한국지능정보시스템학회
02 춘계학술대회 논문집, pp. 134~139

인터넷쇼핑몰에서 웹로그 분석에 대한 개선방안 연구

김 남 호*

A Study on the Improvement of Web-log Analysis in Internet Shopping-Mall

Kim, Nam-Ho*

요 약

인터넷쇼핑몰 서버에의 고객의 상품에 대한 접근을 추적하여 고객의 성향을 추출하기 위한 웹마이닝에서는 웹서버가 생성하는 로그에서 필요한 정보를 수집하였다. 그러나 웹서버가 생성하는 로그는 단순 페이지 액세스의 정보만을 포함하고 있어, 현재 데이터베이스와 연동되어 동작하는 CGI 및 서버스크립트(JSP, ASP, PHP)등을 이용한 시스템에서는 CGI나 스크립트 파일명만 로그로 기록되고 분석시 가장 중요한 상품코드 및 상품 카테고리는 포함되지 않는다. 제안한 모델에서는 기존 쇼핑몰 시스템과의 연동 및 성능을 고려하여 웹서버에 분석전용 가상로그를 기존의 로그파일에 발생시키는 방법을 제안하였다. 이 방법으로 기존 사이트에 복잡한 코드를 추가할 필요 없이 간단한 로그발생코드 한 줄을 추가함으로써 해결할 수 있었다. 또한 유효 로그 필터링 및 클리닝에 걸리는 시간은 일반로그 분석대비 30%정도 향상되었으며 일반 로그에서는 불가능한 고객이 접근한 상품정보코드 및 카테고리코드 등의 정보를 쉽게 추출할 수 있었다.

Key words : 로그분석, 데이터마이닝, 웹마이닝, 인터넷쇼핑몰

1. 서론

인터넷비즈니스의 영향으로 기업의 경영활동에 변화가 일고 있다. 기업의 경쟁력은 정보의 공유로 인해 제품 품질의 차이로 인한 우수성보다는 고객관리에 관심이 모아지고 있다. 기업의 경영활동 중 고객 접점포인트라고 할 수 있는 마케팅, 판매, 대고객서비스 영역에서의 인터넷의 활용은 중요성이 날로 더해가고 있다. 특히 인터넷마케팅은 가상의 공간에서 소비자와의 관계형성 및 실시간 상호작용이 가능한 쌍방향 커뮤니케이션을 통한 일대일 개별화 마케팅이 실현될 수 있는 최적의 환경을 제공하고 있다. 개별화 웹 마케팅은 본질적으로 고객중심의 패러다임이다. 개별고객의 필요한 사항들을 파악해서 각각의 고객에게 차별화된 서비스를 제공하는 것이 그 핵심이다. 이를 통해서 기업은 개별고객과의 관계증진을 통해 고객 유지율을 상승시키고 해당 회사에 대한 충성심을 유도하여 이익을 극대화 하고자하는 것이 고객관계관리 (CRM, Customer Relationship Management)이다.

인터넷상에서의 개별화 마케팅을 위한 데이터는 주로 사용자가 제공하는 회원정보, 웹 서버에 기록되는 사용자 방문경로, 판매정보, 마케팅활동에 대한 반응정보를 주로 이용하여 추출하였다. 이때 사용하는 로그파일에는 사이트의 방문자수, 최다 접근 파일, 시간대별 동시 접속자수 등과 같은 통계데이터를 이용하여 안정된 웹서버를 운영할 수 있는 정보를 취할 수 있다. 뿐만 아니라 로그파일에 데이터마이닝의 연관규칙과 같은 기술을 이용하게 되면 고객행동의 패턴을 파악하여 미래의 행동예측을 통한 마케팅 활동에 활용할 수 있다.[1]

하지만 웹서버가 생성하는 로그는 단순 페이지 접근 정보만을 포함하고 있어, 현재 데이터베이스와 연동되어 동작하는 CGI 및 서버스크립트(JSP, ASP, PHP)등을 이용한 시스템에서는 CGI나 스크립트 파일명만 로그로 기록되고 분석 시 가장 중요한 상품코드 및 상품 카테고리 코드는 포함되지 않는다. 이러한 문제는 사용자의 상황을 파악하는데 있어서 빈약한 정보를 제공하게 되어 고객관리에 있어서 정확성의 한계를 가져오게 된다.

