# 스팸로봇으로부터 웹컨텐츠 보호에 대한 연구

윤승현*, 김창수**
*부경대학교 전산정보학과
**부경대학교 전자컴퓨터정보통신공학부
e-mail:yoonsh@pknu.ac.kr

# A Study on Protection for Web Contents against a Spam-Robot

Seung-Hyun Yoon*, Chang-Soo Kim**
*Dept of Computer and Information,
Pukyoung National University
**Division of Electronic Computer and Telecommunication Engineering, Pukyoung National University

요 약

인터넷의 빠른 성장과 함께 다양한 웹 컨텐츠들이 사용자들에게 서비스되고 있다. 특히 상업적 목적으로 만든 사이트나 사용자들이 많은 컴뮤니티 사이트 같은 경우, 웹 컨텐츠의 정보 가치가 높기 때문에 스팸로봇에 의해 정보가 유출되는 경우가 발생한다. 이는 대부분의 웹컨텐츠가 HTML문서로 작성되어 있어 스팸로봇과 같은 도구들로부터 보안이 취약하다. 본 연구는 스팸로봇으로부터 웹컨텐츠를 보호하기 위해 이미지 보호를 위한 이미지 경로 은닉화 방법을 제시하고, 텍스트와 전자우편 메일을 보호하기 위한 방법을 제시한다.

## 1. 서론

인터넷을 통해서 이미지, 오디오, 동영상 데이터의 인터넷 전송이 일반화되었다. 그러나 텍스트, 이미지, 동영상 등 디지털 정보는 컴퓨터에 의해서 다량의 복사가 가능하며 복사 후 원본과 동일하게 유지되므로 누구든지 인터넷의 저자 동의 없이 복사, 배포할 수 있는 문제점을 가지고 있다. 따라서 이러한 정보에 대해서 불법 복제나 비합법적인 사용, 부분적인 조작을 통한 자료의 변형 또는 도용 등의 불법적 유통을 막기 위한 방법이 필요하다[1][2].

본 논문의 제한 범위는 스팸 로봇이 웹사이트 컨텐츠에 대한 도용 및 불법 복제를 막는데 한정되어 있다. 만약 사람의 수작업으로 인위적인 작업을 한다고 할 때에 대한 방법은 아니다. 스팸로봇은 HTML 태그에 대해서 일정한 양식으로 출력되는 내용을 정규표현식이나 위치정보를 내용을 추출한다.

본 연구는 서론에 이어, 2장은 문제점과 관련연구에 대해서 논의하며, 3장은 웹페이지 보호에 대해서 설명하고 구현된 방법으로 나타난 결과에 대해서 논의한다. 마지막 4장은 결론 및 향후 연구에 대해서 언급하고자 한다.

## 2. 문제점

### 2.1 스팸메일과 스팸로봇

스팸 로봇에 의해 추출되어, 무작위 사용자에게 보내는 스팸메일은 현재 많은 사용자들과 업계에 엄청난 문제점으로, 2003년 한해 최대의 골칫거리이다. 스팸메일로 인한 폐해는 엄청나다. 소비자가 원하지 않는 광고성 메일이 하루에도 수십 건씩 몰려와 지

우는 데 애를 먹는 것은 물론 불특정 다수를 대상으로 하고 있어 미성년자들에게 유해한 광고도 여과없이 전달된다. 이 과정에서 불법적인 인터넷 아이디(ID) 추출은 예사다. 무엇보다 스팸메일은 인터넷 트래픽의 상당부분을 점유하고 있어 문제다. 인터넷과 e메일의 사용이 늘면서 스팸메일은 올해 전체 메일량의 40%에 달할 전망인데 이는 인터넷 병목현상이 나타나는 일부 국가에서는 단순한 번거로움 차원을 넘어서고 있다. 금전적으로 환산하면 스팸메일의 그늘은 한층 두드러진다. 그리고 스팸메일 발송비용은 고스란히 메일 서버 운영업체로 전가된다. 메일을 발송하는 비용보다 메일 서버의 증설 및 운영에 더욱 많은 비용이 필요하기 때문이다. 여기에다 최근에는 휴대폰 스팸메일도 극성을 부리고 있어 본격화되는 모바일 웹시대, 스팸에 대한 대책은 아무리 강조해도 지나치지 않다는 지적이다[3].

