저작권 보호를 위한 폐쇄형 OSP 콘텐츠 자동 검색 방법

박경수*, 김원겸*, 김상진*, 유원영**
*(사)한국저작권단체연합회 저작권보호센터
**한국전자통신연구원
e-mail:{kspark, wgkim, sjkim}@cpcmail.or.kr, zero2@etri.re.kr

Automatic Contents Tracking on Closed-type OSP for Copyright Protection

Kyoungsoo Park*, Won-Gyum Kim*, Sangjin Kim*, Won Young Yu**
*Copyright Protection Center
**Contents Research Division, ETRI

요 약

본 논문에서는 디지털 콘텐츠의 저작권 보호를 위해 인터넷상의 웹하드나 P2P와 같은 폐쇄형 OSP(Online Service Provider)에서 불법으로 유통되고 있는 콘텐츠를 자동으로 검색하고 다운로드 하는 방법을 제안한다. OSP의 웹 페이지를 파싱(Parsing)하여 태그(Tag)와 유일한(Unique)한 속성 정보를 이용하여 필요로 하는 콘텐츠의 게시 정보를 얻거나 액션을 취해 웹 페이지를 제어한다. 또한, IE(Internet Explorer) 이벤트(Event) 함수에 포함되는 URL 정보를 이용하여 각 제어 단계의 성공 여부를 판단한다. 윈도우 기반 어플리케이션은 해당 윈도우의 컨트롤ID 및 기타 윈도우 속성 정보를 이용하여 제어한다.

1. 서론

인터넷 환경이 급속도로 발전함에 따라 멀티미디어 데이터의 범람과 사용자들의 유료 콘텐츠 사용에 대한 인식 부족으로 디지털 콘텐츠의 지적재산권 침해가 빈번하게 발생하고 있다. 특히 웹하드와 P2P 같은 특수유형의 OSP를 통한 불법 콘텐츠들의 무분별한 공유는 디지털 콘텐츠 산업의 발전을 저해하는 심각한 문제로 대두되고 있다. 과거의 디지털 콘텐츠 유통 구조는 암호화 기반의 DRM(Digital Rights Management) 시스템 환경에서 콘텐츠의 다운로드 서비스가 이루어졌지만, 상호호환성 측면에서 많은 불편함을 주고 있어 DRM-free 서비스 형태로 변하고 있다. 이러한 DRM-free 환경은 콘텐츠의 불법 확산을 부추기고 있으며, 현재 웹하드나 P2P와 같은 폐쇄형 OSP를 통해 무수히 많은 콘텐츠들이 불법으로 공유, 유통되고 있는 실정이다.[1]

저작권자들로부터 위임받은 저작물들에 대해 모니터링 요원이 직접 OSP들을 모니터링하고 삭제조치하는 방법을 통해 저작권자들의 저작물을 보호해주고 있다. 이 방법은 웹하드나 P2P와 같은 OSP에 접속하여 직접 콘텐츠를 검색하고 다운로드하여 불법복제물임을 확인하면, 해당 OSP에게 삭제요청 메일을 발송 및 회신하여 콘텐츠의 삭제여부를 확인하는 방법이다. 이러한 수동적인 방법은 많은 인력과 시간이 소요되기 때문에 콘텐츠들을 자동으로 검색하고 다운로드해서 불법 여부를 자동으로 식별할 수 있는 기술을 필요로 하게 되었다.

기존의 검색 기법으로는 매크로(Macro) 프로그램의 스크립트를 이용한 자동 검색 방법이 있다.[2] 이 방법의 기본 개념은 수동 검색 시나리오를 레코딩(Recording)하여 스크립트화 한 다음 이 스크립트를 프로그램을 이용해 반복하는 방법이다. 즉 사용자가 로그인, 검색어 입력, 콘텐츠 검색 및 다운로드 등의 일련의 과정을 스크립트 언어를 통해 기록하고, 이를 해석하고 동작 시킬 수 있는 인터프리터(Interpreter)에 의해 반복 실행되는 방식이다. 하지만, 이 방법은 레코딩한 시나리오에 대해서만 처리가 가능하고 그 밖에 여러 가지 예외 상황에 대한 처리가 불가능하다. 또 다른 방법으로는 윈도우 GUI 기반 자동화 스크립트 언어인 오토잇(Autoit)을 이용한 방법이 있는데, 독립적이고 다양한 인터페이스를 가진 OSP들을 제어하는데 근본적으로 문제점을 가지고 있다.[3]

