낚시성 인터넷 신문기사 검출을 위한 특징 추출

Feature Extraction to Detect Hoax Articles

저자 (Authors)
허성완, 손경아
Seong-Wan Heo, Kyung-Ah Sohn

출처 (Source)
정보과학회논문지 43(11), 2016.11, 1210-1215 (6 pages)
Journal of KIISE 43(11), 2016.11, 1210-1215 (6 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/Article/NODE07049999

APA Style
허성완, 손경아 (2016). 낚시성 인터넷 신문기사 검출을 위한 특징 추출. 정보과학회논문지, 43(11), 1210-1215.

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 43, No. 11, pp. 1210-1215, 2016. 11
http://dx.doi.org/10.5626/JOK.2016.43.11.1210

낚시성 인터넷 신문기사 검출을 위한 특징 추출

(Feature Extraction to Detect Hoax Articles)

허 성 완 †　　손 경 아 ††
(Seong-Wan Heo)　　(Kyung-Ah Sohn)

요 약 스마트 기기의 발달로 많은 사람들이 인터넷 신문기사를 이용하고 있다. 하지만 인터넷 언론사 간의 치열한 경쟁으로 조회수를 올리기 위한 낚시성 기사가 범람하고 있다. 낚시성 신문기사는 제목을 통해 올바른 기사의 줄거리가 제공되지 않았을 뿐만 아니라, 독자로 하여금 잘못된 내용을 떠올리게 한다. 낚시성 신문기사는 핵심에서 벗어난 유명인사 인용, 애매한 문장의 마무리, 제목과 내용의 불일치 등의 특징을 갖는다. 본 논문에서는 이러한 낚시성 기사를 분류하기 위한 특징을 추출하고 성능을 검증해 본다. 기사에 달린 댓글의 키워드를 활용하여 대용량 학습데이터를 생성하고 이를 기반으로 다섯 가지 분류 특징을 추출하였다. 추출된 특징들은 서포트 벡터 머신 분류기를 이용한 실험에서 92%의 정확도를 보여 낚시성 인터넷 신문기사를 분류하는데 적합하다고 판단된다. 뿐만 아니라 제목과 본문의 일관성을 측정하기 위한 전처리 방법으로 고안한 선택적 바이그램 모델은 낚시성 인터넷 신문기사 분류 외에도 일반적인 단문 분석을 위한 전처리 방법으로 유용할 것으로 기대된다.

키워드: 낚시성 기사, 특징 추출, 문서 분류, 텍스트 마이닝, 선택적 바이그램

Abstract Readership of online newspapers has grown with the proliferation of smart devices. However, fierce competition between Internet newspaper companies has resulted in a large increase in the number of hoax articles. Hoax articles are those where the title does not convey the content of the main story, and this gives readers the wrong information about the contents. We note that the hoax articles have certain characteristics, such as unnecessary celebrity quotations, mismatch in the title and content, or incomplete sentences. Based on these, we extract and validate features to identify hoax articles. We build a large-scale training dataset by analyzing text keywords in replies to articles and thus extracted five effective features. We evaluate the performance of the support vector machine classifier on the extracted features, and a 92% accuracy is observed in our validation set. In addition, we also present a selective bigram model to measure the consistency between the title and content, which can be effectively used to analyze short texts in general.

Keywords: hoax article, feature extraction, document classification, text mining, selective bigram

· 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2014R1A1A3051169)

† 비 회 원 : 아주대학교 소프트웨어특성화학과
pwshoot@naver.com
†† 종신회원 : 아주대학교 소프트웨어학과 교수(Ajou Univ.)
kasohn@ajou.ac.kr
(Corresponding author임)

논문접수 : 2016년 3월 15일
(Received 15 March 2016)
논문수정 : 2016년 8월 3일
(Revised 3 August 2016)
심사완료 : 2016년 8월 9일
(Accepted 9 August 2016)

정보과학회논문지 제43권 제11호(2016. 11)

