콘텐츠 기반 음란물 사이트 검출에 관한 연구 동향

*한유나 *박상성 *신영근 *장동식
고려대학교 정보경영공학부
*chunyouna@korea.ac.kr

A Study on Pornographic Web Site Detection Based on Contents

*Han, You-Na *Park, Sang-sung *Shin, Young-Guen *Jang, Dong-Sik
Korea University Division of Information Management Engineering

요약

인터넷의 발전은 온라인상의 무한한 정보 이용을 가능케 하였다. 인터넷 상의 무한한 정보에는 유익한 정보도 있지만 그렇지 못한 유해한 정보도 있다. 대표적인 유해 정보인 음란물 사이트는 청소년들에게 쉽게 노출되어 심신건강에 악영향을 끼친다. 최근에는 이러한 음란물 사이트를 차단하기 위한 알고리즘 개발이 활발히 진행되고 있다. 본 논문에서는 음란물 사이트 차단을 위해 연구되고 있는 대표적인 알고리즘들의 성능 비료를 통하여 각 알고리즘의 취약점을 보완할 수 있는 새로운 방법론을 제시한다.

키워드 : 음란물, 검출, 베이즈 정리, MPEG-7
Keywords : pornographic, detection, Bayes' Rule, MPEG-7

1. 서론

인터넷의 지속적인 발전으로 인하여 현재는 전 세계 사람들이 정보를 공유하고 또 새로운 정보도 얻을 수 있는 거대한 데이터베이스 역할을 하고 있다. 인터넷은 또 사람들은 클릭 몇 번으로 찾고자 하는 사이트, 자료, 음악, 영화 등을 전 세계 데이터베이스에서 마음대로 찾고 보고 듣고 할 수 있는 편리함을 제공해 주었다. 하지만 유용한 사이트뿐 만 아니라 음란물사이트나 폭력, 자살, 마약 등 사이트도 동시에 증가하면서 클릭 한번 잘못하면 전혀 엉뚱한 사이트로 안내하는 곳도 엄청 많아졌다. 이러한 사이트들은 특히 분별력이 약한 청소년들에게도 너무 쉽게 노출되어 있다는 게 큰 문제점이다. 청소년들은 사이트를 둘러보다가 예쁘게 디자인해 놓은 링크가 걸려있으면 호기심에 링크를 클릭한다. 만약 그 링크주소가 합법적인 사이트면 다행이지만 음란물처럼 불법사이트면 청소년들의 심신건강에 큰 영향을 미칠 것이다. 이러한 불법사이트들을 차단하여 청소년들에게 유용한 컨텐츠만을 접하게 하기 위하여 현재까지 국내외 연구진들이 많은 노력을 기울여왔다.

현재까지 연구된 음란물 사이트 검출 방법은 크게 3가지다. IP기반 블랙리스트 차단방법, 텍스트내용 기반 검출방법[1], 이미지내용 기반 검출방법[2] 으로 나눌 수 있다. 먼저 IP기반 블랙리스트 차단방법은 우선 유해인터넷 사이트의 주소리스트를 만들어 저장한다. 만약 접근하려는 사이트의 주소가 블랙리스트에 포함 되어 있으면 그 사이트는 차단되어 나타나지 않는다. 하지만 인터넷상에서의 콘텐츠는 너무 다양하고 또 많은 음란사이트들은 주소를 자주 바꾸기 때문에 매일 접근 불가능한 블랙리스트들을 업데이트 한다는 것은 거의 불가능한 일이다. 두 번째로 텍스트내용 기반 검출 방법은 텍스트 내용에 따라서 그 사이트가 음란물사이트인지 아닌지 판단한다. 이 방법은 머신러닝이나 데이터마이닝 같은 기술을 이용하여 성인사이트에서 자주 등장하는 용어나 구절을 추출해낸 다음 용어묶음으로 데이터베이스에 저장한 한다. 만약 접근하려는 사이트에 데이터베이스에 저장된 용어가 얼마 만큼 들어 있는가에 따라서 그 사이트가 음란물 사이트인지 아닌지를 판단한다. 마지막으로 이미지기반 검출방법은 스킨색상을 이용하여 피부영역을 검출한 다음, 피부색상이 전체 이미지에서 차지하는 비율을 계산하여 음란물 사이트인지 아닌지를 판단한다.

