제23회 한국정보처리학회 춘계학술발표대회 논문집 제12권 제1호 (2005. 5)

하이퍼링크 연관성을 이용한 유해사이트의 자동분류

장영헌
고려대학교 컴퓨터과학기술대학원 정보통신학과
andsimpson@korea.ac.kr

Automatic Harmful Website Rating System Based on Hyperlink Relationship

Young-Hun Jang
Dept. of Information and Communication
Graduate School of Computer Science and Technology, Korea University

요 약

인터넷의 발전과 함께 유해사이트의 급속한 증가로 유해사이트 분류의 신뢰도를 높일 필요성이 높아지고 있다. 기존의 유해사이트 분류방식에는 텍스트 기반의 분류방식과 Skin-Color Detection 알고리즘을 이용한 이미지 기반 방식이 있으며, 현재 텍스트 기반의 사이트 분류방식이 보편적으로 사용되고 있다. 본 논문은 기존 유해사이트 분류의 신뢰도를 높이기 위하여 유해사이트에 포함된 링크 정보를 기반으로 유해사이트 분류의 정확성을 검증할 수 있음을 증명하였다.

1. 서론

인터넷의 급속한 발전과 확장으로 인해 웹에서 얻을 수 있는 정보의 종류와 수는 급진적으로 증가하고 있다. 이러한 정보들은 많은 도움과 편의성을 제공하며 대부분 인류의 삶을 풍요롭고 긍정적으로 바꾸고 있지만 그 이면에는 인터넷의 익명성, 상업성으로 인해 비윤리적이고 부도덕한 행위가 비일비재하게 일어나고 있으며 유해한 정보 역시 급속도로 증가하고 있는 추세이다.

이에 안전하고 건전한 인터넷 활용을 위해 국내 및 해외의 음란·도박사이트를 차단하고 인터넷 음란사이트 DB를 구축하여 관리할 필요성이 증대되고 있으며 정보통신윤리위원회 등 관련기관에서는 음란·폭력정보 등급DB를 구축하여 제공하고 있다[1].

현재 유해 사이트 분류는 검색요원이 직접 분류하는 방식과 웹 검색 로봇(에이전트)이 컨텐츠의 텍스트 혹은 이미지를 추출해 패턴 매칭하는 방식으로 분류를 하고 있다. 사람이 직접 분류하는 방식은 정확도는 높으나 속도와 비용이 많이 들고 웹 검색 로봇에 의한 방식은 속도는 빠르지만 정확성은 그다지 높지 않아 비교적 높은 오분류의 소지를 가지고 있다.

이에 본 논문에서는 컨텐츠에 포함된 하이퍼링크 정보를 이용하여 기분류된 유해사이트 분류의 정확성을 검증할 수 있음을 증명하고자 한다.

본 논문은 2장에서 유해사이트 분류와 웹 구조분석에 관련된 기존연구에 대해 서술하고 3장에서 유해사이트 분류의 정확도를 높일 수 있는 알고리즘과 시스템을 제안하였다. 4장에서 실험 방식과 결과를 서술하고 5장에서 결론 및 향후 연구과제를 제시하였다.

2. 관련 연구

2.1 유해사이트 분류 알고리즘

유해사이트 분류방식으로는 웹페이지의 텍스트에 포함된 유해단어를 검색하여 분석하는 텍스트 기반의 분류방식과 Skin-Color Detection 알고리즘을 이용한 이미지 기반의 분류방식이 있다.

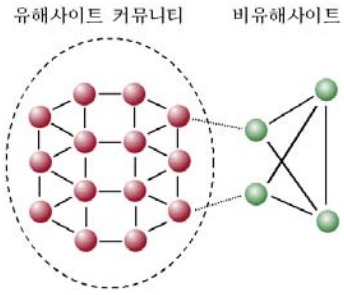
텍스트 기반의 분류방식은 미리 정의된 유해단어 사전의 키워드와 유해사이트에서 추출된 텍스트를 비교하여 유해단어를 검색해 내는 방식이다. 유해단어사전에 등록된 키워드에는 유해정도에 따라 Weight 값을 주고 각각의 키워드가 검색되는 빈도에 비례하여 사이트의 유해정도를 판단하게 된다. 이러한 문서분류 알고리즘은 시간과 비용 측면에서 효율적이나 추출된 텍스트(단어)만을 분석하므로 성상담 사이트 혹은 성교육 사이트까지 유해사이트로 분류하는 오분류의 소지를 가지고 있다.

