제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

단어 빈도 가중치를 이용한 자동 문서 분류

노현아*, 김민수**, 김수형*, 박혁로*
*전남대학교 전산학과
**전남대학교 통계학과
e-mail : hyuna@dal.chonnam.ac.kr

Automatic Document Classification Based on Word Frequency Weight

Hyun-A Noh*, Min-Soo Kim**, Soo Hyung Kim*, Hyuk-Ro Park*
*Dept. of Computer Science, Chonnam National University
** Dept. of Statistics, Chonnam National University

요 약

본 논문에서는 범주 내의 키워드 빈도에 의해 문서를 자동으로 분류하는 방법을 제안한다. 문서 자동분류 시스템에서는 문서와 문서를 비교하기 위해서 분류 자질(feature)에 적절한 가중치를 부여할 필요가 있다. 본 논문에서는 수작업으로 분류된 신문기사를 이용하여 자질의 가중치를 학습하는 방법을 사용하였다. 기존의 용어가중치 방법은 각 범주별로 가장 많이 등장한 명사부터 순서대로 추출하여 가중치를 주는 방법을 사용한 것에 비해 본 논문에서는 명사의 출현 횟수뿐만 아니라 출현위치를 함께 고려하여 가중치를 계산하는 방법을 제안한다. 또한 단어 빈도 가중치 방법의 변형된 방식을 사용함으로써 기존의 단어 빈도 가중치 방법과 비교하여 분류 정확도 측면에서 9%이상 성능 향상을 있음을 보인다.

1. 서 론

온라인상에서 텍스트 문서가 급격하게 증가하여 인간의 수작업으로 모든 문서를 처리할 수 없게 된 현 상황에서 자동 문서분류(Automatic Document Classification) 혹은 텍스트 범주화(Text Categorization)의 중요성은 점차 증대되고 있다.

자동 문서 분류 시스템은 입력된 문서의 내용을 분석하여 미리 정의되어 있는 범주들 중 가장 적합한 범주를 컴퓨터가 자동으로 할당하는 작업으로서, 효율적인 정보검색 및 관리를 가능하게 하는 동시에 전통적으로 문서 분류를 위해 요구되어 왔던 방대한 양의 수작업을 감소시킬 수 있다는 점에서 그 의미가 있다. 또한 최근에는 문서 필터링, 전자 우편 범주화 등의 응용 분야에 확대 적용되고 있다.

자동 문서분류 작업에 있어서는 자질선정이나 자질추출, 자질생성 방법이 범주화 성능에 결정적인 영향을 준다.

자동 문서분류에는 문서 내에 나타나는 단어의 출현 횟수나 분포, 확률 등을 이용하는 통계적인 방법과 자연어 처리를 통하여 문서 내에 있는 문장의 의미(semantic)나 구문(syntactic)을 분석하는 의미 분석 방법이 있다[2,4]. 보다 정확한 문서의 분류를 위해서는 자연어 처리를 통하여 문서의 내용을 파악하는 것이 바람직 하나, 자연어 자체의 모호성 때문에 문장의 의미 분석이 매우 어려워 의미 분석 방법은 한정된 영역에서 사용하기에 적합하다. 한편 통계적 방법은 간단히 구현할 수 있다는 장점 외에도 학습이 가능하기 때문에 충분한 학습 데이터가 주어졌을 경우 의미 분석 방법에 버금가는 결과를 낼 수 있다. 따라서 최근 많은 연구들이 통계적인 방법을 사용하고 있다 [4,10,11].

통계적인 문서분류 방법은 이미 분류된 학습 문서 집합(training document set)을 이용하여 새로운 문서가 분류될 가능성이 가장 높은 범주를 찾아내는 방법이다.

본 논문에서도 이와 같이 프레임워크에 기초하여 수작업으로 분류된 신문기사 학습 문서 집합을 이용하여 입력된 문서를 자동 분류하였다. 이를 위하여

학습 문서 집합 중에서 입력 문서와 가장 유사한 문서들을 추출한 뒤, 추출된 문서들이 속하는 분류 체계 및 분류 체계에 대한 소속 가중치를 이용하여 새로운 문서의 분류 체계를 계산하였다.

