날짜 정보를 이용한 가중치 계산 방법을 적용한 자동 문서분류

심보준 박진우 서정연

㈜다이퀘스트 연구소 ㈜다이퀘스트 연구소 서강대학교 컴퓨터학과
simbj@diquest.com jwpark@diquest.com seojy@sogang.ac.kr

Term Weighting Using Date Information and Its Appliance in Automatic Text Classification

Bojun Shim Jinwoo Park Jungyun Seo

Diquest Inc., R&D lab. Diquest Inc., R&D lab. Department of Computer Science, Sogang University

요 약

문장을 구성하는 단어들은 문장의 의미를 표출하는 데에 있어서 모두 같은 크기의 중요도를 갖지는 않는다. 따라서, 정보검색 분야에서는 오랫동안 단어에 부여할 서로 다른 가중치를 구하는 다양한 전략을 연구해 왔다. 매우 일반적인 기능어들은 불용어로 분류하여 고려 대상에서 제외하기도 하고, 개체명 추출기를 이용하여 고유명사에 높은 가중치를 부여하거나, TF-IDF와 같이 단어가 문서 집합에 출현하는 양상과 빈도를 고려하여 가중치를 구하는 전략을 사용하기도 한다. 이와 같은 연구들에서는 같은 단어라면 어떤 상황에서도 변하지 않는 가중치를 가지게 된다.

본 논문에서는 같은 단어라 할지라도 날짜에 따라서, 어떤 날짜에는 중요한 단어이므로 높은 가중치를 받지만, 다른 날짜에는 낮은 가중치를 부여하는 전략을 제안하고 있다. 이 방법은 모든 정보검색 작업에서 사용할 수 있는 범용적인 전략이다.

본 연구에서는 특히, 문서분류 작업에 제안 방법을 적용했을 때, 제안 방법을 적용하지 않은 기본 시스템보다 분류 정확성이 더 향상되는 것을 실험을 통해서 확인하였다.

1. 서론

정보검색에서는 문장의 구성 요소인 단어의 가중치에 차이를 두는 방법이 오랫동안 연구되고 이용되어 왔다. 다음의 예와 같이 문장간의 관련도를 구하는 문제를 생각해 볼 수 있다.

질의 : 노무현-부시 정상 회담 진행
문장 1 : 노무현 대통령 워싱턴 도착 이모저모
문장 2 : 독-프 정상 회담 다음 주 파리에서

실제의 의미로는 문장 1 이 질의와 유사한 사건을 다루는 문장이지만, 단어의 가중치를 고려하지 않고, 공통으로 출현하는 단어의 개수를 세는 단순한 방법을 사용한다면 문장 2 가 문장 1 보다 질의와 더 관련도가 높은 문장으로 판가름 된다. 이 예에서 보다 정확한 결과를 구하기 위해서는 '노무현' 과 같은 고유명사에 더 높은 가중치를 부여하는 전략이 필요하다.

정보검색에서는 위의 예와 같은 문제를 해결하기 위해 단어에 일정하지 않은 가중치를 부여하는 방법, 또는 중요한 단어와 중요하지 않은 단어를 가려내는 방법이 오랫동안 연구되어 왔다.

TF-IDF를 이용한 가중치 부여 방법, 정보량이 매우 낮은 단어를 고려 대상에서 제외하는 불용어 사용 전략, 고려 대상으로 사용할 단어를 선택하는 각종 자질 선택(Feature Selection) 방법 등은 모두 단어의 정보량에 차이가 있다는 가정을 가지고 접근하는 방법들이다.

본 논문에서는 선행되었던 연구에서 밝혀졌던 것처럼 단어에 따라 중요도가 변화할 뿐만 아니라, 같은 단어라 할지라도 시간이 흐름에 따라 그 중요도가 변화한다는 것을 밝히고 있다. 이러한 현상은 시사성이 있는 신문기사 문서 집합에서 두드

러지게 나타난다. 예를 들어, 2006년 6월 기사에 나타난 '월드컵'과 2006년 12월에 나타난 '월드컵'은 그 중요도가 매우 다르다.

