# 문서분류 기법을 이용한 웹 문서 분류의 실험적 비교

이상순, 최정민, 장근, 이병수
인천대학교 컴퓨터공학과
e-mail : sslee.gcgc.ac.kr,{cjm,mischang,bsl}@incheon.ac.kr

# Empirical Analysis & Comparisons of Web Document Classification Methods

Sang-Soon Lee, Jung-Min Choi, Chang Gun, Byoung-Soo Lee
Dept. of Computer Engineering, University of Incheon.

## 요 약

인터넷의 발전으로 우리는 많은 정보와 지식을 인터넷에서 제공받을 수 있으며 HTML, 뉴스그룹 문서, 전자메일 등의 웹 문서로 존재한다. 이러한 웹 문서들은 여러가지 목적으로 분류해야 할 필요가 있으며 이를 적용한 시스템으로는 Personal WebWatcher, InfoFinder, Webby, NewT 등이 있다. 웹 문서 분류 시스템에서는 문서분류 기법을 사용하여 웹 문서의 소속 클래스를 결정하는데 문서분류를 위한 기법 중 대표적인 알고리즘으로 나이브 베이지안(Naive Baysian), k-NN(k-Nearest Neighbor), TFIDF(Term Frequency Inverse Document Frequency)방법을 이용한다. 본 논문에서는 웹 문서를 대상으로 이러한 문서분류 알고리즘 각각의 성능을 비교 및 평가하고자 한다.

## 1. 서론

최근 인터넷은 급속도로 빠르게 발전해 나가고 있다. 매일 평균 20억 이상의 웹 문서가 증가하고 있으며, 우리는 다양한 정보를 인터넷상에서 접할수 있게 되었다. 이러한 정보는 HTML을 사용하여 웹 문서의 형식으로 인터넷상에 존재하게 된다. 따라서 웹 문서의 분류를 위한 분류기에서는 문서분류 기법을 사용하여 웹 문서의 클래스를 결정한다. 문서분류를 위한 이론 중 대표적인 알고리즘으로는 나이브 베이지안 , k-NN, TFIDF를 들 수 있으며, 나이브 베이지안 기법은 각각의 주어진 클래스에 따른 문서의 통계적 확률을 이용하여 클래스를 결정하는 기법이고, K-nn 기법은 메모리 기반 추론에 기반을 둔 기법으로서 관련 문서들 간의 근접도를 이용하여 문서분류가 이루어지는 기법이다. TFIDF방법은 문서 속에 있는 단어의 빈도수를 이용하여 문서를 분류하는 기법이다. 본 논문에서는 이러한 문서분류를 위한 세 가지 대표적인 기법을 웹 문서를 대상으로 적용하였을 때 각각의 기법에 따른 성능을 비교, 평가 하였다.

본 연구는 한국과학재단 지정 인천대학교 동북아전자물류센터의 지원에 의한 것임

## 2 기존의 문서분류 응용 시스템

웹 문서를 대상으로 문서분류 기법을 적용한 대표적인 시스템은 뉴스기사 분류 시스템, 검색엔진 시스템 등 이 있다. 다음은 이러한 문서분류 기법을 적용하여 만들어진 시스템의 사례에 대하여 기술한다.

문서분류 기법을 이용한 분류 시스템으로는 대표적으로 카네기 멜론 대학의 Personal WebWatcher가 있다. 이 시스템은 사용자의 행동을 웹 브라우저를 통해 모니터링하여, 사용자의 관심영역을 학습한 뒤, 브라우징하는 웹 문서내의 링크들에 대해 사용자 관심영역에 속하는 것들과 그렇지 않은 것들을 분류하여 관심있는 링크들만을 제안 해주는 시스템 이다. 또한, 앤더슨 컨설팅 연구실에서 만든 InfoFinder 역시 사용자의 관심 프로파일을 바탕으로 온라인 문서들에 대한 분류작업을 통해 사용자가 관심을 가질 문서들을 찾아주는 시스템이다. 이외에도 MIT 대학에서 만든 전자우편을 분류하는 Maxims, 엔터테인먼트 선별을 위한 Ringo, 뉴스기사 분류 시스템인 NewT 등이 모두 문서 분류기법을 이용한 대표적인 관련 시스템이다.

