유전 알고리즘을 이용한 음란사이트 식별

한수경
강릉대학교 컴퓨터공학과
skhan@kangnung.ac.kr

Genetic Algorithm for Lewdness Web Site Detection

SooKyung Han
Dept. of Computer Science & Engineering, Kangnung National University

요 약

오늘날 인터넷은 의식주와 더불어 삶에 유용한 다양한 정보를 제공하는 생활 필수품이다. 의식주가 인간의 육체적인 건강을 담당한다면, 인터넷은 정신적인 삶의 질을 담당한다. 그런데 음란사이트는 아직 정신적으로 미숙한 청소년들에게 선별 없이 개방되고 쉽게 노출될 수 있다. 이 논문에서는 웹사이트의 문서가 음란 문서인지, 비음란 문서인지를 바르게 판정하기 위하여 유전 알고리즘을 이용하여 단어에 가중치를 배정하는 문제에 대하여 연구한다. 실험 결과 이렇게 배정된 가중치를 이용하여 평균 93.84%의 인식률로 음란 문서와 비음란 문서를 식별할 수 있었다. 여기서 문서의 음란여부를 판정하기 위하여 가중치를 배정하는 단어는 Zipf's law에 기반 하여 선정하였다.

1. 연구배경

인터넷의 성장으로 삶의 질은 높아졌지만 인터넷 유해 음란정보의 유통에 쉽게 노출되어 청소년들은 원치 않는 음란사이트의 공격을 받을 수 있다. 또한 인터넷을 통하여 음란물을 주고받는 행위는 법에 저촉되지만 사이버 공간을 현실세계의 법으로 규제하는 것은 한계가 있다. 하물며 외국 음란사이트는 국내법을 적용할 방법이 없으므로 인터넷의 국제적 성장 추세를 감안할 때, 외국 음란사이트의 증가는 문제를 더욱 심각히 가중시킨다. 본 논문은 영문 웹사이트가 음란사이트인지 아닌지를 판정할 수 있는 방법을 제안한다.

검색엔진에 몇 개의 질의어를 주고 사이트의 음란 여부를 판정하는 경우, 제한된 양의 질의어로 방대한 양의 사이트를 찾을 수 있지만 질의어의 추가 여부에 따라 검색되는 사이트의 성격이 달라지는 양상을 보인다. 즉 음란, 음란오인, 일반 문서가 규칙 없이 검색되므로, 동일한 단어지만 의미 혹은 사용 목적이 전혀 다른 경우에는 사용자가 원하는 결과와 전혀 상반되는 결과를 얻게되는 문제점이 있다. 예를 들어 "sex"라는 단어는 성에 관한 일반적인 문서에도 자주 등장하므로 이 단어는 문서의 음란 여부를 판정하는데 그다지 중요한 단어로 볼 수는 없다. 따라서 어떤 단어들에 대하여 그것이 문서의 음란 여부를 판정하는데 얼마나 중요한 것인가를 나타내는 가중치를 배정하여 음란 문서를 식별할 수 있을 것이다. 이러한 생각을 기반으로 하여 정보검색의 기술을 이용한 관련 연구들이 있으나 그러한 연구들은 단어의 가중치를 배정하기는 하지만, 그것을 통하여 계산한 유사도 값으로 문서의 음란 여부를 판정할 기준이 되는 임계값threshold을 정하는 기준이 없으므로 신뢰성이 떨어진다. 본 논문에서는 유전 알고리즘을 이용하여 단어의 가중치 및 임계값을 동시에 생성하는 방법을 사용하여 이러한 문제점을 해결하였다.

본 논문은 2004년도 두뇌한국21사업에 의하여 지원되었음.

본 논문의 구성은 다음과 같다. 2장에서는 가중치를 배정할 단어의 선정에 대하여 살펴보고, 3장에서는 가중치 배정을 위한 유전 알고리즘에 대하여 설명하며, 4장에서는 실험을 통하여 결과를 분석하고, 마지막으로 5장에서 결론을 맺는다.

