

논문 2004-41SP-5-11

핑거프린팅을 이용한 이미지 콘텐츠 추적시스템 설계 및 구현

(Design and Implementation of Image Content Tracking System using Fingerprinting)

서 영 호*, 김 원 겸*, 이 선 화*, 황 치 정**

(Young Ho Suh, Won-gyum Kim, Seon Hwa Lee, and Chi Jung Hwang)

요 약

본 논문에서는 웹 로봇과 핑거프린팅 기법을 이용하여 인터넷에 배포되어 있는 이미지 콘텐츠를 자동으로 수집하고 원 구매자를 식별하는 기법을 소개한다. 핑거프린팅 기법은 워터마킹 기법의 확장 기술로 구매자의 정보를 콘텐츠에 삽입/추출하는 저작권보호 기법의 하나이다. 콘텐츠 내에 삽입된 핑거프린트는 콘텐츠를 식별하는 식별자의 역할을 할 뿐만 아니라 원 구매자의 정보도 가지고 있어서 콘텐츠의 저작권 보호에도 활용된다. 본 논문에서는 이미지 검색 엔진과 핑거프린팅 기법을 혼합한 능동적 콘텐츠 추적 시스템을 제안한다. 제안된 시스템은 콘텐츠가 배포되기 전에 원 구매자의 정보를 핑거프린트로 삽입하고 구매기록을 저장한다. 배포된 콘텐츠는 웹 로봇에 의해 자동으로 수집되고 동시에 핑거프린트 정보가 추출되어 서버에 저장된다. 콘텐츠의 식별은 추출된 핑거프린팅 정보와 배포 시점에 저장된 판매기록을 비교하여 이루어진다.

Abstract

This paper proposes a new method for tracking image contents, which are re-distributed on the Internet, using a web-robot and fingerprinting scheme. The fingerprinting is an extended technique of watermarking and embeds identity of the customer into the content as a fingerprint in a way that is very difficult to erase. The fingerprint in the content can be used widely in the field of the copyright protection technique because it identifies the content itself and includes the information about the original customer. In this paper we discuss an active image tracking system using a web-robot and fingerprinting scheme. In the proposed system the information of the original customer is embedded as a fingerprint into the image content before distribution and also stored into DB. The distributed contents are collected by the web-robot and fingerprinting information is extracted from the contents. Finally the identifying of contents is done by comparing the extracted fingerprint with the selling history of the content.

Keywords : Copyright protection, Fingerprinting, Watermarking, Content tracking

I. 서 론

최근 디지털 기술과 인터넷 환경의 급속한 발전으로 인해 디지털 콘텐츠의 제작과 판매가 활발해지고 있다. 그러나 손실 없이 대량 복제가 가능한 디지털 콘텐츠의 특성과 사용자들의 유료 콘텐츠 사용에 대한 인식 부족으로 디지털 콘텐츠의 지적재산권 침해가 빈번히 발생

하여, 콘텐츠 산업 발전을 저해하는 심각한 문제로 대두되고 있다.

디지털 콘텐츠에 대한 저작권을 보호하기 위한 주요 기술로는 DRM(Digital Rights Management), 워터마킹(watermarking), 핑거프린팅(fingerprinting) 기술 등이 있다. 일반적인 DRM 시스템은 사용자가 콘텐츠를 요청하면 암호화된 콘텐츠를 보내주고, 실행 시 지불 시스템과 연결하여 지불이 완료되면 암호를 풀 수 있는 키를 다시 보내주어 사용자가 콘텐츠를 사용할 수 있게 하는 구조로 되어 있다. 암호화에 근간을 두고 있는 DRM은 콘텐츠에 대하여 지불을 하지 않은 일반 사용자들의 접근자체를 봉쇄하지만, 암호화가 풀린 이후의

* 정회원, 한국전자통신연구원 디지털콘텐츠연구단
(Digital Content Research Division, ETRI)

** 정회원, 충남대학교 정보통신공학부
(Chungnam National University)

접수일자: 2003년7월19일, 수정완료일: 2004년9월15일

상황에서는 DRM기술만으로는 콘텐츠의 저작권보호가 사실상 어렵게 된다. 실제로 음악파일이나 동영상 같은 멀티미디어 콘텐츠의 경우 복호화 된 상태나 혹은 콘텐츠의 재생 시 다시 캡처(capture)되어 암호화되지 않고 배포되는 경우가 대부분이다.

DRM이 사전적 의미의 보호 시스템이라면 워터마킹이나 핑거프린팅 기법은 사후적 의미의 저작권 보호 기법이라고 할 수 있다. 즉 이미지나 음악 파일 같은 멀티미디어 콘텐츠에 인간의 시각이나 청각으로는 식별이 어렵도록 저작권정보를 삽입하여 배포하고 저작권 분쟁이 발생하였을 경우 이를 추출하여 저작권을 보호한다. 이때 삽입되는 저작권정보가 워터마크이다.

핑거프린팅 기법은 모든 콘텐츠에 동일한 저작권 정보를 삽입하는 워터마킹 기법과는 달리 콘텐츠마다 각기 다른 구매자정보를 삽입함으로써 불법복제 및 유통행위가 발견되었을 때 불법 배포자를 추적하고자 하는 기술로, 저작권 정보만을 이용하는 워터마킹보다 보다 적극적인 의미의 보호 기법이라 할 수 있으며 부정자 추적(traitor tracing) 기술로도 논의될 수 있다.

본 논문에서는 인터넷상에서 다량으로 배포된 이미지 콘텐츠를 이미지 검색 로봇을 이용하여 수집하고 이 콘텐츠로부터 원 배포자를 추적하기 위한 핑거프린팅 알고리즘과 콘텐츠 추적시스템을 제안한다. 본 논문의 구성은 다음과 같다. II장에서는 불법 복제 추적 기법과 관련된 연구들을 살펴보고, III장에서는 공모공격에 강한 핑거프린팅 기법을 제안한다. IV장에서는 본 논문에서 제안한 콘텐츠 추적 시스템에 대하여 설명하고, V장에서는 인터넷의 공개사이트에 대한 실험결과와 저작권보호에의 활용방안을 고찰한다. 마지막으로 VI장에서 결론을 맺는다.

II. 관련연구

원 배포자를 추적하기 위한 핑거프린팅 기술은 워터마킹 기술의 확장으로 동일한 소유권정보를 모든 콘텐츠에 삽입하는 워터마킹과는 달리 콘텐츠의 판매 시 구매자의 정보를 워터마크로 삽입하며 이를 핑거프린트라 한다. 그러나 같은 콘텐츠라도 핑거프린팅 후의 콘텐츠가 조금씩 다르다라는 특성 때문에 구매자들이 서로 그 차이를 공모하여 삽입된 핑거프린트를 지우려고 하는 공모공격(collusion attack)이 존재하게 한다. 이런 이유로 일반적인 워터마킹 알고리즘이 핑거프린팅에 그대로 적용되기에는 무리가 있다. 핑거프린팅 기술은 이 공모

공격에 강인하도록 개발되어야 한다.

