등급에 따른 웹 유해 문서 분류 기술

김 영 수[†] · 남 택 용[††] · 원 동 호[†††]

요 약

웹의 개방성은 사람들로 하여금 언제, 어디서든 손쉽게 유용한 정보를 획득할 수 있게끔 하였다. 하지만 인터넷은 유용한 정보의 손쉬운 활용이라는 순기능과 더불어 사회적으로 통제를 필요로 하는 유해한 정보 역시 인터넷을 이용하는 이용자들에게 무차별적으로 제공함으로써 역기능을 발생시키고 있다. 성인 컨텐츠 같은 정보들은 모든 사용자들, 특히 청소년들에게 악영향을 미칠 수 있다. 또한, 변태적인 성인 사이트들이 담고 있는 컨텐츠들은 성인들의 정신 건강에도 해를 미치게 된다. 한편, 인터넷은 전 세계적으로 연결된 개방망이므로 유해정보 제공자를 각국의 법적, 제도적 장치를 이용하여 규제하는데 한계가 있다. 또한, 유해 사이트, 유해성 스팸 메일, P2P 등 다양한 경로를 통해 유해 정보를 접할 수 있기 때문에, 어떤 시스템에 특화된 유해정보 분류기술을 개발하는 것은 바람직하지 않다. 따라서, 유해정보의 내용 자체에 기반하여 유해 여부를 자동으로 판별할 수 있는 유해정보 판별 핵심 기술의 연구 및 개발의 중요성이 점차 부각되고 있다. 이에 본 논문에서는 내용 기반 기술을 이용한 효율적인 유해 웹 문서 텍스트 판별 시스템을 제시한다.

키워드 : 텍스트 판별, 유해 사이트 차단, 기계 학습, 키워드 매칭

A Distinction Technology for Harmful Web Documents by Rates

Youngsoo Kim[†] · Taekyong Nam[††] · Dongho Won[†††]

ABSTRACT

The openness of the Web allows any user to access almost any type of information easily at any time and anywhere. However, with function of easy access for useful information, internet has dysfunctions of providing users with harmful contents indiscriminately. Some information, such as adult content, is not appropriate for all users, notably children. Additionally for adults, some contents included in abnormal porn sites can do ordinary people's mental health harm. In the meantime, since internet is a worldwide open network, it has a limit to regulate users providing harmful contents through each countrie's national laws or systems. Additionally it is not a desirable way of developing a certain system-specific classification technology for harmful contents, because internet users can contact with them in diverse way, for example, porn sites, harmful spams, or peer-to-peer networks, etc. Therefore, it is being emphasized to research and develop context-based core technologies for classifying harmful contents. In this paper, we propose an efficient text filter for blocking harmful texts of web documents using context-based technologies.

Key Words : Text Classification, Harmful Web Text Filtering, Machine Learning, Keyword Matching

1. 서 론

21세기 정보화 혁명을 주도하고 있는 인터넷은 사람들로 하여금 시간과 공간을 뛰어 넘어 언제, 어디서든 손쉽게 유용한 정보를 획득할 수 있게끔 하였다. 하지만 인터넷은 유용한 정보와 손쉬운 활용이라는 순기능과 더불어 사회적으로 통제를 필요로 하는 유해한 정보 역시 인터넷을 이용하는 이용자들에게 무차별적으로 제공함으로써 역기능을 발생시키고 있다. 사회적인 보호를 받아야 하는 청소년을 비롯한 판단력과 절제력이 부족한 인터넷 이용자들이 인터넷의 유해 정보를 아무런 제재 없이 접근할 수 있게 되어서 개인뿐 아니라 사회적인 문제가 되고 있다. 한편, 인터넷은 전 세계적으로 연결된 개방망이므로 유해정보 제공자를 각국의 법적, 제도적 장치를 이용하여 규제하는데 한계가 있다. 또한, 유해 사이트, 유해성 스팸 메일, P2P 등 다양한 경로를 통해 유해 정보를 접할 수 있기 때문에, 어떤 시스템에 특화된 유해정보 분류기술을 개발하는 것은 바람직하지 않다. 따라서, 유해정보의 내용 자체에 기반하여 유해 여부를 자동으로 판별할 수 있는 유해정보 판별 핵심 기술의 연구 및 개발의 중요성이 점차 부각되고 있다. 이에 본 논문에서는 내용 기반 기술을 이용한 효율적인 유해 웹 문서 텍스트 판별 시스템을 제시한다. 제2장에서 제안하는 시스템을 구성

