제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

# 블로그 포스트 자동 품질 평가를 위한 기계학습 기법 비교 연구

한범준*, 김민정**, 이형규**
*고려대학교 컴퓨터 정보통신대학원
**고려대학교 컴퓨터·전파통신공학과
e-mail:bjhan93@gmail.com, mjkim@nlp.korea.ac.kr, hglee@nlp.korea.ac.kr

# A Comparison of Machine Learning Techniques for Evaluating the Quality of Blog Posts

Bum-Jun Han*, Min-Jeong Kim**, Hyoung-Gyu Lee**
*Graduate School of Computer & Information Technology, Korea University
**Dept. of Computer & Radio Communications Engineering, Korea University

요 약

블로그는 다양한 주제 분야에 대한 내용을 자유롭게 표현할 수 있는 일종의 개인 웹사이트로, 많은 양과 다양성으로 매우 중요한 정보원이 될 수 있다. 블로그는 생산속도가 매우 빠르므로 보다 고품질의 블로그를 선별하는 것이 중요하다.

본 논문에서는 블로그의 본문을 담고 있는 포스트를 대상으로 기계학습 기법을 이용하여 문서의 품질을 자동으로 평가하고자 하였다. 학습을 위한 자질로는 모든 블로그에 공통적으로 적용할 수 있도록 형태소 분석에서 추출한 동사, 부사, 형용사의 내용어만을 선택하였다. 성능 비교를 위해 수작업으로 약 4,600개의 정답 집합을 구축하고, 적합한 기계학습 기법을 찾기 위해 다양한 학습 기법을 사용하여 비교 실험하였다. 실험 결과 Bagging 기법의 성능이 79% F-measure로 가장 좋음을 보여주었다. 한정된 자질을 사용했을 때와 정답 집합의 문서 수 비율이 불균등할 경우 단순함, 유연성, 효율성의 특징을 지닌 Bagging 기법이 적합할 것으로 보인다.

## 1. 서론

블로그는 개인의 일상생활에서부터 정치, 경제, 사회, 기술 등 다양한 주제 분야에 대한 정보 전달 및 개인적 견해를 자유롭게 표현할 수 있는 일종의 개인 웹사이트이다. 많은 양과 다양성, 그리고 유행이나 사람들의 관심사를 직접적으로 반영하기 때문에 블로그는 매우 중요한 정보원이 될 수 있다.

2009년 Technorati의 “State of the Blogosphere 2008 report”에 따르면 매일 90만개 이상의 블로그가 포스트 되고 있다[17]. 이처럼 블로그는 생산 속도가 매우 빠르므로 보다 고품질의 블로그를 선별하는 것이 중요하다. 그러나 웹 문서 평가와 관련하여 지식 질의응답 문서 품질에 관한 연구[2],[11]는 있었으나, 블로그의 품질 평가에 대한 연구는 상대적으로 많이 이루어지지 않았다.

품질 평가 연구는 문서 분류의 문제로 접근할 수 있다. 기계학습 기법을 이용한 문서 분류를 하기 위해서는 문서를 자질로 표현하고, 추출된 자질을 적절한 기계 학습 기법을 이용하여 분류하는 과정이 필요하다[5].

블로그는 내용이나 형식이 매우 다양하므로 모든 블로그에 공통적으로 적용할 수 있는 자질이 한정되어 있다. 본 논문에서는 문서의 주제 분류가 아니고, 문장의 유창함과 논리 정연함 등을 평가하기 때문에 모든 블로그에 적용할 수 있는 동사, 부사, 형용사를 자질로 사용한다.

기존의 지식 질의응답 문서 품질에 관한 연구에서는 Maximum Entropy, Support Vector Machine과 같은 특정 기계학습 기법만을 적용하였다[6],[7],[14]. 본 논문에서는 제한된 자질을 보다 효과적으로 적용할 수 있는 기계학습 기법을 찾기 위해 이전 연구보다 다양한 기계학습 기법을 적용, 비교해 보고자 한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 살펴보고, 3장에서는 제안하는 방법의 자질 추출, 문서 색인, 자질 선택과 기계학습 기법에 대해 간략히 소개한다. 4장에서는 실험의 결과를 토대로 분석을 하고, 마지막 5장에서는 결론 및 향후 연구에 대해 언급한다.