보다 구체적으로 본 연구에서는 단어가 기사에 출현하는 빈도의 양상을 고려하여, 같은 단어의 가중치를 날짜에 따라 다르게 구하였고, 그 가중치를 이용하여 자동 문서 분류에 적용하는 실험을 수행함으로써 이 전략의 유용성을 확인하였다.

2. 관련연구

정보검색에서 단어와 구에 가중치를 부여하는 연구는 이전에도 다양하게 진행되었다. 영어권에서는 [1], [2], [3] 등에서 용어 가중치를 계산하는 연구가 진행되었으며, 국어권에서도 또한 [4], [5] 등의 연구가 진행되어 왔다.

[4]에서는 특히 복합명사의 가중치를 구하는데 있어서 복합명사의 재출현 양상과 복합명사의 역할변화에 따른 가중치 부여 방법을 제안하고 있으며, [5]에서는 경험적 방법에 의해 용어 가중치(term weight)을 구하는 방법에 대해 연구하였다.

본 연구의 중요도가 높은 단어를 찾는 연구는 주제어 추출 연구와 밀접한 관계를 갖는다.

[5]의 연구도 용어 가중치를 주제어 추출에 이용하였다. [6]의 연구에서는 선형대수학의 주성분 분석 모델을 이용한 주제어 추출 방법을 연구하였다.

주제어 추출 기법을 정보검색의 다양한 분야에 이용한 연구도 진행되었다.

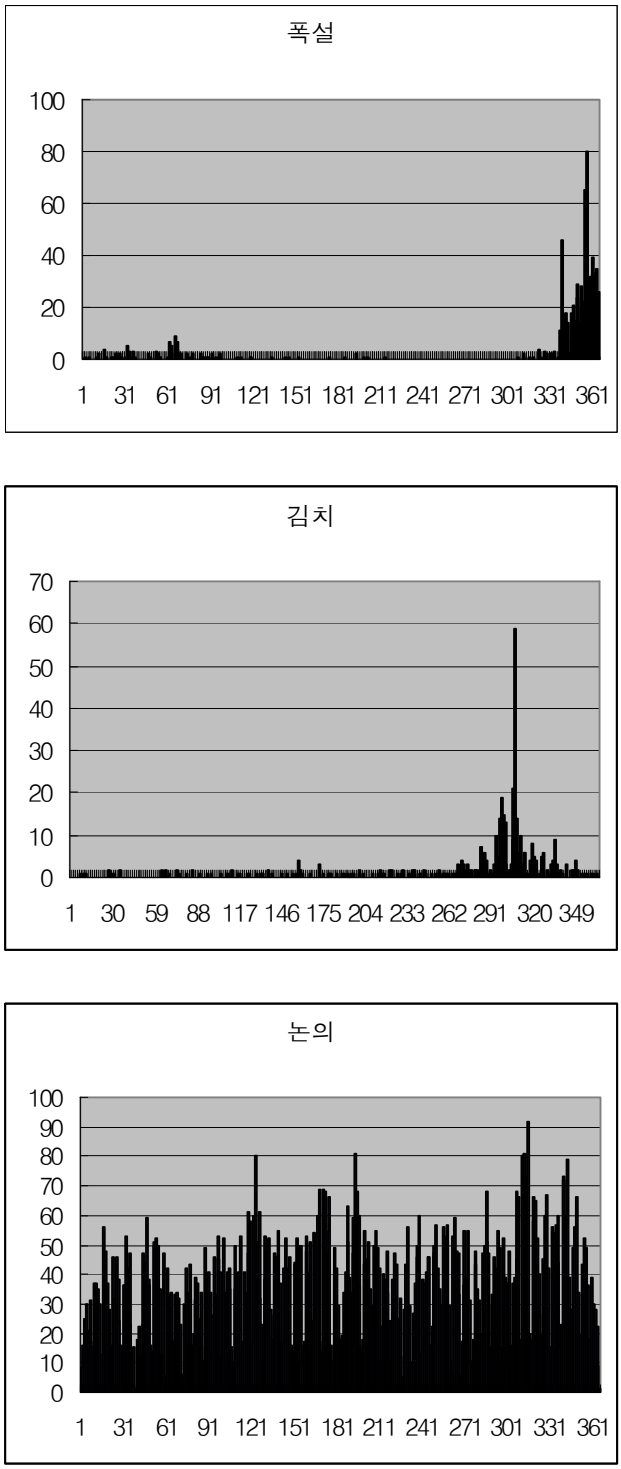
[7]에서는 주제어를 이용하여 문서 클러스터링 알고리즘에 적용하였고, [8]에서는 주제어 추출 결과를 문서 분류에 적용하였다.

