DOI: 10.3745/KIPSTB.2009.16-B.1.85

유해어 필터링과 SVM을 이용한 유해 문서 분류 시스템

이 원 휘[†] · 정 성 종[††] · 안 동 언[†††]

요 약

오늘날 웹이 일반화되면서 사람들은 원하는 정보를 웹을 통해 얻고, 또한 제공하고 있다. 웹이 다양한 정보의 제공과 습득의 장이라는 편의성을 제공하고 있지만, 반면에 너무 많은 정보, 무분별한 유해 정보의 범람 등 여러 가지 문제를 내포하고 있다. 현재 유해 웹 문서를 분류하기 위한 다양한 방법이 연구되고 사용되고 있다. 그러나 각각의 방법들이 갖는 단점들로 인해 획기적인 성과를 내지 못하고 있다. 본 논문에서는 유해 정보로부터 사회적으로 보호를 받아야 할 사용자들을 보호하기 위한 수단으로 유해 웹 문서 차단 방법에 대해 제안하고자 한다. 본 논문에서는 키워드 필터링과 SVM 알고리즘을 이용한 2단계 분류 과정을 통해 분류의 정확률을 높이고자 하였다.

키워드 : 유해어, 분류, SVM, 필터링, 웹 문서

Harmful Document Classification Using the Harmful Word Filtering and SVM

WonHee Lee[†] · SungJong Chung[††] · DongUn An[†††]

ABSTRACT

As World Wide Web is more popularized nowadays, the environment is flooded with the information through the web pages. However, despite such convenience of web, it is also creating many problems due to uncontrolled flood of information. The pornographic, violent and other harmful information freely available to the youth, who must be protected by the society, or other users who lack the power of judgment or self-control is creating serious social problems. To resolve those harmful words, various methods proposed and studied. This paper proposes and implements the protecting system that it protects internet youth user from harmful contents. To classify effective harmful/harmless contents, this system uses two step classification systems that is harmful word filtering and SVM learning based filtering. We achieved result that the average precision of 92.1%.

Keywords : Harmful Word, Classification, SVM, Filtering, Web Document

1. 서 론

오늘날의 정보환경은 웹(World Wide Web)이 대중화 되면서 웹을 통한 정보가 홍수를 이루는 환경을 이루고 있다. 사람들은 원하는 정보를 웹을 통해 얻고, 제공하고자 하는 정보 또한 웹을 통해 제공하고 있다. 이렇게 웹은 다양한 정보의 제공과 습득의 장이 되고 있다. 그러나 이러한 웹의 편리성은 그 이면에 무분별한 정보의 제공으로 인한 여러 가지 문제를 내포하고 있다. 너무 많은 정보의 제공으로 인한 정보 검색의 부담과 무분별한 유해 정보의 범람은 정보 사회를 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. 게시자의 자발적 등급 결정에 기반을 둔 인터넷 내용 선택에 대한 플랫폼(PICS)[5], 유해 사이트 목록에 의한 사이트 차단 방법(URL 차단), 제공되는 영상정보의 스킨컬러(skin color)에 기반을 둔 연구[4], 유해한 단어나 어구에 기반하여 필터링하는 키워드 필터링[15], 신경망 이론을 응용한 지능적 내용 분류 연구, 이미지 정보를 이용한 연구[10,11] 등이 그것이다. 그러나 이러한 연구나 제도들이 가지는 한계로 인하여 극히 저조한 성능을 보이고 있다[2].

본 논문에서는 웹 문서의 텍스트 정보를 이용한 필터링을 구현하였다. 본 논문은 분류의 정확률을 높이기 위하여 2단계의 분류를 수행한다. 먼저, 시스템은 유해어 필터링을 통

† 정 회 원 : 전북대학교 컴퓨터공학과 박사수료
†† 정 회 원 : 전북대학교 전자정보공학부 교수
††† 종신회원 : 전북대학교 전자정보공학부 교수
논문접수 : 2007년 12월 27일
수 정 일 : 1차 2008년 1월 25일, 2차 2008년 11월 20일
심사완료 : 2008년 11월 20일

해 유해 문서와 무해 문서를 분류한다. 이를 위해 유해어 사전을 사용한다. 유해어 사전에는 유해어의 주변 단어들에 대한 정보를 포함한다. 2단계에서는 1단계에서 유해문서로 판정된 문서들을 대상으로 SVM 알고리즘을 이용한 분류를 통해 등급을 분류한다.

본 논문의 구성은 2장에서 웹 문서 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 문서 필터링 연구는 크게 인터넷 내용 선택에 대한 플랫폼, URL 차단, 키워드 필터링, 인공지능 내용 분석, 이미지 기반 필터링으로 분류할 수 있다[2,11].

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), 신체 노출(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 문서의 유해성을 묘사하기 위하여 11개의 카테고리를 사용한다[2,8,5,10].