## 2. 관련 연구

문서 품질에 관해 [6],[7]은 신뢰도 자질을 이용한 지식 검색 문서의 품질 평가에 관한 연구에서 기존의 비텍스트 정보를 이용한 품질 평가와 다르게 내용의 신뢰도 측면에서 평가하는 방법을 제안하였다. 제안한 신뢰도 자질은 연결어의 출현 비율, 가치판단어의 출현 비율, 웹사이트 주소 출현 여부 등이며 어휘사전을 기반으로 측정하였다. 문서 품질 평가 모델로서 Maximum Entropy 기법을 이용하여 비텍스트 자질에 비해 약 1% 정도의 성능을 개선하였다. 이러한 방법은 충분한 양의 사전을 구축하는 데에 많은 시간을 필요로 한다는 단점이 있다.

[9]는 의학 분야의 고품질 문서 검색을 위한 텍스트 분류 모델에 관한 연구에서 ACP Journal Club에서 작성한 고품질의 문서와 내용에 대한 라벨이 있는 문서를 학습시켜 품질 필터를 구축하였다. 이를 PubMed clinical 질의 필터와 비교, 실험한 결과 Polynomial Support Vector Machine Models이 가장 좋은 성능을 보여주었다. 그러나 이미 검증된 고품질 문서가 있고, 정해진 주제 분야가 있다는 점에서 다양한 주제에 대해 자유로운 형식을 가지고 있는 블로그 문서의 품질 평가에는 적합하지 않다.

문서 분류에 관해서는 온라인상의 텍스트 집합으로부터 의견이나 감정 정보를 추출하여 문서 감정 분류와 의견 분류에 관한 연구가 많이 이루어지고 있다. [8]은 형태소 분석을 통한 명사, 동사 등의 내용어 자질과 긍정, 부정의 감정 자질과의 비교 실험을 하였다. Support Vector Machine을 사용한 실험에서 감정 자질을 사용했을 때 형태소 분석을 통해 추출된 내용어 자질보다 더 좋은 성능 향상이 있었다. 하지만 긍정과 부정이라는 명확한 기준이 있는 자질은 의견 문서 분류에는 적합하지만 문서에 대한 품질 평가에는 적용하기 어렵다.

## 3. 제안하는 방법

제안하는 방법은 크게 학습 단계와 문서 분류 단계로 나눠진다. 학습 단계는 자질 추출 및 선정, 문서 색인, 자질 선택 과정으로 세분된다.

### 3.1. 자질 추출 및 선정

문장의 내용이나 특징을 잘 반영하는 단어를 내용어라고 한다. 본 논문에서는 문서의 주제 분류가 아닌 문장의 유창함과 논리 정연함 등을 평가하기 때문에 모든 블로그에 적용할 수 있는 동사, 부사, 형용사를 자질로 사용한다.

내용어로 가장 많이 등장하는 명사를 제외한 이유는 문서의 주제 분류에는 명사가 유용하겠지만, 품질 평가에는 적합하지 않다고 판단했기 때문이다. 또한 논리적이고 성실한 문서는 '그리고, 그러나, 그러므로'와 같은 부사, '나타내다, 덧붙이다' 등의 동사, '다르다, 어렵다' 등의 형용사를 많이 사용할 것이라는 가정을 하였기 때문이다.

따라서 본 논문에서는 형태소 분석 결과 동사, 부사, 형용사의 내용어를 자질로 사용한다.

### 3.2. 문서 색인

선택된 자질을 사용하여 문서를 표현하기 위한 일반적인 방법은 벡터 공간 모델이다. 본 논문에서는 가장 많이 알려진 방법 중 하나인 수식 (1)의 TF-IDF 가중치 방법을 적용하여 문서를 표현한다[1].

$$tf-idf_{t,d} = tf_{t,d} \times \log\frac{N}{df_t} \qquad (1)$$

$tf_{t,d}$ : 문서 d 내 특정 단어 t의 빈도수
N : 말뭉치 내 문서 수
$df_t$ : 말뭉치 내에서 단어 t가 출현한 문서 수

### 3.3. 자질 선택

시스템의 성능을 고려하여 전체 내용어 중 수식 (2)의 카이제곱통계량을 이용하여 자질을 선택한다[16].

$$x^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \qquad (2)$$

A : 범주 c에 속해 있는 문서 중 용어 t를 포함하는 문서의 수
B : 범주 c외의 범주에 속해 있는 문서 중 용어 t를 포함하는 문서의 수
C : 범주 c에 속해 있는 문서 중 용어 t를 포함하지 않는 문서의 수
D : 범주 c외의 범주에 속해 있는 문서 중 용어 t를 가지고 있지 않은 문서의 수
N : 전체 학습 문서의 수

### 3.4. 기계학습 기법

본 논문에서는 서론에서 언급한 바와 같이 블로그 포스트 자동 품질 평가에 적합한 기계학습 기법을 모색하기 위해 다양한 기법을 사용한다. 각 기법에 대한 간략한 내용은 <표 1>과 같다.