본 연구에서는 이에 대한 해결책을 제공하여 현재 인터넷쇼핑몰을 비롯한 로그분석을 필요로 하는 인터넷비즈니스 업체가 공통적으로 직면하고 있는 문제에 대한 해결책을 제시하고자 한다.

논문구성은 1장에서는 연구배경과 목표에 대하

* 호남대학교 정보기술원

여, 2장에서는 이와 관련된 연구를, 3장에서는 활용하고자하는 시스템을 설계하고, 4장에서는 새롭게 제안한 로그분석에 대한 개선안과 성능평가를 제시하고, 5장의 결론에서는 연구의 의의를 제시하였다.

2. 관련연구

2.1 데이터마이닝

2.1.1 개념 정의

데이터 마이닝은 대량의 데이터로부터 유용한 지식을 추출하여 이해하기 쉬운 형태로 변환한 후 의사결정과정에 적용하는 모든 과정으로 정의된다. 웹의 확산과 더불어 인터넷을 이용한 사업이 다양하게 전개되면서 인터넷에서의 접속자와 웹사이트를 관리할 수 있는 전략을 수립하는 것이 중요하다. 웹 사이트를 운영하는 웹 서버는 로그파일이라는 대용량의 데이터를 수록하고 있다. 로그파일에는 접속한 컴퓨터의 IP주소, 접속한 시각, 접근방법, 요청한 파일, 프로토콜, 에러코드, 전송된 파일의 크기 등과 같은 정보가 들어있다. 따라서 이러한 정보를 분석함으로써 인터넷 사업에 유용한 정보를 도출해 낼 수 있다. 이와 같이 웹 사이트로부터 로그파일과 같은 데이터를 분석하고 전략을 수립하는 일을 웹 마이닝(Web Mining)이라고 한다. 대부분의 로그파일은 표준 로그파일 형식(Common Logfile Format)[2]을 따르고 있다. 최근 들어 Windows NT서버가 사용되면서 확장 로그파일 형식(Extended Logfile Format)으로 생성되기도 하지만 큰 규모의 서버를 사용하는 경우는 보통 표준 로그파일 형식을 따라 기록하게 되어 있다.

2.1.2 기존 연구

데이터 마이닝의 개념은 Mobasher[3]등이 처음으로 제안하였으며, 웹 트래픽분석을 제공하는 Marketwave사[4]는 추가적으로 웹상의 데이터마이닝의 결과물을 응용할 수 있는 방법을 제안하였다. 웹상의 데이터마이닝의 결과로 나타나는 연관규칙, 군집분석, 패턴분석 등은 인터넷마케팅에 유용하게 사용할 수 있다. Buhner와 Mulvenna[5]은 OLAP을 이용해 웹상의 데이터마이닝을 수행하고, 그 결과를 인터넷마케팅에 활용하기 위해 고객유도, 고객유지, 교차판매 등과 같은 전략을 제시하였다. 또한 Borges, Levene[6]와 Spiliopoulou[7]등은 데이터마이닝 기법을 사용하여 웹사이트를 분석하고, 사이트의 구조와 링크관계를 분석함으로써 접속자들이 편리하게 탐색할 수 있는 웹 사이트의 디자인 방안을 제시하였다.

웹상의 데이터마이닝의 모델을 제시한 논문들은 다음과 같다. Cooley[8]등은 표준 로그파일 중 access, referrer, agent 로그파일을 통합한 후, 데이터의 사전처리, 마이닝 기법적용, 패턴분석의 세 단계를 거치는 WebMINER 시스템을 제안하였다. Zaiane[9]등은 접근로그파일에 대해서 데이터의 사전처리를 수행하고, 사용자 행동을 정의한 후 이것을 로그파일의 필드로 추가시켜 OLAP기법을 적용하고 연관성 규칙을 발견해 내는 모델인 Web Log Miner를 제안하였다. 한편 Wu[10]등은 로그파일의 형식에 제약을 두지 않는 모델인 Speed Tracer를 제안하였다.