2.2. 유해 사이트 목록에 의한 사이트 차단(URL 차단)

유해 사이트 목록에 의한 차단 방법은 유해사이트를 원천적으로 차단하기 위하여 유해사이트라고 판단된 사이트 목록을 작성하여 사용자가 해당 사이트에 접근하고자 할 경우 접근을 차단하는 방법이다. 그러나 이 방법은 사이트를 원천적으로 차단함으로 인해 해당 사이트 내에 포함되어 있을 유해하지 않은 내용까지도 차단한다는 단점과, 사용자의 저항이 다른 차단 방법에 비해 크다는 단점을 가지고 있다.

2.3. 키워드 필터링

이 방법은 문서 내의 유해한 단어나 어구의 발생에 기초해서 웹 문서를 차단한다. 웹 페이지에서 검색된 단어나 어구는 금지된 단어와 어구로 이루어진 키워드 사전상의 단어들과 비교되고 임계치 이상의 유해 단어나 어구가 발생하는 경우 차단이 이루어진다. 임계치는 유해단어가 출현한 문서상의 위치나 문서 전체에서 출현한 단어에 대한 비율 등에 따라서 조절된다[6].

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 문서까지 차단하는(over-blocking) 단점을 지니고 있다[2,11,15]. 예를 들어, 많은 검색엔진에서 “거유”라는 단어를 입력하면 성인인증 과정을 거치도록 하고 있다. 이는 “거유”라는 단어 자체를 모두 巨乳로 인식해 다른 의미(去油, 巨儒, 據有 등)의 문서까지도 차단하게 된다.

2.4. 지능 내용 분석

웹 필터링 시스템은 자동적인 웹 문서의 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[10-12].

3. 구 현

3.1. 시스템 구조

본 논문에서 제안하고 구현한 시스템은 분류의 정확률을 높이고 판정의 시간을 단축하기 위하여 유해어 필터링 부분과 SVM 학습의 두 단계의 필터링을 수행한다. 시스템의 구성은 다음 (그림 1)과 같다.

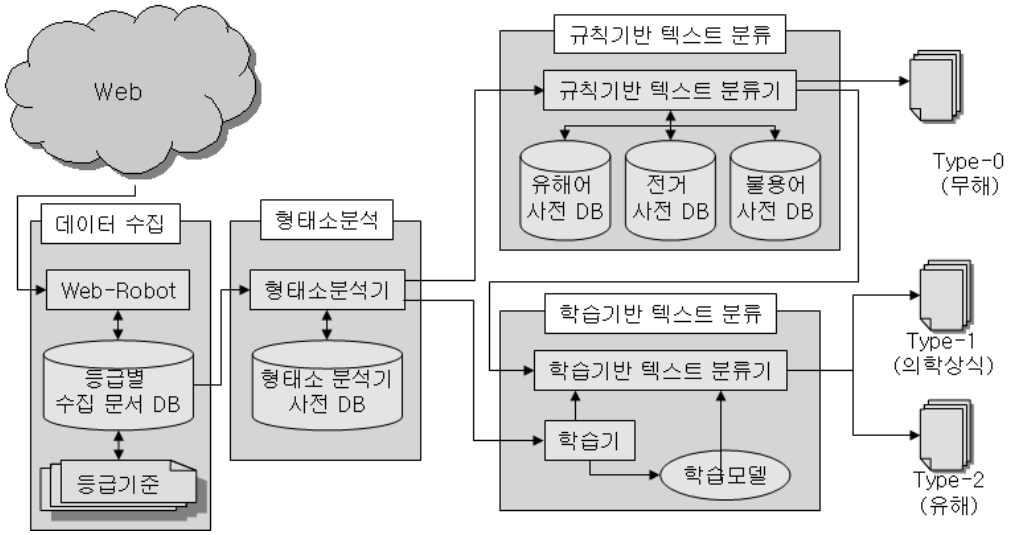
3.2. 유해어 필터링(Harmful word filtering)

유해어 필터링에서는 키워드 필터링을 통해 유해어가 포함된 문서를 분류한다. 이 필터링을 수행하기 위하여 유해어 사전(harmful word dictionary)과 전거 사전(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 유해어의 주변 단어 정보를 추가하였다. 주변 단어는 동일 문장에서 출현하는 인접어와 동일 문서에서 출현하는 비인접어로 구성된다.

3.2.1. 인접어와 비인접어

유해어 사전은 유해어 후보와 해당 유해어 후보 주변 단어들로 인접어와 비인접어로 나누어진다.