3. 가중치 계산 방법 : 날짜에 따라 변화되는 가중치

3.1 날짜에 따른 단어의 중요도 변화

앞 절에서 언급하였듯이 본 논문에서는 같은 단어라 할지라도 어느 시기에 출현하는지에 따라 그 중요도가 달라진다는 점을 밝히고 있다. 이러한 특징을 파악하기 위한 첫 단계는 단어가 날짜에 따라서 출현 양상이 어떻게 변화하는지 관찰하는 것이다.

<그림 1>은 3개의 단어가 연합뉴스 2005년 기사에서 출현한 빈도를 1일 단위로 측정한 그림들이다.



<그림 1> 단어의 날짜에 따른 출현 빈도 변화

그림에서 '폭설'과 '김치'는 특정 날짜에 집중적으로 출현빈도가 급증하는 것을 알 수가 있다.

'폭설'이라는 단어는 특히 2005년 3월 하순의 영동지역 폭설 시기와 연말의 호남지역 폭설 시기에 기사에 출현하는 빈도가 급증하고 있다.

'김치' 라는 단어는 9월 말에 중국산 김치 기생충 파동이 이슈화 됨에 따라 기사에 자주 출현하였다.

그림의 예에서 보는 바와 같이 특정 시기에 갑자기 출현 빈도가 증가하는 단어는 그 시기의 이슈를 반영하는 단어일 확률이 매우 높고, 이러한 단어들은 다른 단어들 보다 더 높은 중요도를 가지게 된다.

반면 '논의' 라는 단어는 비록 자주 출현하는 단어이지만, 특정 시기에 집중되는 것이 아니기 때문에 특정 시기에 중요도가 증가하는 단어로 볼 수가 없다.

본 연구는 '폭설', '김치'와 같은 단어에 그 중요도가 증가하는 시기에 더 높은 가중치를 부여하는 전략을 제안하고 있다.

3.2 특정 날짜, 특정 단어의 가중치

특정 날짜에 특정 단어의 가중치를 계산하는 식은 두 가지 사항을 고려하여 만들어졌다.

첫째, 그 날짜에 단어의 출현 빈도가 높다면 높은 가중치를 준다.

둘째, 단어가 모든 시기에 걸쳐 골고루 출현하고 있다면 즉, 출현하는 날이 많다면 가중치를 낮춘다.

이 두 가지 고려사항으로 다음과 같은 가중치 계산식이 만들어졌다.

$$Weight_w(d,y) = Count_w(d) \times Log\left[\frac{365}{Count_w(y)}\right] + 1 \qquad (1)$$