웹상의 데이터마이닝에서 접속자를 구별하는 방법과 접속시간(Session)을 정의하는 방법은 정해진 정답이 없기 때문에, 보통 발견적 방법(Heuristic Method)을 사용한다.

이에 따라 Murray와 Durrell[11]은 접속자들에 대한 인구 통계학적 정보를 알 수 없는 경우에 대하여, 접속자를 구별하고 접속자 정보를 도출해 내는 방법을 제안하였다. Fu등[12]은 접속 지속시간을 정의하고 접속시간이 유사한 접속자들을 군집화하는 방안을 제안하였다. Cooley[8]등 역시 접속자 구별방법과 접속 지속시간을 정의하였다.

Cooley[8]등은 access 로그파일, referrer 로그파일, agent 로그파일을 웹 마이닝에 사용하였지만, Zaiane[9]등은 페이지간의 연관성 규칙분석을 위한 데이터로 access 로그파일만을 사용하였다. Access 로그파일에는 파일간 이동경로가 나타나 있지 않기 때문에, 일단 access 로그파일로부터 접속자를 구별하고 각 접속자별로 일련의 페이지간 이동 경로를 분석해 내야한다. 그런데 접속자를 구별하는 방법과 접속자의 방문시작과 끝을 나타내는 접속 지속시간에 대한 정의가 연구마다 다른 상황에서 access 로그파일을 통해 페이지간 이동경로를 결정하는 것은 객관성이 부족하다고 할 수 있다. 페이지 간 1차 이동경로가 나타나 있는 referrer 로그파일을 페이지간 연관성 규칙을 발견하는데 사용하게 되면 접속자 구별과 접속 지속시간에 대한 분석을 생략할 수 있기 때문에 연관성규칙 분석을 훨씬 적은 노력으로 수행할 수 있다.

2.2 로그분석

2.2.1 로그분석 의미

사용자의 웹사이트 이용에 대한 기록이 로그라는 형태로 흔적이 남는다. 로그분석이란 이 데이터를 기반으로 다양한 정보를 추출해 내는 것이라 할 수 있다. 이러한 로그분석은 사용자에 따라 단지 로그정보를 분석하는 것에 한정시키기도 하고 때론 로그정보를 기반으로 한 보다 다양한 정보를 분석하는 확장된 개념으로 확대시키기도 한다. 이를 살펴보면 다음과 같다.

일반적 의미의 로그분석은 로그 데이터를 이용하여 트래픽을 파악하고, 이 트래픽이 지닌 의미를 분석해 나가는 것이라고 할 수 있다. 로그 데이터를 이용하여, 웹사이트의 페이지뷰, 사용자별 페이지뷰, 접속장소 및 방식, 시간별 페이지뷰, 방문자수 등에 대한 현황 및 추세를 분석하는 것이다. 웹사이트의 클릭흐름을 분석하는 것 역시 이 범주에 들어간다. 사용자가 웹사이트를 방문하는 경로와 서핑하는 경로에 대한 분석을 통하여 웹사이트가 지닌 문제점을 찾고, 사용자가 웹사이트에서 무엇을 원하는 지를 보다 구체적으로 파악하는 것이다.

이에 반해, 확장된 의미의 로그 분석은 단지 로그 데이터뿐 아니라, 웹사이트에서 보유하고 있는 고객등록정보, 구매정보, 외부환경정보 등을 복합적

으로 사용하는 분석을 말한다. 이러한 분석을 통하여 사용자 특성별로 웹사이트의 이용, 구매에 대한 보다 폭넓은 분석이 가능해지며 이는 개인화 된 서비스의 기반이 된다. 실제 사례로는 다음과 같다.

