잠재의미색인(LSI) 기법을 이용한 kNN 분류기의 자질 선정에 관한 연구

Evaluation of the Feature Selection Function of Latent Semantic Indexing(LSI) Using a kNN Classifier

박부영, 연세대학교 문헌정보학과, booyoung77@naver.com
정영미, 연세대학교 문헌정보학과, ymchung@yonsei.ac.kr

Boo-Young Park, Dept. of Library and Information Science, Yonsei University
Young-Mee Chung, Dept. of Library and Information Science, Yonsei University

텍스트 범주화에 관한 선행연구에서 자주 사용되면서 좋은 성능을 보인 자질 선정 기법은 문헌빈도와 카이제곱 통계량 등이다. 그러나 이들은 단어 자체가 갖고 있는 모호성은 제거하지 못한다는 단점이 있다. 본 연구에서는 kNN 분류기를 이용한 범주화 실험에서 단어간의 상호 관련성이 자동적으로 유도됨으로써 단어 자체 보다는 단어의 개념을 분석하는 잠재의미색인 기법을 자질 선정 방법으로 제안한다.

1. 서론

텍스트 범주화(text categorization)는 문헌의 내용을 바탕으로 미리 정의된 범주를 문헌에 부여함으로써 문헌을 자동으로 분류하는 기법이다. 텍스트 범주화에 적용되는 대부분의 학습 알고리즘 과정은 문헌을 표현하는 단어 집합(bag-of-words)에 의존한다. 학습문헌에 출현하는 단어는 수만에서 수십만에 이르기 때문에 모든 단어가 자질로 추출되기 어렵다. 그러므로 텍스트 범주화 성능에 영향을 주지 않는 범위 내에서 여러 가지 자질 선정 기준을 이용하여 학습문헌에 나타나는 단어들의 정보량을 계산하고, 정보량이 큰 단어들만을 자질로 선정하여 사용하게 된다. 그러나 텍스트에 출현한 단어에 기반한 자질 선정 방법들 즉, 문헌빈도, 카이제곱 통계량, 상호정보량 등은 각 자질이 갖고 있는 모호성은 제거하지 못한다는 단점이 지적되어 왔다.

본 연구에서는 텍스트 범주화 과정에서 중요한 단계인 자질 선정 방법에 있어 잠재의미색인 기법을 적용함으로써 자질들의 지역적 문맥(local context) 정보를 이용해 자질들간의 전역적(global) 의미관계를 찾을 수 있다고 가정하였다. 따라서 잠재의미색인을 이용한 자질 선정을 통해 kNN 분류기의 성능을 향상시키는 방법을 제안하였고 또한 문헌 분류 실험을 통해 잠재의미색인, 문헌빈도, 카이제곱 통계량의 자질 선정 성능을 비교·평가하였다.

2 이론적 배경

2.1 문헌빈도(Document Frequency)

문헌빈도는 문헌집단에서 특정 단어가 출현한 문헌의 수를 의미하며 일정 빈도 이상의 문헌에 출현한 단어들을 자질로 추출하는 기법을 말한다. 이 기법의 기본 가정은 저빈도어는 문헌의 범주 예측에 거의 영향을 미치지 않으므로, 저빈도어를 제거하여 자질의 차원을 줄이는 동시에 범주화의 정확률을 높일 수 있다는 것이다.

2.2 카이제곱 통계량($\hat{A}^2$Statistic)

카이제곱 통계량은 단어 t와 범주 c 사이의 의존성을 측정하는 것으로 다음과 같이 계산된다.

$$\hat{A}^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

A는 범주 c에 속해 있는 문헌 중에 단어 t를 포함하고 있는 문헌의 수이고, B는 범주 c 외의 범주에 속해 있는 문헌 중에 단어 t를 포함하고 있는 문헌의 수이다. C는 범주 c에 속해 있는 문헌 중에 단어 t를 포함하지 않는 문헌의 수이며, D는 범주 c외의 범주에 속해 있는 문헌 중에 단어 t를 가지고 있지 않은 문헌의 수이다. 그리고 N은 전체 학습문헌의 수이다.

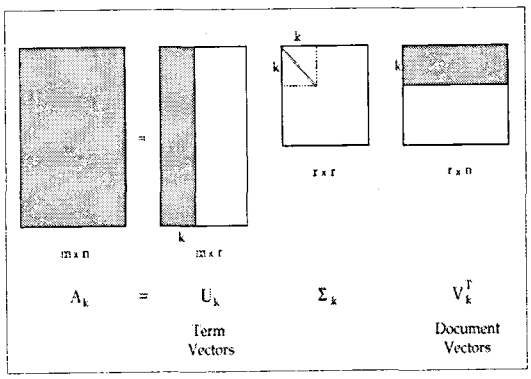
각 범주에 대해서 단어의 카이제곱 통계량을 계산한 후에 전체 학습문헌집단에서의 카이제곱 통계량을 계산하기 위해서 평균 카이제곱 통계량 $\hat{A}^2_{avg}(t)$이나 최대 카이제곱 통계량 $\hat{A}^2_{max}(t)$ 중에서 하나를 선택하여 사용한다.

