https://doi.org/10.5392/JKCA.2017.17.01.039

효율적인 문헌 분류를 위한 시계열 기반 데이터 집합 선정 기법

Time-Series based Dataset Selection Method for Effective Text Classification

채영훈*, 정도헌**
과학기술연합대학원대학교(UST) 빅데이터과학과*, 한국과학기술정보연구원(KISTI)**

Yeonghun Chae(proin@ust.ac.kr)*, Do-Heon Jeong(heon@kisti.re.kr)**

요약

인터넷 기술이 발전함에 따라 온라인상의 데이터는 급격하게 증가하고 있고, 증가하는 데이터에 대해 점진적인 기계학습 기법을 통해 효율적으로 학습하기 위한 연구가 진행되고 있다. 온라인상의 문서는 대부분 게시일, 출판일과 같은 시계열적 정보를 포함하고 있고, 이를 분류에 반영한다면 효율적인 분류가 가능할 것이다. 본 연구에서는 웹 문서상에서 나타나는 어휘의 시계열적 변화를 분석하였고, 분석한 시계열 정보를 기반으로 데이터 집합을 분할하여 효율적인 분류 학습 기법을 제안한다. 실험 및 검증을 위해 온라인상의 뉴스 기사 100만 건을 시계열 정보를 포함하여 수집하였다. 수집된 데이터를 바탕으로 데이터 집합을 분할하여 Naïve Bayes 및 SVM 분류기를 사용하여 실험을 진행하였고, 각 모델에서 전체 데이터 집합 학습 대비 최대 2.02% 포인트, 2.32% 포인트의 성능 향상을 확인하였다. 본 연구를 통해 시계열적 어휘의 변화를 분류에 반영하여 분류의 성능을 향상시킬 수 있음을 확인하였다.

■ 중심어 : | SVM | 나이브베이즈 | 시계열분석 | 기계학습 | 분류 |

Abstract

As the Internet technology advances, data on the web is increasing sharply. Many research study about incremental learning for classifying effectively in data increasing. Web document contains the time-series data such as published date. If we reflect time-series data to classification, it will be an effective classification. In this study, we analyze the time-series variation of the words. We propose an efficient classification through dividing the dataset based on the analysis of time-series information. For experiment, we corrected 1 million online news articles including time-series information. We divide the dataset and classify the dataset using SVM and Naïve Bayes. In each model, we show that classification performance is increasing. Through this study, we showed that reflecting time-series information can improve the classification performance.

■ keyword : | SVM | Naïve Bayes | Time-Series Analysis | Machine Learning | Classification |

I. 서 론

문헌 분류는 온라인 및 오프라인 상에 존재하는 대량의 문헌을 사용자가 원하는 범주로 분류해주는 작업이다. 과거에는 대부분 사람이 수작업을 통해 정보를 구축하였으나, 컴퓨팅 기술이 발전함에 따라 이를 기계적

접수일자 : 2016년 11월 07일
수정일자 : 2016년 12월 15일
심사완료일 : 2016년 12월 19일
교신저자 : 정도헌, e-mail : heon@kisti.re.kr

으로 구축하는 방법이 가능해졌고, 대부분의 문서 분류가 자동으로 진행되고 있다[1].

오늘날 인터넷 기술이 발전함에 따라 온라인상에 다양한 정보가 폭발적으로 증가하고 있다. CISCO는 전 세계 모바일 데이터 트래픽이 '18년에는 1억6140만 테라바이트까지 증가할 것이라고 전망하고 있다[2]. 데이터가 증가함에 따라 대규모의 데이터를 축소하여 성능을 높이거나 추가되는 데이터에 대해 점진적으로 학습하는 등 이를 효율적으로 처리하는 문제 역시 중요한 문제로 제기되고 있고, 이를 해결하기 위한 다양한 연구가 진행되고 있다[3][4].

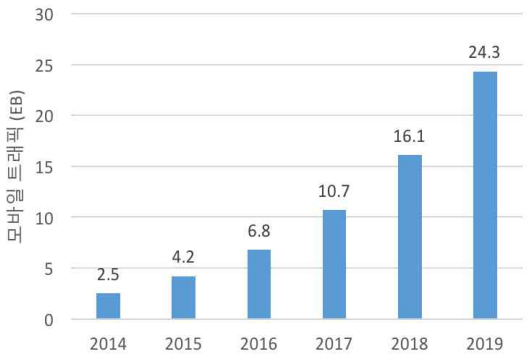

그림 1. 전 세계 모바일 데이터 트래픽 전망

온라인상의 문서들은 대부분 게시일, 출판일 등 시계열 정보를 포함하고 있다. 이를 통해 문서에 포함된 각 단어들의 분포를 시계열적으로 분석이 가능하고, 시계열에 따른 단어의 분포는 웹 문서 분류와 같은 작업을 할 때 분류 성능에 영향을 미칠 수 있다. 따라서 본 연구에서는 증가하는 데이터를 효율적으로 분류하기 위해 시계열적 요소를 반영하여 데이터 집합을 분할하고 이를 효율적으로 선택하여 분류의 효율을 높이는 방법을 제안한다. 2장에서는 본 연구에서 사용된 선행 연구를 위주로 기존 연구들을 소개하고, 3장에서는 시계열적 변화 패턴을 분석하여 분류에 적용 할 수 있는 방법에 대해 고찰한다. 4장에서는 3장에서 소개한 패턴을 바탕으로 분류에 적용하는 방법을 제시하고 이를 검증한다. 마지막으로 5장에서는 본 연구에서 제시한 방법론의 우수성 및 향후 연구 방향에 대해 언급한다.