로그분석을 통해 인터넷의 웹사이트에서는 접속한 고객에 따라 고객별로 상이한 웹 페이지를 보여줄 수도 있다. 물론 고객들이 원하는 내용을 분석한 결과를 분석한 데이터를 기반으로 구성된 페이지다. 쇼핑몰 사이트인 경우 고객에 따라 차별화된 가격을 제시할 수도 있다. 예를 들어 일반고객에게는 일정한 가격을 보여주고, 물건을 많이 구매한 경험이 있는 구매력이 우수한 고객에게는 낮은 가격을 보여줌으로써 고객별로 다른 가격 전략을 펼칠 수도 있다는 뜻이다. 배너 광고의 경우도 마찬가지다. 각 고객마다 반응을 보이는 광고 종류를 분석하여 그 고객이 방문했을 때 반응률이 높은 광고만 노출시킬 수 있다. 또한 고객의 구매를 유도하는 E-mail 마케팅을 전개할 때에도 웹사이트를 통해 수집한 고객 정보를 기반으로, 사용자의 필요성을 자극하는 보다 효과적인 마케팅 활동이 가능해진다.

2.2.2 로그분석의 문제점

로그 데이터는 부정확하며 분석할 수 있는 정보 또한 많지 않다. 예를 들어 로그 데이터에 저장되는 클라이언트 IP는 사용자가 인터넷을 시작한 위치 정보에 해당된다. 클라이언트 IP정보를 통해 웹사이트에 접속하는 사용자들의 정보를 파악할 수 있다. 하지만 사용자가 유동 IP를 사용할 경우, 또는 Proxy Server를 이용할 경우 등에서는 정확한 클라이언트 IP를 파악할 수 없다. 또 캐쉬를 통할 경우 사용자는 페이지를 봤지만 로그에는 그 데이터가 저장되지 않는다. 그리고 뒤로 가기(BACK)를 통해 이동할 경우도 그 행위들은 로그에 기록되지 않는다. 사용자들의 정보도 마찬가지다. 쿠키를 이용할 경우 한 명의 사용자가 한 PC를 계속 사용한다는 경우를 가정해야만 그 정확도에 대해서 어느 정도 만족할 수 있지만 실제 그런 경우는 많지 않다. 그리고 동일한 IP를 가지고 접속했다고 할 지라도 사용자가 계속해서 정보를 이용하는 지 아니면 재 접속해 들어온 건지에 대한 판단을 하기가 쉽지 않다.

따라서 로그 데이터가 웹 사이트에 필요한 모든 정보를 정확히 줄 수 있다고 생각하기보다는 우리가 로그 데이터로부터 필요한 정보를 뽑아 내야 한다. 따라서 로그 데이터가 정확하지 않을 수 있다라는 한계를 극복하고 필요한 정보를 얻기 위해서는 다양한 웹 사이트 특성에 맞는 다양한 알고리즘의 개발이 요구된다.

2.3 웹상의 데이터마이닝

2.3.1 과거의 로그파일 분석

과거의 로그파일 분석은 웹 트래픽 분석이라고 하며 웹 마스터나 시스템 운영자가 사이트에서 어떤 일들이 일어나는지 얼마만큼의 접속이 실패하고 어떤 종류의 에러가 일어나는지 등을 알기 위해 로그파일을 분석하는 것을 의미한다[4]. 웹 트래픽 분석을 통해 알 수 있는 것은 다음과 같은 내용들이다.

- 가장 많이 방문하는 페이지는 어디인가?
- 가장 적게 방문하는 페이지는 어디인가?
- 이 사이트로 들어오기 전에 방문했던 페이지는 어디인가?
- 가장 많이 접속한 기관은 어디인가?
- 가장 많은 접속을 한 사용자는 누구인가?
- 요일별로 접속횟수는 어떻게 되는가?
- 시간대별 접속횟수는 어떻게 되는가?

2.3.2 웹기반의 데이터마이닝

웹상의 데이터마이닝(Web based Datamining - Advanced Web Traffic Analysis)은 일반적으로 앞에서 설명한 웹 트래픽 분석에서 사용되는 로그파일에 다른 데이터를 추가시켜 분석하는 것을 의미한다[4]. 다른 데이터란 고객데이터, 회계데이터, 전자상거래 자료 등이 해당될 수 있다.