2. 색인어 추출 및 가중치 계산

자동색인의 초기 연구가 룬 Luhn 은 "단어의 빈도는 그 단어의 중요성을 측정하는 유용한 수단이 될 수 있다. 그리고 문장 내 단어의 위치도 단어의 중요성을 결정하는 유용한 측정수단이 될 수 있다. 이 두 측정치를 조합하면 어떤 문장의 중요성(significance)을 알 수 있다"라고 주장하였다. 즉 어떤 문서를 대표하는 단어 혹은 문장을 추출하는데 있어, '빈도(frequency)'가 유용한 수단이 된다는 것과 어떤 문서의 내용을 살펴보지 않아도 단어들의 출현 횟수 만 안다면 해당 문서가 어떤 분야에 대해 다루고 있는지 충분히 추측할 수 있다는 것을 주장하였다 [2,13].

이러한 가설을 바탕으로 각 문서들은 형태소 분석을 통한 명사 추출 과정을 전처리 과정으로 사용하여 단어들의 집합(bag-of-words)으로 표현된다. 또한 각 단어들은 해당 문서 내에서의 중요도를 표시하는 가중치를 부여 받게 된다.

본 연구에서는 문서를 자동으로 분류하기 전, 먼저 대상 문서로부터 키워드를 추출하는데, 키워드 추출 후 대상 문서는 다음과 같이 (단어, 빈도수) 쌍의 리스트 형태로 표현된다.

$$\text{문서} = \{(\text{단어}_1,\ \text{빈도수}_1), (\text{단어}_2,\ \text{빈도수}_2), \cdots\cdots, (\text{단어}_n,\ \text{빈도수}_n)\}$$

각 범주들도 문서와 같이 (단어, 가중치) 쌍의 리스트로 표현된다. 범주를 표현하기 위해서는 범주를 대표할 수 있는 단어들을 선정해야 하는데(자질 생성), 본 연구에서는 해당 단어가 나타나는 문서빈도를 이용하여 단어를 선정하였다. 이렇게 생성된 자질(단어)들은 학습 문서 집합을 이용하여 가중치를 부여 받는다.

단어에 가중치(weight)를 부여하는 목적은 한 문서가 취급하고 있는 개념들의 주제적 요소로서의 중요도에 따라 색인어로서 상대적 가치를 표현하기 위함이다. 즉, 각각의 개념을 나타내는 용어마다 일정한 범위 내의 값(가중치)을 부여함으로써 동일한 색인어라도 각각의 문서에 따라 중요도가 다를 수 있음을 나타낸다.

단어 가중치 계산에 주로 반영하는 요소는 어휘 빈도수(TF : term frequency), 역문서 빈도수(IDF : inverse document frequency), 문서 또는 질의 길이에 대한 정규화(normalization) 세가지가 있다.

본 연구에서는 이재윤의 연구[1]에서와 같이 이들 중에서 단어의 출현빈도가 높을수록 그 단어가 문헌의 주제를 대표할 확률이 높다는 통계적 기법의 기본적인 가설에 근거하여 단어의 출현빈도만을 이용하여 가중치를 계산하였다. [표 1]은 [1]에서 제시한 단어빈도를 이용한 가중치 계산 공식이다.

이 름	공 식
단순	$TF = tf$
로그	$TF = 1 + \log(tf)$
더블로그	$TF = 1 + \log(1 + \log(tf))$
루트	$TF = \sqrt{tf}$
Okapi	$TF = \frac{tf}{2 + tf}$
더블로그 2	$TF = 1 + \log_2(1 + \log_2(tf))$
루트직선	$TF = \frac{tf + 3}{4}$

[표 1] 단어빈도 가중치 공식

[1]의 연구에서는 [표 1]의 다양한 단어빈도 가중치 계산 공식을 이용하여 성능 비교한 같이 사용한 공식에 따라 약간의 성능의 차이를 보였으며, 단순 방법이 가장 좋은 성능을 보였으며, 루트직선, 루트, 로그, 더블로그 2, 더블로그, Okapi 방법 순으로 분류 성능이 나타났다.

한편, 본 연구에서는 단순 TF 방법이 tf 가 1 인 단어에 너무 낮은 가중치를 부여하며, tf 가 높은 단어에 지나치게 의존적인 점을 고려하여 [1]이 제안한 공식인 로그 TF 를 수정한 단어빈도 가중치 계산 방법을 채택하였다. 본 논문의 실험 결과는 단순 tf 방법보다 log TF 를 약간 수정한 방법이 더 우수한 성능을 보였다.

