소셜 데이터와 텍스트 분류 기술을 이용한 개인 맞춤형 학습 시스템

김선표, 김은상, 전영호, 이기훈[1)]
광운대학교 컴퓨터공학과
e-mail: kihoonlee@kw.ac.kr

A Personalized Learning System Using Social Data and Text Classification Techniques

Sun-Pyo Kim, Eun-Sang Kim, Young-Ho Jeon, and Ki-Hoon Lee
Dept. of Computer Engineering, Kwangwoon University

요 약

정보통신 기기의 발달에 따라 스마트 러닝으로 교육방법이 진화하고 있다. 스마트 러닝에 있어서 학습자의 관심분야에 맞는 적절한 콘텐츠의 제공이 필수적이다. 본 논문에서는 텍스트 분류 기술을 이용하여 학습자의 SNS 데이터로부터 관심분야를 자동적으로 파악해내는 시스템을 제안한다. 텍스트 분류를 위해 카테고리 별로 기 분류되어있는 데이터를 수집하여 기계 학습을 수행하였다. 텍스트 분류의 정확도 향상을 위해 카테고리 분류 단위 크기를 변화시키면서 정확도를 측정하고 분석하여 실제 서비스에 적용 가능한 수준으로 판단되는 82.5%의 정확도를 얻었다.

1. 서론

정보통신 기기의 발달과 스마트 기기의 보급 확대에 따라 전통적 교육에서 스마트 교육으로 교육방법이 진화하고 있다[1]. 이에 대한민국 정부는 '지능적 개인맞춤 학습 관리 및 운영기술'을 정보통신 분야 국가핵심기술로 지정하였다[2]. 또한 최근 첨단 디지털 기기를 활용한 스마트 러닝이 태블릿 PC에서 로봇까지 다양해지고 있으며, 서비스 운영 및 알고리즘에서도 여러 시도들이 계속되고 있는 등 관련 서비스가 지속적으로 생겨나는 추세이다.

스마트 러닝을 위해서는 학습자의 관심분야에 맞는 적절한 콘텐츠의 제공이 필요하다. 본 논문에서는 학습자별 관심분야(음악, 여행, IT 등)를 자동적으로 파악해내는 시스템을 제안한다. 학습자의 SNS 데이터에 나타난 단어들을 보고 관심분야가 무엇인지를 파악해야 하는데, 이 문제는 텍스트 분류 기술을 이용하여 해결할 수 있다. 텍스트 분류를 위해서는 먼저 각 카테고리 별로 기 분류되어있는 텍스트 데이터가 필요한데, 본 논문에서는 인터넷 뉴스 사이트, 음악 사이트, 여행상품 사이트 등에서 기 분류된 데이터를 수집하여 사용한다. 텍스트 분류를 위한 기계학습 알고리즘으로 널리 쓰이는 지지 벡터 기계(SVM: Support Vector Machine)[3]를 사용한다. 정확도 향상을 위해 카테고리 분류 단위 크기를 조정하면서 실험 결과를 분석하였고 최대 82.5%의 정확도를 얻었다.

2. 관련 연구

기 분류된 메타 블로그 게시물을 이용하여 기계 학습을 한 후에 미 분류 블로그 게시물을 자동으로 분류하는 연구[4]가 진행되었다. SVM를 이용하여 텍스트 분류를 한다는 점은 유사하나, 본 논문에서는 학습자 관심분야 파악에 텍스트 분류를 적용하였다는 점에서 차이가 있다.

3. 제안하는 시스템

스마트 러닝을 이용하는 학습자에게 적절한 콘텐츠를 제공하기 위해서는 학습자의 관심 분야를 알아낼 수 있어야 한다. 그림 1은 인터넷에 수집한 뉴스 데이터와 학습자의 페이스북 데이터를 이용하여 관심분야를 파악하는 시스템을 보여준다. 뉴스 데이터는 이미 카테고리 별로 체계적인 분류가 되어 있으므로 텍스트 분류를 위한 학습 데이터로 활용 가능하다.

