비즈니스 인텔리전스를 위한 지능적 웹 로거

임윤선[0] 정안모 김 명
이화여자대학교 컴퓨터학과
{lys96, jam96, mkim}@ewha.ac.kr

An Intelligent Web Logger for Business Intelligence

Yoonsun Lim[0] Anmo Jeong Myung Kim
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

웹 로그는 웹 서버를 통해 이루어지는 작업들에 관한 기록으로써, OLAP이나 데이터 마이닝과 같은 비즈니스 인텔리전스 기술로 분석되어 고부가가치 창출에 사용되는 중요한 자료이다. 웹 로그에는 파일 이름과 같은 물리적인 데이터가 저장되는데, 이러한 데이터는 분석에 사용되기 전에 정제과정을 통해 의미 있는 데이터로 변환되거나 불필요한 경우에는 삭제된다. 웹 로그 데이터의 분량을 적정선으로 유지하면서 데이터 정제 작업의 일부가 해결되도록 하는 방법으로 웹로그 생성단계에서 시스템이 제공하는 필터를 쓸 수 있다. 그러나, 필터로는 웹 페이지의 내용이 동적으로 변경되는 경우 그 상황을 즉시 반영하기가 쉽지 않다. 본 연구에서는 웹 로그가 '지능적 웹 로거'를 통해 생성되도록 하여 이러한 문제를 해결하였다. '지능적 웹 로거'를 통해 불필요한 데이터의 생성을 막고, 물리적인 데이터를 신속하게 의미 있는 데이터로 변환하도록 하였다. 웹 페이지의 변경 내용을 웹 로그 생성에 즉시 반영하여 의미 있는 데이터 생성에 이용함으로써, 웹 로그 생성 후에 실행되던 데이터 정제작업 자체를 단순화시켰고, 웹사이트 관리자가 편리한 사용자 인터페이스로 로그 규칙을 만들어 적용할 수 있도록 하였다.

1. 서 론

웹 서버는 웹 사이트를 방문하는 사용자의 정보 및 페이지 에러, 사용 브라우저, 사용 운영체제 그리고 방문시간, 다운로드 파일, 방문 페이지, ISP 업체 등의 정보를 파일 형태로 저장하게 되는데, 이를 웹 로그(web log)라고 부른다. 이러한 웹 로그 데이터는 사용자의 선호도 및 행동양식을 파악하여 마케팅 전략을 수립하는데 기반이 되는 중요한 데이터이다[1]. 특히, 전자상거래 시스템에서는 웹 로그 분석을 통해 개인화 된 서비스[2]를 할 수 있으며, 이러한 서비스의 제공 여부는 인터넷 혁명 시대에서의 전자상거래의 승패를 좌우한다.

웹 로그 데이터는 분석에 사용되기 전에 정제 과정을 거치게 된다. 분석에 관한 사전 지식이 없이 저장되는 웹 로그 데이터는 파일 이름, 방문 페이지 등과 같이 물리적인 형태로 저장될 뿐 아니라, 웹 서버에서 이루어지는 모든 작업이 기록되는 것이므로 그 양이 방대하다. 데이터 정제 과정에서 주로 이루어지는 일로는 '파일 이름'과 같은 물리적인 데이터를 '상품 이름'과 같이 의미 있는 데이터로 변환하거나, 분석에 불필요한 웹 페이지 내의 그림 정보 등을 제거하여 보다 가치있는 정보를 추출하는 것이다.

웹 로그 데이터는 생성 단계에서 필터를 써서 부분적으로 정제할 수 있고, 분량을 조절할 수 있다. 현재 상용화된 마이크로소프트의 IIS 서버[3] 또는 Apach 서버[4]와 같은 웹 서버나 웹 트렌즈[5]와 같은 웹 로그 분석 툴 들은 웹 로그 데이터를 생성 단계에서 필터링하는 기능을 제공한다. 그러나 이들은 웹 로그의 분석에 무의미한 파일들을 파일 형식을 기준으로 제거하는 수준에 그친다. 예를 들어, 웹 페이지에 포함된 gif, jpg와 같은 포맷의 파일들은 제거할 수 있으나, 이러한 이미지 파일 중에서 특정 파일을 선택적으로 제거하도록 하지는 않는다. 또한 필터를 변경하려면 필터 프로그램을 수정하여 이를 DLL로 만들어 넣어야 하는 불편함이 있다. 이러한 정도의 기능을 하는 필터를 제외하고는, 웹 로그를 생성과정에서 정제하여 저장하는 연구는 미비하다.