1. 서 론

한국인터넷진흥원(KISA)의 인터넷 이용 실태 조사에 따르면 2015년 7월 전국 만 3세 이상 인구의 인터넷 이용률은 85.1%로 대부분의 국민이 인터넷을 이용하는 것으로 조사되었다. 인터넷 이용자의 92.3%는 하루 1회 이상 인터넷을 사용하고, 일평균 약 2시간 동안 인터넷을 이용하는 것으로 조사되었고, 최근 3년간 인터넷을 통한 미디어 이용률은 꾸준히 증가하는 추세이다[1]. 통계청 사회조사 결과에 의하면 인터넷신문 구독률은 2009년에 종이신문을 추월했다. 2015년 사회조사에 따르면 종이신문을 보는 비율은 43.1%인 반면 인터넷 신문은 86.0%로 인터넷 신문의 이용률이 종이신문의 두 배로 나타났다[2]. 이는 특히 스마트폰 사용의 확산으로 인한 결과로 보인다.

인터넷 신문 이용률이 증가함에 인터넷 신문을 제공하는 언론사들은 트래픽 경쟁에 돌입하였다. 종이 신문과는 달리 인터넷 신문은 트래픽을 기준으로 광고 매출을 창출하기 때문이다. 이러한 과도한 트래픽 경쟁은 독자의 클릭을 유도하기 위한 목적으로 생산된 자극적이고 선정적인 낚시성 기사의 범람으로 이어지고 있다[3]. 종이 신문의 구조와 달리 인터넷 신문은 기사의 제목과 기사가 동시에 노출되지 않는다. 따라서 인터넷 신문기사를 보는 독자는 기사의 제목에 높은 의존도를 보인다. 이러한 인터넷 신문기사의 특성을 이용하여, 많은 인터넷 언론에서는 제목과 실제 기사의 내용이 불일치하는 낚시성 신문기사를 작성하여 이용자의 클릭을 유도하고 있다[3,4].

인터넷 낚시성 신문기사는 '의문유발형', '감정적,선동적, 외설적, 과장적 표현사용', 그리고 속보, 긴급, 단독 등의 색인을 추가하는 등의 유형으로 나누어 연구된 바 있다[3]. 해당 연구에 따르면 의문유발형의 형태의 낚시성 기사가 가장 많은 것으로 조사되었고, 이는 종이 신문에 작성된 기사를 편집한 것으로 30%의 기사가 사후 편집을 통해 낚시성 기사로 둔갑한다고 밝히고 있다.

본 연구에서는 텍스트 마이닝 기법을 활용하여 낚시성 인터넷 신문기사와 일반 기사를 분류하는 방법을 제안한다. 국내 대형 포털 사이트에서 인터넷 신문기사를 수집하였고, 기사에 달린 댓글을 분석하여 낚시성 기사와 일반기사를 구분한 데이터셋을 구축하였다. 낚시성 기사 분류에 효율적인 다섯 가지 주요 특징을 선정하여 추출하였고 분류 실험을 통해 이의 성능을 검증하였다.

2. 실험 방법

실험은 데이터 전처리, 특징 추출, 분류 분석의 세 단계로 진행된다. 2.1절에는 실험에 사용한 데이터를 소개하고, 2.2절에서는 데이터 전처리 방법과 댓글을 분석하여 낚시성기사와 일반기사로 구분된 데이터셋 구축 방법에 대해 설명한다. 2.3절에서는 특징을 추출하는 방법에 대해 설명하고, 2.4절에서는 성능 검증 실험 방법을 소개한다.

2.1 실험 데이터

실험을 위해 포털 사이트 'nate.com'의 뉴스 카테고리에서 2009년부터 2014년까지의 기사 중 사용자가 작성한 댓글이 존재하는 신문 기사를 수집하였다. 일반 기사와 낚시성 기사가 구분된 데이터셋을 자동으로 구축하기 위해 사용자의 댓글이 달린 신문 기사만 수집하여 활용하였다. 실험에 사용된 데이터에 대한 간략한 개요는 표 1과 같으며, 총 82,321편의 신문기사가 수집되었다.

