2001년도 대한전기학회/대한전자공학회 시스템및제어분야 합동 추계학술대회 논문집 (2001.11.24)

패킷 정보를 이용한 웹사용 분석에 관한 연구

곽미라*, 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과

Web Usage Analysis Using Packet Data

Mira Kwak*, Dong-sub Cho
Dept. of Computer Science and Engineering EIST Ewha Womans University

Abstract - 웹사용자들의 서비스 사용 분석에 관한 기존의 연구는, 웹서버가 기록하는 표준로그파일과 쿠키정보를 분석하는데 바탕을 두고 이루어져왔다. 이러한 방법으로는 웹사이트 방문자의 행동에 관한 상세한 정보를 파악할 수 없다. 본 연구에서는 네트웍 상에서 오고가는 패킷들을 캡쳐하고 HTTP에 관한 패킷들을 걸러내어 저장함으로써 웹서버로그파일이나 쿠키정보만으로는 파악할 수 없는 웹사용정보를 취한다. 또한 얻어낸 정보들에 대해 데이터마이닝 기법을 적용하여 웹사용자의 행동을 분석하여 VRM(visitor relationship management)이 가능하게 한다. 이를 위해 패킷의 캡쳐 및 필터링 기법, 패킷정보로부터 웹사용정보의 구성방법, 실시간으로 증가하는 정보들의 저장 및 처리기법, VRM을 위한 데이터마이닝 기법을 고안하였다.

1. 서 론

웹 사용의 증가와 더불어 웹 사이트의 성공적인 운영을 위한 방문자들의 행위 파악의 중요성이 부각되었다. 이에 따라 방문자들의 행위를 파악하고 분석하는데 관한 연구가 활발히 진행중이다. 현재까지 많은 연구들이 웹서버가 생성하는 로그파일, 쿠키나 적절히 제작된 웹브라우저가 서버로 전송하는 사용자 행위관련 내용을 분석하는데 바탕을 두어왔다. 하지만, 이런 방법들은 충분한 정보의 제공이 불가능하거나 실용화되기 힘들다는 문제점을 가진다. 본 연구에서는 이러한 문제를 해결하기 위하여, 네트웍 패킷을 모니터링하고 필요한 내용을 저장하여 웹로그와 함께 사용자 행위를 분석하는데 사용하고자 하였다. 또한 키워드 추출, 연관규칙탐사, 순차패턴탐사 등의 데이터마이닝 기법을 적용하여 지금까지 통계적 수준에서 많이 벗어나지 못하고 있는 웹사용 분석 기능을 지능적인 수준으로 발전시키고자 하였다.

2장에서 이와 관련한 연구내용들을 소개하고, 3장에서 본 논문이 제안하는 웹사용 분석 시스템의 전체적인 설계를 보인다. 4장에서는 이 시스템을 위한 기술들을 설명하고, 5장에서 결론과 향후 연구할 과제들을 이야기함으로써 논문을 맺는다.

2. 관 련 연 구

2.1 웹사용 정보의 수집

웹사이트에 대한 사용자들의 방문과 행위를 파악하기 위한 정보의 수집에 다음과 같은 방법들이 사용되고 있다[3].

- 웹서버 로그
- 특수하게 제작된 웹브라우저 등

이러한 각 방법들이 고유의 장점을 가지지만 기록될 수 있는 정보의 상세한 정도에 있어 한계를 가진다.

이 논문은 2001년도 두뇌한국21사업에 의하여 지원되었음

웹서버는 접속한 클라이언트와 상호작용을 하나 이상의 로그 파일이나 데이터베이스로 로깅하거나 로그 정보를 다른 애플리케이션으로 실시간 파이핑 할 수 있다[7]. 웹서버가 기록하는 로그데이터 내용의 요소들은 웹서버 로그의 표준인 CFL(Common Log Format)이나 ECLF(Extended Common Log Format)이 포함하는 파라미터들을 기본으로 웹서버의 종류에 따라 다른 파라미터들로 구성된다. 이러한 웹로그데이터는 웹사이트 방문자들과 그 행위에 관한 기본적인 정보를 기록하지만, 기본 HTTP 프로토콜에 포함된 정보 외의 것들은 기록할 수 없다는 한계를 가진다.