또 다른 의미로는 로그파일을 분석할 때 단순한 통계분석만 하는 것이 아니라 데이터마이닝 기법을 적용함으로써 좀더 유용한 정보를 얻고자 하는 분석을 가리킨다. 웹상의 데이터마이닝을 통해 얻을 수 있는 정보는 다음과 같다.

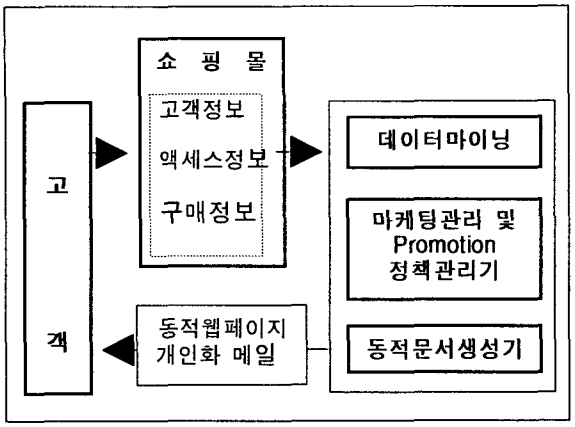
- 누가 우리의 사이트를 방문하는가?
- 어떤 사이트를 거쳐서 우리의 사이트를 방문하는가?
- 방문자의 인구통계학적 생활양식분류 정보와 사이트 방문 형태와는 어떤 관련이 있는가?
- 사이트에서 얻을 수 있는 수익은 무엇이며 얼마나 되는가?
- 어떤 광고배너가 가치 있는 고객들을 우리 사이트로 오게 만드는가?
- 어떤 페이지가 실제로 구매를 하는 사람의 수를 증가시키는데 기여하는가?

3. 로그분석 시스템

본 시스템은 쇼핑몰 사이트 고객 개인의 구매성향을 파악하여 차별화된 서비스를 제공하는 것을 기본 목표로 한다.

CRM을 위한 분석용 데이터베이스를 구축하는 과정에서 첫번째 작업은 인터넷 사용자의 접점 데이터인 고객구매정보 항목들 중 분석에 필요한 항목들을 통합하는 모델을 작성하는 것이다. 이를 원시데이타의 전처리라고 하며 쇼핑몰 서버에서 고객의 상품에 대한 접근기록인 고객이 접근한 상품의 웹 액세스 로그와 고객의 Session ID, 고객 ID를 기준으로 한다.

데이터마이닝 모듈인 구매패턴분석기에는 전처리기의 트랜잭션에 포함된 접근한 상품정보가 항목으로 입력된다. 이때 웹 상품정보에 연관규칙(Association Rule)을 적용하여 접근한 상품들 간의 패턴지식을 찾아내고 패턴 데이터베이스에 저장한다.



[그림1] 시스템 구조

마켓팅관리 및 프로모션 정책관리기에서는 웹상에서 관리자가 조회 조건을 생성할 수 있는 인터페이스를 제공하여 고객의 성별, 연령, 관심도, 직업, 마일리지 등의 관련 조합에 의해 조회식을 용이하게 생성할 수 있도록 지원한다. 관리자는 생성된 조회식을 통해 마켓팅 타겟을 결정할 수 있으며, 마켓팅 정책에 의해 프로모션하는 상품을 정의하고 작성된 발송 메일 양식을 등록한다. 또한 발송일등을 스케쥴링하여 자동발송 될 수 있도록 한다.

개인화 메일 발송 시스템은 마켓팅 타겟으로 결정된 고객 및 고객 집단에 대해 매핑된 특정 메일을 대량으로 발송할 수 있으며, 스케쥴링을 통해 자동으로 발송 처리한다.

4. 분석전용 로그생성

4.1 기존 로그분석의 한계

쇼핑몰 서버에의 고객의 상품에 대한 접근을 추적하여 고객의 성향을 추출하기 위해서는 사이트에 대한 액세스 로그와 로그를 고객의 접근 세션 단위로 필터링할 수 있는 고객의 Session ID와 고객 ID 및 접근 시간등에 대한 저장이 되어 있어야 한다. 여기서 고객의 세션은 고객이 사이트에 처음 접속한 시점부터 현재의 브라우저를 종료한 시점까지를 하나의 세션으로 본다. 그러나 만약, 로그 온했던 고객이 로그오프 한 후 계속해서 사이트에 접속중이라면 로그오프 이전과 이후는 별도의 세션으로 구분된다.