표 1 수집된 데이터
Table 1 Collected data

Article-written period	2009 - 2014
Number of collected articles	82,321
Average number of replies per article	7.5

2.2 데이터 전처리 및 데이터셋 구축

본 절에서는 수집한 인터넷 신문기사 텍스트를 형태소 분석을 통해 분석하기 용이한 형태로 가공하는 과정과, 기사에 달린 댓글을 분석하여 일반 기사와 낚시성 신문 기사를 구분한 데이터셋 구축방법에 대해 설명한다.

2.2.1 형태소 분석을 통한 데이터 전처리

이 연구에서는 인터넷 신문기사의 텍스트 부분만 수집하여 분석에 활용한다. 형태소 분석을 통해 문장 요소 중 조사, 관형사, 접두사, 접미사 등 의미 분석에 불필요한 요소를 제거하였다[5]. 형태소 분석을 위해는 MeCab 엔진을 사용한 오픈소스 기반의 한국어 형태소 분석기[6]를 사용하였다.

2.2.2 댓글 키워드 분석을 통한 학습 및 검증 데이터셋 구축

낚시성 신문기사 특징을 추출하기 위해서는 일반 기사와 낚시성 기사로 구분되어 모델 구축에 이용되는 학습 데이터가 필요하다. 수작업으로 각 기사가 일반 기사와 낚시성 기사 중 어느 그룹에 속하는지 정하는 방식으로는 대용량의 학습 데이터를 얻기 어렵다. 이를 극복하기 위해 본 연구에서는 수집된 인터넷 기사에 달린 댓글을 활용하여 기사가 속한 그룹을 정하였다.

표 2의 키워드는 임의로 뽑은 낚시성 기사 100편의 댓글에서 공통적으로 자주 등장하는 단어를 뽑은 것이다. 기자를 속된 말로 표현한 '기레기'와 같은 기자를 지칭하는 단어와 '낚이다'라는 표현이 낚시성 기사에 자주 등장하였다. 본 연구에서는 일반 기사와 낚시성 기사를

표 2 낚시성 기사의 댓글 키워드
Table 2 Keywords from replies to a hoax article

Id	Keyword	Id	Keyword
1	기레기	5	속다,속이다
2	기자	6	낚다,낚이다
3	쓰레기	7	제목+다르다
4	낚시	8	미치다,미친

표 2의 키워드의 등장 여부로 판단하여 구분하였다. 댓글에 동의한다는 의미의 추천이 많은 3개의 댓글(베플)에서 낚시성 기사 키워드가 등장하는지 여부와 전체 댓글의 30% 이상의 댓글에서 키워드가 발견되는지를 확인하여 낚시성 기사를 정하였고 나머지는 일반 기사로 구분하였다.

댓글 키워드 분석을 통한 기사 레이블 결정 방식의 정확도를 알아보기 위해 임의로 200개의 기사를 선정하고 수작업을 통해 키워드 분석을 통한 구분이 어느 정도의 정확도를 보이는지 확인했다. 임의로 선정된 200개의 신문기사 중 일반기사는 191개, 낚시성 기사는 9개가 포함되었다. 키워드 분석을 통해 191개의 일반 기사는 전체 모두 일반 기사로 11개의 낚시성 기사는 모두 낚시성 기사로 정확하게 구분되었다.

따라서 위의 방식으로 전체 데이터셋을 레이블링 하였으며, 이는 일반 기사 80,000편, 낚시성 기사 5,500편으로 구성되었다. 이 중 일반 기사 40,000편, 낚시성 기사 2,500편을 학습데이터로 사용하였고, 나머지는 검증 데이터로 활용하였다.