자세한 사용자 정보의 수집을 목적으로 별도 제작된 브라우저는 서비스 제공자 측에서 필요로 하는 정보의 수집을 충분하게 할 수 있다. 하지만 넷스케잎사의 네비게이터나 마이크로소프트사의 인터넷익스플로러 등 널리 사용되는 웹브라우저에 익숙한 사용자로 하여금 특정 사이트에 접속할 때마다 전용 브라우저를 사용하도록 하는 데에는 무리가 있다.

따라서, 웹사용자가 인식하지 않도록 하면서 추가적인 웹사용 정보를 파악하기 위한 방법이 필요하다.

2.2 웹사용 정보의 분석

현재 웹사용 분석을 위하여 널리 사용되고 있는 제품들은 표1과 같다.

표 1. 현재 제공되고 있는 웹로그 분석 솔루션들

제품명	제품형태	회사명
웹로그 등	패키지, ASP	(주)웹로그
웹트렌즈	패키지	(주)애드버네트
웹몬스터	패키지	스콥정보통신(주)
파워씨 클릭 애널라이저	패키지	(주)소프트다임
AccessWatch	패키지	Maher Consulting Corporation
웹아날리	패키지	(주)엠엔에스닷컴
카운트보이	ASP	(주)퍼슨앤퍼슨
LogASP	ASP	삼정데이타서비스(주)

패키지 형태로 배포되는 제품들 대부분은 웹로그 파일을 입력으로 사용하여 통계적 처리를 한 결과를 웹사이트 관리자에게 리포팅하는 방식으로 동작한다. ASP(Application Service Provider) 형태로 제공되는 분석 서비스들은 분석하고자 하는 사이트의 로그서버를 입력데이터로 사용하는 것은 아니지만, HTTP 프로토콜이 제한하는 범위 내의 정보들을 기본정보로 하여 분석기가 결과를 출력한다는 점에서 패키지 제품과 기능면에서 큰 차이를 보이지 않는다.

이들 패키지가 제공하는 분석 내용은 표2와 같다.

표 2. 웹로그 분석툴들이 제공하는 보고내용

종합 현황	**방문객 현황**
기본 보고서	방문객 시스템현황
종합 보고서	방문객 브라우저목록
기간별 분석	방문객 OS목록
월별분석	등록사용자현황
사용현황-종합,	등록사용자 방문현황
접속현황, 방문현황,	등록사용자 요금계산
에러현황, 데이터량현황	방문 형태현황
요일별분석	접속횟수순, 방문횟수순,
사용현황-종합,	데이터량순, 접속일자순,
접속현황, 방문현황,	방문횟수별 방문객수
에러현황, 데이터량현황	**방문객 분석**
평일/주말비교표	방문객 정보분석
일별분석	국가, 시도, 업종별 방문분석
사용현황-종합,	가나다순 분류
접속현황, 방문현황,	전체 방문객목록
에러현황, 데이터량현황	ISP 방문분석
시간대별 분석	접속횟수순, 방문횟수순,
사용현황-종합,	데이터량순, 접속일자순 방문객수
접속현황, 방문현황,	전체 ISP목록
에러현황, 데이터량현황	
고객 분석	**페이지 분석**
고객 정보분석	디렉토리 분석
국가별 고객 방문분석	TOP 디렉토리목록
시도별 고객 방문분석	2nd 디렉토리목록
고객 ISP 방문분석	3rd 디렉토리목록
접속횟수순 고객분석	총 디렉토리목록
방문횟수순 고객분석	메뉴 분석
데이터량순 고객분석	TOP LEVEL
접속일자순 고객분석	2nd LEVEL
전체 고객 ISP 목록	3rd LEVEL
고객 방문형태 현황	메뉴분석-종합
접속횟수순 고객	파라미터 분석
방문횟수순 고객	파라미터 분석
방문시간순 고객	**시스템 분석**
데이터량순 고객	사이트 분석
접속일자순 고객	소개사이트 목록
전체 고객 목록	기간별 에러분석
페이지현황	월별 발생현황
사용형태별 현황	요일별 발생현황
접속횟수순 검색현황	시간대별 발생현황
소요시간순 검색현황	발생횟수별 에러분석
접속일자순 검색현황	일일 발생현황
방문형태별 현황	페이지별 발생현황
시작페이지 목록	방문객별 발생현황
종료페이지 목록	없는 페이지목록
종류별 현황	
종류별 페이지목록	
전체 페이지목록	