인터넷 뉴스를 크롤링을 통해 수집한 후, 형태소 분석을 하여 단어들을 추출한다. 카테고리 별로 추출된 단어 데이터를 이용하여 SVM 기계 학습을 수행한다. 다음으로 학습자의 페이스북 게시물을 읽어 형태소 분석을 한 다음 게시물이 어느 카테고리에 속해있는지를 SVM을 통해 알아낸다.

1) 교신저자

(그림 1) 소셜 데이터와 텍스트 분류 기술을 이용한 관심 분야 파악 시스템

4. 실험 및 평가

4.1. 실험 방법

텍스트 분류를 위한 학습 데이터는 일반 뉴스 사이트인 '네이버 뉴스'[5]와 스포츠 뉴스 사이트 '네이버 스포츠'[6], 음악 사이트 '엠넷'[7]의 '월간 Top 100', 여행 상품 사이트 'tourcabin'[8]으로부터 크롤링하여 수집하였다. 뉴스 기사의 제목, 음악의 제목, 여행 상품의 제목 등에 해당 카테고리를 표현하는 주요 단어들이 포함되어 있다고 가정하고, 형태소 분석기를 사용해 제목에서 명사를 추출하였다. 형태소 분석기는 오픈 소스인 hannanum[9]을 사용하였다. 텍스트 분류 프로그램은 제목과 같이 짧은 길이의 문장을 분류하는데 적합한 LibShortText[3]를 이용하여 구현하였다. 2013년도 1년 간의 데이터를 크롤링한 후에 그 중에서 100만여 개의 데이터를 무작위로 추출하여 90%의 데이터는 학습 데이터로, 10%의 데이터는 테스트 데이터로 사용하였다. 전체 테스트 데이터 중에서 정확히 분류된 데이터의 비율인 정확도를 측정하였다. 카테고리 분류 단위 크기가 정확도에 영향을 미칠 수 있으므로 크기를 달리하면서 정확도를 측정하였다. 향후 서비스 운영을 통해 페이스북 데이터가 수집되면 페이스북 데이터를 이용한 실험을 수행할 계획이다.

4.2. 소단위 분류 실험 결과

네이버 뉴스의 경우 뉴스의 카테고리를 대분류와 소분류의 2단계로 세밀하게 나누고 있다. 표 1은 소단위로 세분화된 각 카테고리의 정확도를 보여주고 있다. 정확도가 80% 이상으로 높은 카테고리(증권, 골프 등)가 있는 반면, 정확도가 50%도 되지 않는 카테고리(인터넷/SNS, IT일반 등)도 있다. 정확도가 낮은 카테고리들을 살펴보면 하나의 제목이 여러 카테고리에 해당될 수 있기 때문에 사람이 판단하더라도 카테고리 분류가 애매한 경우이다. 표 2는 정확도가 낮은 '인터넷/SNS' 카테고리의 예를 보여준다. 예를 들어 '멜론'이라는 단어는 '인터넷/SNS'와 '연예가화제' 카테고리에 모두 포함될 수 있기 때문에 정확도가 낮아지게 된다. 이러한 문제점은 카테고리를 지나치게 작게 세분화했기 때문에 발생한다고 볼 수 있다.

