사례 발표

실시간 대용량 Web-DB 구축과 eCRM

이 성 백*, 강 민 형**

● 목 차 ●

1. 서 론

CRM은 최근 2~3년간 가장 크게 주목받고 있는 분야 중 하나로서 그 배경에는 마케팅 개념의 변화와 정보기술의 발전이 큰 역할을 하고 있다. 기업간 경쟁이 심화되고 있는 상황에서 기업들은 제품 위주의 마케팅에서 탈피하여 고객에 주목하기 시작하였으며 이러한 경향을 시스템적으로 지원해줄 정보통신기술이 발전함으로써 CRM이 관심사로 부상하게 된 것이다.[1] 오늘날 온라인 비즈니스의 발달과 기존 오프라인 기업들의 온라인 시장 개척이 활발해짐에 따라 전자 상거래와 온라인 마케팅에 주안점을 둔 eCRM이 각광을 받고 있다.

eCRM의 첫 단계로서 국내외의 수많은 CRM 솔루션 업체들이 웹 로그 분석기를 제공하고 있다. 과거 웹 로그 분석은 페이지뷰 집계 정도의 기능을 수행하는 데에 그쳤으나 최근에는 이러한 단순 통계치를 넘어서서 경로 분석, 선호도 분석, 방문객 행동 분석 등 웹 데이터 마이닝의 기초 자료가 되는 클릭스트림 분석의 중요성이 부각되고 있다.[2]

특히, 최근 대규모 온라인 커뮤니티의 활성화 및 미디어 포탈과 인터넷 쇼핑몰의 증가 추세에 따라 방대한 양의 로그 데이터를 실시간으로 분석하고 해석하여 마케팅 활동으로 연결하는 서비스의 즉시성이 요청됨에 따라 실시간 대용량 웹 로그 분석의 필요성이 대두되고 있다.

본 논문에서는 웹 로그 분석기의 기술적인 현황을 검토하고, 패킷 스니핑 방식에 의한 실시간 웹 로그 추출 기법을 예시하고, 실제로 웹 서비스가 이루어지고 있는 사이트에 대한 실시간 대용량 웹 로그 분석 사례를 소개함으로써 실시간 대용량 웹 로그 분석 기술의 구현 가능성과 기술적 완성도 및 유용성을 검증하였다.

2. 실시간 대용량 웹 로그 분석

실시간 웹 로그 분석을 이용하면 최근의 로그 상황을 얻을 수 있으므로 eCRM 측면에서 다양한 응용이 가능하며 특히 다음과 같은 경우에 유용하다.[3]

- 신문사 웹 페이지의 경우 시시각각 새로운 기사가 추가되며 기사 배열이 바뀌므로 접속자의 반응 분석 결과가 한두 시간 이내에 갱신되어

* (주)한국정보공학 CRM 사업본부장
** (주)한국정보공학 CRM 사업부 기획팀장

야 한다. 특히 속보성 기사의 반응 분석을 통한 기사 배열 순서 결정에 실시간 웹 로그 분석이 유용하다.

- 웹 마케터가 캠페인이나 이벤트에 대한 반응을 실시간으로 분석하여 보완 캠페인을 수행하거나 이벤트 내용을 수정해야 하는 경우 실시간 웹 로그 분석기가 필요하다.
- 게임 사이트의 경우 일정 시간 이상 사용자에게 부가적인 서비스 시간을 제공하거나, 쇼핑몰의 경우 일정 금액 이상의 상품을 쇼핑 카트에 담는 순간 할인 쿠폰을 제공하는 등 즉시성 서비스의 구현을 위해서 실시간 웹 로그 분석이 요구된다.

