유해어의 공기정보를 활용한 유해 웹문서 필터링

안형근° 이원휘 안동언 정성종
전북대학교 대학원 컴퓨터공학과
{ahk007°, wony, duan, sjchung}@chonbuk.ac.kr

Harmful Web-document Filtering using Harmful word Co-occurrence

HyungKeun An° Wonhee Lee, Dongun An, Sungjong Chung
Dept. of Computer Engineering, Chonbuk National University

요 약

웹 환경이 일반화되고 웹을 통해 획득할 수 있는 정보가 다양하고 풍부하다. 이 다양하고 풍부한 정보는 유익한 정보 뿐만 아니라 청소년들을 비롯한 사회적으로 보호를 받아야 할 웹 이용자들의 정신건강을 해치는 정보들도 다수 포함되고 있어 사회적 문제가 되고 있다. 본 연구에서는 웹 문서를 필터링하는 수단으로 공기정보를 포함하고 있는 유해어 사전을 활용한다. 유해어 사전 구축은 단순히 유해어 리스트만으로 사전을 구축하지 않고, 유해어 주위의 공기 단어의 정보를 포함시킴으로써 유해어의 중의성에 의한 오분류를 해소하고자 하였다. 즉, 유해어 후보가 1개 이상의 의미를 가지며 각 의미가 유해 정도가 다를 때, 유해어 후보의 등급을 결정하기 위하여 해당 유해어와 같은 문장 혹은 같은 문서에 출현하는 다른 단어 정보를 활용한다. 이렇게 함으로써 문서의 유해 등급을 결정하게 된다.

1. 서 론

오늘날의 정보환경은 웹(World Wide Web)이 대중화되면서 웹은 다양한 정보의 제공과 습득의 장이 되고 있다. 그러나 이러한 웹의 편리성은 그 이면에 무분별한 정보의 제공으로 인한 여러 가지 문제를 안고 있다. 너무 많은 정보의 제공으로 인한 정보 검색의 부담과 무분별한 유해 정보의 범람은 정보사회를 살고 있는 우리에게 커다란 문제가 되고 있다. 특히 음란, 폭력, 자살 등의 유해 정보는 사회적으로 보호를 받아야 할 청소년들을 비롯한 판단력과 절제력이 부족한 인터넷 이용자들에게 심각한 사회적 문제를 야기하고 있다.

따라서, 이러한 문제를 해결하기 위한 제도 및 연구가 다양한 방법으로 이루어지고 있다. 게시자(publisher)의 자발적 등급 결정에 기반한 인터넷 내용 선택에 대한 플랫폼(PICS)[3], 제공되는 영상정보의 스킨컬러(skin color)에 기반한 연구[2], 유해한 단어나 구에 기반하여 필터링하는 키워드 필터링, 신경망 이론을 응용한 지능적 내용 분류 연구, 이미지 정보를 이용한 연구[7, 8] 등이 그것이다. 그러나 이러한 연구나 제도들이 가지는 한계로 인하여 극히 저조한 성능을 보이고 있다[1].

본 연구에서는 웹 컨텐츠의 텍스트 정보를 이용한 필터링을 구현하였다. 본 논문은 유해 문서를 분류하기 위한 방법으로 단순한 유해어 리스트만을 포함하고 있는 유해어 사전이 아닌 공기정보가 포함된 유해어 사전을 활용함으로써 분류의 정확율을 높이고자 하였다.

본 논문의 구성은 2장에서 웹 컨텐츠 분류를 위한 다양한 기존 연구들에 대하여 살펴보고 정리한다. 3장에서는 제안한 시스템의 알고리즘 및 구현을 다루고, 4장에서 실험 및 평가를 하고 5장에서 결론을 맺도록 한다.

2. 관련 연구

웹 컨텐츠 필터링 연구는 크게 인터넷 내용 선택에 대한 플랫폼, URL 차단, 키워드 필터링, 인공지능 내용 분석, 이미지 기반 필터링으로 분류할 수 있다[1, 8].

2.1. 인터넷 내용 선택에 대한 플랫폼(PICS)

PICS(Platform for Internet Content Selection)은 웹 페이지의 내용에 관한 정보가 기술된 메타 정보를 컴퓨터의 소프트웨어를 통해 인식하고 선별할 수 있는 기술규격이다. PICS는 주로 부모나 교사 등이 미성년자의 인터넷 접속을 지도하고 통제하기 위한 자녀 통제 장치에 이용되어 왔다. 일반적인 PICS 내용 등급 시스템은 RSACi와 SafeSurf가 있다. RSACi(Recreational Software Advisory Council)는 거친 언어(harsh language), 신체노출(nudity), 성행위(sex), 폭력(violence)의 네 가지 카테고리를 사용하며, 각각의 카테고리는 다시 유해정도를 나타내는 0(무해)에서 4까지 5개의 등급으로 분류된다. SafeSurf는 좀 더 상세한 내용 등급 시스템이다. SafeSurf는 나이 그룹에 대한 웹 컨텐츠의 유해성을 묘사하기 위하여 11개의 카테고리를 사용한다[1, 3, 6, 7].