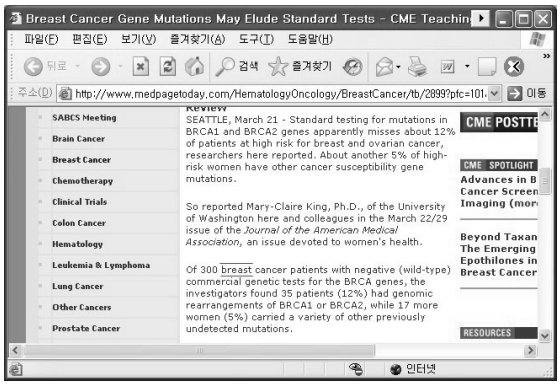
인접어란 유해어 후보와 동일한 문장에 출현하여 해당 유해어 후보의 유해여부나 유해정도를 결정지을 수 있도록 하는 단어를 의미한다. 예를 들어 “가슴”이라는 단어는 유해할 수도 있고, 무해할 수도 있다. 또한 이 단어가 유해다면, 유해정도가 심하거나 미미하기도 할 수 있다. 예를 들어, “가슴”과 같은 문장에 “암”이라는 단어가 출현한다면 해당 문



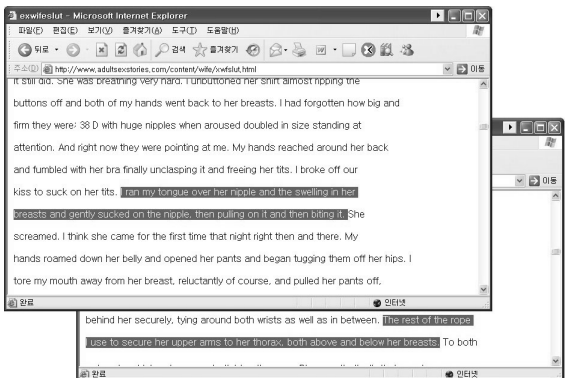
(그림 1) 시스템 구조

서는 무해할 가능성이 높다. 반면, 같은 문장에 "혀"라는 단어가 출현한다면 해당 문서가 유해할 가능성이 높다. 또한, "채찍"과 같은 단어가 동일 문장에 출현한다면 이 문서는 변태적인 내용을 포함하는 유해정도가 심한 문서일 가능성이 높다.

비인접어는 유해어 후보와 동일한 문장에는 출현하지 않지만 동일한 문서에 출현함으로써 해당 유해어 후보의 유해여부를 판정하거나 유해정도를 결정지을 수 있도록 하는 단어를 의미한다. 예를 들어 앞에서 예로 든 "암", "혀", "채찍"들이 "가슴"라는 유해어 후보와 동일한 문장에 출현하지는 않지만 동일한 문서에 출현한다면 해당 문서의 유해성을 판정하는데 어느 정도 기여를 할 수 있을 것이다.



(그림 3) 무해문서의 예(건강)



(그림 2) 유해문서의 예(야설)

이때 주의할 것은 인접어나 비인접어에 포함되는 단어들은 그 자체로는 유해성을 가지지 않는 단어들이라는 것이다. 유해성을 가지지 않는 단어들이 유해할 수도 있는 단어들과 만났을 때 해당 유해어 후보의 유해/무해를 판정하거나 유해정도를 결정지을 수 있는 역할을 하는 단어들이 인접어와 비인접어에 포함된다. 또한 모든 유해어가 인접어와 비인접어 리스트를 갖는 것은 아니다. 유해어 후보 자체가 유해성이 확실하고 유해정도가 확실하다면 인접어와 비인접어를 비교해야 되는 연산을 수행함으로써 발생되는 계산시간을 줄일 수 있기 때문이다.

3.2.2 전거 사전

표준어가 문서 제작자의 의도적 또는 실수로 변형되어 표기되거나 약자로 표기된 경우 이들을 하나의 표준어로 치환하기 위해 사용된다. 그렇지 않을 경우 자질 생성 시 출현빈도수에 영향을 주어 유해/무해 판정에 영향을 미치게 되기 때문이다.

예를 들어 "가슴"라는 용어는 상황 등에 따라 "슴가", "유방", "가심" 등으로 쓰일 수 있다. 그러나 이러한 단어들은 유사하거나 같은 의미로 사용된 경우 이 단어들을 각각의 단어로 계산하게 되면 빈도수가 분산되어 유해/무해 판단에 영향을 미칠 수 있다. 따라서 이들 단어들을 "가슴"으로 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 사용되

〈표 1〉 전거 예

대상어		표준어
...	...	...
브라자	=>	브래지어
사디스트	=>	새디스트
크리	=>	클리토리스
페티시	=>	패티쉬
빤쓰	=>	팬티
페라치오	=>	펠라치오
...	...	...

고, “fuck”의 경우 음란한 의미로 사용되어 “coitus”, “copulate”, “firk” 등과 유사하게 사용된다. 이들을 구분지어서 관리할 필요가 있다.

3.2.3 사전에 기반한 분류

유해어 사전과 전거사전을 이용하여 대상 문서를 유해와 무해 문서로 분류한다.

분류 과정은 다음과 같이 진행된다.

[단계 1] 문서 전처리 단계

전처리 단계에서는 대상 문서에 대하여 태그 제거와 형태소 분석 작업을 수행한다.