본 연구에서는 이러한 필터 기능을 강화하는 '지능적 웹 로거'를 설계하였다. 웹 로그는 '지능적 웹 로거'를 통해 생성된다. 웹 로거는 신속하게 처리될 수 있는 다양한 기능들을 제공하는데, 이를 통해 불필요한 데이터의 생성을 막을 수 있고, 매핑 테이블을 사용하여 물리적인 데이터를 의미 있는 데이터로 신속하게 변환시킬 수 있다. 편리한 인터페이스를 통해 로그 규칙을 변경할 수 있기 때문에 웹 페이지가 변경될 때마다 이를 즉시 웹 로그의 생성에 반영할 수 있고, 이러한 기능은 정밀한 필터링없이 저장한 웹 로그 데이터 상에서 정제 작업을 하는 것에 비해 작업을 단순화시키는 역할을 한다.

본 논문의 구성은 다음과 같다. 2절에서 본 연구에서 제안하는 '지능적 웹 로거'를 소개하고, 3절에서 본 연구에서 제시한 방법으로 웹 로그 분석 시스템(Web Log Analyzer)을 구축한 예를 설명하고, 4절에서 결론을 맺는다.

* 본 연구는 2000년도 한국과학기술부 여자대학교 연구기반확충사업 (과제번호:00-B-WB-06-A-02) 지원에 의해 수행되었음.

2. 지능적 웹 로거

일반적으로 사용자가 웹 사이트를 방문하면 웹 서버는 그림 1에서와 같이 요청된 웹 페이지를 전달하면서 그와 관련된 모든 로그 데이터를 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. 본 연구에서는 이와 같은 전통적 시스템 구조에 추가적으로 그림 2와 같이 '지능적 웹 로거'를 추가한 시스템 구조를 제안한다. 웹 로거는 로그 규칙 관리자와 정의된 로그 규칙이 저장된 매핑 테이블로 구성된다. 로그 규칙 관리자는 웹 서버 관리자가 규칙을 생성하거나 변경하는 일을 수월하게 해 주고, 웹 서버가 서비스를 할 때마다 생성되는 원시 웹 로그 데이터를 로그 규칙을 참조하여 정제하는 일을 한다.

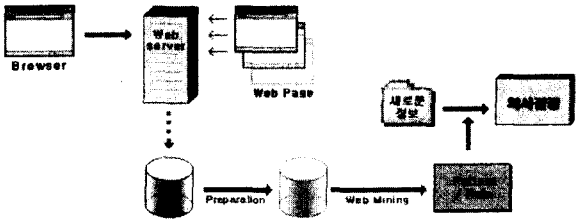

그림 1. 기존 웹 로그 분석 시스템.

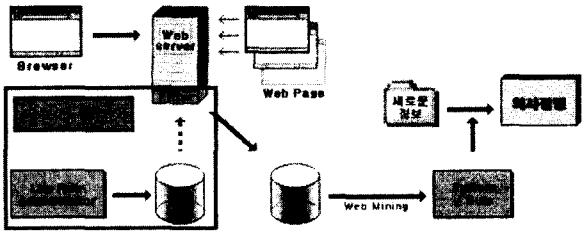

그림 2. 지능적 웹 로거를 이용한 웹 로그 분석 시스템.

우선 일반적인 웹 로그 데이터를 소개하고 지능적 웹 로거가 하는 일을 소개하기로 한다. 사용자가 웹 서버에 접속하면 그 이후의 모든 작업들에 관한 명세는 웹 서버가 정의한 항목마다 파일에 저장된다. 초기에는 웹 서버마다 로그 항목들이 달랐으나, 현재 대부분의 웹 로그는 CERN과 NCSA에서 HTTP 프로토콜로 규정한 Common Log Format(CLF)[6]을 따른다. 이 형식에서는 사용자의 IP 주소, 데이터 전송 프로토콜, 에러 코드, 전송된 데이터 길이 등이 포함되고 이 들 항목들간의 구분은 일반적으로 스페이스를 사용한다. 값이 없는 항목들은 '-'로 기록된다. 그림 3은 마이크로소프트사의 IIS 웹 서버에 저장되는 웹 로그 항목들이다. 웹 서버마다 CLF이외의 풍부한 로그 항목들을 제공하고 있으며, 사용자는 필터를 통해 이를 선택하여 저장할 수 있다.