## 3. 기계학습을 위한 웹 문서 전처리 과정

### 3.1 HTML 태그 제거

웹 상에 있는 웹 문서는 HTML형식에 따라 만들어진다. 따라서 HTML 태그에 대한 처리가 필요하다. 이것은 우리가 관심있어 하는 것이 웹 문서의 태그를 제외한 내용 이므로 웹 문서에서 HTML 태그를 제거하는 작업을 해야한다. HTML 태그는 '<', '>'로 둘러싸여 있으며, 웹 문서에서 정확한 키워드의 추출을 위하여 태그의 제거 작업이 필요하다.

### 3.2 불용어 제거

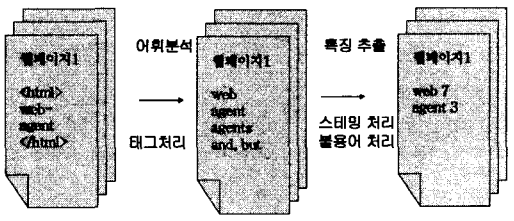
불용어는 영어의 경우 'the', 'of', 'to', 'and'는 발생 빈도가 높아서 색인어로서의 가치가 없는 단어를 말한다. 이러한 단어들은 대부분의 문서에 있어서 큰 비율을 차지한다. 문서를 색인 하는 초기 작업에서 이러한 불용어를 제거하게 되면, 계산 속도를 향상시킬 수 있고 색인에 요구되는 메모리의 용량을 줄일 수 있다. 불용어를 제거하기 위해서 주로 사용하는 방법은 불용어 사전을 만들고 문서를 색인 하는 과정에서 불용어 사전에 등록된 키워드를 모두 제외하는 방법을 사용한다.

### 3.3 스테밍(Stemming)

스테밍이란 각 키워드의 어형론적 변형을 찾아내어 동일한 의미의 여러 키워드를 하나의 키워드로 변환하는 처리이다. 영문에 있어서 단어들은 일정 의미를 갖는 어근인 스템(stem)과 단어의 형태 변화 타입인 서픽스(suffix)로 구성되어 있다. 스테밍을 하는 이유는 단어대신 어근을 저장하면 색인 파일의 크기를 줄일 수 있기 때문이다. 또한 동일한 키워드에 대해서 다르게 표현되는 것을 방지하므로 키워드의 중요도를 보다 정확하게 계산할 수 있다.

### 3.4 특징 추출(Feature selection)

특징 추출은 학습 자원의 중요 속성들을 자원이 구분된 클래스별로 다시 한번 중요도를 정의하는 특징 추출 가중치 설정 기법이다. 이를 위하여 각 학습 자원들의 특징을 고려하여 구분된 클래스들을 대상으로 일련의 구별 작업을 두어 이를 기반으로 한 속성 추출 작업을 수행할 필요가 있다. 이러한 가중치 설정 작업은 해당 키워드가 속해있는 클래스의 정보를 고려하여 이루어지며 이로써 클래스, 즉 각 카테고리를 대표하는 키워드에게 더욱 높은 가중치가 설정된다. 이러한 특징추출에 대한 기계학습 방식은 서로 다른 두 카테고리가 존재하는 경우, 각각의 카테고리 별 키워드에 가중치를 주는 것이다.




[그림 1] 문서의 전처리 과정

## 4. 실험대상 학습기법

### 4.1 나이브 베이지안

나이브 베이지안 기법은 일종의 확률 모델로서 이미 알고있는 지식을 사전 지식으로 사용하여 학습 목표인 조건부 확률을 계산하는 베이스 정리(Bayes theorem)에 그 기초를 두고 있다. 특히 이 기법은 주어진 클래스 $c_j$에 대한 각 단어 $w_t$의 출현은 서로 독립이라는 가정을 바탕으로 한다. 즉, $p(w_1, w_2, \ldots w_{|v|} \mid c_j) = \prod_{t=1} p(w_t \mid c_j)$ 이다. 따라서 클래스 $c_j$에 대한 문서 $d$의 조건부 확률은 다음과 같다.

$$P(d_i|c_j) = \prod_{t=1}^{|V|} (B_{it}P(w_t \mid c_j) + (1 - B_{it})(1 - P(w_t \mid c_j)))$$

- $B_{it}$ : 문서 $d_i$를 위한 벡터의 값 ($d_i$ = 0,1)
- $p(d_i \mid c_j)$ : 클래스 $c_j$에 문서 $d_i$가 나올 확률
- $P(w_t \mid c_j)$ : 클래스 $c_j$에 단어 $w_t$가 나올 확률

