Web Page Fingerprinting을 위한 Web Server 구현

박수빈○ 조동섭
이화여자대학교 컴퓨터학과
subin.ewha@gmail.com, dscho@ewha.ac.kr

Web Server Design For Web Page Fingerprinting

Su-bin Park○, Dong-sub Cho
*Dept. of Computer Science and Engineering, Ewha Womans University

요 약

디지털 핑거프린팅(Digital Fingerprinting) 기술은 구매자의 정보를 인지할 수 없는 방법으로 삽입하는 방법이다. 본 논문에서는 핑거프린팅 기법을 사용하여 웹 서버에 요청이 들어온 모든 웹 페이지에 핑거프린팅 기술의 조건을 충족시키는 방법으로, 정보를 삽입하여 보여 지는 웹 페이지의 변화 없이 사용자의 IP정보를 웹 페이지에 포함시켜 전송할 수 있는 알고리즘을 제시하고 웹페이지 핑거프린팅을 해주는 서버 이용의 장점을 알아보도록 한다.

1. 서 론

최근 네트워크의 발달과 함께 디지털 이미지나 비디오, 음악, 문서 등 디지털 콘텐츠의 불법적인 복제나 재배포로 인한 지적 재산권 문제가 심각하다. 불법 다운로드가 만연한 현재 콘텐츠 사용자들은 이러한 현상을 심각하게 생각하지 않는 경향을 보이지만 콘텐츠 제작자의 창작 의욕을 상실시키고 또한 경제적 손실을 입히므로 불법적인 복제를 막고 저작권을 효과적으로 보호하기 위한 콘텐츠 보호기술이 요구된다[1].

이러한 요구에 의해 등장한 기술인 디지털 워터마킹(Digital Watermarking)은 인간의 의식체계 또는 감지 능력으로는 검출할 수 없도록 저작권자 또는 판매권자의 정보를 멀티미디어 콘텐츠 내에 삽입하여 추후 발생하게 될 지적 재산권 분쟁에서 정당함을 증명하는 데 사용하기 위한 기술이다. 이는 사전적 의미의 보호 시스템인 DRM과 상호 보완적인 수단으로 활용될 수 있다[2].

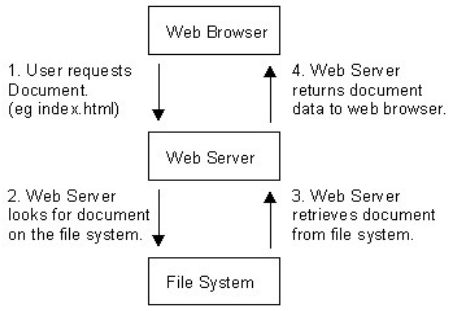
워터마크의 한 분야로 디지털 핑거프린팅(Digital Fingerprinting)이 있는데 이 기법은 기밀 정보를 디지털 콘텐츠에 비가시적으로 삽입하는 측면에서는 디지털 워터마킹과 동일하다고 볼 수 있으나 저작권자나 판매자의 정보가 아닌 콘텐츠를 구매한 사용자의 정보를 삽입함으로써 콘텐츠를 불법으로 배포한 자를 추적(trace traitor)할 수 있도록 한다는 점에서 워터마킹과 차별화된 기술이다. 즉, 디지털 워터마킹을 사용했을 때는 판매되는 모든 콘텐츠에 삽입되는 정보가 그 콘텐츠의 제작자 정보로 동일한 반면, 핑거프린팅을 사용했을 때는 판매되는 콘텐츠를 구매한 사용자들마다 각각 다른 정보를 가지므로 만약 콘텐츠가 불법적으로 재배포 된다면 해당 콘텐츠 내에서 핑거프린팅 정보를 추출하여 어떤 구매자에게 판매된 콘텐츠임을 식별할 수 있게 되어 법적 조치를 가할 수 있게 된다. 이러한 핑거프린팅 기술은 소유권에 대한 인증뿐만 아니라 개인 식별 기능까지 제공해야 하므로 기존의 워터마킹이 갖추어야 할 요구사항인 비가시성, 견고성, 유일성과 더불어 공모 허용, 비대칭성, 익명성, 조건부 추적성 등이 부가적으로 필요하다.