Boneh와 Shaw^[3]는 삽입가정(Marking Assumption)을 바탕으로 공모공격에 강한 핑거프린팅 코드를 제안하였다. 마크는 다른 삽입위치를 갖는 제한된 수의 코드조합이고 핑거프린팅 코드는 이 마크의 집합이며, 같은 위치에 모두 같은 마크가 삽입되지 않았을 경우, 그 마크는 검출가능(detectable)하다라는 것이 삽입가정이다.

공모에 강한 핑거프린팅 코드를 이미지에 적용한 연구는 Dittmann^[2]에 의해 제안되었다. Dittmann은 3명의 배포자중 2명이 서로 공모하였을 때 공모자를 추출할 수 있는 핑거프린팅 코드를 제안하고 이를 이미지의 DCT계수에 적용하였다. 이 방법에서는 공모자의 수가 d명일때 모든 공모자를 색출할 수 있는 유한사영기하학에 기반을 둔 d-검출코드(d-detecting code)를 사용하였다.

Sebe와 Domingo-Ferrer는 이중이진해밍코드(dual binary Hamming code)를 이용하여 공모에 강한 핑거프린팅 코드를 설계하고 이를 정지영상에 적용한 알고리즘을 제안하였고^{[4][5]}, Trappe와 Min Wu는 멀티미디어 데이터를 위한 비공모코드 (ACC:Anti-Collusion Code)를 이용하여 공모자를 색출할 수 있는 알고리즘을 제안하였다^[6]. Wang은 정지영상을 웨이블릿(wavelet) 분해한 후 각기 다르게 선택된 부밴드(subband)에 정보를 삽입하는 핑거프린팅 알고리즘을 제안하였다^[7].

이상의 방법들은 2~3명 이내의 사용자들이 공모하여 만들어 낸 영상으로부터 최소 한 명 이상의 공모자를 추출하기 위한 방법들이다. 단점으로 구매자가 많을수록 필요로 하는 코드의 길이가 기하급수적으로 증가하기 때문에 구매자가 많은 인터넷 환경에서 제한된 크기의 콘텐츠에 적용하기에는 어려움이 있다. 또한 코드의 형태도 단순하여 공격자에 의해 쉽게 예측이 가능한 점도 큰 단점이다.

동영상을 대상으로 한 기법으로 Kirovski^[8]등이 개발한 미디어플레이어 기반의 워터마크/핑거프린팅 시스템이 있다. 이 시스템은 구매자에게 서로 다른 검출키를 배포하는 방식으로 2시간의 고화질 동영상에 대하여 900,000명의 사용자까지 검출키에 대하여 공모강인성을 나타냄을 보였다.

콘텐츠의 저작권보호를 위한 추적시스템에 대한 연구로는 디지털마크(Digimarc)사의 마크스파이더(MarcSpider)^[1]가 있다. 마크스파이더는 검색된 이미지들 중 워터마크가 삽입되어 있는 이미지만을 추출하여 결과로

보여줌으로써 저작권자가 자신의 이미지가 어디에서 어떻게 사용되고 있는지 알 수 있도록 해준다. 마크스파이더는 수백 개의 스파이더가 각각 개별적으로 동작하며, 한달 평균 약 5천만건의 워터마킹 된 이미지를 검색해내는 성능을 보이고 있다. 또한 한달에 한번 데이터베이스에 있는 페이지를 주기적으로 방문하여 해당 페이지의 업데이트 여부를 확인하며 페이지 방문주기는 정책에 따라 조정한다. 그러나 마크스파이더가 본 논문에서 제안하는 콘텐츠 추적 시스템과 다른 점은 추출되는 정보가 소유자 정보만을 나타내는 단순한 워터마크라는 것이다.

III. 핑거프린팅 알고리즘

본 장에서는 제안하는 콘텐츠 추적 시스템이 포함하고 있는 핑거프린팅 알고리즘에 대해 간단히 설명한다. 핑거프린팅 기법은 워터마킹 기법과 같이 삽입과 추출과정으로 구분할 수 있다. 구매자 정보 같은 다양한 삽입정보를 수용하기 위하여 삽입되는 비트의 수가 64비트 이상이 되도록 다중비트 핑거프린팅(Multi-bits finger-printing) 기법으로 설계하였다.

또한 평균화공격(averaging attack) 같은 공모공격에 강인하도록 하기 위하여 충분히 긴 무작위순열(random sequence)을 사용하였다. 무작위순열은 그 길이에 비례하여 공모공격에 대해 어느 정도 강인성을 갖는다는 것은 이미 알려져 있다^[9]. 본 논문은 무작위순열의 이러한 성질을 이용하여 공모공격에 어느 정도 강인한 핑거프린팅 알고리즘을 제안한다.

3.1 핑거프린트 삽입

삽입할 핑거프린트의 구조는 크게 비밀키(Secret Key), 핑거프린팅 정보, 주기적덧검사(CRC), 3가지로 구성된다. 비밀키는 핑거프린팅 데이터를 표현하기 위해 발생하는 무작위순열에 대한 초기값(seed)을 나타내며, 핑거프린팅 정보는 콘텐츠를 구매하는 실 실구매자의 정보나 거래정보를 나타내는 부분이며 64비트 이상으로 구성된다. 마지막으로 비밀키와 핑거프린팅 정보의 16비트 CRC를 추가적으로 삽입한다. 주기적덧검사는 핑거프린트의 추출정확도를 높이기 위해 추출과정에서 사용된다. 삽입될 핑거프린팅 구조는 그림 1과 같다. 총 88비트의 정보가 핑거프린트로 삽입된다.

기존의 핑거프린팅 방식과는 달리 무작위순열을 이용해 의미 있는 아스키 값을 표현하기 위하여 본 논문

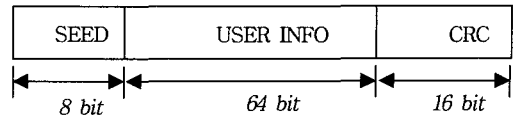


그림 1. 핑거프린팅 정보 구조
Fig. 1. Fingerprinting data structure.

에서는 Joseph이 사용한 DS-CDMA방법^[10]을 사용한다. 삽입할 핑거프린팅 정보 즉, 심볼(symbol), N 개를 $S=\{S_1, S_2, \dots, S_n\}$ 라 하고 각각의 심볼에 해당하는 무작위순열을 $r=\{r_1, r_2, \dots, r_m\}$ 이라 할 때 삽입할 핑거프린팅 신호, F 는 식(1)과 같이 표현된다.

$$F = \sum_{i=1}^N SHIFT_{t_i}(r_i) \quad (1)$$

는 각 심볼에 대한 쉬프트(shift) 파라미터이고 r_i 은 삽입할 콘텐츠의 크기에 따라 일정한 길이를 갖는 무작위순열이다. 예를 들어 삽입할 심볼 S 가 'Peterson'이라 하면 N 은 8이며 콘텐츠의 크기에 비례하는 길이의 무작위순열, r_i 이 각각의 심볼에 따라 8개가 생성된다. 각각의 순열들에 대해 심볼의 아스키 값에 비례하여 쉬프트파라미터가 결정되고 각각의 무작위순열들은 이 값에 따라 쉬프트 된 후 하나의 순열로 더해진다. 이 더해진 최종의 순열값이 삽입할 핑거프린팅 정보, F 이다. 따라서 F 는 -8과 +8사이의 값을 갖는 무작위 순열이다.