† 정 회 원 : 한국전자통신연구원 정보보호연구단 선임연구원
†† 정 회 원 : 한국전자통신연구원 정보보호연구단 보안게이트웨이연구팀 팀장, 과학기술연합대학원대학교(UST) 정보보호공학 교수
††† 종신회원 : 성균관대학교 정보통신공학부 교수(교신저자)

논문접수 : 2006년 6월 2일, 심사완료 : 2006년 8월 23일

하는 텍스트 판별 기술을 구분하고, 제3장에서는 이 기술들을 이용하여 인터넷 웹 문서의 유해 여부를 판별할 수 있는 효율적인 유해 텍스트 판별 시스템을 제시한다. 그리고, 제4장에서는 전장에서 제시한 시스템에 대한 구현을 보이고, 제 5장에서는 장기간에 걸쳐 수집한 웹 문서들을 대상으로 하여 구현한 시스템을 시험하고 그 결과를 분석한 후, 제 6장에서 결론을 맺는다.

2. 텍스트 판별 기술

텍스트 판별은 범주 정형화, 자질 추출 기술 그리고 텍스트 판별 기술로 구분된다. 범주(category) 정형화는 판별 대상이 되는 범주와 각 범주를 구분하는 기준을 정하는 단계이고, 자질(feature) 추출 기술은 문서에 나타나는 내용어 중 문서 판별에 유용하게 사용될 만한 내용어를 선택하는 단계이다. 그리고, 텍스트 판별 기술은 추출된 자질을 통해 텍스트를 판별하는 단계로 기계 학습 분야에서 사용되는 알고리즘들이 사용된다.

2.1 자질 추출

문서에 나타나는 내용어 중 문서 판별에 유용하게 사용될 만한 내용어를 선택하는 단계이다. 학습 문서에 나타나는 내용어의 수는 수만에서 수십만에 이르기 때문에, 모든 내용어가 자질(feature)로 선택될 경우 학습 및 판별 시간이 매우 오래 걸리게 되며 성능도 보장할 수 없다. 성능 저하 없이 자질의 수를 줄이기 위하여, 학습 문서에 나타나는 내용어의 정보량을 계산하고 정보량이 큰 내용어만을 자질로 선택하려는 연구가 활발히 진행 중에 있다[1].

① **TF(Term Frequency) 이용** : 전체 텍스트 집합 중 특정 용어가 출현한 빈도

② **DF(Document Frequency) 이용** : 전체 텍스트 집합 중 특정 용어가 출현한 문서의 수를 의미하며, 일정 빈도(임계치) 이하의 텍스트에서 출현하는 용어들은 텍스트에서의 중요도가 낮다고 판단하고 이를 제거한다. DF를 이용할 때의 기본 가정은 DF가 아주 작은 용어는 특정 주체 범주를 대표할 만한 충분한 정보가 되지 못하고 전체적인 성능에도 큰 영향을 미치지 못한다는 것에 있다.

③ **MI(Mutual Information) 이용** : 두 용어 중 한 용어가 다른 용어에 대해 갖고 있는 정보량을 이용한다. 즉, 두 용어 중 한 용어가 출현했다는 사건이 다른 용어의 출현 여부를 예측하는 데 기여하는 정도를 수치적으로 나타내는 것이다.

④ **IG(Information Gain) 이용** : 특정 단어의 출현 여부가 문서 분류에 기여하는 정도를 계산하여 기여도가 높은 자질만을 선택하는 것으로, 모든 용어들의 정보 획득량을 계산하여 일정 임계값 이상의 값을 갖는 용어들만을 자질로 선택하게 된다. 문서에서의 출현 빈도뿐 아니라 출현하지 않은 빈도까지 고려하여 각 범주에서의 용어 정보량을 계산한다.

⑤ **CHI(x2 statistics) 이용** : 용어와 범주간의 의존성을 측정해 용어의 중요도를 구하는 방법으로, 두 값의 차가 클수록 용어가 자질로 선정될 확률이 높아진다. 문서 빈도를 사용해 범주별 발생분포가 일반적인 단어들의 발생분포와 다른 정도를 계산하고, 그 차이가 특정 값 이상인 단어를 자질로 선정하게 된다. 저빈도 용어에 대해서는 신뢰할 수 없다고 알려져 있다.