[단계 2] 표준어 치환

형태소 분석 결과에서 명사, 형용사, 동사만을 대상으로 단어 리스트를 작성한다. 작성된 단어리스트에서 비표준어가 존재한다면 전거사전을 참조하여 표준어로 변환한다.

[단계 3] 유해/무해 판단

단어 리스트에서 유해어 후보를 찾는다. 찾아진 유해어 후보는 인접어와 비인접어를 이용하여 최종적으로 유해어인지 아닌지를 판단한다. 유해어로 판정된 유해어가 존재하지 않는다면 무해문서로 판정한다. 유해로 판정된 문서는 다음 단계인 학습기반 텍스트 분류단계로 넘긴다.

3.3 학습기반 텍스트 분류

학습기반 텍스트 분류에서는 3단계로 구성이 된다. 자질어의 추출 및 색인 부분과 학습 모델의 생성부분, 실제 텍스트 분류 부분으로 구성된다.

3.3.1 자질어 추출

자질어 추출은 학습 기반 텍스트 분류에서 사용될 자질어의 목록을 생성하는 부분이다. 자질어 추출은 학습대상 문서에서 태그를 제거하고 형태소 분석 단계를 거친 후의 데이터를 이용한다. 문서가 전처리 과정을 거치면서 조사나 형용사 등의 불용어 제거와 전거 사전을 참조하여 비표준어의 표준어 치환 과정을 수행하게 된다. 자질어 추출 알고리즘으로는 DF(Document Frequency), IG(Information Gain), x^2을 각각 적용하여 최선의 방법을 찾고자 하였다. 각각의 알고리즘은 다음과 같다.

DF(Document Frequency)는 전체 문서 집합 중 특정 단어가 출현한 문서의 수를 의미한다. 본 시스템은 이 DF를 이용하여 자질어를 추출하고 일정 임계치 이하의 문서에서 출현하는 용어를 제거한다. 이 때 “문서 빈도가 아주 낮은 용어는 특정 주제 범주를 대표할 만한 충분한 정보가 되지 못하고 전체적인 성능에도 큰 영향을 미치지 못 한다”는 기본 가정을 가지고 출발한다. 이 알고리즘은 매우 간단하고 계산량이 적다는 장점을 가지고 있는 반면 정보검색 분야에서 전통적으로 문서 빈도 값이 낮을수록 색인어로서의 가중치를 높게 할당하는 것과 대치되는 단점을 가지고 있다[18].

IG(Information Gain)는 특정 단어의 출현 여부가 문서 분류에 기여하는 정도를 계산하기 위하여 기여도가 높은 자질만을 선택하는 알고리즘으로 모든 용어들의 정보 획득량을 계산하여 일정 임계치 이상의 값을 갖는 용어들만을 자질로 선택하게 된다. 이 방법은 문서에서의 출현 빈도뿐만 아니라 출현하지 않은 빈도까지 고려하여 각 범주에서의 용어 정보량을 계산한다. 범주 집합이 $\{C_1, C_2, ..., C_n\}$일 때 IG의 알고리즘은 다음과 같다. $\Pr(C_i)$는 전체 문서에 대한 범주 i의 비율이며, $\Pr(C_i|t)$는 범주 i의 문서 중 단어 t가 출현한 문서의 비율, $\Pr(C_i|\bar{t})$은 범주 i의 문서 중 단어 t가 출현하지 않은 문서의 비율이다[18].

$$IG(t) = -\sum_{i=1}^{m}\Pr(C_i)\log\Pr(C_i) + \Pr(t)\sum_{i=1}^{m}\Pr(C_i|t)\log\Pr(C_i|t) + \Pr(\bar{t})\sum_{i=1}^{m}\Pr(C_i|\bar{t})\log\Pr(C_i|\bar{t})$$

x^2는 용어 t와 범주 c간의 의존성을 측정해 용어의 중요도를 구하는 방법으로 t와 c 두 값의 차가 클수록 용어 t가 자질로 선정될 확률이 높아진다. 또한 문서 빈도를 사용해 범주별 발생분포가 일반적인 단어들의 발생분포와 다른 정도를 계산하고, 그 차이가 특정 값 이상인 단어를 자질로 선정하게 된다. 최종 x^2 통계량을 구하기 위해서는 각 용어 및 범주에 대해 통계값을 계산한 후 각 용어마다 계산된 카이제곱 통계량의 평균이나 최대값을 구한다. 알고리즘은 다음과 같이

	범주 c가 할당된 문서	범주 c가 할당되지 않은 문서
단어 t가 출현한 문서	A	B
단어 t가 출현하지 않은 문서	C	D

일 때

$$x^2(t,c) = \frac{N*(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$
$$x^2avg(t) = \sum\Pr(c_i)x^2(t,c_i)$$
$$x^2\max(t) = Max_{i=1}^{m}\{x^2(t,c_i)\}$$

이다. 이 때 N=A+B+C+D이다[18].

