자질의 범주 모호성 해소를 위한 Naive Bayes 분류기 설계

유현숙 정영미
연세대학교 문헌정보학과
cdyhs@lis.yonsei.ac.kr, ymchung@yonsei.ac.kr

A Naive Bayes Classifier for Category Disambiguation of Features

Yu Hyeon-Sook Chung Young-Mee
Dept. of Library and Information Science, Graduate School of Yonsei University

요 약

문서 범주화는 전자 정보환경에서 매우 유용한 정보처리 도구로서, 다양한 문서 범주화 기법 및 성능향상을 위한 연구들이 지속적으로 이루어지고 있다. 그러나, 대부분의 연구들은 문서 범주화의 대상이 되는 단어 자질 공간의 차원축소 문제에만 집중되었을 뿐, 학습단계에 큰 영향을 미치는 다범주 단어 자질의 범주 모호성은 고려하지 않았다. 본 연구에서는, 다범주 자질의 범주 모호성을 해소함으로써 문서 범주화의 성능향상을 유도하는 범주 모호성 해소 가중치 W를 제시하고 이를 실험을 통해 증명하였다. 실험에서는 Naive Bayes 분류기와 가중치 W를 적용한 Naive Bayes-W 분류기를 직접 구축하여 문서 범주화의 성능향상 여부를 비교하는데 사용하였다. 도출된 실험결과를 통해, 가중치 W는 현재의 분류기가 가지고 있는 자질 표현의 범주 모호성이라는 단점을 보완하고 분류기의 성능향상을 유도함으로써 정보검색시스템의 검색효율을 높이는 데 활용될 수 있음이 증명되었다.

1. 서론

문서 범주화에 적용되는 대부분의 학습 알고리즘은 문서를 표현하는 단어집합에 의존한다. 그러나, 기존의 문서 범주화 기법들은 학습집단 문서들을 단지 불용어를 제거한 단어들로 표현하여 주제범주를 할당할 뿐 학습단계에 큰 영향을 미치는 다범주 단어 자질의 범주 모호성(category ambiguity)은 크게 고려하지 않았다.

하지만, 문서를 표현하는 자질이 여러 범주에 출현하는 범주 모호성을 가지는 경우 분류기의 분류 정확도가 낮아지게 되고, 결국 이 문서는 부적합한 범주에 속하게 되어 정보검색시 잡음을 일으키는 원인이 될 수 있다.

그러므로, 자질이 속한 여러 범주들 중 어떤 범주가 자질로 표현된 문서에 할당되어야 하는가를 결정하는 범주 모호성 해소(category disambiguation) 작업이 문서 범주화에 적용되어야 한다. 또한 여기에는 자질과 문맥 및 범주에 대한 다양한 지식이 사용되어야 할 것이다.

자질의 범주 모호성을 해소하는 것은 문서 범주화 기법을 적용한 정보검색 분야 외에도, 의미태깅, 기계번역, 중의성 해소, 대화분석과 같은 여러 자연언어 처리 응용 프로그램에서 활용가능하다는 점에 의의가 있다.

이에 본 연구에서는 다범주 자질의 범주 모호성을 해소하는 가중치 W를 고안하고, 이를 분류기에 적용함으로써 분류기의 성능향상 여부를 실험으로 검증하였다.

2. 문서 범주화 실험

2.1. 실험설계

다범주 자질의 범주 모호성 해소 가중치를 적용한 문서 범주화 실험에는, 연세대학교 문헌정보학과에서 구축한 KFCM-CL 중 대분류 3범주와 각각의 대분류 아래에서의 중분류 3범주들로부터 선별된 문서 216건이 사용되었다. 학습집단과 검증집단의 비율은 3:1이 되도록 구분했으며, 자질들이 문서 범주화에 미치는 영향을 최대한으로 고려하기 위해 각 범주별로 문서 수는 동일하게 수집했다.

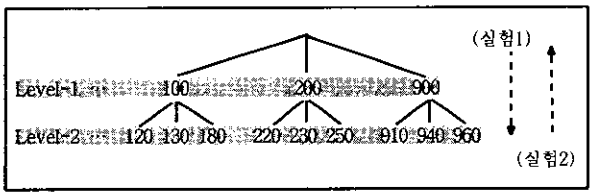
또한, 대분류 3범주는 Level-1, 중분류 9범주는 Level-2가 되어 계층적인 분류체계를 구성하도록 하였다. 검증집단의 문서 범주 할당은 계층적 분류체계를 따라, 하향식(top-down) 실험 1과 상향식(bottom-up) 실험 2 과정으로 나뉘어 수행되었다. 다음의 <표 1>은 본 실험의 문서집단 구성을, <그림 1>은 분류체계를 제시한 것이다.

