# 정규화 용어빈도가중치에 의한 자동문서분류

김수진*0 김민수** 백장선*** 박혁로*
전남대학교 전산학과*, 통계학과***, 한국과학기술원 전산학과**
{ suchin@dal*0, kimms@**, jbaek@**, hyukro@cs* }.chonnam.ac.kr

# Automatic Text Categorization by using Normalized Term Frequency Weighting

Su-Jin Kim*0 Min-Soo Kim** Jang-Sun Baek*** Hyuk-Ro Park*
Department of { Computer Science, Statistics }, Chonnam National University,
Department of Computer Science, KAIST

요 약

본 논문에서는 문서의 자동 분류를 위한 용어 빈도 가중치 계산 방법으로 Box-Cox변환기법을 응용한 정규화 용어빈도 가중치를 정의하고, 이를 문서 분류에 적용하였다. 여기서 Box-Cox 변환기법이란 자료를 정규분포화 할 때 적용하는 통계적인 변환방법으로서, 본 논문에서는 이를 응용하여 새로운 용어빈도 가중치 계산법을 제안한다. 문서에서 등장한 용어 빈도는 너무 많거나 적게 등장할 경우, 중요도가 떨어지게 되는데, 이는 용어의 중요도가 빈도에 따른 정규분포로 모델링 될 수 있다는 것을 의미한다. 또한 정규화 가중치 계산방법은 기존의 용어빈도 가중치 공식과 비교할 때, 용어마다 계산방법이 달라져, 로그나 루트와 같은 고정된 가중치 방법보다는 좀더 일반적인 방법이라 할 수 있다. 신문기사 8000건을 대상으로 4개의 그룹으로 나누어 실험 한 결과, 정규화 용어빈도가중치 계산방법이 모두 우위의 분류 정확도를 가져, 본 논문에서 제안한 방법이 타당함을 알 수 있다.

## 1. 서론

문서 자동 분류 방법에는 크게 통계적방법과 의미 분석방법으로 구분할 수 있다. 후자의 경우 자연어 자체의 모호성 때문에 그 사용이 어렵고 한정되어 있는 반면, 통계적 방법은 간단히 구현할 수 있고 학습이 가능하여, 충분한 학습 데이터가 주어졌을 경우 의미 분석방법에 버금가는 결과를 낼 수 있다. 따라서 최근 많은 연구들이 통계적인 방법을 사용하고 있다[1,2,3].

통계적 문서 분류 방법에서는 문서를 대표하는 용어와 이것의 가중치(대표성)를 결정하는 방법이 필요하다. 용어의 가중치의 계산에 용어의 문서 내 빈도를 고려하는 경우가 있는데, 이것을 용어빈도 가중치라고 한다. 본 논문에서는 이러한 용어빈도 가중치를 계산하는 새로운 방법으로서 Box-Cox변환을 응용한 정규화 가중치 계산방법을 제안한다.

본 방법은 용어빈도의 분포를 정규분포에 근사하게 변환하여, 변환된 빈도를 가중치 계산에 적용하는 방법으로, 문서에 나타나는 용어 중에 출현 빈도가 너무 높거나 낮은 것 보다는 중간 빈도로 나타나는 용어가 문서의 내용을 더 잘 대표한다는 직관적인 논리에 근거하여, 중간 빈도의 용어일수록 가중치를 높게 부여한다.

본 논문에서 제안한 정규화 가중치 계산방법을 기존 가중치 계산 방법과 비교해 실험해 본 결과, 단순TF 방법보다 약 4~6%의 성능 개선을 보였으며, 4개 그룹 모두에서 항상 우위의 성능을 보였다. 따라서 정규화 용어빈도 가중치가 다른 가중치 방법보다 일반적이고 효과적임을 알 수 있었다.

이후 본 논문의 구성은 다음과 같다. 2장에서는 기존의 용어빈도 가중치 계산방법, 3장에서는 정규화 용어빈도 가중치 계산방법, 4장에서는 실험 및 결과, 마지막으로 5장에서는 결론 및 향후연구에 대해 기술할 것이다.

## 2. 기존의 용어빈도 가중치 계산방법

용어빈도 가중치 계산방법은 문헌 내 출현 어부만을 반영하는 이진 값이나 출현빈도 자체를 가중치로 사용할 수도 있으며, 이외의 다양한 공식이 제안되었다. 이렇게 용어에 가중치를 부여함은 한 문서가 취급하고 있는 개념들의 주제적 요소로서의 중요도에 따라, 색인어로서 상대적 가치를 표현하기 위함이다. 기존의 용어빈도 가중치 계산방법들은 아래와 같으며, 여기서 tf (Term Frequence)란 용어가 한 문서 내에서 나온 빈도수를 의미한다[1].