3.3.2 색인

자질어 추출 과정에서 추출된 자질들이 문서에서 차지하는 비중을 계산하는 가중치 부여 단계가 필요하다. 즉 한 문서의 특징을 표현하기 위해 가중치가 부여된 자질어를 이용하여 문서를 벡터화하는 단계이다. 색인 및 가중치를 부여하기 위해 TF, TF-IDF, TF-ICF의 알고리즘을 사용하였다.

TF에 의한 가중치 계산 방법은 단순히 용어가 한 문서 내에서 나온 빈도수, 즉 Term Frequency를 사용한다. TF에 의한 가중치 부여 방법에는 단순TF, 이진TF, 로그TF 등이 사용되는데, 본 연구에서는 출현빈도가 너무 낮거나 높은 단어들의 영향력을 보완하기 위해 로그TF를 사용하였다.

$$TF = 1 + \log(tf)$$

여기서 tf는 용어가 한 문서 내에서 나온 빈도수이며, TF는 적절히 변형된 빈도수를 의미한다.

TF만으로는 고빈도의 용어가 항상 문서를 대표하지 않고 대부분의 고빈도 용어는 기능어로써 많이 등장하지만, 그 문서의 내용을 나타내지 못하는 맹점을 가지고 있다. 그래서 TF-IDF(Inverse Document Frequency)와 TF-ICF(Inverse Category Frequency)를 사용한다.

TF-IDF는 적은 수의 문서에 나타난 자질에 대해 높은 가중치를 부여하는 방법으로 알고리즘은 다음과 같다.

$$IDF = \log N - \log DF_i + 1$$
$$Weight = TF * IDF$$

여기서 DF_j는 자질어 w_i를 포함하는 문서의 개수이며 N은 총 문서의 개수가 된다[19].

ICF는 소수의 범주에 많이 나오는 용어에 높은 가중치를 부여하고, 여러 범주에 고르게 나오는 용어에는 낮은 가중치를 부여하는 방법으로 알고리즘은 다음과 같다.

$$ICF = \log M - \log CF_j + 1$$
$$Weight = TF * ICF$$

여기서 M은 범주의 총 수이며, CF_j는 자질어 w_i를 포함하는 범주 수이다. 범주는 type-0(무해), type-1(성상식), type-2(유해) 등 3개 범주로 구성하였다.

색인 과정을 거친 가중치 값은 SVM의 성능을 높이기 위하여 값들을 정규화 하는 과정이 필요하다. 정규화는 가중치 값이 -1과 1사이의 값으로 정규화 하였다[20].

3.3.3 학습 모델 생성

학습 모델 생성은 추출된 자질어와 정규화된 가중치 값으로 SVM(Support Vector Machines) 학습모델을 생성하는 과정이다. SVM은 고차원 특징 공간상에서 선형함수의 가설공간을 사용하는 학습 시스템으로, 최적화 이론에서 나온 학습 알고리즘으로 훈련되었는데 학습 알고리즘은 통계 학습 이론에서 유도된 학습 바이어스를 구현하는 것이다. 뱁닉(Vapnik)과 그의 동료들에 의해 도입된 이 학습 방법은 몇 년 동안 광범위한 응용분야에서 다른 대부분의 시스템을 능가하는 매우 강력한 방법이다. SVM은 이차원 데이터 분류문제에서 가장 최적의 초평면(Hyperplane)을 구하여 이를 결정경계면으로 선택한다. 최적의 초평면은 선형 분리가 가능한 두 집단에 대해 집단을 구분 지으며, 마진을 최대로 한다. 하지만 실제 문제의 경우 선형적으로 구성되는 예가 적기 때문에 커널 함수를 이용하여 비선형적 특징공간을 선형적 특징공간으로 매핑한 후에 선형 SVM으로 분류하게 된다. 본 논문에서 사용한 SVM 타입과 커널은 각각 C-SVC (C-Support Vector Classification), RBF(Radial Basis Function)을 사용하였다.

C-SVC의 결정 함수는

$$sgn(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b)$$

이다. 이 때, $y_i \in \{1, -1\}$, $0 \le \alpha_i \le C$, $K(x_i, x)$는 커널이다. C는 Cost이다.

커널 함수는

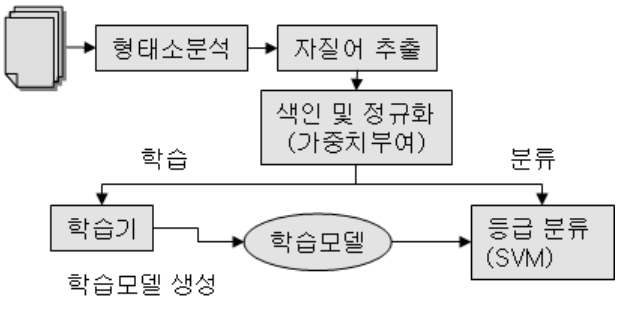
$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)$$

이다[16]. C와 γ는 학습과정에서 구하게 된다.