본 논문에서는 여러 가지 방법들 중에서 최근에 많이 사용되는 이론들을 중심으로 소개하고 새로운 연구방향을 제시하고자 한다.

2. 주요 음란물사이트 검출 기법

음란물 사이트 검출 방법 중에서 가장 중요한 두 가지는 텍스트와 이미지이다. 텍스트를 이용한 방법은 음란물사이트 뿐만 아니라 자살, 테러, 마약 사이트 같은 불법 사이트도 검출할 수 있다는 게 가장 큰 장점이다. 이미지를 이용한 검출방법은 음란물사이트가 텍스트는 줄이고 이미지나 동영상을 늘리는 최근 추세에 따라 함께 발전되어 왔다. 이 방법은 텍스트기반 검출방법으로 추출하지 못하는 음란물 사이트를 사이트에 존재하는 이미지를 분석함으로써 검출할 수 있다는 큰 장점이 있다.

이러한 텍스트기반 음란물 사이트 검출방법과 이미지기반 음란물 사이트 검출 방법에 대해 알아보기로 한다.

2.1. 텍스트기반 음란물 사이트 검출방법

텍스트기반 음란물 사이트 검출방법에서 가장 보편적으로 사용되

는 알고리즘은 Bayes′ Rule[3]이다. Bayes′ Rule는 음란물 사이트와 비음란물 사이트에서 가장 많이 추출된 용어를 이용하여 해당 사이트가 음란물 사이트인지 아닌지를 구분하는 방법이다. Bayes′ Rule에 따라서 미확인 문서 D는 n가지 용어로 구성된 용어벡터$D=(t_1,...t_n)$로 구성되었다.. 미확인문서 D는 음란물과 비음란물 카테고리를 모두 계산하여 높은 조건부 확률을 가지는 클래스Ci에 할당된다.

Bayes′ Rule에 따르면 어떤 문서가 주어진 클래스에 들어 있을 확률은 어떤 용어가 그 문서에 나타났으며 또한 다른 문서에도 나타나는 것을 관측한 빈도를 그 클래스의 대표멤버들이라고 하며 아래의 식(1)을 이용하여 구한다.

$$p(C_i|D)=\frac{\prod_{j=1}^{j=n}p(t_j|C_i)*p(C_i)}{p(D)} \quad (1)$$

2.1.1 평가 매트릭스

음란물 사이트 분류결과를 평가하는 매트릭스는 다음과 같다.

진짜부정(True Negative): T.N. 음란물 사이트인데 비음란물 사이트로 분류된 경우.

가짜긍정(False Positive) :F.P. 비음란물 사이트인데 음란물 사이트로 분류된 경우.

매크로 리콜 비율(Macro Recall Rate) : 음란물 사이트가 음란물 사이트로 분류된 것과 비음란물 사이트가 비음란물 사이트로 분류된 것들의 평균인데 식(2)와 같다.

$$\frac{1}{2}*(\frac{Porn-T.N.}{Porn}+\frac{Non-F.P.}{Non})*100 \quad (2)$$

여기서 Porn은 테스트 언어자료에서 전체 음란물 웹페이지수를 말하고 Non은 테스트언어자료에서 전체 비음란물 사이트의 웹페이지수를 말한다.

매크로정확도(Macro Precision) : 음란물 사이트로 분류된 웹페이지가 진짜 음란물 사이트인 것들과 비음란물 사이트로 분류된 웹페이지가 진짜 비음란물 사이트인 것들의 평균은 식(3)과 같다.

$$\frac{1}{2}*(\frac{Porn-T.N.}{Porn-T.N.+F.P}+\frac{Non-F.P.}{Non-F.P.+T.N.})*100 \quad (3)$$

만약 Porn이 테스트 언어자료에서 전체 음란물 아이템이거나, Non이 테스트 언어자료에서 전체 비음란물 아이템이면 테스트 언어자료에서 장확하게 웹페이지를 분류한 정확성(Acc)은 식 (4)와 같다.

$$\frac{Porn+Non-F.P-T.N}{Porn+Non}*100 \quad (4)$$

2.2 이미지기반 음란물 사이트 검출방법

현재까지 많은 이미지기반 음란물 사이트 검출방법이 연구되어 왔다. 이중 가장 보편적으로 사용되는 MPEG-7 Descriptor를 이용한 검출방법을 소개하도록 한다.