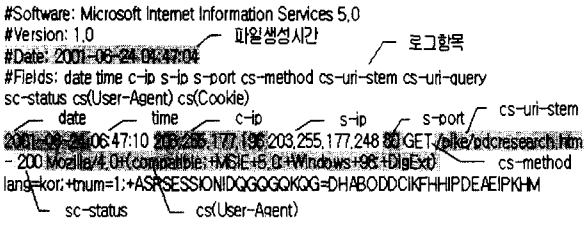

그림 3. IIS 서버에 생성된 로그 파일 예.

2.1 불필요한 데이터의 제거

사용자가 특정 웹 페이지를 요구하면 웹 서버는 해당 웹 페이지 뿐 아니라 그 페이지에 포함되어 있는 이미지 파일이나 동영상 파일 등에도 접근하여 웹 로그에 그 내역을 기록한다. 는 고객의 행동 양식 파악에 무관한 웹 사이트 장식 데이터인 경우가 대부분이고, 그 분량은 전체 데이터의 75%에 이른다고 한다[7]. 그러나 이미지 파일이 전자 상거래 사이트의 특정 상품인 경우는 다르다. 그 상품 이미지를 클릭하였다면 이는 분석에 중요한 정보가 된다. 본 연구에서 제안하는 지능적 웹 로거는 기존의 필터들과는 달리 객체들을 객체 단위, 또는 객체 그룹 단위로 웹 로그에 남길 것인가를 결정한다. 제거할 객체들은 로그 규칙으로 설정되어 매핑 테이블에 기록되고 매핑 테이블은 해쉬을 통해 엑세스되어 신속하게 처리가 가능하다.

2.2 물리적 데이터를 의미있는 데이터로 변환

일반적으로 사용자가 전자상거래 웹 사이트에서 관심 있는 물품을 나타내는 이미지를 선택하면 웹 로그는 해석하기 힘든 암호와 같은 값을 저장한다. 이 값은 웹 로그 분석 이전에 의미 있는 데이터로 변환되어 사용된다. 대부분의 웹 로그 분석은 필터링 단계에서 데이터를 변환하여 저장하지 않고 데이터 정제 과정에서 현재 구축된 웹 사이트를 파악하여 분석할 수 있는 데이터로 변환한다. 따라서, 데이터 정제가 되지 않은 상태에서 웹 사이트가 개편이 된다면 기존 웹 사이트를 파악할 수 없게 되므로 웹 로그 파일에 저장되어 있는 중요한 데이터는 쓸모 없는 데이터가 된다. 본 연구는 그림 4와 같이 필터링 과정에서 의미 있는 데이터로 변환할 수 있도록 로그 규칙을 만들어 설정할 수 있다.

URI Stem	Delete	/product2/greeting.gif	-
URI Query	Exchange	id=item012837	Color TV

그림 4. 로그 규칙 매핑 테이블.

이제 본 연구에서 제안한 지능적 웹 로거의 효율성에 관해 살펴보기로 한다. 우선 웹 로거의 성능에 관해 설명하고, 웹 로거의 사용을 통한 데이터 정제 작업의 효율성 증진에 관해 설명하기로 한다.

2.3 웹 로거의 성능

지능적 웹 로거는 위에서 언급한 두 가지 기능에서 확장하여 각 웹사이트에 맞는 로그 규칙을 적용할 수 있다. 그리고 이를

효율적으로 처리하기 위해 매핑 테이블을 사용한다. 매핑 테이블에는 각 객체를 선택하여 웹 로그에 포함시킬 것인가와 만약 포함된다면 어떤 의미 있는 데이터로 변환하여 저장할 것인가가 기록되어 있다. 매핑 테이블은 웹 사이트의 규모에 따라 크기가 커질 수 있으나 각 항목을 해쉉(hashing) 기법[8]으로 신속하게 찾기 때문에 웹 서비스와 동시에 큰 오버헤드 없이 웹 로그의 생성이 실시간에 이루어질 수 있다.

2.4 데이터 정제 작업의 효율성 증진

전자상거래 시스템과 같이 상품의 종류가 지속적으로 추가, 삭제되는 경우는 웹 사이트의 웹 페이지들, 이미지들을 포함한 모든 정보들이 동적으로 수시로 변경된다. 웹 로그 데이터를 오프라인으로 나중에 정제하려 한다면 시간 정보를 기준으로 특정 객체가 무엇을 의미하는가에 관한 기록을 가지고 이를 반영하여 물리적 정보를 의미 있는 정보로 변환해야 한다. 그러나, 웹 페이지들이 변경될 때 이러한 정보를 미리 로그 규칙으로 만들어서 웹 로그 관리자에게 등록시켜 놓는다면 생성되는 웹 로그는 항상 의미 있는 데이터로 변환되어 저장된다.