$$\hat{A}^2_{avg}(t) = \sum_{i=1}^{m} Pr(c_i)\hat{A}^2(t,c_i)$$

$$\hat{A}^2_{max}(t) = MAX\hat{A}^2(t,c_i)$$

2.3 잠재의미색인(Latent Semantic Indexing)

잠재의미색인은 용어간의 상호 관련성이 자동적으로 유도됨으로써 색인어 자체 보다 개념을 분석한다. 특히 동의어나 다의어가 포함된 문헌에서 그 내용은 서술된 색인어 자체보다는 그 안에 내포된 개념에 더 관련되어 있으므로 색인어 대신에 개념에 기반해야 한다는 것이 잠재의미색인의 기본 개념이다.



U_k : orthogonal, unit-length columns

V_k^T : orthogonal, unit-length columns

Σ_k : 고유치들의 diagonal 행렬

m : 행렬 A_k의 행 수 (용어 수)

n : 행렬 A_k의 열 수 (문헌 수)

r : 행렬 A_k의 rank ($\leq \min(t,d)$)

[그림 36] SVD의 표현

잠재의미색인은 문헌 전체에 사용되고 있는 단어의 패턴에는 잠재하는 의미 구조가 있다는 가정 하에 이 구조를 파악하기 위해 SVD (Singular Vector Decomposition) 방법을 사용한다.

[그림 1]과 같이, 행렬 A를 m × n ($n \ll m$, $k \ll \min(m,n)$)이고 rank(A)=r인 행렬이라 한다면, A에 대한 SVD는 다음과 같이 정의된다.

$$A = U\Sigma V^T$$

원 데이터 행렬 A를 SVD로 분해 한 후 형

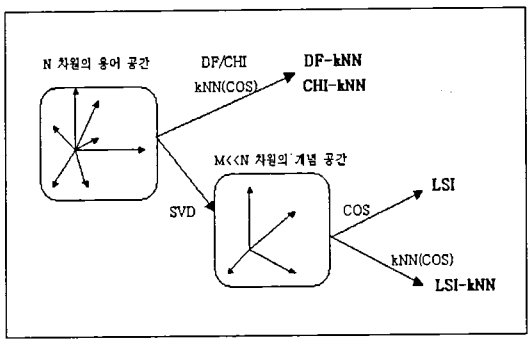
성된 유사 행렬 A_k와 완전히 동일한 분해 $U_k\Sigma_k V_k^T$ 를 얻기 위해서는 각 행렬들이 모두 최대의 랭크(full rank)를 가져야 한다. 최대의 랭크를 모두 사용하지 않고 그보다 적은 k (k<m)개 만큼의 벡터만을 사용한다면 그것은 원래 행렬 A_k에 대한 근사치가 된다.

3 kNN 분류기를 이용한 텍스트 범주화 실험

3.1 실험의 설계

잠재의미색인 기법을 이용한 kNN 분류기 실험은 KTSET 1.0을 실험문헌집단으로 사용하였고, 학습문헌과 검증문헌의 비율을 7:3으로 하여 실험하였다. KTSET 1.0은 기타 문헌집단과는 달리 복수분류범주를 할당하고 있기 때문에 단일분류범주의 경우와 복수분류범주 모두를 사용한 경우로 나누어 예비 실험을 실행하였고, 그 결과 복수분류범주를 할당하였다. 또한 kNN 분류기의 적용에 있어 최적의 k 값 선정을 위한 실험을 실시한 결과 k=5를 선택하였다.

범주화 성능 평가 방법으로는 평균정확률과 평균재현율, F_1 척도를 사용하였다. 본 연구의 실험 개요는 [그림 2]와 같다.



[그림 2] 실험의 개요

3.2 실험결과 분석

잠재의미색인 기법에 의한 LSI-kNN 분류기의 성능을 문헌빈도 기법을 적용한 DF-kNN 분류기 및 카이제곱 통계량 기법을 적용한 CHI-kNN 분류기와 비교한 결과는 다음과 같다.

LSI-kNN 분류기는 평균정확률에서 DF-kNN 분류기의 0.491, CHI-kNN 분류기의 0.392보다 각각 5.4%, 24.5% 높은 성능을 보여 주었다. F_1 값에서는 DF-kNN 분류기의 성능이 가장 좋았고 그 다음으로 CHI-kNN 분류기의 성능이 좋았으며 LSI-kNN 분류기가 제일 낮은 성능을 보였다. 그러나 LSI-kNN 분류기가 DF-kNN 분류기나 CHI-kNN 분류기와 다른 점은 문헌빈도나 카이제곱 통계량을 자질 선정 방법으로 이용한 경우는 4개의 범주 C, D, H, I에만 적합문헌을 분류하고 나머지 범주에는 적합문헌을 분류하지 못했으나, LSI-kNN 분류기는 차원의 값(k)이나 용어가중치에 상관 없이 이들이 적합하게 분류하지 못했던 B, G, Z 범주에 문헌들을 할당했다는 점이다. 즉, 단순 tf 가중치를 부여한 LSI-kNN 분류기에서는 차원의 값에 상관 없이 모두 Z 범주에 적합문헌을 할당하였고, $tf \cdot idf$ 가중치를 부여한 LSI-kNN 분류기에서는 차원의 값이 150인 150LSI-kNN 분류기의 경우는 B 범주에, 차원의 값이 50, 250인 50LSI-kNN 분류기와 250LSI-kNN 분류기의 경우는 G 범주에 적합문헌을 할당하였다.