표2의 내용에서 보이는 보고내용들은 대부분 기본적인 통계처리의 수준에 머무른다. 웹사용 분석이 사이트 방문자의 행위를 파악하여 그의 사이트 네비게이션을 돕고 필요한 정보를 찾기 쉽도록 하는 등의 VRM(Visitor Relationship Management)나 CRM(Customer Relationship Management) 수준의 요구를 충족하는데 도움이 되는 분석시스템은 현 로그분석기보다 지능적인 기능을 가져야 한다.

2.3 본 연구의 접근방법

불충분한 정보를 입력으로 사용하고 분석기능의 수준이 낮은 기존 웹사용 분석 연구의 한계를 극복하기 위해 본 연구에서는 입력 정보의 구성에 TCP 패킷정보를 활용하고 데이터마이닝 기법들을 분석과정에 적용하였다. 3장과 4장에서는 본 연구에서 설계한 분석 시스템을 자세히 설명한다.

3. 웹사용 분석 시스템의 설계

그림 1은 본 논문이 제안하는 웹사용 분석 시스템의 설계내용을 보이고 있다.

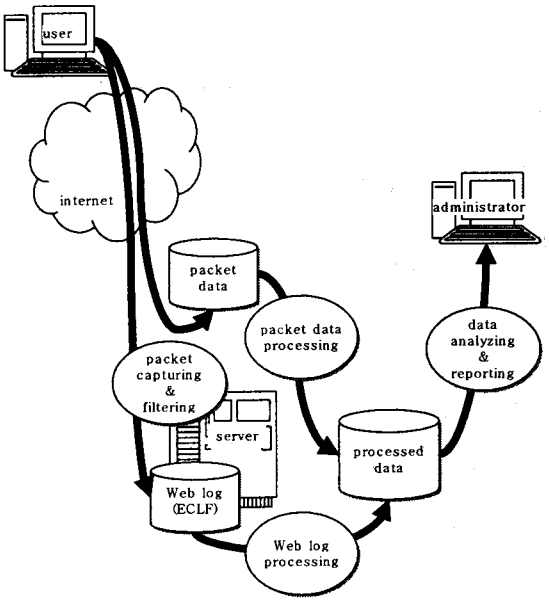

그림 1 웹사용 분석 시스템의 설계

웹 사이트에 접속한 사용자와 웹서버 사이에 일어난 모든 사건의 패킷들은 패킷을 수집하고 필터링 하는 모듈에 의해 수집되어 HTTP에 관한 내용들만 걸러져 로그서버의 파일시스템에 저장된다. 이 내용의 양이 방대하기 때문에 이후 분석에 적절한 형태로 실시간 변환하여 저장하는 것은 불가능하며, 텍스트형태로 특정 디렉토리에 저장된다. 이렇게 저장된 패킷 정보들과 웹서버의 기본적인 로그정보들 중 일부가 함께 로그 서버의 데이터베이스에 적절한 형태로 저장된다. 이것은 이후 분석 작업을 효율적으로 처리하기 위한 과정이며, 주기적으로 수행된다. 데이터베이스화된 로그데이터는 웹사이트관리자의 분석 요청에 결과를 보인다. 4장에서는 이 시스템을 구성하는 각 기술요소들을 자세히 설명한다.

4. 정보의 수집과 분석 기술

4.1 패킷 정보의 수집

본 연구에서는 패킷 정보를 수집하는데 libpcap을 사용하였다. 패킷 캡쳐와 필터링 라이브러리인 libpcap은 시스템환경에 구애받지 않고 사용자수준에서 패킷을 다룰 수 있게 하며, 이를 사용하여 구현한 패킷 처리 모듈은 이에 바탕을 둔 winpcap이라는 라이브러리를 사용한 윈도용 패킷 처리 모듈로 변환하기 쉬우므로, 현재 유닉스와 리눅스를 기반으로 설계된 본 시스템의 윈도용 버전을 개발하는데 유리하다.