학습 절차는 먼저 학습 대상 문서 집합을 결정하고, 결정된 학습 대상 문서 집합을 대상으로 학습이 이루어진다. 다음 단계로, 지정된 범위 내에서 각각의 최적 파라미터 Gamma(ɣ)와 Cost(C)를 찾게 된다. 이렇게 찾아진 최적의 파라미터와 SVM 커널을 이용하여 학습 모델을 생성한다. 학습모델 생성이 완료되면 생성된 해당 모델을 저장하여 등급 분류 테스트나 등급 분류 시에 사용한다.

3.3.4 학습모델을 이용한 등급 분류

앞 단계에서 생성된 학습 모델을 이용하여 실제 문서에 대한 등급분류를 수행한다. 등급을 분류하고자 하는 문서는 태그 제거와 형태소 분석을 과정을 통해 자질어를 생성하게 된다. 생성된 자질어를 이용하여 색인 및 정규화를 수행한다.



(그림 4) 학습기반 텍스트 분류

학습 단계에서는 학습기를 통해 3.3.3에서 설명된 알고리즘을 통해 학습모델을 생성한다. 분류 단계는 학습 단계에서 생성된 학습모델을 이용하여 분류를 수행한다.

4. 실험 및 평가

사용된 데이터 셋은 2005년에 ETRI에서 수집한 유해 문서 집합인 EHDS-20000(ETRI Harmful Data Set)을 이용하였다. EHDS-20000는 신문기사와 의학상식, 성인사이트의 야설문서 등을 수집한 문서 셋으로 <표 2>의 수집데이터와 같다. 위 문서 셋에서 실험 및 평가에 사용된 학습 데이터 및 테스트 데이터는 <표 2>의 학습데이터와 테스트 데이터와 같다.

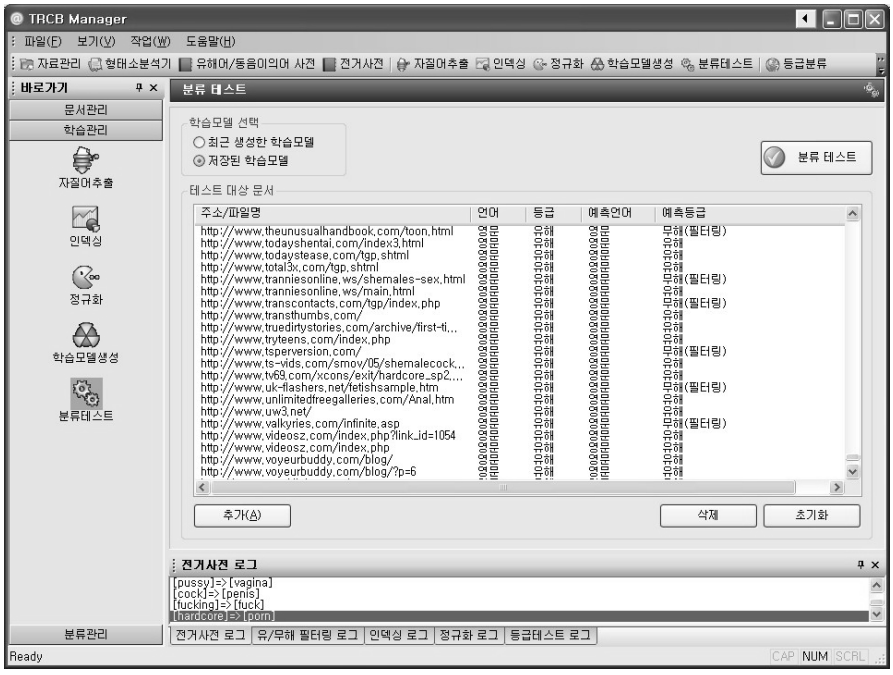
실험에 앞서 먼저 최적의 추출 알고리즘과 색인 알고리즘을 찾기 위하여 자질어 추출 알고리즘(logTF, IG, x^2)과 색

〈표 2〉 데이터 셋

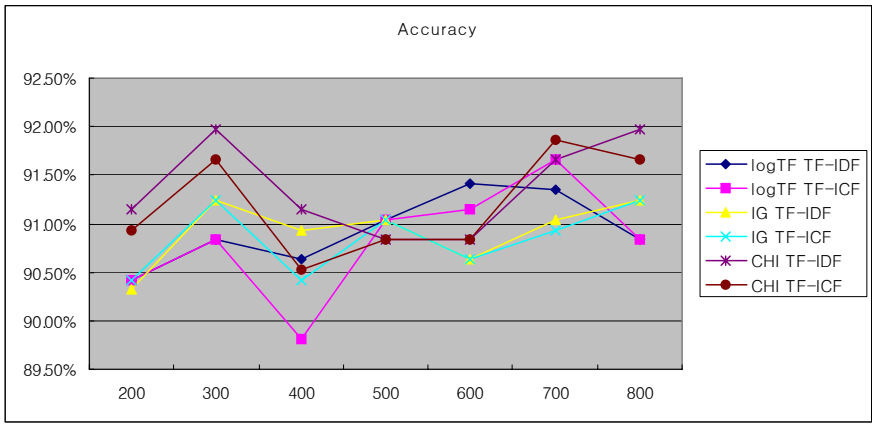
		수집데이터		학습 데이터		테스트 데이터	
		무해	유해	무해	유해	무해	유해
문서	한글	2126	2572	588	1164	462	509
	영문	3340	12250	694	936	533	449

