자동 문서분류에서의 정규화 용어빈도 가중치방법

김수진*, 박혁로**
전남대학교 자연과학대학 전산학과

Normalized Term Frequency Weighting Method in Automatic Text Categorization

Suchin Kim*, Hyukro Park**
Computer Science Department, Chonnam National University
E-mail : *sj8052@magicn.com, **hyukro@chonnam.ac.kr

Abstract

This paper defines Normalized Term Frequency Weighting method for automatic text categorization by using Box-Cox, and then it applies automatic text categorization. Box-Cox transformation is statistical transformation method which makes normalized data. This paper applies that, and suggests new term frequency weighting method. Because Normalized Term Frequency is different from every term compared by existing term frequency weighting method, it is general method more than fixed weighting method such as log or root. Normalized term frequency weighting method's reasonability has been proved though experiments, used 8000 newspapers divided in 4 groups, which resulted high categorization correctness in all cases.

I. 서론

문서 자동분류란, 미리 정의되어 있는 범주를 문서의 내용에 기반 하여 컴퓨터가 자동으로 할당하게 하는 작업이다[3,7,8]. 문서 자동분류를 위해서는 문서를 대표하는 용어와 이것의 가중치를 결정하는 방법이 필요한데, 이때 가중치의 계산에 용어의 빈도를 사용하는 것을 용어빈도 가중치라고 한다. 본 논문에서는 이러한 용어빈도 가중치를 계산하는 정규화 가중치 계산방법을 제안한다. 본 방법은 용어의 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[4]. 또한 정규화 가중치 계산방법은 데이터에 따라 계산식이 바뀜으로써 기존에 제안한 방법에 비해 좀더 일반적인 방법이 될 수 있다.

본 논문에서 제안한 정규화 가중치 계산방법을 기존 가중치 계산 방법과 비교해 실험해 본 결과, 항상 우위의 성능을 보여, 제안한 방법이 일반적이고 효과적임을 알 수 있었다.

II. 기존의 용어빈도 가중치 계산방법

용어에 가중치를 부여함은 한 문서가 취급하고 있는 개념들의 주제적 요소로서의 중요도에 따라, 색인어로서 상대적 가치를 표현하기 위함이다. 기존에 소개된 용어빈도 가중치 계산방법들은 아래 <표 1>과 같으며, 여기서 tf 란 용어가 한 문서 내에서 나온 빈도수를 의미하며, TF 는 변환된 tf 를 가리킨다[2,6].

<표 1> 기존의 가중치 계산 방법

이 름	공 식
단순 TF	$TF = tf$
이진 TF	$TF = 1\ (if\ tf > 0),\ 0$
로그 TF	$TF = 1 + \log(tf)$
더블로그 TF	$TF = 1 + \log(1 + \log(tf))$
루트 TF	$TF = \sqrt{tf}$
보정 TF	$TF = (1-w) + w \times \frac{tf}{\max\_tf}$
Okapi TF	$TF = \frac{tf}{2 + tf}$

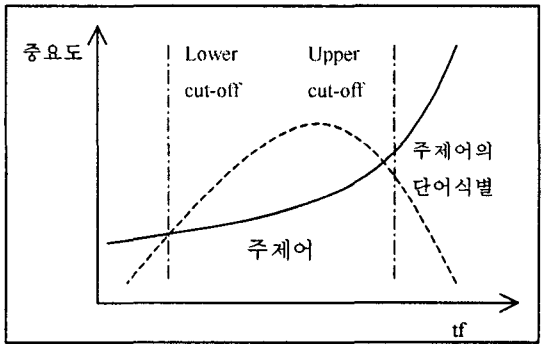
더블로그 2TF	$TF = 1 + \log_2(1 + \log_2 tf)$
루트직선 TF	$TF = \frac{tf + 3}{4}$

하지만 위의 여러 방법들은 어떤 경우에 좋은 성능을 보이는지에 대한 기준이 없다. 따라서, 연구자가 자료의 형태를 보고 경험적으로 함수를 취해서 수행하는 경우가 많은데, 이에 본 논문에서는 정규화 용어빈도 가중치 계산 방법을 제안한다.