실제 범주 할당 과정에서는 실험 1, 실험 2 각각에 대해 일반적인 Naive Bayes 분류기와 다범주 자질의 범주 모호성 해소 가중치를 부여한 Naive Bayes-W 분류기를 사용하여 범주를 할당하고 성능향상 여부를 관찰하였다.

범주 할당에 사용된 분류기들은 일반인에게 공개된 소프트웨어가 아니라, 본 연구의 실험을 위해 직접 프로그래밍된 것들이다.

대분류	중분류	범주별 실험문서 수	학습집단		검증집단
			문서 수	자질 총 출현빈도	문서 수
100	120	24	18	4328	6
	130	24	18	4357	6
	180	24	18	3132	6
100번주 소계		72	54	11817	18
200	220	24	18	3086	6
	230	24	18	3036	6
	250	24	18	3220	6
200번주 소계		72	54	9342	18
900	910	24	18	2565	6
	940	24	18	2725	6
	960	24	18	2314	6
900번주 소계		72	54	7604	18
합계		216	162	28763	54

<표 1> 실험문서집단 구성



<그림 1> 문서 범주화를 위한 문서집단의 분류체계

2.2. Naive Bayes 분류기

Naive Bayes 분류기는 범주가 출현하는 사전 확률을 기반으로 하여 특정 범주가 문서에 할당될 확률과, 특정 단어가 문서에서 발생할 조건확률을 계산한다. 분류하려는 문서 D_k에 단어자질 T_i가 출현한 경우, 문서가 범주 C_j에 분류될 확률은 다음과 같다.

$$P(C_j \mid D_k) = \arg\max P(C_j) \cdot \prod_i P(T_i \mid C_j)$$

이 때 $P(C_j \mid D_k)$은 문서 D_k가 범주 C_j에 할당될 확률이고, $P(C_j)$와 $P(T_i \mid C_j)$는 실험집단으로부터 다음과 같이 계산할 수 있다.

$$P(C_j) = \frac{C_j \text{에 할당된 학습문서 수}}{\text{모든 학습문서 수}}$$

$$P(T_i \mid C_j) = \frac{C_j \text{에 할당된 문서에서 } T_i \text{가 발생한 횟수}}{C_j \text{에 할당된 문서에 나타난 모든 단어의 발생 횟수}}$$

따라서, Naive Bayes 분류기는 검증문서가 범주에 할당될 확률을 각 범주별로 계산한 다음, 최대의 값을 가지는 범주에 최종적으로 문서를 할당한다.

2.3. Naive Bayes-*W* 분류기

단어의 출현빈도는 문서 범주화에서 매우 중요한 부분이지만, 자질들이 전체 문서집단 내에서 특정 범주에 치우친 정도를 고려하여 줌으로써 다범주 자질을 범주 모호성을 가진 단어가 아닌 범주 특정적인 단어로 만들어 주는 과정이 필요하다.

이처럼 범주 모호적인 자질이 각 범주에서 차지하는 영향을 파악하기 위해, 각 범주의 자질 T에 대해서 다음과 같은 범주 모호성 해소 가중치 W가 산출된다.

$$W = \frac{\text{범주 } C_j \text{ 내 } T_i \text{ 출현 횟수}}{\text{학습집단 내 } T_i \text{ 출현 횟수}} \times \frac{\text{전체 범주 수}}{T_i\text{가 출현한 범주 수}}$$

가중치 W의 왼쪽 항은 상대빈도에 의해 자질 T가 학습집단 전체에서의 빈도와 비교하여 어느 정도 치우쳐 그 범주의 자질로서 출현하는가를 나타내는 것이며, 이 값이 크면 클수록 그 범주에서 중요한 자질이 된다. 오른쪽 항은 자질의 범주 집중도를 반영하는 것으로, 한 범주에 치우친 자질들에 대해서는 높은 가중치를 부여하고, 여러 범주에 고르게 출현한 다범주 자질들의 가중치는 더 낮추는 것이다.