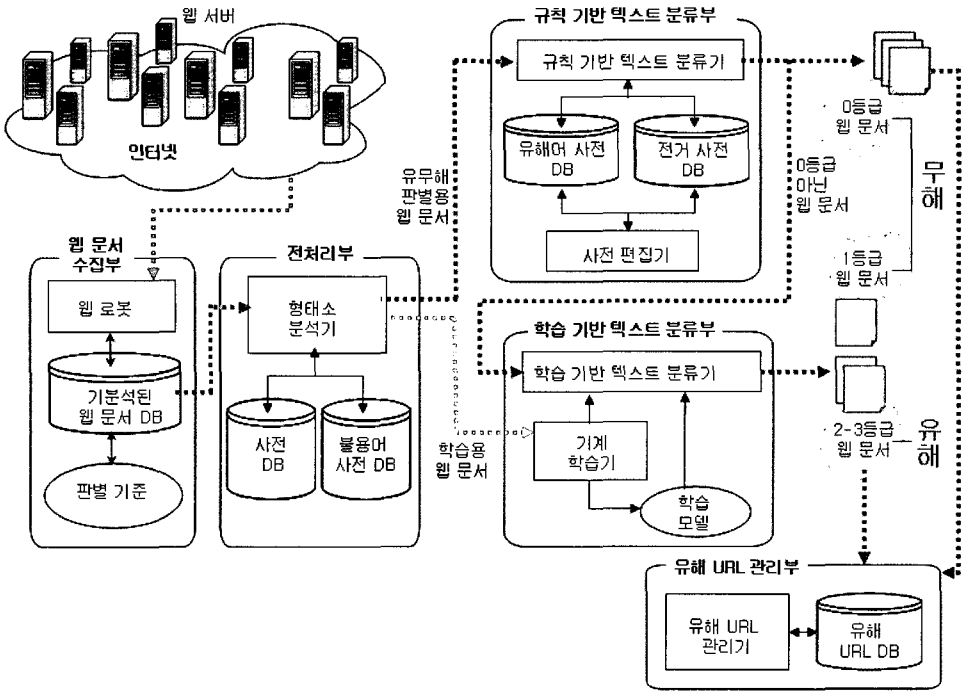
각 학습 대상 문서를 학습에 적합한 형태로 표현하기 위해서는 선택된 자질에 가중치를 부여해야 한다. 각 문서들은 전 단계에서 선택된 자질들의 값으로 표현된다. 전 단계에서 선택된 내용어(자질어)들은 자질이 되고 문서 내에서의 빈도수 등을 이용한 단어 가중치가 값이 된다. 즉, <자질:값> 형태의 표현법이 사용되며, 가장 일반적으로 사용되는 문서 표현 방법은 벡터 공간 모델이다. 이것은 문서 전체에 나타난 각 자질의 출현 빈도(TF)를 이용하여 문서를 하나의 벡터로 표현하는 것이다. 고빈도 용어는 대부분 기능어로서 많이 등장하지만 그 문서의 내용을 나타내지는 못하므로, TF와 역문헌빈도(Inverse Document Frequency)를 함께 고려하여 가중치를 부여하는 방법이 주로 사용된다. 각 자질의 가중치는 해당 문서에서 각 자질의 빈도(TF)와 역문헌빈도(IDF)의 곱으로 나타내어진다[2]. 문서 간 분리도 대신 범주 간 분리도가 높은 용어에 높은 가중치를 주는 방법으로 TF-ICF(Term Frequency-Inverse Category Frequency) 가중치 부여 방법이 있다. 이 방법은 범주 분리 능력이 우수한 색인어에 높은 가중치 부여하는 방법으로, 소수 범주에 많이 나온 용어에 높은 가중치를 주고, 여러 범주에서 고르게 나오는 용어에 대해 낮은 가중치를 부여한다.

2.2 텍스트 판별

많은 양의 문서를 관리하고 이를 효율적으로 검색하기 위한 문서 분류 모델에는 기계 학습 분야에서 사용되는 알고리즘들이 사용되는데, 여기서 제안하는 시스템은 지지 벡터 기계(SVM: Support Vector Machine)를 적용한다. 지지 벡터 기계는 두개의 범주를 구분하는 문제를 해결하기 위해 1995년에 소개된 비교적 최근의 학습 기법으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리할 수 있는 결정면(decision surface)을 찾는 모델이다[3]. 다양한 기계 학습 알고리즘 중 지지 벡터 기계를 사용한 이유는, 학습된 데이터 중 일부만을 사용하여 분류하기 때문에, 즉 지지 벡터만을 사용하여 분류하므로 분류 시간이 적게 소요되는 장점이 있고, 또한 입력 특징 벡터를 고차원으로 매핑하여 분류하므로 분류 성능이 좋고 일반화 성능도 뛰어나기 때문이다[4].