2. 가중치를 배정할 단어의 선정

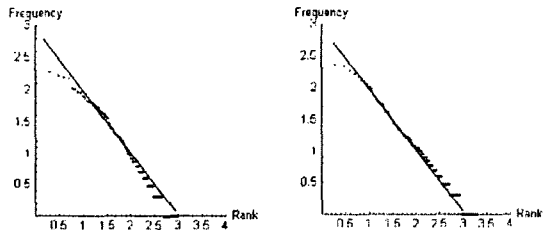
임의의 웹 문서의 음란 여부를 판정하려면 문서에 포함된 단어의 가중치를 산정하고, 배정된 가중치를 이용하여 판정값을 계산한 후, 임계값을 기준으로 음란여부를 판정한다(3.1절 참조). 우선 문서의 음란 여부를 판정하는데 사용할 가중치를 부여할 단어를 선정해야 하는데 본 연구에서는 Zipf's Law에 기반 하여 선정하였다. Zipf's Law는 문서에 나오는 단어들을 모두 세어 단어의 빈도 수와 순위의 관계를 조사한 것으로 모든 문서는 소수의 상위 순위 단어들이 집중적으로 사용되며, 순위가 내려갈수록 사용빈도수가 기하급수적으로 떨어진다는 법칙이다. 사용 빈도수를 y, 순위를 r이라 하면, 다음의 관계식을 만족한다.

$$y = cr^{-b} \qquad \text{(식1)}$$

이때 c와 b는 상수이고, b는 근사적으로 1을 갖는다. 이 식의 log-log 그래프는 -b를 기울기로 갖는 직선의 그래프이다.

웹 문서에서 어휘분석을 하여 웹 태그를 제거한 후 단어의 분포에 대한 그래프를 그려보면 음란 문서나 음란오인 문서, 일반 문서 모두 Zipf's Law를 따르고<그림 2.1>, 각각의 문서에서 스테밍(불용어 제거)한 결과도 Zipf's Law을 따른다. Zipf's Law를 따르는 시스템은 특정한 몇몇 개체에 대부분의 숫자가 몰려 전체에 영향을 미치고, 대다수를 차지하는 나머지의 역할은 미약하다는 공통된 특징을 가지고 있다. 즉 빈도수가 높은 단어가 그 문서의 특징을 반영한다. 따라서 Zipf's law

의 CDF분포와 PDF분포의 상관관계[1]에 의하여, 음란문서 10개와 음란오인 문서 10개의 상위 35%를 추출하여 105개의 단어를 선정하였다. 결국 본 연구에서 단어의 가중치 배정 문제는 경우의 수가 2^{105}인 최적화 문제이므로 유전 알고리즘이 매우 효율적으로 응용될 수 있다[2].



(a) (b)

<그림 2.1> (a)음란 문서 (b)음란오인 문서에서 단어의 도수분포를 log-log scale로 그린 그래프

3. 가중치 산정을 위한 GA

유전 알고리즘은 자연선택과 유전자에 기초를 둔 일종의 탐색 알고리즘이다. 이것은 탐색공간에 대한 어떠한 지식도 사용하지 않으면서도 알고리즘이 간단하고, 강하며(robust), 또한 일반적이기 때문에 많은 분야에서 응용되고 있다. 자연은 제한된 자원에 대한 개체간의 경쟁시에 적응성이 강한 개체가 더 많이 살아남을 수 있게 하는 것으로 알려져 있다. 또한, 그들의 상대적인 우수성은 유전자로 특징 지워지며 이들은 다음 세대에 전달되어 결국 우수한 개체의 우수한 특성이 세대를 건너가며 유지되는 것이다.

