# 다중 등급 유해문서 분류를 위한 워크벤치 프로그램 구현

이원휘, *조윤정, 정성종, 안동언
전북대학교 컴퓨터공학과
e-mail : {enomi, wony, duan, sjchung}@chonbuk.ac.kr

# Implementation of Workbench Program for Multi-Level Harmful Document Classification

Won-Hee Lee, *Yun-Jeong Cho, Sung-Jong Chung, Dong-Un An
Department of Computer Engineering,
Chonbuk National University

## Abstract

유해 문서를 분류하기 위한 고정된 등급에 의한 분류가 아닌 사용자의 필요에 의해 다양한 등급으로 분류할 수 있는 분류기를 구현하였다. 자질 생성을 위해 $\chi^2$, IG, DF, ICF를 이용하였으며, 분류를 위해 나이브 베이지언, C4.5, kNN, SVM을 이용하였다.

## I. 서론

오늘날 정보화 사회가 진행되면서 정보의 제공 및 습득이 용이해지면서 청소년 등 사회적으로 보호 받아야 할 사용자들에게 해가 되는 문서 또한 급증하고, 쉽게 접할 수 있다. 이로 인한 많은 사회문제가 발생하고 있다.

이런 문제를 해결하기 위해 이미 인터넷 내용 선택에 대한 플랫폼, 영상정보나 텍스트에 기반한 내용 분류 연구 등 다양한 연구가 이루어지고 있다. 그러나 기존 연구나 제도들은 정해진 등급체계에 고정되어 있어 다양한 등급 체계를 필요로 조직이나 환경에 적절히 대처할 수 없다.

본 연구에서는 다양한 등급 체계를 필요로 하는 조직이나 사용자들에게 다양한 등급 체계를 가지고 학습하고, 학습된 결과를 토대로 문서를 분류할 수 있는 워크벤치 프로그램을 설계하고 구현한다.

## II. 관련연구

### 2.1 분류기술

분류를 위해 나이브 베이지언, C4.5, kNN, SVM을 이용한다.

나이브 베이지언 분류자는 속성값들의 결합으로 이루어진 각 인스턴스(instance) $x$와 특정 유한 집합 $V$에서 어떤 값을 갖는 목적함수 $f(x)$가 존재하는 학습 단계에 적용된다[1].

C4.5 결정트리(Decisiom Tree)는 트리형태의 규칙을 생성하여 분류를 수행하는 기계학습 방법론으로 분류규칙을 분석하기가 용이하며, 수치 벡터뿐 아니라 비수치적 데이터에 대해서도 분류규칙을 생성해 낼 수 있다[1].

kNN,(k-Nearest Neighbor)은 미리 분류된 데이터들로 참조집합을 구성하고, 새로 들어온 데이터와 가장 유사한 참조집합내의 데이터 k개의 데이터가 속한 부류의 개수를 세어 투표방식으로 결정하는 방법, 벡터 유사도를 가중치로 사용하는 방법 등을 비롯한 다양한 방법이 부류 결정 방법으로 사용된다[1].

SVM(Support Vector Machine)의 구조적 리스크 최소화를 통해 벡터공간에서의 최적의 결정경계영역을 찾아내는 것으로 이진분류문제를 푸는 방법으로 이용되고 있다. 최대 마진(margin)을 가지고 부정예제로부터 긍정예제를 분류해 낼 수 있는 결정면(decision surface)을 찾아내는 선형 분류 모형이다[1].

### 2.2 자질생성 방법

자질생성에는 $\chi^2$, IG, DF, ICF를 이용한다.

$\chi^2$는 어떤 단어 $w_i$가 문서 $d$에 출현했다는 (혹은 출현하지 않았다는) 정보가 이 문서가 범주 $C_j$에 속하는지 여부를 결정하는데 있어서 얼마나 유용한가를 측정하는 방법이다[2].

IG(Information Gain)는 특정 단어의 출현 여부가 문서 분류에 기여하는 정도를 계산하기 위하여 기여도가 높은 자질만을 선택하는 알고리즘이다[2].