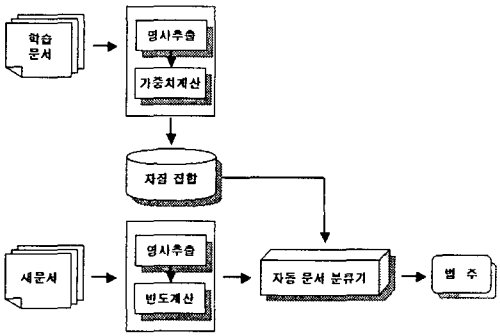
3. 문서 범주 할당

분류될 문서의 단어 추출(명사와 그 출현 빈도수 계산)을 통하여 주제어 후보들을 생성한 후 가중치 계산을 통하여 주제어 추출 작업이 끝나면, 입력 문서와 각 범주 사이의 소속 정도를 계산한다. 여러 범주들 중 입력 문서의 소속 정도가 가장 높은 범주가 해당 문서의 범주로 배정된다.

문서 분류의 정확성을 높이기 위해서는 문서들에서 각 문서를 대표할 수 있는 주요 단어들(keywords)을 추출하고 각 단어의 문서에서의 빈도수를 계산하여야 한다. 그러나, 모든 문서에서 골고루 나타나는 단어는 어떤 특정 문서 집합을 특별히 대표할 수 있는 단어로서는 부적절 하다고 판단할 수 있으므로, 단어의 빈도수 만으로는 주요 단어를 추출하기에는 문제점이 많다.

그러므로, 어떤 단어가 주요 단어인지를 평가하기 위해서는 전체 문서 집합에서 어떤 특정한 문서들에서 많이 등장한 단어가 그 문서 집합을 대표한다고 할 수 있다. 따라서 각 단어의 영향력은 단어의 빈도수 및 문서 전체집합에 대해 그 단어가 등장한 문서 집합의 상대비율을 계산할 수 있다. 문서의 특징이 될

수 있는 단어들을 추출하는 것은 분류 성능에 결정적인 영향을 주기 때문에 매우 중요하다.



[그림 1] 자동 문서 분류 모형화

자질집합을 생성되면, 각 자질에 대한 가중치를 계산하여야 한다. 본 논문에서는 기존의 연구와는 달리 각 자질(단어)의 출현빈도를 고려하여 가중치를 계산하는 방법을 제안하였다.

(방법 1) 단순 TF 를 이용한 일반적인 방법

단순 TF 에서 집단 i 의 j 번째 명사 x_j 에 대한 가중치를 아래와 같이 계산한다.

$$W(x_j \mid 집단=i) = \frac{\sum_{k=1}^{p} TF_{x_j}}{\sum_{k=1}^{p}\sum_{l=1}^{n_k} TF_{x_l}}$$

이때, n_k 는 k 번째 문서의 키워드 개수

p 는 i 번째 집단의 문서 개수이다.

(방법 2) 로그 TF 를 이용한 방법

$$W(x_j \mid 집단=i) = \frac{\sum_{k=1}^{p} (1+\ln TF_{x_j})}{\sum_{k=1}^{p}\sum_{l=1}^{n_k} (1+\ln TF_{x_l})}$$

(방법 3) 로그 TF 를 변형시킨 방법

본 연구에서는 방법 2 의 로그 TF 의 가중치를 약간 조정하여 아래와 같은 가중치를 제안한다.

$$W(x_j \mid 집단=i) = \frac{\ln \sum_{k=1}^{p} (TF_{x_j}+1)}{\sum_{k=1}^{p} \left[\ln \sum_{l=1}^{n_k} (TF_{x_l}+1) \right]}$$

그리고, x_j 를 가지고 있는 문서의 집단 = i 에 대한 점수는

$$\text{Score}(집단=i \mid x_j) = (1+\ln TF_{x_j}) \times W(x_j \mid 집단=i)$$

와 같이 계산된다.

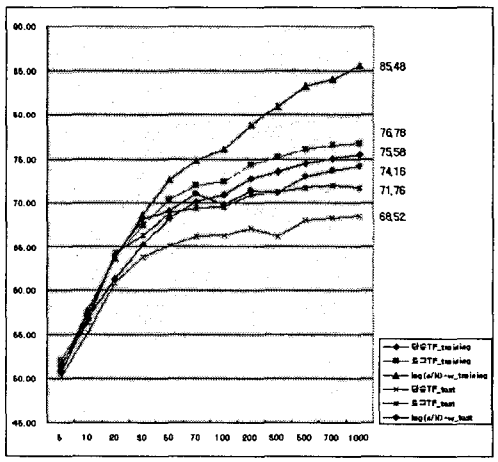
이 방법은 일반적인 로그 TF 와 비교하여 $\text{Score}(집단=i \mid x_j)$ 의 계산 시 가중치에 곱해지는 $1+\ln TF_{x_j}$ 는 같다. 하지만, 가중치 $W(x_j \mid 집단=i)$ 의 계산을 달리하고 있다.