일반적으로 단순 유전 알고리즘은 한 개의 모집단을 구성하는 것으로 시작된다. 여기서 모집단내의 개체는 문제에 대한 잠정적인 해를 나타내는 것이며, 초기 모집단 내의 개체는 임의로 정한다. 초기 모집단이 생성되면 각 개체를 평가하여 그들의 상대적 적합도를 계산하고, 그 값에 근거하여 다음 세대에 자손을 남길(즉, 자신의 유전자를 다음 세대에 넘길) 개체를 선택한다. 선택된 개체들은 자신의 유전자를 모두 전달하거나, 교차를 통하여 다른 개체와 부분적으로 혼합된 유전자를 전달하거나, 돌연변이를 통하여 자신의 것과 다른 유전자를 포함하도록 전달한다. 그렇게 만들어진 새로운 개체들로 구성된 모집단이 다음 세대를 형성하는 것이다. 이러한 진화 과정을 되풀이하면 모집단의 개체들은 적합도가 높은 쪽으로 수렴하게 된다. 그리고 최종적으로 가장 적합도가 높은 개체를 최적해로 사용한다.

3.1 가중치 배정을 위한 개체의 표현

모집단내의 각 개체는 선정된 단어와 문서의 음란 여부를 판정할 임계값을 나타내는 것으로 각각 106개의 유전자를 가지며, 105개의 유전자는 해당 단어의 가중치를 나타내고, 마지막 106번째 유전자는 음란 문서 여부를 판정할 임계값을 갖는데, 이 값은 가중치와 함께 진화하는 실수 값이다. 결국 한 개체는 다음과 같은 실수 벡터의 형태로 나타낸다<그림 3.1>.

1	2	3	4		i		105	106
0.92	0.83	-0.1	0.38	···	w_i	···	-0.9	t_v

<그림 3.1> 개체표현

이때 w_i는 i번째 유전자의 가중치로 범위 $-1 \le w_i \le 1$의 실수값, t_v는 임계값을 나타낸다.

선정된 단어가 특정 웹 문서에 포함되는지의 여부는 <그림 3.2>와 같이 배열로 표현한다.

1	2	3	4		i		105
1	0	0	1	···	e_i	···	0

<그림 3.2> 추출단어의 문서포함여부에 대한 표현

여기서 e_i는 i번째 단어에 대한 값으로 다음과 같다.

$$e_i = \begin{cases} 1 & , \text{단어} i\text{가 웹 문서에 포함} \\ 0 & , \text{단어} i\text{가 웹 문서에 포함 되지 않음} \end{cases}$$

특정 문서의 음란 여부는 다음과 같은 판정값 det_v를 사용하여 결정한다.

$$det_v = \frac{\sum_{i=1}^{n} w_i\, e_i}{n} \qquad \text{식(2)}$$

여기서 $n=105$이고, det_v와 t_v의 크기 비교로 주어진 문서가 음란 문서인지 아닌지를 다음과 같이 판정한다.

$$\text{문서} = \begin{cases} \text{음란 문서} & , det_v \ge t_v \\ \text{비음란 문서} & , det_v < t_v \end{cases} \qquad \text{식(3)}$$

3.2. 적합도 및 유전 연산자

각 개체의 적합도는 단어에 배정된 가중치와 임계값을 이용하여 웹 문서의 음란 여부를 정확하게 판정하는지의 여부로 산정한다. 즉, 진화에 사용되는 문서들에 대하여 음란 문서를 음란 문서로, 음란오인 문서와 일반 문서를 비음란 문서로 바르게 판정한 개수의 합을 적합도로 한다.

유전 알고리즘을 적용할 때 사용하는 진화 연산자 즉, 선택 연산자 및 교차와 돌연변이로 대표되는 유전 연산자의 종류는 매우 다양하며, 각각의 방법은 나름대로의 특징을 갖는다[3]. 본 연구에서는 가장 일반적으로 사용되는 적합도 비례 선택을 사용하였다. 또한, 교차는 일점 교차(one point crossover)를 사용하였고, 돌연변이는 임의의 유전자 값을 임의로 변경하였다.