MPEG-7 Descriptor를 이용한 검출방법은 다음과 같다. 1) 이미지의 배경영역을 제거한 다음 스킨유사 측정방법을 이용하여 우리가 원하는 영역(ROI)을 얻는다. 2) MPEG-7의 세가지 Descriptor인 Color Descriptor, Texture Descriptor, Compactness Descriptor를 이용하여 유사이미지 검색을 수행한다. 3) 이미지가 입력되면 음란물 이미지와 비음란물 이미지를 미리 저장해둔 데이터베이스를 찾아서 100개 정도의 유사 이미지를 검출한다. 4) 검출된 음란물 이미지의 수가 임계치- T_{ad}보다 많으면 입력 이미지는 음란물로 간주되고 반대로 비음란물로 판단된다.

2.2.1 배경 제거

우리가 원하는 영역(ROI)를 얻기 위해 먼저 스킨색상과 다른 배경영역을 제거한다. 스킨유사 픽셀은 Cb, Cr보다 Y부분과 더 많은 연관이 있기에[3] 컬러 영역을 RGB에서 YCbCr로 바꾼다. 만약 변환된 픽셀의 Cb, Cr 영역이 아래의 공식을 만족하면 피부색으로 분류한다.[4]

$$Cr\geq\max\{-2(Cb+24),-(Cb+17),-4(Cb+32),\\ 2.5(Cb+\theta_1),\theta_3,0.5(\theta_4-Cb)\} \quad (5)$$

그리고

$$Cr\leq\min\{(220-Cb)/6,4(\theta_2-Cb)/3\}, \quad (6)$$

여기서 $\theta_1,\theta_2,\theta_3,\theta_4$ 은 다음과 같다.

$$\theta_1=\begin{cases}-2+(256-Y)/16, & \text{if } Y>128,\\ 6, & otherwise,\end{cases} \quad (7)$$

$$\theta_2=\begin{cases}20-(256-Y)/16, & \text{if } Y>128,\\ 12, & otherwise,\end{cases} \quad (8)$$

$$\theta_3=\begin{cases}6, & \text{if } Y>128,\\ 20-(256-Y)/16, & otherwise,\end{cases} \quad (9)$$

$$\theta_4=\begin{cases}-8, & \text{if } Y>128,\\ -16+Y/16, & otherwise,\end{cases} \quad (10)$$

상기식을 수행한 결과 이진 이미지를 얻을 수 있다. 스킨유사 영역을 검출한 후, 스킨유사영역은 흰색으로 나태내고 스킨유사영역이 아닌 부분은 검정색으로 나타낸다. 이후 이 이진이미지를 32X32 개의 중복되지 않는 블록으로 분할한다. 만약 매 블록에서 절반이상의 픽셀이 스킨유사영역 픽셀이면 그 블록은 스킨유사영역 블록으로 간주한다. 다음 3X3 구조로 닫힘 연산을 수행하여 가짜 스킨유사영역들을 제거하면 그룹화 된 스킨유사영역들을 얻을 수 있다. (Gonzalez and Woods,2002) 그리고 추출해낸 스킨유사영역을 따라서 미니멈 정사각형 바운딩 박스로 바운딩 표시하고 나머지 영역은 배경으로 간주한다. 만약 스킨유사영역이 검출되지 않았거나 우리가 얻고자 하는 영역의 넓이/높이가 50픽셀보다 작으면 얻고자 하는 영역이 전체 이미지로 간주된다.