Skin-Color 알고리즘을 이용한 분류방식은 사람의 피부가 컬러 성분 중 Red 성분이 많은 비율을 차지

하는 것을 이용하는 것으로, 색상 공간 변환을 거친 이미지의 RGB 성분 비율 중 각각의 비율이 특정 임계치 사이에 존재하면 살색으로 간주하고 그렇지 않으면 살색이 아닌 것으로 판별한다[2]. 이 방식은 이미지 데이터가 많은 유해사이트의 특성상 분류에 유리한 장점이 있지만 누드예술이나 아기 돌사진을 유해정보로 판단하는 단점이 있다.

2.2 웹 구조 분석

A.Broder는 웹 구조 분석을 위한 그의 연구에서 월드와이드웹이 강력하게 연결된 코어(Strongly Connected Component : SCC)를 가지고 있으며 코어 내부에서는 모든 페이지들이 하이퍼링크를 통해 직접 연결되어 있음을 검색로봇을 이용한 실험을 통해 입증하였다[3]. G.W.Flake는 그의 연구에서 월드와이드웹은 스스로 유기적인 체계를 조직하는 특성을 가지며 (Self-Organization) 하이퍼링크를 통해 같은 주제를 공유하는 커뮤니티를 형성하게 된다고 지적하였다. 그는 또한 수많은 커뮤니티를 구분하는 방법의 하나로 구성요소들 간의 링크의 밀도(Link Density)를 제시하였다[4].



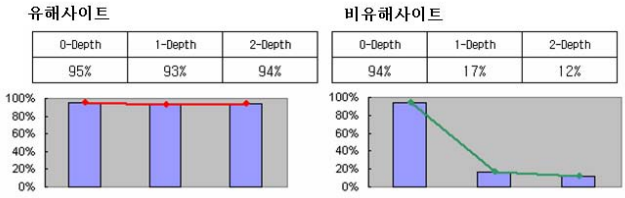
[그림1] 유해사이트 커뮤니티의 링크 밀도

3. 제안 시스템

3.1 시스템 원리

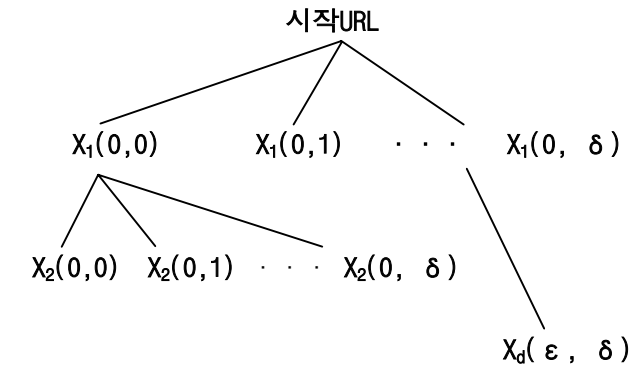
유해·음란사이트의 경우 거의 100% 하이퍼링크를 포함하고 있으며 링크된 사이트들 역시 같은 유해사이트일 확률이 매우 높다. 그러므로 링크된 사이트들의 유해지수가 높게 나타나며 이것은 1-Depth, 2-Depth로 따라가더라도 마찬가지이다.

이에 비해 오분류된 비유해사이트의 경우, 비록 시작 사이트 자체는 유해지수가 높게 나오더라도 링크된 사이트는 유해지수가 낮게 나타나는 것이 보통이다. 따라서 1-Depth, 2-Depth로 들어갈 때 유해지수가 급격히 낮아지면 시작 사이트(0-Depth)는 유해사이트가 아닐 확률이 높다.