시프트파라미터, t_i 는 추출과정에서 비밀키에 의해 생성된 무작위순열의 상관계수 (cross-correlation)를 계산했을 때 가장 일치하는 지점, 즉 피크(peak)점의 위치를 나타낸다. 이 위치정보로 아스키 값을 표현하게 된다.

본 논문에서는 핑거프린팅 정보 F 를 정지영상의 공간영역에 삽입/추출하였다. 삽입정보는 영숫자가 조합된 아스키 8자의 구매자ID를 사용하였고 비밀키와 CRC 정보를 포함하여 총 11개의 무작위순열이 사용되었다. 위에서 설명한 바와 같이 8비트의 정보를 삽입하기 위하여 하나의 무작위 순열을 사용하였다. 비밀키 정보를 숨기기 위한 무작위순열은 하나의 전역비밀키(global secret key)에 의해 생성되었다.

잘림(cropping)같은 정지영상에 대한 공격법에 강인하게 하기 위하여 정지 영상을 단위블럭으로 나누어 삽입한다. 단위블럭의 크기는 핑거프린팅 정보 F 의 길이를 결정지으며 또한 공모공격에의 강인성을 결정하는 요소이기 때문에 적절히 선택되어야 한다. 이미지를 단위블럭으로 나누어 여러 번 삽입하게 되면 두 가지의 장점을 얻을 수 있다. 하나는 반복삽입에 의한 강인성

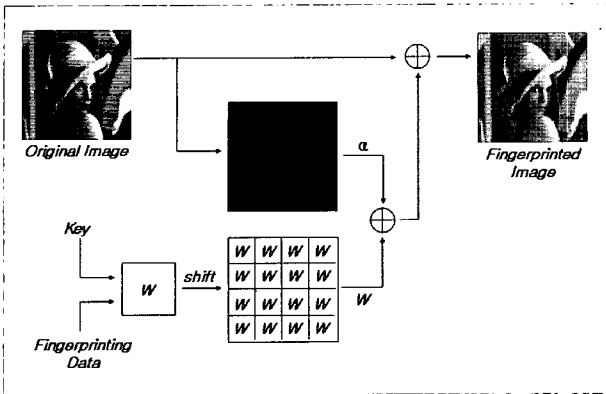


그림 2. 핑거프린팅 정보 삽입 과정
Fig. 2. Embedding process of fingerprinting.

이고 다른 하나는 반복적으로 삽입된 핑거프린팅 정보가 RST공격에 대비한 템플릿(template)으로 작용할 수 있다는 점이다. 그림 2는 본 논문에서 제안하는 핑거프린팅 기법의 삽입과정을 보인 것이다

핑거프린팅 정보의 삽입 시 삽입강도를 이미지에 적응적(adaptive)으로 삽입하기 위해 마스킹(masking) 함수를 사용한다. 그림 2의 중간에 본 논문에서 사용한 마스킹 함수의 결과를 보여주고 있다. 삽입과정을 간단히 하기 위하여 본 실험에서는 3x3크기의 서브윈도우안의 픽셀값에 대한 분산을 마스킹 함수로 이용하였다. 마스킹 함수에 의해 에지(edge)부분에는 강하게, 플랫폼(flat) 부분에는 약하게 삽입되는 효과를 얻었다.

단위화 된 핑거프린팅 정보는 이미지의 크기에 따라 여러 번 반복되어 삽입 되어진다. 그림 2의 경우 512x512 이미지에 128x128의 단위블럭을 사용하여 16번 핑거프린팅 정보를 반복하여 삽입하였다.

3.2 핑거프린팅 정보 추출

핑거프린팅 정보 추출과정은 삽입과정의 역순서로 진행된다. 먼저 핑거프린팅 된 이미지에서 핑거프린팅 정보를 예측한다. 핑거프린팅 정보는 워터마킹과 같이 가우시안 분포를 따르는 노이즈로 간주되기 때문에 필터링에 의해 예측할 수 있다. 본 논문에서는 고대역통과필터(high-pass filter)를 사용하였다.

다음으로 전역비밀키에 의해 무작위순열을 생성한다. 이는 비밀키를 추출하기 위한 것이다. 생성된 무작위순열과 예측된 핑거프린팅 계수들간의 상관계수(cross correlation)를 식 (2)와 같이 구한다

$$Corr_{i,j}(n) = \frac{1}{N^*N} \sum_i \sum_j HPF(X'_{[i,j]}) * W_{[i+n,j+n]} \quad (2)$$

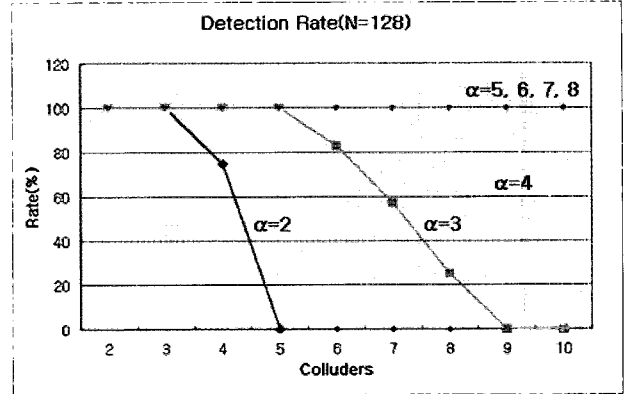


그림 3. 평균화 공모 공격에 대한 검출율
Fig. 3. Detection rate for averaging attack.

X'는 필터를 사용하여 예측한 핑거프린트 정보이며 W는 키에 의해 생성된 무작위순열이다. 피크의 위치값이 핑거프린팅 정보를 생성한 실 비밀키이다. 추출된 비밀키로 다시 10개의 무작위순열을 생성한 후 같은 방법으로 상관계수값을 계산한다. 이중 8개는 구매자정보를 위한 것이고 2개는 주기적검사를 위한 것이다. 각각의 피크의 위치는 다시 아스키값으로 변환된다. 마지막으로 추출된 주기적검사항과, 비밀키와 구매자정보로부터 구해진 주기적검사항을 서로 비교하여 추출된 정보를 검증한다.