<표 1> 소단위 분류의 정확도

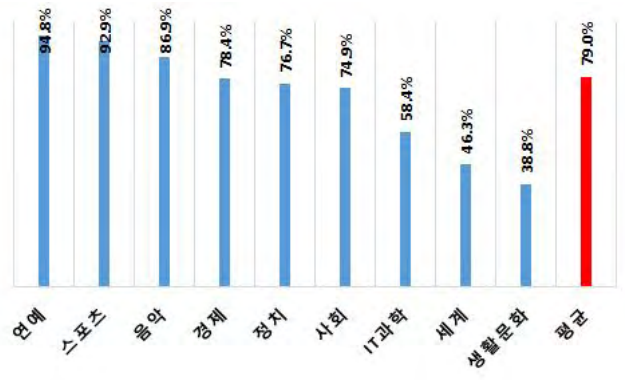
대분류	소분류					
정치	청와대	국회정당	북한	행정	국방외교	정치일반
	22.1%	41.5%	29.5%	4.2%	22.6%	61.8%
경제	금융	증권	기업/재계	부동산	글로벌 경제	
	51.1%	83.8%	42.4%	66.5%	51.6%	
	생활경제	경제일반				
	11.3%	33.8%				
사회	사건사고	교육	노동	언론	환경	인권복지
	13.3%	40.1%	16.8%	45.9%	13.9%	38.3%
	식품/의료	지역	인물	사회일반		
	9.1%	64.6%	42.7%	50.2%		
생활 문화	건강정보	자동차 시승기	도로/교통	여행/레저	음식/맛집	패션/뷰티
	30.0%	42.3%	42.5%	17.7%	79.8%	35.2%
	공연/전시	책	종교	날씨	생활문화 일반	
	22.1%	28.9%	61.0%	80.8%	24.3%	
세계	아시아/호주	미국/중남미	유럽	중동/아프리카	세계일반	
	10.5%	15.6%	16.9%	37.3%	38.0%	
IT/과학	모바일	인터넷/SNS	통신/뉴미디어	IT일반	보안/해킹	컴퓨터
	8.6%	12.1%	5.2%	39.2%	36.6%	16.4%
	게임/리뷰	과학일반				
	67.6%	17.2%				
연예	연예가 화제	방송/TV	드라마	영화	해외연예	
	71.5%	47.1%	29.8%	68.5%	39.4%	
스포츠	국내야구	일본야구	MLB	국내축구	축구 대표팀	해외축구
	92.8%	74.9%	93.0%	78.0%	80.7%	89.5%
	국내농구	NBA	배구	골프	e스포츠	
	84.1%	88.8%	78.7%	89.1%	82.8%	
음악	92.6%					
평균	56.3% (57448/101977)					

<표 2> '인터넷/SNS' 카테고리 예측 결과

기 분류 카테고리	예측 카테고리	형태소 분석이 끝난 제목
인터넷/SNS	연예가화제	멜론 음원 가격 소문
인터넷/SNS	기업/재계	사회 공헌 티몬 현성 대표
인터넷/SNS	게임/리뷰	강민경 게임 신곡 온라인 채널 서비스
인터넷/SNS	IT일반	다음 로드뷰 플러스 오픈
인터넷/SNS	IT일반	유스트림 소비자 가전 전시회 생중계
인터넷/SNS	사회 일반	이근철 20 영어 노하우 토크 리시
인터넷/SNS	IT일반	카메라 애플리케이션 카메라 2000 다운로드
인터넷/SNS	생활문화일반	문자 답장 엄마 문자
인터넷/SNS	IT일반	SK플래닛 日KDDI 소프트뱅크 도쿄 NFC존 구축
인터넷/SNS	세계일반	카망베르 치즈 리스테리아균 사망
인터넷/SNS	IT일반	카카오톡 불통 대란 옛말
인터넷/SNS	IT일반	KISA 인터넷·정보보호 10 이슈 선정
인터넷/SNS	지역	NHN NEXT 학교 입학식 개최