사용자의 이메일이나 아이디 추출은 여러 가지 방법에 의해 유출된다. 이렇게 유출된 이메일로 이메일 광고가 무분별하게 무차별적으로 전송되고 특히 그 광고내용 자체가 불법적인 것들이 점차 증가함에 따라, 이제는 거의 모든 사람들이 스팸메일로 인한 피해를 호소하고 있으며, 이메일마케팅 자체는 사람들로부터 외면을 당하고 있다[6]. 이메일 피로로 인한 마케팅 효과 저하는 이제 이메일마케팅 업체들로 구성된 한국인터넷마케팅협의회에서 옵트인(이용자의 사전 동의를 받았을 경우에만 광고메일을 발송하도록 규정. 현행법은 옵트아웃으로 이용자가 수신거부를 한 이후에만 광고메일 발송을 규제하고 있음) 제도를 주장할 정도로 커져가고 있다. 이메일 유출을 차단하는 프로그램은 사람과 컴퓨터를 구별하여 무분별한 컴퓨터 프로그램(인공지능, AI)에 의한 반복적인 작업을 방지하기 위한 알고리즘이 적용하고 있다.

### 2.2 검색로봇과 스팸로봇의 차이점

웹 검색 사이트도 기본적으로 로봇을 통해 각 웹페이지에서 링크와 텍스트 정보를 추출하여 인덱스시키고 주요 키워드에 대한 정보를 저장한다[4][5]. 그림 1과 같이 간단히 설명되어 질 수 있다. 한때 웹 검색 로봇에 의한 정보 유출도 법적으로 문제가 된 적이 있었으나, 현재는 검색 로봇에 의한 정보 수집과 정보는 예외적인 일이다. 그러나 스팸 로봇은 특정 사이트를 지정한 후 사용자 메일 계정, 이미지, 텍스트 내용 등을 추출하여, 이 정보를 가지고 다시 상업적 웹사이트를 구축하는데 문제가 있다. 이러한 문제는 현재 구인구직, 쇼핑몰, 커뮤니티 같은 사이트에서 발생되고 있다. 예를 들어 어떤 구인구직 사이트에 많은 정보가 있다고 가정할 때, 이 정보들을 체계적으로 스팸 로봇을 이용하여 다른 구인구직 사이트를 만들 수 있다.

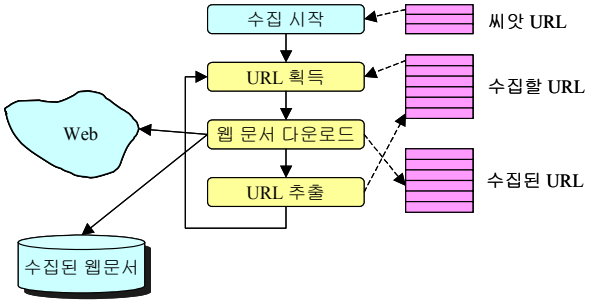


그림 1. 일반적인 검색 로봇의 검색 방법

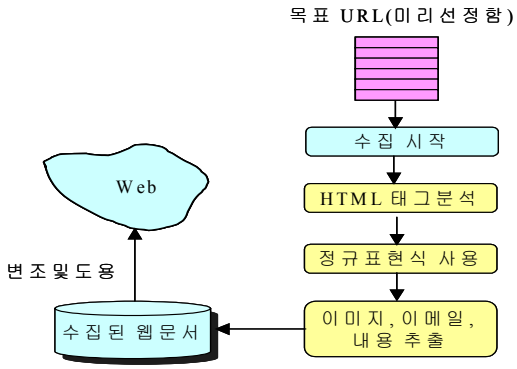


그림 2. 스팸 로봇의 검색 방법

### 3. 웹페이지 보호 구축 방안

웹페이지는 HTML문서로 작성되어 있다. 이러한 문서는 클라이언트의 웹브라우저 엔진에 의해 볼 수가 있다. 그리고 HTML문서로 된 페이지는 정적문서이다. 그래서 동적으로 정보를 제공할 필요성이 있어 서버 사이드 스크립트가 이러한 동적 정보를 제공하기에 적합하다. ASP, PHP, JSP, PYTHON 등 여러 가지 서버용 스크립트 언어가 있다. 그러나 이러한 서버용 스크립트는 서버에서 동작하고 클라이언트에서는 동작하지 않는다. 이와 같은 기본동작 원리 때문에 스팸 로봇과 같은 경우는 해석되어진 페이지에서 정보를 추출한다. 결론적으로 웹페이지가 완전히 노출된 상태로 볼 수 있다.