3. 효율적인 유해 웹 문서 판별 시스템

본 장에서는 판별 대상을 음란 유해 웹 문서로 한정한 유해 웹 문서 판별 시스템을 제안한다. 본 시스템은 전 장에



(그림 1) 효율적인 유해 웹 문서 판별 시스템

서 서술한 요소 기술 중 유해 웹 문서 판별에 적합한 기술들을 취하여 효율을 극대화하였다.[1] (그림 1)은 본 논문에서 제안하는 효율적인 유해 웹 문서 판별 시스템 구성과 흐름을 나타낸 그림이다.

3.1 시스템 구성

본 시스템은 웹 문서 수집부, 전처리부(형태소분석부), 규칙 기반 텍스트 분류부, 학습 기반 텍스트 분류부, 유해 URL 관리부 등 총 다섯 부분으로 구성된다.

① **웹 문서 수집부(Web-documents collection)** : 웹 문서 수집부는 학습이나 분류 테스트 시 사용되는 웹 문서들을 수집하고 이를 유해 URL DB에 저장하는 기능을 한다. 웹 로봇이 특정 웹 사이트를 방문하여 모든 웹페이지들을 데이터베이스에 기록하고 각 웹페이지를 구성하는 텍스트와 이미지들을 수집하여 특정 파일 내에 저장한다.[2]

② **전처리부(Preprocessor)** : 전처리부의 핵심 기능은 형태소 분석(Morphological Analysis)으로, 형태소 사전(morpheme dictionary)을 통하여 웹 문서의 내용을 형태소 단위로 구분한다. 모든 웹 문서들은 HTML 태그를 포함하므로 형태소분석기를 통하여 여러 개의 형태소로 나누기 전에 HTML 파싱(parsing) 작업이 필요하다. 이러한 파싱 작업과 함께 형태소분석을 수월하게 하기 위해 불용어 사전(Stop-words dictionary)을 이용한 기호(symbols) 및 불용어 제거 작업도 수행된다. 불용어 사전은 형태소 분석의 결과인 형태소 중에 "이제", "내가" 등의 의미 없는 단어들(stop-word)이나 "로그인", "버튼" 같이 너무 빈번하게 출현하는 단어들을 포함하고 있다. 이러한 의미 없는 단어들이나 출현 빈도가 높은 단어들은 향후 자질어로 추출되더라도 각 문서의 특징을 나타내는데 별로 기여하지 못하므로, 불용어 사전을 통해 별도로 관리함으로써 자질어 추출에서 배제되도록 한다.

③ **규칙 기반 텍스트 판별부(Rule-based text classification)** : 규칙 기반 텍스트 판별부는 유해어 사전과 전거 사전을 활용하여 무해 문서(특히, 0등급 웹 문서)를 가려내는 기능을 한다. 특정 웹 문서가 유해어 사전 상의 유해 단어를 포함하고 있지 않거나 또는 1~2개 정도의 유해어를 포함하고 있으면 무해한 문서로 간주하고 이를 가려낸다.[3] 규칙 기반 텍스트 판별을 위해서는 유해어 사전과 전거 사전이 필요하다. (제 3장 제 2절 참조)

④ **학습 기반 텍스트 판별부 (Learn-based text classification)** : 학습 기반 텍스트 판별부는 유해 문서를 판별하기 위하여 SVM 학습 알고리즘을 사용한다[5]. 학습 기반 텍스트 판별은 학습 과정과 분류 과정으로 크게 나눌 수 있다. 학습 과정은 자질 추출, 인덱싱, SVM 전처리, 그리고 학습 모델 생성으로 이루어진다.

1) 자질 추출과 인덱싱의 경우 다양한 알고리즘들을 모두 적용하여 구현한 후 시험을 통하여 비교하였다. 비교 결과는 제 5장 참조

2) 데이터베이스에는 페이지 리스트와 함께, 유무해 판별 기준(제 5장 <표 1> 참조)에 의하여 사람이 직접 부여한 (웹페이지) 등급도 기록된다.