4.3. 대단위 분류 실험 결과

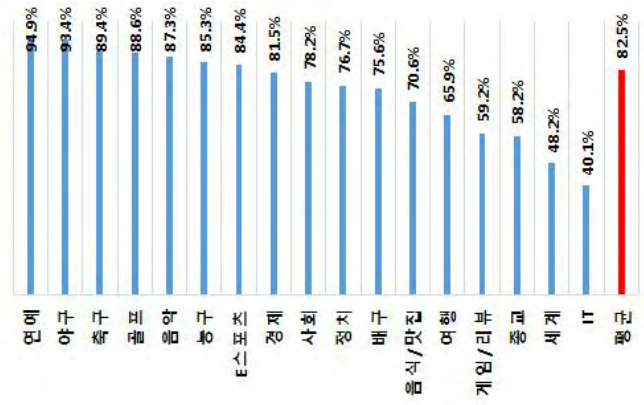
카테고리를 지나치게 세분화하면 정확도가 낮아지는 문제를 해결하기 위해 네이버 뉴스의 대분류만을 이용하여 카테고리를 큰 단위로 분류한 다음 정확도를 측정하였다. 그림 2를 보면 표 1의 소단위 분류 결과보다 정확도가 향상된 것을 확인할 수 있다. 그러나 이렇게 큰 단위로 카테고리를 분류하면 각 카테고리가 너무 포괄적이어서 학습자의 관심분야에 정확히 부합하는 콘텐츠의 제공이 어려울 수 있다. 예를 들어, 학습자의 관심분야가 '축구'라고 하자. 학습자의 게시물에 나타난 '메시'라는 단어를 바탕으로 '스포츠'라는 카테고리를 관심분야로 파악해낸다면 학습자가 관심 없을 수도 있는 '농구'나 '야구'의 콘텐츠가 제공될 수 있다.



(그림 2) 대단위 분류의 정확도

4.4. 중단위 분류 실험 결과

소단위 분류에서 단어들이 여러 카테고리에 걸쳐서 나타남으로 인해 정확도가 낮은 카테고리들을 병합하여 중간 크기로 카테고리를 재분류하였다. 그리고 소단위 분류에서 '여행/레저' 카테고리에 실제 여행 정보와는 관련 없는 기사들이 많은 것을 발견하고 여행 카테고리는 'tourcabin'의 여행 상품 제목을 이용하도록 수정하였다. 그 결과 평균 정확도가 82.5%로 상승한 것을 그림 3에서 확인할 수 있다. 카테고리 분류 또한 대단위 분류보다는 세분화되어 사용자의 관심사를 좀 더 명확히 파악할 수 있게 되었다.



(그림 3) 중단위 분류의 정확도

5. 결론

본 논문에서는 스마트 러닝을 위한 개인 맞춤형 콘텐츠 제공 방법을 제안하였다. 학습자의 관심분야 파악을 위해 기 분류된 데이터를 수집하고 SVM 기반의 기계학습을 수행하였다. 그리고 카테고리 분류의 단위 크기를 조정하여 정확도를 향상시켰다. 본 논문에서 제안한 시스템은 지능형 콘텐츠 제공 엔진으로 활용 가능할 뿐 아니라 맞춤형 광고 솔루션으로 활용해 광고 효과를 증대시킬 수 있다. 이 기술을 상용화하기 위해서는 정확도를 좀 더 높일 필요가 있으며, 새로운 학습 데이터를 주기적으로 추가하여 기계학습하는 알고리즘의 개발이 필요하다. 또한 빅 데이터 처리를 위한 Hadoop과의 연계 방법도 고려해볼 사항이다.

감사의 글

본 연구는 중소기업청에서 지원하는 2014년도 산학연협력 기술개발사업(No. C0187264)의 연구수행으로 인한 결과물임을 밝힙니다.

참고문헌

[1] 한국교육학술정보원(KERIS), 스마트교육 콘텐츠 품질관리 및 교수학습 모형 개발 이슈, 2011년.
[2] 산학연협력 기술개발사업 관리지침, 2014년.
[3] H.-F. Yu, C.-H. Ho, Y.-C. Juan, and C.-J. Lin. LibShortText: A Library for Short-text Classification and Analysis, 2013년 10월
[4] 이상진, 이승훈, 박종헌, "메타블로그 사이트에서의 선형 SVM기반 텍스트 부류 방법론 연구," 대한산업공학회 춘계공동학술대회 논문집, pp. 191-197, 2011년 5월.
[5] NAVER 뉴스, http://news.naver.com
[6] NAVER 스포츠, http://sports.naver.com
[7] Mnet 종합차트 http://www.mnet.com
[8] No. 1 여행포털, 투어캐빈, http://www.tourcabin.com
[9] HanNanum: Korean Morphological Analyzer, http://kldp.net/projects/hannanum, 2009년 5월