인 알고리즘(TF-IDF, TF-ICF)의 각 조합으로 자질어 수를 200과 800사이의 값을 100개씩 증가시켰을 때 분류의 정확도를 측정하였다. 분류의 정확도는 유해문서를 유해 문서로, 무해 문서를 무해문서로 바르게 분류하는 정도이다. 실험 결과 각각의 측정값은 다음 그림과 같이 나타났다. 결과에서 자질어 수는 800에 x^2와 TF-IDF 조합이 가장 좋은 성능을 나타냄을 알 수 있다((그림 6) 참조).

실험은 유해어 필터링 단계만을 이용한 분류와 유해어 필터링과 학습을 이용한 2단계 분류로 나누어 실시하여 성능 개선을 평가하였다. 우선 유해어 필터링만을 이용한 실험결



(그림 5) 실행 화면



(그림 6) 자질어 추출 알고리즘 비교

과는 다음과 같다.

〈표 3〉 유해어 필터링 결과

Non-Harmful(995)		Harmful(958)		Overall accuracy
correctly	incorrectly	correctly	incorrectly	
511	484	933	25	74.35%
(51.3%)	(48.7%)	(97.4%)	(2.6%)	

위 결과에서 보듯이 유해의 경우 정확하게 분류된 경우가 97.4%라는 정확도를 얻을 수 있었다. 그러나 무해의 경우 51.3%라는 낮은 정확도가 나타났다. 원인을 분석한 결과 바르지 않은 것으로 나타난 48.7%의 문서 내에는 성상담이나 의학상식 등의 문서들이 다수 포함되어 있는데, 내용 자체는 무해하나 문서 내의 유해 단어들이 영향을 미친 것으로 나타났다. 위 결과를 개선하기 위하여 문서 셋을 세분화하도록 하였다. 우선, 무해 문서 셋을 두 개의 유형(type-0과 type-1)으로 분류하였다. 유해 단어가 전혀 출현하지 않는 문서를 type-0이라고 정의하고, 유해 단어가 출현하는 성상담이나 의학상식 등의 문서를 type-1이라고 정의하였다. 또한 유해 문서 셋은 type-2로 정의하였다. type-0은 유해어 필터링 단계에서 분류하고 학습기반 분류 단계에서 type-1과 type2를 분류하도록 하였다.

실험을 통해 얻어진 결과는 다음 표와 같다.

〈표 4〉 유해어 필터링과 SVM 기반 필터링 결과

result / input data	type-0	type-1	type-2
type-0(524)	503(95.9%)	14(2.7%)	7(1.4%)
type-1(471)	8(1.7%)	428(90.9%)	35(7.4%)
type-2(958)	25(2.6%)	76(7.9%)	857(89.5%)

위의 결과에서 평균 92.1%의 정확률을 보이고 있다. 무해 문서의 경우 type-0이 95.9%의 정확도, type-1이 90.9%의 정확도로 평균 정확도가 93.4%로 증가됨을 볼 수 있다. type-2(유해)의 경우에는 이전에 97.4%의 정확도에서 오히려 감소된 정확도가 나타났는데 감소한 7.9%를 살펴본 결과 유해 문서 내에 포함되어 있던 성상담이나 의학상식에 관한 내용으로 확인되었다. 본문은 무해하나 본문 이외의 영역(메뉴, 광고 텍스트 등)에 포함되어 있는 유해어가 판정에 영향을 미쳤다. 또한 type-1과 type-2가 type-0으로 판정된 1.7%(type-1), 2.6%(type-2)의 경우는 페이지가 주로 이미지로 이루어져있고 텍스트가 제목 등 극히 일부의 역할을 담당하는 페이지인 경우로 확인되었다.

5. 결 론

현재 인터넷 환경은 정보의 홍수와 함께 음란, 폭력, 자살 등의 유해 정보가 범람하는 환경으로 되고 있다. 이들을 해결하기 위하여 PICS로 대표되는 등급시스템, 키워드 필터링, 지능 분석 시스템 등 다양한 방법이 제안되고 연구되고 있다. 그러나 단순히 이들 방법으로는 유해 정보를 차단하는데 한계가 있음을 알 수 있다.