3) 몇 개의 유해 단어를 포함하는 문서까지 무해 문서로 보느냐는 시험을 통한 결과 분석을 통해 결정할 수 있다.

전장에서 기술한 여러 가지 자질 추출 알고리즘 중 TF의 변형인 logTF, IG, 그리고 CHI를 적용하였다. (성능 비교 결과는 제 5장 참조) 가중치를 부여하기 위한 척도로는 TF-IDF와 TF-ICF을 모두 사용하였다. 인덱싱 과정이 끝나면, 전체 웹 문서들의 각 자질 벡터들을 정규화(normalization)한 후, SVM 기계 학습 방법을 이용하여 학습 모델을 생성한다. 분류 과정은 학습 과정에서 생성된 학습 모델과의 비교를 통하여 특정 웹 문서의 유무해 여부를 판별하는 과정으로 학습 모델과의 비교를 위하여 인덱싱 및 정규화 작업이 필요하다. 제안하는 시스템은 한글 학습 모델과 영어 학습 모델을 별도로 생성하고, 한글과 영어 웹 문서를 각각의 학습 모델과 비교하여 유무해를 판별한다.

⑤ **유해 URL 관리부 (Harmful URL management)** : 유해 URL 관리부는 유해 URL 데이터베이스를 구축하고 관리하는 기능을 한다. 이 데이터베이스는 시리얼번호, 도메인 이름, IP 주소, 패스, 제목, 문서 등급, 등급 부여날짜, 등급부여시간, 유효성검사일 등의 필드를 포함한다. 정기적으로 데이터베이스에 저장된 URL들의 유효성을 검사하는데, 이는 해당 URL이 사라지거나 또는 내용이 변경되었을 경우 이를 폐기하거나 등급을 재부여할 필요가 있기 때문이다. 이러한 유해 URL들은 유해 URL 리스트를 사용하는 다른 차단 도구나 제품에 배포되어 사용될 수 있다.

3.2 유해어 사전 및 전거 사전

3.2.1 유해어 사전 (Harmful word dictionary)

규칙기반 필터링에 반드시 필요한 요소가 단어 사전이다. 유해 웹 문서를 대상으로 할 경우에 각 언어별로 유해 단어들을 포함하는 유해어 사전이 필요하다. 일반적인 유해 단어뿐 아니라 음란한 의미로 사용되는 비속어나 은어도 포함되어야 한다. 또한, 최근 인터넷 사용자가 급격히 늘어나면서 인터넷 공간에서만 사용하는 인터넷 신조어 중에서도 음란한 의미로 사용되는 단어들이 생겨나고 있다. 판별 대상이 웹 문서인 만큼 이러한 음란성 인터넷 신조어도 반드시 포함되어야 한다. 제안하는 시스템을 위해 약 3만건의 웹 문서로부터 발췌한 한글과 영어 단어들로 유해어 사전을 구성하였다.

3.2.2 전거사전 (authority file)

자질 추출의 결과로서 선택된 자질어들은 각 문서의 특징을 나타내는 요소(factors)로 사용된다. 그러므로 이러한 자질어들을 잘 선택하기 위해서는 전거사전이 반드시 필요하다[6]. 전거 사전은 대표어와 전거단어가 매핑되어 있는 형태로 구성된다. 예를 들면, "성교"라는 대표어에는 "떡치기", 빠구리, 빠굴, 섹스, 쎅스, 쎽스, 쌕스, 씹, 씹질, 오입, 오입질, 정사, 좆질, 퍽킹, 피스톤질" 등의 전거단어가 연결되어 있다. 의미가 동일한 단어들만이 전거 사전의 대상은 아니다. 철자법이 틀렸거나 차단을 피하기 위하여 고의적으로 변형된 형태로 표기한 단어들도 그 대상이 된다. 전거 사전의 목적은 동일한 의미로 사용되는 여러 단어들이 하나의 대표어로 매핑되어 자질 선택시에 실제로 각 문서에서 중요한 의미를 갖는 단어들이 자질어로 추출될 수 있도록 하기 위함이다.

4. 구 현

제안하는 유해 웹문서 판별 시스템을 자료 관리 블록(Data Management Block), 학습 관리 블록(Learning Management Block), 그리고 등급 관리 블록(Rating Management Block) 등 3개의 블록으로 구분하여 구현하였다.