[그림2] 유해사이트 검증 원리

3.2 유해성 검증방식



[기호설명]

- d = Depth
- ε = d-Depth의 ε 번째 그룹
- δ = ε 번째 그룹의 δ 번째 링크
- $X_d(\varepsilon, \delta)$ = d-Depth의 ε 번째 그룹 내 δ 번째 링크의 유해지수

$n_d(\varepsilon)$ = d-Depth의 ε 번째 그룹의 링크정보 개수, p = 임의로 결정한 링크정보의 개수라고 하면 d-Depth의 ε 번째 그룹에서 선택된 링크정보의 개수 $Y_{p,d}(\varepsilon, \delta) = \min(p, n_d(\varepsilon))$가 된다. 또한 d-Depth의 ε 번째 그룹 내의 선택된 링크의 유해지수는 $X_{d,j}(\varepsilon, \delta)$, j=1, · · · $Y_{p,d}(\varepsilon, \delta)$로 표현된다. 이럴 때 한 Depth 전체의 유해지수의 평균 h_d를 구하는 식은 다음과 같다.

$$h_d = \frac{\sum_{k=1}^{\varepsilon} \sum_{j=1}^{Y_{p,d}(\varepsilon,\delta)} X_{d,j}(k,\delta)}{\sum_{k=1}^{\varepsilon} Y_{p,d}(k,\delta)} \quad where$$

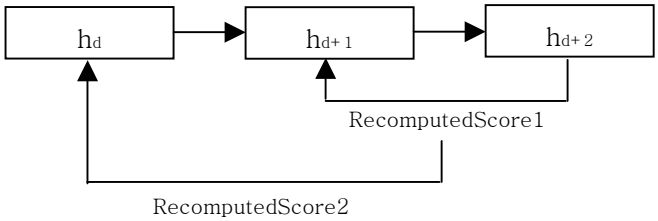
$d,\ p,\ n_d(\varepsilon) \in N$ 이고 $p \neq 1$

위와 같은 방식으로 산출된 유해지수 h_d, h_{d+1}, h_{d+2}, …, h_{d+n}에서 계산의 효율성을 위하여 d-Depth의 하위 2개 Depth의 유해지수 h_{d+1}, h_{d+2}를 이용해 d-Depth 사이트의 유해성을 검증한다.

① h_{d+1}과 h_{d+2}을 곱한 값을 RecomputedScore1이라 한다.

② RecomputedScore1과 h_d 을 곱한 값을 RecomputedScore2라 한다.

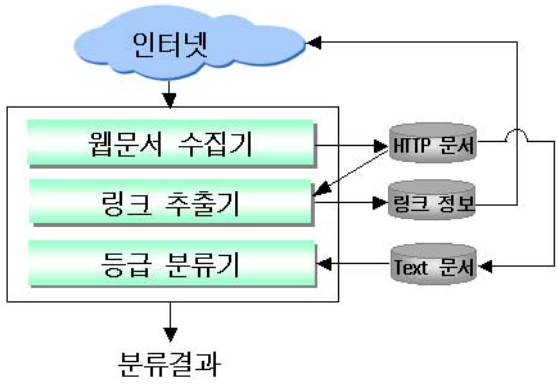
③ h_d 와 RecomputedScore2을 비교하여 RecomputedScore2가 h_d 값에서 50% 이하(h_d /2)로 떨어질 경우 d-Depth사이트는 유해성이 없는 것으로 판정한다.



[그림3] 유해성 검증방식

3.3 시스템 아키텍처

본 논문에서 제안하는 시스템은 웹문서 수집기, 링크 추출기, 등급 분류기로 구성된다.



[그림4] 시스템 아키텍처

웹문서 수집기는 입력된 URL에서 HTTP문서를 수집하여 HTML과 Text형태로 저장하는 모듈이다.

링크 추출기는 웹문서 수집기를 이용해 저장한 HTTP문서에서 <a href=http://*URL*> 태그에 포함된 링크를 추출하는 모듈이다. 이렇게 추출된 링크를 이용해 연결된 사이트의 유해등급을 평가하게 된다. 효율적인 실험을 위해 추출된 링크들 중 일정개수를 무작위로 선택하여 조사한다.

등급분류기는 기존의 유해사이트 분류 알고리즘 중에서 비교적 수행속도가 빠르고 정확한 텍스트 기반의 분류방식을 사용하였다. HTTP문서에서 텍스트를 추출한 후 유해단어DB(사전)와 비교해 해당사이트의 유해정도를 판단한다. 유해등급은 품사별로 미리 정의한 범주와 등급별 색인어의 존재 빈도에 의하여 점수를 매겨 판정이 이루어진다. 유해지수가 높으면 해당사이트를 유해사이트로 분류한다.