[식-1]

### 4.2 k-NN

문서분류 기법으로 대표적인 또 다른 기법으로는 k-NN 기법이 있다. 이 방법은 분류 시에 분류할 문서 Y와 저장된 클래스별 훈련 예제 X와의 거리를 식(2)에 의해서 계산한 다음, 분류대상 문서 Y와 가장 가까운 k개의 훈련 예제 X를 선정한다.

$$D_{XY} = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)2}$$

[식-2]

그리고 선정된 k개 중에서 가장 많은 개수의 훈련예제가 소속된 클래스로 분류대상 문서 Y가 분류된다. 여기서 i는 클래스의 종류이며 n은 클래스의 개수이다. k값은 k-NN기법의 성능을 최적화하기 위하여 일반적으로 교차검증(Cross Validation) 기법을 사용하여 사전에 결정하며, k=1인 경우를 NN 기법이라고 한다.

### 4.3 TFIDF

TFIDF 기법은 문서에서 단어들을 추출하여 단어목록과 가중치로 구성된 테이블을 만들고, 이 것을 이용하여 문서를 분류하는 방법이다

$$W = tf \cdot idf$$

[식-3]

- $W$ : 키워드의 가중치
- $tf$ : 현재문서의 키워드 빈도수
- $idf$ : 키워드가 포함된 문서들의 빈도수의 역

TF(term frequency)는 한 키워드가 속해있는 문서에서 나타나는 횟수를 말하며, IDF(inverse document frequency)는 DF의 역으로서, DF는 키워드가 발견된 문서들로부터 몇 개의 문서에서 나타나는가를 측정한 수치이다. IDF의 수치가 클수록 변별력이 크다는 것을 의미하며, 한 키워드의 가중치를 구하는 식은 다음과 같다.

## 5. 성능실험 및 분석

### 5.1 실험 목적 및 방법

문서 분류는 문서들의 내용을 파악하여 문서가 속한 정확한 클래스를 정하는 작업으로, 사전에 정해져 있는 클래스들 중에서 어떤 클래스에 문서가 속하는지 판단하는 것이다. 문서를 분류하기 위한 대표적인 기법으로 나이브 베이지안, TFIDF, k-NN 방법이 있다. 우리는 이러한 세가지 기법을 이용하여 웹 문서가 속한 클래스를 분류하는 성능실험을 하고, 각각의 기법에 따른 분류 성능분석을 통해 웹 문서가 분류 대상이었을 경우 어떤 기법이 가장 우수한 분류성능을 가졌는지에 대하여 알아보고자 한다. 분류 성능실험을 위한 시스템은 300MHz 펜티엄II 프로세서와 256MB의 주 기억공간을 사용하는 리눅스 환경에서 실험이 이루어졌으며, 실험을 위한 데이터는 웹상의 디렉토리 서비스를 제공하는 야후(yahoo) 검색 사이트의 분류 클래스에 따른 웹 문서를 받아서 사용하였다. 전체 클래스는 14개이며, 각각의 클래스에서 50개씩, 총700개의 웹 문서를 실험에 사용하였다. 실험은 50개씩 문서를 가지고있는 14개의 클래스에서 임의로 클래스 당 10개의 웹 문서를 분류대상 웹 문서로 발췌한다. 그리고, 나머지 클래스 당 40개의 문서들 중에서 훈련예제의 개수를 20개, 30개, 40개로 증가시키면서 세가지 기법을 비교하고, 정확도 값을 구하게 된다.

### 5.2 실험 결과

본 논문에서는 웹 문서를 대상으로 한 문서 분류 알고리즘인 TFIDF, 나이브 베이지안, k-NN 기법의 실험에 대한 결과를 [표 1]에서 나타내었다.