4. 실험 및 분석

음란문서를 식별하기 위하여 선정된 단어에 가중치를 부여하는 문제는 유일한 해를 갖는 것이 아니다. 즉, 유전 알고리즘을 적용하여 학습에 사용한 문서를 모두 바르게 식별할 수 있도록 하는 단어의 가중치 배정은 여러 개가 존재할 수 있다. 따라서 이 논문에서는 초기 모집단 구성에 사용하는 랜덤넘버의 seed값을 달리하여, 학습에 사용한 문서를 모두 바르게 식별하는, 10가지의 가중치 및 임계값 배정을 구하여 평균적인 인식성능을 평가하였다. 학습에 사용한 문서는 음란 문서 20개, 음란오인 문서 10개 및 일반 문서 10개를 사용하였으며, 모든 경우에 모집단의 크기는 50, 교차율은 0.7, 돌연변이율은 0.05, 최대 세대수는 10000으로 하였다. 학습에 사용한 문서를 모두 바르게 식별할 수 있더라도 다른 문서들을 항상 바르게 식별 할 수 있는 것도 아니다. 따라서 선정된 단어에 배정된 가중치 및 임계값에 의한 문서의 인식

률을 평가하기 위하여 음란문서 7개, 음란오인 문서 20개, 일반 문서 6개를 추가하여 총 73개의 문서에 대한 인식률을 평가 분석하였다. 본 논문에서는 음란문서는 아니지만 음란문서로 오인되기 용이한 음란오인 문서를 바르게 식별하는가를 중요하게 고려하여 20개의 음란오인 문서를 평가에 사용하였다.

다음 <표 4.1>은 초기Seed값을 2000부터 92000까지 10000씩 달리하여 10번의 실험으로 얻은 임계값과 그 임계값으로 판정한 문서의 인식율이다. 이 결과를 보면 학습하지 않은 문서에 대해서도 평균 91.51% 이상의 높은 인식율을 보임을 알 수 있다.

<표 4.1> 음란 문서 인식율

임계값	합계		인식율	
	학습문서 포함	학습문서 제외	학습문서 포함	학습문서 제외
0.01	68	48	93.15%	90.57%
0.026	67	47	91.78%	88.68%
0.03	67	47	91.78%	88.68%
0.022	67	47	91.78%	88.68%
0.018	69	49	94.52%	92.45%
0.02	72	52	98.63%	98.11%
0.034	70	50	95.89%	94.34%
0.012	67	47	91.78%	88.68%
0.02	71	51	97.26%	96.23%
0.01	67	47	91.78%	88.68%
평균	68.5	48.5	93.84%	91.51%

다음으로 음란, 음란오인, 일반 문서를 유형별로 바르게 인식한 문서 수와 평균 인식율을 조사하였다<표 4.2>. 이 경우 음란 및 일반 문서에 비하여 음란오인 문서의 인식율이 상대적으로 떨어지는 것을 볼 수 있다. 이런 경우는 실험에서 웹 문서에 포함된 단어의 개수가 상대적으로 적은 경우, 성 문제를 다루는 성인 사이트의 웹 문서를 음란문서로 오인식하는 경우들 때문임을 알 수 있었다.

<표 4.2> 유형별로 바르게 인식한 문서 수와 평균 인식율

임계값	유형별 바르게 인식한 문서수		
	음란(27)*	음란오인(30)	일반(16)
0.01	26	27	15
0.026	24	28	15
0.03	26	26	15
0.022	26	26	15
0.018	26	27	16
0.02	26	30	16
0.034	25	30	15
0.012	27	24	16
0.02	26	29	16
0.01	26	27	14
평균	25.9	27.6	15.4
유형별 평균 인식율	95.56%	91.33%	95.63%

*()는 문서수

다음 <표 4.3>은 유전 알고리즘으로 생성한 단어의 가중치 및 임계값을 나타내는 한 개체의 예이다. 이때 생성된 임계값은 0.02이다. <그림 3.1>에서 단어의 가중치 w_i를 양수로 택하면, 음란 문서에 많이 포함되는 단어를 갖는 음란오인 문서를 바르게 인식하지 못한다. 예를 들어, 화가 피카소에 대한 한 사이트의 경우 음란 사이트가 아님에도 가중치의 범위를 양수로 하는 경우 바르게 식별하지 못한다. 이와 같은 오인식을 막기 위하여 본 연구에서는 가중치 w_i의 범위를 $-1 \le w_i \le 1$로 하였다.