본 논문에서는 다양한 인터페이스를 가진 웹하드나 P2P와 같은 폐쇄형 OSP에서 콘텐츠를 자동으로 검색하고 다운로드 하는 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 콘텐츠 자동 검색 및 다운로드에 대한 기본 기술 설명 및 제어 절차에 대해 설명한다. 다음으로는 각 제어 단계별로 제어 방법을 살펴본 뒤 결론 및 향후 연구 방향을 제시한다.

2. 콘텐츠 자동 검색 및 다운로드 방법

웹으로 구성된 OSP에서 자동으로 콘텐츠를 검색하고 다운로드하기 위해서는 기본적으로 웹 UI를 자동으로 제어할 수 있어야한다. 이를 위해서 웹페이지를 접근하고 분석할 수 있는 기술과 분석된 정보를 이용하여 웹을 제어할 수 있는 기술이 필요하다. 이러한 기술은 일반적으로 IE 같은 웹 브라우저(Web Browser)를 구현하는 기술에 포함되어 있다.[4]

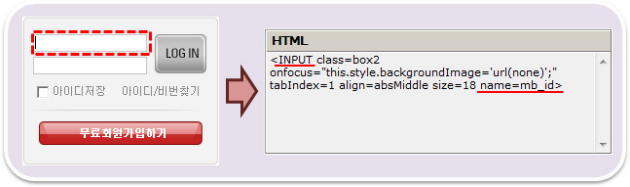
본 논문에서는 웹을 자동으로 제어하기 위해서 MS의 트라이던트(Trident) 기술을 이용한다. 이 기술은 인터넷 익스플로러가 사용하고 있는 레이아웃 엔진의 이름으로 MSHTML로 알려져 있다. 레이아웃 엔진(Layout engine)은 웹 콘텐츠(HTML, XML, 그림파일 등)와 포맷정보(CSS, XML 등)를 파싱해 와서 화면에 해당 콘텐츠를 정리하여 보여 주는 소프트웨어이다. MSHTML은 인터넷 익스플로러의 하부에 위치해서 HTML의 파싱과 렌더링을 담당한다. 이 트라이던트 기술을 이용하면 웹으로 구성된 OSP를 제어하기 위한 태그 분석이나 속성 파악 및 분석된 태그를 통해서 정보를 추출하거나 제어에 필요한 액션을 가할 수 있다.

제안하는 자동 검색 방법은 웹하드나 P2P와 같은 폐쇄형 환경을 대상으로 한다. 폐쇄형 OSP는 회원가입과 콘텐츠를 다운로드하기 위해 사이버 머니 충전이 필수이기 때문에 일반 웹크롤러(WebCrawler)로는 검색 및 다운로드가 불가능하다. OSP에서 콘텐츠를 자동으로 검색, 다운로드하기 위해서는 일반적으로 (그림 1)과 같은 제어 절차로 반복되어 진행된다. 사람이 수동으로 다운로드 받는 절차를 그대로 자동화한 것이다.

(그림 1) OSP 제어 절차

본 논문에서는 OSP 웹 페이지의 사용자 환경(User Interface)을 분석하여 각 제어 단계에서 필요로 하는 태그와 유일한 속성 정보를 추출하고, 제어하고자 하는 액션을 해당 태그에 부여하는 방법과 각 제어 절차가 성공적으로 완료되었음을 판단하기 위해 URL 정보를 이용하는 방법을 제안한다. 우선 웹 페이지를 파싱하여 태그들을 구조적으로 보여주고 분석할 수 있게 해주는 웹 분석툴과 MS 기반의 윈도우를 분석할 수 있는 Spy++와 같은 윈도우 분석툴, 두 개 의 응용 프로그램을 사전에 개발하여 이용하였다.