(1) 단순 : $TF = tf$

(2) 이진 : $TF = 1(if\ tf > 0), 0$

(3) 로그 : $TF = 1 + log(tf)$

(4) 더블로그 : $TF = 1 + log(1 + log(tf))$

(5) 루트 : $TF = \sqrt{tf}$

(6) 보정 : $TF = (1-w) + w \times \frac{tf}{max\_tf}$

(7) Okapi : $TF = \frac{tf}{2 + tf}$

(8) 더블로그2 : $TF = 1 + log_2(1 + log_2(tf))$

(9) 루트직선 : $TF = \frac{tf+3}{4}$

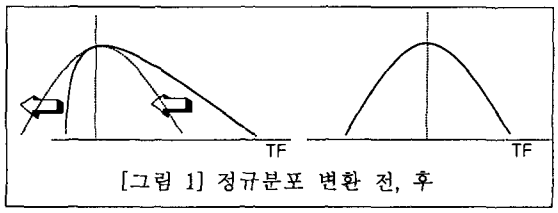
하지만 이런 방법들은 어떤 경우에 좋은 성능을 보이는지에 대한 명확한 기준이 없기 때문에, 연구자가 자료의 형태를 보고 경험적으로 함수를 취해서 수행하는 경우가 많다. 이에 본 논문에서는 일반적인 가중치 부여 방법으로 정규화 용어빈도 가중치 계산 방법을 제안한다.

## 3. 정규화 용어빈도 가중치 계산방법

### 3.1 용어빈도의 정규분포

용어빈도 자료는 문서에서 너무 높거나 낮을 경우 중요도가 떨어지며, 적정 수준에서의 용어빈도 값의 중요도가 높다. 다시 말해, 보통 문서에서 한 용어가 등장하면 그 용어는 문서의 분류에 도움이 되지만, 그 빈도수가 너무 많거나 적다면 그 중요도는 낮아진다는 것이다. 예를 들어 한 용어가 너무 많이 등장한 경우, 이 용어는 전체적인 분산을 증가시켜 변별력을 감소시켜 분류의 정확도를 떨어뜨릴 수 있다.

하지만 실제 문서분류에서 이용되는 용어빈도 자료는 보통 0의 빈도가 가장 많고 다음으로 1의 빈도, 2의 빈도 순서로 분포하는, 오른쪽으로 긴 꼬리를 갖는 왜도가 양수인 분포이다.




[그림 1] 정규분포 변환 전, 후

이러한 자료를 위 [그림 1]과 같이 정규화 시킨다면, 한쪽으로 치우친 분포에서 발생할 수 있는 이상점을 보다 평균방향으로 오도록 하여 그 이상점의 영향력을 줄일 수 있게 된다.

기존의 용어빈도 가중치 계산방법들 중, 단순TF에 비해 로그TF나 루트TF 등이 좋은 수행력을 보였던 것도 같은 맥락이라 할 수 있다. 즉, 기존의 단순 TF를 변형시킨 여러 방법들은 용어빈도가 낮거나 높은 경우 사이의 가중치 차이를 어떻게 줄 것인가를 판단하는 문제를 루트를 씌우거나 로그를 취해 해결하고 있다. 그러므로 용어빈도가 높은 오른쪽 부분을 줄여주고 용어빈도가 낮은 왼쪽 부분을 늘려주는 정규분포 개형으로 만드는 방법과 비슷하다 말할 수 있다.

### 3.2 정규화 용어빈도 가중치 계산방법

본 논문에서 제안한 정규화 용어빈도 가중치 부여방법의 계산식은, 자료를 정규분포로 만드는 Box-Cox변환기법을 토대로 만들어 졌으며, 그 식은 아래와 같다[4].

$\lambda$가 연속일 때 아래와 같은 변환을 정의하고,

$$TF^{(\lambda)} = \begin{cases} \frac{(tf+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(tf+1), & \lambda = 0 \end{cases}$$

주어진 자료 $tf_1, tf_2, \cdots, tf_n$ 에 대해, 아래 식을 최대로 하는 모수 $\lambda$를 선택한다.

$$l(\lambda) = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{j=1}^{n}(TF_j^{(\lambda)} - \overline{TF^{(\lambda)}})^2\right] + (\lambda-1)\sum_{j=1}^{n}\ln(tf_j+1)$$

여기서 $\overline{TF^{(\lambda)}} = \frac{1}{n}\sum_{j=1}^{n}TF_j^{(\lambda)} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{tf_j^{\lambda}-1}{\lambda}\right)$ 이다.

위 식에서 $\lambda$값에 따른 용어빈도 값의 변화를 살펴보면, $\lambda$=0 일 경우는 로그TF, $\lambda = \frac{1}{2}$ 인 경우는 루트TF, $\lambda$=1인 경우에는 단순 TF가 된다. 즉, 정규화 용어빈도 가중치 계산방법은 기존의 여러 가중치 계산 방법들을 포함한 TF의 멱승변환 중에서, 현재 자신이 가지고 있는 용어빈도자료를 가장 정규화 시켜주는 가중치를 찾는 방법이라고도 말할 수 있다.

## 4. 실험 및 결과

### 4.1 실험 환경 및 실험 방법

문서 분류의 실험 대상 문서 집단으로는 동아일보 신문기사에서 임의로 추출한 8개의 카테고리로 이루어진 8000건의 기사를 이용하였다. 그리고 이를 4개 그룹으로 나눠, 각각 75%는 실험 집단으로 25%는 검증집단으로

실험하였다. 또한 실험에 사용한 용어는 모든 카테고리에서 용어빈도가 많이 나오는 순서로 상위 1000개를 선택하였다.