III. 정규화 용어빈도 가중치 계산방법

3.1 용어빈도의 정규분포

Luhn 은 문서에 출현한 단어들은 문서의 내용 분석을 위해 사용될 수 있으며, 또한 단어의 출현빈도가 단어의 주제어로서의 중요성을 측정하는 기준이 된다고 말했다[4]. 그의 연구에서, 단어빈도와 그들의 순위에 따른 그래프는 다음과 같다.



<그림 1> 단어빈도와 단어빈도순위의 관계

Luhn 은 아래와 같이 두 개의 upper, lower cut-off 를 설정해, 비주제어를 배제한다. 즉, upper cut-off 를 초과한 고빈도어는 일반적인 단어로 , lower cut-off 아래의 저빈도어는 문서에 중요하게 기여하지 않는다고 볼 수 있다. 또한 주제어의 단어식별은 두 개의 양쪽 cut-off 사이에서 최고점에 도달하여 문서 내용의 식별력이 크다고 말하고 있다. 직관적으로 보통 문서에서 한 용어가 등장하면 그 용어는 문서의 분류에 도움이 되지만, 그 빈도수가 너무 많거나 적다면 그 중요도는 낮아진다. 즉 중간빈도 자료가 높은 중요도를 가진다는 Luhn 의 연구에 기반해, 용어빈도 자료가 정규분포 모형을 보일 때 좋은 성능을 나타낸다.

정규분포를 만드는 것은 다음과 같은 효과가 있다. 먼저 tf 가 높은 부분을 낮게 조절하여, 지나치게 높은 빈도의 색인어에 대한 영향력을 줄인다. 일반적으로 고빈도 용어는 문서에서 흔하게 쓰여, 다른 문서와의 변별력에 도움을 주지 못한다. 두 번째, 중간 빈도의 tf 를 높게 조절하여, 중요한 용어의 가중치를 올려준다. 마지막으로 저빈도의 용어에 낮은 tf 값을 부여하여, 문서를 표현하는데 적절하지 않는 값의 중요도를 낮춘다. 이런 방식으로 원래의 tf 데이터를 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, 단순 TF 에 비해 로그 TF 등이 좋은 수행력을 보였던 것도 같은 맥락이라 할 수 있다. 즉 용어빈도가 낮거나 높은 경우 사이의 TF 차이를 어떻게 줄 것인가를 판단하는 문제를 루트를 씌우거나 로그를 취해 해결한다. 그렇지 않으면 문서의 표현 시 높은 빈도의 용어에 편향되어, 상대적으로 그 문서를 잘 나타내는 중간빈도 용어의 영향력을 줄이게 된다. 하지만 <그림 1>과 같이 tf 자료를 변형시킨다면, 고빈도 용어가 문서에 끼치는 영향력을 줄임과 동시에 중간빈도의 용어를 부각시킬 수 있게 된다.

또한 기존의 로그 TF 나 루트 TF 등 고정된 계산방법은 연구자가 자료의 형태를 보고 경험적으로 함수를 취하기 때문에, 많은 계산 방법들 중에서 적절한 방법을 찾을 수 없다. 하지만 정규화 용어빈도 가중치 계산방법의 경우, 자료의 형태에 따라 계산방법이 달라져, 현재 가지고 있는 자료에 가장 적절한 방식으로 변형시킬 수 있다.

3.2 정규화 용어빈도 가중치 계산방법

본 논문에서 제안한 정규화 용어빈도 가중치 부여방법의 계산식은, 자료를 정규분포로 만드는 Box-Cox 변환기법을 응용하였다[1].

Box-Cox 변환기법은 1964 년 Box 와 Cox 에 의해 제안된 반응변수의 변환모형이다. 이 방법은 자료의 정규성을 검토해, 만일 자료들이 정규분포로부터 많이 벗어나 있는 경우 다음과 같은 변환을 시행한다.

$$Y' = \begin{cases} Y^k & (if, k \neq 0) \\ \ln Y & (if, k = 0) \end{cases}$$

위의 형태로 Y 를 변환시켜 주는 것을 Box-Cox 변환이라고 하며, k 값은 자료의 분포에 따라 결정되는 파라메터이다.