2.2.2 특징 추출

ROI가 입력 이미지로부터 분할된 후에 있어 미리 입력된 데이터베이스에서 유사이미지를 찾는다. 내용기반 이미지검색에 일반적으로 제일 중요한 특징은 Color이다. Color를 이용하여 먼저 스킨유사영역을 추출한다. 이후 Texture 특징으로 음란물 이미지와 비음란물 이미지를 구분할 수 있는데 많은 음란물 이미지는 스킨이 많이 노출되기

때문에 부드러운 Texture특징을 갖고 있는 반면에 비음란물 이미지는 샤프한 에지를 가지고 있기 때문이다. 또한 스킨영역의 Shape 특징은 음란물이미지를 스킨유사영역과 배경을 분류하는데 중요한 도움을 준다. 그러므로 여기서는 Color, Texture, Shape 3가지 특징을 이용하여 음란물 이미지를 검출하려고 한다.[5]

2.2.2.1 Color 특징 추출 Scalable color descriptor(SCD) :

Color 특징을 추출하기 위해 사용하는 SCD를 구현하려면 공식 (11),(12),(13)을 이용하여 RGB컬러 공간을 HSV컬러 공간으로 바꾼다.[6]

$$H = \cos^{-1}\left\{\frac{[(R-G)+(R-B)/2]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right\}, \quad (11)$$

$$S = \frac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)}, \quad (12)$$

$$V = \frac{\max(R,G,B)}{255} \quad (13)$$

HSV컬러공간은 처음에 16개색상, 4개채도, 4개명도의 256컬러빈으로 양자화되고 각각의 이미지에 대해서 컬러히스토그램은 매 컬러빈에 대한 픽셀개수이다. 최종적으로 SCD는 다음과 같이 나타낸다.

$$scd = [s[1], s[2], \ldots, s[256]], \quad (14)$$

여기서 $s[i], 1 \le i \le 256$은 i번째 빈의 확률이다.

2.2.2.2 Texture 특징 추출 Edge histogram descriptor(EHD) :

MPEG-7에서 EHD는 입력 이미지에서 지역에지를 분류하는데 사용된다. 입력된 이미지는 먼저 4X4 개의 서브 이미지로 나눈다. 매 서브이미지에 대한 지역에지분류는 하나의 히스토그램으로 나타낼 수 있다. 히스토그램을 생성하려면 서브이미지는 블록기반 에지추출방법을 이용하여 다시 5가지 에지타입 : 수평, 수직, 45° 대각선,135° 대각선, 그리고 무방향으로 나뉜다. 에지 방향을 얻기 위해서 각 서브이미지를 다시 미리 정의된 수많은 겹치지 않는 정사각형 이미지블록으로 나눈다. 각 이미지에 대해 이미지 블록이 총 1100개 정도이면 제일 좋은 에지 방향특징을 추출할 수 있다.

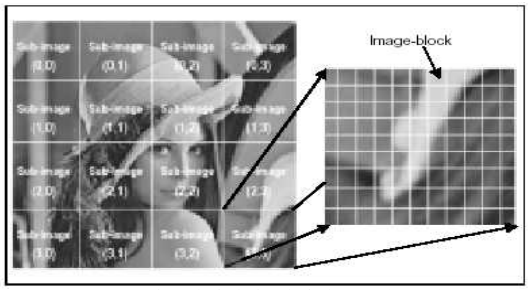

그림 1. 서브이미지와 이미지블록의 정의

다음 적당한 에지검출방법을 이용하여 에지길이를 측정한다. 만약 최대 에지길이가 주어진 임계치보다 크면 그에 해당하는 이미지 블록의 에지위치가 결정된다. 아니면 이미지는 무방향으로 간주된다.16개의 서브이미지가 존재하기 때문에 총 80(16X5)개의 히스토그빈을 얻을 수 있다.

H_E	Semantics
h(1)	Vertical edge of sub-image at (0,0)
h(2)	Horizontal edge of sub-image at (0,0)
h(3)	45° edge of sub-image at (0,0)
h(4)	135° edge of sub-image at (0,0)
h(5)	Nondirectional edge of sub-image at (0,0)
h(76)	Vertical edge of sub-image at (3,3)
h(77)	Horizontal edge of sub-image at (3,3)
h(78)	45° edge of sub-image at (3,3)
h(79)	135° edge of sub-image at (3,3)
h(80)	Nondirectional edge of sub-image at (3,3)

< 표1 >

$$ehd = [e[1], e[1], \ldots e[80], \mathrm{g}e[1], \mathrm{g}e[2], \ldots \mathrm{g}e[5]], \quad (15)$$

여기서 $e[i], 1 \le i \le 80$ and $\mathrm{g}[e], 1 \le j \le 5$는 각자 지역에지 히스토그램과 전역에지히스토그램을 나타낸다.