핑거프린팅 정보가 단위블럭에 반복적으로 삽입되어 있기 때문에 추출과정은 주기적검사항이 일치하는 구매자정보가 나올 때까지 반복한다

3.3 성능분석

제안된 핑거프린팅 알고리즘은 512x512 크기를 갖는 Lena영상의 공간영역에 실험되었다. 본 장에서는 핑거프린팅 알고리즘에 있어서 최우선으로 고려해야 되는 공모공격에 대한 강인성만을 보인다. RST나 필터(filter)처리 같은 기타 공격에 대한 강인성 확보는 향후 연구방향으로 언급된다. 실험을 간략히 하기 위하여 삽입되는 핑거프린팅 정보는 'Perterson', 'wgkim123' 같이 영숫자가 조합된 8자의 구매자 ID를 사용하였다

삽입강도(α)는 2부터 8까지 6단계로 실험하였고 5일 경우 PSNR이 40dB가 되도록 조정하였다. 그림 2와 같이 단위블럭의 크기는 128x128이고 테스트영상에 16번 반복하여 삽입되었다. 그림 3의 그래프는 10명까지의 공모공격 후 검출률을 보인 것이다. 삽입강도가 2일 경우 공모자가(PSNR=38dB) 5명 이상일 때에는 핑거프린팅 정보를 검출할 수 없었다. 반면 삽입강도가 5 이상일 경우 공모자가 10명 일 때까지 핑거프린팅 정보를

모두 검출하였다

IV. 콘텐츠 추적 시스템

본 장에서는 III장에서 제안한 핑거프린팅 알고리즘을 포함하는 콘텐츠 추적 시스템에 대해 고찰한다. 제안한 콘텐츠 추적 시스템은 크게 다음 3가지로 구성된다. 콘텐츠의 원구매자를 증명하기 위하여 삽입되는 핑거프린팅 데이터 구조와 핑거프린트를 삽입/추출하는 서버, 콘텐츠의 거래정보를 관리하고 능동적으로 콘텐츠를 검색하기 위한 웹 검색 로봇을 포함하는 콘텐츠 웹 서버가 3가지 구성요소이다.

4.1 핑거프린팅 데이터 구조

첫 번째로 원 구매자를 추적하기 위한 핑거프린팅 데이터구조는 다음과 같다

1. 서버정보
 - A. 16bit로 구성
 - B. 총 65000여 개의 서버 관리 가능
 - C. 카테고리별로 서버번호 할당관리 (eg. 지역, 특성)
2. 거래정보
 - A. 32bit로 구성
 - B. 서버당 43억여 개의 거래정보 관리 가능
 - C. 거래정보를 이용하여 실질적인 원 구매자를 추적함
3. 미래추가정보
 - A. 16bit로 구성
 - B. 추가정보를 위한 여유공간
 - C. SDMI와 DVDCCA와 같은 표준에서의 권고 사항
4. 에러체크정보
 - A. 16bit로 구성
 - B. CRC 사용
 - C. 추출된 정보의 신뢰성 보장

거래 정보는 원 구매자를 추적하는데 있어 가장 중요한 정보로, 거래번호로 관리되며 중복되지 않도록 관리되어야 한다. 거래정보DB는 거래일자, 사용자명, 거래내용, 사용자정보를 저장한다.

그림 4는 제안한 핑거프린팅의 데이터 구조가 어떻게 활용되는지는 보여주고 있다. 콘텐츠 판매 업체에 의해서 판매된 콘텐츠는 임의의 사용자에게 의해서 복제되어 불법 서버에서 사용된다. 뒤에서 설명될 검색 웹 로봇은 인터넷상의 콘텐츠를 수집하여 핑거프린팅 정보를 추출하고 이를 분석하여 원구매자를 추적한다. 그 과정은 다음과 같다.

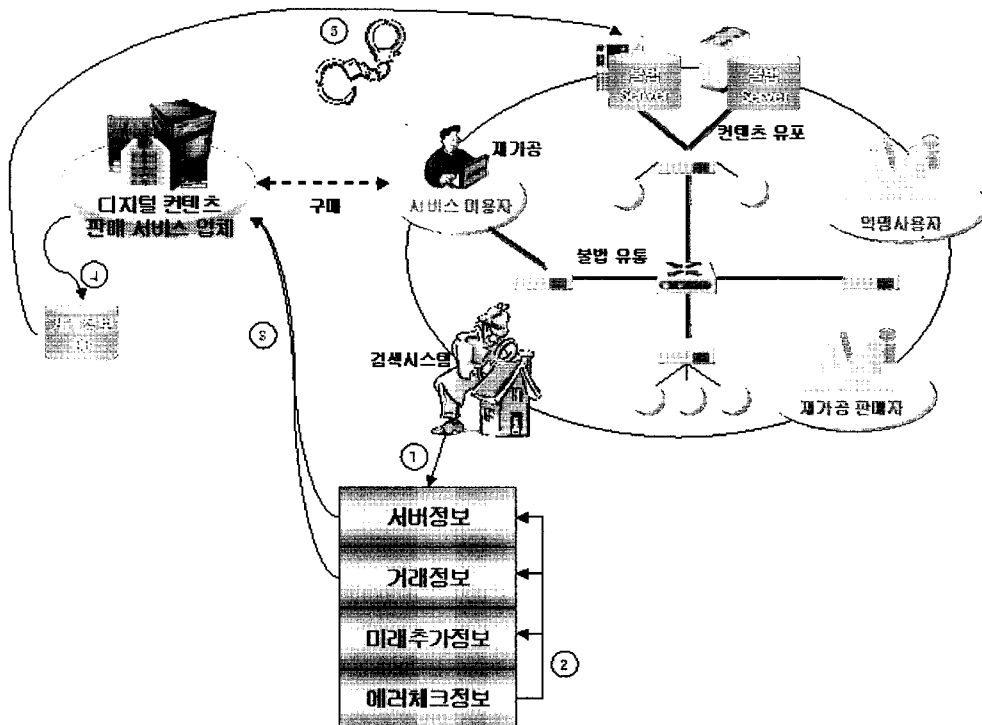


그림 4. 핑거프린팅 데이터 구조와 삽입 과정
Fig. 4. Fingerprinting data structure and embedding process.