<표 1> 기계학습 기법

| 학습 기법 | 내 용 |
|---|---|
| NaiveBayes | 간단하면서도 좋은 성능을 내는 알고리즘으로 문서를 이루는 각 단어들이 서로 조건부 독립(Conditionally Independence)이라는 가정을 전제로 한다[12]. |
| SVM (Support Vector Machine) | 두 개의 범주를 구분하는 문제를 해결하기 위해 두 개의 클래스의 구성 데이터들을 가장 잘 분리할 수 있는 결정면을 찾는 모델이다[1]. |
| KNN (k-Nearest Neighbor) | 실험 문서가 주어졌을 때 학습 문서 중에서 테스트 문서와의 유사도가 가장 높은 k개의 문서를 추출하여 각 후보 범주의 순위를 매기는 방법이다[1]. |
| AdaBoost | 학습 집합 중 학습 결과의 성능 저하를 유발시키는 특정 패턴에 가중치를 부여하여 다시 학습시킨 후 결과를 단계별로 결합하는 방법이다[4]. |
| Bagging (Boostrap Aggregating) | 학습 집합으로부터 샘플 데이터를 추출하여 여러 개의 복사본을 생성하고 이들을 각각 학습시켜 결과를 결합시키는 알고리즘이다. 학습 데이터의 작은 변화가 분류 결과에 큰 영향을 미치는, 불안정한 학습 알고리즘(decision tree 등)을 사용할 경우 정확도를 향상시킬 수 있는 기법이다[10]. |

| 학습 기법 | 내 용 |
|---|---|
| J48(C4.5) | Quinlan에 의해 제안된 대표적 의사 결정 트리 기반의 기법으로 범주형 속성에만 사용가능한 ID3 알고리즘을 개선하여 수치형 속성 처리, missing value, 노이즈 데이터 처리, 가지치기 등의 기능을 추가한 것이다[15]. |
| Random Forest | 많은 결정 트리들로 구성된 분류기이며, 개별 트리들의 결과 클래스들의 최빈값(Mode) 클래스를 출력으로 내보낸다[3]. |

## 4. 실험 및 결과

4.1. 실험 환경

실험을 위해 웹에서 임의로 수집한 약 4,600개의 블로그 포스트를 정보성, 가독성을 고려하여 수작업으로 A(상), B(중), C(하) 평가하여 <표 2>와 같이 정답 집합을 구축하였다. 학습 집합에서 추출한 내용어 자질의 개수는 3,105개이다.

<표 2> 정답 집합

| 등급 | A | B | C | 계 |
|---|---|---|---|---|
| 문서수 | 918 | 206 | 3,480 | 4,604 |
| 비율 | 20% | 4% | 76% | |

기계학습에 사용된 도구는 WEKA이며, 구축한 정답 집합을 WEKA의 입력 포맷인 (그림 1)과 같은 ARFF (Attribute Relation File Format) 형태로 표현하였다.

```
@relation 'BlogData
@attribute docnum numeric
@attribute class {A,B,C}
…
@attribute 그러나 numeric
@attribute 나타나 numeric
@attribute 뛰어나 numeric
…
@data
{0 38,1 C,77 0.693147,87 0.693147,110 0.693147,130
0.693147,204 0.693147,211 0.693147,274 0.693147,295
0.693147,296 0.693147,297 0.693147,298 0.693147,299
0.693147}
{0 749,1 C}
{0 585,3 0.693147,8 0.693147,51 0.693147,72
0.693147,87 0.693147,89 0.693147,102 0.693147,151
0.693147,153 0.693147,172 0.693147,193 0.693147,204
0.693147,208 0.693147,231 0.693147,256 0.693147,382
0.693147,386 0.693147,442 0.693147,816 0.693147,919
0.693147,968 0.693147,1141 0.693147,1142 0.693147,1143
0.693147}
…
```

(그림 1) ARFF 포맷

학습 및 실험은 10-fold cross validation 방법을 사용하고, 평가 척도는 기계 학습을 이용한 문서 분류에서 일반적으로 사용하는 정확도(Precision)와 재현율(Recall)을 사용하였다. 최종 성능은 정확도와 재현율을 하나의 값으로 표현해주기 위한 F-measure를 사용하였다. 각각의 공식은 수식 (3)과 같다.

$$p = \frac{\text{정확하게 긍정으로 분류된 문서 수}(TP)}{\text{긍정으로 분류된 전체 문서 수}(TP+FP)}$$

$$r = \frac{\text{정확하게 긍정으로 분류된 문서 수}(TP)}{\text{정답 문서 수}(TP+FN)}$$

$$f = \frac{2pr}{p+r} \quad (3)$$

4.2. 실험 결과 및 분석

자질 선택 비율을 다양하게 실험해 본 결과 카이제곱 통계량 상위 10%의 자질을 사용하는 것이 최적이었으며, 이 때 실험 결과는 <표 3>과 같다.