Box-Cox 는 적당한 상수에 대해 변환을 시행하면, 변환한 결과는 정규분포에 가깝다고 제안했다.

본 논문에서 제안한 정규화 가중치 계산방법은 앞서 설명한 Box-Cox 변환기법을 토대로 만들어 졌으며, 그 식은 아래와 같이 정의할 수 있다.

변환 파라메터 λ 가 연속일 때 아래와 같은 변환을 정의하고, $TF^{\lambda} = \frac{(tf+1)^{\lambda}-1}{\lambda}(tf, \lambda \neq 0), \ln(tf+1)$

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

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

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

여기서 λ 는 원래의 자료를 정규분포로 가장 근접하게 해주는 값으로, λ 의 값이 커지면 자료의 분포를 넓게 퍼지도록 해주고 값이 작아지면 분포를 좁게 만든다.

정규화 가중치 계산 방법에서 λ 값에 따른 TF 값 변화에 대한 몇 가지 예를 들어보면 아래와 비슷한 형태로 자료를 변형시킨다.

- λ =-1 일 때, $TF^{(\lambda)} = \frac{1}{tf}$
- λ =0 일 때, $TF^{(\lambda)} = \ln tf$ (정의)
- λ =0.5 일 때, $TF^{(\lambda)} = \sqrt{tf}$
- λ =1 일 때, $TF^{(\lambda)} = tf$
- λ =2 일 때, $TF^{(\lambda)} = tf^2$

정규화 가중치 계산은 λ =0 일 경우는 로그 TF, λ =0.5 인 경우는 루트 TF, λ =1 인 경우에는 단순 TF 와 같은 형태를 보인다. 즉, 본 방법은 기존의 여러 가중치 계산 방법들을 포함한 TF 의 멱승변환 중에서, 현재 자신이 가지고 있는 용어 빈도자료를 가장 정규화 시켜주는 가중치를 찾는 방법이다.

3.3 역문헌빈도 및 역카테고리빈도

일반적으로 많이 사용하는 가중치 계산방법은 색인어의 빈도와 역문헌빈도를 이용하는, 즉 TF*IDF 이다. 역문헌빈도(IDF)란 적은 수의 문서에 나타난 색인어에 대해 높은 가중치를 주는 것으로, 색인어 Wi 의 역문헌빈도는 다음과 같이 계산된다[5].

$$IDF = \log N - \log DF_j + 1$$

여기서 N 은 총 문서의 개수, 그리고 DF_j는 색인어 W_i를 포함하는 문서의 개수이다.

IDF 는 문서간의 분리도가 높은 단어에 높은 가중치를 주는 특징이 있다. 하지만, 문서 분류일 경우 문서간 분리도 대신 카테고리간 분리도가 높은 단어에 높은 가중치를 주는 역카테고리빈도가 제안되었다[5]. 색인어 W_i 의 역카테고리빈도(ICF)는 다음 같이 계산되며,

$$ICF = \log M - \log CF_j + 1$$

여기서 M 은 총 카테고리의 개수, CF_i는 색인어 W_i를 포함하는 카테고리 개수이다.

ICF 는 IDF 와 기본원리는 비슷하나, 분류를 위해 카테고리 분리 능력이 우수한 색인어에 높은 가중치를 준다. 분류의 경우 카테고리간 구분에 도움이 되는 색인어가 중요도가 높으므로 ICF 는 IDF 보다 의미 있는 계산방법이 된다. 이를

아래 <표 2>는 TF, TF*IDF, TF*ICF 를 각각 비교 실험해 본 결과이다.

<표 2> IDF 와 ICF 를 적용했을 경우

	TF		TF * IDF		TF * ICF	
	train	test	train	test	train	test
단순	88.91	78.25	88.91	78.60	88.91	79.10
이진	91.80	81.50	91.80	81.55	91.80	81.95
로그	91.58	82.75	91.58	82.85	91.58	83.30
더블로그	92.60	83.05	92.60	83.00	91.61	83.55
루트	91.90	81.80	91.90	81.50	91.9	82.20
보정	88.91	78.25	88.91	78.60	88.91	79.10
Okapi	92.30	83.00	93.00	82.80	92.30	83.50
더블 2	91.23	81.40	91.23	81.25	91.23	81.80
루트직선	88.91	78.25	88.91	78.60	88.91	79.10
정규화	92.53	83.00	92.53	83.10	92.53	83.50

<표 2>와 같이, TF < TF*IDF < TF*ICF 의 결과이지만, 그 차이는 미미하였다. 이는 본 실험에서 사용한 카테고리 수가 적고, 게다가 카테고리 구조가 평면구조였기 때문으로 보인다. 이에 본 논문에서는 TF 값만을 이용하여 실험하였다.