- ① 검색시스템(로봇)이 인터넷상의 콘텐츠를 수집하여 핑거프린팅 정보를 추출한다.
- ② 추출된 정보의 정확성 여부를 확인하기 위하여 CRC를 체크한다.
- ③ 추출된 정보가 정확하면 서버정보의 정보에 의해서 해당서버로 거래정보를 전송한다.
- ④ 거래정보를 이용하여 해당 사용자가 원 구매자인지를 확인한다.
- ⑤ 확인 후 원 구매자가 아닐 경우, 원 구매자가 누구인지를 거래정보 DB를 검색하여 추적한다

4.2 핑거프린트 삽입 모듈

핑거프린트 삽입 모듈은 콘텐츠가 구매자로 다운로드될 때 핑거프린팅 정보를 삽입하는 기능을 수행한다. 구매자는 콘텐츠를 다운로드 받기 위해 콘텐츠공급자(Contents Provider)에 접속한다. 이때 사용자 ID를 입력하여 접속하게 되고 요청된 콘텐츠는 데이터베이스에서 바로 가져오지 않고 핑거프린트 삽입서버에 요청한다. 핑거프린트 삽입서버는 요청된 콘텐츠를 호출한 후 거래정보 데이터베이스에 기록되는 정보에 대한 거래정보를 다운로드 될 콘텐츠에 핑거프린트로 삽입한다. 삽입되는 정보는 위에서 설명된 데이터구조를 가지며 원 구매자를 추적하는 기본 자료가 된다. 그림 5는 핑거프린트 삽입서버의 동작을 나타내고 있다. 콘텐츠DB는 판매하고자 하는 콘텐츠를 저장하고 있는 DB이다

4.3 콘텐츠 웹 서버

콘텐츠 웹 서버는 ASP와 액세스(ACCESS)의 mdb를 이용하여 구성된다. ASP를 이용하여 웹 서버상에서 데이터베이스에 접속하여 사용자를 등록/삭제/갱신하며

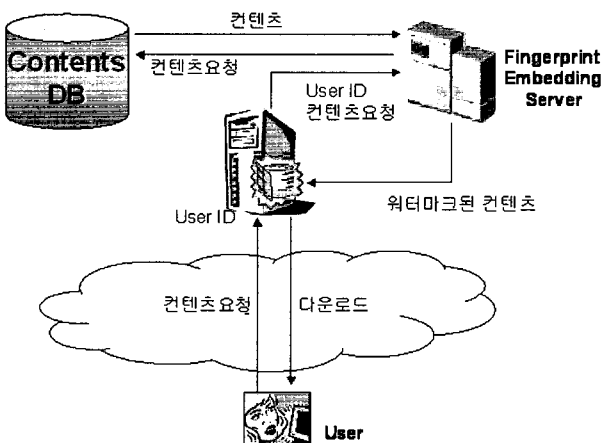


그림 5. 핑거프린트 삽입 서버
Fig. 5. Fingerprint embedding server.

거래가 발생할 때마다 거래정보 데이터베이스 거래정보를 추가하고, 콘텐츠 데이터베이스로부터 콘텐츠를 가져와서 정보를 삽입하는 모든 과정을 통제한다. 콘텐츠 웹 서버의 기능 및 동작은 다음과 같이 개념적으로 설명할 수 있다.

● 계정관리

거래정보 데이터베이스와 연결되어 사용자의 등록/삭제/갱신을 관리한다. 또한 사용자의 로그인을 통제함으로써 구매자 추적에 방해가 되는 익명접속을 허용하지 않는다.

● 콘텐츠 관리

로그인한 사용자가 원하는 콘텐츠를 콘텐츠DB로부터 호출하여 정보 삽입 모듈로 전달한다.

● 거래정보 관리

거래가 발생할 때 마다 거래정보 DB를 갱신하여 효율적인 구매자 추적을 하는데 이용된다.

● 정보삽입 관리

정보 삽입에 필요한 거래정보와 콘텐츠를 정보삽입 모듈에 전달하여 정보삽입 모듈을 동작시켜 핑거프린팅된 콘텐츠를 다운로드 할 수 있게 해 준다.

4.3.1 콘텐츠 검색 웹 로봇

본 논문에서는 원 구매자를 추적하기 위하여 인터넷 상에 배포되어 있는 이미지 콘텐츠를 효율적으로 검색하기 위해 검색 웹 로봇을 이용한다. 웹 로봇은 인터넷 상에서 외부로부터 접근 가능한 컴퓨터 서버들을 방문하면서 콘텐츠들을 핑거프린트 추출 서버로 전송한다.

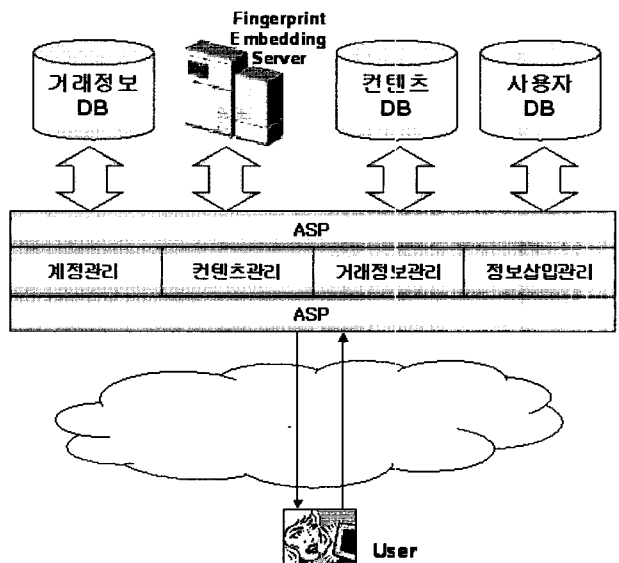


그림 6. 콘텐츠 웹 서버
Fig. 6. Contents web-server

일반적인 홈페이지나 혹은 게시판을 운영하는 사이트가 방문대상이다. 검색 로봇은 크게 링크 정보 검색 모듈과 콘텐츠 수집 모듈, 콘텐츠 분석 모듈로 나누어진다.

가) 정보검색 모듈

정보 검색 모듈은 검색로봇이 인터넷상에서 정보를 수집할 링크 사이트들을 검색해서 링크정보DB에 저장하는 기능을 한다. 초기의 웹 주소를 입력하면, 그 웹 주소를 시작으로 연결된 사이트들을 조사하고 DB에 저장한다. 또한 연결된 사이트로 접속하여 그 사이트와 연결된 또 다른 사이트들을 지속적으로 조사하고 DB에 저장함으로써, 본 시스템이 정지하지 않고, 계속해서 자동으로 DB에 저장된 사이트들을 찾아다니며 콘텐츠를 수집할 수 있게 해준다. 그림 7은 정보 검색 모듈이 정보를 수집하고 저장하는 과정을 보여주고 있다

실제적으로 웹 상에서는 그림 7과 같이 단순히 처리하는 것이 아니라 더 복잡하다. 중복된 사이트를 계속해서 접속할 경우, 무한 반복과 같은 오류가 발생할 수 있기 때문이다. 따라서 정보 검색모듈은 다음과 같은 기능들을 가져야 한다.