본 논문에서는 이러한 특성을 가진 핑거프린팅 기법을 웹 문서에 적용시키는 기능을 서버에서 추가하여 웹 페이지의 내용 안에 사용자의 IP를 포함함으로써 사용자의 정보를 담을 뿐만 아니라 부가적으로 얻을 수 있는 효과에 대하여 알아보도록 한다.

2. Web Server 동작

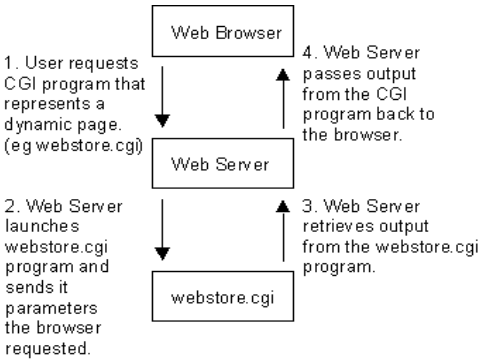
웹 서버는 HTTP를 통해 웹 브라우저에서 사용자가 요청하는 HTML 문서나 오브젝트(이미지 파일 등)를 전송해주는 서비스 프로그램으로 기본적인 레벨에서 웹 브라우저에게 정적인 콘텐츠를 제공한다.

아래의 (그림 1)에서 나타나 있듯이 사용자가 요청한 HTML 파일이 파일 시스템 내의 어딘가에 존재한다면 웹 서버는 이를 읽어 웹 브라우저에 보내준다.



(그림 1) 웹 서버 동작 개념도 (정적 콘텐츠)[3]

이러한 전체 교환 과정은 Hypertext Transfer Protocol (HTTP)을 이용하여 서로 대화하는 웹 브라우저와 웹 서버에서 조정하는 것이다. 이와 조금 다른 경우로 사용자의 입력에 대한 응답으로 직접적, 혹은 간접적으로 생성되는 웹 페이지들을 처리하기 위해 CGI(common gateway interface)를 사용한다.



(그림 2) 웹 서버 동작 개념도 (동적 콘텐츠)[3]

사용자가 정적인 페이지가 아닌 CGI 프로그램에 대한 URL을 클릭하면 웹 서버는 사용자의 요청이 CGI 프로그램에 대한 것임을 URL을 통해 알아내어 CGI 프로그램을 시작하고 동시에 전달받은 요청 메시지도 함께 넘겨준다. 그러면 CGI 프로그램은 요청 메시지에 들어있는 폼 데이터를 이용해서 HTML 페이지를 생성하고 생성한 페이지를 웹 서버에게 넘겨주고 종료한다. 웹 서버는 전달받은 페이지를 그대로 클라이언트에게 전달한다(그림 2).

3. Web Page Fingerprinting을 위한 웹 서버

3.1 핑거프린팅 웹 서버의 동작

기존 웹 서버의 동작과정에 사용자가 서버에 원하는 페이지를 요청할 때 서버에서는 요청 시 전달받은 사용자의 IP 주소를 기억해놓고 file system을 거쳐 읽어오거나 CGI 프로그램을 통해 읽어온 페이지를 클라이언트에게 전달할 때 앞에서 기억해놓은 사용자 IP를 변환하여 첨가하는 기능을 추가해주어 핑거프린팅 된 웹 문서를 사용자에게 전송한다. 이 웹 서버의 동작 시나리오를 살펴보면 다음과 같다.

1. 서버에 자원 요청 – Client
2. 클라이언트의 IP 주소를 추출, 저장
3. 저장된 IP 주소를 이진수로 변환
4. 자원획득 후 클라이언트에게 전달 시 전달할 페이지에 IP 첨부 – Server

3.2 구현

HTML 문서의 소스는 대소문자의 구분이 없고 엔터나 스페이스, 탭은 인정하지 않으며 순차적으로 실행되는 특징을 가지고 있다. 핑거프린팅 기법은 원래의 내용에 영향을 미치지 않으면서 제 3의 정보 즉, 사용자의 정보를 입력해야 하기 때문에 핑거프린팅 된 웹 페이지도 원래 웹 페이지와 다르지 않게끔 하여야한다. 위에 언급한 HTML의 이러한 특징들을 이용하면 이진수로 변환한 사용자의 IP주소를 핑거프린팅하여 여러 가지 방법으로 첨가해 보낼 수 있다. 이를 실행하는 곳은 웹 서버인데 웹 서버의 소스는 보통 C나 C++, JAVA 등의 언어로 이루어져 있기 때문에 대소문자 구분 없음과 공백, 엔터, 탭 무시 기능은 허용되지 않는다.