2.3 특징 추출

본 연구에서는 낚시성 신문 기사의 행태[3,4]와 신문기사의 구조적 특징[7-11]을 분석한 기존 연구를 바탕으로 낚시성 신문기사 분류를 위한 다음의 다섯 가지 특징을 제안한다.

- 주제와 본문의 일치도 (f1)
- 제목, 본문의 글자 수 비교 (f2)
- 제목에 말줄임표 등장 빈도 (f3)
- 선정적 단어 및 단순 감탄사 사용 빈도 (f4)
- 제목의 마지막 단어의 품사 (f5)

다음의 각 절에서 제안된 분류 특징을 수치적 특징값으로 계산하기 위한 구체적인 방법을 설명한다.

2.3.1 주제와 본문의 일치도

낚시성 신문기사의 가장 뚜렷한 특징은 주제와 본문이 일치하지 않는다는 것이다. 독자는 신문 기사의 제목을 보고 본문을 유추하지만 낚시성 신문기사는 독자의 클릭을 유도하기 위해 실제 작성된 본문의 내용과 상이한 제목을 가지고 있는 경우가 많다. 이러한 특징을 추출하기 위해 주제와 본문의 일관성을 측정하는 것이 필요하다.

그림 1 낚시성 신문 기사의 예
Fig. 1 An example of a hoax article

주제와 본문의 일관성을 효과적인 특징으로 사용하기 위해서는 데이터의 추가적인 전처리 과정이 필요하다. 그 이유는 그림 1의 예로 설명하겠다. 그림 1의 기사는 '아이유'라는 이름의 축구선수와 가수 '아이유'를 헷갈리게 하여 독자의 클릭을 유도하는 낚시성 기사이다. 우리는 주제와 본문에 모두 등장하는 '아이유'라는 단어가 주제와 본문에서 다른 의미로 사용되었음을 나타내야 한다. 하지만 주제와 본문에 나타난 두 문자는 정확하게 일치한다. 따라서 이를 일반적인 유사도 측정 방법으로는 해결하기 어렵다.

가수 '아이유'와 축구선수 '안드레 아이유'를 구분하기 위한 방법으로 본 연구에서는 선택적 바이그램 모델을 제안한다. 선택적 바이그램 모델은 단어가 함께 사용된 앞뒤 단어와 관계를 고려하여, 여러 문서에서 함께 등장한다면 두 단어를 결합시켜 의미변화를 최소화하는 방법이다. 문장에서 연속적으로 나타나는 모든 바이그램, 즉, 두 연속단어 중, 전체 데이터 셋에서의 발생 빈도가 기준값 이상인 것들을 추가적으로 분석 대상에 포함시킨다. 선택적 바이그램 모델을 위의 낚시성 기사에 적용하면, 본문의 '아이유' 라는 단어는 '안드레아이유' 로 안드레와 아이유가 결합되어 그 형태가 변하게 된다. 수집된 데이터 셋에 '안드레'라는 단어와 '아이유' 라는 단어는 동시에 등장하는 경우가 많아 선택적 바이그램 모델이 두 단어를 결합시키기 때문이다.

선택적 바이그램 모델은 그림 2와 같이 앞뒤 단어 간의 관계를 고려하여, 서로 밀접한 관계(다른 문서에서도 같은 순서로 등장하는 빈도가 잦음)를 갖는다면 두 단어를 서로 결합시킨다. 그림 2에서 ABCDEFG는 입력값으로 문서에 포함된 단어를 나타낸다. A는 문서의 맨

Input	A	B	C	D	E	F	G
w/ prev.	0	1	0	0	1	1	0
w/ next	1	0	0	1	1	0	0
Output	AB		C	DE EF			G