4.1 자료 관리 블록

자료 관리 블록은 웹 문서에서 텍스트 등급 분류에 불필요한 태그 정보를 제거하고 웹 문서 내의 텍스트만 걸러내는 HTML 필터 기능과 텍스트 문서에서 한글/영문의 형태소를 분석하는 형태소 분석 기능을 갖는다. 또한, 규칙기반 텍스트 분류를 위해 유해어 사전 및 전거 사전을 관리하고 이를 유해어 사전 및 전거 사전 데이터로 변환하는 유해 관련 사전 관리 기능과 형태소 분석 시에 사용되는 불용어 사전 및 형태소 사전을 관리하고, 형태소 사전 및 불용어 사전 데이터로 변환하는 형태소 사전 및 불용어 사전 관리 기능을 갖는다.

4.2 학습 관리 블록

학습 관리 블록은 학습기반 텍스트 분류에서 사용될 자질어의 목록을 생성하는 자질 추출 기능과 자질어가 각각의 학습 문서에서 특정 가중치를 갖도록 자질어의 가중치를 생성하는 인덱싱 기능을 갖는다. 또한, .인덱싱 기능에 의해서 생성된 자질어의 가중치 값을 학습기반 텍스트 분류에서 분류의 정확도를 높이기 위해 전체 자질어의 가중치 값을 -1과 1사이의 값으로 정규화 하는 가중치 정규화 기능과 학습기반 텍스트 분류 시 필요한 학습 모델을 생성하는 기능을 갖는다.

4.3 등급 관리 블록

등급 관리 블록은 유해어 사전을 기반으로 하여 텍스트 문서에 유해한 단어가 포함되어 있는지 검사하여 문서의 무해와 유해 가능성을 검사하는 규칙 기반 텍스트 분류 기능과 규칙기반 텍스트 분류 기능에서 유해 가능성이 있다고 판단된 텍스트 문서를 학습모델 생성 기능에 의해서 생성된 학습 모델을 이용하여 텍스트 문서의 유/무해 등급을 분류하는 학습 기반 텍스트 분류 기능을 갖는다. 그리고, 규칙기반 텍스트 분류 기능과 학습기반 텍스트 분류 기능을 이용하여 텍스트 문서의 등급을 결정하고, 시스템 통합 블록에 등급을 전달하는 기능도 수행한다.

5. 시험 및 결과 분석

본 장에서는 우리가 제안한 효율적인 유해 판별 시스템에 대하여 수집된 유무해 웹 문서들을 대상으로 성능을 시험한 후 그 결과를 분석한다.

5.1 시험 환경

5천 여건의 한글 및 영어 웹 문서를 수집한 후 웹 문서 등급 기준(<표 1>)에 맞게 등급을 부여하였다. 사용된 학습 문서는 한글 문서가 1,752장(무해 588장, 유해 1164장)이고 영어 문서가 1,630장(무해 694장, 유해 936장)이다. 시험 문서의 경우는 한글 문서가 971장(무해 462장, 유해 509장)이고 영어 문서가 982장(무해 533장, 유해 449장)이다.

〈표 1〉 웹 문서 등급 기준

유무해	등급	범주	세부 범주
무해	0	무해 사이트	
	1	의학 사이트	비뇨기과
			산부인과
		성상담 사이트	성상담 전문 사이트
			의학 사이트 내 성상담
		성교육/성클리닉 사이트	청소년 대상
		성관련 신문기사	스포츠 신문 사이트
유해	2	성교육/성클리닉 사이트	성인대상
		성상담 게시판	
		성교육 게시판	
		야설	정상적 성행위 묘사
	3	19세 이상 인증페이지	
		일반 성인 사이트	
		비디오 리뷰	
		야설	변태적 성행위 묘사

5.2 성능 평가 기준

얼마나 정확하게 분류되었는지를 판단하기 위하여 다음과 같은 4가지 항목을 판단 기준으로 사용하였다[7].

① Accuracy : 전체 테스트 데이터 중에서 옳게 분류된 데이터의 개수가 몇 개인지를 비율로 나타낸 것으로 <표 2>에서 **(A+D)/(A+B+C+D)**로 계산할 수 있다.

② Recall (재현율) : 실제 유해 웹 문서 중에서 제안하는 시스템이 유해로 판단한 웹 문서의 개수가 얼마나 되는지를 비율로 나타낸 것으로, <표 2>에서 **A/(A+C)**로 계산할 수 있다. Recall은 무해 웹 문서와는 상관이 없는 것을 수식으로 알 수 있다.