$Weight_w(d, y)$: y 년도, d 날짜의 w 단어의 가중치

$Count_w(d)$: d 날짜에 w 단어의 출현 문서 수

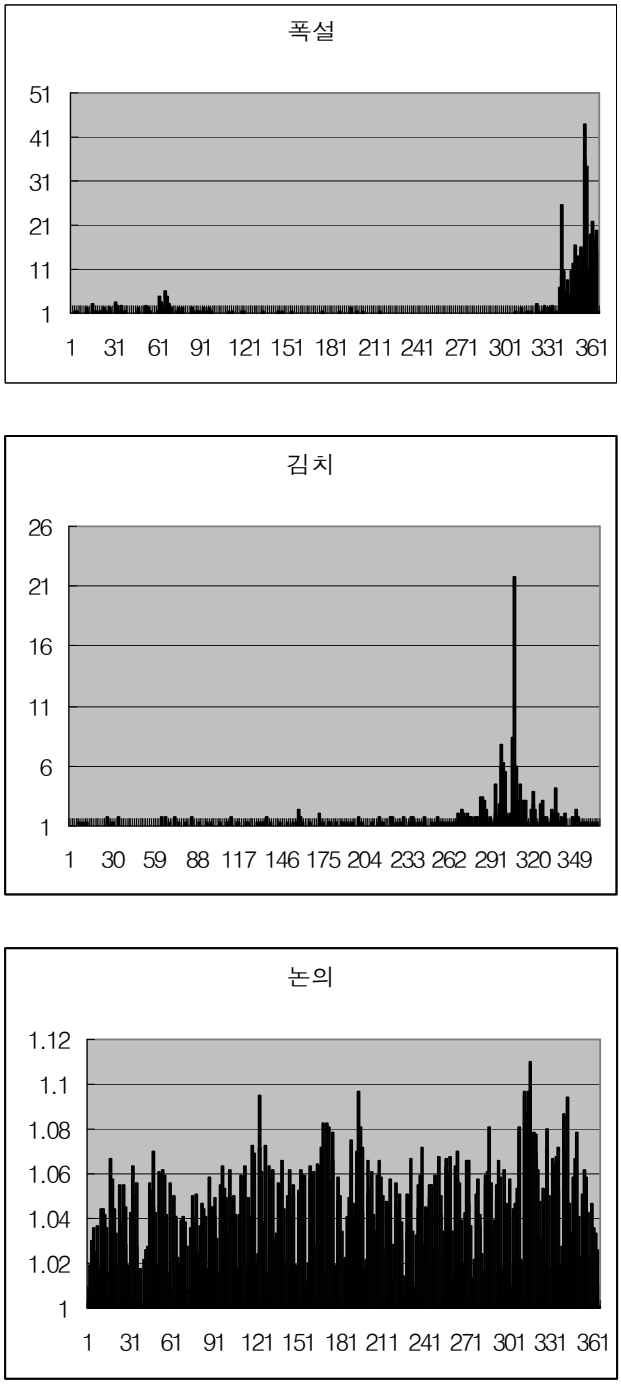
$Count_w(y)$: y년도에 w 단어가 출현한 날짜 수

위 수식은 tf-idf 가중치 계산식을 참고하여 만들어졌으며, 수식의 오른쪽 텀은 Log 함수를 스무딩(smoothing)에 이용하였다.

<그림 2> 는 <그림 1> 에 나타낸 각 단어를 수식 (1) 을 통해서 날짜별 가중치를 구한 그래프 이다.

'폭설', '김치' 와 '논의'의 가중치를 비교해 보면 <그림 1> 에서 그 출현 빈도수가 비슷하거나 낮은 경우에도 <그림 2> 에서 '폭설', '김치'의 가중치가 '논의'의 가중치 보다 훨씬 높게 나타나는 것을 볼 수 있다.

<표 1> 은 2005년 12월 8일을 예로 들어, 각 단어의 출현 빈도와 가중치를 보여주고 있다. <표 1> 에서도 '폭설' 과 '김치' 가 출현빈도는 낮지만 가중치는 '의논' 보다 높은 것을 볼 수 있다.



<그림 2> 단어의 날짜에 따른 가중치 변화

단어	2005년 12월 8일 출현 문서수	2005년 출현 날짜수	가중치
폭설	9	106	5.83
김치	6	250	1.99
논의	79	364	1.09

<표 1> 2005년 12월 8일 출현빈도와 가중치

4. 문서 분류 실험

단어의 가중치를 계산하는 일은 정보검색 분야에서 기본이 되는 작업이므로 정보검색 전 영역에 걸쳐 두루 이용할 수 있다. 본 연구에서는 많은 영역 중에서도 문서 분류 작업에 제안한 방법을 적용하여 그 유용성을 실험하였다.

4.1 연합뉴스 문서집합

제안한 방법은 시간의 경과에 따라서 꾸준하게 문서가 생성되고, 중요 단어가 시기에 따라 급변하는 신문기사 영역에서 매우 유용한 방법이다. 따라서, 연합뉴스에서 생산된 2005년 국문 글기사 집합을 문서 분류 실험에 이용하였다. 연합뉴스 기사는 기사 작성시 마다 기자가 분류 카테고리를 1개 이상 등록하도록 하고 있어서, 문서 분류 실험을 위해서 매우 좋은 문서 집합이다. 분류 카테고리는 내용, 지역, 속성에 따른 분류 기준이 있는데, 본 실험에서는 내용 카테고리만 사용하였다. 또한 카테고리는 대, 중, 소 분류로 계층화 되어 있는데, 본 실험에서는 중분류 카테고리 73개를 사용하였다.