그러나 실시간 대용량 웹 로그 분석을 원활하게 수행하려면 다음과 같은 여러 가지 조건이 충족되어야 한다. 첫째, 로그 추출 즉시 실시간으로 클릭스트림 DB에 적재가 가능해야 한다. 이 기능 없이는 실시간 로그 분석 자체가 불가능하다. 둘째, 웹 로그 추출 작업으로 인해 네트워크와 웹 서버에 과다한 부하를 유발해서는 안된다. 트래픽이 높고 대용량 웹 로그를 발생시키는 사이트에서 웹 로그 추출 작업으로 인해 네트워크와 웹 서버에 추가적인 부하가 발생하는 경우 로그 추출 작업을 안정되게 수행시킬 수 없을 뿐 아니라 웹 서비스 자체에도 위협이 될 수 있기 때문이다. 셋째, 멀티 서버에 대한 대책이 마련되어 있어야 한다. 대용량 웹 로그를 발생시키는 사이트는 대부분 멀티 서버로 구성되어 있어 로그 정보 취합의 문제가 발생하게 되며 특히 두 가지 이상의 플랫폼이 혼재하는 경우 복잡한 문제를 일으킬 수 있기 때문이다. 넷째, 유지/관리 비용이 저렴해야 한다. 활발하게 서비스가 이루어지는 대용량 사이트의 경우 서버 구성 변화나 서버 증설 등이 필요한 경우가 많으며 신규 등록 또는 수정/삭제되는 컨텐츠와 페이지의 양이 막대하므로 유지/관리비 절감이 중요한 문제가 된다. 다섯째, 보안성을 반드시 고려하여야 한다. 대규모 웹 서비스가 이루어지고 있는 사이트의 경우 보안 문제의 발생으로 인한 피해 역시 사이트의 규모에 비례하여 심각해질 수 있기 때문이다. 여섯째, 컨텐츠/카테고리 분석 기능이 필수적이다. 대규모 사이트에서는 컨텐츠와 페이지의 양이 많은 관계로 카테고리별 분류 기능 없이는 의미 있는 분석이 불가능한 경우가 대부분이기 때문이다. 일곱째, 데이터 분석 시간이 짧고 운영자의 수작업이 가능한 한 적게 투입되어야 한다. 매일 수 GB~수십 GB씩 쌓이는 로그 데이터를 처리해야 하기 때문에 ETT에 너무 많은 시간이나 운영자의 수작업이 요구된다면 관리 비용이 급증하고 실시간 로그 분석의 의미도 희석되기 쉽다.

3. 웹 로그 추출 기법

웹 로그 추출 방법은 로그 파일 분석 방식, 스크립트 방식, 패킷 스니핑 방식으로 분류할 수 있다. 각각의 방식에는 나름대로 장단점이 있으며 최근에 출시되는 웹 로그 분석 솔루션 중에는 두 가지 이상의 방식을 조합하여 구현된 예도 적지 않다. 이상의 3가지 방식 중 로그 파일 분석 방식을 제외하면 나머지 2가지 기법에서는 원리적으로 실시간 웹 르그 추출이 가능하다.[2]

로그 파일 분석 방식은 웹 서버에 쌓이는 로그 파일을 분석하는 방법으로서 가장 기본적이면서도 확실한 방법이다. 웹 서버에서 기본적으로 제공되는 자료를 이용하므로 구현이 쉽고 로그 수집 모듈이 문제를 일으킬 가능성이 적으며 일단 쌓인 로그를 잃어버리는 일은 없다. 그러나 일단 저장된 텍스트 파일을 분석해야 하므로 실시간 고객 분석에 한계가 있으며 큰 사이트의 경우 로그 저장 자체가 웹 서버에 부하를 줄 수 있다. 또한, 멀티 서버인 경우 여러 개의 로그 파일을 통합하여 분석해야 하는 부담이 있으며 서버 증설이나 사이트 구성의 변화가 있을 때마다 일일이 조정 작업을 해줄 필요가

있으므로 대규모의 사이트 분석에 적합한 방식이라고는 할 수 없다. 웹로그(www.weblog.com)와 웹트렌즈가 이 방식으로 구현되었다.