ID(Document Frequency)문서 빈도에 근거한 자질 선정은 적어도 일정 개수 $\theta$ 이상의 문서에 나타나는 단어들만 자질로 사용하는 방법이다[2].

ICF(Inverted Category Frequency) 가중치 계산 방법은 어떤 범주 $C_i$에 대한 단어 $w_j$ 가중치를 계산하는데 범주 $C_i$에서 단어 $w_j$가 갖는 가중치는 다음과 같은 식으로 표현된다[2].

$$weight_{ij} = fr_{ij} * \log \frac{N}{cf_j}$$

## Ⅲ. 구현

시스템은 자질생성 모듈, 학습 모듈, 분류 모듈로 구성된다. 다음 그림은 시스템의 구성도이다.

자질생성 모듈에서는 형태소 분석기가 분석한 형태소 데이터들로부터 학습에 사용될 자질을 생성하는 모듈로서 $\chi^2$, DF, ICF, IG 기법 각각에 대하여 자질을 생성한다. 자질 생성 알고리즘과 자질 개수 등을 임의로 설정할 수 있도록 하였다.

학습 모듈에서는 자질생성 모듈에서 생성된 자질들을 이용하여 학습을 수행하게 된다. 학습은 Naïve Bayesian, C4.5, kNN, SVM 알고리즘 각각에 대하여 사용자가 원하는 알고리즘을 선택하여 학습을 수행하여 학습 모델을 생성한다.

분류 모듈에서는 학습 모듈에서 만들어진 학습 모델에 의하여 문서를 분류하여 유해 문서의 등급을 판정한다. 사용자는 생성된 학습 모델을 선택한 후 분류를 수행하게 된다. 분류는 실험을 위해 대량 문서를 분류할 수도 있고 특정 문서 하나에 대한 분류를 수행할 수 있다.

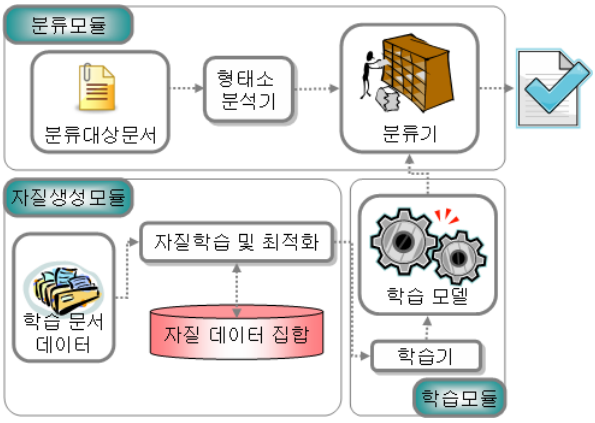


그림 1. 시스템 구성도

## Ⅳ. 결론 및 향후 연구 방향

실험은 로이터 데이터 셋을 이용하여 자질 개수 변화에 따른 결과를 얻기 위한 실험과 자질생성과 분류 기술의 조합에 의한 결과를 얻기 위한 실험을 진행하였다. 실험을 통해 자질 수를 1000으로 했을 때가 가장 좋은 성능을 보였으며, 알고리즘 조합에서는 SVM과 $\chi^2$ 조합이 가장 좋은 성능을 보여주었다.

본 연구를 통해 구현된 워크밴치 프로그램은 다양한 등급의 유해문서를 상황에 맞게 등급 수를 설정하여 분류할 수 있으며, 다양한 분류 기술과 자질 생성 알고리즘 조합을 통해 분류 시스템에 적용될 최적의 알고리즘 조합을 얻어 낼 수 있다.

향후 다양한 알고리즘을 추가하여 다양한 알고리즘 조합을 얻을 수 있는 워크밴치 프로그램으로 개선하고자 한다.

## 참고문헌

[1] Yiming Yang, Xin Liu, "A re-examination of text categorization methods", 22nd Annual International SIGIR

[2] 정정훈, 이원휘, 이신원, 정성종, 안동언, "유해어 필터링을 위한 자질어 추출 알고리즘에 관한 연구", 정보과학회하계학술대회, Vol 33, No 01, 7~9,2006.

[3] Mitchell, T.M, "Machine Learning", McGraw-Hill, 1997.