웹페이지 보호를 위해 암호화하는 방법이 있다. 그러나 웹페이지 전체를 암호화하고 다시 복호화 한다

는 것도 많은 부하가 걸린다. 그리고 검색 사이트의 검색 로봇은 암호화된 웹 컨텐츠 내용을 인식하지 못한다. 검색 엔진 로봇에 인덱싱 되어지지 않는다는 것은 많은 사용자들에게 알려지지 못하는 문제점이 있다. 본 논문에서는 웹에서 보호시킬 부분만 보호시키고, 나머지 기본 페이지는 그대로 사용을 하고 스팸 로봇에 대해서는 정보를 보호하도록 구현하였다.

### 3.1 이미지 보호

웹페이지에서의 이미지는 오른쪽 마우스의 등록정보와 다른 이름 저장이라는 메뉴에 의해 아주 쉽게 조정되어 질 수 있다. 그래서 쇼핑몰이나 이미지 관련 사이트 같은 경우 힘들게 작업한 이미지가 쉽게 스팸 로봇에 의해 추출되어 질 수가 있다. 이런 문제점을 해결하기 위한 방법은 여러 가지가 있으나 본 논문은 더 쉽게 사용하고, 프로그래머가 쉽게 구현할 수 있는 방법으로 이미지 경로를 추출하지 못하게 경로를 변환하는 방법으로 구현하였다.

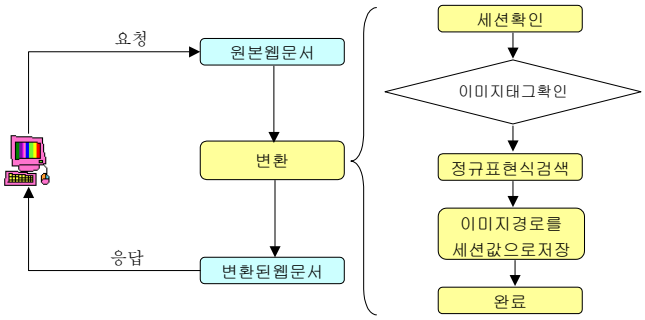


**그림 3. 이미지 경로 은닉화 구성도**

그림 3과 같이 이미지 경로를 변환하여 스팸 로봇에 의해 이미지 경로를 추적하지 못하도록 하였다. 그림 4와 같이 변환알고리즘에 의해 이미지 경로가 기본 디렉토리 경로에서 파일 명 이름으로 변경되었다. 그림 4의 결과에 있는 img.php?id=숫자는 서버 시스템의 메모리에 있는 세션값하고 매칭 시켜서 클라이언트에 결과물이 제공된다. 이 고유 세션값은 서버만이 알기 때문에 스팸로봇은 이미지 경로를 찾을 수 없다. 그리고 php뿐만 아니라 asp, jsp로도 구현이 가능하다.

그리고 별도로 http://호스트명/img.php?id=0을 하여도 이미지는 나타난다. 그러나 일반적으로 웹에서 이미지 상단에서 오른쪽 마우스를 클릭하면 저장에 관련된 메뉴가 나온다. 하지만 본 논문에서 구현된 알고리즘으로 하였을 경우, 오른쪽 마우스를 클릭하여도 이미지 경로가 나오지 않는다.

```
<img SRC="./images/b0.jpg" width="60">
<img SRC="./images/b1.jpg" width="60">
<img SRC="./images/b2.jpg" width="60">
                ↓
<img SRC="img.php?id=0" width="60">
<img SRC="img.php?id=1" width="60">
<img SRC="img.php?id=2" width="60">
```

**그림 4. 변환된 이미지 경로**


**그림 5. 변환된 이미지에서 오른쪽 마우스 사용의 경우**

### 3.2 텍스트와 이메일 보호

현재 스팸 메일 때문에 많은 심각한 문제가 발생되었다. 이러한 문제점은 웹페이지의 텍스트 보호에 대해서 매우 경솔하게 생각했기 때문이다. 현재는 웹페이지에서 메일 보호 프로그램이 있으나, 아직도 홍보 부족으로 사용이 보편화되지 않았으며, 특정 웹 구축 방법에 따라 일부 적용하기 힘든 경우도 있다. 그래서 본 연구는 어떠한 서버 스크립트를 사용해도 기본적으로 웹페이지에서 텍스트와 이메일 계

정의 보호가 가능하도록 하였다.