웹 서버에서 정보를 삽입하는 알고리즘은 일정한 패턴을 가져 삽입하여 핑거프린팅 된 웹 페이지에서 다시 추출이 가능해야 한다. 또한 삽입 정보의 길이도 고려를 해야 하기 때문에 웹 브라우저가 HTML 소스를 읽어 들여 해석할 때 결과가 변하지 않는 방법으로 ①공백을 무시하는 HTML의 특징을 이용, 특정 문자를 기준으로 삼고 좌우에 공백을 주어 코드를 할당하는 방법, ②특정 알파벳만을 대문자로 표현하는 방법, 그 외에도 다양한 응용이 가능하다. ①번은 0 또는 1의 두 가지 숫자로 나타나기 때문에 숫자 뿐 아니라 BCD code를 이용한 문자의 표현이나 이미지 등 다른 다양한 컨텐츠의 표현이 가능하다. ②번과 같은 경우는 문자를 삽입할 때 BCD code 변환 과정이 필요 없고 삽입 공간이 많이 앞의 방법에 비해 많이 요구되지 않기 때문에 짧은 웹 페이지의 내용이 문자를 삽입할 시 효과적이다. 본 논문에서는 사용자 IP를 나타내기 효과적인 ①의 방법을 사용하였다.

이 때 소스코드 중 특정문자로 선택될 수 있는 것은 여러 가지가 있다. 보다 더 효과적인 알고리즘을 위해 사용 빈도수가 높고, 공백을 주어도 내용이 변하지 않는 '='을 선택하여 '=' 기준 공백문자의 유무에 따라 0과 1로 정보 코드를 할당하였다. 서버에서 읽어온 string type의 사용자 IP를 0과 1로 이루어진 코드로 변환하는 방법으로는 BCD코드 변환법과 이진수 변환법이 있다.

이진수 변환	BCD code 변환
(-) class로 나눈 후 이진화 -> 길이 4의 배열 필요 (+) class당 8bit*4class = 32bit의 IP 저장공간 필요	(+) IP를 그대로 변환 -> 1개의 변수 필요 (-) 한 자리 4bit * 12자리 = 48bit의 IP 저장공간 필요

(표 1) 이진코드 변환 법 비교

BCD코드 변환과 이진수 변환은 위 (표 1)에서 알 수 있듯 저장 공간 할당의 측면에서 각각 장단점을 가지고 있다. 본 논문에서는 직접 핑거프린팅되는 변환 후 코드의 길이가 짧은 이진수 변환을 선택하였다.

이진수로 변환한 총 32bit의 코드를 두 자리씩 나누어서 HTML 소스 코드에서 제일 처음 읽어들인 '='부터 순차적으로 삽입하여준다. 공백의 유무에 따른 두 자리의 이진수를 표현하는 코드는 다음에 제시된 (표 2)과 같다.

소스 코드	표현 코드
A=B	00
A= B	01
A =B	10
A = B	11

(표 2) 소스 코드에 따른 표현 코드

다음 코드를 삽입하기 전처리 과정으로 기존의 문서에서 '=' 양쪽의 공백을 제거 후 코드에 따른 공백의 삽입을 해주어야한다. 사용자가 203.209.177.8의 IP로 이 웹 서버에 페이지를 요청하면 웹 서버는 사용자의 IP를 이진수로 변환하여 11001011 11010001 10110001 00001000을 배열에 각각 저장 후 2bit씩 끊어서 11_00_10_11, 11_01_00_01, 10_11_00_01, 00_00_10_00으로 만들어 웹 페이지에 위에서부터 '='을 만날 경우 공백을 삽입하여준다. 이 때 유의할 점은 수식의 '='은 포함시키지 않아야 한다는 것이다. '!='나 '==', '>=', '<='와 같은 경우에는 공백문자를 포함시키는 경우 제대로 동작하지 않으므로 위와 같은 수식을 만날 경우에는 고려하지 않고 넘어가는 알고리즘을 추가하여준다.

```
<meta http-equiv = "content-type" content = "text/html; charset = euc-kr">
                   11                    00                    10
<title>linux.ewha.ac.kr</title>
<meta name = "generator" content = "Namo WebEditor(Trial)">
             11                  11
<script language = "JavaScript">
                   01
<!--
function na_restore_img_src(name, nsdoc)
{
  var img = eval((navigator.appName.indexOf('Netscape',0)!=-1)?nsdoc+'.
          00
  if (name == '')
    return;
  if (img && img.altsrc) {
    img.src = img.altsrc;
              01
    img.altsrc = null;
                 10
  }
}
```