2.2. 키워드 필터링

이 방법은 컨텐츠 내의 유해한 단어나 구의 발생에 기초해서 웹 컨텐츠를 차단한다. 웹 페이지에서 검색된 단어나 어구는 금지된 단어와 어구로 이루어진 키워드 사전상의 것들과 비교되고 일정한 한계 이상의 일치가 발생하는 경우 차단이 이루어진다.

이 내용 분석 방법은 만약 웹 페이지가 잠재적으로 유해한 내용을 가지고 있다면 빠르게 결정할 수 있다. 그렇지만 단어나 어구의 중의성 등에 의해 유해한 내용을

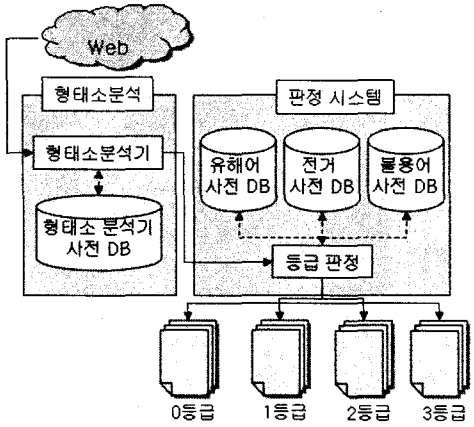
포함하지 않는 웹 컨텐츠까지 차단하는(overblocking) 단점을 지니고 있다[1, 8].

2.3. 지능 내용 분석

웹 필터링 시스템은 자동적인 웹 컨텐츠의 분류를 위해 지능 내용 분석을 이용할 수 있다. 이러한 것 중의 하나가 트레이닝 케이스에 따라 적용되고 학습할 수 있는 신경망(artificial neural networks)이다. 이런 학습과 적응 과정은 포르노와 비-포르노 웹 페이지에 다양하게 나타나는 "sex"처럼 문맥 의존적 단어에 의미를 줄 수 있다. 분류의 높은 정확도를 이루기 위해 다양한 학습 이론이 활용되고 있다. 하지만 이 방법은 학습하는데 시간이 많이 소요된다는 단점을 가지고 있다[7, 8, 9].

3. 시스템의 설계 및 구현

본 시스템은 입력된 문서의 대표 유해어를 추출하고 추출된 대표 유해어의 중의성을 파악하기 위해 해당 대표 유해어 주위의 정보가 같이 저장되어 있는 유해어 사전을 활용하여 좀 더 명확한 판정을 하고자 한다. 시스템의 전체 구성도는 다음 그림과 같다.



[그림 1] 시스템 구성도

3.1. 유해어 필터링(Harmful word filtering)

유해어 필터링에서는 기본적으로 키워드 필터링을 수행한다. 이 필터링을 수행하기 위하여 유해어 사전(harmful word dictionary)과 전거사전(authority word dictionary), 불용어 사전(stop word dictionary)을 구축하고 활용한다. 유해어 사전은 단순히 유해 단어의 리스트만으로 사전을 구성하게 되면 기존의 키워드 필터링에서 발생되는 과도한 분류 문제(over-blocking)를 야기할 위험이 높다. 따라서 본 시스템에서 사용되는 유해어 사전은 단순히 유해한 단어의 리스트뿐만 아니라 유해어 주변 단어 정보를 추가하였다. 추가된 주변 단어는 유해어의 문맥 정보를 반영하여 중의성으로 인한 오분류율(잘못 분류하는 비율)을 낮추는 역할을 한다.

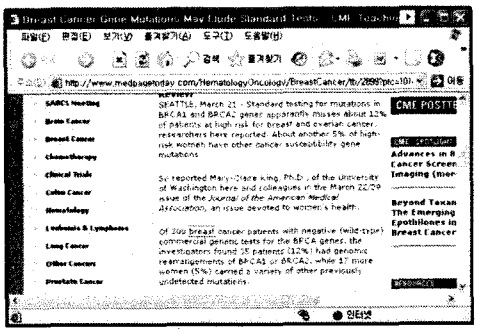
3.2. 인접어와 비인접어

유해어 사전은 유해어 후보와 해당 유해어 후보의 인접어와 비인접어로 구성된다.