특정 OSP의 URL 정보를 이용하여 브라우징하고, 선행되어야 하는 것이 사용자 인증 과정이다. 사용자 ID, 비밀번호 입력창, 로그인 버튼에 해당하는 태그와 유일한 속성 정보를 추출하기 위해서 웹 분석툴을 이용하여 추출한다. (그림 2)와 같이 사용자 ID 입력창은 <INPUT> 태그 이름과 name=mb_id라는 유일한 속성을 갖는다.



(그림 2) 사용자 인증 태그 분석

실 제어 프로그램에서는 위에서 추출한 태그 이름과 속성 정보를 이용하여 로그인 관련 태그들을 찾고, 사용자 ID와 비밀번호 입력창에 텍스트를 정보를 입력하고 로그인 버튼에 해당하는 태그에 클릭(Click) 액션을 가하는 방식으로 프로그래밍 된다. 이때 로그인이 성공적으로 되었는가를 판단하기 위해 기존 방법에서는 일정시간을 대기한 후, 성공했다고 가정하고 다음 제어 단계로 진행하는 방식을 사용한다. 하지만, 이 방식은 에러가 발생하였을 경우 다음 제어가 불가능하며 에러 상황을 검출하기도 어렵다는 단점이 있다. 따라서 본 논문에서는 이러한 단점을 보완하기 위해 IE 이벤트 함수 중에 OnDocumentComplete()를 이용하여 각 제어 단계의 완료를 판단하였다. 이 함수는 웹 페이지의 로딩(Loading)이 완료되면 호출되는 함수인데 로딩이 완료된 URL이 인자로 넘어온다. 이 URL이 각 제어 절차를 성공적으로 수행했다는 지시자(Indicator)로 동작하는 것이다. 따라서 로그인이 성공적으로 되었을 때의 URL을 웹 분석툴을 이용하여 사전에 조사한 후, 제어 프로그램에서 IE 이벤트 함수로 해당 URL이 넘어오면 성공, 그렇지 않으면 실패로 판단한다.

사용자 인증을 완료하고 나면 콘텐츠를 검색하거나 자료실로 직접 이동해야 한다. 대부분의 OSP가 금칙어 설정과 같은 기술적인 보호 조치를 해놓고 있어 본 논문에서는 자료실로 직접 이동 하였다. 자료실 이동은 (그림 3)과 같이 사용자 인증과 유사한 방법으로 진행한다. 즉, 영화 자료실로 이동하는 메뉴에 해당하는 <A> 태그와 href 속성 정보를 추출하고, 완료 URL을 사전에 조사한 후, 제어 프로그램에서 해당 영화 메뉴에 해당하는 태그를 찾아 클릭을 하고, URL 정보를 이용하여 자료실 이동의 성공 여부를 판단한다.

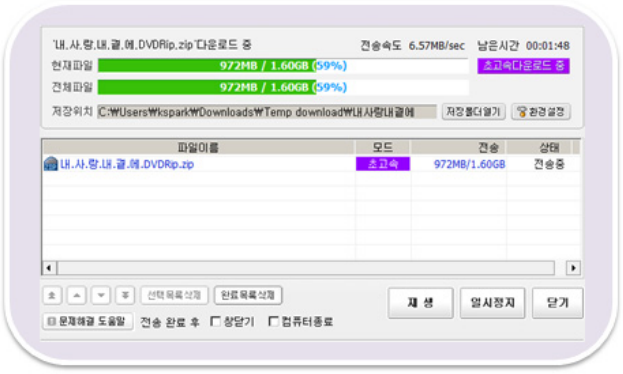
(그림 3) 자료실 이동 태그 분석

자료실로 이동하고 나면 콘텐츠를 획득하기 위해 첫 번째 게시물부터 마지막 게시물까지 클릭하면서 다운로드를 반복 진행한다. 일반적으로 게시물 목록은 대부분 테이블(<TABLE>) 태그 안에 포함되어 있다. 따라서 각각의 게시물을 클릭하기 위해서는 테이블 태그를 반드시 찾아야 하는데, 이 테이블 태그는 다른 제어 절차의 태그와 달리 유일한 속성을 포함하고 있지 않는 경우가 많다. 따라서 상위의 부모 태그들 중에 유일한 속성을 가진 태그를 먼저 찾은 후, 테이블 태그까지 찾아 내려가는 방법을 이용한다. 게시물의 개수는 테이블 태그의 자식 <TR> 태그, 즉 행의 수를 카운트하여 파악한다. 각 게시물의 <TR> 태그의 자식 <TD> 태그에는 게시물의 부가적인 정보들을 담고 있는데, 게시물 번호, 게시물 제목, 파일 크기, 장르, 업로더(Uploader) ID, 가격 정보 등을 추출하여 삭제 요청 메일에 증거 목록으로 이용하게 된다.