$1+\ln TF_{x_j}$ 의 의미는 어떤 명사가 임의의 문서집단에서 전체적으로 똑같이 m 번 등장했어도 특정문서에서만 너무 많이 등장하고 다른 문서에서는 거의 나타나지 않았다면 가중치를 줄여주는 효과를 준다. 또한 본 연구에서 제안한 새로운 가중치는 일반적인 로그 TF 에 비해 1 번이라도 등장한 명사에 대해서는 가중치를 크게 해주는 효과가 있다.

4. 실험 및 결과 분석

본 논문에서는 동아일보 2001 년 기사를 수집하여 경제, 국제, 문화생활, 방송연예, 사람속으로, 사설칼럼, 사회, 스포츠, 정치, IT 등 10 개의 범주로 분류하는 실험을 하였다. 각 범주에 각각 1,000 개씩 10,000 개의 문서를 수집하여, 이 중 5,000 개 문서는 학습 집합으로 사용하고, 그리고 나머지 5,000 개 문서는 테스트 집합(Test Set)으로 사용하였다.

추출된 문서들을 먼저 본 연구실에서 개발한 명사 추출기를 사용하여 명사를 추출한 후, 추출된 명사들을 자질 후보로 사용하였다. 자질 추출방법을 통해 자질 후보들 중 각 분야별로 자주 나타나는 1,000 개의 키워드를 추출하였다.



[그림 2] 분류 정확도 실험 결과 그래프

추출된 키워드에 가중치를 부여하는 방법을 사용하여 본 논문에서 제안한 문서 분류 방법인 방법 3 과 방법 1, 방법 2 의 분류 성능 비교를 수행하였다. 분류

성능에 대한 비교를 위해 문서 표현을 위한 특징 값의 수를 결정하여야 하는데, 본 연구에서는 비교 실험에서 사용된 단어 수를 1,000 개로 고정하였다. 이는 실험 결과 1,000 개 단어를 사용하는 경우가 가장 좋은 성능을 보였기 때문이다.

[그림 2]는 테스트집합에 대한 분류 성능을 나타내는 그래프이다. 그래프에서 나타나듯이 단순 TF 보다는 로그 TF 를 취하는 방법이 더 나은 결과를 보였고, 또 로그 TF 보다는 로그 TF 를 변형시켜 가중치를 부여한 분류가 다른 분류 방법에 비해 상당한 성능 개선을 보임을 알 수가 있었다.

실험 결과 그래프에 의해 방법 3 의 분류 정확도를 평가하기 위해서 오분류행렬(Confusion Matrix)을 사용하여 전체 표시하였다. 오분류행렬(Confusion Matrix)은 항목별 분류 정확도 평가 방법으로 오분류들의 공간적 분포를 파악하고 그 결과의 신뢰도 및 적합도를 판단하는 근거를 가질 수 있는 지표가 된다.

[표 2] 학습문서 오분류행렬

	경제	IT	국제	문화생활	방송연예	사람속으로	사설칼럼	사회	스포츠	정치
경 제	448	27	3	1	0	5	1	8	2	5
I T	15	452	5	4	2	7	4	5	5	1
국 제	18	15	415	15	14	3	4	5	4	7
문화생활	0	19	10	393	35	20	5	12	5	1
방송연예	0	11	6	5	451	13	2	5	3	4
사람속으로	0	3	6	0	1	479	1	1	3	6
사설칼럼	26	2	19	17	13	5	343	17	9	49
사 회	13	22	4	9	3	30	15	382	4	18
스 포 츠	2	1	0	2	1	10	0	3	481	0
정 치	4	1	12	0	0	7	9	7	0	460