스크립트 방식은 분석 대상 페이지마다 스크립트를 삽입하여 고객의 행동을 감지하는 방법으로서 실시간 고객 분석이 가능하다. 스크립트 자체의 크기는 매우 작으므로 웹 서버의 용량에는 큰 부담을 주지 않는다. 그러나 방문자가 페이지에서 페이지로 움직일 때마다 웹 서버와 분석 서버간의 교신이 필요하므로 방문자수가 많은 사이트의 경우 웹 서버의 성능 저하와 네트워크 부하를 야기할 수 있다. 또한, 로그 추출 동작을 웹 서버가 해주어야 하므로 로그 분석 모듈에 문제가 생길 경우 로그를 잃어버릴 수 있으며 웹 서비스에 지장을 초래할 수 있다. 로그 파일 분석 방식과 마찬가지로 멀티 서버인 경우 각 서버마다 로그 추출 모듈을 설치 관리해야 한다는 약점이 있으며 분석하고자 하는 모든 페이지에 스크립트를 삽입해야 할 뿐 아니라 로그 분석기 설치 이후에도 새로이 생성되는 모든 페이지에 일일이 스크립트를 삽입하기 위한 관리 비용이 발생한다. 이 방식을 채택한 웹 로그 분석기에는 카운트보이(www.countboy.com)와 HitBox(www.hitbox.com)가 있다.

패킷 스니핑 방식은 서버가 아니라 네트워크 스위치나 허브에 설치하여 그곳을 통과하는 모든 패킷을 분석하는 방식으로서 실시간 고객 분석이 가능하고 서버에 전혀 부하를 주지 않는다. 로그 추출 모듈이 문제를 일으킬 경우 로그를 잃게 되는 점은 스크립트 방식과 같지만 서버에는 아무런 영향을 미치지 않으므로 웹 서비스에 지장을 주지는 않으며 유실된 로그 정보는 분석 서버의 로그를 분석함으로써 보충할 수 있다. 멀티 서버나 서버 증설, 사이트 구조 개편 등의 경우에도 모듈 추가를 비롯한 운영 및 관리 비용이 추가되지 않는다. 그러나 웹 서버나 네트워크에 부하를 주지 않는 대신 스위치와 분석 서버는 트래픽에 비례하여 높은 사양의 하드웨어를 요구한다는 문제점이 있다. 이 방식으로 구현된 웹 로그 분석기는 클릭마인드(www.clickmind.co.kr)와 Accrue Insight(www.accrue.com)가 있다.

<표 1> 웹 로그 추출 방식의 장단점 비교

	로그 파일 분석 방식	스크립트 방식	패킷 스니핑 방식
실시간 로그 분석	불가능	가능	가능
멀티 서버 분석	싱글 서버보다 복잡	싱글 서버 보다 복잡	멀티 서버 투명성 제공
네트워크 부하	높다	높다	없다
웹 서버 부하	높다	높다	없다
보안성	낮다	낮다	높다
유지/관리 비용	높다	높다	낮다
하드웨어 사양	낮다	낮다	부하에 따라 다르다

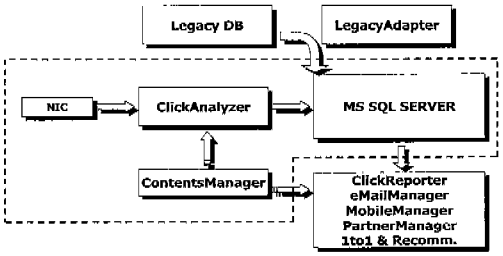
<표 1>에서 3가지 방식의 장단점을 정리하였다. 일단 실시간 웹 로그 분석이 불가능한 로그 파일 분석 방식을 제외하고, 2장에서 설정한 실시간 대용량 웹 로그 분석기의 요구 사항에 따라 나머지 두 방식의 장단점을 평가해보면 네트워크 부하와 웹 서버 부하가 적고 멀티 서버 투명성을 제공하며 보안성이 높고 유지/관리 비용에서도 강점을 가진 패킷 스니핑 방식이 실시간 대용량 웹 로그 분석에 적합한 것을 알 수 있다. 또한, 웹 로그 분석 방식의 차이에 의해서라기보다는 구현 기법에 의해 결정되는 요소이므로 <표 1>에는 포함되지 않은 컨텐츠/카테고리 분석 기능, 데이터 분석 시간, 운영자의 수작업 여부 등이 구현 단계에서 추가적으로 고려되어야 한다.