다범주 자질의 범주 모호성을 해소하는 가중치 W는 Naive Bayes 분류기의 $P(T_i \mid C_j)$에 곱해져 사용되었다.

3. 문서 범주화 실험결과 평가 및 분석

본 연구에서는 Naive Bayes 분류기 및 Naive Bayes-*W* 분류기의 문서 범주화 실험결과 평가를 위해 정확도(accuracy), 정확률(precision), 재현율(recall)을 주된 척도로 사용하고 필요시 부적합률(fallout)을 살펴보는 과정을 취하였다.

3.1 분류 순서에 따른 분류기 성능 비교

분류 순서를 기준으로 분류기의 문서 범주화 실험 결과를 비교해 보면, 하향식, 상향식 모두에서 Naive Bayes-*W* 분류기가 Naive Bayes 분류기 보다 높은 성능을 보였다.

<표 2>와 <표 3>은 각각 하향식, 상향식 분류를 수행한 실험 1의 결과로서, Level-1(대분류 수준)과 Level-2(중분류 수준) 모두에서의 Naive Bayes 분류기와 Naive Bayes-*W* 분류기의 성능 평가 및 성능향상 정도를 보여준다.

분류 수준 (범주 수)	Level-1 (3)			Level-2 (9)		
분류기	Naive Bayes	Naive Bayes-W	성능향상 정도	Naive Bayes	Naive Bayes-W	성능향상 정도
정확도	90.12	95.06	+4.94	92.59	96.30	+3.71
정확률	87.19	94.21	+7.02	92.39	96.30	+3.91
재현율	83.33	90.74	+7.41	87.04	94.44	+7.4

<표 2> 실험 1(하향식 분류)의 범주 할당 (단위: %)

분류 수준 (범주 수)	Level-2 (9)			Level-1 (3)		
분류기	Naive Bayes	Naive Bayes-W	성능향상 정도	Naive Bayes	Naive Bayes-W	성능향상 정도
정확도	92.59	94.24	+1.65	83.95	89.51	+5.56
정확률	79.82	75.29	-4.53	78.28	85.14	+6.86
재현율	66.67	74.07	+7.4	75.93	83.33	+7.4

<표 3> 실험 2(상향식 분류)의 범주 할당 (단위: %)

3.2 분류 수준에 따른 분류기 성능 비교

분류 수준을 기준으로 한 Naive Bayes 분류기 및 Naive Bayes-*W* 분류기의 문서 범주화 결과는 <표 4>와 <표 5>에 제시되어 있다.

아래의 결과에서 알 수 있듯이, 실험 1, 실험 2의 거의 모든 수준에서 Naive Bayes-*W* 분류기는 Naive Bayes 분류기보다 높은 분류 정확도 성능을 보이고 있다.

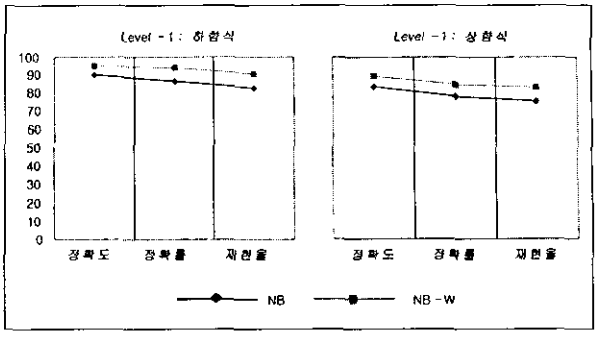
분류 순서 (방향)	실험 1 (하향식)			실험 2 (상향식)		
분류기	Naive Bayes	Naive Bayes-W	성능향상 정도	Naive Bayes	Naive Bayes-W	성능향상 정도
정확도	90.12	95.06	+4.94	83.95	89.51	+5.56
정확률	87.19	94.21	+7.02	78.28	85.14	+6.86
재현율	83.33	90.74	+7.41	75.93	83.33	+7.4

<표 4> Naive Bayes 분류기와 Naive Bayes-*W* 분류기의 성능 비교 : Level-1(대분류 수준) (단위: %)