(그림 3) Fingerprinting된 HTML 소스

위 (그림 3)은 Portable Web HTTP Server by Ron Logan version 2.0 미니 웹 서버 위에 올려진 웹 페이지의 소스이다. 앞에 언급한 미니 웹 서버에 핑거프린팅을 위한 기능을 추가하여 구현하였다. 이 웹 서버를 통한 웹 페이지에는 다음과 같은 요청자의 IP가 포함되어있다. 위에서부터 순차적으로 입력이 되고 밑줄 친 '!='과 '==' 의 경우에는 공백문자의 삽입 없이 다음 '='을 찾는다. IP 정보 삽입의 경우 총 32bit가 2bit씩 삽입되므로 16개의 이진코드가 포함된 '='를 통해 사용자는 자신이 요청한 웹 페이지에 자신의 흔적이 찍힌 웹 페이지를 받아볼 수 있다.

3.3 기능 및 장점

핑거프린팅은 단순 사용자의 정보를 삽입하는 것이 아니라 서론에서 언급하였던 비가시성, 견고성, 유일성, 공모 허용, 비대칭성, 익명성 등을 충족시켜주어야 한다.

1) 비가시성 : 웹 페이지에 가시적으로 삽입되어있지 않고 원래의 웹 페이지와 핑거프린팅 된 웹페이지가 웹 브라우저에 보여 지는 화면에도 차이가 없다.
2) 견고성 : Text 형식의 문서이므로 필터링, 압축, 재샘플링 등의 일반적인 신호처리 및 포맷 변환 등을 가한 후에도 삽입정보가 그대로 유지된다.
3) 유일성 : 삽입 정보가 사용자의 IP주소이므로 웹 페

이지를 요청한 사용자를 확정지을 수 있다.

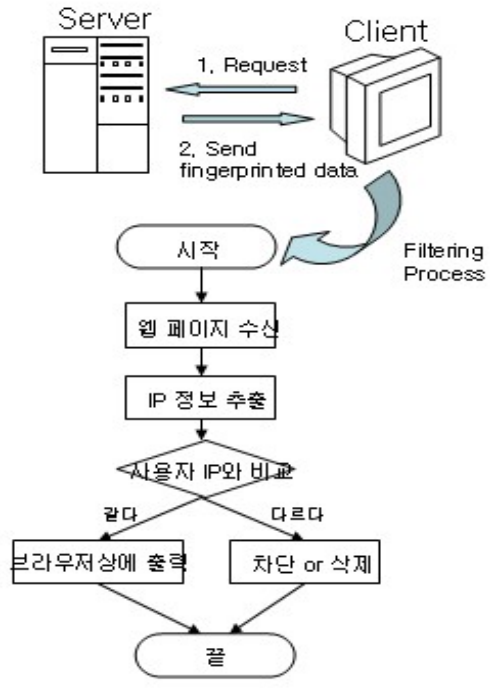
4) 공모 허용 : 핑거프린팅 처리가 되어 웹 브라우저로 보내주므로 웹 서버에서 수정하지 않는 한 사용자가 정보를 수정하거나 삭제할 수 없다.

5) 비대칭성 : 웹 서버가 자원을 획득한 후 브라우저로 사용자에게 전송하기 바로 전 단계에 핑거프린팅 처리를 하므로 사용자만이 자신의 IP가 핑거프린팅 된 문서를 받아볼 수 있다.

6) 익명성 : IP가 삽입된 웹 페이지는 서버에 페이지를 요청한 사용자에게 전달된다.