사용자가 요청한 페이지는 서버에서 변환되어서 클라이언트에 출력된다. 이러한 기본동작에서 스팸메일의 목표페이지는 요청한 페이지에 따른 응답 페이지가 된다. 일반적으로 이런 응답에 대한 웹 페이지는 자바 스크립트가 가장 효율적이다. 그러나 자바 스크립트 또한 클라이언트 웹 페이지에 표출되기 때문에 아주 큰 문제점이 있다.

<script Language="JavaScript" src="/script/javascript/banner.js"></script>

↓

**세션에 따라 동적으로 파일 경로 변환**

① <script Language="JavaScript" src="./1f3870be274f6c49b3e31a0c6728957f/banner.js"></script>

② <script Language="JavaScript" src="./b798abe6e1b1318ee36b0dcb3fb9e4d3/banner.js"></script>

**그림 6. 변환된 파일 경로**

그림 6과 같이 자바스크립의 경로를 변환하여 스팸 로봇이 접근하지 못하게 하였다. 그리고 텍스트와 이메일 관련 인코드, 디코드 함수를 자바스크립트에 정의하였다.

정의된 함수에 의해 페이지 표현은 그림 7과 같이 “한글”, “1000원”, “hong1004@mail.com"과 같이 출력된다. 하지만HTML 소스보기를 하면 아래 그림과 같이 표현된다. 이렇게 함으로써 스팸 로봇에 의해 읽혀지는 값이 출력되기 때문에 안전하게 표현할 수 있다.

**HTML문서 소스**

```
<SCRIPT><SCRIPT>dF("*zI%3A%3AH*zFJ555");</SCRIPT>
<BR>
<SCRIPT>dF("2111%26vD7E11");</SCRIPT><BR>
<SCRIPT>dF("ipoh2115Anbjm/dpn1");</SCRIPT><BR>
```

↓

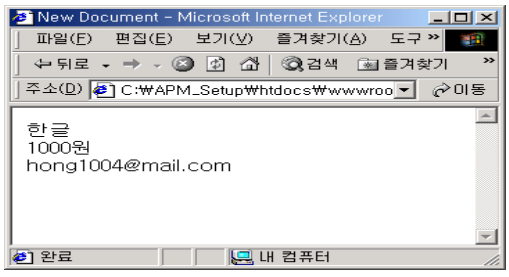


**그림 7. 텍스트의 암호화와 복호화**

## 4. 결론

스팸로봇에 의해 정보가 추출되어 다른 사이트에서 도용 및 변질되어 사용되고 있는 사이트가 많다. 본 논문은 이러한 웹 컨텐츠는 보호를 위해 구현된 방법으로 이미지경로 변환과 텍스트와 이메일 변환을 통해 해결하고자 논의하였다. 첫 번째로 이미지 경로를 추출하지 못하게 경로변환 방법을 구현하였으며, 두 번째는 어떠한 서버 스크립트를 사용해도 기본적으로 웹페이지에서 텍스트와 이메일 계정을 스팸로봇은 암호화된 값으로 출력하기 때문에 안전하게 표현할 수 있으며, 보호가 가능하다. 향후 과제로는 다양한 방법으로 접근하는 스팸 로봇에 대해서, 웹컨텐츠를 보호할 수 있는 추가적인 연구가 필요하다.

## 참고문헌

[1]차주언, 윈도우 웹 서버보안, 대림출판사, 2002

[2]정태명, 서광현, 이동현, 인터넷정보보호, 영진닷컴, 2002

[3]프로그램심의조정위원회, 전자상거래시대의 프로그램 저작권의 보호방안, 2000

[4]W.Niblack,RBarber,Wequitz.M.Fickner.E.Blasman, "The QBIC Project:Querying images by content using color, texture, and shape" SPIE vol 1908, February 1993.

[5] Han.J, Y.Huang, N.Cercone, and Y.Fu, "intelligent Query Answering by Knowledge Discovery Techniques." IEEE Trans. Knowledge and Data Engineering, Vol.8.No 3, 373-390

[6] htt://dci.sppo.go.kr/ 대검찰청인터넷범죄수사센터