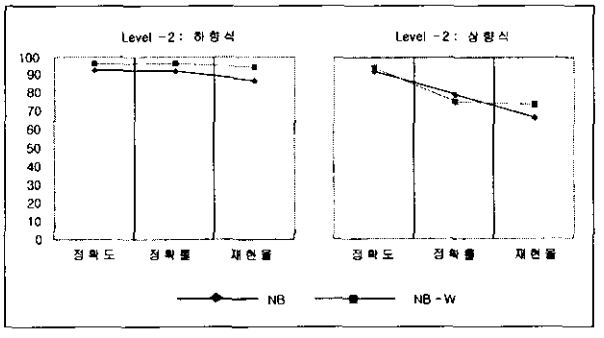
분류 순서 (방향)	실험 1 (하향식)			실험 2 (상향식)		
분류기	Naive Bayes	Naive Bayes-W	성능향상 정도	Naive Bayes	Naive Bayes-W	성능향상 정도
정확도	92.59	96.30	+3.71	92.59	94.24	+1.65
정확률	92.39	96.30	+3.91	79.82	75.29	-4.53
재현율	87.04	94.44	+7.4	66.67	74.07	+7.4

<표 5> Naive Bayes 분류기와 Naive Bayes-*W* 분류기의 성능 비교 : Level-2(중분류 수준) (단위: %)

<그림 2>과 <그림 3>은 분류 순서 및 분류 수준에 따른 Naive Bayes 분류기와 Naive Bayes-*W* 분류기의 성능 평가 결과를 종합하여 그래프로 표현한 것이다.



<그림 2> Level-1(대분류 수준)에서의 Naive Bayes 분류기와 Naive Bayes-*W* 분류기의 성능 비교



<그림 3> Level-2(중분류 수준)에서의 Naive Bayes 분류기와 Naive Bayes-*W* 분류기의 성능 비교

4. 결론 및 제언

본 연구에서는 문서 범주화의 성능향상을 위한 방법으로서 다범주 자질의 범주 모호성을 해소하는 가중치 *W*를 제시하고, 이를 Naive Bayes 분류기에 적용하여 비교 검증하는 실험을 수행하였다.

실험결과, 가중치 *W*를 적용한 Naive Bayes-*W* 분류기가 거의 모든 경우에서 Naive Bayes 분류기보다 월등한 성능을 보임으로써 다범주 자질의 범주 모호성 해소 가중치가 Naive Bayes 분류기의 성능향상을 유도한 것으로 증명되었다.

예외적으로, 실험 2의 Level-2 범주 할당시 Naive Bayes 분류기가 더 높은 정확률 결과를 보인 바 있지만, 이 경우에도, Naive Bayes 분류기의 부적합률은 4.17%인 반면 Naive Bayes-*W* 분류기는 3.24%로 Naive Bayes-*W* 분류 알고리즘이 부적합한 범주를 할당하는 정도가 더 낮은 것으로 드러났다. 이는 Level-2 수준의 범주 할당시에는 부적합한 범주가 할당되었다 할지라도, Level-1 수준에서는 적합한 범주가 할당됨으로써 Naive Bayes-*W* 분류기에 전체적인 범주 할당 성능향상 효과를 가져오기 때문이다.

그리고, 실험 1(하향식)과 실험 2(상향식)의 범주 할당 결과를 통해서는, 실험 1의 하향식 범주 체계, 즉 Level-1(대분류 수준)에서 Level-2(중분류 수준)로 세분하여 범주를 할당하는 과정이 보다 정확한 문서 범주화 결과를 낳는다는 사실을 알 수 있었다.

마지막으로, Naive Bayes-*W* 분류기가 지니는 일부 제한점을 극복하기 위해서는 자질 표현의 집단화 및 구체화, 새로운 범주와 자질의 추가 문제를 지속적으로 고려해야 할 필요성이 있다. 또한 다양한 매체로 표현된 대이터의 처리 및 다변하는 웹환경 반영 등의 문제가 Naive Bayes-*W* 분류기의 적용 분야 확대를 위한 향후 연구 과제가 될 수 있을 것이다.

5. 참고문헌

[1]. W.A. Gale, K.W. Church and D. Yarowcky. "A method for disambiguating word senses in a large corpus." Computers and the Humanities, 26: 415-439. 1992.

[2]. T.M. Mitchell, "Machine Learning", McGraw Hill. 1997.

[3]. C.D. Manning, and H. Schütze. "Foundations of Statistical Natural Language Processing." 1999.