기존에는 웹서버가 생성하는 로그에서 필요한 정보를 수집하였다. 그러나 웹서버가 생성하는 로그는 단순 페이지 액세스의 정보만을 포함하고 있어, 현재 데이터베이스와 연동되어 동작하는 CGI 및 서버스크립트(JSP, ASP, PHP)등을 이용한 시스템에서는 CGI나 스크립트 파일명만 로그로 기록되고 분석 시 가장 중요한 상품코드 및 상품 카테고리코드는 포함되지 않는다.

이러한 문제는 사용자의 성향을 파악하는 데 있어서 빈약한 정보를 제공하게 되어 고객관계관리에 있어서 정확성의 한계를 가져오게 된다.

초기에 웹서버가 발생하는 단순 로그를 대상으로 고객의 인터넷 쇼핑몰에서의 패턴분석시스템을 개발하면서 일반적인 웹 로그만으로는 현재와 같은 동적인 사이트에서 고객의 성향을 분석할 수 없다는 한계에 부딪치게 된다.

4.2 새로운 로그분석방법 제안

4.2.1 필요성

웹서버가 발생하는 로그는 고객이 접근한 페이지에 대한 정보만 가지고 있다. 이 정보는 현재 데이터베이스와 연동되어 동적으로 서비스하고 있는 CGI기반의 웹 어플리케이션이나 ASP, JSP, PHP와 같은 서버측 스크립트와 같은 웹어플리케이션인 경우에도 해당 파일명만을 포함한다. 따라서 실제 넘어가는 상품의 코드와 카테고리 코드와 같은 정보와 같이 중요한 정보를 얻기 위해서는 클라이언트 쿠키를 사용하는 방법 등이 동원되고 있다.

로그정보 중 분석처리에 유효한 정보는 5%이내이다. 따라서 방대한 데이터 필터링과 클리닝 작업이 초기 전처리 단계에서 필요하게 되어 시스템의 요구사항을 증대시킨다.

데이터베이스에 대한 별도의 로그발생시스템을 개발할 경우 기존 사이트에 대한 수정이 발생하게 되고, 또한 전반적으로 3%정도의 시스템 부하가 증가하게 된다.

4.2.2 일반 웹 로그 형식

항목: date | IPaddress | page URL | browser type
로그 예:
2001-02-10 02:33:19 211.247.83.79
- 211.174.58.39 80 GET /shop/productdetail.jsp
- 200 Mozilla/4.0+(compatible;+MSIE+5.0;+Windo ws+98;+DigExt)

4.2.3 분석전용 로그생성

일반적인 웹 로그와 기초 패턴은 동일하나 접근 페이지에 서버측 스크립트 또는 CGI상에서 존재하지 않는 가상페이지를 요청하는 방식으로 분석을 위해 특정한 형태로 정의된 가상로그를 발생시킨다.

그러나 이 경우 다음과 같은 제약조건이 있다.

첫째, 스크립트 페이지나 CGI또는 HTML문서를 호출 시 별도페이지 호출없이 호출된 기 페이지 내에서 호출될 수 있는 형태이어야 한다.
둘째, 서버시스템에 부하를 주지 않아야 한다.
셋째, 클라이언트 시스템에 별도의 플러그인이나 애플릿, 컴포넌트등을 설치하지 않아야 한다.
넷째, 분석에 유효한 정보를 충분히 포함할 수 있어야 한다.

이와 같은 제약을 만족하는 방법으로 다음과 같은 방식으로 가상로그를 생성시키는 기법을 완성시켰다.

하나의 HTML 파일이나 동적인 CGI 또는 스크립트가 호출될 때 동시에 호출되어질 수 있는 파일 포맷들 중에서 가장 일반적이면서 쉽게 적용이 가능한 것이 이미지 파일이다.