Ⅱ. 관련연구 및 연구배경

2.1 문헌 분류

2.1.1 Naïve Bayes

Naïve Bayes 분류는 특성 사이의 독립을 가정하는 베이지안 정리를 적용한 확률 분류기의 일종으로 분류 작업에서 광범위하게 사용되고 있다[5]. 문헌 분류에서 Naïve Bayes 분류기는 단어의 빈도를 바탕으로 통계적인 계산을 하는 단순한 알고리즘이지만, IBM의 연구 보고서에 따르면 적절한 전처리를 하면 SVM 등 우수한 방법과 비교해도 성능 면에서 경쟁력을 보인다는 것을 확인 할 수 있다[6].

Naïve Bayes 분류기의 학습 과정은 범주 c_k에서 자질 f가 출현할 확률인 $P(c_k|f)$을 학습하는 것이다. 이를 구하는 방법은 아래와 같다.

$$P(c_k|f) = P(f|c_k) \times P(c_k) \quad (1)$$

$$P(f|c_k) = \prod_{i=1}^{N} P(f_i|c_k) \quad (2)$$

여기서 $P(c_k)$은 전체 문서에서 범주 c_k가 나타날 확률이고, $P(f|c_k)$은 분류 대상 문서에 포함된 자질 f_i가 범주 c_k에 출현할 확률인 $P(f_i|c_k)$을 곱하여 구할 수 있다. 분류 대상 문서에서 각 범주별로 $P(c_k|f)$을 계산한 후 비교하여, 가장 높은 확률을 가지는 범주 c_k를 현재 문서의 범주로 분류한다. 본 연구에서는 분류의 성능을 높이기 위해 Naïve Bayes의 범주 확률 값에 시계열적 요소를 반영하였다.

2.1.2 Support Vector Machine (SVM)

SVM은 1995년 Vladimir Vapnik이 제안한 지도 학습 모델로 이미지, 텍스트 등 다양한 분야에서 분류 문제를 해결하고 있다[7]. SVM 분류 모델은 초평면으로 구성되어 있고, 그 중 가장 큰 폭을 가지는 경계를 찾는 알고리즘이다. SVM 알고리즘은 선형 분류 모델로 고안되었지만, 커널 트릭을 사용하여 비선형 분류 문제에도 사용되고 있다[8].

Hirotoshi Taira와 Masahiko Haruno의 연구에서 자질 선택 방법을 통해 자질을 축소하여 SVM의 성능을 측정하는 실험을 하였다[9]. 해당 실험에서는 전체 자질의 수 15,000개에서 300개까지 축소하여 실험한 결과 300개의 자질만으로도 안정적인 성능을 보임을 확인하였다. 이를 통해 SVM 분류기는 적은 양의 데이터로도 안정적인 성능을 보인다는 것을 확인 할 수 있다. 하지만 kNN, Naïve Bayes와 같은 단순한 알고리즘과 비교했을 때, SVM은 학습 시간이 오래 걸리고, 대규모 데이터를 학습 할 때 많은 컴퓨팅 자원이 필요하다[10]. 본 연구에서는 시계열 요소를 고려하여 데이터 집합을 나눔으로써 SVM을 안정적이고 효율적인 분류 방법으로 사용하였다.

2.2 자질 선택

자질 선택은 SVM 등의 분류 모델을 학습하기 이전에 전처리 과정에서 수행되는 작업으로, 분류 성능을 향상시키고 속도, 컴퓨팅 자원 사용을 감소시킨다. 많은 연구들에서 TF-IDF, AdaBoost 등의 기법을 통해 자질 선택을 하여 SVM의 성능을 향상시키는 것을 확인하였다[11][12]. 본 연구에서는 Document Frequency(DF) 기법과 Ambiguity Measure(AM) 알고리즘을 사용하였다.

2.2.1 Document Frequency (DF)

Document Frequency(DF)를 통한 자질 선택은 계산이 간단하면서 보편적으로 많이 사용되는 방식이다. DF를 통한 자질 선택은 특정 자질이 나타난 문서의 빈도를 계산한 후, 이 값이 특정 임계값을 넘지 않은 단어를 모두 제거하는 방식이다. 이 방식은 적은 문서에서 사용된 단어는 전체 분류에서 영향을 미치지 않는다는 전제에서 사용된다. Yiming Yang과 Jan O. Pedersen의 연구에 따르면, DF를 사용하여 대략 90%의 자질을 제거하여도 문서의 분류에는 영향이 없는 것을 볼 수 있다[13]. 이를 통해, 문서 빈도가 낮은 단어를 제거하여 학습한다면, 분류 성능은 유지하면서 분류기에서 사용되는 컴퓨팅 성능 및 학습 시간을 단축 할 수 있는 것을 확인할 수 있다.

2.2.2 Ambiguity Measure (AM)

AM 알고리즘은 단어의 범주별 빈도를 통해 단어가 가지는 모호성을 계산하는 알고리즘이다[14]. 예를 들면, 뉴스 기사에서 “인공지능”이라는 단어는 IT 분야의 기사에서 주로 사용되는 용어이다. 통계적 분류 모델에서 이와 같은 용어들은 영향력이 크다. 반면 “뉴스”, “한국”과 같이 보편적으로 사용되는 단어들은 모든 분야에서 균등하게 사용되기 때문에 영향력이 낮다. AM에서 단어의 모호성을 계산하는 방법은 아래와 같다.

$$AM(t_k,c_i) = \frac{tf(t_k,c_i)}{tf(t_k)} \quad (3)$$

$$AM(t_k) = \max(AM(t_k,c_i)) \quad (4)$$

2.3 데이터 집합 선정 기법

데이터 집합 선정 기법은 주로 이기종 데이터 간 통합 검색 환경을 구축하기 위해 사용한다. 새로운 데이터가 입수되었을 때, 다양한 데이터 집합 중 최적의 데이터 집합을 선정하여 분류 정확도를 향상시킨다.