4. 패킷 스니핑에 의한 실시간 웹 로그 분석

본 논문에서는 패킷 스니핑 기법을 응용하여 구현된 (주)한국정보공학의 클릭마인드를 이용하여 고속 대용량 환경 하에서의 실시간 웹 로그 분석을

수행하였다. 클릭마인드의 구성은 (그림 1)과 같다.

클릭마인드는 웹 로그 추출 엔진인 Click Analyzer, 컨텐츠와 카테고리의 정의 및 관리를 담당하는 ContentsManager, Legacy DB 연동을 위한 LegacyAdapter, 추출된 데이터를 이용한 활용 모듈인 ClickReporter, eMailManager, MobileManager, PartnerManager 등으로 구성된다. ClickAnalyzer는 네트워크 스위치로부터 네트워크 카드(NIC)를 거쳐 추출한 패킷을 처리하여 MS SQL SERVER에 적재하며 이때 ContentsManager가 정의한 데이터 구조에 따라 클릭스트림 데이터가 재구성된다. 고객 프로파일이나 트랜잭션 이력 등의 레거시 데이터는 LegacyAdapter에 의해 MS SQL SERVER로 전달되어 통합된다. 이렇게 구축된 통합 DB를 자료로 하여 웹 로그 분석 결과 리포터인 ClickReporter와 다양한 채널별 마케팅 활동을 관리하는 eMail-Manager, MobileManager, PartnerManager 등을 운용할 수 있다.



(그림 1) 클릭마인드 구성

본 논문에서 소개할 실시간 웹 로그 분석 시스템에서는 클릭마인드의 전체 패키지 중에서 (그림 1)에서 점선으로 표시된 ClickAnalyzer와 Contents-Manager 부분만을 구축하였으며 그 부분의 상세한 과정은 다음과 같은 6단계로 구성된다.[4]

- Session Identification : 네트워크 스위치를 지나가는 패킷을 분석하여 패킷 세션을 조합하고 에러 패킷, 에러 세션을 구분하여 의미 있는 자료를 추출한다.
- Filtering Session : 분석에 필요하지 않은 아이템에 대한 방문 기록을 삭제한다. 일반적으로 gif, jpg, swf 등의 로그를 제외하게 되며 이 과정에 의해 데이터 용량이 1/10에서 1/40까지 축소된다.
- User Identification : 각 세션에 대해 unique id를 부여하여 신규 방문자를 구분하고 기존 고객에 대해서는 cookie나 login 정보를 바탕으로 id를 추출한다.
- Contents & Event Identification : 고객의 행동에 기반하여 웹 페이지에 논리적인 의미를 부여하고(Contents Identification), 분석하고자 하는 고객의 행동을 유형별로 구분한다(Event Identification).
- Visit Completion : 이미 부여된 unique id를 바탕으로 사용자의 방문 한번에 대한 경로를 파악한다.
- Formatting : 데이터 마이닝에 적합한 형태로 전환시킨다.

<표 2> 옥션 사이트 부하 상황

Network Load	
Peak bandwidth	280 Mbps
Peak time average HTTP packets	36,000 packets/sec
Peak time	19:00~23:00
Log Data Amount	
Pageviews	16 million pageviews/day
Peak time average pageviews	430 pageviews/sec
Raw log amount	30GB/day (in file)
Log amount	4GB/day (in Database)

5. 클릭마인드™를 이용한 구현 사례

대표적인 국내 경매 사이트인 옥션(www.auction.co.kr)은 회원수 410만 명, 등록 컨텐츠 155만 건, 거래 금액 1천억 원에 달하는 초대형 사이트이며 <표 2>에 보인 바와 같이 고속/대용량/고부하 상황에서 운영되고 있다.

옥션 사이트의 전체 시스템은 HP 서버 150대로 구성되었으며 이중 5대는 DB 서버(HP LX 8500), 30대는 웹 서버(HP LP), 나머지는 메일과 커뮤니티에 배치되었다(HP LP). OS는 windows 2000 Advanced Server, 스토리지 장비로는 EMC 시메트릭스를 채택하였다.