따라서, 이미지 파일 포맷 중 Jpg포맷을 사용하여 서버 측에 존재하지 않는 가상의 페이지를 아래와 같은 형태의 포맷으로 구성된 파일명으로 정의하고 웹서버에 요청하도록 함으로써 아래와 같은 독특한 형태의 분석 전용 로그가 웹서버 측에 남는다.

접근페이지 내의 분석전용 로그 패턴

: 분석식별자 | 세션아이디 | 고객아이디 | 카테고리코드 | 상품코드 | .jpg

로그 예:

/^^|20020410023313435344 3|cus0001|10003|100002|.jpg

구분자	세션아이디	고객아이디	카테고리	상품코드	파일확장자

4.3 성능평가

고객의 정보를 정확하게 수집하기 위한 또 하나의 방법으로 데이터베이스상에 직접 분석전용 로그를 남기는 방법을 시도해 보았다.

그러나 이때에는 데이터베이스에 별도의 부하를 가져와 3%정도의 전체적인 성능저하를 가져왔으며, 기존의 사이트와 적용 시 상당 부분 기존 사이트를 수정하여야 하는 단점이 제기 되었다. 반면에 분석과정에서 전처리기에 의한 필터링과 클리닝 등의 공정이 제거되어 전체적으로 분석과정에 소요되는 시간은 기존 웹로그 분석에 비해 50%정도 성능향상을 가져왔다.

<표1>분석방식에 따른 성능 및 비용비교

성능 체크	일반 로그 분석	DB 로그 생성	분석전용 로그생성
사이트 성능	성능저하 없음	3%성능저하	1%이하 성능저하
전처리과정 성능	낮음	높음	중간
분석성능	낮음	높음	높음
기존사이트와 적용성	저비용	고비용	저비용
시스템 구축 비용	고비용	저비용	저비용
정확성	낮음	높음	높음
확장성	높음	낮음	높음

제안한 모델에서는 기존 시스템과 융화 및 성능을 고려하여 웹서버측에 분석전용 가상로그를 기존의 로그파일에 추가하였다.

이 방법으로 앞서 시도한 웹서버에 별도로 데이터베이스에 로그를 남기므로 데이터베이스에 대한 직접적인 로그 발생으로 생성된 시스템 부하를 줄일 수 있었으며, 아울러 기존 사이트에서는 복잡한 코드를 추가할 필요 없이 간단한 로그발생 코드 1줄의 추가로 개발 적용을 할 수 있게 되었다. 또한, 유효 로그 필터링 및 클리닝에 걸리는 시간은 일반 로그 분석대비 30%정도 향상되었으며 일반로그에서는 불가능한 고객이 접근한 상품정보코드 및 카테고리코드 등의 정보를 쉽게 추출할 수 있었다.

<표2>분석방식에 따른 성능 및 비용 우선순위

성능 체크	우 선 순 위	비 고
사이트 성능	A > C > B	높을수록 우수
전처리과정 성능	B > C > A	높을수록 우수
분석성능	B > C > A	높을수록 우수
기존 사이트와 적용성	A > C > B	높을수록 우수
시스템 구축 비용	A > B > C	낮을수록 우수
정확성	B = C > A	높을수록 우수
확장성	A > C > B	높을수록 우수

A: 일반 웹로그 분석, B: 데이터베이스 로그생성, C: 일반 웹로그 파일에 분석전용 로그생성

5. 결론

정적인 웹문서에 대한 로그분석 방식은 현재의 CGI와 서버측 스크립트에 의한 동적인 웹사이트에서는 고객의 성향을 파악하는데 적용될 수 없음을 확인하여 고객성향 분석전용 로그발생 및 분석 알고리즘을 개발하게 되었다.

또 다른 방법으로 관련 데이터베이스에 직접 로그기록을 하는 방법도 있으나 이 경우 기존의 사이트에 많은 손질이 필요하며 확장성이 부족하여 상품화하는데 한계점이 있었다.

제안한 방법은 웹 서버측에 존재하지 않는 파일이라 하더라도 웹브라우저 클라이언트의 요청에 의해 웹 액세스 로그는 발생한다는 점에 착안하였다.