③ Precision(정확도) : 제안하는 시스템이 유해라고 판단한 웹 문서 중에서 실제 유해한 웹 문서의 개수가 얼마나 되는지를 비율로 나타낸 것으로, <표 2>에서 **A/(A+B)**로 계산할 수 있다. 실제로 무해한 웹 문서 중 제안하는 시스템이 유해로 판단한 웹 문서의 개수(<표 2>에서 B)가 많게 되면, 분모가 커지게 되므로 precision이 그만큼 낮아지게 된다. 그러므로 precision과 recall이 모두 높아야 정확도가 높은 시스템이라고 할 수 있다.

〈표 2〉 성능 평가 기준 개념 설명

		제안하는 시스템	
		유해	무해
실제	유해	A	C
	무해	B	D

④ F-measure: Recall 값과 Precision 값의 조화 평균을 의미하며 **(2*recall*precision)/(recall+precision)**로 계산할 수 있다.

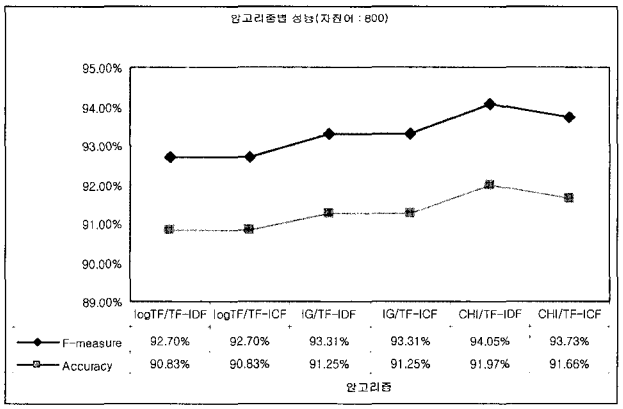
5.3 시험 방법

수집한 웹 문서를 이용하여 한글과 영어 학습 모델을 각각 생성하고, 시험 웹 문서를 제안하는 시스템에 입력하여 전 절의 4가지 항목에 대하여 성능을 측정하였다. 다음과 같은 다양한 매개변수 조절을 통하여 총 84개(주 : 자질어개수(7)*자질추출 알고리즘(3) *인덱싱 알고리즘(2)*한글/영어(2)=84)의 학습 모델을 생성하였다.

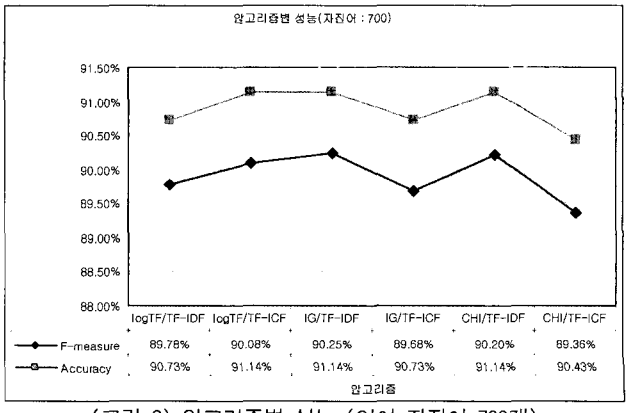
- 자질어 개수: 200개~800개 (100개 단위)
- 자질추출 알고리즘: logTF, IG, CHI 적용
- 인덱싱 알고리즘: TF-IDF와 TF-ICF 적용

5.4 시험 결과 분석

한글 웹 문서의 경우, (그림 2)와 같이 자질어 개수를 800개로 하고 CHI와 TF-IDF 알고리즘을 사용하였을 경우에 accuracy와 f-measure가 각각 91.97%와 94.05%로 가장 높



(그림 2) 알고리즘별 성능 (한글 자질어 800개)



(그림 3) 알고리즘별 성능 (영어 자질어 700개)

게 나타났다. 그리고, 자질어 개수를 조절하더라도 CHI와 TF-IDF 알고리즘으로 만든 학습 모델을 적용하였을 경우가 대체적으로 성능이 높게 나타났다.