4.2 가중치 계산을 위한 준비

본 연구에서 제안하는 방법으로 단어의 날짜별 가중치를 구하기 위해서는 단어가 날짜별로 출현하는 횟수를 가지고 있어야 한다. 실험에서는 다음의 순서로 단어의 추출과 가중치 계산을 실행하였다.

1단계 : 2005년 모든 글기사의 제목과 본문에 대해 형태소 분석을 실행하고, 명사 단어를 모두 추출한다.

2단계 : 추출된 단어의 날짜별 출현빈도와 2005년 한해 동안 출현한 날짜의 수를 카운팅하여 기록해 놓는다.

3단계 : 특정 날짜, 특정 단어의 가중치를 2단계에서 기록해 놓은 두 가지 정보를 수식 (1)에 대입하여 구했다. 이 과정은 모든 단어에 대해 미리 계산해 놓을 수도 있고, 문서분류 실행 단계에 계산할 수도 있다.

4.3 문서 분류기

자동 문서 분류에는 사용 가능한 기계학습 알고리즘이 여러 가지가 있다. 본 연구에서는 Naïve Bayes 문서 분류 알고리즘 위에 제안하는 가중치 부여 방법을 적용하여 성능이 향상되는지 실험하였다. 기본적인 Naïve Bayes Classifier 위에 제안한 가중치 부여 방법을 적용한 변형된 수식은 다음과 같다.

$$C = \arg\max_{c} p(C=c)\prod_{i=1}^{n} weight_i(y,d)\,p(w_i \mid C=c) \quad (2)$$

C : 출력 카테고리
c : 선택 가능한 카테고리(73개중 1개)
w_i : 문서를 구성하는 i 번째 단어
$weight_i(y, d)$: 수식 (1) 에 의해 계산된 y 년도, d 날짜의 i 번째 단어의 가중치

4.4 자질 선택

Yiming Yang은 [9]에서 여러 가지 자질 추출 방법을 사용하여 실험을 한 결과 카이 제곱 통계량과 정보 획득량을 사용하는 것이 효과적임을 보였다. 본 논문에서는 이를 바탕으로 비교적 구현이 쉽고 고빈도 단어에 친화적인 카이 제곱 통계량을 사용하여 카이제곱 통계량이 상위인 명사를 자질로 사용하였다. 자질 선택에서 배제된 형태소는 분류과정에서 완전히 배제 하였다.

4.5 실험 집합

학습 집합과 테스트 집합의 모집합은 연합뉴스의 2005년 국문 글기사 집합이고, 전체 219,078개의 문서를 무작위로 20%는 테스트 집합으로, 80%는 훈련집합으로 분류하여 학습과 테스트를 실행하였다.

5. 실험 결과

5.1 결과

실험에서는 모든 단어에 대해 균등한 가중치를 부여하는 시스템을 기본 시스템으로 하고, 제안한 가중치 계산 방법을 사용한 시스템을 제안 시스템으로 실험하였다. 또한, 각각에 대해 자질 선택을 상위 5,000개, 10,000개로 나누어 실험을 진행하였다.

진행한 실험의 결과는 <표 2> 와 같다.

구분	학습 문서	테스트 문서	5000 자질	10000 자질
기본 시스템	175263건	43815건	70.20%	69.00%
제안 시스템			71.09%	70.13%

<표 2> 실험결과

실험 결과 사용된 자질의 개수와 상관없이 제안 시스템이 기본 시스템 보다 더 높은 정확률을 기록하였다.

5.3 결과 검토

진행한 실험에서 제안한 방법이 성능을 향상시킨 두 가지 예를 들어보도록 하겠다.