그림 2 선택적 바이그램 모델
Fig. 2 Selective bigram model

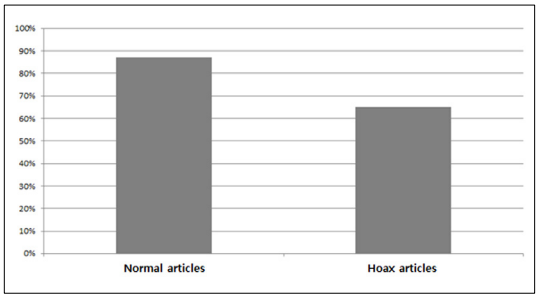


그림 3 기사의 제목과 본문의 유사도
Fig. 3 Similarity between the headline and body

처음 등장하는 단어로 앞에 다른 단어가 등장하지 않는다. 따라서 그림 2에서 A와 앞 단어는 관계없음을 0으로 표현하였다. A와 다음 단어 B의 관계를 계산해 보면 A 다음에 B가 등장하는 문서가 전체 데이터 셋에서 15회 이상 관찰되었다고 가정하면, 이는 A와 B가 서로 연관이 큰 단어 조합이라고 판단하여 1로 표현하였다. 같은 방식으로 문서의 모든 단어를 앞뒤 단어와의 관계를 고려하여 서로 관련이 깊은 단어를 결합하였다. 본 연구에서는 두 단어가 동시에 등장하는 문서의 수가 15개 이상이면 서로 관련이 깊다고 판단하였지만 분석할 데이터에 따라 그 기준값을 정하여야 한다.

선택적 바이그램 모델을 적용시킨 후 일반기사와 낚시성기사의 주제와 본문의 유사성을 측정한 결과를 그림 3으로 나타내었다. 유사성은 주제를 이루는 단어가 본문에도 등장하는지 여부를 백분위로 나타내었다. 낚시성 신문기사는 제목에 등장한 단어가 본문에 나타나지 않는 경우가 많음을 알 수 있다. 이로 인해 독자들이 제목과 본문의 내용이 달라 낚시성 기사라고 느낄 수 있다.

2.3.2 기사의 제목과 본문의 글자 수 비교

두 번째 특징은 신문기사의 제목과 본문의 글자 수를 비교하여 추출하였다. 표 3에 나타나듯, 낚시성 신문 기사는 일반 기사에 비해 적은 글자 수로 이루어진 본문으로 구성되는 것으로 분석되었다. 따라서 제목과 본문의 글자 수를 비교하여 낚시성 기사와 일반 기사를 분류하는 특징으로 사용할 수 있다.

2.3.3 기사의 제목에 말줄임표 등장 여부

세 번째 특징은 기사의 제목에 말줄임표 여부를 나타낸다. 표 4에 나타난 실험결과를 보면 알 수 있듯 낚시성 신문기사에는 일반기사에 비해 말줄임표가 더 빈번하게 나타난다. 기사의 제목에 말줄임표가 등장하면 1, 말줄임표가 사용되지 않았으면 0값을 부여하여 일반 기사와 낚시성 기사를 구분하는 특징으로 사용하였다.

표 3 제목과 본문의 글자 수 비교
Table 3 Comparison of the number of characters in the headline and body

	# characters in headline	# characters in body	# sentences in body
Normal articles	33.1	1201.3	11.7
Hoax articles	25.0	567.6	4.5

표 4 제목에 말줄임표 등장 여부
Table 4 Comparison of the existence of an ellipsis in the headline

	ellipsis appearance rate
Normal articles	13%
Hoax articles	27%

2.3.4 기사의 제목에 선정적 단어 등장 여부

네 번째 특징은 기사의 제목에 선정적 단어가 등장하는지 여부를 나타낸다. 낚시성 신문기사의 특징 중 하나는 선정적인 단어의 사용이 빈번하다는 것이다. 특히 기사의 제목에 자극적인 단어를 사용하여 사용자의 클릭을 유도하는 경향이 있다. 학습 데이터의 낚시성 기사로부터 수동으로 선정적 단어를 선정하여 목록화하였다. 노출, 누드, 섹시, 글래머, 속옷 등 총 17개의 선정적 단어를 선정하였고, 기사의 제목에 선정적 단어가 등장하면 1, 그렇지 않으면 0값을 부여하였다.