인접어란 유해어 후보와 동일한 문장에 출현하여 해당 유해어 후보의 유해여부나 유해정도를 결정지을 수 있도록 하는 단어를 의미한다. 예를 들어 "breast"이라는 단어는 유해할 수도 있고, 무해할 수도 있다. 또한 이 단어가 유해다면, 유해정도가 심하거나 미미하기도 할 수 있다. 예를 들어, "breast"와 같은 문장에 "cancer"라는 단어가 출현한다면 해당 문서는 무해할 가능성이 높다. 반면, 같은 문장에 "tongue"라는 단어가 출현한다면 해당 문서가 유해할 가능성이 높다. 또한, "rope"와 같은 단어가 동일 문장에 출현한다면 이 문서는 변태적인 내용을 포함하는 유해정도가 심한 문서일 가능성이 높다.

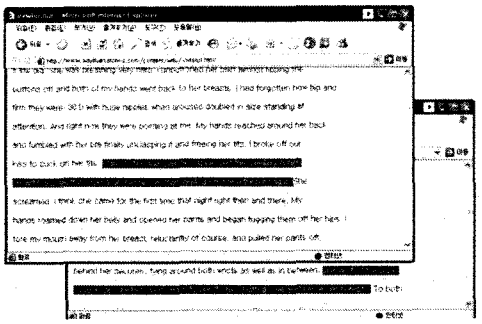
비인접어는 유해어 후보와 동일한 문장에는 출현하지 않고 동일한 문서에 출현하여 해당 유해어 후보의 유해여부를 판정하거나 유해정도를 결정지을 수 있도록 하는 단어이다. 예를 들어, 앞에서 예로 든 "cancer", "tongue", "rope"들이 "breast"라는 유해어 후보와 동일한 문장에 출현하지는 않지만 동일한 문서에 출현한다면 해당 문서의 유해성을 판정하는데 어느 정도 기여를 할 수 있을 것이다.

이때 주의할 것은 인접어나 비인접어에 포함되는 단어들은 그 자체로는 유해성을 가지지 않는 단어들이라는 것이다. 인접어와 비인접어는 유해성을 가지지 않는 단어들이 유해할 수도 있는 단어들과 만났을 때 해당 유해어 후보의 유해/무해를 결정하거나 유해 정도를 결정지



[그림 2] 무해 문서의 예(건강)

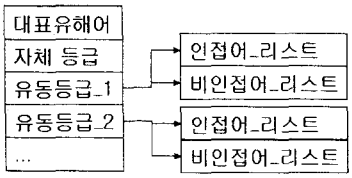
을 수 있는 역할을 하는 단어들이다. 또한 모든 유해어가 인접어와 비인접어 리스트를 갖는 것은 아니다. 유해어 후보 자체가 유해성이 확실하고 유해정도가 확실하다



[그림 3] 유해문서의 예(야설)

면 중의성 제거를 위한 인접어와 비인접어가 필요없기 때문이다.

다음 [그림 4]는 유해어 사전의 한 노드의 구조이다.



[그림 4] 유해어 사전 구조

노드 구조에서 "자체 등급"은 대표 유해어가 독립적으로 쓰였을 때의 등급이다. "유동등급"은 대표 유해어가 특정 인접어 또는 비인접어와 출현했을 때 변동될 수 있는 등급이다. 이 "유동등급"은 0개 이상 존재 할 수 있다. 0개 이상인 이유는 대표 유해어 자체가 유해정도(등급)이 명확할 경우이다. 각 "유동등급"은 인접어 리스트와 비인접어 리스트를 갖는데 이는 대표 유해어를 해당 등급으로 결정짓게 하는데 영향을 미치는 단어의 리스트이다.

3.3. 전거 사전(Authority word dictionary)

표준어가 컨텐츠 제작자의 의도 또는 실수에 의해 변형되어 표기되거나 약자로 표기된 경우 이들을 하나의 표준어로 치환해야 한다. 그렇게 하지 않으면, 약어나 변형어(abbreviated word or metamorphosed word)가 출현 빈도수에 영향을 주어 유해/무해 판정에 영향을 미치기 때문이다.

예를 들어, "breast"라는 용어는 상황 등에 따라 "maracas", "udder", "molehill" 등으로 쓰여질 수 있다. 그러나 이러한 단어들은 유사하거나 같은 의미로 사용된 경우에 이 단어들을 각각의 단어로 카운팅을 하게 되면 빈도수가 분산되어 유해/무해 판단에 영향을 미칠 수 있다. 따라서 이들 단어들을 "breast"로 치환하여 빈도수에 반영해야 정확한 빈도수를 구할 수 있다.