게시물 목록을 클릭하면, 일반적으로 (그림 4)와 같이 콘텐츠 등록 정보가 있는 팝업 웹 페이지가 출력되는데, 위와 같은 방법으로 초고속다운 버튼을 찾아 클릭해서 실제 다운로드를 진행한다.

(그림 4) 콘텐츠 등록 정보 페이지

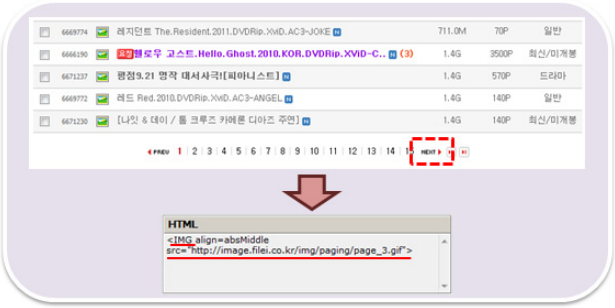
대부분의 OSP에서 다운로드는 (그림 5)와 같은 자체 제공하는 윈도우 기반의 클라이언트(Client) 프로그램, 즉 다운로드 윈도우를 통하여 이루어진다. 다운로드가 진행되는 동안 메인 웹 페이지, 콘텐츠 등록정보 페이지, 표준시계, 다운로드 윈도우가 모두 보이는 화면을 캡처(Capture), 저장해서 삭제 요청 메일에 증거 자료로 활용한다.



(그림 5) 다운로드 윈도우

다운로드 윈도우의 분석은 Spy++와 유사한 형태로 개발된 윈도우 분석툴을 이용한다. 해당 윈도우 컨트롤(Control)의 클래스 이름, 컨트롤 ID, 위치 정보 등을 이용해서 핸들(Handle) 값을 얻어 텍스트 정보를 얻거나 클릭 액션을 취한다. 예를 들어, 다운로드 윈도우가 실행되면 일정 시간동안 클래스, 캡션(Caption) 이름 정보를 이용하여 다운로드 윈도우를 찾고, 컨트롤 ID나 위치 정보를 이용하여 우측 하단에 위치한 다운로드 버튼을 찾아 클릭하여 다운로드를 시작한다. 파일 다운로드 경로는 저장위치 에디트 컨트롤을 찾아 텍스트 정보를 추출한다. 그리고 리스트 컨트롤을 찾아 파일이름 컬럼에서 다운로드 받는 파일 이름을 얻고, 상태 컬럼의 다운로드 상태를 일정 시간 주기로 검사하여 텍스트가 완료로 바뀌면 다운로드 완료로 판단한다. 영상 파일인 경우 다운로드 시간이 오래 걸리기 때문에 사용자가 정의한 파일 크기만큼 다운로드 하는 부분 다운로드 기능도 수행한다. 또한, 다운로드한 파일이 압축 파일인 경우 압축 파일임을 인지하고 자동으로 압축을 해제한다. 다운로드가 완료되면 다운로드 윈도우를 종료하고 다음 콘텐츠 목록을 클릭하여 위 다운로드 과정을 반복 진행한다.

이렇게 동일 페이지 내의 마지막 게시물까지 다운로드를 완료하면 다음 페이지로 이동한다. 페이지 이동은 (그림 6)과 같이 게시물 목록 페이지의 하단에 위치한 페이지 이동 버튼 또는 페이지 숫자에 해당하는 태그를 직접 찾아서 클릭을 한다.