2.3.5 제목 마지막 단어의 품사

마지막 특징은 기사 제목의 마지막 단어의 품사에 관한 내용이다. 그림 4와 그림 5는 일반 기사와 낚시성 기사의 제목 마지막 단어의 품사의 분포를 나타낸다. 대부

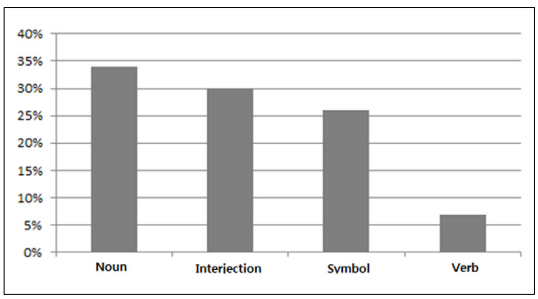


그림 4 낚시성기사의 제목 마지막 단어 품사 분포
Fig. 4 Distribution of the part of speech for the last word in the headline of hoax articles

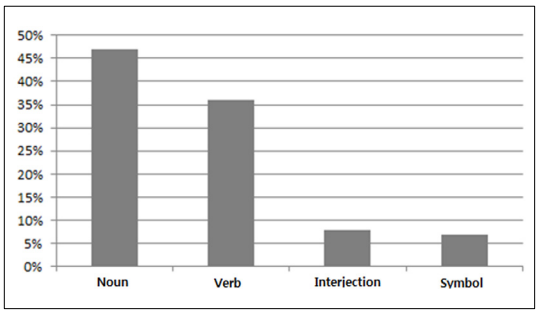


그림 5 일반기사의 제목 마지막 단어 품사 분포
Fig. 5 Distribution of the part of speech for the final word in the headline of hoax normal article

표 5 제목의 마지막 단어의 가중치
Table 5 Weight for the last word of the headline

	noun	interjection	symbol	verb
Weight	0	-0.23	-0.20	0.28

분의 기사는 명사로 제목이 끝나지만 낚시성 기사에는 감탄사와 기호가 더 빈번히 사용된 것을 알 수 있다. 따라서 아래 표 5와 같이 낚시성 기사는 음의 수로, 일반기사는 양의 수로 가중치를 줘 그 둘을 구분하는 특징으로 사용하였다.

2.4 추출된 특징 성능 검증

앞서 구축한 데이터셋을 활용하여 일반기사와 낚시성 기사를 분류하는 실험을 통해 추출된 특징을 검증하였다. 추출된 특징과 Support Vector Machine (SVM) 분류기를 활용하여 일반기사와 낚시기사를 분류하여, 특징의 정확도를 측정한다. SVM은 기계학습 기법 중 분류 분석을 위한 대표적 지도학습 모델이다. 이렇게 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되며, SVM 알고리즘은 가장 큰 폭을 가진 경계를 찾아 성능 향상을 추구하는 알고리즘이다[12].

선형 SVM은 주어진 데이터 xi들의 이진 분류를 위한 결정 초평면을 구하여 식 (1)과 같은 분류 함수의 성능을 극대화한다. 하지만 실제 데이터는 선형분류가 불가능한 경우가 대부분이다. 따라서 오분류를 인정하는 소프트 마진 기법과 비선형 SVM을 통한 방법이 사용된다. 식 (2)는 소프트 마진 기법을 통해 두 범주 사이의 마진을 최대화하는 초평면을 구하기 위한 과정을 수식화하고 있다. 식 (2)에서 ξ는 각 샘플마다 오분류에 대한 슬랙변수(slack variable)이고, C는 오분류된 샘플에 대한 패널티이다. C는 사용자 지정변수로 학습 데이터를 통해 최적의 C 값을 찾아 분류 정확도를 높여야 한다.