데이터 집합 선정 기법 중 하나로 최대 개념강도 인지 기법(Maximal Concept-Strength Recognition Method, MCR)이 있다[15]. MCR 기법을 통해 신규 문헌을 가장 잘 분류할 수 있는 분류기를 선택하여 최적의 분류 결과를 얻을 수 있다. MCR 제안 논문의 실험에서는 GTB, SOC, NDS, BIST의 4가지 데이터 집합을 사용하여 각 데이터 집합을 교차 검증하고 MCR 기법의 성능을 비교하였다. 이를 통해 MCR 기법이 최적의 데이터 집합을 선정하여 이기종 데이터 간 안정적인 분류 성능을 보임을 확인하였다.

본 연구에서는 데이터 집합을 시계열에 따라 분할하여 Inner Product 기반의 데이터 집합 선정 기법을 사용하여 최적의 시계열 구간을 선택하여 분류의 성능을 높이는 방법을 제안한다.

2.4 시계열 분석

시간 또는 시대에 따라 어휘 등 실제 세계에 존재하는 요소들은 변화한다. 예를 들면, 특정 사람이 사용하는 이메일 주소는 시간이 지남에 따라 졸업, 취직, 이직

등 소속의 변화에 따라 변화한다. 이메일 주소의 변화는 개인을 식별 할 때 도움을 줄 수 있다[16]. 이메일 주소뿐만 아니라 어휘 또한 시간의 변화에 따라 의미가 변화한다. Derry와 Reyyan의 연구에서는 K-means 클러스터링 기법을 사용하여 시대에 따른 어휘의 의미 변화를 기계적인 방법으로 분석하는 시도를 하였다[17]. 해당 연구를 통해 시대에 따라 어휘의 의미가 변화하는 것을 자동으로 식별 할 수 있었다. Jeong과 Song의 연구에서는 토픽 모델링 기법을 사용하여 시간차(Time Gap)을 분석하는 시도를 하였다[18]. 이와 같은 연구를 바탕으로 본 연구에서는 어휘의 시계열적 변화를 분석하여 분류 기술에 적용하는 시도를 하였다.

2.5 어휘의 의미 변화

뉴스 기사와 같은 온라인상의 문서는 게시일, 출판일 등 시계열적 속성을 지니고 있다. 이를 통해 뉴스 기사에서 나타나는 단어가 시간에 따라 변화하는 과정을 AM 알고리즘을 활용하여 분석할 수 있다.

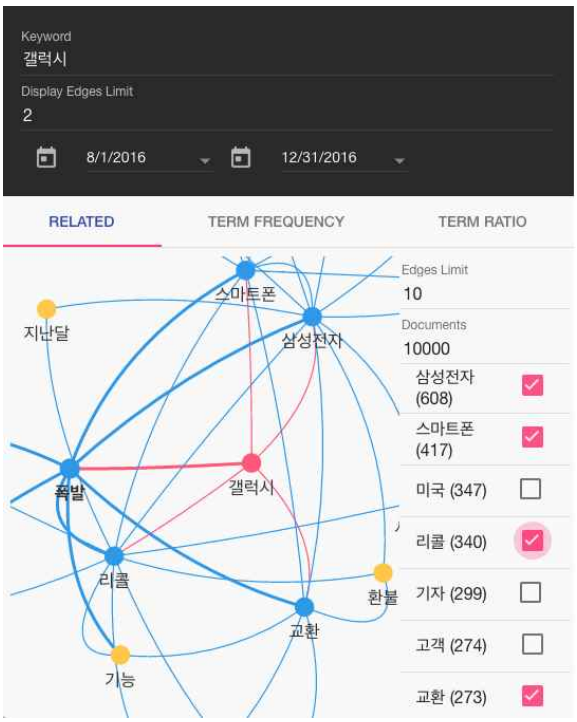

그림 2. 어휘 트렌드 분석 시스템

본 연구에서는 시계열에 따른 어휘의 의미 변화를 분석하기 위해 [그림 2]와 같은 시각화 시스템을 구축하였다. 시스템 구축을 위해 국내 포털 "네이버"로부터 약 100만 건의 뉴스 기사를 수집하였다. 수집된 데이터를 바탕으로 기간에 따라 범주별 AM을 측정하여 시간에 따른 범주 비중을 계산하였고, 단어가 문서에서 동시 출현한 빈도를 기간별로 계산해서 연관 단어를 추출하였다. 추출된 데이터의 시각화를 위해 Node.js 기반의 웹 서버를 구축하고, Javascript 기반의 시각화 라이브러리인 Vis.js 및 Highchart를 사용하여 그래프로 표현하였다[19-21]. 구현한 시스템을 사용하여 시계열적 특성을 살펴보고, 시간에 따른 어휘의 범주 변화 패턴을 분석하여 분류에 적용하는 방법에 대해 고찰하였다.

2.5.1 관심도의 변화

뉴스 기사와 같은 웹 문서는 사회에서 주요 이슈가 되는 내용을 포함하고 있다. 따라서 웹 문서 상에 시간에 따라 출현하는 단어의 빈도수는 사람들의 관심도를 반영하고 있다.