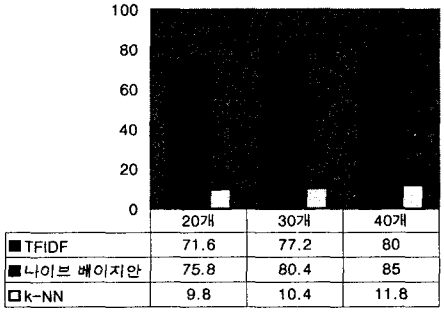
[표 1] 실험 결과 ; 분류 정확도(%)

| 훈련예 \ 학습법 | 시도 | Naive Bayesian | TFIDF | k-NN |
|---|---|---|---|---|
| 20개/클래스 | 1차 | 72 | 70 | 14 |
| | 2차 | 80 | 75 | 9 |
| | 3차 | 75 | 72 | 10 |
| | 4차 | 77 | 70 | 8 |
| | 5차 | 75 | 71 | 8 |
| 30개/클래스 | 1차 | 81 | 72 | 10 |
| | 2차 | 80 | 78 | 11 |
| | 3차 | 78 | 83 | 9 |
| | 4차 | 82 | 81 | 15 |
| | 5차 | 81 | 72 | 7 |
| 40개/클래스 | 1차 | 83 | 84 | 12 |
| | 2차 | 89 | 78 | 10 |
| | 3차 | 84 | 80 | 13 |
| | 4차 | 83 | 79 | 10 |
| | 5차 | 86 | 82 | 14 |

(분류대상 웹 문서 개수 = 10개/클래스)

[표 1]은 분류대상의 웹 문서 개수를 클래스 당 10개씩 140개로 정하고, 학습기법에 따른 분류성능을 훈련예제의 개수를 변화하면서 측정한 결과값이다. 웹 문서는 사전에 14개 클래스에 각각 속해 있는 50개의 웹 문서 중에서 임의로 선택되어 지기 때문에 실험 조건에 따라 각각 5번에 걸친 반복 실험을 통한 평균을 구하여 실험의 오차를 줄였다.

### 5.3 성능 비교 및 평가




| | 20개 | 30개 | 40개 |
|---|---|---|---|
| ■TFIDF | 71.6 | 77.2 | 80 |
| ■나이브 베이지안 | 75.8 | 80.4 | 85 |
| □k-NN | 9.8 | 10.4 | 11.8 |

[그림 2]학습기법에 따른 분류성능 비교

[그림 2]은 실험결과에 따른 각각의 기법에 대한 비교를 나타내는 그래프이다. 웹 문서를 대상으로 한 분류실험 결과는 TFIDF와 나이브 베이지안 기법이 k-NN에 비하여 월등히 우수한 성능을 보이고 있으며, TFIDF와 나이브 베이지안은 비교적 비슷한 성능을 보이고 있다. 그러나 약 5%정도 나이브 베이지안 기법이 더 나은 성능을 보이고 있으며, TFIDF에 비하여 [표 1]에서 보는 것 처럼 같은 개수의 훈련예제에서 다섯번의 반복실험 결과, 나이브 베이지안 기법이 비교적 낮은 오차의 성능을 유지하였다. 그리고, 웹 문서를 분류하는데 걸린 시간은 메모리 기반 기법인 k-NN 방법이 가장 오래 걸렸으며 나머지 두 방법은 서로 비슷한 시간이 소요되었다.

## 6. 결론 및 향후 연구 방향

본 논문에서는 문서분류 기법 중 대표적인 세가지 기법에 대하여 분류대상을 일반문서가 아닌 웹상의 웹 문서를 대상으로 하였을 때 분류 성능에 있어서 어떤 기법이 우수한 결과를 보이는지에 대하여 측정하고 비교하였다. 실험결과에 따르면 가장 우수한 성능을 보인 기법은 나이브 베이지안 기법이다. 웹 문서는 대부분 HTML 형식으로 만들어져 있으며, 일반문서와는 다른 형태로 되어 있으므로, 문서분류 기법을 위한 문서의 전처리 과정이 일반문서와는 다르게 처리된다. 따라서 웹 문서의 분류 성능과 정확도를 높이기 위해서 앞으로 시행되어야 할 향후 연구로서 요구되는 부분은 웹 문서의 전처리 과정에 대한 연구이며, 나아가 본 논문의 분류성능 측정을 통하여 분류 응용 시스템에서 보다 향상된 분류성능의 발전이 기대된다.

[참고문헌]

[1]임윤택, 윤충화 ' 고정 분할 평균법에 기반한 점진적 알고리즘' 정보처리학회 가을 학술발표논문집 제6권 제1호, pp.559 1999

[2] Jeffrey M. Bradshaw " Software Agent" AAAI Press/The MIT Press pp151-161

[3] McCallum, A. Nigam, K. 1998 " A Comparison of Event Models for Naï ve Bayes Text Classification" In AAAI-98 Workshop on Learning for Text Categorization., *http://www.cs.cmu.edu/~mccallum.*