IV. 실험 및 결론

4.1 실험환경 및 결과

문서 분류의 실험 대상 문서 집합으로는 동아일보 신문기사에서 8 개의 그룹으로 이루어진 8000 건의 기사를 이용하였다. 그리고 이를 4 개 그룹으로 나눠, 각각을 75%는 실험 집단으로 25%는 검증집단으로 이용하였다.

<표 3> Group1, Group2 분류실험 정확도 결과

용어 가중치 계산방법	Group1		Group2	
	train	test	train	test
단순 TF	88.9	78.2	89.2	78.3
이진 TF	91.8	81.5	91.6	81.5
로그 TF	92.6	82.8	92.9	82.7
더블로그 TF	92.3	83.1	92.3	83.0
루트 TF	91.9	81.8	92.1	83.1
보정 TF	89.0	78.3	89.2	78.3
Okapi TF	92.3	83.0	92.0	82.2
더블로그 2 TF	91.2	81.4	91.5	80.8
루트직선 TF	89.0	78.3	89.2	78.3
정규화 TF	92.5	83.0	92.9	83.2

<표 4> Group3, Group4 분류실험 정확도 결과

용어 가중치 계산방법	Group3		Group4	
	train	test	train	test
단순 TF	88.2	78.4	86.7	78.1
이진 TF	91.7	81.5	90.7	82.3
로그 TF	91.8	82.4	90.7	81.3
더블로그 TF	92.0	82.1	92.0	82.6
루트 TF	92.7	83.4	90.6	81.1
보정 TF	88.2	78.4	87.8	77.4
Okapi TF	92.9	83.2	91.7	81.4
더블로그 2 TF	91.3	80.4	89.5	78.9
루트직선 TF	88.0	78.1	87.8	77.1
정규화 TF	92.8	83.4	91.7	82.5

위 실험결과를 보면 정규화 용어빈도 가중치 계산 방법이 모두 상위의 정확도 그룹에 속해있다. 이는 용어빈도가 정규분포일 때 분류 성능이 향상될 것이라는, 본 논문의 가정이 옳음을 입증하는 것이다. 또한 다른 가중치 계산방법은 실험 그룹에 따라 다른 순위를 보이지만, 제안한 방법은 항상 상위 그룹에 속해있다. 이는 용어빈도 데이터에 따라 다른 가중치부여방법을 가져야 하기 때문에, 고정된 가중치 방법보다 λ 에 따라 변환식이 달라지는 정규화 가중치 계산방법이 좀더 일반적인 가중치 계산방법이 될 수 있다는 것을 의미한다.

4.2 결론

본 논문에서는 문서의 자동 분류 시 사용하는 색인어에 대한 새로운 가중치 계산 방법으로 정규화 가중치 계산방법을 제안하였다. 본 방법은 중간빈도 용어들이 키워드로서 더 적당하다는 언어적 직관에 근거하여, 용어빈도의 분포를 정규분포로 변환하여 적용하는 방법이다.

실험에서 본 논문의 가중치 계산 방식은 다른 가중치 계산 방법들에 비해 항상 좋은 결과를 보였다. 이것으로 보아 정규화 가중치 계산 방법은 언어적 직관에 비교적 잘 들어맞는 모델로 볼 수 있을 것이다.

참고문헌

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

[2] G. Salton. And C. Buckly, "Term Weighting Approaches in Automatic Text Retrieval," Information Processing and Management, 1998

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

[4] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBMJRD 2(2), 1958

[5] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류," 정보과학회 학술발표논문집, 4, 1997

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

[7] 정영미, 정보검색론, 구미무역 출판부, 1993

[8] 강승식, 한국어 형태소 분석과 정보검색, 홍릉과학 출판사, 2002