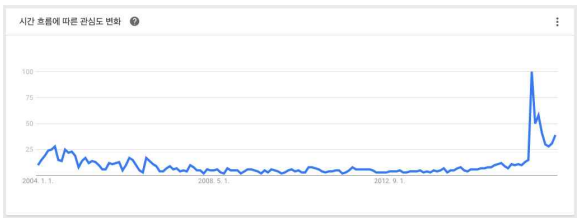

그림 3. "인공지능" 구글 트렌드 검색 결과

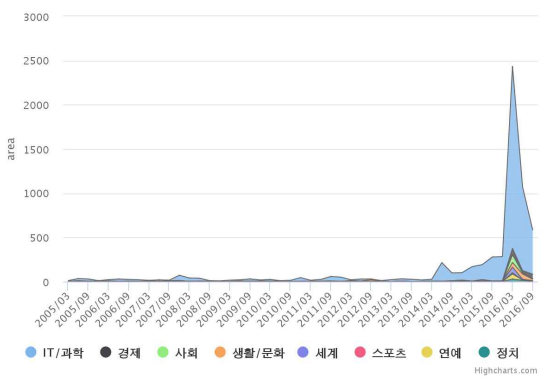

그림 4. 뉴스 빈도로 측정한 "인공지능"의 트렌드

[그림 3]과 [그림 4]는 각각 구글 트렌드에서 인공지능을 검색한 결과와 시간에 따른 뉴스의 빈도를 그래프로 표현한 그림이다. 구글 트렌드는 사용자 검색 로그를 분석하여 키워드의 관심도를 계산한다. 동일 기간의 두 그래프를 보면 관심도의 변화의 추세가 유사한 것을 확인 할 수 있다. 이를 통해 사용자의 검색 로그를 통한 관심도와 마찬가지로 뉴스 기사 또한 사회적인 관심도를 반영한다는 것을 알 수 있다.

온라인상의 문서에서 어휘의 변화는 사람들의 관심도를 반영하기 때문에 실제 세계에서 어휘의 변화와 시기적 차이가 존재할 수 있다. 본 연구에서는 실제 어휘의 변화를 분석하는 것이 아닌 웹 문서상의 어휘 변화에 초점을 맞추어 연구를 진행하였다.

2.5.2 신조어

기존에 존재하지 않던 단어가 새로 생겨나면 이는 분류에 영향을 준다. 예를 들면, “빅데이터”는 2001년에 처음 등장하였지만, 대중적으로 사용된 시점은 2010년부터이다. 뉴스 기사의 빈도는 대중의 관심도와 연관이 있기 때문에 뉴스에서는 “빅데이터”라는 키워드가 2010년 이후 일종의 신조어와 같다. 만약 2010년 이전까지의 데이터 집합을 학습하여 분류를 수행한다면, “빅데이터”라는 키워드가 등장한 2010년 이후 데이터를 분류할 때 오류를 발생 시킬 수 있다. 따라서 웹 문서 분류 작업은 주기적으로 새로운 문서를 추가하여 학습해서, 새로 등장한 어휘에 대한 학습이 가능하도록 해야 한다.

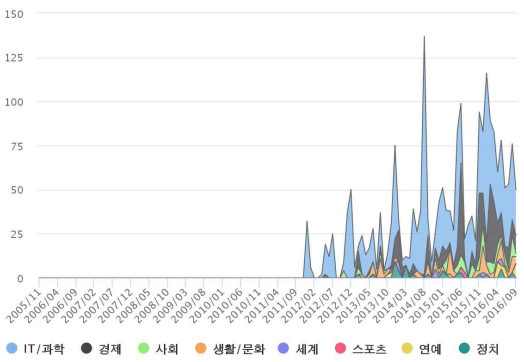

그림 5. “빅데이터” 빈도 그래프

2.5.3 동형이의어

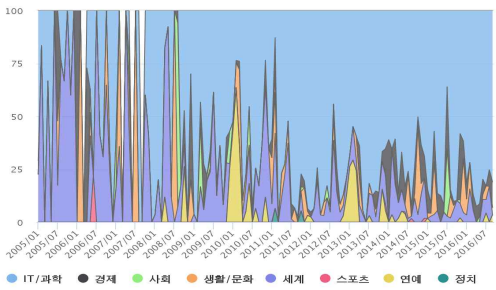

그림 6. “아마존”의 범주 비율 변화

동형이의어는 형태는 같지만 의미가 다른 단어를 뜻한다. 예를 들어, “아마존”이라는 단어는 남아메리카에 있는 강을 지칭하는 지명과 온라인 쇼핑몰을 운영하는 기업인 “아마존닷컴”의 두 가지 의미를 가지고 있다. [그림 6]은 시간에 따른 “아마존”의 범주 변화를 나타낸다. 전체 기간의 범주별 빈도의 경우, “아마존”은 “IT/과학” 분야에서 사용되는 것으로 나타나지만, 기간을 나누어 분석한다면 “아마존”의 범주가 서로 다른 것을 확인 할 수 있다. 따라서 전체 기간의 빈도를 바탕으로 분류를 수행한다면 시간에 따른 범주 변화가 반영되지 않아 실제 의미와 관계없이 “IT/과학”으로 인식되어 잘못된 분류를 할 확률이 높다.

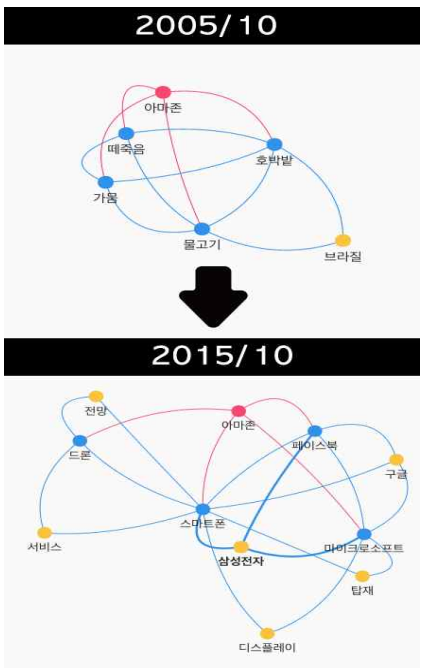

그림 7. “아마존” 연관단어 네트워크

[그림 7]은 아마존이 나타난 문서에 동시 출현한 단어를 위주로 시기에 따라 연관단어 네트워크를 구성한 것이다. 2005년에는 연관 단어가 호박밭, 물고기, 브라질 등이 나오는 것으로 보아 아마존 강을 지칭한다는 것을 알 수 있다. 하지만 2006년 이후 아마존닷컴에서 아마존 웹 서비스(AWS)를 제공하기 시작하였고, 이후 클라우드 컴퓨팅 기술이 각광받기 시작하면서 IT 기술 분야에서 관심을 받기 시작하였다. 아마존닷컴의 관심도 증가에 따라 아마존의 연관 단어 또한 "삼성전자", "마이크로소프트" 등 "IT/과학" 분야의 단어들로 변화하였다.