[표1] 유해단어사전 샘플

단어	가중치	단어	가중치
eros	60	erotic	70
porno	90	adult	60
nude	70	nudity	50
topless	80	bottomless	80
orgasm	70	virgin	70
penis	70	dildo	90

4. 실험 및 평가

4.1 실험환경 및 평가방법

본 시스템의 구현을 위해 Dual CPU 1.2GHz, Main Memory 1 Giga byte의 Linux 운영체제 위에서, 웹문서 수집기, 링크 추출기, 그리고 등급 분류기는 JAVA(J2SE 1.4.2)로 구현하였다.

웹문서 수집기의 출발정보 URL은 Altavista 검색엔진(htpp://www.altavista.com)에서 "sex"라는 키워드로 검색하여 나온 URL 리스트 100개를 사용하였다.

등급분류시 사이트의 메인페이지와 Sub 페이지에서 추출된 텍스트를 기반으로 유해지수를 평가하였다. 그리고 시작사이트(0-Depth)의 유해지수를 평가한 후 추출된 링크정보 중 무작위로 5개를 골라 1-Depth 사이트의 유해지수 평균을 계산하였다. 다음 단계로 다시 각각의 1-Depth 사이트의 링크 5개를 골라 2-Depth 사이트의 유해지수 평균을 계산하였다.

실험결과 평가방법으로는 얼마나 정확히 분류되었는지 성능을 측정하기 위하여 등급분류기의 분류결과와 링크정보를 이용한 분류결과를 아래와 같은 수식으로, 정확도(Precision)와 재현율(Recall) 그리고 F-Measure 측정식을 이용하여 평가하였다[5].

$$\text{Precision} = \frac{\text{Categories assigned by the system and correct}}{\text{Total Categories assigned}}$$

$$\text{Recall} = \frac{\text{Categories assigned by the system and correct}}{\text{Total Categories correct}}$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

4.2 실험결과

실험결과를 보면 등급분류기만을 이용하여 분류한 결과는 86%의 정확도를 보였고 링크정보를 이용해 분류한 결과는 96%의 정확도를 보여 등급분류에 링크정보를 이용하는 것이 분류의 신뢰도 향상에 크게 기여함을 알 수 있다. 전체적인 성능비교를 위하여 F-Measure 값을 측정하였으며, 이 또한 링크정보를 이용한 분류가 신뢰도가 높음을 보여주었다.

[표2] 성능 측정 결과표

	유해등급기 분류결과	링크정보 검증결과
RI	94	89
RL	81	86
NRD	5	10
Recall	0.931	0.989
Precision	0.862	0.966
F-measure	0.895	0.977

T1 : 전체 Data(100개의 사이트)

TL : 전체 Data에서 실제로 유해하다고 판단된 Data(87개의 사이트)

RI : 전체 Data에서 구현한 시스템이 유해하다고 분류한 Data

RL : 구현한 시스템이 유해하다고 분류한 Data 중 실제로 유해한 Data

NRD : 구현한 시스템이 유해하지 않다고 분류한 Data 중 실제로 유해하지 않은 Data

5. 결론 및 향후 연구

본 논문은 유해사이트 컨텐츠에 내재된 하이퍼링크의 연관성을 이용하여 기존 유해사이트 분류의 정확도를 약 10% 높여 신뢰도를 향상시킬 수 있다는 것을 증명하였다.

그러나 본 논문에서 구현한 알고리즘은 분류된 사이트의 링크 사이트를 일일이 접속하여 유해성을 검증하는 것으로 시간과 비용 측면에서 비효율적인 단점이 있다. 그러므로 차후 이 알고리즘이 효율성을 갖기 위해 시간과 비용을 절감하는 방법을 연구하는 노력이 필요하리라 생각한다.

6. 참고문헌

[1] 정보통신윤리위원회, "2003년 정보통신윤리통계", 2004년 1월

[2] P.Peer, F.Solina, "An Automatic Human Face Detection Method", Proc.4th Computer Vision Winter Workshop, pp. 122~130, 1999

[3] A. Broder, "Graph Structure in the Web", In Proceeding of the 9th International World Wide Web Conference(WWW9), pp. 309-320, 2000

[4] G. W. Flake, S. Lawrence, C. L. Giles, F. Coetzee, "Self-Organization and Identification of Web Communities", IEEE Computer, pp. 66-71, March 2002

[5] Bekkerman R., El-Yaniv R., Tkshby N., Winter Y., "On Feature Distributional Clustering for Text Categorization", Proc. SIGIR 2001, SIGIR Conference, pp.146~153, 2001