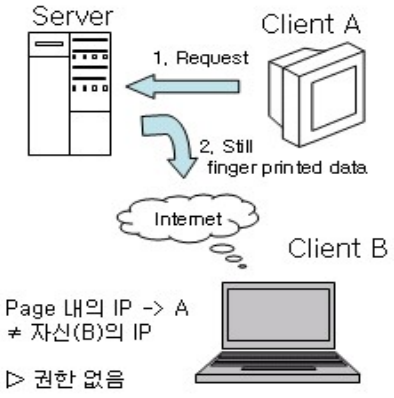
위 서버에서 생성된 IP정보가 핑거프린팅 된 웹 페이지는 본래의 핑거프린팅의 목적인 사용자 추적 등은 물론 좀 더 다양한 방향으로 사용될 수 있다.

핑거 프린팅 된 웹 페이지의 경우 인터넷에서의 스파이웨어나 악성코드 팝업창 같은 경우 웹 브라우저 상에서 필터링 해줄 수 있다. 사용자가 웹 서버에 요청한 페이지에는 사용자의 IP가 들어있기 때문에 웹 브라우저가 웹 서버로부터 전송되어온 문서에서 삽입 정보를 추출한 후 사용자 IP와의 일치여부를 판단하고 요청하지 않은 페이지인 경우 자동으로 필터링을 할 수 있다. (그림 4)에서 정상적으로 요청된 페이지와 사용자가 요청하지 않은, 외부에서 넘어온 페이지의 처리 프로세스를 살펴볼 수 있다. 결국 사용자는 자신이 선택한 웹 페이지만 볼 수 있게 되는 것이다.



(그림 4) 웹 페이지 필터링

또한 웹 서버로 요청메시지를 보낼 때 자신이 받아볼 페이지에 자신의 도장과 같은 기능을 하는 IP를 삽입함으로써 사용자가 요청한 웹 페이지가 중간에 제 3자에 의해 강탈당했을 때 추적 가능하며 사용자가 요청한 페이지는 당사자만 볼 수 있게끔 브라우저상에 설정할 수도 있다(그림 5).



(그림 5) 강탈당한 웹 페이지 처리

4. 결론 및 발전 방향

지금까지 Fingerprinting의 특성을 살펴보고 멀티미디어 콘텐츠의 사용자 추적 위주로 사용되는 핑거프린팅 기법을 웹 문서에 적용시켜 웹 서버를 통하여 웹 문서에 사용자의 IP를 삽입시켜 핑거프린팅을 가능하게 하는 알고리즘과 그로인해 얻을 수 있는 효과들을 살펴보았다.

이러한 idea를 발전시키면 충분히 긴 HTML 문서를 표현하게 되는 다양한 콘텐츠를 가진 웹 페이지를 요청할 경우에는 IP뿐만 아니라 길이가 긴 제 3의 정보를 문서의 코드 안에 삽입시켜서 보낼 수 있다. 길이가 긴 문서의 경우가 아니어도 정보를 삽입하는 패턴은 다양하게 표현될 수 있으므로 위에 언급한 '=' 문자를 기준으로 양옆에 공백을 넣어주는 방법 외에 공백문자의 개수에 따른 코드 할당이 등의 방법을 통하여 표현할 수 있는 bit 수를 늘릴 수 있다.

송신자와 수신자가 같은 패턴 정보를 가지고 있는 Key값을 가지고 있다고 가정한다면 Key값을 가지고 있는 특정 수신자만이 볼 수 있게 설계하여 특정 사용자만이 알아볼 수 있는 문서를 웹 페이지로 포장하여 전송할 수도 있고, 같은 웹 페이지를 요청했어도 사용자의 권한이나 접속 지역(IP 정보로 구분 가능)에 따라 다른 콘텐츠가 웹 브라우저에 나타나게 구현하는 등 웹 서버를 통한 웹 페이지 핑거프린팅은 다양한 응용이 가능하다.

참고문헌

[1] G.C. Langelaar, I. Setyawan, and R.L. Lagendijk, "Watermarking digital image and video data. Astate-of-the-art overview." IEEE Signal Processing Magazine, vol. 17, no. 5, pp. 20-46, Sept. 2000.
[2] I. J. Cox, M. L. Miller, and J. A. Bloom, "Digital Watermarking" Morgan Kaufmann Publishers, 2002.
[3] http://koreainternet.com