어휘의 연관 단어 변화는 문서 집합의 빈도 분포를 바탕으로 분류 대상 문서와 데이터 집합의 유사도를 계산하여 최적의 데이터 집합을 선정함으로써 효율적인 분류를 하도록 반영 할 수 있다.

Ⅲ. 데이터 집합 선정 모델

3.1 가정

본 연구에서는 3장의 시계열 분석을 분류에 반영하는 방법을 제안한다. 본 실험을 위한 가정은 아래와 같다.

- 학습 데이터 집합은 시계열 정보를 포함하고 있다.
- 분할된 각 데이터 집합은 분류하기에 충분한 수의 자질을 가지고 있다.
- 분할된 각 데이터 집합은 해당 시기를 대표한다.

3.2 모델 구조

본 실험에서는 어휘의 시계열적 요소를 반영하기 위해 데이터 집합을 시계열을 기준으로 분할하여 별도로 학습 한 후, 데이터 집합 선정 모델을 통해 최적의 집합을 찾는 방법을 제안한다.

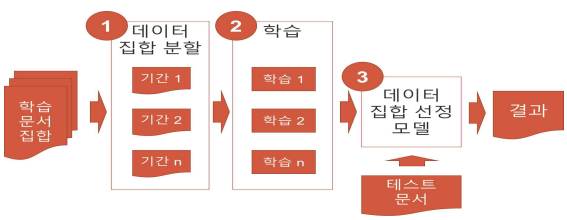

그림 8. 데이터 집합 선정 모델 구조

본 실험은 [그림 8]과 같이 데이터 집합 분할, 학습, 데이터 집합 선정의 과정으로 진행된다. 데이터 집합 선정에서는 Score Rank Model과 Inner Product Model의 두 가지 모델을 제안한다. 실험의 과정은 아래에서 과정별로 설명하고, 제안하는 데이터 집합 선정 모델은 3.3절을 통해 설명한다.

① 데이터 집합 분할

분류기를 통해 학습하기 전에 전체 문서에서 시계열을 기준으로 데이터 집합을 분할한다. 본 실험에서는 2004년 7월부터 2014년 12월까지 전체 126개월의 데이터를 3개월, 6개월, 12개월, 24개월로 기간을 나누어 분할하는 기간에 따른 정확도의 변화를 측정하였다.

데이터 집합을 시기별로 분할 한 후, AM을 사용한 부스팅을 통해 시계열에 따른 범주의 변화를 반영하였다. 시기별로 분할된 데이터 집합에서 각 구간별 AM을 계산하였고, 구간별로 계산된 AM을 자질의 값에 곱하여 부스팅하였다.

② 학습

각 기간별로 분할된 데이터를 분류기를 통해 독립적으로 학습한다. 본 실험에서는 Naïve Bayes와 SVM 분류기를 사용하여 학습하고, 분류 성능 비교를 통해 본 모델 성능의 경향성을 분석하였다.

분할하는 기간이 커질수록 학습을 위해 더 많은 메모리와 시간이 필요하였다. 하지만 기간별 학습 과정이 독립적으로 수행되기 때문에, 각 학습 과정을 동시해 수행하여 학습 시간을 단축할 수 있었다.

③ 데이터 집합 선정 모델

본 실험에서는 시계열로 분할된 데이터 집합들에서 최적의 데이터 집합 선정을 위해 Score Rank Model과 Inner Product Model의 두 가지 모델을 고안하였다. 데이터 집합 선정 모델은 [그림 8]의 3번 과정에서 수행된다.

3.3 데이터 집합 선정 모델

3.3.1 Score Rank Model

Score Rank Model은 데이터 집합을 분할하여 학습

한 후, 각 학습된 분류기를 통해 계산된 범주들의 점수를 바탕으로 최적의 범주를 선택하는 방법이다. 본 실험에서는 Naïve Bayes 분류기를 통해 Score Rank Model을 실험하였고, 데이터 집합 별 점수를 계산하는 방법은 [수식 5]와 같다.

$$score(c_k,d) = \sum_{i=1}^{N} Log(P(f_i|c_k)) \qquad (5)$$

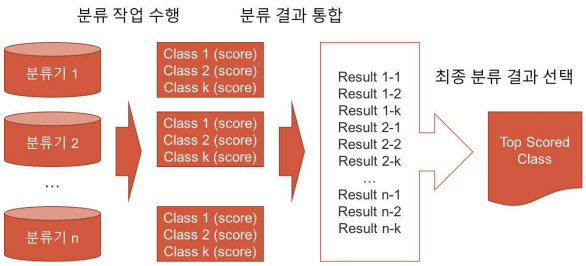

그림 9. Score Rank Model 구조

[그림 9]는 Score Rank Model의 구조이다. 분할된 각 학습 집합별로 분류기를 통해 범주들의 점수를 구하고, 전체 결과를 통합하여 가장 높은 점수를 가지는 범주를 분류 대상의 범주로 선택한다. 분류기에서 계산되는 범주의 점수들은 단어 빈도 분포를 바탕으로 계산하기 때문에 계산된 결과에 학습 집합의 유사도가 반영된다.