3. 구축 사례

그림 5는 로그 규칙을 적용하여 구축한 웹 로그 분석 시스템(Web Log Analyzer) 구조이다. 구축한 환경은 마이크로소프트의 IIS 5.0 웹서버를 사용하고 ISAPI 필터를 통해 로그 규칙을 적용하는 프로그램과 로그 규칙 관리자 프로그램을 작성하였다. 로그 규칙을 적용하여 웹 로그 파일에 저장된 웹 로그 데이터는 SQL Server2000의 DTS(Data Transformation System)를 이용하여 데이터 정제 작업 없이 바로 데이터베이스에 로딩 시킨다. 데이터베이스에 만들어진 웹 로그 데이터는 마이크로소프트 Analysis Services[3]의 OLAP과 데이터 마이닝을 이용하여 분석해서 기업 전략을 수립할 수 있도록 한다.

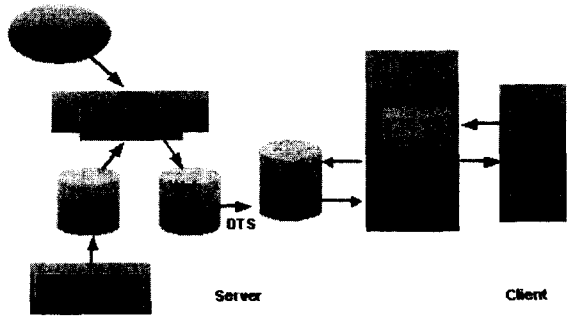

그림 5. Web Log Analyzer 시스템 구조.

Web Log Analyzer는 그림 6과 같이 관리자가 사용하기 편한 인터페이스를 통해 웹사이트에 맞는 로그 규칙을 설정할 수 있도록 하였다.

4. 결론

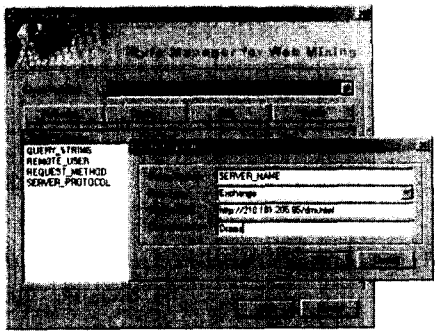

그림 6. 로그규칙 관리자 프로그램.

본 연구에서는 웹 로그를 효율적으로 생성하고 저장하는 방법을 제안하였다. 지능적 웹 로거를 두어 웹 서버가 웹 서비스를 함과 동시에 웹 로그를 분석에 적절한 형태로 동시에 변환하여 저장할 수 있도록 함으로써 웹 로그의 분량을 크게 감소시켰고 데이터 정제 작업 자체를 단순화시켰다. 제안한 지능적 웹 로거는 해쉉 방식으로 접근되는 매핑 테이블을 사용하여 데이터 정제 작업을 하므로, 웹 서비스의 속도를 크게 저하시키지 않는다. 미래의 전자상거래 시스템은 개인화된 서비스를 하기 위해서 사용자가 접속되어 있는 상태에서 데이터 정제는 물론 고객 정보를 온라인으로 분석하고 데이터 마이닝을 하여 사용자의 패턴을 즉시 탐지하고 이에 대처해야 한다. 이러한 일이 모두 온라인으로 처리되어야 할 상황에서 본 연구는 웹 로그 생성과 웹 로그 데이터의 정제를 신속하게 처리했다는 점에 의의가 있다고 본다.

5. 참고 문헌

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

[2] Alexander Pretschner, Susan Gauch
"Personalization on the Web,"
Technical Report ITTC-FY2000-TR-13591-01, 1999.

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

[4] http://www.apache.org

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

[6] Phillip M. Hallam-Baker, Brian Behlendorf
"Extended Log File Format,"
http://www.w3.org/pub/WWW/TR/WD-logfile.html

[7] http://www.i-biznet.com

[8] Karel Driesen,
"Selector Table Indexing & Sparse Arrays,"
In Proc of the ACM OOPSLA'93 Conference, 1993.