<표 3> 비교 실험 결과

| 학습 기법 | 문서 등급 | 정확도 | 재현율 | F-measure | 오분류율(%) |
|---|---|---|---|---|---|
| NaiveBayes | A | 0.518 | 0.735 | 0.608 | 23.0435 |
| | B | 0.185 | 0.107 | 0.136 | |
| | C | 0.893 | 0.818 | 0.853 | |
| | 전체 | 0.787 | 0.770 | 0.773 | |
| SVM | A | 0.130 | 0.003 | 0.006 | 24.5217 |
| | B | 0 | 0 | 0 | |
| | C | 0.758 | 0.995 | 0.86 | |
| | 전체 | 0.6 | 0.755 | 0.653 | |
| KNN | A | 0.546 | 0.177 | 0.268 | 25.6087 |
| | B | 0.08 | 0.078 | 0.079 | |
| | C | 0.79 | 0.931 | 0.855 | |
| | 전체 | 0.711 | 0.744 | 0.704 | |
| AdaBoost | A | 0.502 | 0.453 | 0.476 | 23.2609 |
| | B | 0 | 0 | 0 | |
| | C | 0.825 | 0.894 | 0.858 | |
| | 전체 | 0.724 | 0.767 | 0.745 | |
| **Bagging** | **A** | **0.69** | **0.498** | **0.578** | **18.1739** |
| | **B** | **0.5** | **0.005** | **0.01** | |
| | **C** | **0.84** | **0.95** | **0.891** | |
| | **전체** | **0.795** | **0.818** | **0.79** | |
| J48(C4.5) | A | 0.615 | 0.445 | 0.516 | 20.413 |
| | B | 0.048 | 0.01 | 0.016 | |
| | C | 0.834 | 0.933 | 0.881 | |
| | 전체 | 0.756 | 0.796 | 0.771 | |
| **Random Forest** | **A** | **0.629** | **0.504** | **0.56** | **20.0217** |
| | **B** | **0.14** | **0.039** | **0.061** | |
| | **C** | **0.842** | **0.921** | **0.88** | |
| | **전체** | **0.769** | **0.8** | **0.78** | |

<표 3> 우측의 오분류율은 잘못 분류된 문서들의 비율을 나타낸다. Bagging 기법의 오분류율은 18.1739%, Random Forest 기법은 20.0217%로 다른 기계학습 기법에 비해 낮은 수치를 보여준다. 또한 Bagging과 Random Forest 기법의 F-measure는 각각 0.79와 0.78로 SVM이나 KNN과 같이 최근 성능이 우수하다고 알려진 기계학습 기법에 비해 좋은 성능을 나타낸다.

특히, B등급 문서의 경우 정답 집합의 수가 전체의 약 4%밖에 되지 않지만 Bagging 기법은 0.5%의 정확도로

다른 기법에 비해 월등히 높음을 알 수 있다.

Bagging 기법이 높은 성능을 보이는 이유는 주제 분류와 같이 특정한 패턴이 없이 내용어 자질 만으로 학습을 했기 때문으로 보인다. 즉, Bagging 기법의 특징인 단순함, 유연성, 효율성으로 인해 한정된 자질을 사용했을 때 유의한 결과를 보인다고 판단된다[13].

또 다른 이유는 정답 집합의 등급별 문서 수의 비율 불균형으로 생각된다. Bagging 알고리즘은 관련 연구에서 언급한 바와 같이 예측력 및 정확도 향상을 위해 다양한 복사본을 생성하고 이를 결합하여 오분류된 부분을 보강한다. 이러한 복사본 생성과 결합이라는 부분이 학습 시 정답 집합의 불균형을 해소시켜 다른 기계학습 기법에 비해 좋은 성능을 나타낸 것으로 판단된다.

