이지를 요청하지 않을 때 로그아웃한 것으로 간주한다. 위와 같이 지정된 기간에 연관된 규칙을 얻고 싶다면 트랜잭션 분할 프로세서를 통하여 관리자가 지정한 시간 간격으로 트랜

잭션을 구성한다. 날짜별로 고객이 접근한 페이지가 [표 4]와 같고, 기간을 3일로 주었을 때 각각의 트랜잭션의 구성은 [표 5]와 같다.

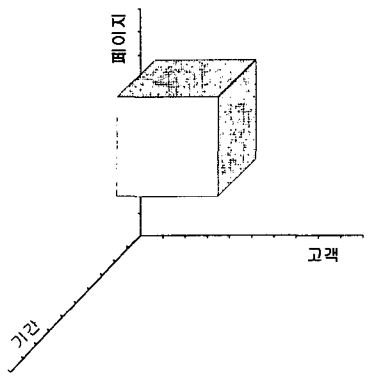
[표 4] 날짜별 접근 로그

Date	URL
17/jun/1999	B, C, F, G, I, X, R
18/jun/1999	A, I, O, P, C
19/jun/1999	C, D, U,
20/jun/1999	O, W, E, R, F, U, A, I, K, X, Y
21/jun/1999	C, O, A, K, L, I
22/jun/1999	V, W, T, A, M, N, O
23/jun/1999	D, U, I

[표 5] 트랜잭션 분할 테이블

TID	URL
1	A, B, C, D, F, G, I, O, P, R, U, X
2	A, C, D, E, F, I, K, O, P, R, U, X, Y
3	A, C, D, E, F, I, K, L, O, R, T, U, X, Y
4	A, C, E, F, I, K, L, M, N, O, R, T, U, V, W, X, Y
5	A, C, D, K, L, M, N, O, Q, T, U, V, W
6	A, M, N, O, T, V, W....
7	D, I, U....

결론적으로, 본 논문이 제안하는 전처리기의 트랜잭션 식별 프로세서는 고객에 따라, 페이지의 속성에 따라, 기간에 따라, 트랜잭션을 각기 달리 식별하여 [그림 3]과 같이 해당하는 범위의 데이터들로써 트랜잭션 테이블을 구축하게 되고, 원하는 연관 규칙을 얻게 된다.



[그림 3] 트랜잭션 식별 프로세서

5. 결론 및 향후 연구 방향

본 논문에서는 웹 로그 분석을 위한 전처리기를 설계하였다. 웹 로그 파일에서 등록된 고객 아이디, IP주소, 접근한 날짜와 시간, 요청한 페이지를 추출하고, 고객 아이디를 이용하여 고객을 식별하였다. 기존의 시스템은 데이터베이스에서 트랜잭션을 동일하게 구분했던 것에 반해, 본 논문에서 제안한 시스템은 원하는 파라미터들을 선택하여 관리자의 요구에 따라 각각의 트랜잭션을 구축함으로써 상황에 맞는 연관 규칙을 획득할 수 있다.

그런데, 로그인하기 전 고객이 접근한 웹 페이지 정보는 식별할 수 없다. IP주소가 동일하게 기록되는 경우는 식별할 수 있는 변수가 없기 때문이다. 그러므로, 고객을 식별하는 방법에 대한 연구가 더 진행되어야 한다.

참고문헌

[1] Cooley, R., Mobasher, B., and Srivastava, J., "Data preparation for mining World Wide Web browsing patterns," J. Knowledge and Information Systems, (1) 1, 1999.

[2] J. Pitkow. "In Search of Reliable Usage Data on the WWW," Proc. Sixth Int'l Conf. World Wide Web p. 451-463, 1997.

[3] Mobasher B, Cooley, R., and Srivastava J, "Web Mining: Information and Pattern Discovery on the World Wide Web," Proc. Ninth Int'l Conf. IEEE on Tools with Artificial Intelligence (ICTAI '97), November 1997.

[4] A. Luotonen. The common log file format. http://www. w3.org/pub/WWW/, 1995.

[5] M. S. Chen, J. S. Park, and P. S. Yu, " Efficient Data Mining for Path Traversal Patterns," IEEE Trans. on Knowledge and Data Engineering, Vol. 10, No. 2, pp. 209-221, March/April 1998.