실시간으로 로그 서버에 적재되는 패킷 정보는 주기적으로 패킷 정보 처리기에 의하여 웹사용 분석의 입력으로 사용되기에 적절한 형태로 변환되어 데이터베이스에 저장된다. 처리의 내용은 HTTP 헤더의 재구성, POST 파라미터의 분석을 통한 사용자 요청내용 재구성, 서버의 응답 컨텐트의 재구성 및 키워드 추출을 포함한다. 그림2는 이를 나타낸다.

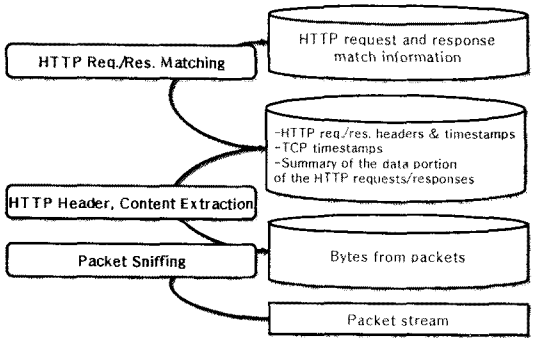

그림 2 패킷정보의 수집과 요청/응답 내용의 구성

수집되고 패킷 정보와 웹서버 로그로부터 재구성된 데이터는 사용자구분, IP, 브라우저, 운영체제, 접속일시, 참조사이트정보, 요청, 응답내용(키워드) 등을 포함한다.

4.2 연관규칙탐사

데이터로부터 숨겨진 패턴을 찾아내는 연구 중 연관규칙탐사에 대해 가장 많은 연구가 이루어져왔다. 연관 규칙탐사는 한 항목그룹과 다른 항목그룹 사이에 존재하는 강한 연관성을 밝힌다.

각 레코드들이 항목집합인 레코드들의 집합이 주어졌을 때, $support(X)$는 항목집합 X를 포함하는 레코드들이 전체 레코드 집합에서 차지하는 퍼센트를 나타낸다. 연관규칙은 $X \rightarrow Y, [c, s]$로 표현한다[6]. 이 때 X와 Y는 각각 항목집합이며, $X \cap Y = \varnothing$, $s = support(X \cup Y)$는 이 규칙의 지지도, $c = \frac{support(X \cup Y)}{support(X)}$는 이 규칙의 신뢰도이다.

본 연구에서는 사용자구분과 요청으로 구성된 레코드집합과 사용자구분과 응답내용으로 구성된 레코드집합에 대해 연관규칙탐사 알고리즘을 적용하였다. 이를 통해 특정 사용자가 주로 검색하고자 하는 내용과 찾아낸 내용을 파악할 수 있다.

4.3 순차패턴탐사

순차패턴탐사는 한 트랜잭션 안에서 발생하는 항목들 간의 연관규칙에 시간의 변이를 추가한 것이다. 연관규칙탐사가 트랜잭션 안에서 함께 발견되는 항목들을 발견하는 트랜잭션 내의 문제라 한다면, 순차패턴탐사는 트랜잭션 간의 문제라 할 수 있다.

시간기록된 사건 레코드들의 집합이 주어지고 각 레코드는 항목집합일 때, 기간 $[t_1, t_2]$는 t_1에 시작하고 t_2에 끝난 사건 레코드들의 시퀀스이다. 기간의 넓이는 $t_2 - t_1$로 정의된다. 항목집합 X가 있을 때, 어떤 기간이 X를 포함하며 그 기간의 어떠한 부분기간도 X를 포함하지 않는 경우 그 기간을 X의 최소발생이라 한다. $support(X)$는 X를 포함하는 최소발생의 모든 사건 레코드들 개수에 대한 비율로 정의된다. 순차패턴규칙은 $X, Y \rightarrow Z, [c, s, w]$로 표현된다[4]. 이 때 X, Y, 그리고 Z는 모두 항목집합들이며, 이들이 함께 한 순차패턴을 이룬다. $s = support(X \cup Y \cup Z)$는 이 규칙의 지지도이며, $c = \frac{support(X \cup Y \cup Z)}{xupport(X \cup Y)}$는 이 규칙의 신뢰도이다. 각 발생의 넓이는 w보다 작아야 한다.