$$d(x) = \langle w, x_i \rangle + b = 0 \tag{1}$$

$$\begin{aligned} &Minimize\ J(w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ &Subject\ to\ \ y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i\ , \\ &\xi_i \geq 0,\ \ i = 1, \cdots, N \end{aligned} \tag{2}$$

본 연구에서는 SVM 파라미터 학습에 필요한 학습 집단과 학습의 정확도를 측정하기 위한 검증 집단을 선정하여 실험하였다. 파라미터 학습을 위해 낚시기사 분류 정확률, 즉, 재현율(recall)을 가장 중요한 평가요소로 판단하였다. 이는 실제 낚시기사가 분류기를 통해 낚시기사로 분류되는지를 나타내는 정확도를 의미한다.

3. 실험 결과

3.1 파라미터 학습

선형커널과 RBF커널을 이용하여 학습데이터에서 교차검증 실험을 실시하였다. 표 6은 선형커널을 통한 실험결과를 나타낸다. 그 결과 선형커널의 C값이 100일 때 낚시기사 분류율이 가장 높은 것을 확인할 수 있었다.

표 6 SVM 선형커널 파라미터 최적화
Table 6 Parameter optimization using linear kernel in SVM

	recall
C = 1	42.2 %
C = 10	33.1 %
C = 100	89.9 %
C = 150	63.5 %
C = 200	30.0 %

표 7 SVM RBF커널 파라미터 최적화
Table 7 Parameter optimization using RBF kernel

	gamma	recall
C=100	Y = 0.01	77.4 %
	Y = 0.05	88.1 %
	Y = 0.1	91.8 %
	Y = 0.5	43.1 %
	Y = 1.0	23.9 %

RBF커널을 이용한 실험에서는 선형커널 실험을 통해 얻은 C값과 감마값을 변화시키며 실험을 하였다. 감마값은 특징간 거리에 얼마만큼의 가중치를 줄지 결정하는 파라미터이다. RBF커널을 이용한 실험의 결과는 표 7에 나타난다. 실험결과, 감마값이 0.1일 때 낚시기사 분류율이 가장 높았다. 이 결과를 바탕으로 검증집단을 통해 특징의 성능을 평가하였다.

3.2 특징 성능 평가

최적화 실험을 통해 선정된 파라미터 C와 감마값을 적용하여 RBF 커널을 이용한 성능을 평가하였다. 표 8은 특징을 하나만 사용했을 때부터 모두 사용했을 경우까

표 8 특징 선택을 통한 성능 평가
Table 8 Performance evaluation through feature selection

No. of features	1	2	3	4	5 (all)
used features	f4	f3+f4	f1+f3+f4	f2+f3+f4+f5	**f1+f2+f3+f4+f5**
precision	0.05	0.08	0.20	0.36	**0.46**
recall	0.387	0.518	0.712	0.862	**0.917**
F-measure	0.09	0.14	0.31	0.51	**0.61**

지 정밀도(precision), 재현율(recall), 그리고 이들의 조화 평균인 F-척도(F-measure)의 변화를 나타낸다. f1-f5는 각각 2. 3절에서 설명한 각 특징을 순서대로 의미한다.

특징을 하나만 사용했을 때는 '선정적 단어 사용여부'(f4)가 가장 좋은 특징의 성능을 보였다. 또한 두 개의 특징을 사용할 때는 '선정적 단어 사용'과 '말줄임표 사용여부' 특징이 가장 좋은 분류 정확도를 보였다. 세 개의 특징을 사용해서 분류하면 앞서 두 개의 특징에 '제목과 본문의 글자 수' 특징이 좋은 성능을 보였다. 하지만 결과를 통해 알 수 있듯 모든 특징을 사용했을 때 낚시기사 분류율(재현율)이나 다른 척도 기준으로 가장 우수한 성능을 보인다.