4 결론

자질 선정 방법에 있어 $tf \cdot idf$ 가중치를 부여한 150LSI-kNN 분류기가 DF-kNN 분류기

[표 1] 자질 선정 기법에 따른 kNN 분류기의 범주별 성능 비교

평가	가중치	범주 / 분류기	B	C	D	G	H	I	X	평균
평균 재현율	tf · itf	DF-kNN	0.000	0.146	0.389	0.000	0.217	0.247	0.000	0.250
		CHI-kNN	0.000	**0.195**	0.362	0.000	0.167	0.276	0.000	0.250
		50LSI-kNN	0.000	0.063	**0.471**	**0.005**	**0.225**	0.236	0.000	0.200
		150LSI-kNN	**0.008**	0.100	0.329	0.000	0.124	0.438	0.000	0.200
		250LSI-kNN	0.000	0.100	0.441	**0.005**	0.205	0.250	0.000	0.200
	단순 tf	50LSI-kNN	0.000	0.130	0.301	0.000	0.161	0.420	**0.016**	0.206
		150LSI-kNN	0.000	0.102	0.249	0.000	0.218	0.426	0.005	0.200
		250LSI-kNN	0.000	0.161	0.217	0.000	0.167	**0.444**	0.011	0.140
평균 정확률	tf · itf	DF-kNN	0.000	0.446	0.550	0.000	0.452	0.515	0.000	0.491
		CHI-kNN	0.000	0.425	0.434	0.000	0.341	0.369	0.000	0.392
		50LSI-kNN	0.000	0.400	0.529	**0.200**	0.430	0.450	0.000	0.402
		150LSI-kNN	**0.400**	**0.455**	**0.586**	0.000	**0.564**	0.589	0.000	0.519
		250LSI-kNN	0.000	0.440	0.497	**0.200**	0.450	0.500	0.000	0.417
	단순 tf	50LSI-kNN	0.000	**0.455**	0.580	0.000	0.517	0.559	0.300	0.482
		150LSI-kNN	0.000	0.444	0.576	0.000	0.506	**0.600**	0.200	0.465
		250LSI-kNN	0.000	0.446	0.557	0.000	0.545	0.571	**0.400**	0.504
F_1 값	tf · itf	DF-kNN	0.000	0.220	0.455	0.000	0.293	0.334	0.000	0.325
		CHI-kNN	0.000	**0.267**	0.395	0.000	0.224	0.316	0.000	0.301
		50LSI-kNN	0.000	0.109	**0.498**	**0.010**	0.291	0.310	0.000	0.244
		150LSI-kNN	**0.016**	0.164	0.421	0.000	0.203	**0.502**	0.000	0.261
		250LSI-kNN	0.000	0.163	0.467	**0.010**	0.282	0.333	0.000	0.251
	단순 tf	50LSI-kNN	0.000	0.202	0.396	0.000	0.246	0.480	**0.028**	0.270
		150LSI-kNN	0.000	0.166	0.347	0.000	**0.305**	0.498	0.011	0.265
		250LSI-kNN	0.000	0.237	0.312	0.000	0.256	0.500	0.021	0.265

와 CHI-kNN 분류기 보다 평균정확률에서 더 좋은 성능을 보였으나 평균재현율이 낮게 나왔기 때문에 결과적으로 단일가 척도인 F_1 값에서는 DF-kNN 분류기와 CHI-kNN 분류기의 성능보다 낮았다. 그러나 LSI-kNN 분류기의 범주별 재현율이 낮음에도 불구하고 높은 정확률을 보이고 있어 잠재의미색인이 정확률 향상에 매우 효과적일 수 있다. 또한 문헌빈도나 카이제곱 통계량을 자질선정 기준으로 사용하였을 때에는 적합문헌을 하나도 분류하지 못했던 범주 B, G, Z에 문헌을 적합하게 할당함으로써 잠재의미색인을 이용한 자질 선정이 자질들 간의 의미를 분석함과 동시에 성능의 향상도 기대할 수 있음을 알 수 있다.

참고문헌

Berry, M. W., Dumais, S. T., and Letsche, T. A. 1995. "Computation methods for intelligent information access". [online]. [cited 2004.06.29] <http://www.cs.utk.edu/~berry/sc95/sc95.html#tinterp>

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. A., 1990. "Indexing by latent sementic analysis". *JASIS*, 41(6):391-407.

Yang, Y., and Pedersen, Jan O. 1997. "A comparative study on feature selection in text categorization". *ICML*:412-420.