3.3.2 Inner Product Model

Inner Product Model은 3장에서 언급한 시기별 연관 단어의 변화를 데이터 집합 선정에 반영한 모델이다. 각 데이터 집합과 분류 대상 문서의 유사도를 내적(Inner Product)을 통해 계산하여 최적 집합을 찾는다.

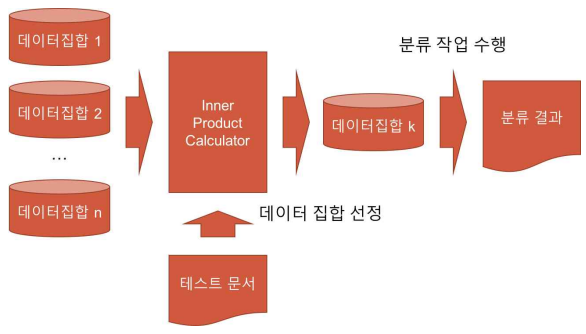

그림 10. Inner Product Model 구조

[그림 10]은 Inner Product Model의 구조이다. 분할된 데이터 집합과 테스트 문서의 유사도를 Inner Product 공식을 통해 계산한 후 최적의 데이터 집합을 선정한다. 해당 데이터 집합을 통해 학습한 분류기를 통해 테스트 문서의 분류 작업을 수행하여 분류 결과를 구한다. 데이터 집합 유사도는 해당 집합의 단어 빈도 분포와 분류 대상 문서의 단어 빈도 분포를 바탕으로 [수식 6]을 통해 유사도를 계산할 수 있다.

$$IP(s_i,d) = \sum_{k=1}^{n} (tf_k(d) \times tf_k(s_i)) \qquad (6)$$

s_i는 분할된 각 데이터 집합을 의미하고 d는 분류 대상 문서를 의미한다. $tf_k(d)$는 분류 대상 문서에 나타난 k번째 단어의 빈도수를 의미하고, $tf_k(s_i)$는 k번째 단어가 데이터 집합 s_i에서 나타난 빈도수를 의미한다. 계산된 유사도를 통해 가장 높은 유사도를 가지는 학습 집합을 선정하여 분류를 진행한다.

Ⅳ. 성능평가

제안하는 모델을 검증하기 위해 Naïve Bayes 및 SVM 분류기를 사용하여 실험을 진행하였다. 각 분류 모델에서 구간을 나누지 않고 학습하여 분류한 결과를 비교 대상으로 설정하고, 데이터 집합을 2년, 1년, 6개월, 3개월의 구간으로 분할하여 학습하면서 시계열에 따른 분할이 어떠한 영향을 미치는지 검증하였다.

성능 평가를 위해 micro averaged precision을 계산하여 각 구간을 비교하였다. micro averaged precision은 [수식 7]을 통해 계산 할 수 있다.

표 1. 평가를 위한 2x2 Contingency table

		Observation	
		Class	Not Class
Prediction	Class	TP	FP
	Not Class	FN	TN

$$Micro\ Averaged = \frac{\sum_{j=1}^{n} TP_j}{\sum_{j=1}^{n} (TP_j + FP_j)} \quad (7)$$

4.1 데이터 집합

본 연구에서는 시계열 정보를 포함한 데이터를 사용하기 위해 네이버에서 주제별로 제공하는 일자별 주요 뉴스를 수집하였다. 수집된 데이터의 수는 [표 2]와 같다.

표 2. 범주별 네이버 뉴스 문서 수

범주	문서 수	비율 (%)
IT/과학	136,524	12.72
경제	136,515	12.71
사회	136,521	12.72
생활/문화	136,537	12.72
세계	136,391	12.70
스포츠	136,505	12.71
연예	118,040	11.00
정치	136,559	12.72
전체	1,074,043	100.00

수집된 데이터 집합에서 2004년부터 2014년까지 126개월의 문서 919,500건을 학습 데이터로 사용하였고, 2015년 이후 수집된 문서 154,543건을 검증 데이터로 사용하였다.

4.2 결과 분석

4.2.1 데이터 집합 선택 모델 비교

Score Rank Model과 Inner Product Model의 성능을 비교하기 위해 Naïve Bayes 분류기를 사용하여 두 모델을 비교하였다. 두 모델 모두 데이터 집합을 시계열로 분할하여 학습하였을 때 분류 성능이 향상되었다.

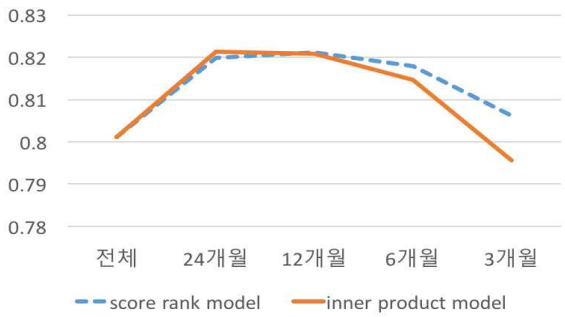

그림 11. 데이터 집합 선택 모델 비교 그래프

[그림 11]은 Score Rank Model과 Inner Product Model의 분할 기간별 Micro Precision 값을 비교한 그래프이다. Score Rank Model의 경우 데이터 집합을 12개월 단위로 분할하였을 때 82.11%로 가장 높은 정확도를 보였고, 이는 분할하지 않고 학습한 결과보다 1.87% 포인트 향상된 성능이다. Inner Product Model의 경우 24개월 크기로 데이터 집합을 분할하였을 때 82.13%로 가장 높은 정확도를 보였다. 이는 전체 집합을 학습한 결과인 80.11%보다 2.02% 포인트 향상된 성능이다.