5. 결론

본 논문에서는 웹 사이트 문서의 음란 여부를 바르게 판정하기 위하여 웹 문서에 포함된 단어를 Zipf's law에 기반을 두어 선정하고, 유전 알고리즘을 이용하여 개체의 유전자인 단어의 가중치 및 임계값을 함께 진화시킨 후, 그렇게 배정된 값으로 임의의 웹 문서가 음란문서인지 아닌지를 판정할 수 있는 방법을 제안하였다. 실험결과 웹 문서에 대한 음란여부를 판정하면 평균 93.84%의 높은 인식율을 보였다. 그러나 인터넷 사이트의 웹 문서에 포함된 단어의 개수가 상당히 적은 경우, 음란오인 문서 중에서 음란이 아닌 성인 사이트의 웹 문서를 음란문서로 오인식하는 비율이 다른 유형의 문서에 대한 오인식율 보다 높다. 향후 과제로 단어의 빈도수와 단어간의 상관관계를 분석하여 보통의 성인 사이트를 음란사이트로 오인식하는 비율을 낮추는 연구를 진행 중에 있다. 또한 한글 음란 사이트는 당연히 그 이용자의 절대 다수가 한국인이므로 앞으로 한글 웹 문서에 대한 연구가 이루어져야 할 것으로 본다. 더 나아가 인터넷 유해 음란정보의 유통 문제에 대한 대응방식으로 웹 상에서 음란사이트를 차단하는 시스템을 개발해야 할 것으로 본다.

<표 4.3> 단어의 가중치 리스트(스테밍한 단어[4])

credit	0.954	penetr	0.51	girl	-0.274
adult	0.95	real	0.486	health	-0.296
pictur	0.932	straight	0.442	realiti	-0.312
porn	0.924	director	0.434	free	-0.32
comment	0.918	lesbian	0.408	earli	-0.348
fetish	0.91	cancer	0.36	famili	-0.374
gai	0.88	brunett	0.35	big	-0.466
babi	0.868	reproduct	0.29	symptom	-0.482
ten	0.852	colleg	0.256	research	-0.55
dvd	0.85	enter	0.218	women	-0.562
cam	0.844	thumbnail	0.2	disea	-0.578
webcam	0.834	card	0.194	abort	-0.59
show	0.824	ag	0.172	read	-0.598
teacher	0.824	parent	0.162	pai	-0.646
amateur	0.822	scene	0.154	list	-0.66
blond	0.814	movi	0.106	public	-0.676
articl	0.798	milf	0.078	includ	-0.682
hardcor	0.774	unplan	0.076	sexi	-0.714
resourc	0.772	org	0.048	educ	-0.72
categori	0.77	fertil	0.032	consult	-0.736
law	0.754	click	0.028	galleri	-0.752
link	0.74	pussi	0.024	site	-0.774
xxx	0.732	method	-0.01	chart	-0.798
tit	0.668	tool	-0.032	fuck	-0.802
hot	0.656	matur	-0.112	life	-0.814
asian	0.656	infect	-0.148	good	-0.848
pam	0.636	condom	-0.162	jai	-0.884
therapi	0.616	concept	-0.17	pregnanc	-0.916
hairr	0.594	sexual	-0.194	inform	-0.926
babe	0.58	font	-0.218	top	-0.95
teen	0.57	check	-0.224	dick	-0.954
tip	0.558	expert	-0.24	pleasur	-0.956
video	0.55	birth	-0.24	nude	-0.962
contracept	0.532	page	-0.242	rate	-0.966
sex	0.512	control	-0.266	bodi	-0.982

6. 참고 문헌

[1]. R. Gunther, L. Levitin, B. Shapiro, P. Wagner, "Zipf's law and the effect of ranking on probability distributions", International Journal of theoretical Physics, 35(2), pp. 395-417, 1996

[2]. D. E. Goldberg, "Genetic Algorithms in Search, Optimization & Machine Learning, pp. 1-25, 1989.

[3]. Hancock, Selection Methods for Evolutionary Algorithms, in Practical Handbook of Genetic Algorithms, Vol.2, pp. 67-92, 1995

[4]. http://www.tartarus.org/~martin/PorterStemmer/