4장에서 설명한 바와 같이 클릭마인드 전체 패키지 중에서 ClickAnalyzer와 Contents Manager만으로 실시간 웹 로그 추출 시스템을 구성하였으며 상세한 구축 사양은 <표 3~4>와 같다.

<표 3> 구현된 클릭마인드 시스템 - Components

ClickAnalyzer	
Environment	
OS	Windows 2000 Advanced Server
DB	MS SQL Server 2000 Enterprise Edition
DB Connection	OLE DB
Build	Visual C++ NT Service executable file
Module Feature : Gigabit Network Card를 위하여 구현	
ContentsManager	
Environment	
OS	Windows Family
DB	MS SQL Server 2000 Enterprise Edition
DB Connection	ADO
Build	Visual Basic executable file
Module Feature : C/S 환경, Windows Family OS	

<표 4> 구현된 클릭마인드 시스템 - 하드웨어

분석 서버(클릭마인드 설치) : HP Lxr 8500	
CPU : PIII Xeon 700MHz * 8 Memory : 6GB (256MB * 24) HDD : 54GB (HP 18GB * 3) NIC : HP Gigabit NIC Card(Optical fiber), HP 10/100 NIC Card Performance : 43046 tpmC	
DB Server : HP LH 6000	Switch : CISCO Catalyst 3508 XL
CPU : PIII 800MHz * 4 Memory : 4GB (256MB * 16) HDD : 90GB (HP 18GB * 5) Performance : 37596 tpmC	Gigabit 8 port Layer 2 switch Throughput : 5.4GB

구축 완료 후 운용 결과는 다음과 같다

- DB log amount : 4GB / days
- System Resource 사용량

분석 서버 : 2.5CPU used (5 CPU idle)

DB 서버 : Half of 4 CPU used

구축된 실시간 대용량 웹 로그 분석기는 실시간 방문객 정보 업데이트(추천을 위한 기초 자료로 활용 중), 실시간 고객 행동 데이터로부터 유용한 클릭스트림 정보 추출, 백업을 위한 페이지뷰 요약 데이터 작성 등의 3가지 용도로 운용중이며 현재 옥션 사이트의 서비스에 원활하게 이용되고 있다.

6. 결 론

본 논문에서는 웹 로그 분석기의 기술적인 현황을 정리하고, 패킷 스니핑 기술로 구현된 웹 로그 분석기인 클릭마인드TM를 이용하여 옥션 사이트(www.auction.co.kr)에 대한 실시간 대용량 웹 로그 분석 시스템 구축함으로써 실시간 대용량 웹 로그 분석 기술의 기술적 완성도 및 유용성을 검증하였다.

참고문헌

[1] 오정숙, 국내외 CRM 시장의 현황, 전망 및 문제점, KISDI IT Focus, 2001년 6월호.

[2] 김형택 & 민옥길, 효과적인 인터넷 마케팅을 위한 웹 로그 분석, 도서출판 비비컴, 2001.

[3] 이성백, 실시간 클릭스트림 DB 구축과 분석, 한국데이터마이닝학회 춘계 컨퍼런스, 2001.

[4] 이성백, 웹 사이트의 방문자 행동 분석 및 관리, 컴덱스 코리아 2001 컨퍼런스, 2001.

저자약력

이 성 백

1985년 서울대학교 산업공학과(공학사)
1987년 한국과학기술원(KAIST) 산업공학과(공학석사)
1991년 한국과학기술원(KAIST) 산업공학과(공학박사)
1991년-2000년 한국국방연구원 선임연구원
2000년-2001년 (주)이씨마이너 연구소장
2001년-현재 (주)한국정보공학 CRM 사업본부장
관심분야 : CRM, DW, Mobile Business
E-mail : ibizlee@kies.co.kr

강 민 형

1993년 서울대학교 기계설계학과 (공학사)
1995년 서울대학교 기계설계학과 유공압제어 및 생산자동화 (공학석사)
1995년-2000년 KIST 연구원
2001년-현재 (주)한국정보공학 CRM 사업부 기획팀장
관심분야 : CRM, DW, Mobile Business
E-mail : staire@kies.co.kr