4. 결론 및 향후 연구

본 논문에서는 낚시성 인터넷 신문 기사를 분류하기 위한 특징을 추출하고, 그를 바탕으로 낚시성 기사와 일반 기사를 분류하기 위한 방법을 제시하였다. 낚시성 기사의 주요 특성을 분석하여, 제목과 본문의 일치도, 제목·본문의 글자 수 비교, 제목에 말줄임표 등장 빈도, 선정적 단어 및 단순 감탄사 사용 빈도, 제목의 마지막 단어의 품사 등 다섯 가지의 특징을 추출하였다. 실험을 통해 92%의 분류율 보인 해당 분류 특성들은 낚시성 인터넷 신문기사를 분류하는데 적합하다고 판단된다. 뿐만 아니라 제목과 본문의 일관성을 측정하기 위한 전처리 방법으로 고안한 선택적 바이그램 모델은 낚시성 인터넷 신문기사 분류 목적 외에도 단문을 분석하기 위한 전처리 방법으로 유용할 것으로 예상된다.

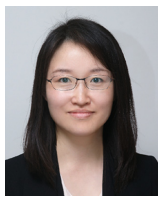
향후 연구 과제로 특징의 추가 개발로 분류율을 더욱 높일 수 있는 방법을 연구하고, 현재 연구는 모두 텍스트 데이터를 기반으로 이루어졌는데 이미지와 같은 다른 데이터를 활용하여 분류율을 높이는 방법을 연구할 계획이다.

References

[1] Korea internet security agency(KISA), "2015 Survey on the Internet usage executive (in Korean)," 2015.
[2] Statistics Korea, "2015 Social Survey results (in Korean)," 2015.
[3] Kin sunjin, "The utilization reality of hooking news titles in portal news service - Focused on major daily newspapers in naver newscast(in Korean)," *Korea Digital Design Council, Study on digital design*, Vol. 10, No. 4, pp. 283-293, 2010.
[4] Jeong kyihong, "Is internet hoax articles good as it is?" (in Korean), Kwanhun Club, *Journal of Kwanhun*, 115, pp. 129-141, 2010.
[5] Lim Kwon-Mook, "Morphological Ambiguity Resolution by using Semantic Information(in Korean)," *Korean Institute of Information Scientists and Engineers, Conference Proceedings*, Vol. 21, pp. 649-652, 1994.
[6] eunjeonhanip project, [Online]. Available: http://eunjeon.blogspot.kr/
[7] Lee joon-ho, "Implications and Theories of Headline Journalism : Focusing on Headlines of Newspapers and Portal Sites(in Korean)," *Society for journalism and communication studies, Journalism Research*, pp. 249-280, 2015.
[8] Jang min-jeong, "Analysis of newspaper headlines among Korea, USA and Japan (in Korean)," *International Association of Language & Literature Conference Proceedings*, pp. 25-37, 2012.
[9] Janggeun Oh, "Semiotic analysis of the daily press Edit(in Korean)," *Korean Association for Visual Culture, Visual culture*, pp. 108-121, 2007.
[10] Rhee, June Woong, "Direct Quotations in newspaper headlines in the coverage of the local election on May 31, 2006," *Korea Journal of Jounalism & Communication studies*, pp. 64-90, 2007.
[11] Yu, Hong-Sik, "Effects of Quantitative Distortion in the Equalized Exemplification on Readers' Issue Perception(in Korean)," *Korea Journal of Jounalism & Communication studies*, pp. 346-373, 2008.
[12] Corinna cortes, "Support-vector networks," *Machine Learning*, Vol. 20, Issue 13, pp. 273-297, 1995.

허 성 완

2012년 전남대학교 전자공학과 학사 2016년 아주대학교 소프트웨어특성화학과 석사. 관심분야는 데이터 마이닝, 기계 학습, 빅 데이터



손 경 아

2000년 서울대학교 수학과 학사. 2003년 서울대학교 컴퓨터공학과 석사. 2011년 Carnegie Mellon Univ. 박사. 현재 아주대학교 소프트웨어학과 조교수. 관심분야는 기계 학습, 데이터 마이닝, 의생명정보학