Inner Product Model에서 최대 82.13%로 가장 높은 성능을 보였지만, Score Rank Model에서 또한 최대 82.11%의 성능으로 두 모델 모두 유사한 성능 향상을 보인다. 또한 분할 단위에 따른 성능의 경향이 유사하다. 이를 통해 시계열 요소를 분류에 반영할 경우 유의미한 성능 향상이 가능하다는 것을 확인 할 수 있다. 하지만 두 모델 모두 데이터 구간을 일정 기준 이상 세밀한 크기로 분할할수록 분류 성능이 감소하였다. 이는 세밀하게 분할 할 경우 자질의 수가 분류하기에 적합하지 않을 정도로 적어지기 때문이다. 따라서 효율적인 분류 성능을 위해서는 전체 데이터 집합의 크기에 따라 적당한 분할 구간을 설정해주어야 할 것으로 생각된다.

두 모델 모두 분류 성능은 유사한 수준으로 향상되지만 분류 시간에는 차이가 있었다. 학습 시간의 경우 동일한 과정을 거치기 때문에 차이가 없지만, Score Rank Model의 경우 분할된 크기에 따라 모든 집합에서 분류 작업을 수행해야하기 때문에 분류 작업 시 많은 시간이 필요하였다. 반면 Inner Product Model은 분류 작업에 비해 계산량이 적은 내적 계산이 선행되어 데이터 집합을 선택하기 때문에 비교적 적은 시간을 사용하였다.

4.2.2 분류기에 따른 비교 (SVM)

분류기에 따른 성능의 경향성을 보기 위해 Inner Product Model을 SVM 사용하여 Naïve Bayes와 분류 성능을 비교하였다.

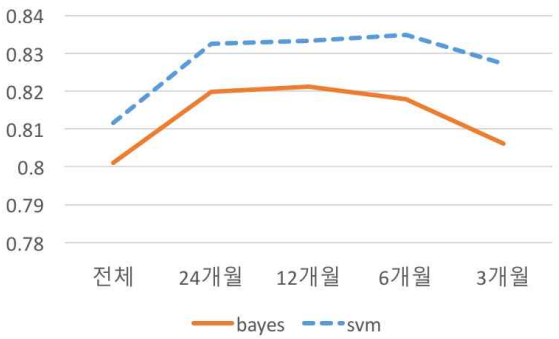

그림 12. 분류기별 비교 그래프

[그림 12]는 SVM 통해 Inner Product Model의 정확도를 Naïve Bayes에서의 결과와 비교한 그래프이다. 데이터 집합을 6개월의 단위로 분할하여 학습한 분류 정확도가 83.49%로 가장 높은 성능을 보였다. 전체 구간을 학습한 경우 81.17%로 가장 낮은 성능을 보였고 6개월 단위로 분할할 때 까지 성능이 점차 향상됨을 볼 수 있다. SVM에서도 마찬가지로 3개월 이하로 구간을 분할할 경우 성능 저하가 생기지만, Naïve Bayes 모델에 비해 세밀하게 구간을 나눌 수 있음을 확인하였다. 이는 작은 크기의 데이터 집합에서도 안정적인 성능을 보이는 SVM의 특성 때문이라고 생각 된다[9].

4.2.3 자질 수에 따른 비교

자질 수에 따른 정확도 감소를 확인하기 위해 전처리 과정에서 DF를 사용하여 자질수를 감소시킨 후 정확도를 측정하여 본 연구에서 제안하는 실험과 비교하였다. [표 3]은 전체 데이터 집합에서 DF를 사용하여 자질을 축소하였을 때 정확도를 나타낸 표이고, [표 4]는 Inner Product Model의 분할 기간별 자질의 수와 정확도를 나타낸 표이다.

표 3. DF에 따른 자질 수 감소 및 분류 정확도

DF	Unique Words (%)	Words(%)	정확도
1	100.00	100.00	80.11
3	32.54	98.66	80.11
5	21.07	98.03	80.11
10	12.70	97.10	80.11
20	8.00	95.99	80.04
50	4.46	94.01	79.50

표 4. 분할 기간에 따른 자질 수 감소 및 분류 정확도

period	Unique Words (%)	Words(%)	정확도
126 (전체)	100.00	100.00	80.11
24	32.90	19.20	82.13
12	21.16	9.82	82.08
6	13.63	4.95	81.46
3	8.82	2.49	79.56
1	4.43	0.83	75.77

DF를 통한 자질 축소의 경우 벡터의 크기(Unique Words)는 크게 감소하지만 전체 자질의 수(Words)는 감소폭이 작았다. [표 3]을 통해 자질의 수가 전체의 95% 미만으로 내려갈 때 정확도 감소의 폭이 커지는 것을 볼 수 있다. 이를 통해 분류 정확도의 경우 벡터의 크기보다는 전체 자질의 수에 영향을 받는 것을 확인할 수 있다.

반면 [표 4]를 보면, Inner Product Model의 경우 데이터 집합을 시계열 단위로 균등하게 분할하였기 때문에, 분할 구간의 크기에 따라 벡터의 크기와 자질의 수가 모두 큰 폭으로 감소한 것을 확인 할 수 있다. Inner Product Model의 경우 6개월 미만의 단위로 데이터 집합을 분할하였을 때 비교 정확도에 비해 성능이 낮아지는데, 벡터 차원의 크기와 자질의 수 감소로 인한 성능 저하라는 것을 확인 할 수 있다.

따라서 데이터 집합 분할을 할 때 각 분할된 집합의 벡터 크기 및 자질의 수를 고려하여 최적의 분할 기간을 설정해 줄 필요가 있다.

V. 결 론

본 연구를 통해 시계열 정보를 포함하는 문서를 데이터 집합 분할 및 선택 방법으로 효율적인 학습을 할 수 있음을 증명하였다. 연구의 우수성을 요약하면 다음과 같다.

- 시계열 요소를 반영하여 분류 효율을 향상 시킬 수 있음을 확인하였다.
- 데이터 집합 분할을 통해 적은 컴퓨팅 성능으로 효과적인 분류가 가능하다.