[표 3] 실험문서 오분류행렬

	경제	IT	국제	문화생활	방송연예	사람속으로	사설칼럼	사회	스포츠	정치
경 제	403	32	4	6	0	22	15	10	3	5
I T	23	408	10	22	7	16	5	4	4	1
국 제	25	20	339	21	20	9	22	18	7	19
문화생활	7	27	4	326	47	17	27	33	9	3
방송연예	2	15	13	46	402	6	7	4	3	2
사람속으로	5	1	8	12	9	421	13	19	1	11
사설칼럼	47	20	50	53	18	5	201	31	3	72
사 회	12	27	8	26	11	41	30	314	2	29
스 포 츠	2	5	3	9	4	11	4	3	459	0
정 치	10	3	15	2	5	13	42	17	0	393

위의 Confusion Matrix 를 보면, 로그 TF 를 변경시켜 가중치를 부여한 방법 3 을 이용한 결과 대부분의 분야가 아주 양호하게 분류됨을 알 수 있다. 그러나 사설칼럼 분야의 경우 가장 오분류율이 높은데 사설칼럼의 특성상 경제, 정치, 국제, 사회, 문화생활 등 여러 분야에 대한 칼럼들이 실려있기 때문에 단어 출현 빈도수를 이용한 통계적인 방법이 문서를 분류하는데 있어 크게 도움이 되지 않음을 알 수 있었다.

5. 결 론

본 논문에서는 단어 출현 빈도를 이용한 새로운 자동 문서 분류 방법을 제안하고 실험하였다. 제안된 방법의 특징은 먼저 분류 자질 생성에 있어서 단어의 출현 횟수 뿐만 아니라 문헌빈도를 함께 고려한다는 점을 들 수 있다. 또한 각 자질의 가중치 계산에 있어서도 기존의 로그 TF 방법을 수정한 새로운 방법을 제시하였다.

10,000 건의 신문기사 및 10 개의 범주를 이용하여 분류 실험을 수행한 결과 본 논문에서 제안한 방법이 기존의 단순 TF 방법보다 9% 성능 개선을 보여 본 논문에서 제시한 방법이 타당함을 알 수 있었다.

한편, 사설칼럼의 경우 다양한 주제를 다루고 있기 때문에 단어 출현 빈도수에 가중치를 부여한 통계적인 방법만을 이용해서는 한계가 있음을 알 수 있다. 이는 보다 더 정확한 실험을 위해서는 서로 겹치지 않고 명확하게 분리된 범주 집합 및 해당 문서 집합이 필요하다는 것을 보여주고 있다.

참고문헌

[1] 이재윤, 최보영, 정영미, "문헌 자동분류에서 용어 가중치 기법에 대한 연구", 제 7 회 한국정보관리학회 학술대회 논문집, pp.41-44, 2000.

[2] 정영미, 정보검색론, 구미무역 출판부, 1993.

[3] 김상범, 임해창, "범주간의 관계를 통한 자동 문서 범주화의 개선", 고려대학교 컴퓨터학과 석사학위논문, 1999.

[4] 한정기, 박민규, 조광제, 김준태, "구문 패턴과 키워드 집합을 이용한 통계적 자동 문서 분류의 성능 향상", 한국정보처리학회 논문지, 제 7 권 제 4 호, pp.1150-1159, 2000.

[5] 최성환, 정영미, "용어가중치 결합이 검색 효율성에 미치는 영향 연구", 한국정보과학회 학술발표 논문집, 제 29 권 제 1 호, pp.481-483, 2002.

[6] 허준회, 고수정, 김태용, 최준혁, 이정현, "문서의 주제어별 가중치와 말뭉치를 이용한 한국어 문서의 자동분류: 베이지안 분류자", 한국정보과학회 가을 학술발표 논문집, Vol.26, No.2, pp.154-156, 1999.

[7] G.. Salton. And C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, 24(5), pp.512-23, 1988.

[8] Y. Yang, J.O. Pederson, "A Comparative study on feature selection in text categorization", In Proceeding of the 24th International Conference on Machine Learning, 1997.

[9] Y. Yang, "An evaluation of statistical approaches to text categorization", Information Retrieval Vol 1, No. 1/2, pp. 69-90, 1999.

[10] L.Larkey. and W.Croft, "Combining classifiers in text categorization", SIGIR'96, 1996.

[11] D.Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task", SIGIR'92, 1992.

[12] T. M. Mitchell, "Machine Learning", McGraw-Hill companies, Inc., 1997.

[13] H.P.Luhn, "The Automatic Creation of Literature Abstracts," IBMJRD 2(2), pp.159-165, 1958.