2005년 1월 1일자 기사
제목 : “지진, 해일 참사 이모저모"
본문 : 영국, 새해 맞이 행사서 구호기금 모금… 캐나다, 미국 주도 해일 구호연합 참여…

기본 시스템에서는 위 기사가 영국, 캐나다, 미국 등 나라 이름들이 출현하는 데에 영향을 받아 ‘외교’ 카테고리로 분류되는 오류를 일으켰다. 하지만, 제안 시스템에서는 ‘지진’, ‘해일’ 이라는 단어들이 1월 1일 즈음에 집중적으로 발생함에 따라서 높은 가중치가 부여되었고, ‘자연재해’ 라는 카테고리로 옳게 분류가 되었다.

2005년 3월 30일자 기사
제목 : “유엔 다르푸르 제재 결의안 통과”
본문 : 유엔 안전보장이사회는 29일 수단 다르푸르 지역에서…

위 기사 역시 ‘다르푸르’ 라는 문장 내 핵심적인 키워드에 높은 가중치가 할당됨에 따라서 분류 결과가 개선된 예이다. 기본 시스템에서는 ‘유엔’, ‘안전보장이사회’ 등의 단어들에 의해 ‘외교’ 카테고리로 잘못 분류되었으나, 제안 시스템에서는 ‘다르푸르’ 라는 단어에 높은 가중치가 부유됨에 따라 ‘국제’ 카테고리로 옳게 분류되었다.

위의 두 가지 전형적인 예에서 보는 바와 같이 제안한 방법은 문서의 주제어에 높은 가중치를 부여함으로써 정보검색의 성능을 향상시키는데 의미있는 역할을 하고 있음을 알 수 있다.

6. 결론 및 향후 과제

본 연구에서는 같은 단어라 할지라도 출현하는 날짜에 따라서 다른 가중치를 부여하는 방법을 제안하였고, 그 유용성을 확인하기 위해 신문기사의 자동 문서 분류 실험에 제안된 방법을 적용하여 성능의 향상을 확인하였다.

본 연구는 또한, 다음과 같은 향후 연구 과제를 가지고 있다.

첫째, 본 연구에서는 Naïve Bayes 문서 분류방법만을 이용하여 실험을 하였지만, SVM 이나, KNN 알고리즘 등 더 많은 기계학습 알고리즘을 이용하여 제안 방법이 범용적으로 유용한 방법인지 확인할 필요가 있다.

둘째, 본 연구에서는 문서 분류 분야에서만 실험을 진행하였지만, 더 많은 정보 검색 작업에 제안 방법을 적용해 볼 수 있을 것이다. 특히, 이벤트 추적, 이슈 추적 등의 작업은 본 연구의 특성과 매우 잘 부합하는 성격의 작업이라고 판단된다.

참고 문헌

[1] Salton, G.(1986) : “Recent Trends in Automatic Information Retrieval”, Proceedings of 1986 ACM Conference on Research and Development in Information Tetrieval

[2] Salton, G. and C. Buckley(1988) : “The Term-Weighting Approaches in Automatic Text Retrieval”, Information Processing and Management

[3] Spark Jones, K(1973) : “Indexing Term Weighting”, Information Storage and Retrieval, vol. 9, no. 11

[4] 손기준, 이상조(2004) : “특허 문헌 검색에서 복합명사 가중치 부여 방법”, 한국정보과학회 2004년도 봄 학술발표논문집 제31권 제 1호(B)

[5] 강승식, 이하규, 손소현, 홍기채, 문병주(2001) : “조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법”, 한국정보과학회 2001년도 가을 학술발표논문집 제28권 제 2호(II)

[6] 이창범, 김민수, 이기호, 이귀상, 박혁로(2002) : “주성분 분석을 이용한 문서 주제어 추출”, 정보과학회논문지 : 소프트웨어 및 응용 제29권 제 9.10호

[7] 장성호, 강승식(2002) : “주제어 기반 문서 클러스터링 알고리즘”, 한국정보과학회 2002년도 봄 학술발표논문집 제29권 제 1호(B)

[8] 안희국, 노희영(2005) : “문서 분류를 위한 문자 응집도와 주어주도의 주제어 추출”, 한국정보과학회 2005 한국컴퓨터종합학술대회 논문집(B)

[9] Yang, Y., Pedersen, J.O.(1997) : “A Comparative Study on Feature Selection in Text Categorization”, Proceedings of The 14th International Conference on Machine Learning