영어 웹 문서의 경우, (그림 3)과 같이 자질어 개수를 500개로 하고 CHI와 TF-ICF 알고리즘을 사용하였을 경우와 자질어 개수를 600개로 하고 CHI와 TF-ICF 알고리즘을 사용하였을 경우에 accuracy와 f-measure가 각각 91.34%와 90.46%로 가장 높게 나타났다. 그러나, 전체적으로는 자질어 개수가 많아질수록 IG와 TF-IDF와 CHI와 TF-IDF 알고리즘으로 만든 학습 모델을 적용하였을 경우가 성능이 높은 것으로 나타났다.

6. 결 론

지금까지 효율적인 유해 텍스트 판별 시스템을 제시하고 이를 구현한 후 성능 측정 기준을 제시하고 다양한 시험을 통해 성능을 측정하여 비교 분석하였다. 시험 결과에서 보듯이 자질어의 개수, 자질 추출 방법, 인덱싱 방법 등 관련 파라미터 변화에 따라 어느 정도의 성능 차이가 있음을 알 수 있다. 그러나, 유해 웹 문서 텍스트 판별이라는 특정한 애플리케이션에 대하여 가장 성능이 좋은 자질어 개수나 자질 추출 방법, 또는 인덱싱 방법을 선택하기는 힘들 만큼의 성능 차이를 보이고 있다. 또한, 전 장의 시험 결과는 학습이나 시험을 위한 웹 문서의 개수와 내용에 따라서 그 결과가 달라지게 되므로, 장기간에 걸친 꾸준한 학습 샘플 수집과 학습 모델 생성이 이루어져야만 성능의 신뢰도를 높일 수 있다. 지속적인 피드백을 통한 지능 향상이 이루어질 경우 본 시스템의 성능은 꾸준히 증가할 것으로 기대된다.

참 고 문 헌

[1] Y.Yang and J.O.Pederson, A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp.412-420, 1997.

[2] W.Frakes and R.Baeza-Yates, Information Retrieval: Data Structures and Algorithms, Chapter7, Prentice-Hall, 1992

[3] M.Shin and C.Park, A Radial Basis Function Approach to Pattern Recognition and its Applications, ETRI Journal, Vol.22, No.2, pp.1-10, 2000.

[4] T.Joachims, Estimating the Generalization Performance of a SVM Efficiently, Proceedings of the International Conference on Machine Learning, 2000.

[5] G.Siolas, Support Vector Machines based on a semantic kernel for text categorization, IJCNN 2000, Vol.5, pp.205-209, 2000.

[6] 시소러스, http://www.minjung.net/bbs/zboard.php?id=hk221a&page=1&sn1=&divpage=1&sn=off&ss=on&sc=on&select_arrange=hit&desc=asc&no=294&PHPSESSID=91520360f59f5ba41270dc082eaf5b21

[7] 강승식, 한국어 형태소 분석과 정보 검색, 홍릉과학출판사, 2002.

김 영 수

e-mail : blitzkrieg@etri.re.kr

1998년 성균관대학교 정보공학과(학사)

2000년 성균관대학교 컴퓨터공학과 (공학석사)

2000년~현재 한국전자통신연구원 정보보호연구단 선임연구원

관심분야 : 암호학, 네트워크보안, 개인정보보호, Trusted Computing 등

남 택 용

e-mail : tynam@etri.re.kr

1987년 충남대학교 계산통계학과(학사)

1990년 충남대학교 계산통계학과(석사)

2005년 한국외국어대학교 전자정보공학과 (공학박사)

1987년~현재 한국전자통신연구원 정보보호연구단 보안게이트웨이 연구팀 팀장

2004년~현재 과학기술연합대학원대학교(UST) 정보보호공학 교수

관심분야 : 개인정보보호, 네트워크보안, 인터넷기술, 차세대네트워크구조 등

원 동 호

e-mail : dhwon@security.re.kr

1976년~1988년 성균관대학교 전자공학과 (학사, 석사, 박사)

1978년~1980년 한국전자통신연구원 전임연구원

1985년~1986년 일본 동경공업대 객원연구원

1988년~2003년 성균관대학교 교학처장, 전지전자 및 컴퓨터공학부장, 정보통신 대학원장, 정보통신기술연구소장, 연구처장

1996년~1998년 국무총리실 정보화추진위원회 자문위원

2002년~2003년 한국정보보호학회장

현 재 성균관대학교 정보통신공학부 교수, 한국정보보호학회 명예회장, 정보통신부지정 정보보호인증기술연구센터 센터장, IT보안성평가연구회 위원장

관심분야 : 암호이론, 정보이론, 정보보호