그러나 이 때 주의할 것은 전거사전에서는 사전적인 단어로 표준어를 설정하고 치환하는 것이 아니라 의미적인 단어로 표준어를 설정하고 치환해야 된다는 것이다. 예를 들어 "sex"라는 단어와 "fuck"라는 용어는 유사한 의미를 내포하지만 실제 사용되는 의미에서는 "sex"는 주로 의학적인 의미의 단어로 사용되어 "boff", "bonk" 등과 유사하게 사용된다. 반면, "fuck"의 경우에는 음란한 의미로 사용되어 "coitus", "copulate", "firk" 등과 유사하게 사용된다. 이들을 구분지어 관리할 필요가 있다.

[표 1] 전거 예

대상어		표준어
...	...	...
pussy	=>	vagina
showing	=>	show
blonde	=>	blond
panties	=>	pant
undress	=>	striptease
cock	=>	penis
blowjob	=>	suck
latina	=>	latin
...	=>	...

3.4. 유해 판정 기준

본 연구에서 분류한 문서의 등급은 0등급~3등급의 4개의 등급으로 분류하였다. [정보통신 윤리위원회의 SafeNet 등급기준(http://www.safenet.ne.kr)]의 경우 0~4등급의 5개 등급으로 분류하였으나, 실제에서는 유해 정도의 경계가 모호함을 염두해 4개의 등급으로 분류하였다. 다음 표는 본 연구에서 사용한 등급표이다.

[표 2] 등급 및 판정기준

등급	판정기준
0등급	무해
1등급	학생들을 대상으로 한 성교육 교재 등의 수준
2등급	정상적인 성행위나 일반적 연애소설에 등장하는 성해위 수준
3등급	포르노물 등의 노골적, 변태적 성행위 수준

3.5. 유해어 필터링 알고리즘

유해어 필터링은 다음과 같은 절차로 수행된다.

단계 1: 대상 문서에 대하여 태그 제거와 형태소 분석 작업을 수행한다.

단계 2: 형태소 분석된 단어 리스트를 전거링하여 표준어로 변환한다.

단계 3: 단어 리스트에서 유해어 후보를 찾는다. 찾아진 유해어 후보는 인접어와 비인접어를 이용하여 최종적으로 유해어인지 아닌지를 판단한다. 유해어로 판정되었다면 등급은 몇등급에 해당되는지 결정하게 된다.

단계 4 : 유해어 여부 및 등급이 판정되면 문서 내에 출현하는 각 유해어의 등급을 확인하여 가장 높은 등급을 해당 문서의 등급으로 설정한다.

4. 실험 및 평가

4.1. 실험 데이터

실험에 사용된 데이터는 일반 사이트, 의학 정보 사이트, 성인 사이트 등을 통해 아래의 [표 3]과 같이 각 등급별로 약 4,000개의 문서를 수집하고, 0등급은 다양한 유해 문서에 비해 존재하는 데이터가 많음을 감안하여 약 1,000개의 문서를 더 수집하여 실험에 반영하였다. 분류하기 부적절한 경우, 분류에서 제외하였다.

[표 3] 각 등급별 수집 문서 수

등급	0등급	1등급	2등급	3등급
문서 수	5479	4060	4022	4078

실험은 유해어 가중치만을 이용한 필터링과 공기 정보를 적용한 유해어 사전을 이용한 경우로 분리하여 수행하였

다.

4.1.1. 유해어 가중치만 이용

유해어의 가중치만을 이용하여 문서 분류를 수행하였다.

[표 4] 실험 결과

	0등급		1등급		2등급		3등급	
분류불가	447	8.2%	195	4.8%	6	0.2%	0	0%
0등급	3583	65.4%	583	14.4%	349	8.7%	83	2.0%
1등급	1213	21.1%	2831	69.7%	353	8.8%	28	0.7%
2등급	295	5.4%	468	11.5%	1753	43.6%	725	17.8%
3등급	135	2.5%	46	1.1%	1567	39.0%	3237	79.4%

4.1.2. 공기정보 적용

공기정보를 이용하여 유해어의 가중치를 적용한 결과는 아래와 같다.