본 논문에서는 웹상에 급속히 확산되고 있는 유해 문서에 대한 정확한 분류를 통해 청소년 등 사회로부터 보호를 받아야 될 인터넷 사용자들이 유해한 정보로부터 보호될 수 있는 시스템을 제안하고 구현하였다. 본 논문에서는 효과적인 유해/무해 문서의 분류를 위하여 규칙기반 텍스트 분류와 학습기반 텍스트 분류라는 2단계 분류시스템을 사용하였다. 문서의 자질 측면에서도 logTF, IG, CHI, TF-IDF, TF-ICF 등 다양한 알고리즘에 대하여 실험하여 정확도를 높일 수 있는 알고리즘 조합을 선택하였다. 그 결과 평균 92.1%의 정확률을 얻을 수 있었다.

향후 이미지나 소리, 문서 구조 등 다양한 문서 요소의 분석을 통해 정확률을 좀 더 향상할 필요가 있다.

참 고 문 헌

[1] Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin, “A Practical Guide to Support Vector Classification,” http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[2] 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, pp.287-294, April 2000.

[3] Dequan Zheng, Yi Hu, Tiejun Zhao, Hao Yu and Sheng Li, “Research of Machine Learning Method for Specific Information Recognition on the Internet,” IEEE International Conference on Multimedia Interfaces(ICMI), pp, October 2002.

[4] Huicheng Zheng, Hongmei Liu and Mohamed Daoudi, “Blocking Objectionable Image : Adult Images and Harmful Symbols,” IEEE International Conference on Multimedia and Expo(ICME), pp.1223-1226, June 2004.

[5] Jae-Sun Lee and 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.

[6] KwangHyun Kim, JoungMi Choi and JoonHo Lee, “Detecting Harmful Web Documents Based on Web Document Analyses,” Communication of KIPS Review, Vol.12-D, No.5, pp.683-688, Oct. 2005.

[7] 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.

[8] 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.

[9] Nello Cristianini and John Shawe-Taylor, “An Introduction to Support Vector Machines and other kernel-based learning methods,” Cambridge university press, 2000.
[10] P.Y. Lee and S.C. Hui, “An Intelligent Categorization Engine for Bilingual Web Content Filtering,” IEEE Transaction On Multimedia, Vol.7, No.6, December 2005.
[11] P.Y.Lee, S.C.Hui and A.C.M. Fong, “Neural Networks for Web Content Filtering,” IEEE Intelligent Systems, pp.48-57, Sept./Oct. 2002.
[12] Qing Yang and 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, pp.3787-3791, Aug. 2005.
[13] Seung-Man Lee, Young-Hun Jang and Jung-Hwan Lim, “Implementation of a Harmful Website’s Automatic Classification System based on Morphological Analysis and Skin-Color Distribution’s Human Detection Algorithm,” KISS Spring Conference Vol.31, No.1, pp.601-603, Apr. 2004.
[14] Thorsten Joachims, “Learning to Classify Text using Support Vector Machines,” Kluwer Academic Publishers, 2002.
[15] Yun-Jung Jang, Taehun Lee, Kyu Cheol Jung and Kihong Park, “The Method of Hurtfulness Site Interception Using Poisonous Character Weight,” KIPS Spring Conference, Vol.10, No.01, pp.2185-2188, May 2003.
[16] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM:a Library for Support Vector Machines,” http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html
[17] 김영수, 남택용, 원동호, “등급에 따른 웹 유해 문서 분류 기술”, 한국정보처리학회논문지C, 제13C권 7호, pp.859-864, 2006.
[18] 김영택 외, “자연언어처리”, 생능출판사, 2003.
[19] 권용진, 황수찬 역, “정보검색개론”, 도서출판 미래컴, 2003.
[20] Reed J.W, Jiao Yu, Potok T.E, Klump B.A, Elmore M.T and Hurson A.R, “TF-ICF, A New Term Weighting Scheme for Clustering Dynamic Data Streams,” Machine Learning and Applications, 2006. ICMLA ’06. 5th International Conference on Dec. 2006 Page(s), 258-263.

이 원 휘

e-mail : wony0603@chonbuk.ac.kr
1997년 전주대학교 경영학과 졸업(학사)
1999년 전주대학교 컴퓨터공학과 졸업(공학석사)
2007년 전북대학교 컴퓨터공학과 박사 수료
관심분야 : 정보검색, 문서분류, 문서요약

정 성 종

e-mail : sjchung@chonbuk.ac.kr
1975년 한양대학교 전기공학과 졸업(학사)
1981년 휴스턴대학교 전자공학과 졸업(공학석사)
1988년 충남대학교 전산공학과 졸업(공학박사)
1985년~현 재 전북대학교 전자정보공학부 교수
관심분야 : 정보검색, Grids

안 동 언

e-mail : duan@chonbuk.ac.kr
1981년 한양대학교 전자공학과 졸업(학사)
1987년 KAIST 전산학과 졸업(공학석사)
1995년 KAIST 전산학과 졸업(공학박사)
1995년~현 재 전북대학교 전자정보공학부 교수
관심분야 : 정보검색, 한국어정보처리, 문서분류, 문서요약