가중치 부여 방법으로는, 기존의 여러 용어빈도 가중치 공식들과 정규화 용어빈도 가중치 계산방법을 비교하기 위해, 모두 10가지 방법으로 실험하였다. 이때, 제안한 정규화 용어빈도 가중치 방법은 모든 용어에 대해서 $\lambda$를 계산해서 변환해야 하지만, 본 실험에서는 가장 빈도가 높은 용어에서의 $\lambda$값을 모든 자료에 적용하였다.

### 4.2 실험 결과

동아일보 기사를 대상으로 분류를 실험한 결과를 아래 표에 나타내었다. 왼쪽의 (1)~(9)는 기존의 용어빈도 가중치를 나타낸 것이며, (10)은 본 논문에서 제안한 정규화 용어빈도 가중치 계산 방법이다.

| | Group 1 | | Group 2 | | Group 3 | | Group 4 | |
|---|---|---|---|---|---|---|---|---|
| | train | test | train | test | train | test | train | test |
| (1) | 88.9 | 78.2 | 89.2 | 78.3 | 88.2 | 78.4 | 86.7 | 78.1 |
| (2) | 91.8 | 81.5 | 91.6 | 81.5 | 91.7 | 81.5 | 90.7 | 82.3 |
| (3) | | | | | 91.8 | 82.4 | 90.7 | 81.3 |
| (4) | | | | | 92.0 | 82.1 | | |
| (5) | 91.9 | 81.8 | | | | | 90.6 | 81.1 |
| (6) | 89.0 | 78.3 | 89.2 | 78.3 | 88.2 | 78.4 | 87.8 | 77.4 |
| (7) | | | | | | | 91.7 | 81.4 |
| (8) | 91.2 | 81.4 | 91.5 | 80.8 | 91.3 | 80.4 | 89.5 | 78.9 |
| (9) | 89.0 | 78.3 | 89.2 | 78.3 | 88.0 | 78.1 | 87.8 | 77.1 |
| (10) | | | | | | | | |

[표 1] 동아일보 분류실험 정확도 결과 (단위: %)

위 실험결과를 보면 정규화 용어빈도 가중치 계산 방법이 모두 상위의 정확도 그룹에 속해있음을 알 수 있다. 이는 용어빈도 분포개형이 정규분포라면 분류 성능이 향상될 것이라는, 본 논문의 가정이 옳음을 입증하는 것이다. 또한 항상 상위 그룹에 속한 정규화 가중치 방법과 달리, 다른 가중치 계산방법은 실험 그룹에 따라 다른 순위를 보임을 볼 수 있다. 이는 용어빈도 데이터에 따라 다른 가중치부여방법을 가져야 하기 때문에, 고정된 가중치방법보다 $\lambda$에 따라 변환식이 달라지는 정규화 가중치 계산방법이 좀더 일반적인 가중치 계산방법이 될 수 있다는 것을 의미한다.

## 5 .결론

본 논문에서는 문서의 자동 분류 시 사용하는 색인어에 대한 새로운 가중치 계산 방법으로 Box-Cox변환을 응용한 정규화 가중치 계산방법을 제안하였다.

본 방법은 문서에서 많이 나오거나 적게 나오는 용어는 문서를 잘 대표하는 키워드로서 부적당하며, 적정 수준 등장하는 용어들이 키워드로서 더 적당하다는 언어적 직관에 근거하여, 용어빈도의 분포를 정규분포로 변환하여 적용하는 방법이다.

실험에서 본 논문의 가중치 계산 방식은 다른 가중치 계산 방법들에 비해 좋은 결과를 보이고 있는데, 이것으로 보아 Box-Cox변환 정규화 가중치 계산 방법에서 사용한 정규화 방법은 언어적 직관에 비교적 잘 들어맞는 모델로 볼 수 있다.

앞으로 분류의 성능을 더욱 높이기 위해 각 색인어에 따라 $\lambda$값을 다르게 적용 시키는 방법과, 용어빈도 외의 자질을 정규화 용어빈도 가중치 기법과 함께 사용했을 때의 문서 분류방법 등에 대한 연구가 필요할 것이다.

## 참고문헌

[1] 이재윤, 최보영, 정영미, "문헌 자동분류에서 용어 가중치 기법에 대한 연구," *제 7회 한국정보관리학회 학술대회 논문집*, 2000

[2] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류," *정보과학회 학술발표 논문집, 4,* 1997

[3] *강승식, 한국어 형태소 분석과 정보검색*, 홍릉과학 출판사, 2002

[4] R.A.Johnson and D.W.Whichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 1987

[5] Y.Yang, J.O.Pederson, " An evaluation of statistical approaches to text categorization," *In Proceeding of the 24th International Conference on Machine Learning*, 1997

[6] 김상범, "*범주간의 관계를 통한 자동 문서 범주화*," 고려대 이학 석사학위논문, 1999

[7] *정영미, 정보검색론*, 구미무역 출판부, 1993