본 연구에서는, 연관규칙탐사를 적용한 경우와 마찬가지로, 사용자구분과 요청으로 구성된 레코드집합, 사용자구분과 응답내용으로 구성된 레코드집합에 대해 순차패턴탐사를 적용하였다. 이를 통해 특정사용자의 한 세션 동안의 탐색경로를 파악하는데 그치던 기존 웹로그 분석 시스템의 한계를 벗어나, 사용자들의 시간에 따른 탐색 패턴까지 파악할 수 있다.

4.4 키워드 추출

웹사용 분석에는 웹사용자의 요청에 대한 서버의 응답 페이지 내용을 기록하여 요청과 응답의 쌍을 구성하고 특정 사용자가 주로 요청한 내용들을 파악하는 기능이 요구된다. 서버가 사용자에게 제공한 모든 컨텐트 내용을 기록하는 것은 시간적, 공간적으로 많은 소모를 가져오며, 이후의 분석에도 효율적이지 않다. 본 연구에서는 서버의 응답 컨텐트로부터 키워드들을 추출하여 기록하는 방법을 사용하였다. 이를 위해 HAM(Hangul Analysis Module)을 사용하였다[1].

5. 결 론

본 논문에서는 패킷 정보를 활용하여 웹사용분석을 위해 보다 충분한 정보를 수집할 수 있도록 하였고, 데이터마이닝 기법을 적용하여 웹사용자의 행위를 파악하여 웹 사용자의 요청과 서버의 응답 사이의 관계, 네비게이션의 일반적 혹은 사용자에 따른 개별적 패턴 등을 찾을 수 있는 분석 시스템을 설계하였다.

패킷 정보의 수집은 웹사용 분석 이상의 목적으로 활용될 수 있다. 패킷정보들로부터 텔넷세션들과 셸명령 기록을 구성할 수 있으며, 이들 정보에 대해 데이터마이닝 기법을 적용함으로써 웹서버 공격을 탐지할 수 있다. 본 논문에서 제안한 시스템에 이러한 기능을 추가함으로써 이 시스템은 종합적인 웹서버 관리 시스템으로 확장될 수 있다.

또한 사용자에 대한 서버의 응답 컨텐트로부터 추출한 키워드들을 바탕으로 클러스터링 하여 사용자들을 그 관심 내용별로 그룹화할 수 있다. 이는 향후 웹사이트의 웹어플리케이션 데이터베이스와 연동하여 분석 내용을 바탕으로 웹사용자에게 개인화된 서비스를 제공하는데 도움이 된다.

[참 고 문 헌]

[1] 강승식, "HAM : 한국어 형태소 분석 라이브러리", http://nlp.kookmin.ac.kr/HAM/kor/ham-intr.html
[2] 김형택, 민옥길 , "웹 로그 분석 : 효과적인 인터넷 마케팅을 위한", 비비컴, ISBN: 8-986-52590-9, 2001년 09월
[3] A. Feldmann, "BLT: Bi-Layer Tracing of HTTP and TCP/IP", In Proc. of the Ninth Int. World Wide Web Conference, May 2000
[4] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovering frequent episodes in sequences", In Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining, Montreal, Canada, August 1995
[5] J. Srivastava, R. Cooley, M. Deshpande, and P-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data", SIGKDD Explorations, 1(2):12-23, Jan 2000
[6] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 207-216, 1993
[7] R. Kimball, R. Merz, "The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse", Wiley, ISBN: 0-471-37680-9, January 2000
[8] V. Jacobson, C. Leres, and S. McCanne, "pcap - Packet Capture library." UNIX man page
[9] W. Lee, "A data mining framework for building intrusion detection Models", In IEEE Symposium on Security and Privacy, pages 120-132, Berkeley, California, May 1999