- 매일 데이터가 증가하는 환경에서 다양한 분류 모델을 통해 점진적인 학습이 가능하다.

본 연구를 통해 온라인상 문서에서 나타나는 어휘의 시계열적 변화 패턴을 분석하였고, 이를 분류에 반영한다면 효율적인 분류가 가능하다는 것을 실험을 통해 확인하였다.

SVM과 같은 분류기는 본 실험과 같은 환경에서 학습 데이터 전체를 학습하기 위해 최소 12기가 이상의 메모리를 필요로 한다[22]. 현재 학습 데이터 집합은 2004년부터 2014년까지 일자별 주요 뉴스로 약 90만 건으로, 웹상에 존재하는 모든 뉴스 기사를 학습 데이터로 사용한다면 많은 컴퓨팅 자원이 필요할 것이다. 이와 같은 환경에서 본 연구에서 제안하는 모델은 데이터 집합을 분할하여 독립적으로 학습하기 때문에 효율적인 분류가 가능하다.

또한 SVM 등의 분류기는 새로운 데이터가 추가 될 때 전체 데이터를 처음부터 학습하여야 한다. 이러한 문제는 오늘날과 같이 매일 데이터가 증가하는 환경에서 사용하기에는 비효율적이다. 하지만 본 연구에서 제안하는 모델은 데이터가 증가하면 이를 주기적으로 수집한 후 기간별로 학습시켜 점진적인 학습이 가능하다.

본 연구에서는 시계열적 요소를 분류하기 위해 AM 알고리즘을 통한 부스팅과 내적 계산을 사용하였다. 시기별 연관 단어의 분포를 반영하기 위해 계산량이 적어 속도가 빠른 내적 계산을 사용하였지만, 향후에는 내적 계산 보다 향상된 기법을 적용하여 분류의 성능을 향상시키는 것이 가능할 것이라 생각된다. 또한 어휘 변화 패턴을 클러스터링 기법 등을 통해 심도 있게 분석하여 분류에 적용한다면 유의미한 성능 향상이 가능할 것이다.

참 고 문 헌

[1] B. Croft, "Machine Learning and Information Retrieval," ICML '95, 1995.

[2] E. Jessica, "Forecast: Mobile Data Traffic, Worldwide, 2011-2018," Gartner, 2015.

[3] H. Chih and N. Kulathuramaiyer, "An empirical study of feature selection for text categorization based on term weightage," In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pp.599-602, 2004.

[4] D. Jeong, J. Kim, M. Hwang, S. Song, and H. Jung, "Classification Method by Integrating Feature Property Matrices for Large Scale Data," SMA, 2012.

[5] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI '98, 1998.

[6] Irina Rish, *An empirical study of the naive Bayes classifier*, IBM Research Report, 2001.

[7] C. Cortes and V. Vapnik, "Support-Vector Net - works," Machine Learning, 제20권, 제3호, pp.273-297, 1995.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," COLT '92, 1992.

[9] H. Taira and M. Haruno, "Feature selection in SVM text categorization," AAAI, 1999.

[10] F. Colas and P. Brazdil, "Comparison of SVM and some older classification algorithms in text classification tasks," IFIP, 2006.

[11] Pascal Soucy and Guy W. Mineau, "Beyond TF -IDF Weighting for Text Categorization in the Vector Space Model," IJCAI, 제5권, pp.1130-1135, 2005.

[12] G. Forman, "BNS Feature Scaling: An Improved Representation over TF·IDF for SVM Text Classification," ACM, 2008.

[13] Yiming Yang and Jan O. Pedersen, "A comparative study on feature selection in text categorization," ICML, 제97권, pp.412-420, 1997.

[14] Saket S. R. Mengle and Nazli Goharian, "Ambiguity Measure Feature-Selection Algorithm," Journal of the American Society for Information Science and Technology, 제60권, 제5호, pp.1037-1050, 2009.

[15] 정도헌, "최대 개념강도 인지기법을 이용한 데이터베이스 자동선택 방법에 관한 연구," 정보관리학회지, 제27권, 제3호, pp.265-281, 2010.

[16] J. Gim, Y. Jang, D. Jeong, and H. Jung, “Anayzing Email Patterns with Timelines on Researcher Data,” JIST 2014, 2014.
[17] Derry Tanti Wijaya and Reyyan Yeniterzi, “Understanding Semantic Change of Words Over Centuries,” DETECT, 2011.
[18] Do-Heon Jeong and Min Song, “Time gap analysis by he topic model-based temporal technique,” Journal of Informetrics, 제8권, 제3호, pp.776-790, 2014.
[19] https://nodejs.org
[20] http://visjs.org
[21] http://www.highcharts.com
[22] 정도헌, 정창후, 김장원, 김태홍, *빅데이터 마이닝을 위한 점진적 학습 기술 개발*, KISTI 성과보고서, 2015.

저 자 소 개

채 영 훈(Yeonghun Chae) 준회원

- 2015년 2월 : 고려대학교 전자및정보공학과(학사)
- 2015년 3월 ~ 현재 : UST 빅데이터과학(석사)
- 2015년 3월 ~ 현재 : KISTI 학생연구원

<관심분야> : 기계학습, 빅데이터, 텍스트마이닝

정 도 헌(Do-Heon Jeong) 정회원

- 1997년 2월 : 연세대학교 문헌정보학과(학사)
- 2003년 2월 : 연세대학교 문헌정보학과(석사)
- 2014년 2월 : 연세대학교 문헌정보학과 박사)
- 2003년 6월 ~ 현재 : KISTI 선임연구원
- 2015년 3월 ~ 현재 : UST 빅데이터과학 조교수

<관심분야> : 점진적학습, 기계학습, 빅데이터, 시계열분석, 네트워크분석