(그림 6) 페이지 이동

페이지 이동을 하고 나면 해당 페이지의 첫 번째 게시물부터 다시 위 다운로드 하는 과정을 반복 진행하면서 마지막 페이지까지 모든 콘텐츠들에 대해서 다운로드한다. 이렇게 다운로드한 각 콘텐츠들은 콘텐츠의 게시 정보, 캡처 화면, 파일 정보 등이 데이터베이스에 저장되어 삭제 요청 자료로 활용하게 된다.

3. 구현

제안한 자동 검색 프로그램은 MS 윈도우즈 환경에서 음원, 영상, 어문 등 3개 콘텐츠를 대상으로 모니터링할 수 있도록 구현되었다. 구현된 검색 프로그램은 HP-E 시리즈 서버에 탑재되어 서버당, 5개 OSP를 대상으로 모니터링하고 있다.

<표 1> 콘텐츠 유형별 다운로드 양(점)

콘텐츠 / 기간	음원	영상	어문
3일	5,548	1,707	181,565

<표 1>은 콘텐츠 유형별 평균 다운로드 점수를 나타낸 것이다. 테스트는 10개 웹하드를 대상으로 3일 동안 OSP 검색 프로그램을 운영하여 다운로드된 콘텐츠 점수를 조사하였다. 음원의 크기는 MP3기준 4~6MB를 전체 다운로드하였고, 영상은 다운로드하는데 시간이 많이 소요되기 때문에 해당파일 크기 기준으로 15%의 부분 다운로드를 하였다. 압축 파일인 경우에는 압축 해제를 해야 하기 때문에 전체 다운로드를 하였다. 어문은 스캔된 이미지나 텍스트 파일로 되어 있고, 음원이나 영상에 비해 파일 크기가 작기 때문에 많은 점수를 기록하였다. 테스트 기간 동안 OSP의 UI가 변경되거나 여러 가지 예외 상황 등이 발생하여 유지 보수하는 시간을 포함하였다.

4. 결론

본 논문에서 디지털 콘텐츠의 저작권 보호를 위해 웹하드와 P2P와 같은 폐쇄형 OSP를 대상으로 콘텐츠를 자동으로 검색하고 다운로드하는 방법을 제안하였다. 제안한 방법은 OSP에서 에러 없이 콘텐츠를 자동으로 검색하고 다운로드할 수 있는 장점을 가지고 있다. OSP의 UI가 변경되더라도 스크립트 기반의 기존 방법은 변경된 UI에 기반 하여 새롭게 다시 프로그래밍 해야 하지만, 제안한 방법은 태그와 URL 정보만 같다면 다시 변경할 필요가 전혀 없다. 일반적으로 위치 변환 등의 변경은 제안한 방법에 영향을 주지 않는다. 예를 들어, 로그인을 하기 위한 폼의 위치가 변경되더라도 위치에 상관없이 로그인이 가능하다.

사용자의 수동 검색을 대치하는 제안한 콘텐츠 자동 검색 기술은 OSP의 잦은 UI 변경에 따른 적응성을 가지고 있다. 잦은 UI 변경에도 제어 프로그램을 새롭게 작성해야하는 기존 기술에 비해서 본 제안 방법은 어느 정도 UI 변경에 적응력을 가지므로 콘텐츠 검색 성능이 기존 방법보다 우수하다고 말할 수 있다. 나아가서 콘텐츠의 불법 유통이 가장 심각한 폐쇄형 OSP를 대상으로 자동으로 콘텐츠를 검색한다는 점에서 불법 유통 근절의 효과가 크다고 할 수 있다.

제안한 콘텐츠 자동 검색 기술은 웹하드와 P2P에서 무분별하게 유통되는 콘텐츠를 실시간으로 모니터링하고 있으며, 콘텐츠의 불법 여부를 판단하는 인식기술과 연계하여 자동으로 삭제조치하는 불법저작물 추적관리 시스템의 한 부분으로 운영되고 있다.

참고문헌

[1] 정혜원, 이준석, 서영호, “불법콘텐츠 추적 기술 연구 동향,” 전자통신동향분석 제20권 제4호, pp.120-128, 2005
[2] Macro Express, http://www.macros.com/
[3] Autoit, http://www.autoitscript.com
[4] Microsoft, http://msdn.microsoft.com