- 멀티쓰레드에 의한 동시 수집 기능
- 스케줄링에 의한 수집시간 관리 기능
- HTML파싱과 JAVA script 파싱 기능

나) 콘텐츠 수집 모듈

콘텐츠 수집 모듈은 정보 수집 모듈에 의해서 수집된 링크주소를 사용하여 사이트에 접속하고 실제로 콘텐츠를 다운로드 해서 데이터베이스에 저장하는 기능을 수행한다. 콘텐츠 수집 모듈은 콘텐츠DB와 링크정보DB,

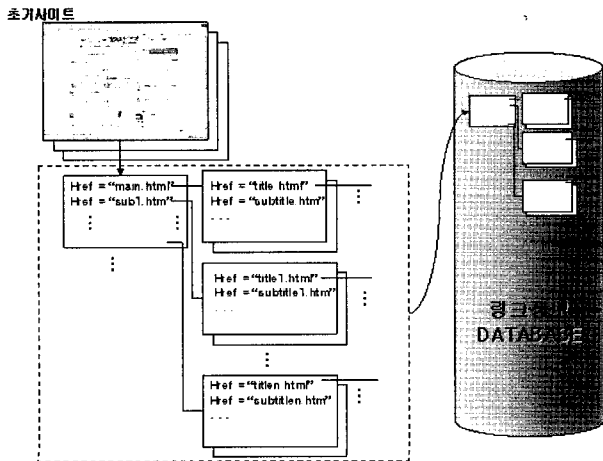


그림 7. 정보검색모듈
Fig. 7. Web search module.

두 개의 DB가 접속되어 있어서 링크 사이트를 관리하고 콘텐츠 DB에 다운로드 된 콘텐츠를 저장한다. 링크 정보 DB는 앞으로 접속해야 할 사이트를 관리하는 미방문주소(Unvisited URLs)와 이미 접속한 사이트를 관리하는 방문주소(Visited URLs)로 구성된다.

콘텐츠 수집 모듈은 미방문 사이트에 접속해서 문서를 받아오고 HTML파서에 의해서 문서를 분석하여 콘텐츠의 주소를 추출한다. 최종적으로 추출된 주소로부터 콘텐츠를 다운로드 하여 콘텐츠DB에 저장한다. 콘텐츠 수집 모듈의 기능을 정리하면 다음과 같다.

- 다양한 콘텐츠 수집/분석 기능
(콘텐츠 타입별 선별적인 수집 기능)
- 주소확인기(URL Checker)에 의한 비유효주소(Dead Link) 체크
 - 주소확인 : 수집 주소의 비유효사이트 체크를 통하여 정보의 상태 체크.
 - 갱신확인(update checker) : 수집 정보의 갱신 정책을 통하여 정보의 갱신체크. 만일 갱신이 되지 않았을 경우, 일정기간이 만료될 때까지는 다시 접속하지 않음.
- 수집결과 자동분류 및 저장 기능
- 원하는 정보만 추출할 수 있는 선별적HTML 문서 파싱 기능

그림 8은 구현된 콘텐츠 수집 모듈의 전체구조를 보여 주고 있다. 방문모듈이 링크정보 DB에서 미 방문된 사이트의 주소를 가져와 접속하고 HTML문서를 다운로드 하면 파서가 주소를 추출하고 콘텐츠 다운로드 모

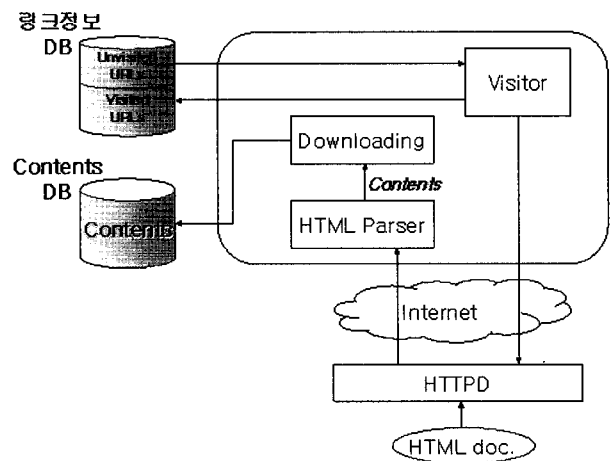


그림 8. 콘텐츠 수집 모듈
Fig. 8. Contents correcting module.

들이 실제의 콘텐츠를 다운로드하여 콘텐츠DB에 저장한다. 일반적으로 멀티미디어 콘텐츠는 대용량이고 많은 링크정보가 사이트 내에 포함되어 있기 때문에 제안된 모듈에서는 수집의 효율화를 위해 다음과 같은 검색조건을 지정할 수 있도록 설계되었다.

- 콘텐츠 크기에 따른 선택적 다운로드
(일반적으로 40KB 이상)
- 네트워크 전송 속도에 따른 선택적 사이트 검색

다) 콘텐츠 분석 모듈

콘텐츠 분석 모듈은 콘텐츠DB에 저장된 콘텐츠에서 핑거프린팅 정보를 추출하는 기능을 수행한다. 추출된 핑거프린팅 정보는 거래정보와 원 구매자 정보, 그리고 배포정보를 포함하고 있으며 로그데이터 형태로 관리된다. 이 정보는 콘텐츠의 판매자나 저작권자에게 콘텐츠가 적법으로 유통되었는지를 판단할 수 있는 근거를 제공한다. 또한 콘텐츠의 배포 상태에 대한 정보도 주기적으로 보고서 형식으로 함께 제공된다.

그림 9는 콘텐츠 분석 모듈의 역할에 대해서 보여주고 있다. 그림에서 로그데이터는 파일이 존재하는 위치를 알려주는 링크주소, 콘텐츠 파일명, 콘텐츠로부터 추출된 핑거프린팅 정보, 검색일자로 구성되어 있다. 이렇게 구성된 로그데이터는 정기적으로 판매자와 저작권자 또는 콘텐츠 관리자에게 원하는 형태의 레포트로 정기적으로 알려줌으로써 콘텐츠들이 어디에 어떻게 분포되어 있는지 관리할 수 있게 한다. 또한, 복제 콘텐츠가 발견될 경우, 핑거프린팅 정보를 추출하여 콘텐츠DB와 사용자DB를 검색하면 누가 원 구매자인지 알 수 있다.

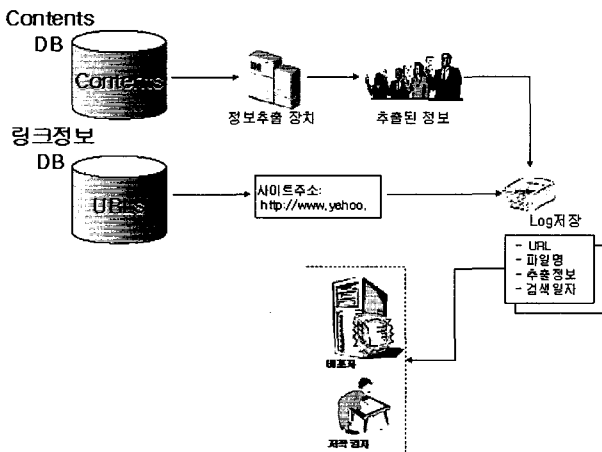


그림 9. 콘텐츠 분석 모듈
Fig. 9. Contents analysis module.

V. 실험결과

제안된 이미지 콘텐츠 추적 시스템은 윈도우 2000의 환경하에서 Visual C++을 이용하여 구현되었고 인터넷상의 bmp와 jpg 형식의 이미지 파일을 검색대상으로 하고 있다. 3개의 데이터베이스를 연결하기 위하여 ODBC를 사용하였고 링크정보를 추출하고 이미지를 다운로드하기 위하여 MSHTML파서와 URL 모니터(moniker)를 사용하였다.