[표 5] 실험 결과

	0등급		1등급		2등급		3등급	
분류불가	447	8.2%	195	4.8%	6	0.2%	0	0%
0등급	4086	74.6%	175	4.3%	409	10.2%	67	1.6%
1등급	915	16.7%	3257	80.2%	129	3.2%	27	0.7%
2등급	31	0.6%	424	10.4%	2058	51.1%	346	8.5%
3등급	0	0%	9	0.2%	1420	35.3%	3638	89.2%

4.2 평가

유해어만을 이용했을 경우는 정확율이 평균 64.5%를 보였다. 공기정보를 이용하여 가중치의 변화를 주었을 때는 정확율이 평균 73.8%로 9%정도의 성능 향상을 보였다. 2등급을 제외한 0, 1, 3등급은 평균 80%이상의 정확율, 특히 가장 유해성이 높은 3등급의 분류에서 90%에 가까운 정확율을 보여주고 있다. 2등급의 문서가 3등급으로 많이 분류된 것은 2등급과 3등급의 경계가 모호하고, 따라서 3등급에서 중요도가 높은 단어들이 2등급에서도 사용이 빈번하여 3등급으로 오분류 되었다.

5. 결론 및 향후 과제

본 연구에서는 웹 문서의 등급 분류에 사용되는 유해어 사전에 공기정보를 이용함으로써 분류의 효과를 높이고자 하였다. 결과로 단순히 유해어 리스트에만 의존한 경우보다 약 9% 정도의 성능 향상을 할 수 있었다.

그러나 두 실험 모두에서 2등급 문서를 3등급으로 분류되는 오분류가 많았다. 이는 모호한 경계로 인한 문제로 모호한 경계를 해결할 수 있는 방법이 필요한 것으로 파악되었다. 또한 분류불가의 경우도 발생하였는데 이 경우를 살펴본 결과 문서가 이미지 등으로 이루어져 있거나, 문서 내에서 충분한 텍스트 정보를 얻을 수 없는 경우였다. 이렇게 충분한 텍스트 정보를 얻을 수 없는 경우의 분류 방법이 병행되어야 할 필요가 있다.

참고문헌

[1] Christopher D. Hunter "Internet Filter Effectiveness : Testing Over and Underinclusive Blocking Decisions of Four Popular Filters", Proceedings of the tenth conference on Computers, freedom and privacy: challenging the assumptions, pp287-294, April 2000

[2] Huicheng Zheng, Hongmei Liu, Mohamed Daoudi, "Blocking Objectionable Image : Adult Images and Harmful Symbols", IEEE International Conference on Multimedia and Expo(ICME), pp1223-1226, Jun 2004

[3] Jae-Sun Lee, Young-Hee Jeon, "A Study on the Effective Selective Filtering Technology of Harmful Website Using Internet Content Rating Service", Communication of KIPS Review, VOL. 09, NO. 02, Oct 2002

[4] KwangHyun Kim, JoungMi Choi, JoonHo Lee, "Detecting Harmful Web Documents Based on Web Document Analyses", Communication of KIPS Review, Vol. 12-D No. 5, pp683-688, Oct 2005

[5] M. Hammami, Y.Chahir, and L.Chen, "WebGuard: Web Based Adult Content Detection and Filtering System", IEEE WIC International Conference. Web Intelligence, pp. 574-578, 2003

[6] Mohamed Hammami, Youssef Chahir, and Liming Chen, "WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis", IEEE Transaction On Knowledge and Data Engineering, Vol. 18, No. 2, February 2006

[7] P.Y.Lee, S.C.Hui, "An Intelligent Categorization Engine for Bilingual Web Content Filtering", IEEE Transaction On Multimedia, Vol. 7, No. 6, December 2005

[8] P.Y.Lee, S.C.Hui, and A.C.M. Fong, "Neural Networks for Web Content Filtering", IEEE Intelligent Systems, pp48-57, Sept./Oct. 2002

[9] Qing Yang, Fang-Min Li, "SUPPORT VECTOR MACHINE FOR CUSTOMIZED EMAIL FILTERING BASED ON IMPROVING LATENT SEMANTIC INDEXING", Proceedings of the Fourth International conference on Machine Learning and Cybernetics, Vol. 6, pp3787 - 3791, Aug 2005

[10] Seung-Man Lee, Young-Hun Jang, Jung-Hwan Lim, "Implementation of a Harmful Website's Automatic Classification System based on Morphological Analysis and Skin-Color Distribution's Human Detection

[11] 정규철, "문자 기반 유해사이트 판별 기법", 한국컴퓨터교육학회 논문지 제7권 제5호, 2004.9

[12] 김광현, "웹 문서 분석에 근거한 유해 웹 문서 검출", 한국정보처리학회 논문지 D, VOL.12-D NO.05, 2005.10