2.2.2.3 Shape 특징 추출 Compactness descriptor(CD) :

CD는 스킨영역을 흰색으로 하고 배경영역을 검정색으로 한 후 ROI를 묘사할 때 사용된다. ROI는 몇 개의 서로 다른 블록으로 나뉜다. 전역블록 1개, 전역블록을 4개 블록으로 분할한 블록 4개, 전역블록을 16개로 분할한 블록 16개, 해서 총 21개의 블록으로 나뉜다. 매 블록에 대한 스킨픽셀의 비율 PSP(Propotion of the Skin Pixels)는 다음과 같다.

$$PSP_i = \frac{N_j}{W_j \times H_j} \quad (16)$$

N_j는 블록 j의 스킨영역픽셀이고 W_j와H_j는 각각 넓이와 높이이다. PSP_1은 전역블록의 **cd**를 나타내고 $PSP_j, 2 \le j \le 5$일 때 4개의 서브블록의 **cd**를 나타내며 $PSP_j, 6 \le j \le 21,$은 16개 작은 사이즈의 **cd**를 나타낸다. 전역블록의 값 PSP_1을 cd의 최초특징값으로 한다면 **cd**는 다음과 같다.

$$cd = [cd[1], cd[1], \ldots cd[21]] \quad (17)$$

2.2.3 이미지검색기법을 이용한 음란물 검출

음란물검출을 이미지검색기법을 이용하여 실행한다. 이미지가 입력이 되면 미리 만들어 놓은 데이터베이스에서 약 100개 정도의 음란물과 비음란물 이미지가 검출된다. 이후 특징벡터 **scd,,ehd** 와 **cd**를 이용하여 입력 이미지 t 와 각각 매칭되는 이미지 s의 거리가 각각 계산되며 이는 d_{SCD}, d_{EHD}, d_{cd}로 표시된다.

$$d_{SCD}(t,s) = \| scd_t - scd_s \| = \sum_{i=1}^{256} |scd_t[i] - scd_s[i]|, \quad (18)$$

$$d_{EHD}(t,s) = \| ehd_t - ehd_s \| = \sum_{i=1}^{80} |e_t[i] - e_s[i]| + \sum_{i=1}^{i=5} |ge_t[i] - ge_s[i]|, \quad (19)$$

$$d_{cd}(t,s) = \| cd_t - cd_s \| = \sum_{i=1}^{21} |cd_t[i] - cd_s[i]| \quad (20)$$

$d_{SCD}(t,s)$에 따라 제일 짧은 거리를 가진 이미지 g(g=100)가 제일 먼저 발견되고 거리 값 증가순서대로 분배한다. i번째 분류된 이미지는 $g-i+1, 1 \le i \le g$등급에 할당된다. 그리고 유사하지 않은 다른 이미지들은 0등급에 할당된다. 보통 짧은 거리를 가진 이미지가 높은 등급을 가진다. 테스트 이미지 t의 각각의 매칭 이미지 s는 SCD 특징에 따라 하나의 등급에 할당되며 $G_{SCD}(t,s)$로 표시한다. 등급 할당 프로세스는 각 매칭 이미지의 EHD, CD의 거리도 같이 계산하여 각각의 등급을 $G_{SCD}(t,s)$, $G_{CD}(t,s)$ 로 표시한다. 결론적으로 각각의 매칭 이미지 s에 대하여 전체적인 등급은 3가지 등급의 합으로 계산된다.

$$G(t,s) = G_{SCD}(t,s) + G_{EHD}(t,s) + G_{CD}(t,s) \quad (21)$$

이 기준에 따라서 높은 등급을 가진 이미지는 제일 유사한 이미지로 간주된다. 여기서 100개의 유사한 이미지가 검출되었을 때 Na를 이 검색집단에서의 총 음란물 이미지라고 하면 Na가 주어진 임계치 Tad보다 크면 음란물 사이트로 판단하고 아니면 비음란물 사이트로 판단한다. 예를 들어서 아이가 인터넷을 사용할 때는 Tad를 작게 설정하여 불필요한 이미지를 모두 제거하고 일반적으로 사용할 때에는 Tad를 좀 크게 설정하여 음란물 이미지가 아닌 것을 음란물 이미지로 분류한 오류들을 크게 줄일 수 있다.