실험은 제안한 이미지 추적 시스템의 검색효율과 핑거프린팅 정보 추출 기능을 검증하기 위하여 일반적인 웹 사이트에 대한 검색효율성 실험과, 핑거프린팅 이미지를 포함하고 있는 테스트 사이트를 구축하여 검색하는 실험을 실시하였다.

표 1은 검색을 시도한 3개의 사이트에 대한 검색이미지 수와 검색시간을 보여주고 있다. 사이트 1, 2는 인터넷상의 일반적인 사이트이며 사이트 3은 실험을 위하여 17개의 핑거프린팅 된 이미지를 게시한 사이트이다. 사이트 1, 2에서 총 이미지 수와 저장된 이미지 수의 차이는 검색의 효율성을 위해서 이미지 크기에 따라 다운로드를 제한한 결과이다. 40KByte 미만의 이미지는 핑거프린팅 정보 추출에서 제외하였다.

사이트 1, 2와 같은 일반적이 사이트에서는 핑거프린팅 정보가 검출되지 않았지만 실험을 위해 구축된 사이트 3의 경우에는 게시된 모든 콘텐츠의 핑거프린팅 정보가 검색되었다. 실제로 핑거프린팅 된 콘텐츠가 인터넷에 유통되는 것이 아니기 때문에 다양한 실험 환경을 구현하는 것은 매우 어렵다. 그러나 핑거프린팅 정보를 포함하는 테스트 사이트에서 모든 콘텐츠에 대한 핑거프린팅 정보를 추출했다는 결과는 제안한 시스템의 추출 기능을 검증한 결과라 할 수 있다. 핑거프린트의 추출에 있어서, RST공격이나 기타 이미지 처리 후에도 추출이 가능한지를 보이는 실험은 본 논문의 향후 실험 과제이다.

다음 실험으로는 이미지 검색과 핑거프린팅 정보 추

표 1. 검색된 이미지 수와 검색 시간
Table 1. The results of searched images and elapsed time.

| Site | Search pages | Total images | Stored images | Elapsed time |
|------|--------------|--------------|---------------|--------------|
| 1 | 626 | 1,406 | 800 | 3:14:10 |
| 2 | 71 | 275 | 207 | 00:03:55 |
| 3 | 3 | 17 | 17 | 00:00:23 |

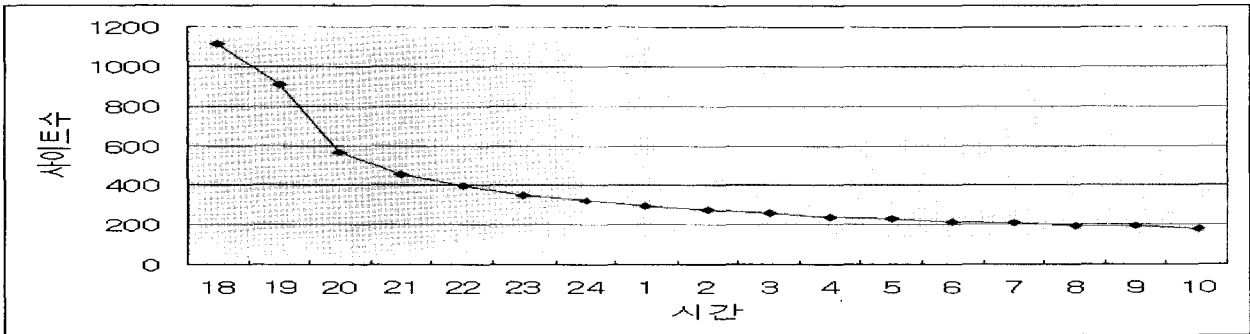


그림 10. 시간별 분석 사이트 수
Fig. 10. The number of visited sites.

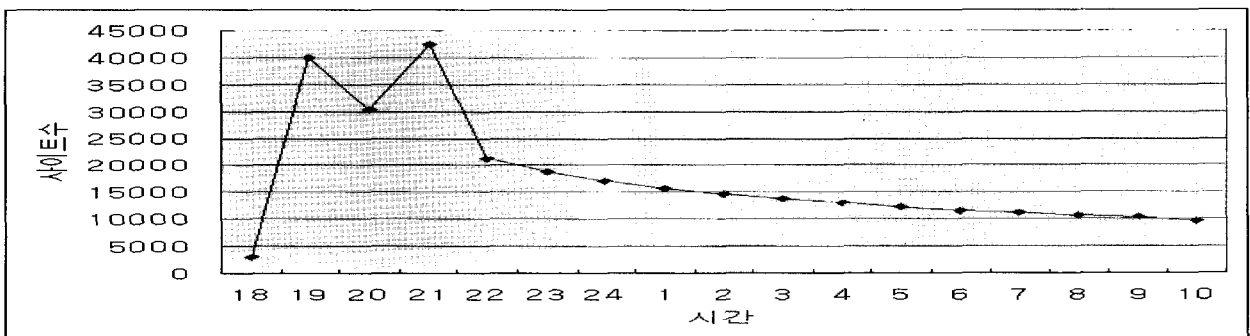


그림 11. 시간별 검색 링크 수
Fig. 11. The number of visited links.

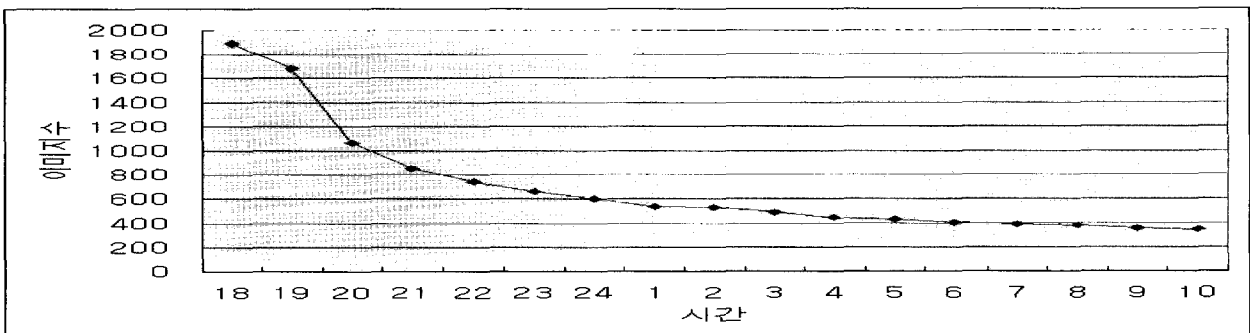


그림 12. 시간별 핑거프린팅 정보 추출 이미지 수
Fig. 12. The number of retrieved images.

출의 효율성을 검증하기 위하여 인터넷의 일반 사이트를 대상으로 17시간 동안 검색을 실시하였다. 검색 결과는 시간별 분석 사이트 수, 시간별 검색 링크 수 그리고 시간별 핑거프린팅 정보 추출 이미지 수로 구분하였다.