## 5. 결론 및 향후 연구

본 논문에서는 문장의 유창함과 논리 정연함을 고려할 수 있는 동사, 부사, 형용사의 내용어를 자질로 하여 블로그 포스트 품질의 자동 평가에 적합한 기계학습 기법을 찾기 위해 다양한 학습 기법을 비교, 실험하였다.

실험 결과는 한정된 내용어만으로 기계학습을 했을 때와 정답 집합의 등급별 비율이 균등하지 않을 경우 Bagging 기법이 가장 적합함을 보여준다.

제안한 방법은 주제와 형식에 구애받지 않고 사용할 수 있는 자질을 사용했기 때문에 블로그뿐만 아니라 웹상의 다른 여러 종류의 문서 평가 분류에도 적용할 수 있을 것으로 생각한다.

추후 조사, 어미와 같은 형식어 자질을 내용어 자질과 비교, 조합해보는 연구와, 등급별 문서 비율을 비슷하게 정답 문서 말뭉치를 구축해서 본 논문의 실험 결과와 상호 성능 평가를 하는 연구가 필요하다.

또한 Bagging 기법 외에 더 좋은 성능을 보이는 기법을 알아보고, 자질 선택 시 본 논문에서 사용한 카이제곱 통계량 외에 다른 자질 선택 방법을 사용하여 필터링 하였을 경우와 비교, 평가하는 연구도 의미가 있을 것이라 생각한다.

## 참고문헌

[1] 고영중, 서정연, “문서관리를 위한 자동문서범주화에 대한 이론 및 기법”, 정보관리연구, Vol. 33, No. 2, pp. 19-32, 2002.

[2] 박소연, 이준호, 전지운, “지식 검색 서비스 개선을 위한 문서의 적합도 및 신뢰도 분석”, 한국문헌정보학회지, 제40권 제2호, pp. 299-314, 2006.

[3] 박수혁, “기계학습 기법을 이용한 문장경계인식”, 고려대학교 컴퓨터정보통신대학원, 2008.

[4] 신현정, 장민, 조성준, 이봉기, 임용업, “앙상블 학습알고리즘의 일반화 성능 비교 : OLA, Bagging, Boosting”, 한국정보과학회 봄 학술발표논문집, 제27권 제1호(B), pp. 226-228, 2000.

[5] 이경찬, 강승식, “자질 중요도 계산 기법에 의한 자동문서 범주화”, 한국정보과학회 봄 학술발표논문집, 제30권 제1호(B), pp. 537-539, 2003.

[6] 이정태, 송영인, 임해창, “신뢰도 자질을 이용한 지식검색 문서의 품질 평가”, 한국정보과학회 언어공학연구회 학술발표 논문집, pp. 62-67, 2007.

[7] 이정태, 송영인, 박소영, 임해창, “텍스트 신뢰도 자질 기반 지식 질의응답 문서 품질 평가 모델”, 정보과학회논문지: 소프트웨어 및 응용, 제35권 제10호, pp. 608-615, 2008.

[8] 황재원, 고영중, “문장 감정 강도를 반영한 개선된 자질 가중치 기법 기반의 문서 감정 분류 시스템”, 정보과학회논문지: 소프트웨어 및 응용, 제36권 제6호, pp. 491-497, 2009.

[9] Yindalon Aphinyanaphongs, Ioannis Tsamardinos, Alexander Statnikov, Douglas Hardin, and Constantin F. Aliferis, "Text categorization models for high quality article retrieval in internal medicine", Journal of the American Medical Informatics Association, Vol. 12, No. 2, pp. 207-216, 2004.

[10] Leo Brieman, “Bagging Predictors”, Machine Learning, Vol. 24, No. 2, pp. 123-140, 1996.

[11] Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park, "A Framework to Predict the Quality of Answers with Non-textual Features", In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 228-235, 2006.

[12] Tom M. Mitchell, "Machine Learning", McGraw-Hill, 1997.

[13] Mrutyunjaya Panda, and Manas Ranjan Patra, "Ensemble of Classifiers for Detecting Network Intrusion", International Conference on Advances in Computing, Communication and Control, pp. 510-515, 2009.

[14] Jun Suzuki, Yutaka Sasaki, and Eisaku Maeda, "SVM answer selection for open-domain question answering", Proceedings of the 19th international conference on Computational linguistics, pp. 1-7, 2002.

[15] Ian H. Witten and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques", 2nd ed., Morgan Kaufmann, 2005.

[16] Yiming Yang, and Jan O. Pedersen, "A comparative study on feature selection in text categorization", Proceedings of the 14th International Conference on Machine Learning, 1997.

[17] http://technorati.com/blogging/state-of-the-blogosphere/