3. 실험결과

Bayes' Rule를 이용한 텍스트기반 성인사이트 검출방법은 HTML의 모든 섹션(타이틀, 메타, 바디)에서 모두 95%가 넘는 정확도를 기록했으며 특히 바디에서는 98.4%나 되는 정확도를 얻을 수 있었다. 타이틀에서 적은 비율을 얻은 원인은 많은 웹페이지의 타이틀의 용어가 보통 10개 이내로 매우 짧기 때문이다. 또한 메타부분에서 적은 퍼센티지를 기록한 원인은 많은 웹페이지들이 메타정보를 포함하고 있지 않기 때문이다. 이 3가지 섹션을 적절하게 조합함으로써 학교나 일반 서치엔진에서 좋은 결과를 나타낼 수 있다.

이미지를 이용한 검출방법에서 만약 입력된 이미지가 넓이나 높이가 50픽셀보다 작으면 아이콘 이미지로 간주하고 비음란물 이미지로 분류된다. 검출비율로 성능 측정하는 방법은 식(22)와 같다.

$$DR = \frac{|Q|}{|B|} \quad (22)$$

여기서 B는 원래 있는 음란물이미지/비음란물이미지의 수이고 Q는 검출된 음란물이미지/비음란물이미지의 수이다. Tad를 10으로 설정했을 때 SCD만을 이용하여 음란물과 비음란물 검출한 비율 DR은 96.43%와 79.27%를 기록하였다. 하지만 SCD, EHD, CD 3가지를 합치면 각각 99.47%와 79.46%로 상승하였다.

각각의 장단점을 비교해보면 다음과 같다.

	텍스트 기반	이미지 기반
장점	음란물사이트뿐만 아니라 자살, 폭탄, 테러, 마약 등 불법 사이트도 대거 검출할 수 있다.	음란물사이트, 특히 이미지뿐만 아니라 음란 동영상도 검출할 수 있는 장점이 있고 정확도가 높다.
단점	이미지가 많고 텍스트가 적은 사이트에서는 검출율이 약하다. 최근 flash 등으로 제작된 사이트에서는 검출이 불가능하다.	조명과 같은 외부환경에 많은 영향받으며, 스킨과 유사한 색상은 모두 검출되기 쉽다. 텍스트 검출보다 속도도 좀 느린 단점이 있다.

< 표2 >

4. 실험결과

본 연구에서는 음란물 사이트 검출방법 중 2가지 일반적인 방법을 자세히 살펴보았다. 향후의 연구과제는 앞서 언급한 다양한 검출방법을 조합하여 텍스트와 이미지뿐 만아니라 각각 인종에 대한 스킨색상을 구별하는 방법뿐 만아니라 한 장의 이미지에 서로 다른 인종의 사람이 같이 있을 때 음란물인지 아닌지를 구별하는 방법이 필요하고 또 한 사람이나 여러 사람이 다양한 포즈를 취하였을 때 어떻게 음란물인지 아닌지를 구별하는 방법도 필요하다. 또한 이 많은 작업을 동시에 수행한다면 검출 속도도 느려질 것이기 때문에 속도를 향상시키는 방법도 향후 연구방향이다.

Reference

[1]W.H.Ho., P.A.Wattrers. "Identifying and Blocking pornographic content".ICDE ,2005

[2]Jau-Ling Shih,Chang-Hsing Lee,Chang-Shen Yang."An adult image identification system employing image retrieval technique",Pattern Recognition Letters 28,2007,pp.2367-2374.

[3]T.Bayes."An essay towards solving a problem in the doctrine of chance",Philosophical Transactions of the royal Society of London,Society of London,53,1763,pp.370-418.

[4]Rafael C.Gonzalez,Richard E.Woods,Steven L.Eddins,"Digital Image Processing Using MATLAB" ,Prentice-Hall.Inc, 2004, pp. 204-206

[5]B.S.Manjunath, Philippe Salembier, Thomas Sikora. "Introduction to MPEG-7", WILEY,2002,pp.179-260.

[6]Rafael C.Gonzalez,Richard E.Woods,"Digital Image Processing",Prentice-Hall.Inc,2002,pp.290-301