고객성향 분석전용 가상로그는 고객이 어떤 분야에 관심을 가지고 있는지, 그리고 어떤 상품에 특별히 관심을 가지고 액세스를 하였는지를 파악할 수 있는 정보를 담고 있으며, 분석 시스템에서는 이 정보를 기존의 정보와 통합적으로 분석하여 고객의 성향과 관심 물품을 최종적으로 산출 할 수 있다.

이러한 기술적 접근은 기존 업계의 단순로그분석에 비해서 고객의 관심분야와 상품까지 파악할 수 있는 차별화된 것이며, 가격대비 성능을 높일 수 있게 되었다. 또한 방대한 양의 로그데이터를 데이터웨어하우스로 관리하려면 대용량의 스토리지가 필요하며 분석에 걸리는 시간과 많은 유지비가 소요된다. 하지만 현재 제안한 시스템은 분석에 필요한 로그만 필터링하여 데이터베이스에 저장 관리함으로 전체적인 고객분석 및 마케팅지원 시스템 개발에 대한 비용을 줄일 수 있다.

쿠키를 이용하는 다른 방식은 사용자가 보안상 자신의 쿠키를 기록하지 못하게 하면 무용지물이 되고 일련의 쿠키를 읽어 서버측 데이터베이스에 저장하는 부하를 발생하게 된다. 그러나 현재 제안한 방식의 경우 사용자가 해당 페이지를 자신의 브라우저에 요청하는 가운데 웹서버측에 자동으로 간단한 분석로그가 발생하여 시스템에 걸리는 부하증

가가 없다.

그리고 기존의 CGI와 서버측 스크립트(JSP, ASP, PHP)로 개발된 시스템과 연동에 있어서 분석과 관련된 페이지에 대해서만 한 줄의 코드 추가로 분석전용 로그가 웹서버 측에 발생하게 됨으로, 개발시 커스트마이징 비용과 시간을 최소화 할 수 있는 장점이 있다. 지금까지의 연구를 바탕으로 앞으로는 무선인터넷의 모바일 환경에서 발생하는 로그 분석과 적용 방안에 대한 연구를 진행하고자 한다.

참고문헌

[1] Dong-Ha Lee, Dong-Yal Seo, Nam-Ho Kim, and Jeon-Young Lee, "Discovery and Application of User Access Patterns in the World Wide Web", *4th World Congress on Expert system 98*, March 16-20, Mexico City.

[2] Luotonen A. "Common Log Format", http://www.w3c.org/Daemon/User/Config/Logging.html, 1995

[3] Mobasher, B., Jain, N., Han, E. H. and Srivastava, J.,"Web mining: pattern discovery from world wide web transactions", Technical Report, 1996

[4] Whitepaper "Web statistics and traffic analysis software", http://www.marketwave.com/ press/whitepaper.htm

[5] Buhner A. G. and Mulvenna, M. D., "Discovering internet marketing intelligence through online analytical web usage mining", SIGMOD Record, vol. 27, No. 4, pp.54-61, 1998

[6] Borges, J. and Levene, M., "Data mining of user navigation patterns", WebKDD'99, 1999

[7] Spiliopoulou, M., Pohle, C. and Faulstich, L. C. ,"Improving the effectiveness of a web site with web usage mining", WebKDD'99, 1999

[8] Cooley, R., Mobasher, B. and Srivastava, J. "Data preparation for mining world wode web browsing patterns", Journal of Knowledge and Information System, Vol. 1, No.1, 1999

[9] Zaiane, O. R., Xin, M. and Han, J. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", IEEE International Forum on Advanced in Digital Libraries 98 Proceedings, pp. 19-29, 1998

[10] Wu, K. J. Yu, P. S. and Ballman, A. "SpeedTracer: A web usage mining and analysis tool", http://www.research.ibm.com/journal/sj371/ wu.txt

[11] Murray, D. and Durrell, K. "Inferring Demographic Attributes of Anonymous Internet Users', WebKDD'99, 1999

[12] Fu, Y., Sandhu, K. and Shin, M. Y. "Clustering of web users based on access patterns", WebKDD'99, 1999