그림 10의 시간별 분석 사이트 수는 18시부터 익일 10시까지 검색 로봇이 방문하여 HTML문서를 완전히 분석한 사이트 수이다. 시간이 지날수록 사이트의 수가 감소하는 이유는 콘텐츠 수집 모듈이 사이트의 방문 시 중복방문을 줄이기 위해 미방문 주소인지를 DB에서 검색하는 시간이 길어지기 때문이다.

그림 11은 시간별 검색한 링크 수이다. 이는 분석된 HTML문서로부터 얻어지는 링크된 사이트의 개수다.

이 결과 또한 시간이 지날수록 줄어드는 결과를 보이고 있는데, 이는 방문한 사이트가 점점 많아지고 또한 중복방문인지를 검색하는 시간이 많이 요구됨에 따라 새롭게 링크된 사이트 정보를 얻어오는 시간이 많이 소비되기 때문이다.

마지막으로 그림 12는 링크된 사이트로부터 실제로 이미지 콘텐츠를 다운로드하고 핑거프린팅 정보를 추출한 이미지의 시간별 수를 나타내고 있다. 시간이 지남에 따라 그림 10과 11에서 보여준 결과와 마찬가지로 검색된 이미지의 수가 감소하지만 평균적으로 시간당 600-1000개의 이미지를 검사하였다. 검색로봇의 사이트 중복방문을 방지하는 효율적 방문 정책을 적용하여 검

색의 효율성을 증대시키는 것은 본 논문의 향후 연구방향이다.

VI. 결 론

본 논문에서는 구매자의 정보와 거래정보를 이미지 콘텐츠 내에 핑거프린팅 정보로 삽입하고 배포한 다음 웹 로봇을 이용하여 콘텐츠를 수집하고 수집된 콘텐츠로부터 구매자 정보를 역추적하여 복제 콘텐츠의 원 구매자를 추적할 수 있는 시스템을 제안하였다. 또한 제안한 이미지 콘텐츠 추적 시스템을 인터넷상에서 일반 사이트와 테스트 사이트를 대상으로 실험하여 콘텐츠 추적 기술을 검증하였다. 제안한 기술은 인터넷상에서 콘텐츠를 구별하는 기능과 함께 거래정보도 함께 추출할 수 있어 콘텐츠의 복제 시 원 구매자를 쉽게 추적할 수 있도록 한다. 이는 제안된 시스템이 향후 불법복제 콘텐츠 추적시스템으로도 활용될 수 있음을 보여주고 있다.

참 고 문 헌

- [1] Digimarc Corporation. <http://www.digimarc.com>
- [2] J. Dittmann, P. Schmitt, E. Saar, J. Schwenk, and J. Ueberberg, "Combining digital Watermarks and collusion secure Fingerprints for digital Images." *SPIE Journal of Electronic Imaging*, vol.9, pp.456-567, 2000.
- [3] D. Boneh and J. Shaw, "Collusion-Secure Fingerprinting for Digital Data," *IEEE Tran. on Information Theory*, vol.44, no.5, pp.1897-1905, 1998.
- [4] J. Domingo-Ferrer and J. Herrera-Joancomarti, "Simple collusion-secure fingerprinting schemes for images," *Proc. Intl. Conf. Inf. Tech.: Coding & Computing 2000*, pp. 128-132, 2000.
- [5] F. Seband J. Domingo-Ferrer, "Short 3-secure fingerprinting codes for copyright protection," *Lecture Notes in Computer Science*, vol. 2384, pp. 316-327, June 2002.
- [6] W. Trappe, M. Wu, and K.J.R.Liu, "Collusion-resistant for multimedia," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, 2002, pp.3309-3312.
- [7] Y. Wang, J. F. Doherty, and R. E. Van Dyck, "A watermarking algorithm for fingerprinting intelligence images," *Proc. Conf. on Information Sciences and Systems*, Baltimore, MD, March 2001.
- [8] D. Kirovski, H. S. Malvar, and Y. Yacobi, "Multi-media Content Screening using a Dual Water-marking and Fingerprinting System," *Proc. ACM Multimedia 2002*, Juan Les Pins, France, 2002.
- [9] J. Kilian, T. Leighton, L. R. Matheson, T. G. Shamoan, R. E. Tarjan, and F. Zane, "Resistance of digital watermarks to collusive attacks," *Tech. Rep. TR-585-98*, Dept. of Computer Science, Princeton University, 1998.
- [10] Joseph. J. K, O Ruanaidh and T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," *Signal Processing*, 66(3), 1998.
- [11] M. Koster, "WWW Robots, Wanderers and Spiders," <http://www.robotstxt.org/wc/robots.html>

저 자 소 개



서 영 호(정회원)
 1979년 경북대학교 전자공학과
 학사 졸업.
 1995년 광운대학교 전자계산학과
 석사 졸업.
 2001년 ~ 현재 충남대학교 컴퓨터
 과학과 박사 수료.

1978년 7월~1998년 5월 시스템공학연구소
 책임연구원
 1998년 5월~현재 한국전자통신연구원 디지털
 콘텐츠연구단 책임연구원
 2000년 12월~현재 DRM포럼 운영위원
 <주관심분야: 영상처리, 워터마킹, 핑거프린팅,
 DRM>



김 원 겸(정회원)
 1992년 충남대학교 전산학과 학사
 졸업.
 1994년 충남대학교 전산학과 석사
 졸업.
 2001년 충남대학교 컴퓨터과학과
 박사 졸업.

1995년 2월~1997년 7월 LG반도체 생산기술
 연구소 주임연구원
 2002년 2월~현재 한국전자통신연구원 디지털
 콘텐츠연구단 선임연구원
 <주관심분야: 이미지/오디오 신호처리, DRM, 워
 터마킹, 핑거프린팅>



이 선 화(정회원)
 2000년 부산대학교 전자계산학과
 학사 졸업.
 2002년 부산대학교 전자계산학과
 석사 졸업.
 2002년 3월~현재 한국전자통신
 연구원 디지털콘텐츠연구
 단 연구원

<주관심분야: 영상/신호처리, 워터마킹, 핑거프린
 팅>



황 치 정(정회원)
 1975년 서강대학교 수학과 학사
 졸업.
 1979년 서강대학교 수학과 석사
 졸업.
 1981년 뉴욕주립대학교 수학과
 석사 졸업.

1985년 코네티컷 주립대학 전산학 석사 졸업
 1987년 코네티컷 주립대학 전산학 박사 졸업
 1980년~1981년 뉴욕주립대학 강사
 1981년~1987년 코네티컷 주립대학 강사
 1987년~1988년 코네티컷 주립대학 객원교수
 1988년~1988년 한국원자력연구소 선임연구원
 1988년 현재 충남대학교 정보통신공학부 교수
 <주관심분야: 영상처리, 패턴인식, 핑거프린팅>